How to Learn Klingon Without a Dictionary: Detection and Measurement of Black Keywords Used by the Underground Economy

Hao Yang[1], Xiulin Ma[1], Kun Du[1], Zhou Li[2], Haixin Duan[1*], Xiaodong Su[3], Guang Liu[3], Zhifeng Geng[3], and Jianping Wu[1]

{yang-h16,xl-ma15,dk15,duanhx}@tsinghua.edu.cn, lzcarl@gmail.com
{suxiaodong, liuguang03, gengzhifeng}@baidu.com, jianping@cernet.edu.cn
[1]Tsinghua University, [2]IEEE Member, [3]Baidu Inc.

***Abstract*—Online underground economy is an important channel that connects the merchants of illegal products and their buyers, which is also constantly monitored by legal authorities. As one common way for evasion, the merchants and buyers together create a vocabulary of jargons (called "black keywords" in this paper) to disguise the transaction (*e.g.*, "smack" is one street name for "heroin" [1]). Black keywords are often "unfriendly" to the outsiders, which are created by either distorting the original meaning of common words or tweaking other black keywords. Understanding black keywords is of great importance to track and disrupt the underground economy, but it is also prohibitively difficult: the investigators have to infiltrate the inner circle of criminals to learn their meanings, a task both risky and time-consuming.**

In this paper, we make the first attempt towards capturing and understanding the ever-changing black keywords. We investigated the underground business promoted through blackhat SEO (search engine optimization) and demonstrate that the black keywords targeted by the SEOers can be discovered through a *fully automated* approach. Our insights are two-fold: first, the pages indexed under black keywords are more likely to contain malicious or fraudulent content (*e.g.*, SEO pages) and alarmed by off-the-shelf detectors; second, people tend to query multiple similar black keywords to find the merchandise. Therefore, we could infer whether a search keyword is "black" by inspecting the associated search results and then use the related search queries to extend our findings. To this end, we built a system called `KDES` (Keywords Detection and Expansion System), and applied it to the search results of Baidu, China's top search engine. So far, we have already identified 478,879 black keywords which were clustered under 1,522 core words based on text similarity. We further extracted the information like emails, mobile phone numbers and instant messenger IDs from the pages and domains relevant to the underground business. Such information helps us gain better understanding about the underground economy of China in particular.

In addition, our work could help search engine vendors purify the search results and disrupt the channel of the underground market. Our co-authors from Baidu compared our results with their blacklist, found many of them (*e.g.*, long-tail and obfuscated keywords) were not in it, and then added them to Baidu's internal blacklist.

*Corresponding author.

I. Introduction

Assume one day you ask a person about the meaning of "溜冰", a Chinese word for "ice skating". The answer will probably be similar to the explanation given by a dictionary. But if asking a merchant or a buyer of the *underground economy*, the answer could be quite different. In fact, " 溜冰" also means enjoying methamphetamine, a very dangerous drug.

Underground economy, an online marketplace that facilitates the transactions between merchants and buyers of illegal products, has been continuously proliferating. The revenue produced from the sales of the underground economy is enormous [2] and disrupting this marketplace will deal great damage to the criminal organizations hiding behind. A big obstacle in tracing the underground economy is to understand how the involved parties are communicating. To escape from law enforcement, jargons referring to the products are invented and added into the criminals' vocabulary. The above example suggests without a good reference, the chance of capturing a jargon's exact meaning is no more than a blind guess (similar to learning *Klingon*, a constructed language in the fictional Star Trek universe, without reference to any dictionary).

Challenges and our solution. Currently, to find and understand those jargons, the primary approach is to infiltrate the underground forum. Such approach is unscalable, when a large of amount of threads have to be reviewed and the jargons are tangled with many irrelevant texts. In this work, we explore the direction of discovering the jargons related to underground economy (called "*black keywords*" in this paper) through an *automated* fashion, in attempt to provide the investigators (or analysts) feed of black keywords and their context. This problem is similar to Name-Entity Recognition (NER), an area that has been advanced radically by many Natural Language Processing (NLP) techniques. Yet, our problem cannot be solved adequately by existing NLP techniques for two main reasons. First, NER systems are domain-specific, hard to adapt to a new domain where the vocabulary is very different [3]. However, the vocabulary of black keywords is rapidly evolved and differs significantly from other commonly

DOI 10.1109/SP.2017.11

used vocabularies. Second, many black keywords are created in ungrammatical or even obfuscated forms (*e.g.*, letter 'l' could be replaced by digit '1' in a word) while the NLP techniques are suitable for well-written text [4].

On the other hand, based on our prior experience in investigating underground economy, we found this challenge can be addressed through a pure *data-driven* approach. Many underground merchants rely on *blackhat SEO* (search engine optimization) to promote their business. Usually, plenty of black keywords are stuffed into one SEO page inside certain HTML tags (*e.g.*, anchor tags) to fool the search engines, yet making themselves distinguishable under content analysis. We could extract them from the SEO pages but the irrelevant texts have to be pruned. It turns out that the search results associated with a candidate keyword can be leveraged to determine whether the keyword is "black": as revealed by our study, querying a black keyword usually returns multiple links alarmed by the existing scanners, so we can use the result as the main indicator. Additionally, we found our list of black keywords can be extended through *related search*, a feature presented by major search engines to correlate similar search terms based on users' searching behaviors. After these steps, a lot of black keywords can be discovered, but a large portion of them are *long-tail keywords* which contain words not of our interest (*e.g.*, words except "heroin" in "where to buy heroin in Beijing"). To extract the *core words* (*e.g.*, "heroin" in the above example), we devised a substring matching algorithm which can process the keywords very efficiently.

We developed KDES (Keywords Detection and Expansion System) and evaluated it on more than 2 million pages related to SEO, porn and gambling. We discovered 478,879 black keywords in total and extracted 1,522 core words (433,335 black keywords are covered). After sampling the detected keywords, we found that the accuracy can achieve 94.3%, suggesting KDES is effective. We applied our findings to Baidu and the feedback was very encouraging. Many of the detected keywords have been added into their internal blacklist.

Discoveries. Our study also sheds light on the underground economy disguised under black keywords. We revealed the web infrastructure supporting the underground market (mostly in China) and unearthed contact information about the merchants (6,620 phone numbers and 7,272 QQ numbers). We traced the geo-locations of the phone numbers and found the merchants are located widely across China (30 provinces of China with at least 12 phone numbers each). The adversaries are able to disseminate the black keywords into the results of major search engines (*e.g.*, Google, Bing and Baidu). Surprisingly, we found that even search ads are tainted (under 17% core words), which should have been rejected by human reviewers. While our study focused on the communication channel established upon blackhat SEO, other online channels, like Baidu Tieba (a Chinese version of Reddit) and Zhidao (a Chinese version of Quora), were also ramped (more than 200K spammed pages detected for each) and the vocabularies are shared. Obfuscation is performed as a method to conceal the sale message and we have identified a list of transformation rules employed by adversaries. In the end, we provide a study about the online drug business based on our data. We identified several new characteristics undocumented by previous works, including the new hosting patterns and the new payment methods.

Contributions. We outline the contributions of this paper below:

- **Automated detection of new black keywords.** We retrofit the features from search engines (labels on search results and related search) to discover black keywords unknown to the public and extract the core words to ease the review process of human analysts.
- **Measurement and new findings.** We revealed the back-end infrastructure powering the underground economy mainly in China, the parties involved and new keyword transformation techniques used for evasion. In the meantime, the impact on search engines is also assessed.
- **Reporting our discoveries.** We collaborated with Baidu and reported our result to their security team.

Roadmap. The rest of the paper is organized as follows. Section II provides background information of our study. Section III elaborates the design of KDES. Section IV shows the implementation and evaluation result. Section V shows interesting discoveries of our study. In section VI we discuss the limitations, dependencies and use cases regarding KDES. Section VII reviews related work and Section VIII concludes the paper.

II. Background

In this section, we first present an example regarding how illegal products are promoted under black keywords. Then, we describe the typical blackhat SEO techniques that can be leveraged to promote sites unethically. Next, we describe the related search feature utilized by our system. Finally, we discuss the policies enforced by search engines.

Black keywords. To promote illegal products while circumventing the legal supervision, the underground merchants tend to abuse the public channels of internet and disguise their business intention under black keywords. Here we show one example of such promotion under blackhat SEO (illustrated in Figure 1). We assume the merchant is selling drugs (*e.g.*, heroin) in China and has created a shopping website (*e.g.*, foo.drug). As advertising such business is prohibited in China, the merchant has to seek irregular channels, *e.g.*, through poisoning the search engine results. As such, the merchant coordinates with blackhat SEOers and let them boost the ranking of foo.drug under certain search keywords (*i.e.*, black keywords), like "海洛因" (Chinese word for heroin). The search engine companies usually keep a close eye on keywords associated with illegal products, therefore, the SEOers also target keywords that are pertinent to "海洛因" but less monitored. To find such keywords, the SEOers either *obfuscate* the original keyword (*e.g.*, "洛因" with the first Chinese character removed), or combine other words to construct *long-tail* keywords [5] (*e.g.*, "北京哪里买海洛因" with the location term prepended to the original keyword).

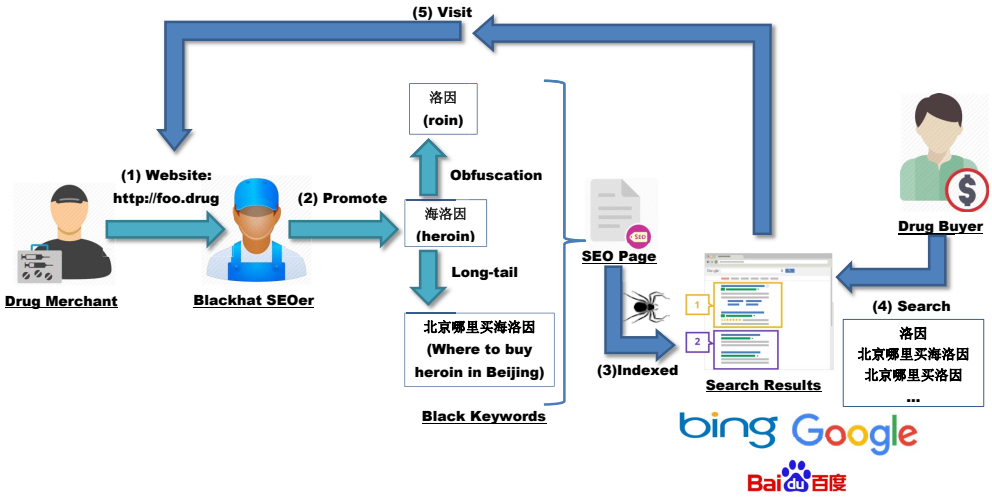

Fig. 1: An example of promotion under black keywords.

These two methods are sometimes utilized together to make the black keywords more elusive. Due to the gap between machine and human in comprehending natural language, those keywords are more likely to be ignored by the detector of search engines but queried by the buyers. Therefore, they could find foo.drug in the results if the SEO campaign runs successfully. Interestingly, because search engines often perform automated transformation on search keywords (*e.g.*, re-ordering words) to offer more relevant search results [6], the sites promoted under less used keywords could be merged with other frequently queried keywords and seen by the users.

