Mingling of Clear and Muddy Water: Understanding and Detecting Semantic Confusion in Blackhat SEO

Hao Yang[1], Kun Du[1], Yubao Zhang[2], Shuai Hao[3], Haining Wang[4], Jia Zhang[1,⋆], and Haixin Duan[5,6]

[1] Tsinghua University, Beijing, China
yang-h16, dk15@mails.tsinghua.edu.cn, zhangjia@cernet.edu.cn
[2] University of Delaware, Newark, DE, USA
ybzhang@udel.edu
[3] Old Dominion University, Norfolk, VA, USA
shao@odu.edu
[4] Virginia Tech, Arlington, VA, USA
hnw@vt.edu
[5] Institute for Network Science and Cyberspace, Tsinghua University
[6] Qi An Xin Group Corp.
duanhx@tsinghua.edu.cn

Abstract. Search Engine Optimization (SEO) is a set of techniques that help website operators increase the visibility of their webpages to search engine users. However, there are also many unethical practices that abuse ranking algorithms of a search engine to promote illegal online content, called blackhat SEO. In this paper, we make the first attempt to systematically investigate a recent trend in blackhat SEO, semantic confusion, which mingles the content of a webpage to deceive existing detection of blackhat SEO. In particular, from a new perspective of content semantics, we propose an effective defense against the semantic confusion based blackhat SEO. We built a prototype of our defense called SCDS, and then we validated its effectiveness based on 4.5 million domains randomly selected from 11 zone files and passive DNS records. Our evaluation results show that SCDS can detect more than 82 thousand blackhat SEO websites with a precision of 98.35%. We further analyzed 57,477 long-tail keywords promoted by blackhat SEO and found more than 157 SEO campaigns. Finally, we deployed SCDS into the gateway of a campus network for ten months and detected 23,093 domains with malicious semantic confusion content, showing the effectiveness of SCDS in practice.

1 Introduction

Search engines are the entrance to the Internet, and search engine optimization (SEO) helps legal service/content providers improve their page ranking to be more visible to Internet users [12,11,9]. However, underground online business

⋆ Corresponding Author

such as gambling and porn may be disallowed to blatantly exhibit their content because it is banned according to the laws in some countries and regions. Thus, they adopt blackhat SEO to promote themselves. Blackhat SEO [8] is a set of unethical practices of search engine optimization, such as content spam [20], and it aims to make a page's ranking rise in a short time, no matter what kind of content is on the page. In the past decades, blackhat SEO has become the tailor-made technique for promoting underground online business.

To avoid being detected by search engines, a spate of blackhat SEO webpages recently disguise themselves by leveraging *semantic confusion*. That is to say, these webpages, on one hand, include promotion contents since their ultimate goal is to promote illegitimate products or services. On the other hand, they also mingle with some legitimate content, which is usually copied from other sources, in order to disguise themselves. Therefore, there exists a clear semantic discrepancy within these blackhat SEO webpages.

In this paper, we investigate this recent trend in blackhat SEO webpages that leverage semantic confusion to cloak illegal topics webpages. We propose an effective defense called Semantic Confusion Detection System (`SCDS`) to detect these blackhat SEO webpages based on semantic discrepancy. In particular, we developed two separate deep-learning classifiers to measure two coordinates with regard to legitimacy and illegitimacy, respectively. With the coordinate system built, we can identify webpages with semantic discrepancy and further check the external hyperlinks on these pages to detect blackhat SEO webpages.

Contributions. We summarize the major contributions of this work as follows:

(1) Understanding of the semantic confusion practice in blackhat SEO. We made the first attempt to systematically investigate the semantic confusion in the context of blackhat SEO. We revealed that semantic confusion has been widely adopted in practice for evading detection, posing a serious security threat to Internet users.

(2) Development of effective defense. We proposed and implemented a novel detection system, `SCDS`, which exploits the mingling of semantic content for accurate detection by recognizing the topics and semantic context of the text on a webpage. The evaluation results show that `SCDS` can effectively detect semantic confusion based blackhat SEO pages.

(3) Deployment and disclosure. We deployed our detection system on the gateway of a campus network and presented what we found to experienced security practitioners from the industry. Our findings have been confirmed and added into blacklists as seed for broader filtration.

The remainder of this paper is organized as follows. In Section 2, we review the background of SEO and its manipulation. In Sections 3 and 4, we detail the architecture of `SCDS` and its implementation and evaluation, respectively. In Section 5, we present the measurement results and analysis of the ecosystem of blackhat SEO. In Section 6, we describe the practical issues of blackhat SEO. In Section 7, we discuss several related issues of this study. In Section 8, we survey related work, and finally, we conclude in Section 9.

