To appear in Proceedings of the 20th USENIX Security Symposium (USENIX Security'11).
San Francisco, CA. August 2011.

Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade

Nektarios Leontiadis
Carnegie Mellon University

Tyler Moore
Harvard University

Nicolas Christin
Carnegie Mellon University

Abstract

We investigate the manipulation of web search results to promote the unauthorized sale of prescription drugs. We focus on *search-redirection attacks*, where miscreants compromise high-ranking websites and dynamically redirect traffic to different pharmacies based upon the particular search terms issued by the consumer. We constructed a representative list of 218 drug-related queries and automatically gathered the search results on a daily basis over nine months in 2010-2011. We find that about one third of all search results are one of over 7 000 infected hosts triggered to redirect to a few hundred pharmacy websites. Legitimate pharmacies and health resources have been largely crowded out by search-redirection attacks and blog spam. Infections persist longest on websites with high PageRank and from `.edu` domains. 96% of infected domains are connected through traffic redirection chains, and network analysis reveals that a few concentrated communities link many otherwise disparate pharmacies together. We calculate that the conversion rate of web searches into sales lies between 0.3% and 3%, and that more illegal drugs sales are facilitated by search-redirection attacks than by email spam. Finally, we observe that concentration in both the source infections and redirectors presents an opportunity for defenders to disrupt online pharmacy sales.

1 Introduction and background

Prescription drugs sold illicitly on the Internet arguably constitute the most dangerous online criminal activity. While resale of counterfeit luxury goods or software are obvious frauds, counterfeit medicines actually endanger public safety. Independent testing has indeed revealed that the drugs often include the active ingredient, but in incorrect and potentially dangerous dosages [48].

In the wake of the death of a teenager, the US Congress passed in 2008 the Ryan Haight Online Pharmacy Consumer Protection Act, rendering it illegal under federal law to "deliver, distribute, or dispense a controlled substance by means of the Internet" without an authorized prescription, or "to aid and abet such activity" [35]. Yet, illicit sales have continued to thrive in the nearly two years since the law has taken effect. In response, the White House has recently helped form a group of registrars, technology companies and payment processors to counter the proliferation of illicit online pharmacies [19].

Suspicious online retail operations have, for a long time, primarily resorted to email spam to advertise their products. However, the low conversion rates (realized sales over emails sent) associated with email spam [22] has led miscreants to adopt new tactics. Search-engine manipulation [47], in particular, has become widely used to advertise products. The basic idea of search-engine manipulation is to inflate the position at which a specific retailer's site appears in search results by artificially linking it from many websites. Conversion rates are believed to be much higher than for spam, since the advertised site has at least a degree of relevance to the query issued.

In this paper, we focus on a particularly pernicious variant of search-engine manipulation involving compromised web servers, which we term *search-redirection attacks*. Analyzing measurements collected over a nine-month interval, we show that search-redirection attacks are fast becoming the search engine manipulation technique of choice for online miscreants.

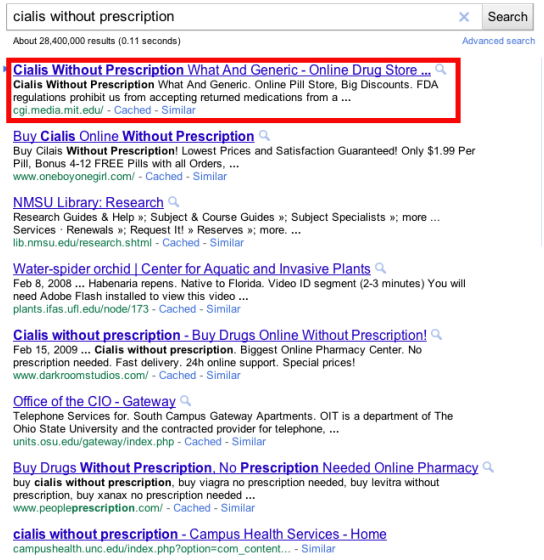


Figure 1: Example of the search-redirection attack. Only two of the results actually belong to online pharmacies. The rest are unrelated `.com` or `.edu` sites that had been compromised to redirect to online pharmacies, or have been populated with spam. The top search result (framed) was still infected at the time of this writing.

1.1 Search-redirection attacks

Figure 1 illustrates the attack. In response to the query "*cialis without prescription*", the top eight results include five `.edu` sites, one `.com` site with a seemingly unre-

lated domain name, and two online pharmacies. At first glance, the `.edu` and one of the `.com` sites have absolutely nothing to do with the sale of prescription drugs. However, clicking on some of these links, including the top search result framed in Figure 1, takes the visitor not to the requested site, but to an online pharmacy store.

The attack works as follows. The attacker first identifies high-visibility websites that are also vulnerable to code injection attacks.[1] Popular targets include outdated versions of WordPress [49], phpBB [38], or any other vulnerable blogging or wiki software. The code injected on the server intercepts all incoming HTTP requests to the compromised page and responds differently depending on the type of request.

Requests originating from search-engine crawlers, as identified by the *User-Agent* parameter of the HTTP request, return a mix of the compromised site's original content plus numerous links to websites promoted by the attacker (e.g., other compromised sites, online stores). This technique, "link stuffing," has been observed for several years [34] in non-compromised websites.

Requests originating from pages of search results, for queries deemed relevant to what the attacker wants to promote, are redirected to a website of the attacker's choosing. The compromised web server automatically identifies these requests based on the *Referrer* field that HTTP requests carry [14]. The *Referrer* actually contains the complete resource identifier (URI) that triggered the request. For instance, in Figure 1, when clicking on any of the links, the *Referrer* field is set to `http://www.google.com/search?q=cialis+without+prescription`. Upon detecting the pharmacy-related query, the server sends an HTTP redirect with status code 302 (Found) [14], along with a *location* field containing the desired pharmacy website or intermediary. The upshot is that the end user unknowingly visits a series of websites culminating in a fake pharmacy without ever spending time at the original site appearing in the search results. A similar technique has been extensively used to distribute malware [40], while web spammers have also used the technique to hide the true nature of their sites from investigators [33].

All other requests, including typing the URI directly into a browser, return the original content of the website. Therefore, website operators cannot readily discern that their website has been compromised. As we will show in Section 4, as a result of this "cloaking" mechanism, some of the victim sites remain infected for a long time.

While each of the components (link stuffing, redirection chains) of the search-redirection attack has been previously observed, to our knowledge, no study has investigated the combined attack itself, its effect on search results, or the potential harm it inflicts.

Three classes of websites are involved in search-redirection attacks. **Source infections** are innocent websites that have been compromised and reprogrammed with the behavior just described; **redirectors** are intermediary websites that receive traffic from source infections; and retailers (here, **pharmacies**) are destination websites that receive traffic from redirectors.

It is not immediately obvious who the victim is in search-redirection attacks. Unlike in drive-by-downloads [40], end users issuing pharmacy searches are not necessarily victims, since they are actually often seeking to illegally procure drugs online. In fact, here, search engines do provide results relevant to what users are looking for, regardless of the legality of the products considered. However, users may also become victims if they receive inaccurately dosed medicine or dangerous combinations that can cause physical harm or death. The operators of source infections are victims, but only marginally so, since they are not directly harmed by redirecting traffic to pharmacies. Pharmaceutical companies are victims in that they may lose out on legitimate sales. The greatest harm is a societal one, because laws designed to protect consumers are being openly flouted.

1.2 Summary of our contributions

Our study contributes to the understanding of online crime and search engine manipulation in several ways.

First, we collected search results over a nine-month interval (April 2010–February 2011). The data comprises daily returns from April 12, 2010–October 21, 2010, complemented by an additional 10 weeks of data from November 15th 2010–February 1st 2011. Combining both datasets, we gathered about 185 000 different universal resource identifiers (pharmacies, benign and compromised sites), of which around 63 000 were infected. We describe our measurement infrastructure and methodology in details in Section 2, and discuss the search results in Section 3.

Second, we show that a quarter of the top 10 search results actively redirect from compromised websites to online pharmacies at any given time. We show infected websites are very slowly remedied: the median infection lasts 46 days, and 16% of all websites have remained infected throughout the study. Further, websites with high reputation (e.g., high PageRank) remain infected and appear in the search results much longer than others.

Third, we provide concrete evidence of the existence of large, connected, advertising "affiliate" networks, funneling traffic to over 90% of the illicit online pharmacies we encountered. Search-redirection attacks play a key role in diverting traffic to questionable retail operations at the expense of legitimate alternatives.

Fourth, we analyze whether sites involved in the phar-

[1] We defer the study of the specific exploits to future work. Our focus in this paper is the outcome of the attack, not the attack itself.

maceutical trade are involved in other forms of suspicious retail activities, in other security attacks (e.g., serving malware-infested pages), or in spam email campaigns. While we find occasional evidence of other nefarious activities, many of the pharmacies we inspect appear to have moved away from email spam-based advertising. We discuss infection characteristics, affiliate networks, and relationship with other attacks in Section 4.