In this work, we consider all keywords pointing to illegal products (*e.g.*, drugs and firearms) and services (*e.g.*, playbacks of adult video) as the target our study, including the original term, the obfuscated version and the long-tail version. With off-the-shelf tools [7] available, long-tail keywords can be generated. Unfortunately, these tools can only generate long-tail keywords. The original terms of the long-tail keyword are created and can only by created manually, without any rules to follow. As such, our problem cannot be solved through collecting and applying their tools.

We look into blackhat SEO activities as the data are easier to crawl and we focus on China's market as blackhat SEO is used extensively by the underground economy there [8]. However, we believe the similar situations also exist in other regions and the black keywords are also prevalent in other channels (as shown in Section V-D).

Blackhat SEO. SEO is a practice for one website to gain high rankings in search results, in order to attract more web users. Search engine companies recommend to improve website's structure and quality, but the blackhat SEO community does not follow such guidelines and advocates abusing the resources (*e.g.*, outbound hyperlinks) of innocent sites (*e.g.*, forums and blogs) or colluding with other SEO sites to boost the search rankings [9].

Our prior work revealed a new blackhat SEO infrastructure (called *spider pool*) that has gained traction especially in China [8]. More than 400K SEO sites have been discovered and many underground merchants relied on this type of SEO model to promote their business. In this work, we followed the same methodology and have collected 2,733,728 spider pool pages as the major source to study the phenomenon of black keywords. Here we briefly overview the model of spider pool.

Different from traditional blackhat SEO infrastructures, spider pool mainly targets long-tail keywords that are less competed by site owners. To reduce the operational cost, blackhat SEOers often purchase cheap domains (*e.g.*, expiring domains) in bulk to set up SEO sites. The spider pool actively traps the search crawlers and makes them visit the SEO sites indefinitely, and finally directs them to their customers' sites. There are two main techniques leveraged by spider pool to avoid triggering alarms of search engines. First, *wildcard DNS* is enabled on the SEO sites to create virtually infinite sub-domains to bypass the dead-loop detection performed by search crawlers. Second, the SEO page is generated dynamically for each crawler's visit to evade the check of content plagiarism. Built on top of these primitives, the customer's site gets much more frequent visits by search crawlers without improving site's quality and reputation. Besides customer's sites, customer's message (*e.g.*, "contact [phone number] to find prostitute in Beijing") is also promoted by spider pool, through a technique called *site free-ride*. Our work looks into the black keywords correlated with both customer's site and message.

Related search. People tend to try multiple times with modified search terms when the search results do not meet their expectation. To save user's time of typing new terms and to guide the user to find the correct terms, a feature called *related search* is provided by major search engines like Google, Bing and Baidu. Figure 2 shows an example of keywords prompted by Baidu's related search when a user queries "world war". Nine keywords are displayed, among which 8 contain the

Fig. 2: An example of related search results when a user queries "world war" in Baidu.

Fig. 3: Search result page with warning.

original term "world war" and 1 is translated from "world war z", the first related term.

The main principle for implementing related search is similar among search engines and here we use Baidu as an example. In fact, Baidu considers two types of input to correlate keywords [10]:

- **Keywords typed by one user in a search sequence.** This is the primary factor to determine the similarity between keywords. When a user queries one search term (say S_A) but does not click any search result, any following keywords with similar characters inputed by the user (say S_B) will be treated as related to S_A.
- **Result links clicked by a user.** From the perspective of search engines, the result links clicked by the user are expected to be relevant to the search keyword (say S_C). Meanwhile, the landing page redirected from the search result has already been analyzed and the representative keywords embedded in the page have been extracted (say $S_{D1}, S_{D2}, ..., S_{Dn}$ in a set S_D). The search engine would consider S_C and a subset of S_D (usually 2 or 3 most prominent keywords) as potentially related. If a pair of keywords shows up together with high frequency (*e.g.*, S_C and S_{D1}), the pair is treated as related.

Given that related keywords are extracted through analyzing user's searching behaviors and it is usual for a user to search multiple times for a single question/item and follow the most accurate search result, the related search offers decent opportunities in finding similar black keywords.

Search engine policies. Though the primary goal of search engines is to make web pages universally accessible, there are circumstances when they want to purify the search results. Below we briefly describe the policies established by popular search engines, including Google, Bing and Baidu, on what types of content are banned.

In the organic search result(The listings on search engine results pages that appear because of their relevance to the search terms), there are several categories considered as harmful by the majority of search engines, like child sexual abuse imagery, content infringing intellectual property, private information of a person [11], [12]. The related search result is either removed upon valid request or pruned automatically. In addition, search engines follow the law of local government if they have local versions. For example, Google removes the content that is at issue under the local laws, *e.g.*, "content that illegally glorifies the Nazi party on google.de" [13]. Another example is Baidu, which more actively censors business content that violates China's law [14]. We learned from the security team of Baidu that it keeps an internal blacklist covering underground business, including bogus purchasing, financial swindle, pornography, lottery, bogus bank site and etc. Sites captured by the list are removed automatically. Occasionally, Baidu prompts a warning page or partially blocks the search results when a user queries black keywords. As an example, Figure 3 shows Baidu's search results when a user queries "methamphetamine". A special warning is shown right above the first search result.

The displayed ads within search results are also examined (we introduce the background of search ads and the security implications in in Appendix A). The banned categories tend to be broader than those from organic search results. For example, Google AdWords bans counterfeit and dangerous product, inappropriate content and content enabling dishonest behaviors [15]. Bing ads lists 17 categories about disallowed content, product and services [16]. Baidu forbids products violating nation's law and requires licenses to be submitted for certain business [17].

To summarize, the attitudes of search engines are consistent against illegal content, or underground economy in particular [1], though the concerned categories might be different slightly. In fact, search engines do inspect black keywords and are striving to purify the indexed pages under them, but their knowledge of black keywords is lagged behind the underground communities. We aim to bridge this gap and for this purpose, we develop a novel system to harvest black keyword, which has been proved to be highly valuable.

III. Design

The ties between black keywords and underground economy call for an effective detection system. In this section, we present

[1]We consider sites delivering child abuse and pornography content also part of the underground economy as they usually ask for payment.

our solution, KDES. We first overview the architecture and then describe the design of each component.

A. Overview

Understanding the meaning of search keywords and identifying the black ones seems to be an obvious solution to our problem. In fact, there have been prominent progresses made in the domain of NLP in analyzing short texts [18], [19], [20], and the techniques were also incorporated by search engines to increase the relevancy of search results [21], [22]. However, such direction is unlikely to succeed in our settings, as a large amount of black keywords are deliberately obfuscated and context-dependent. The transformation rules and context are usually absent for NLP tools. As a result, it is almost impossible to infer keywords' meaning directly.

During our empirical analysis on black keywords, we found they frequently appear in SEO pages, underground forums and merchants' websites. A lot of such sites were indexed by search engines [2] and a noticeable proportion of them trigger alarms of detection systems. Therefore, we could leverage the labels (malicious, fraud and etc.) of the search results associated with one keyword to backtrack and infer whether it is black.

Motivated by the above observations, we developed KDES to detect black keywords unknown to the security tools or analysts. Figure 4 illustrates the system architecture. The *keywords extraction* module analyzes pages related to underground economy and extracts the keywords intended for SEO purposes. It also filters out the irrelevant keywords which is legitimate by using search engines as oracle. The remaining keywords are tunneled to the *keywords expansion* module and we leverage the related-search functionality to find other similar black keywords. A massive amount of keywords would be discovered after these steps, which is a heavy burden for analysts. The majority of the keywords are long-tail ones which are usually concatenations of *"core words"* (directly tied to illegal products or services) and *"filler words"* (less meaningful words like stop words). Core words are much more important and they are discovered by KDES through the core words identification module. Hereby, an analyst could find the interesting black keywords timely and prioritize her investigation.

Data source. As described in Section II, we implemented the same approach of [8] to detect SEO pages employed by spider pool. We targeted spider pool because it is widely used by blackhat SEO community in China nowadays and supports a broad spectrum of underground businesses. We collaborated with Baidu and scanned its indexed pages using our spider pool detection system from Aug 25th to Sep 10th, 2016, which in total yielded 2,733,728 SEO pages. We also obtained 63,424 pages marked as "evil" by Baidu, including 60,000 porn pages and 3,424 gambling pages. The label "evil" means the page content is associated with sex, gambling, dangerous goods, surrogacy, drug, faked sites and etc. These abundant data (2,797,152 pages in total) provided us a comprehensive view of the underground market, but it is also a nontrivial task to identify the unknown black keywords, similar to finding the needle in a haystack.

Our spider-pool detector and Baidu's detector focus on URL-level detection (see Appendix B for more details). The black keywords embedded within the web page are usually tangled with legitimate terms (see Section III-B) and how to pinpoint them accurately is not addressed by existing works. Though Baidu maintains an internal blacklist of keywords, they have made great efforts to analyze the keywords, especially those in the field of phishing sites and malware. Not any in-depth study has been conducted in all fields of black keywords. While machine-learning techniques have been used to generate realistic-looking passwords [23], due to shortage of samples, these techniques cannot be applied for discovering black keywords, especially the ones referring to new underground businesses or obfuscated manually with new rules. Below, we elaborate how we design each module of KDES to detect black keywords.

B. Keywords extraction

SEO practice (both whitehat and blackhat) advocates direct embedding of targeted keywords in the web page, to increase the *relevance* score assigned by search engines. Blackhat SEO differs from whitehat SEO in that it recommends *"term spamming"* [9], which fills a massive amount of keywords (*e.g.*, long-tail keywords under one topic) into the page to get it associated with as many search keywords as possible. As such, extracting black keywords from SEO pages becomes a natural choice.

As summarized by Gyongyi et al. [9], term spamming could happen in the body, title, meta tag, anchor text and hosting URL. We started from extracting keywords from all the above cases but soon gave up this approach, due to the high error rate. Black keywords are usually stitched together or mixed with legitimate ones and how to separate them is unclear (even just splitting a sentence into words is not easy for some languages, like Chinese [24]). After some failed attempts, we found extracting keywords from anchor texts (text inside the HTML tag `a href`) could yield the most promising result. Figure 5 shows one example of SEO page hosted by spider pool, which was dedicated for methamphetamine promotion. It contains 5 consecutive anchors and for each anchor, a black keyword is presented together with a URL pointing to another SEO site. We examined a small corpus of SEO pages and found that for all of them, the black keywords embedded in anchors are "clean-cut", meaning that they are not mixed with other black keywords or legitimate ones. Hence, we decided to narrow our scope to only anchor texts.

Keywords filtering. After extracting the keywords from SEO pages, we removed all duplicate ones, which still left us a gigantic keyword list (details described in Section IV-C). Through manual sampling, we found most of the keywords are beyond our target. One prominent error source is news link. In fact, spider pool often copies the web pages from news sites

[2]There are many underground transactions happening in anonymous marketplace, like Silk Road, which cannot be indexed by search engines. But still a large number of merchants and buyers communicate through visible internet services. We give more details in Section VII.