2 Background

2.1 Search Engine Optimization (SEO)

The purpose of SEO is to increase the exposure of websites to search engine users. Websites provide content with valid semantics in tags such as `<title>` and `<meta>`, provide a valid sitemap to help crawlers retrieve content quickly and refresh the content regularly. Search engines retrieve specific content from HTML pages, and match the content with keywords from users' input. Search engines use PageRank (PR) values to evaluate a website's relevance to keywords. To avoid being abused, a search engine changes its PR algorithm from time to time [24]. Every year, search engine vendors publish guidelines for SEO [13]. Search engine vendors encourage benign SEO because this can help the spider of a search engine to crawl more effectively and help the PR algorithm to rank a page more appropriately. The SEO encouraged by search engine vendors is often called "whitehat SEO".

2.2 Blackhat SEO

While whitehat SEO advocates improving the structure of a website to make it more friendly to search engines, different unethical techniques, called *blackhat SEO*, have been exploited to manipulate ranking results on search engines. The most common practice of blackhat SEO is to directly repeat the keywords or phrases in the webpages to increase their appearance and relevance to certain terms being promoted [20]. However, such an approach can be easily recognized and penalized by search engines through detecting repeated content. Furthermore, to engage more efficient promotion, a link farm has been constructed to accumulate incoming links to a website by exploiting the vulnerabilities of other reputable websites to inject a large number of links [6]. We discuss more details about cloaking pages in blackhat SEO in Section 5.1.

2.3 Semantic-based Techniques

Beyond the straightforward content manipulation, sophisticated, semantic-based content spam tied with the underground business, including illegal online gambling and pornography, has become more pervasive for evading detection. At a lower level (*i.e.*, word level), blackhat SEOers leverage automated spinning to avoid duplicate detection [30]. With the spinning, texts with similar semantic meanings but different appearances are generated and inserted into webpages.

At a higher level (*i.e.*, semantic level), blackhat SEOers fetch a large piece of normal content from legitimate websites, elaborately stuffed with a small piece of content from underground business, and assemble them into one webpage. As such, the page would be treated as a normal webpage, and the promoted content of underground business would be indexed by search engines. We call this more advanced manipulation in blackhat SEO as semantic confusion. In this study, we aim to conduct a comprehensive investigation to understand this new trend of blackhat SEO and explore the detection of semantic confusion.

Fig. 1: A typical webpage of illegal gambling site.

Fig. 2: A typical webpage of mixed content.

3 Semantic Confusion Detection

In this section, we present the architecture of SCDS and detail its components, including data source, data processor, semantic analyzer, and SEO collector.

3.1 System Overview

Our goal is to identify blackhat SEO webpages with semantic confusion, in which the illegitimate content is embedded for promotion while the legitimate content is compiled into the pages for the evasion of detection mechanisms. With regard to the legitimacy of content, we classify webpages into three different categories: normal (completely legitimate), semantic confusion (partially illegitimate), and underground (completely illegitimate).

For example, Figure 1 shows a webpage of underground economy for illegal online gambling, which is designed for attracting visitors to play online gambling games. Figure 2 shows a webpage with semantic confusion. One part of the content is associated with education that is normal, while the other part is associated with illegal online gambling. As mentioned earlier, such a practice has been prevalent in blackhat SEO to circumvent state-of-the-art detection mechanisms such as [8].

To address this emerging threat, we propose a novel approach for classifying blackhat SEO webpages through the detection of semantic confusion. Figure 3 depicts the workflow of our proposed Semantic Confusion Detection System (SCDS), including data processing, semantics analysis, and SEO collection. In the data processing module, we build a crawler to collect the content of webpages on a large scale, and parse those pages to extract semantic-related content for further analysis. Next, such content is fed into the semantics analyzer module to determine the *semantic context (illegal and normal topics)* of webpages.

If more than one semantic contexts are recognized in a webpage at the same time, we consider the webpage possessing semantic confusion and further inspect whether one of semantic contexts of the webpage is associated with an underground business. As a result, we can effectively identify blackhat SEO pages

Table 1: Summary of datasets.

