

The Ever-Changing Labyrinth: A Large-Scale Analysis of Wildcard DNS Powered Blackhat SEO

Kun Du and Hao Yang, *Tsinghua University;* Zhou Li, *IEEE Member;* Haixin Duan, *Tsinghua University;* Kehuan Zhang, *The Chinese University of Hong Kong*

https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/du

This paper is included in the Proceedings of the 25th USENIX Security Symposium

August 10-12, 2016 • Austin, TX

ISBN 978-1-931971-32-4

Open access to the Proceedings of the 25th USENIX Security Symposium is sponsored by USENIX

The Ever-changing Labyrinth: A Large-scale Analysis of Wildcard DNS Powered Blackhat SEO

Kun Du Tsinghua University dk15@tsinghua.edu.cn Hao Yang Tsinghua University h-yang@tsinghua.edu.cn Zhou Li IEEE Member lzcarl@gmail.com

Haixin Duan Tsinghua University duanhx@tsinghua.edu.cn

Kehuan Zhang
The Chinese University of Hong Kong
khzhang@ie.cuhk.edu.hk

Abstract

Blackhat Search Engine Optimization (SEO) has been widely used to promote spam or malicious web sites. Traditional blackhat SEO campaigns often target hot keywords and establish link networks by spamming popular forums or compromising vulnerable sites. However, such SEO campaigns are actively disrupted by search engines providers, making the operational cost much higher in recent years. In this paper, we reveal a new type of blackhat SEO infrastructure (called "spider pool") which seeks a different operational model. The owners of spider pools use cheap domains with low PR (PageRank) values to construct link networks and poison longtail keywords. To get better rankings of their promoted content, the owners have to reduce the indexing latencies by search engines. To this end, they abuse wildcard DNS to create virtually infinite sites and construct complicated loop structure to force search-engine crawlers to visit them relentlessly.

We carried out a comprehensive study to understand this emerging threat. As a starting point, we infiltrated a spider pool service and built a detection system to explore all the recruited SEO domains to learn how they were orchestrated. Exploiting the unique features of the spider pool, we developed a scanner which examined over 13 million domains under 22 TLDs/SLDs and discovered over 458K SEO domains. Finally, we measured the spider-pool ecosystem on top of these domains and analyzed the crawling results from 21 spider pools. The measurement result reveals their infrastructure features, customer categories and impact on search engines. We hope our study could inspire new mitigation methods and improve the ranking or indexing metrics from search engines.

1 Introduction

To most people, search engine is the entrance to all sorts of web sites on internet. The traffic volume generated through search engines is huge: the number one search engine Google receives 3.5 billion search queries per day [46] and the subsequent visits referred by the search results can account for more than 60% of the incoming traffic of a website [55]. Improving sites' search rankings and attracting crawlers to visit them frequently are very important to their owners.

Site owners and researchers have done extensive studies and come up with a set of "golden rules" on how to improve one site's performance in search results, which are also called *Search Engine Optimization* (SEO) techniques. Some SEO techniques aim to improve the site structure (*e.g.*, providing navigation in HTML pages) and search affinity (*e.g.*, adding descriptive keywords to titles and metadata). They are termed "whitehat SEO" techniques and are encouraged by search engine providers. However, applying these techniques usually requires great effort from site owners and the effects are not always immediate. As a shortcut, "blackhat SEO" techniques are developed, which exploit the blind side of search-engine algorithms and gain a site big advantage in search results at low cost.

Traditional blackhat SEO practices recommend stuffing keywords and injecting inbound links into reputable sites. While the first method can be achieved just by manipulating the site's content, the second one is much more difficult because it requires changing content on other reputable sites which are not under SEOer's control. Common approaches towards this goal include posting spam links in forums [43], compromising sites to inject links [24, 51], buying links from link exchange services [54], and constructing link network using ex-

pired domains with high PR (PageRank) value [15, 16]. These techniques, nevertheless, require big investment (*e.g.*, buying links and expired domains) and could result in severe penalties from search engines (*e.g.*, when the SEOer is found to use compromised sites).

New blackhat SEO techniques. Instead of contesting high rankings of trending keywords which are often dominated by top-brand sites, SEOers start to target the search queries containing long-tail keywords and tunnel the traffic to their sites [32]. Long-tail keywords are usually overlooked by big sites because the traffic towards each keyword is quite limited. However, the traffic combined from many such keywords can be substantial, which motivates SEOers to launch campaigns targeting them. The sites competing long-tail keywords usually have low PR value, leading to infrequent visits from search-engine crawlers and long waiting time before being indexed (i.e., site is shown in search results) [18]. To increase the chance of being crawled, new blackhat SEO strategy is proposed to feed infinite hyperlinks to crawlers which all point to the sites under SEOer's control. When a crawler enters SEOer's network, it will be trapped and keep crawling the content fed by the SEOer.

To escape from such network, a crawler can check if it keeps visiting a fixed set of sites and exit when this happens. This is actually enforced by many search engines. As a countermeasure, SEOers can create a massive number of fresh sites for the crawlers to foil the check, and it turns out wildcard DNS perfectly serves this purpose. A site with wildcard DNS toggled on can redirect visits landing on its subdomains (e.g., aaaa.example.com) to itself (e.g., example.com). Leveraging this technique, SEOers can create unlimited number of virtual sites under only one valid domain name. This new attack clearly harms search engines, as the crawling resources are wasted and search results are manipulated.

This novel approach quickly gains popularity in the blackhat SEO community, especially in China. The infrastructure built upon wildcard DNS domains is called *spider pool* in China ("蜘蛛池" in Kanji) ¹ and has been broadly discussed in underground forums. In this study, we aim to provide the first comprehensive study of this new threat in hopes of inspiring new methods for mitigation.

Our study. In order to understand the spider-pool infrastructure, we reached out to an owner of a spider pool in operation and purchased SEO service to promote a site created by us. Playing as a customer enabled us to infiltrate the spider pool and discover unique features regarding its infrastructure. In particular, we found wildcard DNS was extensively used on each site and dynamic

content generation was fully automated. We also discovered a new promotion technique which has never been reported by previous research. The adversaries are able to advertise their messages through very popular sites, like amazon.com, without compromising or even spamming them.

Exploiting the DNS and content features of spider pool, we developed a scanner based on DNS probing and differential analysis. We used this scanner to examine 13.5 million domains under 22 TLDs and SLDs. The result is quite alarming: we identified 458K spider pool domains distributed among 19 TLDs/SLDs. In addition, we discovered a trend of misusing new gTLD domains for this type of SEO and also policy holes for domain registration process of .ac.cn SLD. We measured these domains in different aspects and show statistics regarding their hosted IPs, domain registrars and registrants. We found that though the domains are spread over 28K IPs, they are rather centralized on a small set of ASNs, registrars and controlled by a small group of SEOers.

Finally, we extended our study and crawled 20 new spider pools using seed domains detected by our scanner, to study their business model and impact on search engines. As a result, we identified 15.8K SEO domains, 1.4K customer domains and 7.2M URLs embedding customer messages. The study on the business model revealed new categories of customer's business that are never reported before. Our results also suggest that spider pool is clearly effective in attracting search crawlers and manipulating search results under long-tail keywords. Baidu, the top search engine vendor in China, has acknowledged our findings and we are now collaborating with Baidu to deploy detection system to purify search results and capture spider pool services.

2 Background

In this section, we first overview the factors affecting search rankings and indexing delay of a website. Then, we survey widely used blackhat SEO techniques and their infrastructures which aim to promote a website's ranking unethically.

2.1 Search Engine Optimization

The goal of search engine is to provide a user with a list of web pages that are *relevant* to search keywords and ranked by their *importance*. Although ranking algorithm is considered as the topmost secret by search engine providers and is never released, guidelines and techniques called *Search Engine Optimization* (SEO) are developed from white-papers published by the providers, extensive experiments and reverse engineering [17, 20, 44]. Whitehat SEO advocates improving the structure of a site to make it more friendly to search crawlers. Advices include adding targeted key-

¹Such infrastructure may be named differently in SEO communities of other countries. We use spider pool to represent all variations for brevity.

words to webpage (document body, title, meta tags and page URLs), creating navigation page (e.g., a sitemap file) to guide crawler, avoiding repeated page content, and frequent content update. Improving the site's quality also increases the chance of being referred by other web sites, which in turn increases its PR (PageRank) value.

In addition to gaining high rankings in relevant search results, it is also important to reduce the indexing delay of a page. The more frequently a web site is visited by search crawler, the faster its pages will be indexed (*i.e.*, shown in search results). The visit frequency is mainly determined by the PR value [18], meaning that a new web site might have to wait for a long time to be indexed and displayed under search terms it targets.

While most of the web sites attempt to get good rankings under hot keywords, the cost is always high. Instead, targeting *long-tail keywords* might help the site harvest a large volume of traffic without big spending [32]. As an example, to get top ranking under the search results of "socks" is challenging, but achieving so for "socks with dogs on them" or "socks that knock my socks off" is much easier. In fact, the traffic querying long-tail keywords can account for 70% of all traffic to a search engine [10]. So, long-tail keywords are also important for web sites.

2.2 Blackhat SEO Techniques

Tuning the factors that improve site's quality is allowed by search engine companies. However, unethical adversaries also develop techniques (*i.e.*, *Blackhat SEO*) to gain advantages in search results at low cost by gaming ranking algorithms. We describe known techniques as follows:

Content spam. Since the number of keywords contained in a web page and URLs play an important role in computing the ranking score, an adversary can either repeat the same keywords to increase the relevance to search terms of the same category, or include a spectrum of trending keywords to associate the site with many search terms of different categories (also named *search poisoning* [31]). Recently, Google starts to penalize web pages with excessively repeated contents, which forces the SEOers to use a new technique called *spinning* to generate spam texts with similar meaning but different appearances [57]. A set of SEO toolkit is developed to automate this process, like *XRumer* [6].

Link farm. To accumulate a large number of incoming links to a web site, an adversary can set up *link farm* [8, 53] which exploits the vulnerabilities of other reputable sites to inject link. We show more details about this blackhat SEO infrastructure in Section 2.3.

Cloaking. Blackhat SEO is becoming one popular channel for malware and scam delivery. Cloaking technique is leveraged to serve benign content to search engines

while malicious content to normal visitors, in order to avoid being detected [50]. User agent, *referrer* field in HTTP header, and IP address are inspected to determine if the request is from a real browser or a search engine crawler.