Fifth, we derive a rough estimate of the conversion rates achieved by search-redirection attacks, and show they are considerably higher than those observed for spam campaigns. We present this analysis in Section 5.

Sixth, we consider a range of mitigation strategies that could reduce the harm caused by search-redirection attacks in Section 6.

In addition to these contributions, we compare our study with related work in Section 7, before concluding in Section 8, where we also describe ongoing work tracking the promotion of other types of fraudulent goods.

2 Measurement methodology

We now explain the methodology used to identify search-redirection attacks that promote online pharmacies. We first describe the infrastructure for data collection, then how search queries are selected, and finally how the search results are classified.

2.1 Infrastructure overview

The measurement infrastructure comprises two distinct components: a search-engine agent that sends drug-related queries and a crawler that checks for behavior associated with search-redirection attacks.[2]

The search-engine agent uses the Google Web Search API [2] to automatically retrieve the top 64 search results to selected queries. From manually inspecting some compromised websites, we found that search-redirection attacks frequently also work on other search engines. Every 24 hours, the search-engine agent automatically sends 218 different queries for prescription drug-related terms (e.g., "*cialis without prescription*") and stores all 13 952 ($= 64 \times 218$) URIs returned. We explain how we selected the corpus of 218 queries in Section 2.2.

The crawler module then contacts each URI collected by the search-engine agent and checks for HTTP 302 redirects mentioned in Section 1.1. The crawler emulates typical web-search activity by setting the *User-Agent* and *Referrer* terms appropriately in the HTTP headers. Initial tests revealed that some source infections had been programmed to block repeated requests from a single IP address. Consequently, all crawler requests are tunneled through the Tor network [11] to circumvent the blocking.

[2] All results gathered by the crawler are stored in a *mySQL* database, available from `http://arima.ini.cmu.edu/rx.sql.gz`.

2.2 Query selection

Selecting appropriate queries to feed the search-engine agent is critical for obtaining suitable quality, coverage and representativeness in the results. We began by issuing a single seed query, "*no prescription vicodin*," chosen for the many source infections it returned at the time (March 3, 2010). We then browsed the top infected results posing as a search engine crawler. As described in Section 1.1, infected servers present different results to search-engine crawlers. The pages include a mixture of the site's original content and a number of drug-related search phrases designed to make the website attractive to search engines for these queries. The inserted phrases typically linked to other websites the attacker wishes to promote, in our case other online pharmacies.

We compiled a list of promoted search phrases by visiting the linked pharmacies posing as a search-engine crawler and noting the phrases observed. Many phrases were either identical or contained only minor differences, such as spelling variations on drug names. We reduced the list to a corpus of 48 unique queries, representative of all drugs advertised in this first step.

We then repeated this process for all 48 search phrases, gathering results daily from March 3, 2010 through April 11, 2010. The 48-query search subsequently led us to 371 source infections. We again browsed each of these source infections posing as a search engine crawler, and gathered a few thousand search phrases linked from the infected websites. After again sorting through the duplicates, we got a corpus of 218 unique search queries.

The risk of starting from a single seed is to only identify a single unrepresentative campaign. Hence, we ran a validation experiment to ensure that our selected queries had satisfactory coverage. We obtained a six-month sample of spam email (collected at a different time period, late 2009) gathered in a different context [42]. We ran SpamAssassin [5] on this spam corpus, to classify each spam as either pharmacy-related or otherwise. We then extracted all drug names encountered in the pharmacy-related spam, and observed that they defined a subset of the drug names present in our search queries. This gave us confidence that the query corpus was quite complete.

We further validated our query selection by comparing results obtained with our query corpus to those collected from two additional query corpora: 1) searches ran on an exhaustive list of 9 000 prescription drugs obtained from the US Food & Drug Administration [15], and 2) 1 179 drug-related search queries extracted from the HTTP logs of 169 source websites. The results (in Appendix A) confirm adequate coverage of our 218 queries.

2.3 Search-result classification

We attempt to classify all results obtained by the search-engine agent. Each query returns a mix of legitimate re-

sults (e.g., health information websites) and abusive results (e.g., spammed blog comments and forum postings advertising online pharmacies). We seek to distinguish between these different types of activity to better understand the impact of search-redirection attacks may have on legitimate pharmacies and other forms of abuse. We assign each result into one of the following categories: 1) search-redirection attacks, 2) health resources, 3) legitimate online pharmacies, 4) illicit online pharmacies, 5) blog or forum spam, and 6) uncategorized.

We mark websites as participating in search-redirection attacks by observing an HTTP redirect to a *different* website. Legitimate websites regularly use HTTP redirects, but it is less common to redirect to entirely different websites immediately upon arrival from a search engine. Every time the crawler encounters a redirect, it recursively follows and stores the intermediate URIs and IP addresses encountered in the database. These redirection chains are used to infer relationships between source infections and pharmacies in Section 4.3.

We performed two robustness checks to assess the suitability of classifying all external redirects as attacks. First, we found known drug terms in at least one redirect URI for 63% of source websites. Second, we found that 86% of redirecting websites point to the same website as 10 other redirecting websites. Finally, 93% of redirecting websites exhibit at least one of these behaviors, suggesting that the vast majority of redirecting websites are infected. In fact, we expect that most of the remaining 7% are also infected, but some attackers use unique websites for redirection. Thus, treating all external redirects as malicious appears reasonable in this study.

Health resources are websites such as `webmd.com` that describe characteristics of a drug. We used the Alexa Web Information Service API [1], which is based on the Open Directory [4] to determine each website category.

We distinguish between legitimate and illicit online pharmacies by using a list of registered pharmacies obtained from the non-profit organization Legitscript [3]. Legitscript maintains a whitelist of 324 confirmed legitimate online pharmacies, which require a verified doctor's prescription and sell genuine drugs. Illicit pharmacies are websites which do not appear in Legitscript's whitelist, and whose domain name contains drug names or words such as "pill," "tabs," or "prescription." LegitScript's list is likely incomplete, so we may incorrectly categorize some collected legitimate pharmacies as illicit, because they have not been certified by LegitScript.

Finally, blog and forum spam captures the frequent occurrence where websites that allow user-generated content are abused by users posting drug advertisements. We classify these websites based only on the URI structure, since collecting and storing the pages referenced by URIs is cost-prohibitive. We first check the URI subdomain and path for common terms indicating user-contributed content, such as "blog," "viewmember" or "profile." We also check any remaining URIs for drug terms appearing in the subdomain and path. While these might in fact be compromised websites that have been loaded with content, upon manual inspection the activity appears consistent with user-generated content abuse.

	URIs		Domains	
	#	%	#	%
Source infections	73 909	53.8	4 652	20.2
Active	44 503	32.4	2 907	12.6
Inactive	29 406	21.4	1 745	7.6
Health resources	1 817	1.3	422	1.8
Pharmacies	4 348	3.2	2 138	9.3
Legitimate	12	0.01	9	0.04
Illicit	4 336	3.2	2 129	9.2
Blog/forum spam	41 335	30.1	8 064	34.9
Uncategorized	15 945	11.6	7 766	33.7
Total	137 354	100.0	23 042	100.0

Table 1: Classification of all search results (4–10/2010).

3 Empirical analysis of search results

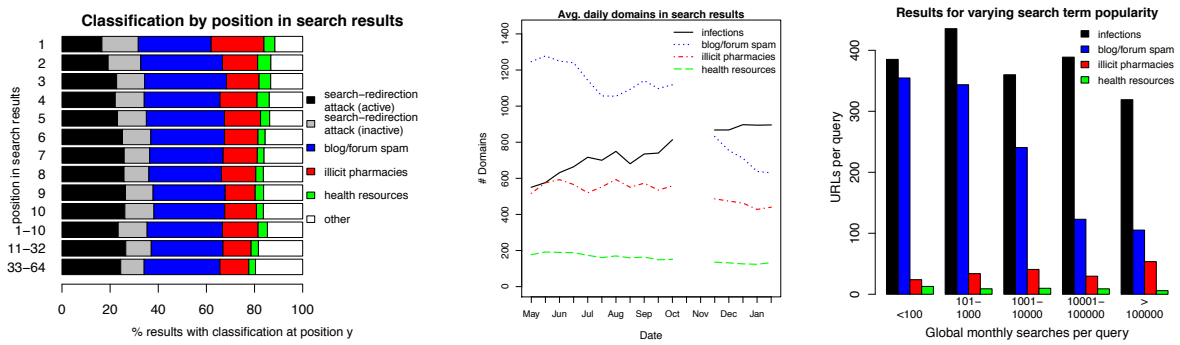
We begin our measurement analysis by examining the search results collected by the crawler. The objective here is to understand how prevalent search-redirection attacks are, in both absolute terms and relative to legitimate sources and other forms of abuse.

3.1 Breakdown of search results

Table 1 presents a breakdown of all search results obtained during the six months of primary data collection. 137 354 distinct URIs correspond to 23 042 different domains. We observed 44 503 of these URIs to be compromised websites (*source infections*) actively redirecting to pharmacies, 32% of the total. These corresponded to 4 652 unique infected source domains. We examine the redirection chains in more detail in Section 4.3.