Data	Source	Purpose	Period	# Count
$Data_{news}$	THUCNEWS	Training I†	2005-2011	740,000
$Data_{nor/ugd}$	Baidu	Training II‡	2019	130,000
$Domain_{zone}$	Verisign & ICANN	Testing	2020/03	3,000,000
$Domain_{pdns}$	Farsight	Testing	2020/02 - 04	1,500,000
$Domain_{All}$	—	Testing	—	4,500,000

† Normal Semantic Coverage. ‡ Normal/Underground Classification.

that aim to promote illegitimate content with semantic confusion to deceive the search engine. Previous studies (*e.g.*, [19]) typically rely on the results from search engines to detect blackhat SEO and only focus on the pages under domain names with specific semantics (*e.g.*, `.edu` and `.gov`). However, the results from search engines may introduce inaccuracy due to outdated entries and may also cause bias when the results of search engines are incomplete or manipulated. Our method identifies blackhat SEO pages mainly based on the page content and extends the detection scope to all domains, which could significantly improve efficiency for blackhat SEO detection. In the SEO extension module, we attempt to collect more blackhat SEO pages by leveraging the elements that may present the correlation of blackhat SEO activities. Specifically, we consider two methods to extensively discover potential SEO pages: (1) we extract the external links from identified blackhat SEO pages, and (2) we use a set of blackhat SEO domains as a seed set and search the URLs under these domains that have been indexed by the search engine. We then recursively check whether these pages are also semantic-confusion-based blackhat SEO pages through the semantic analyzer.

3.2 Datasets

A summary of the datasets used in this work is listed in Table 1.

Training Datasets. The recognition of semantic contexts in webpages requires that our training dataset has complete coverage of various topics to the greatest extent possible. We collected two separate datasets for different training purposes: one for normal semantic context recognition and the other for underground economy detection.

(1) For the training dataset of semantic context classification, we use the THUCNEWS[7] dataset, which collected 740,000 pages from the RSS subscription channel of Sina News, one of the most popular online news sites in China. The dataset, labeled as $Data_{news}$ in our study, covers 14 general news top-

[7] The dataset that has been widely used in text-related studies (`http://thuctc.thunlp.org/` [in Chinese]). Note that although the dataset itself was compiled based on the News pages from 2005 to 2011, the semantics of the language remains significantly stable and the accuracy of text classification holds too.

ics, including sports, entertainment, housing, home decoration, fashion, politics, gaming, society, science, finance, *etc.*

(2) For the training dataset of underground economy detection, under the help of Baidu, we collected 100,000 normal webpages and 30,000 webpages of underground economy that have been explicitly labeled by their search engine (*i.e.*, 15,000 illegal online gambling webpages and 15,000 pornographic webpages). We labeled the second dataset as $Data_{nor/ugd}$.

Testing Datasets. Liao *et al.* [19] reported that webpages with semantic inconsistency were found in all different kinds of top-level domains (TLDs). To obtain good coverage of TLDs, we collected the registered domains from 11 different TLDs from which zone files are available. Since this dataset is only for testing and there is no need to crawl the webpages from all domains, we sampled 1.64% of domains from each zone file and obtained 3,000,000 domains in total, labeled as $Domain_{zone}$. Moreover, to enrich the diversity of the testing datasets, we also retrieve DNS records through Farsight [10] DNS database APIs. Since detecting all domains is time-consuming, we select 1,500,000 domains from all `A` records within 2020, for producing a sufficient coverage within a reasonable time, labeled as $Domain_{pdns}$. The entire testing dataset is labeled as $Domain_{All}$.

3.3 Data Processor

With the domains listed in Table 1, we developed a crawler to obtain the webpages in each domain. We processed the content of webpages in the following steps: (1) In this research, we only consider the text content of HTML pages, so we first stripped JavaScript, CSS, and comments embedded in webpages that are irrelevant to our semantic recognition. (2) We extracted the text from HTML pages (including all text in head and body elements) and removed non-breaking spaces. Note that many pages may play "tricks" with hidden elements (*e.g.*, in `<div>` labels) including underground economy content to evade detection by normal users. Those hidden elements can be collected in this step. (3) Since most of the pages in our training and testing datasets are in Chinese, we performed word segmentation on the extracted text. It is worthwhile to note that our method is also applicable to webpages in English (or other languages) with no need of word segmentation. (4) We removed the stop words, as well as the words that appear only once, which dramatically decreases the scale of words and mitigates the over-fitting problem. (5) For the rest of the words on the page, we put them together to form a new piece of text for semantic analysis.

3.4 Semantic Analyzer

We are interested in two different semantic contexts: *i.e.*, legitimate topics (*e.g.*, sports and finance) and underground or illegal topics (*e.g.*, illegal online gambling and porn). Accordingly, we developed two independent classifiers to measure the semantic affinities of webpages to legitimate topics and underground topics, called a *normal topics* classifier and a *illegal topics* classifier, respectively.