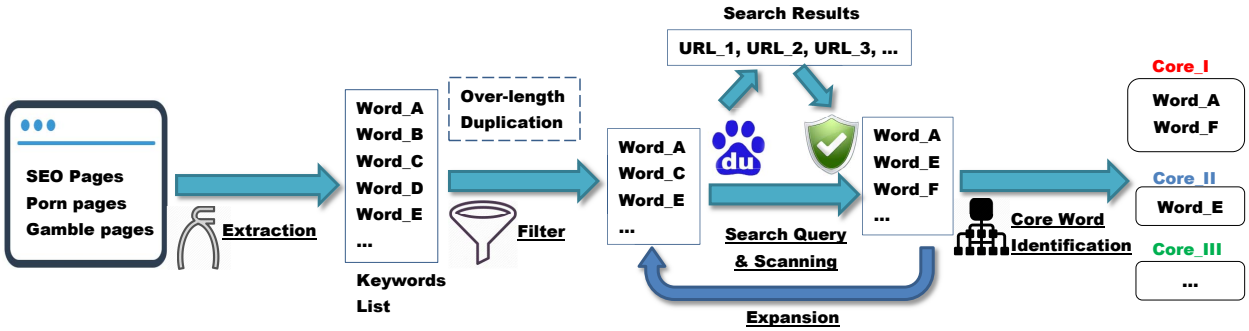

Fig. 4: Architecture of KDES.

and interleaves black keywords with the legitimate news links, in order to bypass the spam checker of search engine. Similar cases also appear in other "evil" pages marked by Baidu. In order to evade detection, the administrators of those sites prefer to add some anchors with legitimate content to their pages. These anchor texts are hard to remove through NLP based analysis, but we discovered a shortcut to quickly remove a large proportion of them: keywords in these anchors are usually much longer than the black keywords, and we could use a threshold of string length (denoted TH_{Len}) for filtering (we elaborate how TH_{Len} is determined in Section IV-B).

Yet, not every keyword remained after the above step is wanted. For instance, some keywords might refer to the titles of other pages under the same site. Instead of making decision based on their *intention*, we solve the problem from the *opposite* direction by looking into the *consequences* led by them. Our approach is both simple and effective: we issue search queries using all remained keywords and scan each link presented by the search result using the existing detectors. We inspect the first five page returned by the search engine and if a decent number of links are alarmed (above a threshold TH_{Flag}), we label the keyword black. Focusing on the first five page can help us detect black keywords more accurately, because legitimate sites have strong motivation to gain good rankings under legitimate keywords while stay away from black keywords [3]. In Section IV-A we give more details on the implementation and how we determine TH_{Flag}.

C. Keywords expansion

We obtained a plenty of black keywords after the above steps but their vocabulary is far from being exhausted, in part to that we did not (and cannot) collect the pages spanning all underground business. To complement the missing part, we leverage the related search feature provided by the search engine to find unknown black keywords similar to what we have discovered. As described in Section II, related search presents similar keywords based on users' querying behavior. Therefore, the keywords shown under related search should

[3]Black keywords might be written in the articles of legitimate sites, but they are less likely to be promoted for SEO purposes.

Fig. 5: An example of SEO page with anchor text highlighted.

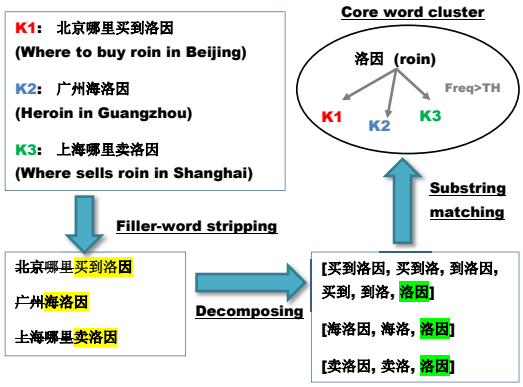

Fig. 6: An example demonstrating core word identification.

have high chance to be black if we seed one confirmed black keyword [4].

Specifically, we query a black keyword and save all the

[4]The adversary might submit many irrelevant queries following a black-keyword query to poison the keywords relations, in order to misguide our system. However, this method requires huge computing resources to override the large volume of search queries from ordinary users. Such anomaly can be spotted easily.

keywords listed in the related search section (say W_{Rel}). If a word W_1 from W_{Rel} has not been examined before, we run it through search engine and use the same detection method of keywords filtering component (count the number of alarmed search results) for classification. While we could continue the expansion by seeding W_1 recursively, the cost is prohibitively high. To demonstrate the overhead, we use Baidu query as an example. Baidu usually (and maximumly) returns 9 related keywords for one query, so one-round expansion equals to 9 search queries. Seeding all of them for the next round leads to 81 additional queries (90 by including the first-round queries). Even with Baidu's internal API, 90 queries would cause notable overhead (at about 2 seconds and this API is only accessible to us at night). In addition, the more iterations we run, the less relevant are the keywords returned.

D. Core word identification

We could report all the keywords produced by the previous steps to analysts but the reviewing procedure would be very painful, if without further processing. In fact, many black keywords are just extension of a limited set of core words (*e.g.*, "heroin" is the core word for the long-tail keyword "where to buy heroin in Beijing") which are much more attractive to the analysts. Hence, we developed a component to identify the core words and cluster similar black keywords under them to curtail analysts' workload.

NLP based text analysis is not a viable solution here. Many black keywords are obfuscated (*e.g.*, letter changed to digit) and some of them are not even processable by NLP. For example, "菠菜"(Bo Cai) means spinach in Chinese, and its pronunciation is exactly the same as "博彩"(Bo Cai) which means gambling. So NLP techniques will identify "Bo Cai" as spinach, rather than gambling, thereby this expression can evade the detection of NLP. Besides, NLP requires a basic set of language elements as input but so far there is no comprehensive reference for black keywords. As such, we cannot identify the core words by analyzing their semantic meanings (like NER). Taking one step back, we could use NLP to just parse the keyword string into separated elements and pick up the elements that are unusual but shown in many detected keywords. Unfortunately, this approach fails too. A lot of black keywords we encountered are in Chinese, and word segmentation for Chinese sentence is a notoriously hard problem [24]. We tried state-of-art NLP tools (e.g., Natural Language Toolkit [25] and nlp-toolkit [26]) on 1000 samples and the false-positive rate is over 10%.

Alternatively, we exploit one prominent feature of black keywords, that most of them are long-tail keywords generated by combining filler words and core words, to address this problem. The filler words are picked from a dictionary, including stop words (like "a" and "where") and places (like "Beijing" and "China").

While the parts except filler words are not always core words, we can identify our targets by picking up the substrings showing up with sufficient frequency, on the grounds that core words are widely shared in underground community. Stripping off filler words is straightforward, but finding the common substring is not so easy. The problem we are dealing with here is similar to finding the longest common substring (LCS) among a list of strings [27]. There are several shortcuts allowing us to design an efficient algorithm: each keyword is short after removing filler words; a core word only needs to be discovered from some keywords to be prominent. As such, our algorithm starts from breaking a keyword into substrings whose length are more than a threshold TH_{SubLen} (set to 2 characters during evaluation). Then, it picks one substring and check its existence in every other keyword. We stop the searching process earlier if the number of matched keywords is above TH_{freq} (10 based on our empirical analysis). The substring passed the check is considered as a core word and we store it together with its connection to the owner keyword [5]. The process goes on until all substrings of all keywords are examined. Figure 6 shows one example processed by our algorithm and Figure 7 shows the pseudo-code for this algorithm. In the end, all core words and their associated black-keyword clusters are sent to the analysts for review.

```
MULTIPLELONGESTCOMMONSUBSTRING(D)
 1  A ← PICKRECORDFROMDATASET(D)
 2  A ← REMOVESTOPWORDS(A)
 3  A ← REMOVETOPONYMWORDS(A)
 4  SubStringSet ← GENERATESUBSTRING(A)
 5  for substring in SubStringSet
 6       do count ← 0
 7          for LongTailKeyword in D
 8               do if substring in LongTailKeyword
 9                    then count ← count +1
10                         if count > threshold
11                           then STORECOREKEYWORD(substring)
12                                break
```

Fig. 7: Pseudocode of Core Word Identification Algorithm.

IV. Implementation and Evaluation

A. Implementing KDES

We bootstrapped KDES by loading the 2,797,152 pages detected by our spider-pool scanner and the evil pages provided by Baidu, as described in Section III-A. Both the crawlers run by us and Baidu downloaded the HTML pages without executing dynamic content (*e.g.*, JavaScript code) or storing pages linked by iframe for efficiency. We found that in most cases black keywords are rendered in the main HTML pages so the volume of overlooked keywords should be relatively small comparing to what we have captured.

Parsing all the pages on a single machine is quite time-consuming. As a result, we implemented the parser on a Apache Hadoop cluster (consisting of 166 machines), which stored all the pages in Hadoop Distributed File System (HDFS) and ran MapReduce jobs for processing. Our parser leverages BeautifulSoup to extract keywords from anchors and we stored

[5] A keyword might be associated with multiple core words. For instance, "heroin" and "methamphetamine" are both core words within "where to buy heroin and methamphetamine"

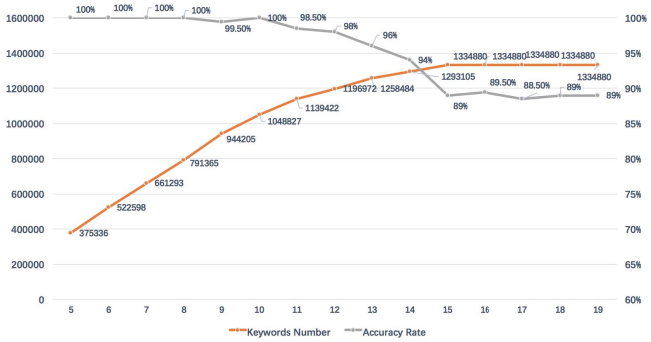

Fig. 8: Statistics on different TH_{Len}.

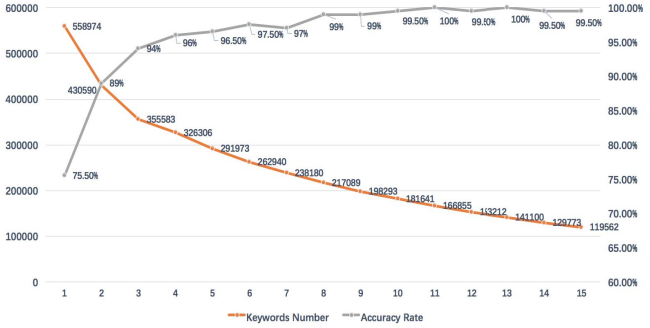

Fig. 9: Statistics on different TH_{Flag}.

them on a workstation (12 core E5 CPU, 128G Memory). In total, we got 8,812,609 keywords in 27 hours were consumed at this stage.