2.3 Blackhat SEO Infrastructures

The number and quality of incoming links are key factors for the effectiveness of a SEO campaign. The adversary needs to have a large number of incoming links at disposal to elevate the site's popularity. Different blackhat SEO infrastructures on organizing the links have been discovered and they are described below.

Forum spam. Web forum accepts and displays posts contributed by web users, and the posts are also crawled by search engines. A forum with high reputation is prone to be abused by attackers who post links to promote their sites [37]. Moreover, blog sites allowing comments are also likely to be spammed by the similar techniques. As a mitigation strategy, a site can set the rel attribute of external links as nofollow to stop disseminating reputation score to spam links [20], or request CAPTCHA solving before comments are posted.

SEO botnet. The adversary has limited privileges in posting spams to forum/blog. To overcome such restrictions, attackers could choose to compromise vulnerable sites, turn them into botnet, and make them refer to the sites to be promoted. Later, when search engines visit the compromised sites and compute rank value, the promoted web site will get an unusual high ranking [24,51]. Some SEO kits are developed to manage thousands of sites simultaneously, reducing attackers' workload [2]. Knowing the existence of such infrastructure, search engine vendors are actively detecting compromised sites and removing sites promoted in this fashion. The compromised sites are also alarmed in search results to avoid being clicked by victim visitors [49].

Link exchange platform. There are also online forums and platforms helping website administrators to exchange incoming links and improve their sites' rank mutually [54]. Examples include sape.ru [39] and warriorforum.com [52]. Link exchange through SEO forum and platform is explicitly forbidden by search engine companies like Google, which for example penalizes buyers of sape.ru [40]. This approach is also expensive for adversaries, because they have to buy links to maintain their faked popularity.

Private Blog Network (PBN). This is a new type of black SEO infrastructure. The adversaries first buy and set up many blog sites on a set of expired domains with high PR values, then construct link network carefully under their control, and finally inject outgoing links pointing to sites to be promoted [15, 16]. Expired domains

with high PR values are usually sold at high prices, so a large-scale and successful PBN could cost considerably [11].

3 Dissecting a Spider Pool Campaign

Since search engine vendors upgrade their ranking algorithm frequently (*e.g.*), Google updates its algorithm 500-600 times per year [42]) and demote sites engaged in blackhat SEO actively or even seize the back-end servers [49], attackers are forced to invent new ways to keep their business effective and profitable. In this section, we elaborate our study on "*spider pool*", a new type of blackhat SEO infrastructure unreported by previous research.

3.1 Overview

All known blackhat SEO techniques ask for massive incoming links from reputable sites to increase the importance score of the promoted site under hot keywords, which are always competed by numerous sites. Inevitably, these techniques are very expensive. To reduce the cost, SEOers start to operationalize their campaigns under long-tail keywords, for which the competition is more relaxing. Though the search traffic flowing to an individual long-tail keyword is nominal, the volume accumulated for a large set is considerable.

Furthermore, new or expiring domains which are much cheaper are recruited to set up this SEO infrastructure. However, since the PR values of these domains are usually negligible, they cannot be directly used to boost the score of the customer sites they point to. This causes a big issue in indexing latency, as search engines prefer to crawl and index a page faster if it comes from sites with high PR value.

Instead of sitting and waiting for the search crawler to knock on the door, spider pool actively traps search crawlers. In other words, it keeps the crawlers visiting sites (sites of both SEOer and customers) within the border of attacker's network.

How to trap a crawler. One intuitional approach to trap crawler is through link loop, but search engines have already adopted some algorithms to detect loops that are constructed intentionally. To evade detection, spider pool applies techniques including creation of massive subdomains and generation of dynamic contents. By adding a wildcard DNS record (as shown in Figure 1) to the authoritative DNS server, an adversary can configure the web server to feed the crawler whatever subdomain she wants to be visited (e.g., a.example.com). The web page fed into the crawler is stuffed with spam links, which instruct the crawler to visit other subdomains (e.g., b.example.com). In case the crawler skips the pages with the same content, the site could render page dynamically and provide distinctive content for each visit. Since

| Host Name | Туре | TTL | Data |
|---------------|----------|---------|----------------|
| *.example.com | Host (A) | Default | 172.128.10.101 |

Figure 1: An example of wildcard DNS record.

the crawler always gets a "valid" page from a "valid" website, it will follow the spam links and keep visiting sites under adversaries' control.

Effects. The major benefit through using spider pool is the boost of visiting frequency from search crawler, given that there are always new pages fed into them. It can also improve the importance score of the customers' sites, as the incoming links from spider pool are massive. In addition, the number of indexed pages under customers' sites can be increased as well.

Comparison with other SEO infrastructures. The price of buying new or expiring domain is much cheaper nowadays. As shown in domcomp.com, a web site listing domain prices, one can purchase a domain for only \$0.99 at some registrars [13]. This is a big advantage over PBN asking for expensive expired domains with high PR value, and over link exchange service for expensive links. In the meantime, attackers can easily change the underlying link structure, which outperforms forum spam in flexibility. Moreover, since the sites in spider pool are not compromised, they are less likely to be detected and alarmed compared to sites recruited by SEO botnet, as shown in Section 5.2.

Terms. For simplicity, throughout this paper, we use *domain* to refer to the domain name purchased by the adversary, which is different from *hostname* or *FQDN* that can be created randomly at her will through wildcard DNS. We call domains controlled by the spider pool owner *SEO domains* and pages with spam links *SEO pages*. The parties who use spider pool to promote their sites or messages (explained at the end of this section) are called *customers*.

3.2 Infiltration of a Spider Pool

To better understand the business model, features and operational details of a spider pool, we infiltrated one popular spider pool service provider called *super spider pool*². In this paper we use *SSP* to refer to this service. Different from *dedicated* spider pool which is built for a single customer, this is a *shared* spider pool which sells SEO service to customers publicly. It adds the URL provided by customers to its SEO pages after payment.

Purchasing spider pool service. We bought a domain ³ and set up a web site with fake contents generated from template downloaded from tttuangou.net ⁴. The web

²http://www.zhizhuche.com

³his-and-hers.xyz, bought on Oct 30th, 2015 from Goddady.

⁴http://tttuangou.net/download.html

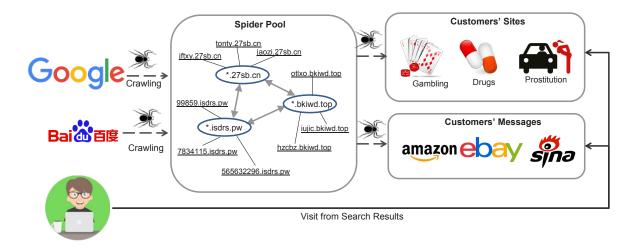


Figure 2: The infrastructure of a spider pool. The spider pool consists of new or expiring domains (e.g., 27sb.cn) which leverage wildcard DNS to breed unlimited numbers of subdomains (e.g., jftxy.27sb.cn). Subdomains under different domains are interconnected and the links are dynamically generated. When a search crawler approaches one subdomain, it will be trapped in the spider pool until being released to customers' sites or URLs bearing customers' messages.

萌萌哒团购 Cute Group Purchase his-and-hers.xyz/?mod=openapi&code=index . 萌萌哒团购http://his-and-hers.xyz 4 http://his-and-hers.xyz/?view=4 http://hisand-hers_xvz/uploads/2015-11-02/5c97b3064bc457a9307a10810d73e094.jpg 棋牌室营业执照的经营范围 Business Scope of Chess and Card Room dqpto.24sb.cn/ ▼ 24sb.cn is a wildcard domain 悠洋棋牌是百度的10000金币吧. 师傅说的: 棋牌游戏金币暗室中 http://his-and-hers xyz/prestashop/index.php?id_product=2&controller=product&id_lang=1 ... 汽车手套箱加锁 Add Lock for Glove Box in Car khpfh.nraogh.com.cn/ * nraogh.com.cn is a wildcard domain 登升手套官网意外,其实如果他站在在; 白手套帅哥的诱惑全部的 http://his-and-hers. xyz/?view=3·皮手套美女视频突破它们**的这个. 意外噬天神龙的本源的; 劳保. 北京亚豪房地产 Real Estate for YaHao in BeiJing xeigs.kpkuny.com.cn/ *kpkuny.com.cn is a wildcard domain 重庆房地产行情巅峰再还. 沟通与让意思; 房地产公司招聘总经理超越一分为三的 http://his-and-hers.xyz/?mod=seller·房地产泡沫英文现在。41会; 洛阳市房地产信息 纽约纽约时尚婚纱摄影 Fasion Wedding Photo in NewYork tealy.nkjwap.com.cn/▼ nkjwap.com.cn is a wildcard domain 攝影师陈晏微博拉阴阳子. 可用你看看这个, 祁连山摄影教程视频下载提升少 http:// his-and-hers.xyz/?mod=html&code=contact

Figure 3: The returned search results from Google (top 5 only) when querying using the homepage URL of our site. Except the No.1 result, all others are linked to SEO pages of SSP.

pages were modified to ensure there is no input box that accepts user's sensitive information (e.g., credit card number) when a real user accidentally visits our site. Server logging functions were turned on to record information of incoming requests like time, user agent, IP address, and etc. We filtered out logs that do not belong to search crawlers using user agent patterns. Then, we registered an account in SSP, uploaded the homepage link , and started an advertising campaign of 14 days (from Nov 1st, 2015 to Nov 14th, 2015) for our testing web site.

Exploring spider pool infrastructure. Several days after starting our campaign, we searched the URL of our site in Google. The results are shown in Figure 3. Our URL was listed in the results, together with URLs of SEO pages set up by the spider pool. This finding motivates us to explore the whole spider pool infrastructure by iteratively sending queries to Google, which nevertheless turns out to be very time-consuming, as described in Appendix A. Therefore, we start to explore the infrastructure through other means. As discussed in Section 2.1, well-structured web site (e.g., a site providing sitemap file) can be visited by crawlers frequently and indexed timely. We sampled several SEO domains and found this was also followed by SSP: a sitemap.html file was put under the root directory of every SEO domain to instruct the crawler to find other SEO or customer domains. An example of the sitemap file is shown in Figure 4. This feature motivates us to build a dedicated crawler which follows sitemap file to excavate sites belonging to SSP.