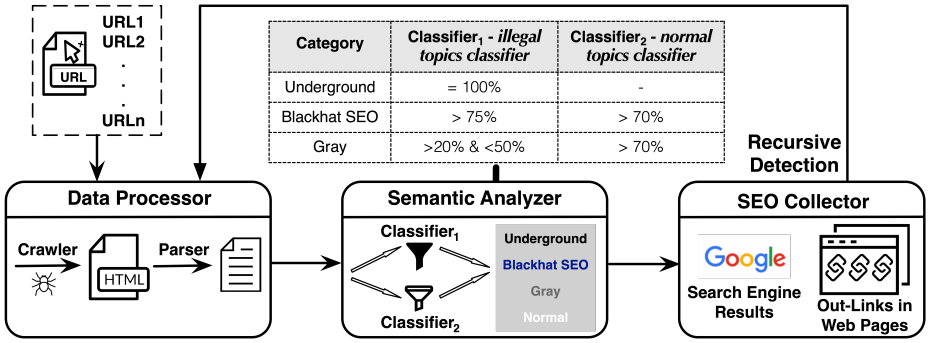

Fig. 3: Architecture of SCDS.

The illegal topics classifier, labeled as Classifier$_1$, is trained with the dataset $Data_{nor/ugd}$, for distinguishing the normal topics from underground topics. It produces the possibility that a webpage could be associated with an underground business. The normal content classifier, Classifier$_2$, is trained with the dataset $Data_{news}$ and is used for identifying the normal topic to which the webpages belong. This classifier outputs the possibilities that a webpage belongs to each normal topic, and the sum of all the possibilities is 100%. The topics include sport, entertainment, housing, lottery, home decoration, fashion, politics, games, society, science and technology, stocks, finance, *etc.* We compare six different classification algorithms, including Naive Bayes, Logistic Regression, Random Forest, RNN, FastText and TextCNN. We use F1-measure for measuring the classification accuracy. Based on the classification results, we select TextCNN to build our classification model.

Based on the outputs of the two classifiers above, we define four categories of webpages in our study:

(1) **Underground economy page**. The content on the page is completely associated with underground economy. In this case, the output of Classifier$_1$ is positive, with a probability of 100%.

(2) **Blackhat SEO page**. The webpage contains both normal semantic context and underground economy semantic context. We found that most pages of this kind leverage semantic confusion for blackhat SEO purposes. In this case, the output of Classifier$_1$ is positive, with a probability of between 50% and 100%; and the output of Classifier$_2$ is that the web page belongs to a specific topic, with a probability greater than 70%. The selection of these threshold values is illustrated in Section 4. Note that this type of webpages is the primary target of our detection system.

(3) **Gray page**. This kind of webpage also has semantic confusion that is similar to blackhat SEO page. Gray page differs from blackhat SEO page in terms of the output of Classifier$_1$, which has a positive probability ranging from 20% to 50%. It indicates that gray pages have less content associated with underground economy compared to blackhat SEO pages. The threshold value we

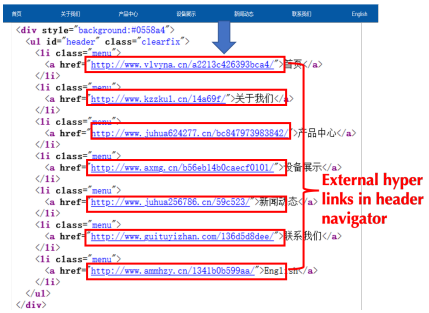

Fig. 4: A blackhat SEO page with many external links.

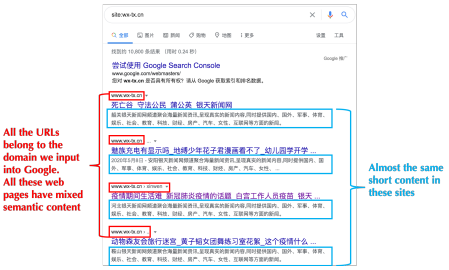

Fig. 5: Search results with a specific keywords in Google.

set for Classifier$_1$ is illustrated in Section 4. The reason why we classify blackhat SEO pages and gray pages into different categories based on the output of Classifier$_1$, lies in that these two kinds of pages have distinct goals. We will further discuss the differences in Section 5.

(4) **Normal page**. If (i) the probability of Classifier$_1$ output is less than 20% and (ii) the Classifier$_2$ indicates the webpage belongs to a normal topic with a probability of greater than 80%, we consider the webpage as a normal page.

Note that the primary targets of our detection are blackhat SEO pages and gray pages, because underground economy pages and normal pages are relatively easy to identify.

3.5 SEO Collector

Next, with the blackhat SEO pages detected in the semantic analyzer, the SEO collector expands the detection by recursively fetching more candidate pages from the identified SEO pages. Those extensive candidate pages are then fed back into the semantic analyzer, which will determine whether the input is a blackhat SEO page. The extension for collecting more candidate pages is achieved through the following two approaches:

(1) **Extension based on hyperlinks on webpages**. Figure 4 shows an example of a typical blackhat SEO page. We can see that the blackhat SEO pages tend to include many hyperlinks pointing to other blackhat SEO pages, which confirms the observation reported in [8]. In this study, we leveraged this feature for expanding the detection of blackhat SEO pages. Specifically, we crawled the external links in the detected blackhat SEO pages and performed the recursive detection to these newly collected pages. Moreover, if the semantic analyzer labels an expended webpage as a blackhat SEO page, it can be further expanded recursively.