The next step is to filter the duplication and overlength keywords, then query the keywords with search engine. Search requests are usually throttled by either CAPTCHA or volume control if queried in large volume. Baidu generously granted us access to its unthrottled version of search API and we finished the job (including related search) in 5 days. We scanned all URLs in search results using an in-house scanner provided by Baidu together with our spider-pool detector and we elaborate this mechanism in Appendix B.

For keywords expansion, we utilized the related search feature from Baidu and queried all related terms (9 maximum) per detected keyword. For core word identification, we also used the Hadoop cluster to parallelize the matching process of substrings, which was completed in 0.5 hours (for 478,879 keywords).

B. Parameter selection

Below we elaborate the process of parameter tuning for TH_{Len} and TH_{Flag}.

Keyword length (TH_{Len}). We use this threshold to filter out anchor texts about legitimate content or news links. We experimented with values from 5 to 19 and counted the number of remained keywords and the ratio of keywords that are unrelated to news (true positives) from 400 random samples. The statistics are illustrated in Figure 8. When TH_{Len} equals to 14, the result is optimal (biggest for the product of the two values).

Alarmed search results (TH_{Flag}). A keyword is deemed black if more than TH_{Flag} search results are alarmed. We changed TH_{Flag} from 1 to 15, and counted the number of keywords associated with at least TH_{Flag} alarmed search results and the ratio of black keywords from 400 random samples. We show the statistics in Figure 9 and we set TH_{Flag} to 3 which leads to the best outcome, for acceptable accurate more than 90%.

C. Evaluation result

We ran KDES on all 2,797,152 pages and 1,293,105 keywords are preserved after deduplication and filter on keyword length (7,519,504 keywords are removed). The remaining keywords were examined by the search engine oracle and 355,583 keywords were marked black in the end.

For keywords expansion, we randomly selected 69,475 black keywords and queried for related keywords. We did not seed all the keywords from the last stage because of the high overhead (9 related keywords have to be queried for one). In this stage, we confirmed 133,738 keywords, adding up to a total number of 478,879 black keywords (there is an overlap of keywords in these two stages).

Reviewing this sheer amount of keywords one by one requires huge manual efforts. It turns out the core word identification component is very helpful: 1,522 core words were recognized and 433,335 (90.4%) black keywords were covered. We found many core words never known before, like the ones obfuscated with unseen rules (more details are given in Section V-C). Based on our discussion with Baidu's security team, we found that a lot of black keywords were not yet covered, although Baidu keeps collecting black keywords through manual inspection.

Accuracy of KDES. Due to the absence of ground truth, we have to manually review the detected keywords and understand its relation with the underground economy. Verifying all keywords is not feasible under limited time, so we sampled 1,000 keywords randomly. While the meaning for some keywords were known and clear (*e.g.*, "海洛因" or "heroin"), we did find many keywords whose meaning were ambiguous (*e.g.*, "溜冰" or "ice skating"). For those keywords, we queried them on Baidu Tieba (a Chinese forum), QQ groups and Baidu search, three popular communication channels for the underground economy, and examined the posts, chat messages, member information and web pages (specifically images and videos) to determine their real meaning. In the end, we confirmed 943 keywords as black (94.3% accuracy).

Presumably, black keywords should lead to artifacts (*e.g.*, web pages) belong to the underground economy at higher chances. To assess this claim, we selected 200 black keywords

and the same amount of random keywords, and compared the *toxicity* of search results (the proportion of results that are malicious, as defined in [28]). The comparison result shows black keywords are indeed good indicators, as the toxicity is 53.5%, 5.6 times of the toxicity from the random keywords (9.5%).

V. Measurement

Based on the collected black keywords, we carried out a comprehensive assessment about the underlying infrastructure, the criminals behind, the impact on search engines, the overlap with other channels and keyword transformation performed by the adversaries. In the end, we bring to light a previously unreported drug equipments marketplace in China.

A. Overall statistics

We manually reviewed the 1,522 core words and divided them into 7 categories: drugs (sales of drugs, drug equipments and etc.), dangerous items (*e.g.*, machine gun, daggers, knives and etc.), gambling (online casino, gambling machines, sports gambling and etc.), sex (pornography, adult forums and prostitute service), blackhat SEO, surrogacy and others. For each black keyword, we also obtained the list of URLs that were alarmed by the scanner and the associated domains. Table I lists the statistics regarding the black keywords in both the extraction and the expansion stage.

During the extraction stage, 1,297 core words were identified and the number is increased to 1,522 after expansion. Interestingly, the core words from the expansion stage covered all core words in the extraction stage, while the expansion core words are from 133,738 keywords and the original core words are from 355,583 keywords, suggesting that related search mainly appends extra characters to a user's keyword. Most of the core words are related to sex (769, 50.5%) and gambling (503, 33%). Comparing to other categories, these two are more popular among ordinary web users. Our substring matching algorithm successfully clustered the majority of the black keywords (88% in the extraction stage and 90% in the expansion stage). The remaining ones cannot be classified because the frequency of such core word is under the threshold TH_{freq}.

Yet, we found some new business categories were not covered by the 6 main categories, including carder (trading of stolen credit card data), financial fraud and medical equipments. These minor categories also need to be inspected and we are improving our algorithm to cover these cases. From the black keywords, we are able to discover around 2 million malicious URLs and 1 million malicious domains, showing that adversaries are well funded to register many domains for their operations.

B. Underground organizations

From the malicious pages and domains associated with the black keywords, we could identify the behind organizations. To make their businesses more visible to the buyers, the merchants usually left contact information, like telephone numbers and QQ (an IM tool popular in China) numbers, in web pages. Most of the domain registrants provide their contact information to registrars in China, due to the strict policies enforced there. As such, we could trace back to the merchants and registrants by analyzing the pages and Whois databases.

However, identifying the contact information from the web page is not a trivial task. The scanners from security companies and search engines are also looking for such information to detect spam. The adversaries are aware of that and they have applied a variety of obfuscation techniques (*e.g.*, "4x4x5" for "445"). Through empirical analysis on a small set of SEO pages, we summarized a set of commonly used techniques and devised the rules to recover the the numbers and their types from obfuscated text. The rules are applied when a short text resembling to a number is discovered. The mostly used rules are removing padded blanks and delimiters (*e.g.*, "x" and "-"), switching letters to digits (*e.g.*, letter "o" to digit "0"), replacing Chinese characters (*e.g.*, "壹" to "1") and homophonic words (*e.g.*, "扣扣" to "QQ").

In the end, we were able to extract contact information from 283,547 pages (15.3% of all malicious pages) and obtained 6,620 phone numbers and 7,272 QQ numbers. It turns out the merchants are aggressively publishing their numbers, as the volume of pages is far more than the amount of phone and QQ number. Then, we queried each phone number in ip138.com, a website providing the owner's location based on the carrier's records and we retrieved 6,331 valid records. Most of the numbers without record start with "400", a proxy number that avoids back-tracking. Figure 10 illustrates the popularity of numbers within each province of China. We found that the merchants behind the underground economy are rather disperse: there are 30 provinces owning at least 12 phone numbers. But in the meantime, the numbers are not evenly distributed: more developed and crowded provinces tend to own more phone numbers (*e.g.*, 1,311 numbers belong to Guangdong, followed by 512 from Shandong). We classified the phone and QQ numbers into the same categories and show the results in Table II and Table III. Similar to the keywords popularity, sex and gambling are the most popular categories.

Next, we looked into the domains registered for the underground economy. Though in average less than two malicious pages were hosted by each domain, there are some domains populating a large amount of pages into the search result. Table IV lists the top 10 domains and we discovered 5,299 URLs for the top 1 domain. By clustering the domains by their TLD, we found `.com` domains are most popular, covering 34.01% of all domains. However, we found there are also many domains registered under new gTLDs, like .top (12.41%). Previous works have shown that new gTLDs are favored by blackhat SEOers [8]. Our result is consistent with the prior findings.

Among all 1,014,688 domains, we obtained the valid Whois record for 424,498 (the remaining ones have expired, according to Whois query). We extracted 32,970 unique email addresses from 361,086 domains and listed the top 10 addresses in Table VI. While some emails with high rankings do not point to the individual registrants (*e.g.*, the user ID

TABLE I: Keywords statistics divided by the 7 categories.

Extraction	Category	Core Words	Keywords	Keywords%	URL	URL%	Domain	Domain%
1	Sex	682	180,641	50.80%	546,239	48.77%	268,010	41.99%
2	Gambling	428	116,894	32.87%	354,604	31.66%	221,272	34.67%
3	Danger	145	9,043	2.54%	80,769	7.21%	17,988	2.82%
4	Surrogacy	5	325	0.09%	6,254	0.56%	1,171	0.18%
5	Blackhat SEO	34	5,986	1.68%	52,392	4.68%	25,774	4.04%
6	Drug	3	7	0.001%	134	0.01%	31	0.001%
7	Other	0	42,687	12.00%	79,798	7.12%	103,975	16.29%
8	Total	1,297	355,583	100%	1,120,091	100%	638,221	100%
Expansion	Category	Core Words	Keywords	Keywords%	URL	URL%	Domain	Domain%
1	Sex	769	61,194	45.76%	329,866	39.90%	194,298	40.42%
2	Gambling	503	52,134	38.98%	287,093	34.73%	175,845	36.58%
3	Danger	161	3,973	2.97%	84,568	10.23%	26,486	5.51%
4	Surrogacy	18	482	0.36%	1,210	0.15%	188	0.04%
5	Blackhat SEO	40	2,600	1.94%	34,964	4.23%	28,196	5.87%
6	Drug	31	56	0.04%	130	0.02%	86	0.02%
7	Other	0	13,299	9.94%	88,853	10.75%	55,613	11.57%
8	Total	1,522	133,738	100%	826,684	100%	480,712	100%
-	All Total	1,522	478,879	100%	1,848,749	100%	1,014,688	100%

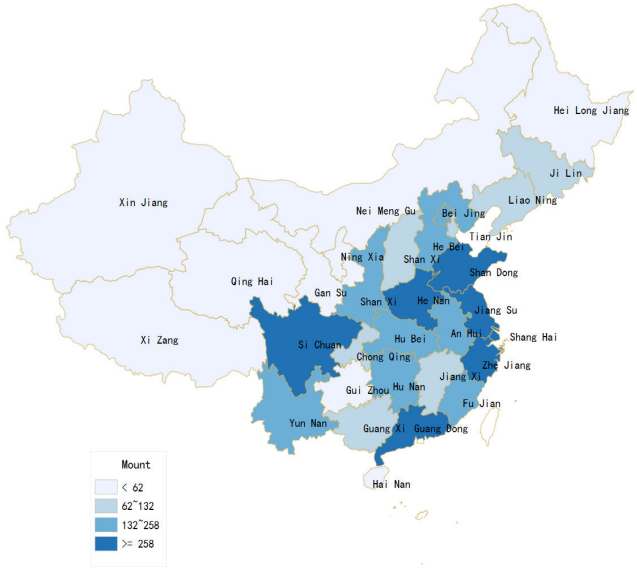

Fig. 10: Heatmap of the phone numbers.