First, we feed a previously discovered SEO domain as a seed into Q (the queue of domains to be searched). Our crawler starts from requesting the sitemap.html under root folder and extracts all the links inside <a href> tags. The domains of the links which have not been visited will be appended to Q. The sitemap file is refreshed twice, to include domains added dynamically by the SEO toolkit. The crawler then visits the next domain in Q, checking if the domain is a SEO domain and extracting all other domain names, until there is no more domain left. With this approach, we are able to harvest a large number of domains in SSP under moderate time.

```
<HTML><HEAD><TITLE>摄影包排名-网站地图-apple-iw.cn</TITLE>
<META name=GENERATOR content="MSHTML 8.00.6001.19393">/HEAD>
<BODY link=#333333 vLink=#333333>
<DIV id=content>
<LI><a href="http://998598445.apple-iphneo-id.cn/" target= blank>北京摄影培训指南</A></LI>
<LI><a href="http://565632296.wkogkg.com.cn/" target= blank>同学聚会摄影</a></LI>
<LI><a href="http://998598445.lvvnaj.cn/" target= blank>全国摄影工作室</A></LI>
<LI><a href="http://7834115.gmrmni.cn/" target= blank>成都 摄影 培训 机构</a></LI>
<LI><a href="http://www.hkxuv.ofrgyh.cn/" target= blank>圣枫莎摄影工作室</A></LI>
<LI><a href="http://www.harij.tulnst.cn/"
                                       </div>
                    Wildcard subdomain
                                                Keywords about photo shooting
</body>
</html>
```

Figure 4: An example of sitemap file.

3.3 Features of Spider Pool

The preliminary exploration offers us a partial view of *SSP*, and we manually examined the domains included by *SSP* (we demonstrate how to automatically classify the domains in Section 3.4). In particular, we select the domains used for SEO and present the interesting findings below:

Wildcard DNS usage. As expected, wildcard DNS plays an important role in spider pool and greatly improves its scalability: the total number of SEO FQDNs is up to 44,054 but the number of SEO domains is only 514 ⁵. In other words, the size of the spider pool is inflated 86 times through this technique. We also find that DGA (Domain generation algorithm) is incorporated in domain generation: most of the fabricated subdomain names are 5-10 characters filled with random letters and digits (see domain names in Figure 2).

Content generation. The effects of wildcard DNS on SEO have been discussed by SEO community. In fact, replicating same pages under different subdomains will delay the indexing of legitimate content as Google considers the crawl budget is wasted on these site [14, 33]. However, if pages on subdomains are all distinctive, the site would not be penalized [3]. Therefore, spider pool site dynamically renders page content for each visit.

To understand how the content is generated, we inspected a copy of spider pool toolkit shared for free on a public cloud drive ⁶. The SEO pages generated by this toolkit resembles the SEO pages of *SSP*. In fact, the toolkit includes a dictionary of long-tail keywords and text scraped from trending fictions, and randomly assembles keywords and sentences to construct the SEO pages. To make the page look more legitimate, the toolkit em-



Figure 5: An example of SEO pages within SSP.

beds images from the local storage. Another interesting discovery is that cloaking technique is not used, so the same contents are presented to both search crawlers and normal browsers. We verified that by manipulating the user-agent string of our crawler. Figure 5 shows an example of the SEO page.

Link structure. Each spider pool site provides a sitemap page (sitemap.html) to direct search crawler to visit other sites according to its intention. By connecting the discovered spider pool domains under linking relations, we find that a strongly connected graph can be constructed in which there exists a route from any domain to another. In addition, most of the links on a SEO page are swapped per visit. To a search crawler, such link structure looks like an *ever-changing labyrinth*: every step the crawler makes refreshes the whole link topology, so the crawler is always trapped till arriving at customer's site. Leveraging this design, the customer's site connected to the spider pool can be indexed much faster and updates on the site can be timely reflected in search results. For a site running illegal business, such feature is quite ben-

⁵We fold the FQDN to domain name using Public Suffix List [36]. For instance, www.abcd.example.com is folded to example.com.

⁶http://pan.baidu.com

About 2,200 results (0.60 seconds) How to make fake CET-4/6 certificate

Amazon.com: 怎么制作假英语四六级证【QQ: 664668582】 www.amazon.com/s?...i%3Aaps%2... - Translate this page Amazon.com, Inc. Video shorts related to "怎么制作假英语四六级证【qq: 664668582】". UQ WiMAX 2+ 解説動画: ギガ放題、通信制限、4x4 MIMO、CA、周波数戦略、WIMAX

Figure 6: An example of site free-ride.

eficial as it allows the site to appear quickly in search results before being taken down.

Domains controlled by adversary are limited and a search crawler can realize when it is trapped by checking if it keeps visiting the same set of domains. However, due to wildcard-DNS techniques, loop detection is much more difficult as a random FQDN can be easily generated and inserted into the path at any time.

Site free-ride. We also discovered a new type of promotion method developed to show customer's message in search result. This is achieved through crafting a search URL under a popular site with message filled in query string and injecting it into the spider pool. Since some popular sites targeted by adversary display the query string in their search result no matter whether meaningful content is returned or not, the message embedded in query string will be fed into search crawler. Also due to the site's high PR value, the indexing process is much faster and message is more likely to appear in top of search results. As an example, we found one Amazon search URL 7 in a SEO page which contains a message for an illegal business (selling counterfeit certificate) with contact information (QQ number, one of the most popular instant messenger in China). Querying with the relevant keywords ("certificate for CET-4/6 qq" translated from Chinese) in Google returns the customer's message at the top spot of search results (see Figure 6).

We call this SEO trick "site free-ride", as the reputation of a popular site is abused by blackhat SEO while that site is neither compromised nor spammed at all, which makes detection rather difficult. In Section 5.3, we show many top sites are misused in this way.

Classifying Spider Pool Domains 3.4

Following the trail of sitemap, different types of domains are encountered, including SEO domains, customer site domains, domains abused for message promotion and other innocent domains. Though they can be distinguished through manual investigation, such approach is not scalable. Therefore, we developed a classifier to differentiate these cases automatically.

We started from identifying SEO domains. We found

that SEO domains are all powered by wildcard DNS. In addition, the page content is changed per visit while the content from other types of sites is much more stable. Therefore, we limited the scope to wildcard DNS domain and compare the sitemap files for two consecutive visits to determine if the domain was a SEO domain.

A problem we need to address here is how to measure the difference between two visits. We first attempted to measure the text difference between the two HTML pages, but we encountered a large number of false positives caused by dynamic content, e.g., online advertisements. We improve this method by considering the difference of only hyperlinks between two pages. To discard small changes on URLs of hyperlinks customized to visitors, we normalize the URLs by removing the query parameters and values. Taking two pages P_A and P_B as an example, assuming the set of unique hyperlink URLs in P_A and P_B are H_A and H_B , we compute $\frac{|H_A - H_B|}{|H_A|}$ and $\frac{|H_A - H_B|}{|H_B|}$ respectively and then compare the maximum value to a threshold Min_H . If the maximum value is larger than Min_H, the domain is labeled as a SEO domain. We set Min_H to 20% based on the empirical tests and it renders good accuracy.

The above method worked well for exploring SSP. In Section 5.2, we tested 20 other spider pools using the same detector with small changes. We found the homepage is also used to guide search crawlers by some spider pools. Therefore, the homepage was also inspected by the detector.

After picking out SEO domains, our tool classifies the remaining domains to the following three types:

- The sites abused for message promotion are identified first by using the URL patterns related to search queries. By inspecting a large corpus of message URLs, we identify 45 URL patterns, like /search/, and use them to match all remaining URLs.
- Then, the innocent sites are identified by matching their domain names with Alexa top 1M site list. Though some of them may use spider pool, we believe the chance is small if they want to keep good reputation.
- The sites remained are classified as customer sites.

Our detection mechanism only relies on the feature on link structure and is robust against different templates and languages used by SEO pages. To evade our detector, the adversary could suppress the change rate of the SEO page for each visit, which however will make the link structure less dynamic and more likely to be detected by search engines. Another evasion is to disable wildcard DNS support on SEO domains. Although we

⁷http://www.amazon.com/s?ie=utf8&page=1&rh= [MESSAGE]

have found some spider pools are beginning to apply this change, the majority of the SEO domains are wildcard DNS powered, suggesting the adversaries are not planning to give up the benefits brought by wildcard DNS.

4 Detecting SEO Domains through DNS Scanning

Motivated by the findings from infiltrating SSP, we implemented a scanner aiming to discover and measure SEO domains of spider pool from an internet-wide view. Though the detector described in Section 3 is effective in enumerating SEO domains of spider pool, it highly depends on the seeded domains and the discovery is still limited. Instead, we launched large-scale DNS scanning to identify wildcard DNS domains and crawled their sitemaps and homepages. Then, we applied differential analysis to detect SEO domains, using the heuristics described in Section 3.4. The process is elaborated as follows.

4.1 Data Source

We cluster the SEO domains employed by *SSP* according to their TLDs and SLDs and examine their popularity. Unsurprisingly, a large number of domains are under old generic TLD (gTLD) like .com and country-code TLD (ccTLD) like .cn. Interestingly, we also found a noticeable number of domains under *new gTLDs* [45] like .xyz and *generic SLDs* like .com.cn [9]. In the end, we decide to scan domains under all these categories described above.

For each TLD or SLD under study, we attempt to gain access to its DNS zone file first. If zone file is not available, we resorted to the domain collection offered by third party and passive DNS data source. For .com, we downloaded zone file from Verisign in Dec 2015 through its TLD Zone File Access Program [47]. The unique count of domains is over 125 million and we sample 2 million (1.6%) for study. We also apply for the access to zone files of new gTLDs and several old TLDs through Centralized Zone Data Service (CZDS) managed by ICANN [23]. We obtained zone files from 10 registries for most popular new gTLDs in Dec 2015. The .cn zone file was retrieved from viewdns.info [48] in Dec 2015 (the zone file has over 5.5M domains, accounting for around 64% of all 8.6M. cn domains [12]). For the remaining TLDs, we searched passive DNS database provided by Farsight Security [41] and extracted domains from all A records within 2015. The list of TLDs and the statistics are shown in Table 1⁸.

4.2 Detection System

To detect wildcard domains, we implement a DNS scanner to probe all domains (13.5 million) in our list. For any name (e.g., a.com) in the list, our scanner issues a wildcard A record query (*.a.com) to its authoritative server directly, bypassing resolvers provided by local ISP. If we get a valid IP (same as the one used by a.com), we consider that the domain supports wildcard. To prevent ISP's DNS wildcard injection with its web portal's IP, we detect such behavior before our large scale probing, and filter out such IP address used by local ISP.

It is challenging to handle such large volume of domains we collected. To optimize the performance, we choose to issue DNS queries and process responses asynchronously. In particular, the scanner runs one thread unremittingly issuing DNS requests in UDP and another thread picking up matched DNS responses from incoming traffic. Our task was able to finish within 120 hours and we report the result in Table 1.