(2) **Extension based on search engine results**. As mentioned above, the URLs/webpages under an SEO *domain* would also be likely to contain SEO content. Modern search engines provide an advanced feature by which the URLs under one domain can be easily found. For example, one can search

"site:wsj.com" in Google, which will return all URLs under wsj.com that have been indexed by Google. Other search engines like Baidu and Bing also provide similar services. Figure 5 shows the search results of "site:worfwx-tx.cn" in Google, in which the domain wx-tx.cn has been identified as a blackhat SEO domain. We observed that the search results exhibit a number of URLs that have similar content and page layouts. According to the discussion above, the URL under an SEO domain has a high possibility of containing SEO content. Based on the results of search engines, we can expand our detected blackhat SEO pages effectively.

With these two extension methods above, hyperlinks based extension can detect the blackhat SEO pages hosted by different domains, while search engine based extension can detect blackhat SEO pages from the same domain. All of those URLs are also classified as "blackhat SEO pages" by our analyzer, showing that we can effectively expand our detection in practice.

4 Implementation and Evaluation

In this section, we detail the implementation of SCDS, the selection of parameters, and the evaluation of its effectiveness.

4.1 Implementation

Training and Detection We trained Classifier_1 and Classifier_2 with the datasets of $Data_{nor/ugd}$ and $Data_{news}$, respectively. First, we extracted text content from those webpages using BeautifulSoup [1], processed text segment with Jieba [2], and removed the blank words and stop words. Then, we employed the Keras library [3] and TextCNN [5] to train our classifiers with a 3:1 ratio of the training data and testing data. The training process was performed on a server with 64GB memory and 12 cores E5-CPU and took about 45 hours. As a result, Classifier_1 achieves an accuracy of 99.993%, and Classifier_2 achieves an accuracy of 96.73%.

As mentioned earlier, we randomly collected 4,500,000 e2LDs (effective second-level domains) from zone files and passive DNS databases to compose the testing dataset $Domain_{All}$. We divided these domains into 10 groups, and deployed 10 crawlers built with Python Selenium Library [4] to crawl the HTML pages of these domains in 10 independent PCs equipped with 32GB memory and 1TB disk. The crawling of webpages was conducted from April 10 to April 15, 2020. We obtained 3,231,942 valid webpages. Note that we could not crawl valid content from the rest of 1.27 million domains because most of them disabled the standard web ports (*e.g.*, 80, 8080, or 443), and a few of them became expired during our experiments. Finally, we used Classifier_1 and Classifier_2 to recognize the blackhat SEO webpages and gray pages.

4.2 Evaluation

Parameter Selection We describe our selection of key parameters in SCDS.

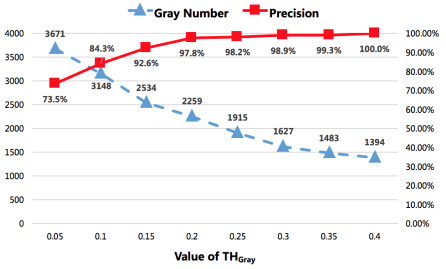

Fig. 6: Impact of different TH_{Gray}.

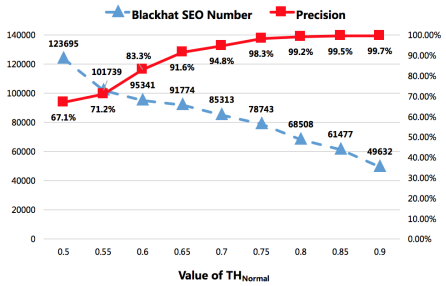

Fig. 7: Impact of different TH_{Normal}.

Gray bound TH_{Gray}. This parameter is the lower bound for Classifier$_1$ of the semantic analyzer. In other words, if the possibility of input text is greater than TH_{Gray}, Classifier$_1$ will determine that the corresponding webpage includes the content of semantics associated with an underground business. To decide a proper threshold, we empirically set a value from 0.05 to 0.4 with a step of 0.05, and randomly selected 1,000 samples to run the classification process and check its effectiveness. The results of precision, along with the number of total gray pages identified by SCDS under the corresponding parameter, are plotted in Figure 6. We set TH_{Gray} to 0.2 since it produces a very high precision (97.8%) while effectively capturing the most gray pages in our dataset. More specifically, as the threshold increases greater than 0.2, the increase of precision flats, but at the same cost of missing more pages that actually belong to the gray category.