TABLE II: Telephone numbers counted by categories.

No.	Category	Telephone Count	Percentage
1	Sex	2,428	38.35%
2	Gambling	1,851	29.24%
3	Danger	1,103	17.42%
4	Blackhat SEO	594	9.38%
5	Surrogacy	334	5.28%
6	Drug	21	0.33%
Total	-	6,331	100%

of `yu*in*pi*a@163.com` means "domain sales"), there are emails seemingly to be registered using individual email account (*e.g.*, the email addresses under `qq.com`). It turns out the malicious domains we captured may be only a tip of the iceberg: we queried reversewhois.domaintools.com to get the number of all domains registered under each top email address

TABLE III: QQ numbers counted by categories.

No.	Category	QQ Count	Percentage
1	Sex	2,956	40.65%
2	Gambling	2,585	35.55%
3	Danger	816	11.22%
4	Blackhat SEO	789	10.85%
5	Surrogacy	92	1.27%
6	Drug	34	0.47%
Total	-	7,272	100%

TABLE IV: Top 10 black domains ordered by the number of captured URLs.

No.	Domain	Category	URL Count	Percentage
1	hqzsx.com	Blackhat SEO	5,299	0.29%
2	xibu.tv	Blackhat SEO	4,027	0.22%
3	308k.com	Gambling	2,810	0.15%
4	cz89.com	Gambling	2,616	0.14%
5	xhlfmc.com	Blackhat SEO	2,543	0.14%
6	ccbkr.com	Sex	2,350	0.13%
7	99083.co	Gambling	2,195	0.12%
8	kantao.net	Blackhat SEO	1,936	0.10%
9	93013.co	Gambling	1,496	0.08%
10	zhiguan muju.com	Blackhat SEO	1,355	0.07%
Total	-	-	26,627	1.44%

and the aggregated number is more than 2 million. Finally, we compared the email addresses to the ones extracted from the spider-pool domains we have captured before (2,731). The Venn diagram of the two sets is shown in Figure 11. This time, we identified much more underground players through `KDES`, which include Faking certificate, Faking Luxury and Drug.

C. Impact on search engines

Search volume. We are interested in which keywords are more likely to be queried by the users. To this end, we asked Baidu and obtained a snapshot of the search volume pertaining to each keyword aggregated from August 27th, 2016 to October 10th, 2016. We matched the black keywords against the search volume list and were able to find many matches. We show the

TABLE V: Top 10 TLDs ordered by the domain count.

No.	TLD	Domain Count	Percentage	Type
1	com	345,092	34.01%	gTLD
2	cn	238,867	29.57%	ccTLD
3	top	125,896	12.41%	gTLD
4	cc	68,375	6.74%	ccTLD
5	net	41,829	4.12%	gTLD
6	win	18,387	1.81%	gTLD
7	xyz	14,375	1.42%	new gTLD
8	wang	12,137	1.20%	new gTLD
9	info	9,329	0.92%	gTLD
10	bid	8,394	0.83%	new gTLD
Total	-	933,681	92.02%	-

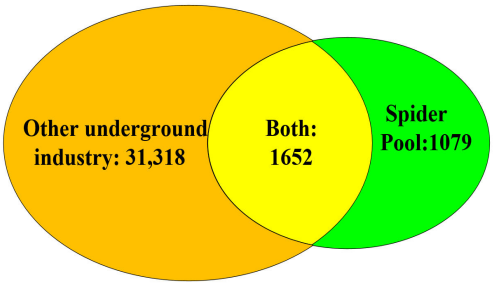

Fig. 11: Comparison of registrants' emails.

top 10 keywords in Table VII. As usual, keywords about sex and gambling received the most queries. Most of the queries were quite specific about the names of software (No.1 and 3) [6], web sites (No.4 and 5) and gambling games (No.2, 5-8 and 10).

Keywords in other search engines. Most of the data (*e.g.*, search results) we obtained came from Baidu. One may wonder whether the same phenomenon exists in other search engines, like Google and Bing. We want to learn the answer as well but due to the deployment constraints (we have no access to the internal APIs of Google and Bing), verifying every keyword will be very difficult. Thus, we sampled 120 keywords (20 per each categories) and queried them with Google and Bing. We checked all returned links of the first page. It turns out that Google and Bing were neither exempted from the issues related to black keywords. Due to the space limit, we only showed 12 keywords in Table VIII.

Search ads. The revenue brought by search ads usually takes the lion's share in the overall revenue of a search engine company. As described in Appendix A and Section II, strict screening is usually performed for each search ad submitted to the advertising platform. However, the platform is not bullet-proof. We show one example about Baidu search ads in Figure 12. To assess the prevalence of this problem, we queried all 1,522 core words on Baidu and found 259 (17%) of them lead to search ads, which should be banned at the very beginning. Table IX shows the number of such keywords under each category.

[6]While these software can play any video, they are mostly welcome for porn video in China.

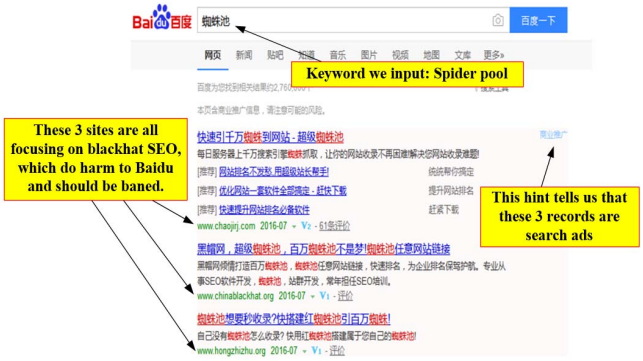

Fig. 12: Search ads under keyword "spider pool".

Fig. 13: Search ads submitted by tmall.com.

One case which draws our attention is shown in Figure 13, which displays a search ad selling one type of knives prohibited in China. The ad came from tmall.com, the leading eCommerce platform in China. Surprisingly, the ad does not link to any particular merchant under tmall.com. It seems that tmall.com just blindly bids Baidu's keywords to attract as much traffic as it can. Such keywords should be avoided by both eCommerce sites and search engines, but the real meanings are usually obscure to even seasoned reviewers. As a countermeasure, these affected parties could deploy our approach to sanitize the search ads.

D. Other channels

We examined the popularity of black keywords disseminated through search engines. We are also interested in whether the same set of keywords is used in other channels. To this end, we matched the black keywords with the pages under Tieba [29] and Zhidao [30], two very popular social network sites in China (like Reddit and Quora) operated by Baidu. Since we have no API access to obtain designated pages from these sites directly, we examined the pages indexed by Baidu listed in the first 5 result pages. In total, we identified 243,859 and 208,869 pages of Tieba and Zhidao separately. We listed the count of URLs under different categories in Table X and Table XI.

While our prior work has demonstrated that popular sites can be abused to promote spam messages through site free-

TABLE VI: Top 10 registrants' email addresses ordered by domain count (ID anonymized).

No.	Email	# Black	All Black %	# Registered	$\frac{\#Black}{\#Registered}$
1	yu*in*pi*a@163.com	33,319	7.85%	214,516	15.53%
2	39*68*22*@qq.com	14,816	3.49%	272,691	5.43%
3	yu*in*@yinsibaohu.aliyun.com	11,291	2.66%	295,090	3.83%
4	zh*nt*nf*999@gmail.com	8,755	2.06%	65,576	13.35%
5	28*6*@qq.com	7,117	1.68%	675,710	1.05%
6	41*76*34@qq.com	6,693	1.58%	37,094	18.04%
7	xi*os*ou*um*ng@163.com	4,730	1.11%	181,746	2.60%
8	a*m*n@juming.com	4,275	1.01%	255,948	1.67%
9	29*23*34*3@qq.com	3,992	0.94%	39,379	10.14%
10	3@th*e*d.cn	3,863	0.91%	14,990	25.77%
Total	-	98,851	23.29%	2,052,740	4.82%

TABLE VII: Top 10 keywords searched by all users.

No.	Keywords	Category	Note	Count
1	影音先锋	Sex	a porn video player	1,904,522
2	大乐透	Gambling	a game of lottery	1,009,248
3	先锋影音	Sex	reverse of No.1	614,372
4	草榴	Sex	a porn community	481,691
5	五月天	Sex	a port community	275,653
6	七星彩	Gambling	a game of lottery	238,671
7	九色腾	Gambling	a game of lottery	190,891
8	天线宝宝	Gambling	a new game of lottery	104,928
9	代孕	Surrogacy	surrogacy in Chinese	52,702
10	管家婆	Gambling	a new game of lottery	50,980

TABLE VIII: Number of links in the first page in Google and Bing for black keywords.

No.	Black Keywords	Category	Google	Bing
1	缅甸腾龙娱乐	Gambling	7	9
2	澳门威尼斯人平台	Gambling	10	10
3	唐刀网购	Danger	8	8
4	黑鳄战刀哪里能购买	Danger	8	3
5	南京代孕价格	Surrogacy	10	10
6	厦门代孕公司	Surrogacy	10	10
7	咕噜咕噜溜冰群	Drug	4	8
8	南阳冰妹怎么找	Drug	10	0
9	四房播播开心色播	Sex	10	10
10	聚色导航	Sex	10	7
11	简单格调灰色SEO优化手机模板	Blackhat SEO	5	2
12	时时彩推广软件	Blackhat SEO	9	10

ride [8], this finding is different. For the prior approach, the adversary only needs to inject the URL (with the spam message attached) of popular sites into the results of search engines, the attack we found here has to post articles to the sites, similar to forum spamming. While the issue has been known long time ago, and there are several defense techniques employed by

TABLE IX: Statistics of search ads under black keywords.

No.	Category	Keywords Count	Ads Count
1	Sex	44	58
2	Gambling	105	155
3	Danger	88	131
4	Blackhat SEO	14	28
5	Surrogacy	1	1
6	Drug	7	7
Total	-	259	380

TABLE X: Classification of Tieba URLs.

No.	Category	Count	Percentage
1	Sex	119,194	48.88%
2	Gambling	83,457	34.22%
3	Danger	16,496	6.76%
4	Surrogacy	573	0.23%
5	Blackhat SEO	1998	0.82%
6	Drug	115	0.05%
7	Other	22026	0.03%
Total	-	243,859	100%

TABLE XI: Classification of Zhidao URLs.

No.	Category	Count	Percentage
1	Sex	109,143	52.25%
2	Gambling	71,072	34.03%
3	Danger	11,675	5.59%
4	Surrogacy	343	0.16%
5	Blackhat SEO	1261	0.60%
6	Drug	30	0.01%
7	Other	15,345	7.35%
Total	-	208,869	100%

Baidu already, like CAPTCHA and ID verification, however, the threat is far from being completely mitigated. We found the adversary either posted a question whose title contains black keywords, or put the keywords to the replies following a question, which might be created by herself as well. Since the hosting sites are assigned with high reputation score, the spam posts easily got high rankings in search results. Despite the difficulty of finding spam posts on behalf of the site owners, our solution provides an alternative way to mitigate this issue. Finally, we reported all the URLs to Tieba and Zhidao. The security team there have acknowledged our discovery and are in the process of deleting all spammed posts.