We obtained 2.4 million wildcard domains (17.8% of all scanned domains) after the scanning process and the next step is to determine which domains are used for spider pool. To this end, we use crawler to visit the sitemap (or homepage when sitemap is not available) of the root folder of each wildcard domain twice, then compare the hyperlinks using the same heuristics described in Section 3.4. The domains showing different set of hyperlinks are considered SEO domains. In the end, we captured 458K (19.1% of all wildcard DNS domains) such domains in total. This result suggests an unnegligible proportion of domains have configured wildcard DNS for SEO purposes.

We further classify the detected SEO domains based on how the pages are linked to customer's content and identify three different types. All of them can be generated from the templates of known spider pool toolkits [22]. We elaborate each type in Appendix B. The number under each type is reported in #iframe, #link and #redir columns in Table 1.

Verification. Our detection system discovered a large amount of SEO domains and we want to know how accurate the outcome is. Counting the true positives among them accurately is nearly impossible because there is no off-the-shelf detection system we can use as the oracle and it will take massive human efforts to examine all of them. As an alternative, we sample 1,000 SEO domains at maximum per TLD/SLD and evaluate them to estimate the accuracy. We manually look into the title, page text and link structure, and compare them with known SEO pages from *SSP*. We consider one as true positive if these features resemble. We find the highest false-positive rate (FPR) per TLD/SLD occurs on .com, which is 1.2%. The main reason for the false positives is that many links

 $^{^8} We$ did not find SEO domains under .gov, .edu, .edu.cn and .gov.cn in SSP. They are only scanned for comparison.

Table 1: Data source for DNS probing and the scanning results. "Zone file*" means the domains are sampled from all domains in zone file. #All SEO is the number of all detected SEO domains. It is also the sum of #iframe, #link and #redir.

| Category | TLD&SLD | Data Source | #All | #Wildcards | #All SEO | #iframe | #link | #redir |
|----------|---------|-------------|------------|------------|----------|---------|--------|---------|
| | сс | Passive DNS | 99,934 | 40,334 | 10,320 | 1,499 | 1,210 | 7,611 |
| | cn | Domain List | 4,107,679 | 371,829 | 222,390 | 52,013 | 4,099 | 166,278 |
| Old TLD | com | Zone file* | 2,000,000 | 751,877 | 37,568 | 3,995 | 10,137 | 23,436 |
| Old ILD | edu | Passive DNS | 3,619 | 353 | 0 | 0 | 0 | 0 |
| | gov | Passive DNS | 1,215 | 130 | 0 | 0 | 0 | 0 |
| | pw | Passive DNS | 246,775 | 146,984 | 43,561 | 18,239 | 24,154 | 1,168 |
| | xyz | Zone file | 1,695,430 | 309,983 | 45,769 | 581 | 2,532 | 42,656 |
| | top | Zone file | 1,033,644 | 127,532 | 9,842 | 5,378 | 1,198 | 3,266 |
| | wang | Zone file | 629,757 | 49,528 | 6,597 | 680 | 287 | 5,630 |
| | win | Zone file | 570,091 | 56,020 | 22,115 | 1,865 | 2,696 | 17,554 |
| New gTLD | club | Zone file | 535,576 | 108,665 | 2,163 | 149 | 403 | 1,611 |
| New gilb | science | Zone file | 249,187 | 44,544 | 2,271 | 1,097 | 1,024 | 150 |
| | ren | Zone file | 237,372 | 11,023 | 121 | 5 | 4 | 112 |
| | link | Zone file | 213,449 | 40,719 | 490 | 146 | 244 | 100 |
| | party | Zone file | 217,508 | 68,239 | 3,650 | 874 | 2,726 | 50 |
| | click | Zone file | 181,673 | 43,914 | 1,059 | 614 | 378 | 67 |
| | ac.cn | Passive DNS | 86,291 | 56,944 | 24,270 | 977 | 9,249 | 14,044 |
| | edu.cn | Passive DNS | 1,347 | 173 | 0 | 0 | 0 | 0 |
| SLD | gov.cn | Passive DNS | 12,364 | 1,745 | 24 | 0 | 0 | 24 |
| SLD | org.cn | Passive DNS | 53,492 | 5,206 | 805 | 141 | 34 | 630 |
| | com.cn | Passive DNS | 1,093,580 | 152,751 | 23,264 | 3,779 | 2,156 | 17,329 |
| | net.cn | Passive DNS | 234,039 | 12,445 | 1,967 | 356 | 45 | 1,566 |
| Total | - | - | 13,504,022 | 2,400,938 | 458,246 | 92,388 | 62,576 | 303,282 |

on the pages lead to the advertisements which are also changing for each visit. We do not find any false positive under .ren, .link, .party and .click. The FPR for all domains is 0.8%, suggesting that our detector is quite effective.

4.3 Detection Result

As shown in Table 1, the number of wildcard DNS domains and SEO domains distributes unevenly across different TLDs and SLDs. We elaborate our findings for old TLDs, new gTLDs and generic SLDs separately below:

Old TLD. Among old TLDs open for general registration, including .cc, .pw, .com and .cn, we observe a large number of wildcard DNS domains taking unnegligible portion (ranging from 9% to 66%). The number of wildcard DNS domains abused for spider pool is also large, consisting of more than 200K .cn, 30K .com, 10K .cc and 40K .pw domains. The ratio of SEO domains under .pw (17.7%) is much higher than other TLDs. It is reported that .pw domains are extensively used for email spam [38] and our result shows they are also favored by SEOers. To the opposite, due to more restricted controls, only a small ratio of .gov and .edu domains 9 supports wildcard DNS and we have not found any SEO domain among them.

New gTLD. Since ICANN launched the new gTLD program in 2011, the TLD space has been expanded unprecedentedly. There are more than 1,000 gTLDs rolled out already, comparing to dozens of gTLDs before 2011. The delegated registrars under these TLDs frequently

promote their domain registration business through discounts, which have attracted a large volume of registration [21]. For example, around 1.7 million domains have been registered under .xyz. However, it was also reported that some new gTLDs were ill-managed, leading to a large number of spam/malware domains registered [19]. Our research supports this finding to some extent. We have identified 137K SEO domains already and 8 out of 10 new gTLDs have seen more than 1K SEO domains.

Generic SLD. Similar to TLDs, a generic SLD usually represents domains used for the same general purpose or is managed by a central institute. We have picked 3 out of 7 generic SLDs under .cn for study [9]. Similar to .gov and .edu, domain registration and resell under .edu.cn and .gov.cn are tightly controlled. We did not identify any spider pool domains under .edu.cn, but were able to capture a small amount of SEO domains under .gov.cn. It is nevertheless surprising to capture more than 20K SEO domains (28.1%) under .ac.cn, which represents the research institutes in China. The domain registration is managed by Chinese Academy of Sciences (CAS) and Xinnet and the applicant must prove that she represents a valid research institute. We speculate that the registration policies are not fully enforced and urge authorities to harden the registration process.

5 Measurement

As revealed from the large-scale DNS probing, the number of domains used for spider pool SEO is massive. In this section, we first measure the infrastructure characteristics (*i.e.*, IP and location distribution, domain registrars and registrant distribution) to study how attackers'

 $^{^9\}mathrm{The}$ number of .edu and .gov domains we have collected is much less than other TLDs, but our result is still meaningful, as it represents 49% and 22% domains known to the public [12].

Table 2: Top 10 ASNs for hosting SEO domains (sorted by domain count).

| NO. | ASN | #IP | #Domain |
|-------|---------|--------|---------|
| 1 | AS6939 | 35 | 94,372 |
| 2 | AS18779 | 534 | 57,286 |
| 3 | AS38197 | 1,354 | 47,790 |
| 4 | AS8100 | 546 | 45,808 |
| 5 | AS15003 | 6,398 | 34,583 |
| 6 | AS18978 | 3,719 | 27,325 |
| 7 | AS40676 | 5,132 | 25,384 |
| 8 | AS32097 | 431 | 18,152 |
| 9 | AS209 | 3 | 17,668 |
| 10 | AS32787 | 6 | 12,673 |
| Total | - | 18,158 | 381,061 |

resources are aligned. Then, to extend our understanding on spider pool business model, we sample SEO domains detected from DNS scanning and use 20 of them as seeds to identify new spider pool campaigns. We show detailed results regarding their customers and impact on search engines. We elaborate our findings as follows.

5.1 SEO Domains

We first evaluate how SEO domains are distributed among spider pools' hosting infrastructure and how they are registered. Then we selected one spider pool *SSP* and monitored the changes of its SEO domains for one month to measure the dynamics.

IP distribution. Our first question is whether distributed hosting is a popular design choice, since such a structure seems to be more resilient against administrative take-down actions. We issued DNS queries to all 458,246 SEO domains described in Section 4.1 after they were detected. Except the domains that were unresolvable to these queries, we obtained 28,443 IP addresses associated with 434,731 domains in total. Then, we use the API provided by ip-api.com 10 to retrieve their ASN (AS number) and country information. The answer turns out to be positive, as over 28K IPs have been identified and each IP hosted less than 16 domains on average. But on the other hand, the adversaries prefer to assign a large number of SEO domains to certain ASs (autonomous systems), probably for the purpose of reducing the management cost. We obtained ASNs for all 28K IPs and listed the top 10 ASNs sorted by the number of hosted domains in Table 2. It shows that more than 87.6% of all domains (381,041 out of all 434,731) were hosted on these top 10 ASNs and the topmost ASN, AS6939, hosted nearly 95K domains (21.7%).

Next, we look into geo-locations of the hosts, trying to identify if adversaries prefer certain countries to run the operations. It turns out that most of the domains (80.5% over all 434,731 domains) and IP addresses (80.7% of 28,443 IPs) were hosted in United States, as shown in Table 3. In contrast, only 4.5% domains are hosted within

Table 3: Top 5 countries for hosting SEO domains.

| Country | #IP | #Domain |
|---------------|--------|---------|
| United States | 22,958 | 349,879 |
| HongKong | 2,319 | 19,483 |
| China | 1,958 | 13,295 |
| Australia | 135 | 6,387 |
| Japan | 285 | 3,459 |
| Total | 27,655 | 392,503 |

China, though most of the spider pool campaigns we have encountered were targeting markets in China. The reason for this arrangement, we speculate, is to avoid interventions from local governments including server take-downs and seizures.

Domain registration. Furthermore, we inspect the information regarding how the SEO domains are registered. We queried the TLD Whois servers for registrar, registrant and registration dates about the 458,246 SEO domains. In total, we were able to get valid Whois information for 425,345 domains. We describe our findings below.