An additional 29 406 URIs did not exhibit redirection even though they shared domains with URIs where we did observe redirection. There are several plausible explanations for why only some URIs on a domain will redirect to pharmacies. First, websites may continue to appear in the search results even after they have been remediated and stop redirecting to pharmacies. In Figure 1, the third link to appear in the search engine results has been disinfected, but the search engine is not yet aware of that. For 17% of the domains with inactive redirection links, the inactive links only appear in the search results after all the active redirects have stopped appearing.

However, for the remaining 83% of domains, the inactive links are interspersed among the URIs which ac-



(a) Distribution of different classes of results according to the position in the search results.

(b) Change in the average domains observed each day for different classes of search results over time.

(c) Search-redirection attacks appear in many queries; health resources and blog spam appear less often in popular queries.

Figure 2: Empirical measurements of pharmacy-related search results.

tively redirect. In this case, we expect that the miscreants' search engine optimization has failed, incorrectly promoting pages on the infected website that do not redirect to pharmacies.

By comparison, very few search results led to legitimate resources. 1 817 URIs, 1.3% of the total, pointed to websites offering health resources. Even more striking, only *nine* legitimate pharmacy websites, or 0.04% of the total, appeared in the search results. By contrast, 2 129 illicit pharmacies appeared directly in the search results. 30% of the results pointed to legitimate websites where miscreants had posted spam advertisements to online pharmacies. In contrast to the infected websites, these results require a user to click on the link to arrive at the pharmacy. It is also likely that many of these results were not intended for end users to visit; instead, they could be used to promote infected websites higher in the search results.

3.2 Variation in search position

Merely appearing in search results is not enough to ensure success for miscreants perpetrating search-redirection attacks. Appearing towards the top of the search results is also essential [20]. To that end, we collected data for an additional 10 weeks from November 15th 2010 to February 1st 2011 where we recorded the position of each URI in the search results.

Figure 2(a) presents the findings. Around one third of the time, search-redirection attacks appeared in the first position of the search results. 17% of the results were actively redirecting at the time they were observed in the first position. Blog and forum spam appeared in the top spot in 30% of results, while illicit pharmacies accounted for 22% and legitimate health resources just 5%.

The distribution of results remains fairly consistent across all 64 positions. Active search-redirection attacks increase their proportion slightly as the rankings fall, rising to 26% in positions 6–10. The share of illicit pharmacies falls considerably after the first position, from 22% to 14% for positions 2–10. Overall, it is striking how consistently all types of manipulation have crowded out legitimate health resources across all search positions.

3.3 Turnover in search results

Web search results can be very dynamic, even without an adversary trying to manipulate the outcome. We count the number of unique domains we observe in each day's sample for the categories outlined in Section 2. Figure 2(b) shows the average daily count for two-week periods from May 2010 to February 2011, covering both sample periods. The number of illicit pharmacies and health resources remains fairly constant over time, whereas the number of blogs and forums with pharmaceutical postings fell by almost half between May and February. Notably, the number of source infections steadily increased from 580 per day in early May to 895 by late January, a 50% increase in daily activity.

3.4 Variation in search queries

As part of its AdWords program, Google offers a free service called Traffic Estimator to check the estimated number of global monthly searches for any phrase.[3] We fetched the results for the 218 pharmacy search terms we regularly check; in total, over 2.4 million searches each month are made using these terms. This gives us a good first approximation of the relative popularity of web searches for finding drugs through online pharmacies. Some terms are searched for very frequently (as much as 246 000 times per month), while other terms are only searched for very occasionally.

We now explore whether the quality of search results vary according to the query's popularity. We might expect that less-popular search terms are easier to manip-

[3] https://adwords.google.com/select/TrafficEstimatorSandbox

ulate, but also that there could be more competition to manipulate the results of popular queries.

Figure 2(c) plots the average number of unique URIs observed per query for each category. For unpopular searches, with less than 100 global monthly searches, search-redirection attacks and blog spam appear with similar frequency. However, as the popularity of the search term increases, search-redirection attacks continue to appear in the search results with roughly the same regularity, while the blog and forum spam drops considerably (from 355 URIs per query to 105).

While occurring on a smaller scale, the trends of illicit pharmacies and legitimate health resources are also noteworthy. Health resources become increasingly crowded out by illicit websites as queries become more popular. For unpopular queries (< 100 global monthly searches), 13 health URIs appear. But for queries with more than 100 000 results, the number of results falls by more than half to 6. For illicit pharmacies, the trends are opposite. On less popular terms, the pharmacies appear less often (24 times on average). For the most popular terms, by contrast, 54 URIs point directly to illicit pharmacies. Taken together, these results suggest that the more sophisticated miscreants do a good job of targeting their websites to high-impact results.

4 Empirical analysis of search-redirection attacks

We now focus our attention on the structure and dynamics of search-redirection attacks themselves. We present evidence that certain types of websites are disproportionately targeted for compromise, that a few such websites appear most prominently in the search results, and that the chains of redirections from source infections to pharmacies betray a few clusters of concentrated criminality.

4.1 Concentration in search-redirection attack sources

We identified 7 298 source websites from both data sets that had been infected to take part in search-redirection attacks – 4 652 websites in the primary 6-month data set and 3 686 in the 10-week follow-up study. (1 130 sites are present in both datasets.) We now define a measure of the relative impact of these infected websites in order to better understand how they are used by attackers.

$$\mathcal{I}(\texttt{domain}) = \sum_{q \in \text{queries}} \sum_{d \in \text{days}} u_{qd} * 0.5^{\frac{r_{qd}-1}{10}}$$

where

u_{qd} : 1 if `domain` in results of query q on day d & actively redirects to pharmacy

u_{qd} : 0 otherwise

r_{qd} : `domain`'s position (1..64) in search results

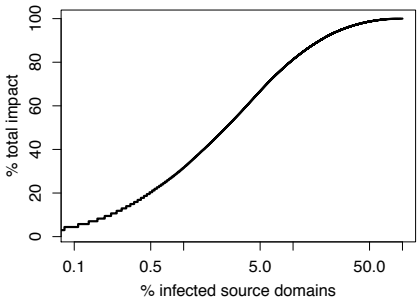


Figure 3: Rank-order CDF of domain impact reveals high concentration in search-redirection attacks.

	.com	.org	.edu	.net	other
% global Internet	45%	4%	$< 3\%$	6%	42%
% infected sources	55%	16%	6%	6%	17%
% inf. source impact	30%	24%	35%	2%	10%

Table 2: TLD breakdown of source infections.

The goal of the impact measure $\mathcal{I}$ is to distill the many observations of an infected domain into a comparable scalar value. Essentially, we add up the number of times a domain appears, while compensating for the relative ranking of the search results. Intuitively, when a domain appears as the top result it is much more likely to be utilized than if it appeared on page four of the results. The heuristic we use normalizes the top result to 1, and discounts the weighting by half as the position drops by 10. This corresponds to regarding results appearing on page one as twice as valuable as those on page two, which are twice as valuable as those on page three, and so on.

Some infected domains appeared in the search results much more frequently and in more prominent positions than others. The domain with the greatest impact – `unm.edu` – accounted for 2% of the total impact of all infected domains. Figure 3 plots using a logarithmic x-axis the ordered distribution of the impact measure $\mathcal{I}$ for source domains. The top 1% of source domains account for 32% of all impact, while the top 10% account for 81% of impact. This indicates that a small, concentrated number of infected websites account for most of the most visible redirections to online pharmacies.

We also examined how the prevalence and impact of source infections varied according to top-level domain (TLD). The top row in Table 2 shows the relative prevalence of different TLDs on the Internet [46]. The second row shows the occurrence of infections by TLD. The most affected TLD, with 55% of infected results, is `.com`, followed by `.org` (16%), `.edu` (6%) and `.net`(6%). These four TLDs account for 83% of all infections, with the remaining 17% spread across 159 TLDs. We also observed 25 infected `.gov` websites and

22 governmental websites from other countries.

One striking conclusion from comparing these figures is how more 'reputable' domains, such as `.com` (55% of infections vs. 45% of registrations), `.org` (16% vs. 4%) and `.edu` (6% vs. $< 3\%$), are infected than others. This is in contrast to other research, which has identified country-specific TLDs as sources of greater risk [26].

Furthermore, some TLDs are used more frequently in search-redirection attacks than others. While `.edu` domains constitute only 6% of source infections, they account for 35% of aggregate impact through redirections to pharmacy websites. Domains in `.com`, by contrast, account for more than half of all source domains but 30% of all impact. We next explore how infection durations vary across domains, in part with respect to TLD.

4.2 Variation in source infection lifetimes

One natural question when measuring the dynamics of attack and defense is how long infections persist. We define the "lifetime" of a source infection as the number of days between the first and last appearance of the domain in the search results while the domain is actively redirecting to pharmacies. Lifetime is a standard metric in the empirical security literature, even if the precise definitions vary by the attacks under study. For example, Moore and Clayton [27] observed that phishing websites have a median lifetime of 20 hours, while Nazario and Holz [32] found that domains used in fast-flux botnets have a mean lifetime of 18.5 days.