Normal semantic bound TH_{Normal}. This parameter sets the lower bound for Classifier$_2$ of the semantic analyzer to check if the input text is normal. For an input text, if one of the topics in CNEWS is more than TH_{Normal}, we checked it as content with a normal topic. First, we set the parameter ranging from 0.50 to 0.9 with a step of 0.05, then sampled 1,000 check results to confirm the precision. The statistics are shown in Figure 7. Based on the same configuration principle of TH_{Gray}, we set TH_{Normal} to 0.75, achieving a very high precision (98.3%). The increase of precision flats afterwards.

SEO Extension We then deployed our detection model on 10 PCs and evaluated with the testing dataset $Domain_{All}$, which contains 3,231,942 webpages. Our detection model identifies 75,288 blackhat SEO pages with gambling content and 3,455 blackhat SEO pages with pornographic content, as well as 2,259 gray pages (each blackhat SEO or gray page is with an individual domain). Because of the time limitation, we could not query all the domains in a search engine. We randomly selected 2,000 detected blackhat SEO domains as the input of the SEO Collector module to expand the SEO detection. These 2,000 domains are divided into two parts. (1) We extracted 1,000 domains and searched them in Google. By collecting the top 10 results of each domain, we obtained 6,117 search results, 5,373 of which are valid webpages. We confirmed that 5,191 of them are blackhat SEO pages (including 4,007 gambling-related webpages and 1,184 porn-related webpages). We manually checked the other 182 pages and found that all

Table 2: Detection result of SCDS.

	# Domain	# URL	#Gambling related	# Porn related	# Type of SEO pages		
					Link	iframe	Cloaking
Blackhat SEO	82,061	100,792	78,079	3,982	70,727	4,161	7,173
Gray	2,259	2,259	1,809	450	-	-	-
Total	84,320	103,051	79,888	4,432	70,727	4,161	7,173

of them are semantic confusion webpages, but the text that causes confusion is also with normal topics and no content of underground economy is included. (2) We extracted 23,317 out-links from the other 1,000 domains and confirmed that 16,857 of them are blackhat SEO pages (including 13,791 gambling-related webpages and 3,066 porn-related webpages), which belong to 3,318 e2LDs.

In total, we expanded our detection results with 100,791 blackhat SEO pages, which belong to 82,061 e2LDs. The detection results are shown in Table 2.

Evaluation Results Due to the lack of ground truth, we manually inspected the results to validate our detection mechanism, which is also a common method used in previous studies [28]. In doing so, we randomly sampled 1,000 pages from the results of blackhat SEO and gray page detection, respectively. In order to ensure precision, two experienced researchers investigated the results independently. The guidelines of our evaluation process are as follows: (1) For gray pages, they inspected whether the page contains underground economy semantics and normal semantics simultaneously. (2) For blackhat SEO pages, they checked the mixed semantics in (1) as well as whether the pages contain hyperlinks that point to external pages, and then checked these domains in search engines to confirm if underground economy content is also indexed by search engines. In our evaluation, when both researchers agreed on a webpage being classified as an SEO page or a Gray page, we determine that the classification result is correct; otherwise, we consider the classification as a false positive case. In the end, we identified 989 blackhat SEO pages and 978 gray pages, with an precision of 98.35%. Furthermore, we deployed the detection system on our campus network for 10 months, and the detection results have been acknowledged by IT security department, justifying the effectiveness of our detection system.

5 Measurement

In this section, we present our measurement results of detecting mixed semantic pages. We first characterize blackhat SEO domains and perform an analysis on both keywords and content extracted from blackhat SEO pages. We then cluster the detected domains by external links to infer the SEO campaigns and describe our real-world deployment.

5.1 Overview

Blackhat SEO pages. With the semantic confusion detection, we explore the features of those blackhat SEO pages and identify three major categories:

```
<head>
    <title>
    Gambling Website_Betting
            Game_Macau Gaming
            Games_Cash Gaming
            Games_[Online
            Entertainment]_Cash games
    </title>
</head>

<body>
    <iframe scrolling="no"
            frameborder="0"
            marginheight="0"
            marginwidth="0" width="
            100%" height="4000"
            allowtransparency="" src="
            http://www.pankou8.com/"
            ></iframe>

    <ul class="nav-list" id="
            J_navlist">
        <li><a href="/">Home</a
            ></li>
        <li><a href="/">Education
            </a></li>
        <li><a href="/">Political<
            /a></li>
        ...
    </ul>
    ...
</body>
```

Fig. 8: HTML of iframe based blackhat SEO page.

```
<script type="text/javascript" style="display:
    none;">
    var strRef=document.referrer;
    var robots=['baidu','google','yahoo','bing
        ','soso','sogou','so','youdao','jike
        ','anquan','360.cn','haosou'];
    var ishave=false;

    for(var t in robots){
        if(strRef.indexOf(robots[t])!=-1){
        ishave=true;
        if(parent.window.opener){
            parent.window.opener.location='
                https://www.tc8806.com/';
    }
</script>
```

Fig. 9: JS Code of cloak-based blackhat SEO page.