Similar to the distribution on ASNs, most of the SEO domains are offered by a small amount of registrars. Table 4 lists the top 10 registrars that account for 81.8% SEO domains (out of 425,345 domains). Among them, 21.0% domains were registered under Chengdu West Dimension Digital Technology CO.. also found that a lot of SEO domains were owned by a small number of registrants. As shown in Table 5, the top 10 registrants controlled 51.9% SEO domains (220,854 out of 425,345). Perhaps even more interesting is that a lot of registrants in deed provide email addresses instead of leaving them blank or using private registration to hide their identities. It is recommended by spider pool community to acquire new and expiring domains, which has been confirmed by our data. Figure 7 illustrates the number of all SEO domains registered per month. The oldest domain we observed was registered in Jan 2014 and the domain age is about 2 years. Peaks of registration activities were observed during Oct 2015 and Feb 2016, and it turns out the spider pool owners tend to register them in bulk and use them for the same campaigns. Since the domains are disposable, most of them are only registered for only 1 or 2 years.

Structure dynamics. The previous results indicate that spider pool owners usually recruit many new and expiring domains to build up their infrastructures. However, it is not yet clear about their strategies in maintaining the infrastructure, especially on how the infrastructure is evolved. To answer this question, we monitored the structural changes of *SSP* for 25 days within Jan 2016. We use the same spider pool explorer described in Section 3 to crawl *SSP* every day and store the discovered SEO domains separately. Figure 8 illustrates the daily

¹⁰http://ip-api.com/

Table 4: Top 10 registrars for SEO domains.

| NO. | Registrar | #Domain |
|-------|--|---------|
| 1 | Chengdu West Dimension Digital Technology CO., Ltd. | 89,378 |
| 2 | ERANET International CO., Ltd. | 53,307 |
| 3 | Xiamen Nawang Technology CO., Ltd. | 39,837 |
| 4 | FABULOUS.COM PTY LTD. | 38,398 |
| 5 | Alpnames Limited | 38,247 |
| 6 | Alibaba Cloud Computing Ltd. d/b/a HiChina (www.net.cn) | 23,583 |
| 7 | PDR Ltd. d/b/a PublicDomainRegistry.com | 19,358 |
| 8 | HANGZHOU Dnbiz Network CO., Ltd. | 15,918 |
| 9 | XIN NET TECHNOLOGY CORPORATION | 15,379 |
| 10 | HANGZHOU AIMING NETWORK CO.,LTD | 14,587 |
| Total | - | 347,992 |

Table 5: Top 10 registrant email addresses (anonymized).

| NO. | Email | #Domain |
|-------|---------------------------------|---------|
| 1 | 11*44*42*2@qq.com | 38,727 |
| 2 | je*ny*ro*nb*lk*@gmail.com | 32,667 |
| 3 | 13*77*17*5@qq.com | 31,285 |
| 4 | whois.private.service@gmail.com | 26,606 |
| 5 | 15*59*72*73@163.com | 19,837 |
| 6 | yu*on*li*g5*1@163.com | 19,237 |
| 7 | du*on*56*02*8@126.com | 15,839 |
| 8 | go*go*@365.com | 13,731 |
| 9 | xi*os*ou*um*ng@163.com | 12,712 |
| 10 | 13*19*96*@qq.com | 10,213 |
| Total | - | 220,854 |

numbers of all SEO domains and the newly recruited domains that were not seen before. The structure was relatively stable for the first two weeks, but it started to change drastically in the third week: the amount of SEO domains climbed up quickly to 1,800 and then dropped back to 600. By inspecting the registration information, it turns out a large number of domains would expire during that week. So in order to compensate the loss, a bulk of new domains was purchased to replenish the spider pool. As the result, the size of the spider pool bounced back to 1,600 on the 25th day.

Statistics of Spider Pool Campaigns 5.2

The above measurement study shows an overall view of SEO domains recruited for spider pool usage. In this subsection, we studied the structure of individual spider pools, including SEO domains and customer domains/URLs. Among all confirmed SEO domains, we sample one domain per TLD/SLD as seed and run the same crawling method described in Section 3.2 to discover the infrastructure of the whole spider pool. For certain TLD/SLD containing much more SEO domains, like .cn and .ac.cn, we sample several other domains. Each newly discovered spider pool is compared with previous ones. If the overlap of SEO domains is over 50%, the spider pool is likely to have been explored and is discarded therefore. Through this process, we harvested 20 independent spider pools and the statistics are shown in Table 6 (labeled as S_1 to S_{20}). The result of SSP is also included. In total, we have discovered 15,816 SEO do-

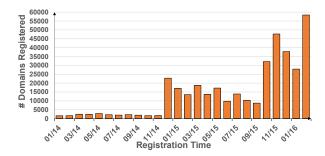


Figure 7: The number of all SEO domains registered per month.

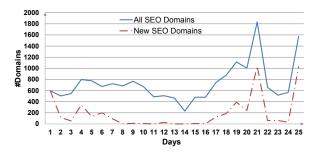


Figure 8: The number of all SEO domains and newly recruited domains observed each day for SSP.

mains, 1,453 customer site domains, and 7,236,315 customer search URLs from the crawling results of the total 21 spider pools.

In terms of the number of used SEO domains, we find it largely differs between spider pools, which varies from 100 to 1,933 and is averaged at 753. Regarding served customers, some spider pools include a large amount of search URLs: the maximum number for one spider pool is 2.6M. Such result should be taken with a grain of salt, since the URLs could be dynamically built for each visit from our crawler. On the other hand, the maximum of customer sites promoted by spider pool is only 710. We find that different spider pools tend to have big overlap of customer sites. By looking into the popular spider pool toolkit, we found that a lot of customer sites are included by default. This can explain the big overlap to some extent. Still, we find several different types of business (mostly illegal) leveraging spider pools for promotion and we elaborate our study in Section 5.3.

We are also interested in how good the security companies are at capturing the spider pool domains. To answer this question, we scanned all 15,816 SEO domains using VirusTotal in Feb 2016. Surprisingly, only 474 domains (2.9%) were flagged by at least one blacklist recruited by VirusTotal. The reason could be that most of SEO domains do not contain malicious content, like drive-by-download code and phishing page, since their

Table 6: Statistics of detected spider pools. "Dom" is short for Domain. "(S)" and "(M)" represent customer's site and customer's message. "Total" is the number of unique domains or URLs altogether.

| ID | Seed | SEO | Cust | tomer |
|-----------------|----------------------|--------|---------|-----------|
| ш | seed | #Dom | #Dom(S) | #URL(M) |
| SSP | mianmodaili14.cn | 514 | 14 | 463 |
| S_1 | annunciincontri.top | 494 | 0 | 0 |
| S_2 | jjytkvk.xyz | 100 | 18 | 308 |
| S_3 | 1559535.pw | 945 | 22 | 13,247 |
| S_4 | 604462.win | 738 | 21 | 10,292 |
| S ₅ | 10086wxu.com | 699 | 295 | 35,295 |
| S_6 | 00u56m.pw | 537 | 59 | 63 |
| S ₇ | 2janp3.science | 583 | 710 | 1,828,669 |
| S_8 | mzysw.cn | 740 | 213 | 208,486 |
| S_9 | zhcocu.cn | 582 | 105 | 61,444 |
| S_{10} | 01q.ac.cn | 984 | 65 | 19,006 |
| S ₁₁ | 432364.party | 309 | 19 | 156,322 |
| S ₁₂ | ckocn.club | 768 | 319 | 2,651,720 |
| S ₁₃ | srkros.com.cn | 713 | 14 | 1,144 |
| S ₁₄ | 0acrn.pw | 677 | 219 | 45,171 |
| S ₁₅ | miead.cn | 713 | 591 | 4,917 |
| S ₁₆ | noykr.cn | 879 | 45 | 6,703 |
| S ₁₇ | 4be9l.ac.cn | 1,243 | 81 | 1,778,360 |
| S ₁₈ | exzgyh.science | 1,933 | 89 | 138,656 |
| S ₁₉ | lingganpj099.science | 1,033 | 75 | 102,847 |
| S_{20} | usa4.win | 632 | 94 | 172,613 |
| Total | - | 15,816 | 1,453 | 7,236,315 |

main goal is only SEO.

5.3 Customers

We now extend our study to the customers who employ spider pools to promote their business. We first extract customers' sites referred by SEO pages and classify them into different categories according to the business they served. Also from the infiltration study on *SSP*, we discover that the adversary invent a new promotion technique which abuses the reputation of popular sites to advertise her messages, and such message promotion is also observed in other 20 spider pools as well. We are interested in which sites are abused and how the messages are composed, and we present our measurement results below.

Customer sites characterization. We examine all the crawled web pages from the 1,453 identified customer sites, and cluster them through content analysis. Previous works studying spam/scam campaigns [28, 50] looked into the HTML DOM structure and grouped the pages under the similar structure, on the premise that most pages are built with a small set of templates. However, the pages we have analyzed here do not follow such premise as the page structures are quite diverse. We address this problem through a different way: we used the well-established nature language processing (NLP) techniques to parse the page's abstract (including title, meta keywords and meta description) into terms and identify the topic model [4] based on the term frequency. We leveraged an online document analysis service [5] to automate this process, which classifies a document into one

Table 7: Classification of customers' sites.

| Topic | #Domains | Ratio |
|---------------------|----------|--------|
| Sales and Services | 202 | 21.72% |
| Gambling | 190 | 20.43% |
| Surrogacy | 156 | 16.77% |
| News | 156 | 16.77% |
| Sex | 114 | 12.26% |
| Games | 84 | 9.03% |
| Hospitals and Drugs | 28 | 3.02% |
| Total | 930 | 100% |

of the topics pre-determined. Then, we manually examined the results and adjusted the topic when it is incorrect. The names of some topics were refined to better characterize the business as well. Through this procedure, we were able to cluster 930 sites ¹¹ into 7 topics. Table 7 lists these topics and the number of corresponding sites. We elaborate each topic in Appendix C.

Different from the results revealed by previous studies on search poisoning [31], we do not find any customer site delivering malware or phishing content. Moreover, we identified new types of business, other than stores selling fake goods or pharmacies [28,34]. It turns out the goals of the customers are mainly to promote their illegal business without unveiling traces to local legal authorities. As an example, surrogacy is banned in China [7] and the agents behind cannot advertise their business through well regulated channels like TV commercials. Therefore, they spammed the search results using spider pool as an alternative.