Calculating the lifetime of infected websites is not entirely straightforward, however. First, because we are tracking only the results of 218 search terms, we count as "death" whenever an infected website disappears from the results or stops redirecting, even if it remains infected. This is because we consider the harm to be minimized if the search engine detects manipulation and suppresses the infected results algorithmically. However, to the extent that our search sample is incomplete, we may be overly conservative in claiming a website is no longer infected when it has only disappeared from our results.

The second subtlety in measuring lifetimes is that many websites remain infected at the end our study, making it impossible to observe when these infections are remediated. Fortunately, this is a standard problem in statistics and can be solved using survival analysis. Websites that remain infected and in the search results at the end of our study are said to be *right-censored*. 1 368 of the 4 652 infected domains (29%) are right-censored.

The survival function $S(t)$ measures the probability that the infection's lifetime is greater than time t. The survival function is similar to a complementary cumulative distribution function, except that the probabilities must be estimated by taking censored data points into account. We use the standard Kaplan-Meier estimator [23] to calculate the survival function for infection lifetimes, as indicated by the solid black line in the graphs of Figure 4. The median lifetime of infected websites is 47 days; this can be seen in the graph by observing where $S(t) = 0.5$. Also noteworthy is that at the maximum time $t = 192$, $S(t) = 0.160$. Empirical survival estimators such as Kaplan-Meier do not extrapolate the survival distribution beyond the longest observed lifetime, which is 192 days in our sample. What we can discern from the data, nonetheless, is that 16% of infected domains were in the search results throughout the sample period, from April to October. Thus, we know that a significant minority of websites have remained infected for at least six months. Given how hard it is for webmasters to detect compromise, we expect that many of these long-lived infections have actually persisted far longer.

We next examine the characteristics of infected websites that could lead to longer or shorter lifetimes. One possible source of variation to consider is the TLD. Figure 4 (left) also includes survival function estimates for each of the four major TLDs, plus all others. Survival functions to the right of the primary black survival graph (e.g., `.edu`) have consistently longer lifetimes, while plots to the left (e.g., other and `.net`) have consistently shorter lifetimes. Infections on `.com` and `.org` appear slightly longer than average, but fall within the 95% confidence interval of the overall survival function.

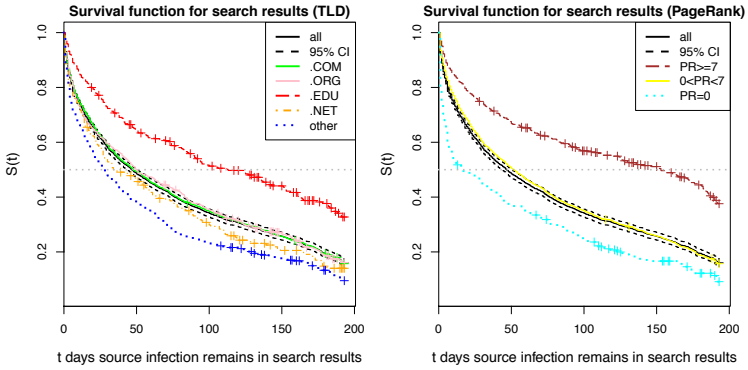
The median infection duration of `.edu` websites is 113 days, with 33% of `.edu` domains remaining infected throughout the 192-day sample period. By contrast, the less popular TLDs taken together have a median lifetime of just 28 days.

Another factor beyond TLD is also likely at play: the relative reputation of domains. Web domains with higher PageRank are naturally more likely to appear at the top of search results, and so are more likely to persist in the results. Indeed, we observe this in Figure 4 (center). Infected websites with PageRank 7 or higher have a median lifetime of 153 days, compared to just 17 days for infections on websites with PageRank 0.

One might expect that `.edu` domains would tend to have higher PageRanks, and so it is natural to wonder whether these graphs indicate the same effect, or two distinct effects. To disentangle the effects of different website characteristics on lifetime, we use a Cox proportional hazard model [10] of the form:

$$h(t) = \exp(\alpha + \text{PageRank}x_1 + \text{TLD}x_2)$$

Note that the dependent variable included in the Cox model is the hazard function $h(t)$. The hazard function $h(t)$ expresses the instantaneous risk of death at time t. Cox proportional hazard models are used on survival data in preference to standard regression models, but the aim



Cox-proportional hazard model
$h(t) = \exp(\alpha + \text{PageRank}x_1 + \text{TLD}x_2)$

	coef.	exp(coef.)	Std. Err.)	Significance
PageRank	−0.085	0.92	0.0098	$p < 0.001$
`.edu`	−0.26	0.77	0.086	$p < 0.001$
`.net`	0.08	1.1	0.084	
`.org`	0.055	1.0	0.054	
other TLDs	0.34	1.4	0.053	$p < 0.001$

log-rank test: Q=158, $p < 0.001$

Figure 4: Survival analysis of search-redirection attacks shows that TLD and PageRank influence infection lifetimes.

is the same as for regression: to measure the effect of different independent factors (in our case, TLD and PageRank) on a dependent variable (in our case, infection lifetime). PageRank is included as a numerical variable valued from 0 to 9, while TLD is encoded as a five-part categorical variable using deviation coding. (Deviation coding is used to measure each categories' deviation in lifetime from the overall mean value, rather than deviations across categories.) The results are presented in the table in Figure 4. PageRank is significantly correlated with lifetimes – lower PageRank matches shorter lifetimes while higher PageRank is associated with longer lifetimes. Separately, `.edu` domains are correlated with longer lifetimes and other TLDs to shorter lifetimes.

Coefficients in Cox models cannot be interpreted quite as easily as in standard linear regression; exponents (column 3 in the table) offer the clearest interpretation. $\exp(\text{PageRank}) = 0.92$ indicates that each one-point increase in the site's PageRank decreases the hazard rate by 8%. Decreases in the hazard leads to longer lifetimes. Meanwhile, $\exp(\texttt{.edu}) = 0.77$ indicates that the presence of a `.edu` domain, holding the PageRank constant, decreases the hazard rate by 23%. In contrast, the presence of any TLD besides `.com`, `.edu`, `.net` and `.org` increases the hazard rate by 40%.

Therefore, we can conclude from the model that *both* PageRank and TLD matter. Even lower-ranked university websites and high-rank non-university websites are being effectively targeted by attackers redirected traffic to pharmacy websites.

4.3 Characterizing the online pharmacy network

We now extend consideration beyond the websites directly appearing in search results to the intermediate and destination websites where traffic is driven in search-redirection attacks. We use the data to identify connections between a priori unrelated online pharmacies.

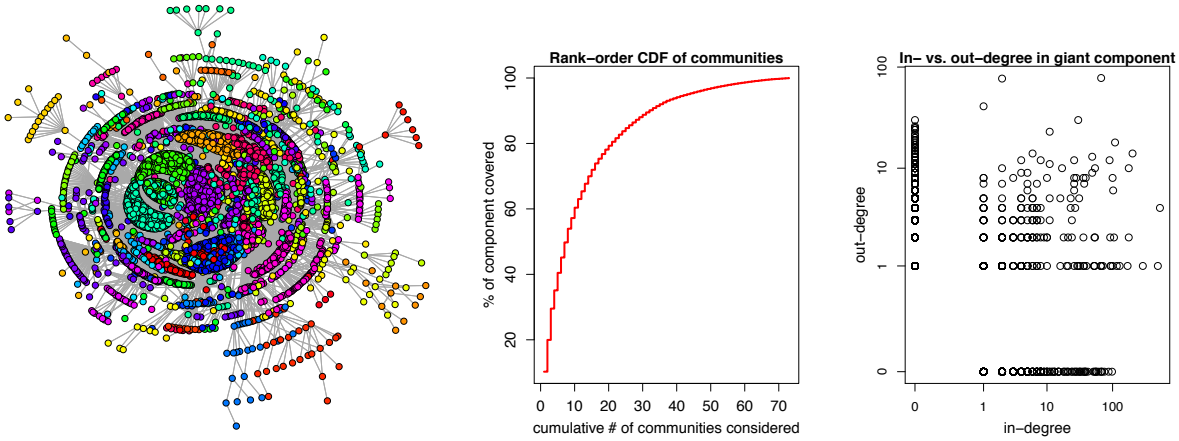
We construct a directed graph $G = (V, E)$ as follows. We gather all URIs in our database that are part of a redirection chain (source infection, redirector, online pharmacy) and assign each second-level domain to a node $v \in V$. We then create edges between nodes whenever domains redirect to each other. Suppose for instance that `http://www.example.com/blog` is infected and redirects to `http://1337.attacker.test` which in turns redirects to `http://www32.cheaprx4u.test`. We then create three nodes $v_1 =$ `example.com`, $v_2 =$ `attacker.test` and $v_3 =$ `cheaprx4u.test`, and two edges, $v_1 \to v_2$ and $v_2 \to v_3$. Now, if `http://hax0r.attacker.test` is also present in the database, and redirects to `http://www.otherrx.test`, we create a node $v_4 =$ `otherrx.test` and establish an edge $v_2 \to v_4$.