E. Keywords construction

Black keywords are continuously created either from scratch or transformed from the existing ones. By manually reviewing the 1,522 core words and searching for the explanations, we discovered 5 ways about keyword construction and we elaborate them below.

1) **New keyword without reference.** For example, "球版" (ball board) is the name for a new type of sports gambling. "三响海豚" (three-shout dolphin) refers to a new gaming machine for gambling. Another example is "出肉" (cut meat) whose real meaning is selling drugs.

2) **New meaning for an existing keyword.** Distorting the meanings of existing keywords can cheat analysts and leverage the high weight assigned by search engines at the same time. We find this is a quite popular approach. "新球代理", a keyword referring to an agent selling balls, now refers to sports gambling. "打保单", a keyword meaning printing an insurance contract, now means printing the result on a paper receipt before the online gambling starts in order to prevent cheating. "咕噜咕噜" is particularly interesting because it was an onomatopoeia (guru guru) at first but now refers to drug equipments, because the same sound could be produced when using the equipments.
3) **Replacing characters with ones of similar pronunciation.** This transformation has been leveraged for Soundsquatting, a practice to register domains with the similar pronunciation [31]. We found there are some black keywords constructed in this way. For example, "博彩" is the Chinese word for gambling and lottery. Underground community used a new keyword "菠菜" (spinach) to represent the same thing, because "博彩" and "菠菜" pronounce quite similar in Chinese Pinyin.
4) **Replacing characters with ones of similar shapes**. Because human is good at recognizing misspelled words but machine is not so, adversaries can construct new keywords by exploiting such gap. During our research, we found an adult keyword "吉彩娱乐" (fortune color entertainment) first, a service for online gambling. And ten days later, we found another adult keyword "杏彩娱乐" (apricot color entertainment) referring to the same service. The first character of these two keywords have the similar shapes. Hence, the latter one also receives search queries from users.
5) **Changing character set**. This is very prevalent in keywords embedding phone numbers and QQ IDs. There are various forms for this transformation. For example, "1234" can be transformed by replacing the digits with simplified Chinese characters ("壹贰三4") or traditional Chinese characters ("12叁肆").

A keyword created initially could receive a large volume of search traffic and the ones evolved from it receives search traffic as well, even though less than the initial one usually. We show one example about Mark Six ("六合彩"), a kind of lottery game forbidden in mainland China but permitted in Hongkong. In fact, we have discovered a lot of keywords apparently transformed from "六合彩". Table XII shows the top 10 related keywords ordered by the search volume, using the snapshot described in Section V-C. The transformation strategies are quite diverse, including removing the last character and replacing the character with the same pronunciation (*e.g.*, "合", "和" and "盒"). "六" can be replaced by "6" or even "⑥" not in any human language. Adversaries also target Pinyin of "六合彩", *i.e.*, "liuhecai". As shown in Table XII, "六合" receives the most traffic other than the original keyword. The traffic towards other keywords are more widespread. Interestingly, though not shown in Table XII, keywords with "⑥" are still queried even though they are very inconvenient for a user to input.

TABLE XII: Top 10 keywords related to "Mark Six".

No.	Core Keyword	Search Count
1	六合彩	17,443
1	六合	16,212
2	六盒	1,560
3	六和	1,339
4	6合	1,142
5	liuhecai	1,110
6	6喝	440
7	六喝	403
8	六禾	327
9	六彩	202
10	六he	135

F. Case study

Drug selling and taking in any form is forbidden in China. Compared to other underground businesses, the legal punishment is much more severe on drug dealers and takers. Still, we found 63 drug keywords promoted by merchants and searched by users through search engines. The online drug market in China is quite mysterious and there is no report before about how this market operates. In this subsection, we describe our primary results of the initial exploration.

First, we inspected the black keywords and their search volumes. Table XIII lists the top 5 keywords queried by users. Besides "溜冰", the keyword meaning drug taking, keywords about the subsidiary business (*e.g.*, "冰妹" about drug-taking companion and "咕噜咕噜" for drug equipments) also receive noticeable volumes of user queries, suggesting the whole marketplace is well developed.

Although the merchants directly selling drugs only leave contact information through forum spamming or site free-ride, we discovered 20 eCommerce sites directly selling drug equipments (*e.g.*, Bongs). The sales of drug equipments are considered as "gray" in China, but we believe tracing from such eCommerce sites could help the legal authorities infiltrate the drug community. We first issued DNS request to obtain the IP for each domain and learned its location by using 360 Passive DNS [7] and IP Location-finding Tool [8]. Then, we obtained the Whois record to extract the registration information. Finally, we queried Passive DNS of 360 to get the aggregated number of DNS requests since the domains were registered. Appendix C lists these results and there are several interesting findings. First, all the sites were hosted in China instead of United States commonly chosen by cyber-criminals in China [8]. This setting however allows legal authorities to take over the servers more easily. Most of the domains were registered prior to 2016, suggesting the business has been running for a long time without legal disruption. The majority of the sites received a decent number of DNS queries (over 10K), and two sites were queried for more than 3 million times (No.19 and No.20).

[7] https://passivedns.cn/
[8] http://ip.chinaz.com/

TABLE XIII: Top 5 drug keywords searched by users.

No.	Keyword	Note	Volume
1	溜冰	Taking drug	9,387
2	冰妹	Prostitute taking drug together	7,255
3	咕噜咕噜	Sound from drug equipments	3,135
4	冰文	Articles about drug experience	72
5	真冰	Drug of high purity	58

We found the home pages of these two sites look legitimate and malicious URLs are on their sub-folders. We speculate the sites were either spammed or compromised.

Since these sites run eCommerce web applications, we were able to obtain the payment information regarding the merchants (in Appendix C). In particular, we added some products into the shopping carts to proceed to the checkout pages, which displayed merchants' payment information. Previous studies have revealed the payment methods employed for abuse-advertised goods (*e.g.*, Viagra) [32]. In our study, we identified several new payment methods, like Alipay, Tenpay and Haipay [9]. To our surprise, some sites showed merchants' names and banking account numbers directly on the web pages (No.3, 8, 12-15).

VI. Discussion

Limitations. Blackhat SEOers can adjust the operational model to keep black keywords hidden from KDES. They can move the black keywords from anchors to other sections under the SEO page to escape from our parser. They can suppress the number of poisoned search results under a black keyword. In addition, they could apply more intensive transformation on the keywords to reduce the frequency of the core words. However, all of these measures will introduce prominent side-effects: shifting the location of black keywords will reduce the relevance score counted by the search engines; poisoning less search result would lead to less incoming traffic and it requires coordination of independent SEOers; more transformation also reduces the incoming traffic, since the transformed keywords are used less often. In the meantime, we will continue to improve KDES to make it more resilient against the evasion performed by adversaries.

Our result mainly covers the black keywords and the underground economy in China, in part to the data we collected. We want to emphasize that the issue is not isolated. The measurement on other popular search engines like Google and Bing shows that their search results were also poisoned (see Section V-C), so they should spend efforts to purify the search result. We will continue our research to study the same phenomena under other languages (*e.g.*, English), regions and channels.

Dependency of implementation. We leveraged the internal APIs and scanners from Baidu to implementat KDES. These components enabled KDES to uncover a large quantity of black keywords in a short time. One would question how KDES performs when deployed by other organizations when the access to these components is not available. Our assessment is that the efficiency might be reduced (*e.g.*, more time will be consumed for search query without unthrottled API) but the effectiveness should not be affected. Each component outside the territory of KDES could be replaced by a component open to the public. For instance, the scanners from Baidu could be replaced by other public scanners, like VirusTotal [33].

Responsible disclosure and deployment. We reported all the detected black keywords (and the associated core words) to the security team of Baidu. The feedback so far is quite positive. Many black keywords have been confirmed and included by Baidu. Following this trail, we will keep KDES running and continuously send reports to Baidu. In summary, we envision four use cases that KDES could benefit search engines and other parties:

1) It helps search engines to regulate search queries containing black keywords.
2) The detected URLs can be divided into more precise categories. As an example, Baidu currently divides the URLs about the underground economy into three categories: lottery, pornography and fraud. The result can be refined with the help of KDES.
3) Search ads are regulated under stricter policies but we found the underground merchants were able to exploit the weakness of the screening process and sneak in illegal ads. KDES could assist auditors in understanding the real business behind the search ads and rejecting the illegal ones.
4) Other parties, like legal authorities, could learn the trend of the underground economy after digesting the black keywords discovered by KDES. Finding illegal activities in other channels will be also easier.

A large amount of search queries have been issued to Baidu during our study, which could incur some noticeable overhead on their servers. We tried to confine the overhead by scheduling the queries running in nighttime when the servers were used less intensively. The search result scanner fetches the page for a URL when there is no match in its cache. If the URL belongs to search ad, the advertiser might be charged unfairly. We excluded the URLs of search ads to avoid such charges.

VII. Related Work

Blackhat SEO. Blackhat SEO has been extensively leveraged to disseminate malicious/fraudulent content of cyber-criminals. A lot of efforts have been spent by search engine companies, security companies and academic institutes towards mitigating this issue. There are mainly two lines of works in this area. First, a plenty of works revealed the strategies of Blackhat SEOers, including constructing SEO botnet [34], spamming forums [35], compromising legitimate sites [36], text spin [37] and cloaking [38], [39]. Through exploiting the difference between promotion models of blackhat and whitehat SEO, the sites manipulating search rankings can be detected in large

[9]These payment methods are all similar to Paypal. While Alipay and Tenpay are supported by Alibaba and Tencent, giant IT companies in China, the provider of Haipay is less known to the public.

scale [36], [37], [40]. The second line of research focused on understanding and measuring the ecosystems of blackhat SEO business. Previous studies showed that pharmaceutical affiliate programs and stores selling fake products largely rely on blackhat SEO to reach potential buyers [41], [42], [43].

Our recent work *et al.* [8] uncovered a new blackhat SEO technique using wildcard DNS (spider pool) to tamper search rankings under long-tail keywords. This work consumed the data (SEO pages) from spider pools but the goal is different here. We aim at detecting *keywords* targeted by adversaries (black keywords), while the main focus of our last work is detecting *sites* boosted by blackhat SEO. As described in Section VI, finding and understanding black keywords could benefit many parties (*e.g.*, legal authorities) in addition to search engine companies. The result here facilitate the investigations in other channels (*e.g.*, social networks) since the vocabulary of black keywords is usually shared.

The underground economy. Understanding and measuring the underground economy is an active research area for a long time. There have been many works providing insights into the operational models and ecosystems of cybercrimes, including email spam [42], pay-per-install malware [44], unwanted software [45], [46], Twitter spam [47], illicit online pharmacies [48], [41] and etc. While the online anonymous marketplace, like Silk Road, has attracted a large number of buyers and sellers since 2011 [49], [50], "traditional" channels that are more "friendly" to general web users like forums [51], spam [52], and Blackhat SEO [43] are still playing important roles in the underground economy, especially in the regions outside of western countries [53], [54], [55]. To skip under the radar of legal authorities or security companies, the adversaries are constantly inventing new terms or obfuscating existing terms for communication. However, our work shows such terms can be effectively captured by their associated search results, even without understanding their semantic meanings.