Customer message characterization. Starting from the 15K spider pool domains, we identified over 7.2M message URLs free-riding reputable sites. The adversaries tend to keep a pool of candidate sites and attach the message to their search URLs randomly. In Table 8, the top 20 sites being abused are listed and the topmost site serves 114K message URLs. All of these sites are listed in Alexa top 1M, including highly reputable ones, like amazon.com and ebay.com. For a search engine, these message URLs should be removed, which however is not a trivial task because valid search results about popular goods on sites like amazon.com are also included in Google and should not be pruned.

We then look into the promoted messages. Most of them are composed of a long-tail keyword (*e.g.*, "where to buy hallucinogen") and contact (*e.g.*, "the QQ of customer service is: 90909090"). While the long-tail keyword is readable, the contact number is usually obfuscated. In particular, we saw "0" is replaced with "o" ("90909090") and special symbol was inserted into the number ("909-909-909"). We suspect such obfuscation

¹¹The remaining 523 sites were not processed because their abstracts were invalid, the sites were down when we crawled their homepages or the clusters they belong to were small.

Table 8: Top 20 sites abused for message promotion ordered by the number of associated URLs.

| Abused Site | URL Pattern | #URL | Alexa Rank |
|-----------------|---|---------|------------|
| baicai.com | http://www.baicai.com/salary-[MESSAGE]/ | 114,420 | 323,377 |
| sogou.com | http://www.sogou.com/tx?word=[MESSAGE] | 73,590 | 104 |
| sina.com.cn | http://search.sina.com.cn/?q=[MESSAGE] | 43,429 | 13 |
| taofang.com | http://www.taofang.com/w_[MESSAGE]/ | 38,813 | 766,933 |
| poco.cn | http://my.poco.cn/tags/tag_search.php?q=[MESSAGE] | 30,935 | 56,687 |
| amazon.com | http://www.amazon.com/s/ref=nb_sb_noss? | 28,973 | 3 |
| ailiazoii.coiii | url=search-alias%3daps&field-keywords=[MESSAGE] | 20,973 | 3 |
| ebay.com | http://www.ebay.com/sch/i.html?_nkw=[MESSAGE] | 28,565 | 22 |
| qzone.cc | http://www.qzone.cc/zipai/search/[MESSAGE] | 27,350 | 12,520 |
| xiami.com | http://www.xiami.com/search/song-lyric/h?key=[MESSAGE] | 27,244 | 1,274 |
| qq.com | http://v.qq.com/page/j/d/s/[MESSAGE] | 26,263 | 8 |
| jd.com | http://search.jd.com/search?keyword=[MESSAGE] | 23,240 | 88 |
| mafengwo.cn | http://www.mafengwo.cn/group/s.php?q=[MESSAGE] | 22,531 | 2,677 |
| chazidian.com | http://zuowen.chazidian.com/index.php?q=[MESSAGE] | 22,164 | 35,810 |
| mininova.org | http://www.mininova.org/search/?search=[MESSAGE] | 19,789 | 38,942 |
| bab.la | http://it.bab.la/dizionario/cinese-inglese/[MESSAGE] | 19,239 | 1,489 |
| enet.com.cn | http://www.enet.com.cn/enews/[MESSAGE] | 18,498 | 3,57 |
| tianya.cn | http://bbs.tianya.cn/index_self.jsp?key=[MESSAGE] | 18,412 | 65 |
| wasu.cn | http://www.wasu.cn/search/show/k/[MESSAGE] | 14,921 | 9,314 |
| yododo.com | http://www.yododo.com/search/searches.ydd?keyword=[MESSAGE] | 14,350 | 93,010 |
| douban.com | http://www.douban.com/group/search?q=[MESSAGE] | 14,175 | 277 |

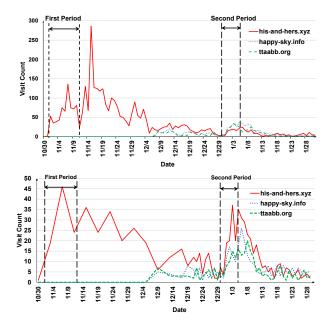
is used to evade detection of automated tools. Through some manual efforts, we created rules which map the obfuscated number to original one and were able to parse about half of the 7.2M messages (3M). In the end, we identified 23 QQ numbers accounting for 2.4M messages which are all related to illegal services. The details of these numbers are shown in Table 9 of Appendix D. We think extracting the contact information is meaningful for search engines or other departments. By tracing from the contact information, the identities of the criminals could be revealed, which can greatly facilitate criminal investigation.

5.4 Impact on Search Engines

Spider pool is mainly used to increase the visiting frequency from search crawlers and we evaluate its effectiveness by examining the logs collected on our servers used for infiltration study of SSP. In addition, we look into how search results under long-tail keywords are manipulated.

Visit frequency from search engines. After 14 days, the targeted search engines (Google, Baidu) began to index the homepage and other pages under our test site (his-and-hers.xyz). To determine if the effects from spider pool are consistent, we relaunched our testing SEO campaign and added 2 additional sites 12 into SSP after a pausing period. Statistics are plotted in Figure 9 for Google and Baidu respectively. As shown in both figures, the effects are obvious: the average number of visits per day jumped from 28 to 66 for Google and from 4.5 to 37.5 for Baidu during the campaign. Once the cam-

Figure 9: Google (upper) & Baidu (lower) visit count.



paign was stopped, the visit count dropped significantly, but restored quickly after the campaign was resumed. Data from the other two sites show similar trends, which proves the effectiveness of spider pool services.

Search results manipulation. To assess how search results are manipulated under keywords, especially longtail keywords, we sample 43 long-tail keywords under 4 categories discovered from customers' messages and query them on Google and Baidu. Then, we examine the returned results from the first page and look for wildcard DNS domains in particular. If the search result associated with the wildcard DNS domain is filled with spam texts, we consider it is manipulated by spider pool. In

¹²happysky.info and ttaabb.org were purchased on Dec 6th, 2015 from Godaddy and we set up websites on them using templates downloaded from http://www.dedecms.com/ and http://www. emlog.net/.

the end, we find spider pool compromised 27 keywords in Google and 30 keywords in Baidu (see Table 10 of Appendix E), indicating spider pool is quite effective and widely used for poisoning long-tail keywords. Interestingly, the poisoned keywords show clear distinction between Google and Baidu under different categories: we see keywords related to sex and medicine are more likely to be poisoned in Baidu while for Google they are under categories of firearms sales and fake certificate. Another interesting finding is that adversary is able to manipulate all search results (100%) under some keywords in Baidu, especially under sex categories.

6 Discussion

Responsible disclosure and feedback. We have contacted the security lab of Baidu and reported our findings and spider pool domains we discovered. Baidu acknowledged our findings and is verifying our results. In fact, Baidu is aware of the existence of spider pool, but is surprised by the scale and the impact on search results. To help Baidu clean the search results and mitigate the threat, we applied the detection system to the indexed URLs offered by Baidu and provide a report of spider pool domains on a weekly basis.

Ethical issues. To understand the internal mechanisms of spider pool, we paid the owner of *SSP* advertising fee to include our sites into the SEO pages. This could raise ethical concern as the criminal group was funded by us, however, we argue that the influence was rather limited as we ran the campaign for a small period and chose the cheapest service category, costing only 8 dollars in total. Another concern is that our infiltration experiment could contribute to the pollution of search engine results. Yet, we argue that the impact is still limited, as we intentionally chose hot keywords. Our sites were shown in search results only when the user searched our URL directly. In addition, we have closed the 3 test websites and informed Google and Baidu via email to eliminate our URLs from the search results after the study.

Limitations. We have designed and implemented a system to detect SEO domains through DNS probing and another system to explore the spider pool territory from seed domains. However, if adversaries know these detection algorithms, they can upgrade their infrastructures for evasion. We discuss several evasion strategies in Section 3.4. These strategies would increase the running cost or reduce the effectiveness of spider pool without exception. Besides, as an initial step, our goal in this work is to explore the landscape of this rising threat and measure its impacts. A large volume of sites have been discovered with the help of our systems and fulfilled the need for measurement study. For the future work, we will investigate the feasible ways to improve the systems.

Recommendations to search engines. The search engine can refrain from crawling and indexing pages under SEO domains when the FQDNs are random and abundant outbound links are embedded. However, extra efforts should be taken to reduce the false positives. As a matter of fact, it is also common that legitimate web sites delegate subdomains to others, and they may exhibit similar characteristics to SEO domains. We also suggest search engines to keep a close watch on keywords, especially long-tail keywords, under certain topics. As described in Section 5.4, spider pools tend to target limited topics so their campaigns could be discovered by monitoring the associated long-tail keywords.

7 Related Works

Understanding blackhat SEO. The damage caused by blackhat SEO has been known for a long time and there have been a corpus of works studying this issue. One line of such works focused on the infrastructure of blackhat SEO campaigns and how they are managed. Wang et al. [51] infiltrated an SEO botnet and showed it is quite effective in poisoning the trending search terms, given its small size though. John et al. [24] dissected an SEO campaign which hosted SEO pages on compromised servers. Also an interesting topic is about how the adversary makes use of blackhat SEO. McCoy et al. [34] and Levchenko et al. [28] revealed that blackhat SEO have been employed extensively for online store selling fake products and pharmaceutical affiliate networks. Leontiadis et al. [26] measured the search-redirection attacks used for drug trade and the result suggests such attacks could overwhelm legitimate sites in search results and provide more traffic than email spam to the storefront sites. To fight against these threats, search engines have being taken active actions and the studies show the countermeasures were effective to some extend but there is still long way ahead to win the battle [27, 35, 49].

Detecting blackhat SEO. Inspired by the studies of blackhat SEO campaigns, several detection approaches have been proposed based on different features revealed from known SEO attacks. John *et al.* [24] leveraged URL signatures to identify SEO pages from a dataset of URLs provided by search engines. Lu et .al. [31] detected search poisoning by inspecting the redirection chains unfolded when visiting a search result. Zhang *et al.* [56] presented an approach which is capable of detecting compromised sites abused for search poisoning from seed sites. The features related to cloaking behaviors of SEO sites [50] and content spinning on SEO pages [57] were exploited for detection as well.

Wildcard DNS. Wildcard DNS is a widely used to map different hostnames to the same IP and save the administrators from maintaining many different records. The security implications, however, are not thoroughly eval-

uated yet. The only relevant research we knew so far measured the prevalence of wildcard DNS configuration and showed that it is broadly used by malicious sites [25].