In the graph G so built, online pharmacies are usually leaf nodes with a positive in-degree and out-degree zero.[4] Compromised websites feeding traffic to pharmacies are generally represented as sources, with an in-degree of zero and a positive out-degree. Traffic redirectors, which act as intermediaries between compromised websites and online pharmacies have positive in- and out-degrees.

The resulting graph G for our entire database consists of 34 connected subgraphs containing more than two nodes. The largest connected component G_0 contains 96% of all infected domains, 90% of the redirection domains and 92% of the pharmacy domains collected throughout the six-month collection period.

In other words, we have evidence that most online pharmacies are connected by redirection chains. While this does not necessarily indicate that a single criminal organization is behind the entire online pharmacy network, this does tell us that most illicit online pharmacies in our measurements are obtaining traffic from a large interconnected network of advertising affiliates. Undercover investigations have confirmed the existence of such affiliate networks and provided anecdotal evidence on

[4]Manually checking the data, we find a few pharmacies have an out-degree of 1, and redirect to other pharmacies.



(a) Structure of the giant component G_0 that links 96% of infected domains. Links between vertices are based on observed traffic redirection chains. Vertices are colored according to their community.

(b) CDF of nodes in the giant component belonging to different communities. The largest 7 (out of 73) communities comprise over half the nodes.

(c) Scatter plot of in- and out-degree of nodes in the giant component. (Log-log scale, where 0 is technically represented as 0.1.)

Figure 5: Network analysis of redirection chains reveals community structure in search-redirection attacks.

their operations [44], but they have not precisely quantified their influence. These affiliate networks consist of a loosely organized set of independent advertising entities that feed traffic to their customers (e.g., online retailers) in exchange for a commission on any resulting sales.

Communities and affiliated campaigns. To uncover affiliate networks, we locate *communities* within G_0, i.e., sets of vertices closely interconnected with each other and only loosely connected to the rest of the graph. Here, each community represents a set of domains in close relationship with each other, possibly part of the same business operation, or in the same manipulation campaigns. Several algorithms have recently been proposed for community detection, e.g., [36,41,43]. We use the spin-glass model proposed by Reichardt and Bornhold [43] (with $q = 500$, $\gamma = 1$) because its stochastic nature allows it to complete quickly even on large graphs like ours, and because it works on directed graphs.

In Figure 5(a), we plot a visual representation of G_0. Different colors denote different communities. The community detection algorithm identifies a total of 73 distinct communities. Most larger communities can be observed in the dense clusters of nodes in the center of the figure, and it appears that less than a dozen of communities play a significant role. More precisely, we plot in Figure 5(b) the cumulative fraction of nodes in G_0 as a function of the number of communities considered. The graph shows that the seven largest communities account for more than half of the nodes in the graph, and that about two thirds of the nodes belong to one of the top twelve communities. In other words, a relatively small number of loosely interconnected, possibly distinct, operations is responsible for most attacks.

Manual inspection confirms these insights. For instance, the third largest community (400 nodes) consists of compromised hosts primarily sending traffic to a single redirector, which itself redirects to a single pharmacy (`securetabs.net`).

Figure 5(c) is a scatter-plot of the in- and out-degree of each node in G_0. A vast majority of nodes are source infections (null in-degree, high out-degree, i.e., points along the y-axis) or pharmacies (low out-degree, high in-degree, i.e., along the x-axis). Redirectors, with non-zero in- and out- degrees are comparatively rare. We identify 314 redirectors in G_0, out of which only 127 have both an in- and an out-degree greater than two. 103 of these 127 redirectors (80%) are *cut vertices* for G_0. That is, removing any of these 103 redirectors would partition G_0. We will discuss these interesting properties in further details in Section 6, where we detail the possible remedial strategies against the search-redirection attacks.

4.4 Attack websites in blacklists

The websites we have identified here have either been compromised (in the case of source infections) or have taken advantage of compromised servers (in the case of redirects and pharmacies). Given such insalubrious circumstances, we wondered if any of the third party blacklists dedicated to identifying Internet wickedness might also have noticed these same websites. To that end, we consulted three different sources: Google's Safe Browsing API, which identifies web-based malware; the `zen.spamhaus.org` blacklist, which identifies email spam senders; and McAfee SiteAdvisor, which tests websites for "spyware, spam and scams".

Figure 6 plots sets of Venn diagrams of the three blacklists for each class of attack domain. Several trends are

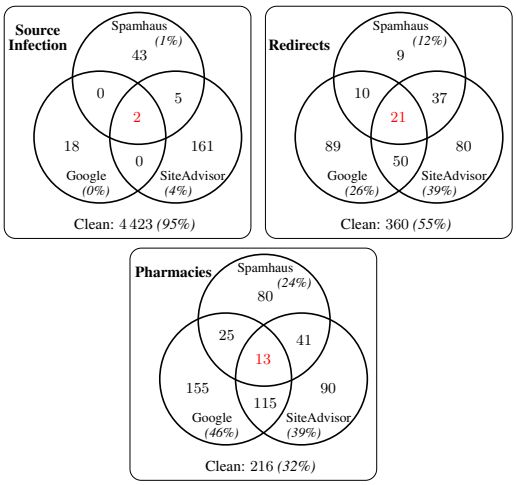


Figure 6: Comparing web and email blacklists.

	Mean	Median	% Searches>0	Total
Main	14 388	1600	73%	2 374 085
FDA drugs	74	0	6%	323 104
Extra queries	46 380	1 300	59%	32 652 121
Total	6 771	0	20%	35 343 610

Table 3: Monthly search query popularity according to the Google Adwords Traffic Estimator.

apparent from inspecting the diagrams. First, source infections are not widely reported by any of the blacklists (95% do not appear on a single blacklist), but around half of the redirects are found on at least one blacklist and over two thirds of pharmacy websites show up on at least one blacklist. Surprisingly, 12% of redirects appear on the email spam blacklist, as well as 24% of pharmacies. We speculate that this could be caused by affiliates advertising pharmacy domains in email spam, but it could also be that the pharmacies directly send email spam advertisements or use botnets for both hosting and spamming.

The level of coverage of Google and SiteAdvisor are comparable, which is somewhat surprising given SiteAdvisor's relatively broader remit to flag scams, not only malware. Google's more comprehensive coverage of pharmacy websites in particular suggests that some pharmacies may also engage in distributing malware. We conclude by noting that the majority of websites affected by the traffic redirection scam are not identified by any of these blacklists. This in turn suggests that relatively little pressure is currently being applied to the miscreants carrying out the attacks.

5 Towards a conversion rate estimate

While it is difficult to measure precisely as an outsider, we nonetheless would like to provide a ballpark figure for how lucrative web search is to the illicit online prescription drug trade. Here we measure two aspects of the demand side: search-query popularity and sales traffic.

For the first category, we once again turn to the Google Traffic Estimator to better understand how many people use online pharmacies advertised though search-redirection attacks. Table 3 lists the results for each of the three search query corpora described in Section 2.2 and Appendix A. The main and extra queries attract the most visitors, with a median of 1 600 monthly searches for the main sample and 1 300 for the extra queries. Several highly popular terms appeared in the results: "viagra" and "pharmacy" each attract 6 million monthly searches, while "cialis" and "phentermine" appear in around 3 million each. By contrast, only 6% of the search queries in the FDA sample registered with the Google tool. The FDA query list includes around 6 500 terms, which dwarfs the size of the other lists. Since over 90% of the FDA queries are estimated to have no monthly searches, the overal median popularity is also zero.

While these search terms do not cover all possible queries, taken together they do represent a useful lower bound on the global monthly searches for drugs. To translate the aggregate search count into visits to pharmacies facilitated by search-redirection attacks, we assume that the share of visits websites receive is proportional to the number of URIs that turn up in the search results. Given that 38% of the search results we found pointed to infected websites, we might expect that the monthly share of visits to these sites facilitated by Google searches to be around 13 million. Google reportedly has a 64.4% market share in search [13]. Consequently we expect that the traffic arriving from other search engines to be $\frac{1-0.644}{0.644} * 13 \text{ million} = 7 \text{ million}$.

We manually visited 150 pharmacy websites identified in our study and added drugs to shopping carts to observe the beginning of the payment process. We found that 94 of these websites in fact pointed to one of 21 different payment processing websites. These websites typically had valid SSL certificates signed by trusted authorities, which helps explain why multiple pharmacy storefronts may want to share the same payment processing website.

The fact that these websites are only used for payment processing means that if we could measure the traffic to these websites, then we could roughly approximate how many people actually purchase drugs from these pharmacies. Fortunately for us, these websites receive enough traffic to be monitored by services such as Alexa. We tallied Alexa's estimated daily visits for each of these websites; in total, they receive 855 000 monthly visits.

We next checked whether these payment websites also offered payment processing other than just for pharmacy websites. To check this, we fetched 1 000 backlinks for each of the sites from Yahoo Site Explorer [6]. Collectively, 1 561 domains linked in to the payment websites. From URI naming and manual inspection, we determined that at least 1 181 of the backlink domains, or

75%, are online pharmacies. This suggests that the primary purpose of these websites is to process payments for online pharmacies.