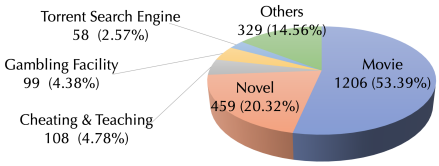

Fig. 10: Category of gray pages.

(1) *Blackhat SEO with links.* To gain better SEO effectiveness, SEOers usually embed numerous external links that point to other blackhat SEO pages or target pages, *e.g.*, the links of gambling websites (see Figure 4).

As discussed in Section 4, we discovered 100,791 blackhat SEO pages from 82,061 different e2LDs and 2,259 gray pages. The statistics of our detection results are listed in Table 2.

(2) *Blackhat SEO with iframe.* Typically in a blackhat SEO page with an iframe, an SEOer will set the width of the iframe to 100% and the height to even more than 2000px, to ensure that the iframe will be displayed in full screen and the content can be displayed to visitors in its entirety. Figure 8 shows an example of mixed semantics, where the text in the title promotes gambling webpages while the text in the body presents a normal page. When users visit this page, they will see content from another website, instead of the original page that is hidden by the full-screen iframe. To recognize this category, we first determine whether a page contains iframe tag. If it does, we check the width/height attributes of the tag. If the width is 100% and the height is higher than 2000 (covering visual areas), it is considered as a blackhat SEO page with iframe.

(3) *Blackhat SEO with cloaking.* Cloaking is a common practice exploited by blackhat SEOers to cheat the search engines [15]. When visited by human users, cloaking pages will bring them to the sites that SEOers want them to visit. However, when crawled by search engines, cloaking pages will feed them with

normal content copied from other popular sites. Figure 9 shows a blackhat SEO page with cloaking, where the script embedded in the page checks if a visitor is a human user or a crawler, and then decide which strategy to adopt. From Table 2, we can see that more than 95% of blackhat SEO pages are link-based SEO, which is apparently due to its simplicity and efficiency. To recognize this category, we fetch all blackhat SEO pages twice. We first retrieve the page content normally. Then, we modify the *referrer* field of the browser to 'baidu.com' to obtain cloaked content. If the page content is different, it is considered as an blackhat SEO page with cloaking.

Gray pages. Different from blackhat SEO pages whose potential target is the crawlers of search engines, gray pages are mainly target human users. Based on the types of their underground content, we classify gray pages into two main categories:

(1) *Gray pages with pornographic content.* These pages do not explicitly advocate unvarnished pornographic content, but offer normal content with a small portion of pornography simultaneously, *e.g.*, a movie website mainly provides normal films but also hosts pornographic videos. In addition, another group of webpages may provide seeds of Magnet URI[8] or BitTorrent[9] often suggest pornographic content to visitors.

(2) *Gray pages with gambling content.* Unlike the gambling blackhat SEO pages that usually direct visitors to gambling sites, these pages often have content about how to play gambling games, how to buy gambling facilities, or even how to cheat while gambling.

Next, we also examine the normal topics that the gray pages belong to. The statistics are shown in Figure 10. We can see that more than 70% of gray pages are with movie or novel, indicating that not only the webpages in these two categories have the broadest visitor base, but they also have the types of content directly related to the pornographic content (*i.e.*, videos and literature).

5.2 SEO Domains

Here we group and characterize the domains hosting blackhat SEO pages with their IP geolocation, TLDs, and domain registration.

IP geolocation. In order to analyze the IP geolocation of blackhat SEO domains, we first used the APIs provided by Farsight's passive DNS database [10] to query the current and historical IP addresses of all blackhat SEO domains. Then, we used the GEOLite2 database from MaxMind [21] to search the AS number and location of each IP address. We observed that most of these blackhat SEO domains are located in United States, followed by Hong Kong. This reflects the practice of blackhat SEOers for avoiding local supervision. We show the country-level IP geolocation in Table 3 and the AS information in Table 4.

[8] `http://magnet-uri.sourceforge.net/`, a URI-scheme in P2P file sharing for enabling resources to be referred to without an available host.

[9] `https://www.bittorrent.com/`, a popular file-sharing P2P tool based on distributed hash table (DHT) method.

Table 3: Top 5 countries/regions for hosting blackhat SEO domains.