Retrofitting search engine for detection. Previous works have demonstrated that search engines are helpful in detecting compromised and malicious sites. Invernizzi *et al.* proposed EvilSeed which automatically generates search queries by analyzing the seeding malicious web pages and the probability of finding malicious sites can be significantly increased [28]. Liao *et al.* studied promotional infection, an attack that injects advertising content into compromised sites [56]. Their study shows the semantics of the injected pages vastly differs from the semantics of hosting sites' sponsored top-level domains (sTLD). By querying the terms irrelevant to sTLD semantics, the promotional infection can be spotted effectively. Zhang *et al.* utilized search engine visibility (whether the site is indexed) as a factor to determine the maliciousness of a site [57]. In this work, we showed that the indexed pages of search engines can be used to discover trending terms used in underground markets. Such direction has never been explored before.

Search query abuse. While the public interface offered by search engines allows the security practitioners to discover sites involved by hackers, it unfortunately gives the hackers a lift to identify their targets. One common type of search query abuse is "Google Dork Query" [58], which models the fingerprint of a website and is submitted by attackers to harvest the URLs of websites built on top of specific website templates (*e.g.*, querying "Powered by WordPress" returns a list of sites developed with WordPress [59]). As a flurry of vulnerabilities have been discovered on website templates, the adversary is able to find vulnerable sites and craft relevant exploits more easily through this method. Although many Google Dorks have been published (*e.g.*, `exploit-db.com` [60]), the study by Zhang *et al.* [61] showed that only a handful of them were frequently exploited. Toffalini *et al.* [62] characterized the known dorks and showed dorks can be created in an automated fashion. Adversaries also leverage bot clients to send search queries at large volume to identify vulnerable site, harvest emails and scrape content from websites [63]. Black keywords and Google Dorks are both considered harmful by search engines but detecting the prior one is much more challenging, as there is no public reference for the ever-changing vocabulary and it is hard to analyze them using classic models like NLP.

VIII. Conclusion

Black keywords are frequently used for communication by underground economy while escaping from being tracked by law enforcement or being regulated by search engines. But capturing these black keywords are rather difficult, because they are quickly evolved and artificially obfuscated. In this work, we present the first approach towards detecting black keywords automatically. Our approach looks into the search results and uses several search engine features to determine their labels (black or not). By running our system, more than 400K black keywords were discovered and we found many keywords previously unknown to the public, indicating this approach is highly effective.

We believe our approach suggests a new direction in tackling security problems deemed hard by conventional approaches, like traditional NLP. In the future, we will go forth on this direction and explore the area of translating black keywords, which we think is feasible through big-data analytics. In the short term, we will collaborate with search engines and online communities, like Baidu, to build additional defense against illegal promotion by the underground economy.

IX. Acknowledgement

This work was supported by the Natural Science Foundation of China (grant No.U1636204 and 61472215) and sponsored by CCF-Venustech Hongyan Research Initiative (016-014). We thank anonymous reviewers for their insightful comments. We also owe a special debt of gratitude to Prof. James Mickens, for his instructive advice on our paper. We are deeply indebted to Jinjin Liang, Fengpei Li and Yiming Gong from Qihoo 360 for providing us passive DNS data for this research. Finally, special thanks should go to our colleagues from Baidu company who provide the platform support and data.

REFERENCES

[1] Heroin.net, "Heroin Street Names," http://heroin.net/about/street-names-for-heroin/, 2016.

[2] K. Thomas, D. Y. Huang, D. Wang, E. Bursztein, C. Grier, T. J. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna, "Framing dependencies introduced by underground commoditization," in *Proceedings of the Workshop on the Economics of Information Security*, 2015.

[3] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. G. Berbís, "Named entity recognition: Fallacies, challenges and opportunities," *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482–489, 2013.

[4] C. Napoles, A. Cahill, and N. Madnani, "The effect of multiple grammatical errors on processing non-native writing," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, 2016, pp. 1–11.

[5] X. Liao, C. Liu, D. McCoy, E. Shi, S. Hao, and R. A. Beyah, "Characterizing long-tail SEO spam on cloud web hosting services," in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 2016, pp. 321–332. [Online]. Available: http://doi.acm.org/10.1145/2872427.2883008

[6] M. A. Hearst, *Search User Interfaces*, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[7] ChinaZ, "One keypress to generate long-tail keywords online (translated)," http://s.tool.chinaz.com/longkeywords, 2016.

[8] K. Du, H. Yang, Z. Li, H. Duan, and K. Zhang, "The ever-changing labyrinth: A large-scale analysis of wildcard dns powered blackhat seo," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 245–262. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/du

[9] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," Stanford InfoLab, Technical Report 2004-25, March 2004.

[10] Baidu, "Related Search," http://baike.baidu.com/view/3429724.htm, 2016.

[11] Google, "Policies. Inside Search," https://www.google.com/insidesearch/howsearchworks/policies.html, 2016.

[12] Bing, "How Bing delivers search results," http://onlinehelp.microsoft.com/en-us/bing/ff808447.aspx, 2016.

[13] Google, "Removal Policies - Search Help," https://support.google.com/websearch/answer/2744324?hl=en, 2016.

[14] Baidu, "Baidu user service center (translated)," http://help.baidu.com/question?prod_en=webmaster&class=123&id=1503, 2016.

[15] Google, "AdWords policies. Advertising Policies Help," https://support.google.com/adwordspolicy/answer/6008942?rd=1, 2016.

[16] Bing, "Restricted and disallowed content policies," https://advertise.bingads.microsoft.com/en-us/resources/policies/restricted-and-disallowed-content-policies, 2016.

[17] Baidu, "Official site of Baidu Editor. Let the customers find you (translated)," http://e.baidu.com/help, 2016.

[18] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 368–378.

[19] P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1625–1628.

[20] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short text understanding through lexical-semantic analysis," in *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 495–506.

[21] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, and C. Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines," *Future Generation Computer Systems*, vol. 37, pp. 468–477, 2014.

[22] J. M. Wissner and N. T. Spivack, "Generating user-customized search results and building a semantics-enhanced search engine," May 19 2015, uS Patent 9,037,567.

[23] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 175–191. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/melicher

[24] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu, "A comprehensive study of named entity recognition in chinese clinical text," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 808–814, 2014.

[25] NLTK Project, "Natural Language Toolkit," http://www.nltk.org/, 2016.

[26] npm, "nlp-toolkit," https://www.npmjs.com/package/nlp-toolkit, 2016.

[27] D. S. Hirschberg, "Algorithms for the longest common subsequence problem," *Journal of the ACM (JACM)*, vol. 24, no. 4, pp. 664–675, 1977.

[28] L. Invernizzi and P. M. Comparetti, "Evilseed: A guided approach to finding malicious web pages," in *2012 IEEE Symposium on Security and Privacy*. IEEE, 2012, pp. 428–442.

[29] Baidu, "Baidu Tieba - The world's largest Chinese online community (translated)," http://tieba.baidu.com/, 2016.

[30] ——, "Baidu Zhidao - The world's largest Q&A forum in Chinese (translated)," https://zhidao.baidu.com/, 2016.

[31] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the Use of Homophones in Domain Squatting," in *Information Security - 17th International Conference, ISC 2014, Hong Kong, China, October 12-14, 2014. Proceedings*, 2014, pp. 291–308.

[32] D. McCoy, H. Dharmdasani, C. Kreibich, G. M. Voelker, and S. Savage, "Priceless: The role of payments in abuse-advertised goods," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, ser. CCS '12. New York, NY, USA: ACM, 2012, pp. 845–856. [Online]. Available: http://doi.acm.org/10.1145/2382196.2382285

[33] VirusTotal, "Free Online Virus, Malware and URL Scanner," https://www.virustotal.com/, 2016.

[34] D. Y. Wang, S. Savage, and G. M. Voelker, "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. [Online]. Available: http://internetsociety.org/doc/juice-longitudinal-study-seo-botnet

[35] Y. Niu, Y.-M. Wang, H. Chen, M. Ma, and F. Hsu, "A quantitative study of forum spamming using context-based analysis," Microsoft Research, Tech. Rep. MSR-TR-2006-173, December 2006. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=70379

[36] J. P. John, F. Yu, Y. Xie, A. Krishnamurthy, and M. Abadi, "deseo: Combating search-result poisoning." in *USENIX security symposium*, 2011.

[37] Q. Zhang, D. Y. Wang, and G. M. Voelker, "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. [Online]. Available: http://www.internetsociety.org/doc/dspin-detecting-automatically-spun-content-web

[38] D. Y. Wang, S. Savage, and G. M. Voelker, "Cloak and dagger: dynamics of web search cloaking," in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 477–490.

[39] L. Invernizzi, K. Thomas, A. Kapravelos, O. Comanescu, J.-M. Picod, and E. Bursztein, "Cloak of visibility: Detecting when machines browse a different web," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2016.

[40] L. Lu, R. Perdisci, and W. Lee, "Surf: detecting and measuring search poisoning," in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 467–476.

[41] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko, "Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs," in *Proceedings of the 21st USENIX Conference on Security Symposium*, ser. Security'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 1–1. [Online]. Available: http://dl.acm.org/citation.cfm?id=2362793.2362794

[42] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage, "Click trajectories: End-to-end analysis of the spam value chain," in *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, ser. SP '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 431–446. [Online]. Available: http://dx.doi.org/10.1109/SP.2011.24