Taken together, we can use all the information discussed above to provide a lower bound on the sales conversion rate of pharmacy web search traffic:

$$\text{Conversion} \approx \frac{0.75 \times 855\,000}{20\,000\,000} = 3.2\% \,.$$

To ensure that the estimate is a lower bound for the true conversion rate, whenever there is uncertainty over the correct figures, we select smaller estimates for factors in the numerator and larger estimates for factors in the denominator. For example, it is possible that the estimate of visits to payment sites is too small, since pharmacies could use more than the 21 websites we identified to process payments. A more accurate estimate here would strictly increase the conversion rate. Similarly, 20 million visits to search-redirection websites may be an overestimate, if, for instance, more popular search queries suffer from fewer search-redirection attacks. Reducing this estimate would increase the conversion rate since the figure is in the denominator.

There is likely one slight overestimate present in the numerator. It is not certain that every single visitor to a payment processing site eventually concluded the transaction. However, because these sites are *only* used to process payments, we can legitimately assume that most visitors ended up purchasing products. Even with a conservative assumption that only 1 in 10 visitors to the payment processing site actually complete a transaction, the lower bound on the conversion rates we would obtain (in the order of 0.3%) far exceeds the conversion rates observed for email spam [22] or social-network spam [17].

While email spam has attracted more attention, our research suggests that more illicit pharmacy purchases are facilitated by search-redirection attacks than by email spam. One study estimated that the entire Storm botnet (which accounted for between 20-30% of email spam at its peak [12, 37]) attracted around 2 100 sales per month [22]. The payment processing websites tied to search-redirection attacks collectively process many hundreds of thousands of monthly sales. Even allowing for the possibility that these websites may also process payments for pharmacies advertised through email spam, the bulk of sales are likely dominated by referrals from web search. This is not surprising, given that most people find it more natural to turn to their search engine of choice than to their spam folder when shopping online. To those who aim to reduce unauthorized pharmaceutical sales, the implication is clear: more emphasis on combating transactions facilitated by web search is warranted.

6 Mitigation strategies

The measurements we gathered lead us to consider three complementary mitigation strategies to reduce the impact of search-redirection attacks. One can target the infected sources, advocate search-engine intervention, or try to disrupt the affiliate networks.

Remediation at the sources. The existing public-private partnership initiated by the White House [19] has so far focused on areas other than search-redirection attacks. Domain name registrars (led by GoDaddy) can shut down maliciously registered domains, while Google has focused on blocking advertisements (but not necessarily search results) from unauthorized pharmacies. Unfortunately, no single entity speaks for the many webmasters whose sites have unknowingly been recruited to drive traffic to illicit pharmacies.

Nonetheless, eradicating source infections at key websites could be effective. As shown in Figure 3, a small number of source infections repeatedly appear towards the top of the search results. Remediating only the most frequently-occurring websites could substantially reduce sales. Furthermore, attackers would likely struggle to adapt to the heightened enforcement. Placing websites at high-ranking search positions through search-engine optimization is a slow process, given that the search engine controls the rankings-update cycle. Second, high-ranking websites that can permeate the top levels of search results are fairly scarce resources, so that any coordinated reduction is likely to be painful for pharmacies.

How might an enforcement agent select which websites to target for remediation? Again, our findings are informative. The survival analysis in Section 4.2 indicates that websites with high PageRank or `.edu` TLDs are more persistent. A simple heuristic, then, would be for an agent to run a few search queries for drug terms and try to clean up any `.edu` or high-ranking website that appears in multiple results.

Search-engine intervention. In the absence of direct law enforcement involvement in remediating source infections, search engines could play a more active role in detecting search-redirection attacks and blocking them from search results. Google already blocks websites that are known to be distributing malware [40], and recently began including warnings on websites believed to be compromised. From anecdotal inspection, several source websites participating in search-redirection attacks now carry the warning. Users are still free to visit the compromised website, however, so those seeking to buy drugs without a prescription may still find willing sellers. We encourage search engines to consider dropping such results altogether, given the illegal activity that is being directly facilitated.

Disrupting the redirection network. The high degree of interconnection of the different sites we observed in Section 4 suggests that monetary profits come from funneling traffic between different affiliates. One can thus conjecture that disrupting the connectivity of the network we observed would have adverse economic consequences for the miscreants. Can this be easily achieved?

As described in Section 4, while the network of pharmacies, sources, and redirectors is almost completely interconnected, there is a comparatively small number of nodes in the network that redirect traffic from one host to the next and play a central role in the drug trade. Specifically, taking down any of 103 redirectors would break up the large network of affiliates we observed, and could have strong disruptive effects on the profits made by advertisers. Of course, we would expect attackers to quickly move redirectors to different hosts after takedowns — and in fact, have, over the long measurement interval we consider, evidence that this sometimes happens. Nevertheless, the currently long lifetime of redirectors indicates that defenders could act more forcefully.

Perhaps even more interestingly, we were able to find BGP Autonomous System (AS) information for 84 of the 127 redirectors with in- and out-degrees greater than two;[5] of these, 53 (or 63%) belong to one of only 11 distinct ASes.[6] In other words, a very limited number of infrastructure providers appear to play an important role in the illicit online drug trade. Likewise, we were able to identify domain name registrars for 73 of the redirector domains; 49 of these domains belong to one of only 5 registrars (ENOM and GoDaddy, which is expected given their market share, but also "A to Z Domains Solutions," "BizCN," and "Directi Internet Solutions," which are far more represented in this sample than their market share would warrant).

Determining whether these hosting providers and registrars are willing participants or simply have lax hosting practices is beyond the scope of our investigation. However, by strengthening their controls, these service providers could probably make it harder to operate redirectors, thereby yielding tangible benefits in combating illicit online drug trade. Should these registrars and hosting providers take action, we would certainly expect the miscreants to adapt, and move to different providers (e.g., bulletproof hosting); but, it is likely that these alternative solutions would be more financially costly than what is currently used, which in turn would reduce the profit margins miscreants enjoy. In the end, making illicit online commerce an unattractive economic proposition could be the strongest deterrent to such activities.

In sum, any subset of source-infection remediation, search-engine filtering, and redirector take-down would make it more difficult for miscreants to conduct their business. Combining these mitigations would likely cause significant hardship to the criminal networks in play and would help thwart the illicit online trade of pharmaceutical drugs (and of other counterfeit goods).

7 Related work

The shift observed in the past decade, from Internet and computer security attacks motivated by fame and reputation to attacks motivated by financial gain [30], has led to a number of measurement studies that quantify various aspects of the problem, and to motivate possible intervention policies by quantitative analysis. Due to the amount of network measurement literature available, we focus here on work most closely related to this paper.

Many studies, e.g., [7, 22, 24, 50], have focused on email spam, describing the magnitude of the problem in terms of network resources being consumed, as well as some of its salient characteristics. Two key take-away points are that spam is a game of very large numbers, and that it is not a very effective technique to advertise products, as observed conversion rates (fraction of email spam that eventually result in a sale) are small. As pointed out earlier, spamming techniques are however evolving and increase their effectiveness by better targeting potential customers, as described by the recent flurry of spam observed in social networks [17].

A very recent paper by Levchenko et al. [24] provides a thorough investigation of the different actors participating spamming campaigns, from the spammers themselves, to the suppliers of illicit goods (luxury items, software, pharmaceutical drugs, ...). The key difference with the present study is that Levchenko et al. are focusing on businesses advertising by spam, while we are looking into search-engine manipulation. The data we gathered (see Section 4.4) seems to suggest that, so far, the two sets of miscreants remain relatively disjoint, but that advertising based on search engine manipulation is on the rise (see Section 3.3).

Measurement studies of spam have also informed possible intervention policies, by identifying some infrastructure weaknesses. For instance, taking down a few servers from suspicious Internet Service Providers [9] can significantly reduce the overall volume of email spam. Infiltration of spam-generating botnets, as suggested by [39], has also been shown to be effective in designing much more accurate spam filtering rules.

A series of papers by Moore and Clayton [27, 29, 31] investigates the economics of phishing, and show interesting insights on the tactics phishers use to evade detection. A further outcome of this line of research is a set of

[5] The remaining 43 redirectors had gone offline when we ran this experiment in February 2011.

[6] Many nodes in a given community are hosted on the same AS, giving additional evidence that the community detection algorithm discussed in Section 4 is quite accurate.

recommended intervention techniques to combat phishing, e.g., applying economic pressure on DNS registrars. The present paper borrows some of the techniques (use of Webalizer data, lifetime computation) used for phishing measurements, as they apply as well to measurement of online pharmacy activity (see Section 3).

A separate branch of research has focused on economic implications of online crime. Thomas and Martin [45], Franklin et al. [16] and Zhuge et al. [51] passively monitor the advertised prices of illicit commodities exchanged in varied online environments (IRC channels and web forums). They estimate the size of the markets associated with the exchange of credit card numbers, identity information, email address databases, and forged video game credentials. Christin et al. [8] mine online forum data to assess the economic impact of a social engineering attack pervasive on Japanese-language websites, and to identify some of the key characteristics of the network of perpetrators behind these scams.