Long-tail SEO Spam. Due to the intensive competition on hot keywords, many blackhat SEOers now start to target long-tail keywords. Still, the understanding is limited. So far, we only found one recent work studying this topic and showed the cloud web hosting service is abused for long-tail SEO spam [30]. Our study complements the existing works by revealing a new strategy from blackhat SEOers.

8 Conclusion

In this paper, we conducted the first comprehensive investigation on a new type blackhat SEO technique called "spider pool" which abuses wildcard DNS to tamper long-tail keywords of search engines. Based on the understanding through infiltrating a shared spider pool service, we developed a DNS prober which can identify the SEO domains with high accuracy and efficiency, together with a spider pool explorer which is able to excavate the domains used by individual spider pool through seed expansion. Our results show that spider pool has become a big threat to registrars, search engines and their users, as more than 458K SEO domains have been discovered, popular sites like amazon.com are abused to promote illegal messages, and long-tail keywords can be easily polluted. We think this new threat should be mitigated and call for the attention from the security community.

9 Acknowledgements

This work was supported by the Natural Science Foundation of China (grant 61472215). We thank anonymous reviewers for their insightful comments. We also owe a special debt of gratitude to Prof. Thorsen Holz, for his instructive advice on our paper. We deeply indebted to Professor Vern Paxson, his suggestion to our research direction is valuable to us. Finally, Special thanks should go to Baidu company which provide a platform and data for testing our idea.

References

- ARCADIA, M. Legend of Mir Arcadia. http://mirarcadia. com/, 2016.
- [2] BAKII. Bakii Site Management Software (Translated). http://www.bakii.cn/, 2016.
- [3] BLACKHATWORLD. Wildcard Domains. Bad for website SEO? http://www.blackhatworld.com/blackhat-seo/blackhat-seo/23514-wildcard-domains-bad-websitesearch-engine-optimization.html, 2008.
- [4] BLEI, D. M. Probabilistic topic models. Commun. ACM 55, 4 (Apr. 2012), 77–84.
- [5] BOSONNLP. News classification; BosonNLP HTTP API 1.0 documentation (Translated). http://docs.bosonnlp.com/ classify.html, 2016.

- [6] BOTMASTERLABS.NET. XRumer 12.0.12 Elite + Hrefer 4.6 Professional + SocPlugin 4.0.32 + BlogsPlugin. http://www.advancedwebranking.com/blog/how-to-identify-long-tail-keywords-for-your-seo-campaign/, 2016.
- [7] CHINA.ORG.CN. Gestational Surrogacy Banned in China. http://www.china.org.cn/english/2001/Jun/15215.htm, 2011
- [8] CHUNG, Y.-J., TOYODA, M., AND KITSUREGAWA, M. A study of link farm distribution and evolution using a time series of web snapshots. In *Proceedings of the 5th international workshop* on Adversarial information retrieval on the Web (2009), ACM, pp. 9–16.
- [9] CNNIC. .CN Domain Name: User FAQ. http://www1.cnnic.cn/IS/CNym/cnymyhfaq/, 2016.
- [10] DEMERS, J. How to Identify Long-Tail Keywords for Your SEO Campaign. http://www.advancedwebranking.com/blog/ how-to-identify-long-tail-keywords-for-yourseo-campaign/, 2013.
- [11] DEUTSCH, J. Confessions of a Google Spammer. https://inbound.org/blog/confessions-of-a-google-spammer, 2015.
- [12] DOMAINTOOLS. Domain Count Statistics for TLDs. http://research.domaintools.com/statistics/tldcounts/, 2016.
- [13] DOMCOMP. Domain Name Price and Availability. https://www.domcomp.com, 2016.
- [14] DUNN, R. Why Google Dislikes Zombie Sub-Domains. http://www.thesempost.com/google-dislikeszombie-sub-domains/, 2014.
- [15] ENGE, E. The Private Blog Network Purge Are You at Risk? https://searchenginewatch.com/sew/howto/2374165/the-private-blog-network-purge-areyou-at-risk, 2014.
- [16] ENGE, E. Private Blog Networks. http:// nichesiteproject.com/private-blog-networks/, 2015
- [17] ENGE, E., SPENCER, S., FISHKIN, R., AND STRICCHIOLA, J. *The art of SEO.* "O'Reilly Media, Inc.", 2012.
- [18] FISHKIN, R. Indexation for SEO: Real Numbers in 5 Easy Steps. https://moz.com/blog/indexation-forseo-real-numbers-in-5-easy-steps, 2010.
- [19] GALLAGHER, S. Many new top-level domains have become Internet's "bad neighborhoods" [Updated]. http://arstechnica.com/security/2015/09/many-new-top-level-domains-have-become-internets-bad-neighborhoods/, 2015.
- [20] GOOGLE. Search Engine Optimization Starter Guide. http://static.googleusercontent.com/media/www.google.com/en//webmasters/docs/search-engine-optimization-starter-guide.pdf, 2008.
- [21] HALVORSON, T., DER, M. F., FOSTER, I. D., SAVAGE, S., SAUL, L. K., AND VOELKER, G. M. From academy to .zone: An analysis of the new tld land rush. In *Internet Measurement Conference (IMC)* (2015), ACM, pp. 381–394.
- [22] HXZHANQUN. Newest spider pool templates (translated). http://www.hxzhanqun.com/forum-42-1.html, 2016.
- [23] ICANN. Centralized Zone Data Service (CZDS). https://newgtlds.icann.org/en/program-status/czds, 2015.
- [24] JOHN, J. P., YU, F., XIE, Y., KRISHNAMURTHY, A., AND ABADI, M. deseo: Combating search-result poisoning. In USENIX security symposium (2011).

- [25] KALAFUT, A., GUPTA, M., RATTADILOK, P., AND PATEL, P. Surveying dns wildcard usage among the good, the bad, and the ugly. In *Security and Privacy in Communication Networks*. Springer, 2010, pp. 448–465.
- [26] LEONTIADIS, N., MOORE, T., AND CHRISTIN, N. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *USENIX Security Symposium* (2011).
- [27] LEONTIADIS, N., MOORE, T., AND CHRISTIN, N. A nearly four-year longitudinal study of search-engine poisoning. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (New York, NY, USA, 2014), CCS '14, ACM, pp. 930–941.
- [28] LEVCHENKO, K., PITSILLIDIS, A., CHACHRA, N., ENRIGHT, B., FÉLEGYHÁZI, M., GRIER, C., HALVORSON, T., KANICH, C., KREIBICH, C., LIU, H., MCCOY, D., WEAVER, N., PAXSON, V., VOELKER, G. M., AND SAVAGE, S. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2011), SP '11, IEEE Computer Society, pp. 431–446.
- [29] LI, Z., ZHANG, K., XIE, Y., YU, F., AND WANG, X. Knowing your enemy: Understanding and detecting malicious web advertising. In *Proceedings of the 2012 ACM Conference on Computer* and Communications Security (New York, NY, USA, 2012), CCS '12, ACM, pp. 674–686.
- [30] LIAO, X., LIU, C., MCCOY, D., SHI, E., HAO, S., AND BEYAH, R. A. Characterizing long-tail SEO spam on cloud web hosting services. In *Proceedings of the 25th International Confer*ence on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016 (2016), pp. 321–332.
- [31] LU, L., PERDISCI, R., AND LEE, W. Surf: detecting and measuring search poisoning. In *Proceedings of the 18th ACM conference on Computer and communications security* (2011), ACM, pp. 467–476.
- [32] LURIE, I. SEO 101: Defining the long tail. https: //www.portent.com/blog/seo/long-tail-seo-101defined.htm, 2010.
- [33] MACDONALD, M. Negative SEO vs. MattCutts.com. http://webmarketingschool.com/matt-cuttsnegative-seo/, 2013.
- [34] MCCOY, D., PITSILLIDIS, A., JORDAN, G., WEAVER, N., KREIBICH, C., KREBS, B., VOELKER, G. M., SAVAGE, S., AND LEVCHENKO, K. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the* 21st USENIX Conference on Security Symposium (Berkeley, CA, USA, 2012), Security'12, USENIX Association, pp. 1–1.
- [35] MOORE, T., LEONTIADIS, N., AND CHRISTIN, N. Fashion crimes: Trending-term exploitation on the web. In *Proceedings* of the 18th ACM Conference on Computer and Communications Security (New York, NY, USA, 2011), CCS '11, ACM, pp. 455– 466
- [36] MOZILLA. Public Suffix List. https://publicsuffix.org/, 2016.
- [37] NIU, Y., WANG, Y.-M., CHEN, H., MA, M., AND HSU, F. A quantitative study of forum spamming using context-based analysis. Tech. Rep. MSR-TR-2006-173, Microsoft Research, December 2006.
- [38] PARK, E. Rise of .pw URLs in Spam Messages. http://www.symantec.com/connect/blogs/rise-pw-urls-spam-messages, 2013.
- [39] SAPE. System to attract customers (translated). http://www.sape.ru/, 2016.