[43] D. Y. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. M. Voelker, "Search+seizure: The effectiveness of interventions on seo campaigns," in *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 2014, pp. 359–372.
[44] J. Caballero, C. Grier, C. Kreibich, and V. Paxson, "Measuring Pay-per-Install: The Commoditization of Malware Distribution," in *Proceedings of the 20th USENIX Security Symposium*, San Francisco, CA, USA, August 2011.
[45] P. Kotzias, L. Bilge, and J. Caballero, "Measuring PUP Prevalence and PUP Distribution through Pay-Per-Install Services," in *Proceedings of the 25th USENIX Security Symposium*, Austin, TX, USA, August 2016.
[46] K. Thomas, J. A. E. Crespo, R. Rasti, J.-M. Picod, C. Phillips, M.-A. Decoste, C. Sharp, F. Tirelo, A. Tofigh, M.-A. Courteau, L. Ballard, R. Shield, N. Jagpal, M. A. Rajab, P. Mavrommatis, N. Provos, E. Bursztein, and D. McCoy, "Investigating commercial pay-per-install and the distribution of unwanted software," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 721–739. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/thomas
[47] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse," in *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 195–210.
[48] N. Leontiadis, T. Moore, and N. Christin, "Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade." in *USENIX Security Symposium*, 2011.
[49] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 213–224. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488408
[50] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, Aug. 2015, pp. 33–48. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska
[51] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068824
[52] C. Kanich, N. Weavery, D. McCoy, T. Halvorson, C. Kreibichy, K. Levchenko, V. Paxson, G. M. Voelker, and S. Savage, "Show me the money: Characterizing spam-advertised revenue," in *Proceedings of the 20th USENIX Conference on Security*, ser. SEC'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 15–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=2028067.2028082
[53] Z. Jianwei, G. Liang, and D. Haixin, "Investigating China's online underground economy," 2012.
[54] G. Guzman, "Hiding in Plain Sight: THE GROWTH OF CYBERCRIME IN SOCIAL MEDIA, PART 1," https://www.rsa.com/content/dam/rsa/PDF/Growth-of-Cybercrime-in-Social-Media_WhitePaper.pdf, 2016.
[55] ——, "Hiding in Plain Sight: THE GROWTH OF CYBERCRIME IN SOCIAL MEDIA, PART 2," http://blogs.rsa.com/wp-content/uploads/2016/04/WP_Hiding_in_Plain_Sight-Part_2_reduced.pdf, 2016.
[56] X. Liao, K. Yuan, X. Wang, Z. Pei, H. Yang, J. Chen, H. Duan, K. Du, E. Alowaisheq, S. Alrwais, L. Xing, and R. Beyah, "Seeking nonsense, looking for trouble: Efficient promotional-infection detection through semantic inconsistency search," ser. IEEE Security and Privacy 2016, 2016.
[57] J. Zhang, X. Hu, J. Jang, T. Wang, G. Gu, and M. Stoecklin, "Hunting for invisibility: Characterizing and detecting malicious web infrastructures through server visibility analysis," in *Proceedings of the 2016 IEEE International Conference on Computer Communications*, ser. INFOCOM 2016. Washington, DC, USA: IEEE Computer Society, 2016.
[58] J. Long, E. Skoudis, and A. v. Eijkelenborg, *Google Hacking for Penetration Testers*. Syngress Publishing, 2004.
[59] WordPress.com, "Create a free website or blog," https://wordpress.com/, 2016.
[60] E. Database, "Google Hacking Database (GHDB)," https://www.exploit-db.com/google-hacking-database/, 2016.
[61] J. Zhang, J. Notani, and G. Gu, "Characterizing google hacking: A first large-scale quantitative study," in *Proceedings of the 10th International Conference on Security and Privacy in Communication Networks (SecureComm'14)*, September 2014.
[62] F. Toffalini, M. Abba, D. Carra, and D. Balzarotti, "Google dorks: analysis, creation, and new defenses," in *DIMVA 2016, 13th Conference on Detection of Intrusions and Malware & Vulnerability Assessment, July 7-8, 2016, San Sebastian, Spain*, San Sebastian, SPAIN, 07 2016. [Online]. Available: http://www.eurecom.fr/publication/4892
[63] J. Zhang, Y. Xie, F. Yu, D. Soukal, and W. Lee, "Intention and origination: An inside look at large-scale bot queries." in *NDSS*, 2013.
[64] ppgwebsolutions, "What is Search Engine Marketing?" http://ppgwebsolutions.com/search-engine-marketing/, 2016.
[65] WordStream, "Google Ads: What Are Google Ads and How Do They Work?" http://www.wordstream.com/google-ads, 2016.

APPENDIX

A. Background about search ads

Search engine marketing (SEM) is the major revenue source for a search engine. It allows site owners to increase visibility through payment. Different from organic search results whose ranks are determined by sites' importance and relevance, the rankings of the paid sites (*i.e.*, search ads) mostly depend on the amount of spendings. To the newer sites whose reputations have not been accumulated, SEM is a convenient approach to boost incoming traffic in a short time. So far, four types of SEM products are offered, including paid inclusion (pay to be included in search index), paid placement (similar to paid inclusion but top rankings are guaranteed), local search ads (sites are shown when users are residing in or querying specific locations) and product listing ads (ads from merchants allowing display of product image and price) [64]. Typically, site owners choose a list of keywords which are relevant to their business as target. For example, a flower store might target the terms "roses", "cheap bouquet" or "wedding flowers". Then they bid on these keywords and make payments based on the number and quality of clicks going through their ads [10], [65].

To avoid dampening its reputation, the search engine checks whether the search ad is relevant to the targeted keywords, ad's quality and the consistency between ad and landing page. Ads failing these basic checks will be asked for adjustment, but ads promoting illegal products are rejected without consideration. However, due to the freedom granted by search engines in keywords selection, promotion under black keywords is possible and we have identified several such cases (described in Section V-C). As one example, surprisingly, we found spider pool services are listed in Baidu's search ads (see Figure 12), and this happens because the marketing personnels are not aware that spider pool belongs to the underground economy, as we learned from them. Auditing the sites behind the search ads is not a promising solution, as adversaries could apply cloaking techniques to conceal the real content. To mitigate this issue, our approach provides an alternative solution which could significantly reduce the delay for discovering illegal search ads.

B. URL scanner

The scanner we integrated into `KDES` runs on Hadoop + MapReduce and uses multiple detectors against a URL. The label of the URL is determined by the combined results from the detectors. Figure 14 illustrates how the detectors work

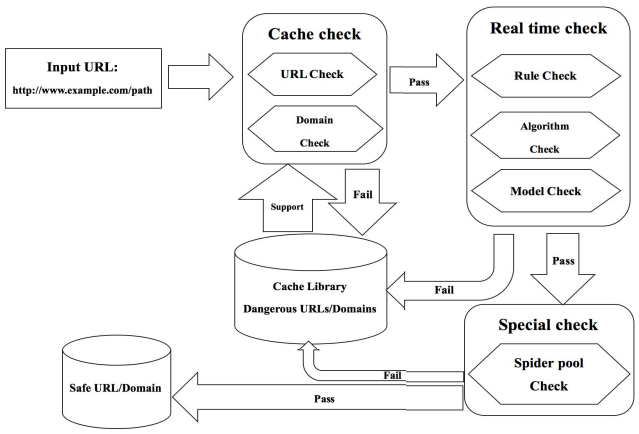

Fig. 14: Flow chart of the URL scanner.

together. We divide the scanning process into 3 phases. In the first phase, the URL is checked against a blacklist of URLs and domains. If the URL is not alarmed, the page of the URL will be fetched and checked by three real-time detectors running sequentially, including rule-based, algorithm-based and model-based detectors. Rule-based detector checks the existence of known black keywords and their frequencies on the page. To notice, the new black keywords identified by `KDES` will be imported to this detector after confirmed by the analysts. Algorithm-based detector renders the page to spot the suspicious behaviors and analyzes all images to capture the porn ones. Model-based detector classifies the page based on a set of models trained through machine learning. The result of this phase depends on the output of all the three detectors. The final phase consists of the spider-pool detector which checks whether the page is built for blackhat SEO purposes. The URLs and their enclosed domains are added to the blacklist used by the first phase in the end.

C. Registration and IP information of sites selling drugs

We queried IP and location of these sites that sold drugs, and list them in Table XIV. From this table we can see that half of them are located in HongKong, and half in China mainland. There are 5 sites sharing the same IP.

TABLE XIV: Sites selling drug equipments.

No.	URL	IP Address	Location	Reg time	Reg Email
1	http://www.kxyj.net/	122.10.114.24	HongKong	May, 14th, 2014	8*8@szsm.net
2	http://www.gdyc.net/	122.10.114.24	HongKong	October, 14th, 2014	8*8@szsm.net
3	http://www.nryj.com.cn/	122.10.114.24	HongKong	March, 27th, 2014	8*8@szsm.net
4	http://www.shopqt.com/	122.10.114.24	HongKong	May, 8th, 2010	p*g*u@vip.qq.com
5	http://liubinghu.cn/	122.10.114.24	HongKong	April, 5th, 2014	8*8@szsm.net
6	http://www.bhzmd.com/	122.10.114.6	HongKong	April, 26th, 2013	16*15*40@qq.com
7	http://yjwg.net/	122.10.114.6	HongKong	April, 11st, 2014	8*8@szsm.net
8	http://www.dupinbinghu.com/	110.173.55.221	HongKong	May, 6th, 2013	11*87*76*0@qq.com
9	http://www.hongyunyj.com/	58.64.206.182	HongKong	Jan, 10th, 2014	93*24*54*@qq.com
10	http://www.gbinghu.com/	58.64.204.41	HongKong	May, 15th, 2015	28*14*75*@qq.com
11	http://www.lfcryj.com/	103.61.241.176	HongKong	July, 18th, 2014	32*85*646@qq.com
12	http://www.binghu6.com/	123.57.216.241	Guangdong China	June, 25th, 2015	14*16*24*1@qq.com
13	http://www.666binghu.com/	121.43.149.146	Chengde China	Feb, 23th, 2016	14*16*24*1@qq.com
14	http://www.xiaogulu777.com/	121.43.149.146	Chengde China	Sep, 15th, 2014	Yu*in*@YinSiBaoHu.AliYun.com
15	http://www.vipbinghu.com/	121.41.13.189	Fuzhou China	Feb, 24th, 2016	14*16*24*1@qq.com
16	http://www.hookahweb.com/	121.41.14.29	Fuzhou China	Oct, 28th, 2015	Yu*in*@YinSiBaoHu.AliYun.com
17	http://www.chihuo.co/	61.160.224.188	Changzhou China	Dec, 3rd, 2013	8*8@szsm.net
18	http://www.sfy888.com/	182.61.64.91	Beijing China	Sep, 25th, 2012	ji*hi*n@163.com
19	http://www.100ye.com/11274746	221.234.43.212	Wuhan China	Sep, 26th, 2009	xu*iq*ng@126.com
20	http://www.yxool.com/kblzfsdi/	125.88.190.22	Guangdong China	April, 17th, 2011	tr*ns*er*e@51hkidc.com

TABLE XV: Payment methods of sites selling drug equipments.

No.	URL	DNS queries	Alipay	Tenpay	Haipay	Online Banking	Wired transfer	Cash on Delivery
1	http://www.kxyj.net/	18,886		√	√			√
2	http://www.gdyc.net/	5,682		√	√			
3	http://www.nryj.com.cn/	14,063		√	√	√	√	
4	http://www.shopqt.com/	21,335	√	√	√			
5	http://liubinghu.cn/	10,083		√	√			
6	http://www.bhzmd.com/	24,723		√	√			√
7	http://yjwg.net/	76,933	√	√	√			
8	http://www.dupinbinghu.com/	2	√				√	
9	http://www.hongyunyj.com/	33,062						
10	http://www.gbinghu.com/	0	√				√	
11	http://www.lfcryj.com/	4,439						
12	http://www.binghu6.com/	14,142				√	√	√
13	http://www.666binghu.com/	1,824				√	√	√
14	http://www.xiaogulu777.com/	19,555	√			√	√	√
15	http://www.vipbinghu.com/	1,592				√	√	√
16	http://www.hookahweb.com/	374						
17	http://www.chihuo.co/	12,670		√	√			
18	http://www.sfy888.com/	12,798	√					√
19	http://www.100ye.com/11274746	11,312,768						
20	http://www.yxool.com/kblzfsdi/	3,658,148						