More closely related to the attack described here, Ntoulas et al. [34] measure search engine manipulation attacks, and Wang et al. [47] show the connection between web and email spam, and online advertisers.

The medical literature has been preoccupied with illicit online pharmacies for a few years, but has mostly looked at smaller data samples, and has solely focused on the retail side rather than the entire infrastructure supporting this commerce. As examples, Henney et al. investigated the credentials of 37 online pharmacies [18]. Littlejohn et al. [25] focused on a slightly larger sample of 275 websites, to primarily inform the socio-economic impact of Internet availability on drug abuse. Likewise, we are not the first to evidence the existence of advertising affiliate networks, which have been previously described informally (see, e.g., [44]).

We believe that the work presented in this paper is the first to provide a detailed analysis of search-redirection attacks, and to substantiate their use with a quantitative analysis of the overall magnitude of the illicit online prescription drug trade. Further, we obtain both an understanding of the structure of the miscreants' networks, and an idea of the conversion rates they can expect. In that respect, our measurements may be a useful starting point for a more thorough quantitative economic analysis.

8 Conclusions and future work

Given the enormous value of web search, it is no surprise that miscreants have taken aim at manipulating its results. We have presented evidence of systematic compromise of high-ranking websites that have been reprogrammed to dynamically redirect to online pharmacies. These search-redirection attacks are present in one third of the search results we collected. The infections persist for months, 96% of the infected hosts are connected through redirections, and a few collections of redirectors are critical to the connection between source infections and pharmacies. We have also observed that legitimate businesses are nearly absent from the search results, having been completely drawn out of the search results by blog and forum spam and compromised websites. We also offer a conservative estimate of between 0.3% and 3% conversion rate of searches for drugs turning into sales, which should motivate the pressing need for countermeasures. Fortunately, we are optimistic that the criminals behind search-redirection attacks could be disrupted with targeted interventions due to the high concentrations we observed empirically.

In terms of immediate future work, there is nothing inherent to the search-redirection attack suggesting it only applies to online pharmacies. Even though counterfeit drugs are the most pressing issue to deal with due to their inherent danger, other purveyors of black-market goods, such as counterfeit software, or luxury goods replicas, might also hire affiliates that manipulate search results with infected websites for advertising purposes.

We ran a brief (12 days) pilot experiment to assess how search-redirection attacks applied to counterfeit software in October 2010. After collecting results from 466 queries, created using input from Google Adwords Keyword Tool, we gathered 328 infected source domains, 72 redirect domains and 140 domains selling counterfeit software. Using the same clustering techniques described earlier in the paper, we discovered two connected components dominating the network, each in its own way: one component was responsible for 44% of the identified infections, and the other was responsible for 30% of the software-selling sites.

We also observed a small but substantial (12.5%) overlap in the set of redirection domains with those used for online pharmacies. Some redirection domains thus provide generic traffic redirection services for different types of illicit trade. However, the small overlap is also a sign of fragmentation among the different fraudulent trading activities. We have begun a longitudinal study of all retail operations benefiting from search-redirection attacks, in order to better understand the economic relationships between advertisers and resellers.

Systematic monitoring of web search results will likely become more important due to the value miscreants have already identified in manipulating outcomes. Indeed, this paper has shown that understanding the structure of the attackers' networks gives defenders a strong advantage when devising countermeasures.

Acknowledgments

We thank our anonymous reviewers for feedback on earlier revisions of this manuscript, and our shepherd, Lucas Ballard, for his help in finalizing this version. This

research was partially supported by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office, and by the National Science Foundation under ITR award CCF-0424422 (TRUST).

References

[1] Alexa Web Information Service. http://aws.amazon.com/awis/.

[2] Google Web Search API. https://code.google.com/apis/websearch/.

[3] Legitscript LLC. http://www.legitscript.com/.

[4] Open Directory project. http://www.dmoz.org/.

[5] The Apache SpamAssassin Project. http://spamassassin.apache.org/.

[6] Yahoo Site Explorer. http://siteexplorer.search.yahoo.com/.

[7] D. Anderson, C. Fleizach, S. Savage, and G. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In *Proc. USENIX Security'07*, pp. 1–14. Boston, MA, Aug. 2007.

[8] N. Christin, S. Yanagihara, and K. Kamataki. Dissecting one click frauds. In *Proc. ACM CCS'10*, pp. 15–26, Chicago, IL, October 2010.

[9] R. Clayton. How much did shutting down McColo help? In *Proc. CEAS'09*, July 2009.

[10] D. Cox. Regression models and life-tables. *J. Royal Stat. Soc., Series B*, 34:187–220, 1972.

[11] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *Proc. USENIX Security'04*, pp. 303–320, San Diego, CA, August 2004.

[12] J. Dunn. Srizbi grows into world's largest botnet. *CSO*, May 2008. http://www.csoonline.com/article/356219/srizbi-grows-into-world-s-largest- botnet.

[13] Experian Hitwise. Experian Hitwise reports Bing-powered share of searches reaches 30 percent in March 2011. http://www.hitwise.com/us/press-center/press-releases/experian-hitwise-reports-bing-powered-share-of-s/. April 2011.

[14] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. RFC2616: Hypertext Transfer Protocol–HTTP/1.1. June 1999.

[15] U.S. Food and Drug Administration. National drug code directory, Nov. 2010. http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm.

[16] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *Proc. ACM CCS'07*, pp. 375–388, Alexandria, VA, Oct. 2007.

[17] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground in 140 characters or less. In *Proc. ACM CCS'10*, pp. 27–37, Chicago, IL, Oct. 2010.

[18] J. Henney, J. Shuren, S. Nightingale, and T. McGinnis. Internet purchase of prescription drugs: Buyer beware. *Ann. Int. Med.*, 131(11):861–862, Dec. 1999.

[19] K. Jackson Higgins. Google, GoDaddy help form group to fight fake online pharmacies. *Dark Reading*, Dec. 2010. http://www.darkreading.com/security/privacy/228800671/google-godaddy-help-form-group-to-fight-fake-online-pharmacies.html.

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

[21] G. Jolly. Explicit estimates from capture-recapture data with both death and immigration – stochastic model. *Biometrika*, 52(1-2):225–247, 1965.

[22] 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'08*, pp. 3–14, Alexandria, VA, Oct. 2008.

[23] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53:457–481, 1958.

[24] 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. IEEE Symp. Sec. and Privacy*, Oakland, CA, May 2011. To appear.

[25] C. Littlejohn, A. Baldacchino, F. Schifano, and P. Deluca. Internet pharmacies and online prescription drug sales: a cross-sectional study. *Drugs: Edu., Prev., and Policy*, 12(1):75–80, 2005.

[26] McAfee. Mapping the Mal Web., 2010. http://us.mcafee.com/en-us/local/docs/Mapping_Mal_Web.pdf.

[27] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In *Proc. APWG eCrime'07*, pp. 1–13, Pittsburgh, PA, Oct. 2007.

[28] T. Moore and R. Clayton. The consequence of non-cooperation in the fight against phishing. In *Proc. APWG eCrime'08*, Atlanta, GA, October 2008.

[29] T. Moore and R. Clayton. Evil searching: Compromise and recompromise of internet hosts for phishing. In *Proc. Financial Crypto'09*, LNCS 5628, pp. 256–272, Barbados, February 2009.

[30] T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *J. Econ. Persp.*, 23(3):3–20, Summer 2009.

[31] T. Moore, R. Clayton, and H. Stern. Temporal correlations between spam and phishing websites. In *Proc. USENIX LEET'09*, Boston, MA, April 2009.

[32] J. Nazario and T. Holz. As the net churns: Fast-flux botnet observations. In *Proc. MALWARE'08*, pp. 24–31, Fairfax, VA, October 2008.

[33] Y. Niu, H. Chen, F. Hsu, Y.-M. Wang, and M. Ma. A quantitative study of forum spamming using context-based analysis. In *Proc. ISOC NDSS'07*. San Diego, CA, Feb. 2007.

[34] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. WWW'06*, pp. 83–92, Edinburgh, Scotland, May 2006.

[35] Department of Justice. Implementation of the Ryan Haight Online Pharmacy Consumer Protection Act of 2008. *Fed. Reg.*, 74(64):15596–15625, 2009.

[36] G. Palla, I. Derény, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, June 2005.

[37] D. Pauli. Srizbi botnet sets new records for spam. *PCWorld*, May 2008. http://www.pcworld.com/businesscenter/article/145631/srizbi_botnet_sets_new_records_for_spam.html.

[38] PhpBB Ltd. PhpBB website. http://www.phpbb.com.

[39] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G.M. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet Judo: Fighting Spam with Itself . In *Proc. ISOC NDSS'10*, San Diego, CA, March 2010.

[40] N. Provos, P. Mavrommatis, M. Rajab, and F. Monrose. All your iFrames point to us. In *Proc. USENIX Security'08*, pp. 1–16, San Jose, CA, Aug. 2008.