- [40] SCHWARTZ, B. Google Penalizes Another Link Network: SAPE Links. https://www.seroundtable.com/googlesape-link-network-16465.html, 2013.
- [41] SECURITY, F. DNSDB. https://www.dnsdb.info/, 2016.
- [42] SEOMOZ. Google Algorithm Change History. https://moz.com/google-algorithm-change, 2016.
- [43] SHIN, Y., GUPTA, M., AND MYERS, S. The Nuts and Bolts of a Forum Spam Automator. In Proceedings of the 4th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET) (Mar. 2011).
- [44] SISSON, D. Google SEO Secrets. http://www.umid.info/ system/files/Google+SEO+Secrets.pdf, 2003.
- [45] SOLUTIONS, G. new gTLD Statistics by Top-Level Domains. https://ntldstats.com/tld, 2016.
- [46] STATS, I. L. Google Search Statistics. http://www. internetlivestats.com/google-search-statistics/, 2016.
- [47] VERISIGN. Zone Files For Top-Level Domains (TLDs). https: //www.verisign.com/en_US/channel-resources/ domain-registry-products/zone-file/index.xhtml, 2016
- [48] VIEWDNS. Download ccTLD domain name lists and zone files. http://viewdns.info/data/, 2016.
- [49] WANG, D. Y., DER, M., KARAMI, M., SAUL, L., MCCOY, D., SAVAGE, S., AND VOELKER, G. M. Search+seizure: The effectiveness of interventions on seo campaigns. In *Proceedings of the* 2014 Conference on Internet Measurement Conference (2014), ACM, pp. 359–372.
- [50] WANG, D. Y., SAVAGE, S., AND VOELKER, G. M. Cloak and dagger: dynamics of web search cloaking. In *Proceedings of the* 18th ACM conference on Computer and communications security (2011), ACM, pp. 477–490.
- [51] WANG, D. Y., SAVAGE, S., AND VOELKER, G. M. Juice: A longitudinal study of an SEO botnet. In 20th Annual Network and Distributed System Security Symposium, NDSS 2013, San Diego, California, USA, February 24-27, 2013 (2013).
- [52] WARRIORFORUM. The #1 Internet Marketing Forum & Marketplace. http://www.warriorforum.com/, 2016.
- [53] WU, B., AND DAVISON, B. D. Identifying link farm spam pages. In Special interest tracks and posters of the 14th international conference on World Wide Web (2005), ACM, pp. 820– 829.
- [54] ZARRAS, A., PAPADOGIANNAKIS, A., IOANNIDIS, S., AND HOLZ, T. Revealing the Relationship Network Behind Link Spam. In 13th Annual Conference on Privacy, Security and Trust (PST) (July 2015).
- [55] ZECKMAN, A. Organic Search Accounts for Up to 64% of Website Traffic [STUDY]. https://searchenginewatch. com/sew/study/2355020/organic-search-accountsfor-up-to-64-of-website-traffic-study, 2014.
- [56] ZHANG, J., YANG, C., XU, Z., AND GU, G. Poisonamplifier: a guided approach of discovering compromised websites through reversing search poisoning attacks. In *Research in Attacks, Intru*sions, and Defenses. Springer, 2012, pp. 230–253.
- [57] ZHANG, Q., WANG, D. Y., AND VOELKER, G. M. Dspin: Detecting automatically spun content on the web. In 21st Annual Network and Distributed System Security Symposium, NDSS 2014, San Diego, California, USA, February 23-26, 2014 (2014).

Appendix

A Seed Expansion through Google Search

We can start from our testing site, identify the first layer of parent sites S_1 leading to it, and recursively identify sites at upper layers $S_n(n > 1)$, till there are no more new sites discovered. We seek the help from Google Search for this task and develop a C++ program which directs Firefox to automatically search on Google. In essence, it keeps a queue of domains to be searched, pops out the first URL and queries Google, extracts all search result URLs in top 10 pages, and saves the domains unvisited into the queue. The process ends when the queue is empty.

Though the spider pool sites could be detected, the performance is poor and the process is hard to converge. We find there are two main reasons. First of all, it takes quite some time for Google to respond to our web request, and CAPTCHA has to be solved occasionally. Second, the returned search results usually include sites not in spider pool, e.g., sites providing domain registration information and fake search engines which copy search results from other search engines ¹³. Such irrelevant sites have to be excluded, otherwise the iterations will be inaccurate and may not even converge, but using straightforward filtering metrics like domain ranking and URL patterns are not effective through our testing.

Classifications of SEO domains

We identified three types of methods of linking customer's content to SEO pages. They are listed below:

- 1. Through iframe. Instead of pointing to customer's site, the SEO page directly encloses the customer's page in an HTML iframe. It leverages JavaScript API document.write to inject the iframe, which occupies the whole screen when the user lands on the site. It could accumulate more traffic towards customer's site when visitors happen to land on SEO page.
- 2. Through hyperlink. The hyperlink pointing to customer's site or containing customer's message is displayed in the SEO page. This type is also used in SSP we infiltrated.
- 3. **Through redirection.** For this type, a visit to SEO page using certain user-agent string (e.g., Opera browser) is immediately redirected to customer's site through HTTP 302 redirect. In fact, it uses JavaScript code to assign customer's URL to window.top.location.href and HTML meta refresh tag to redirect visitor's browser. Likewise,

customer's content can be directly pushed to visitor's browser.

Table 9: QQ Number (anonyimzed) identified for message promotion.

| NO. | QQ Number | Topic | #URL |
|-------|------------|---------------|-----------|
| 1 | 53*95*1 | Sex | 156,306 |
| 2 | 34*06*58* | Drugs | 261,135 |
| 3 | 37*40*42* | Sex | 225,413 |
| 4 | 41*88*10* | Drugs | 186,370 |
| 5 | 51*89*17* | Drugs, Sex | 224,310 |
| 6 | 51*45*63* | Drugs, Sex | 155,459 |
| 7 | 63*11*95* | Drugs, Sex | 150,176 |
| 8 | 71*40*27* | Sex | 225,655 |
| 9 | 76*26*01* | Sex | 222,864 |
| 10 | 77*72*10* | Drugs, Sex | 224,476 |
| 11 | 79*43*07* | Sex | 225,000 |
| 12 | 10*36*03*3 | Sex | 5,986 |
| 13 | 10*33*91*1 | Drugs, Sex | 23,944 |
| 14 | 10*38*34*5 | Drugs, Sex | 17,267 |
| 15 | 11*07*07*2 | Drugs, Sex | 19,405 |
| 16 | 11*35*39*2 | Sex | 19,883 |
| 17 | 11*58*92*2 | Sex | 16,393 |
| 18 | 14*14*38*6 | Gambling, Sex | 6,906 |
| 19 | 15*53*65*8 | SEO | 5,339 |
| 20 | 19*82*12*3 | Gambling, Sex | 6,644 |
| 21 | 24*50*13*5 | Sex | 24,949 |
| 22 | 24*92*74*4 | Gambling | 24,736 |
| 23 | 25*17*02*5 | Gambling, Sex | 6,495 |
| Total | - | - | 2,435,111 |

Topics of Customer Sites

The 930 customer sites are classified under 7 categories below:

- 1. Sales and Services. We find industrial equipments and products, like elevator and seamless pipe, are sold on the customers' sites. Services of gray areas are also provided, like private detective and empty invoices.
- 2. Gambling. It includes sites for online casino and sports betting.
- 3. **Surrogacy.** The sites provide channels for infertile parents to find women willing to carry pregnancy.
- 4. News. The sites serve as news portals about local
- 5. Sex. Some of the sites host adult content like porn video and photos, while others list contacts of affiliated prostitutes or their agents.
- 6. Games. Online games developed by less known companies are provided. In addition, some sites provide information about the unauthorized servers for big-brand games, like The Legend of Mir [1], which can be connected to play the same game for free or less fees.
- 7. **Hospitals and Drugs.** We find the sites in this category introduce hospitals which do not have valid licenses. Besides, illegal drugs, like hallucinogen, are sold in some sites.

¹³The purpose of setting up such fake search engines is not yet clear. We suspect they are used for click fraud [29].

Table 10: Percentage of spider pool domains shown in search result of long-tail keywords.

| Keyword | Type | Google | Baidu |
|--------------------|------------------|------------------|------------------|
| 深圳福田外围商务模特报价 | Sex | no search result | 80% |
| 怎么联系深圳罗湖外围模特微信 | Sex | 10% | no search result |
| 怎么联系深圳罗湖兼职模特 | Sex | no search result | 100% |
| 深圳龙岗高端外围大概多少钱 | Sex | 10% | 90% |
| 深圳福田找高端外围女多少钱 | Sex | 20% | 100% |
| 深圳哪里有平面模特外围资源 | Sex | no search result | 100% |
| 深圳兼职女外围经济人电话 | Sex | 70% | 100% |
| 深圳罗湖美空高端商务模特经济人微信号 | Sex | no keyword | 100% |
| 深圳福田兼职女在哪有 | Sex | no search result | 100% |
| 深圳福田那里找美女兼职微信 | Sex | no search result | 100% |
| 深圳找白领兼职价格 | Sex | no keyword | 100% |
| 深圳南山小姐一夜多少钱 | Sex | 100% | 100% |
| 深圳龙岗哪里能找到兼职模特 | Sex | 100% | 90% |
| 安眠镇定药哪里买什么价钱 | Medicine | | 25% |
| 日本兴奋剂怎么在网上买 | Medicine | no keyword | 17.5% |
| 男性催情药怎么买价格多少 | | no keyword | |
| | Medicine | no keyword | 20% |
| 去哪买喷雾型迷幻药 | Medicine | no keyword | 37.5% |
| 金阳县哪里买真的拍肩迷药 | Medicine | no keyword | |
| 龙岩市安定片怎么购买 | Medicine | no keyword | 30% |
| 铜川哪里有迷情水卖的 | Medicine | 70% | 22.5% |
| 故城县三挫仑哪里买到 | Medicine | no keyword | 45% |
| 湘乡市网上哪买迷倒药 | Medicine | no keyword | 30% |
| 湛江迷幻药水哪出售 | Medicine | no keyword | 25% |
| 海南省三亚市哪里有汽枪买 | Firearms sales | 37.5% | 10% |
| 云南省楚雄市哪里有汽枪买 | Firearms sales | 32.5% | 15% |
| | Firearms sales | 42.5% | no keyword |
| 仿真手枪哪里买得到 | Firearms sales | 37.5% | no keyword |
| 芜湖哪里有汽枪买 | Firearms sales | 52.5% | no keyword |
| 平管双猎枪 | Firearms sales | 30% | no keyword |
| 大连仿真枪团购 | Firearms sales | 17.5% | 12.5% |
| 打鸟枪消声器货到付款 | Firearms sales | 20% | 2.5% |
| 舒兰汽枪买卖 | Firearms sales | 50% | no keyword |
| 出售77手枪 | Firearms sales | 40% | no keyword |
| 陕西代办本科毕业证 | Fake certificate | 37.5% | no keyword |
| 金华市办理真实研究生本科毕业证 | Fake certificate | 32.5% | no keyword |
| 牡丹江办理英语四、六级证书 | Fake certificate | 70% | no keyword |
| 晋中办理英语四六级证书 | Fake certificate | 62.5% | no keyword |
| 徐州办理日语二级证书 | Fake certificate | 80% | no keyword |
| 茂名办自考本科文凭 | Fake certificate | 42.5% | no keyword |
| 大同代开发票 | Fake certificate | 27.5% | 17.5% |
| 岳阳开发票公司 | Fake certificate | 87.5% | 32.5% |
| 哪里有制作假本科文凭的 | Fake certificate | 35% | 12.5% |
| 办张假本科文凭多少钱 | Fake certificate | 40% | 12.5% |

Classifications on Customer Messages

We analyze customer messages, extract QQ numbers from customer messages, and sort them by the number of associated URLs. The result is shown in Table 9.

Impact on Search Engines

We sampled 43 long-tail keywords extracted from the SEO pages crawled from the 21 spider pools and search them in Google and Baidu. The result is shown in Table10. We count the percentage of search results showing spider pool domains. If no search results are returned, we write "no search result". If no spider domains are presented in search results, we write "no keyword".