[41] U. Nandini Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, 2007.

[42] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *Proc. ACM SIGCOMM'06*, pp. 291–302, Pisa, Italy, Sep. 2006.

[43] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74(1):016110, July 2006.

[44] D. Samosseiko. The partnerka – what is it, and why should you care? In *Virus Bulletin Conf.*, 2009.

[45] R. Thomas and J. Martin. The underground economy: Priceless. *;login:*, 31(6):7–16, December 2006.

[46] Verisign. The domain industry brief, 2010. http://www.verisigninc.com/assets/Verisign_DNIB_Nov2010_WEB.pdf.

[47] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: connecting web spammers with advertisers. In *Proc. WWW'07*, pp. 291–300, Banff, AB, Canada, May 2007.

[48] T. Wilson. Researchers link storm botnet to illegal pharmaceutical sales. *Dark Reading*, June 2008. http://www.darkreading.com/security/security-management/211201114/index.html.

[49] Wordpress. Wordpress website, September 2009. http://www.wordpress.org.

[50] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: Signatures and characteristics. *ACM SIGCOMM Comp. Comm. Rev.*, 38(4):171–182, 2008.

[51] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying malicious websites and the underground economy on the Chinese web. In *Managing Information Risk and the Economics of Security*, pp. 225–244. Springer, 2008.

A Additional query-sample validation

We have collected two sets of additional search queries to compare to our main corpus of 218 terms. First, we have derived a query set from an exhaustive list of 9 000 prescription drugs provided by the US Food and Drugs Administration [15]. We ran a single query in the form of "no prescription [drug name]" and collected the first 64 results for each drug in the list. We executed the 9 000 queries over five days in August 2010. About 2 500 of the queries returned no search results. Of the queries that returned results, we observed redirection in at least one of the search results for 4 350 terms.

For the second list, we inspected summaries of server logs for 169 infected websites to identify drug-related search terms that redirected to pharmacies. We obtained this information from infected web servers running The Webalizer,[7] which creates monthly reports, based on HTTP logs, of how many visitors a website receives, the most popular pages on the website, and so forth. It is not uncommon to leave these reports "world-readable" in a standard location on the server, which means that anyone can inspect their contents.

In August 2010, we checked 3 806 infected websites for Webalizer, finding it accessible on 169 websites. We recorded all available data – which usually included monthly reports of activity up to and including the current month. One of the individual sub-reports that Webalizer creates is a list of search terms that have been used to locate the site. Not all Webalizer reports list referrer terms, but we found 83 websites that did include drug names in the referrer terms for one or more months of the log reports. Since we identified the infected servers running Webalizer by inspecting results of the 218 queries from our main corpus, it is unsurprising that 98 of these terms appeared in the logs. However, the logs also contained an additional 1 179 search queries with drug terms. We use these additional search terms as an *extra queries list* to compare against the main corpus.

We collected the top 64 results for the extra queries list daily between October 20 and 31, 2010. When comparing these results to our main query corpus, we examine only the results obtained during this time period, resulting in a significantly smaller number of results than for our complete nine-month collection.

We compare our main list to the additional lists in three ways. First, we compare the classification of search results for differences in the types of results obtained. Second, we compare the distribution of TLD and PageRank for source infections obtained for both samples. Third, we compute the intersection between the domains obtained by both sets of queries for source infections, redirects and pharmacies.

[7] `http://www.mrunix.net/webalizer/`

	FDA drug list				Extra query list			
	Drug list		Main list		Extra list		Main list	
	URIs	dom.	URIs	dom.	URIs	dom.	URIs	dom.
Search result classification								
Source infections	24.7	4.0	43.7	22.4	35.6	14.0	49.3	27.9
Health resources	12.7	7.4	2.8	3.5	4.9	4.2	2.4	3.0
Legit. pharm.	0.5	0.1	0.03	0.07	0.1	0.1	0.02	0.05
Illicit pharm.	6.7	6.9	8.2	13.6	6.1	11.6	6.5	12.0
Blog/forum spam	25.4	23.7	18.6	17.8	26.3	22.7	17.8	17.7
Uncategorized	30.1	57.9	26.7	42.7	27.2	46.9	24.0	39.4
Source infection TLD breakdown								
`.com`		60.0		56.9		56.3		54.6
`.org`		13.8		17.0		15.4		18.0
`.edu`		5.6		8.9		6.2		9.3
`.net`		6.1		5.6		5.6		4.6
other		14.3		11.5		16.5		13.5
Source infection PageRank breakdown								
PR $0 \leq 3$		47.2		35.0		47.5		41.9
PR $3 \leq 6$		41.4		51.3		44.2		46.3
PR ≥ 7		11.4		13.7		8.3		11.8

Table 4: Comparing different lists of search terms to the main list used in the paper. All numbers are percentages.

Table 4 compares the FDA drugs and extra queries lists to the main list. The breakdown of search results for both samples is slightly different from what we obtained using the main queries. For instance, only 25% of the URIs in the FDA results are infections, compared to 44% for the main list during the same time period. 13% of the results in the FDA drug list point to legitimate health resources, compared to only 3% of the main sample. This is not surprising, given that the drug list often included many drugs that are not popular choices for sales by online pharmacies. Illicit pharmacies appear slightly less often in the drugs sample (6% vs. 8%), while blog and forum spam is more prevalent (25% to 19%).

The extra queries list follows the FDA list in some ways, e.g., more blog infections and fewer source infections than results from the corresponding main list. On the other hand, the URI breakdown in health resources is much closer (4.9% vs. 2.4%). In all samples, the number of results that point to legitimate pharmacies is very small, though admittedly biggest in the drugs sample (0.5% vs. 0.1% for the extra queries).

We next take a closer look at the characteristics of the source infections themselves. The TLD breakdown is roughly similar, with a few exceptions. `.com` is found slightly more often in the FDA drugs and extra queries results, while `.org` and `.edu` appear a bit more often in the results for the main sample. The drugs and extra queries list tend to have slightly lower PageRank than the results from the main sample, but the difference is slight.

B Estimating the number of sites involved

We also wish to compare the number of attack domains that can be identified for different sets of queries. Figure 7 compares the overlap between each class of domains for the different samples. The FDA drugs queries identified 1 919 distinct source infections, compared to

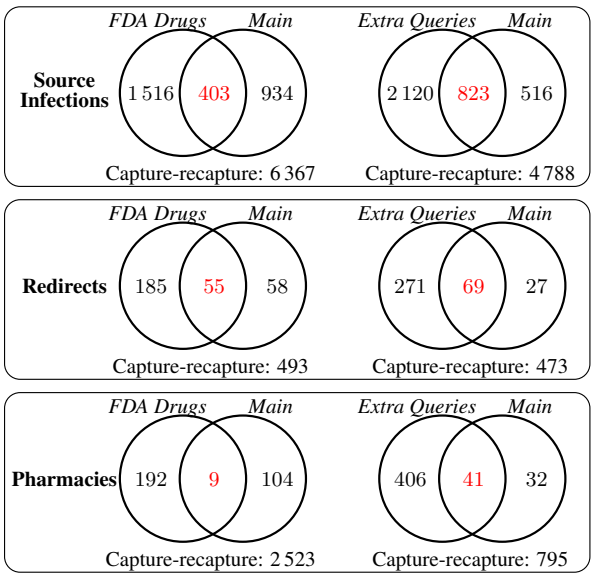


Figure 7: Comparing the source, redirect and pharmacy domains observed for different query lists.

1 337 found in the main sample during the same time period. 403 infected domains appeared in both lists.

It is unreasonable to expect any single query list to be comprehensive and identify all attack websites. In both of our test cases, we compared much larger query corpora to a smaller list (6 500 and 1 179 versus 218). Despite this, in each case many domains were found exclusively in the results of the smaller main sample. This is a common outcome when trying to measure online attacks such as phishing websites [28].

Given the difficulty in getting a truly comprehensive query list, one alternative is to estimate the total number of affected domains to get a better sense of an attack's impact. We apply capture-recapture analysis [21] based on our incomplete samples to get an estimate of the magnitude of the activity studied in this paper.

Capture-recapture analysis uses repeated sampling to estimate populations. In its simplest form, a sample S_1 is taken, then replaced into the population. A second sample S_2 is taken, and the population can be estimated $P = \frac{|S_1| \times |S_2|}{|S_1 \cap S_2|}$.

For the capture-recapture model to be perfectly accurate, a number of assumptions must apply. Notably, the population must be homogeneous and closed (i.e., no new entries). These assumptions do not entirely hold for our analysis: some websites are more likely to appear in the search results than others, and websites can be added and removed frequently. Nonetheless, we have computed the capture-recapture estimate in order to get a first approximation of the greater population size. The results are given in Figure 7. Notably, the estimates for source infections and redirects generated by comparing the different samples are fairly close. Both predict that the true number of redirects to be near 500, and the number of source infections to be around 5 000-6 000. The estimates for the number of pharmacies is more divergent, with one predicting a population size of 2 523 and the other predicting 795.