No.	Country	# IP	# Domain
1	United States	43,456	52,079
2	Hong Kong	6,729	8,877
3	South Africa	2,311	2,529
4	Netherlands	1,927	2,098
5	China	722	1,136
Total	-	55,145	66,719

Table 4: Top 5 SEO-hosting ASNs (sorted by the number of hosted domains).

NO.	ASN	Organizations	# IP	# Domain
1	AS18978	ENZUINC (US)	6,626	10,642
2	AS35916	MULTA-ASN1 (US)	6,607	7,509
3	AS15003	NOBIS-TECH (US)	6,758	7,060
4	AS40676	Psychz Networks (US)	5,096	5,648
5	AS38197	Sun Network (HongKong) Ltd. (HK)	4,846	5,217
Total	-	-	29,933	36,076

TLD distribution. We find that `.com` and `.cn` are still the most popular domains that are abused for blackhat SEO, accounting for 63.12% and 18.49%, respectively. Table 5 lists the statistics of top-level domains and predefined second-level domains.

Table 5: Top 5 TLDs/predefined-SLDs for mixed semantic domains.

No.	TLD/SLD	# Total	# SEO	# Gray	%
1	.com	53,225	51,779	1,446	63.12%
2	.cn	15,588	15,251	337	18.49%
3	.top	7,025	6,935	90	8.31%
4	.net	2,363	2,292	71	2.80%
5	.com.cn	2,349	2,319	30	2.79%
Total	-	80,550	78,576	1,974	95.51%

Domain registration. We retrieved the domain registration information through WHOIS database. Due to the effect of GDPR [14], the registrant information like emails and telephone numbers is mostly hidden now. Finally, we obtained 56,672 valid WHOIS records showing the registrar information. Table 6 lists the top 5 registrars that operate the most domains abused for blackhat SEO. The most common registrars are Alibaba's cloud platform (`aliyun.com`).

Content Analysis. In order to understand the content that blackhat SEOers prefer to use, we collected the normal semantic of each blackhat SEO page. We found that the content from education websites is the most widely used by blackhat SEO pages. This is likely because the education-based content has good reputation and can easily catch the attention of human users and gain their trust to read it. The content distribution is shown in Figure 11.

Table 6: Top 5 registrars for SEO domains.

Registrar	# Count	%
Alibaba Cloud Computing, Ltd.	11,210	13.66%
Chengdu West Dimension Digital Technology Co., Ltd.	9,937	12.11%
Xin Net Technology Corporation	4,416	5.38%
GoDaddy.com, LLC	4,169	5.08%
Bizcn.COM, Inc.	2,785	3.39%
Total	32,517	39.62%

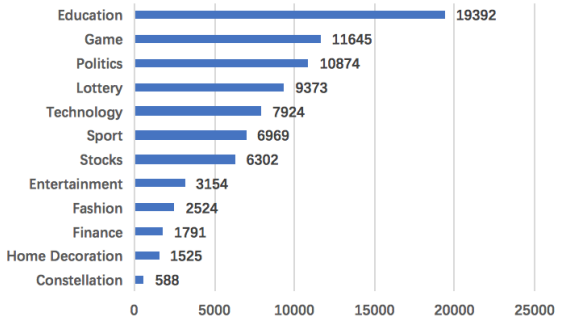

Fig. 11: Normal semantic distribution of blackhat SEO pages.

5.3 SEO Campaigns

Cluster based on external links. We first clustered the domains based on the external links embedded in blackhat SEO pages. *i.e.*, if a group of domains have the blackhat SEO pages that include the same link pointing to a unique target, we grouped them as a cluster. We filtered out the high-reputation domains from Alexa top 10 thousand list. Then, we crawled these target links and used Classifier_1 to detect whether they include underground content. If identified, we recognized it as the promotion target of a blackhat SEO campaign. Through this method, we successfully identified 157 SEO campaigns, covering 21,374 different domains that participate in these campaigns. In particular, the largest SEO campaign we detected includes 1,758 SEO domains, all of which are with iframe pages and are targeted on the domain of `7497999.com`.

Cluster based on statistics ID. To make a profit through the campaigns, blackhat SEOers need to embed their statistics ID in the pages and monitor the visitor count during the campaign. Typically, they insert a snippet of JavaScript into HTML source code to trigger the browser to visit a specific URL and record the visitor count. We examined the HTML source code of detected pages to extract statistics IDs based on the URL patterns presented by various service providers. For example, in an embedded JavaScript code

```
<script language="javascript" src="http://
count1.51yes.com/click.aspx?id=12345678"></script>
```

the URL points to the visitor analytic provider, and the parameter "`id=`**`12345678`**" presents the SEOer's ID. As shown in Table 7, the four most popular service providers for visitor analytics in China are: baidu.com, 51la.com, 51yes.com,

Table 7: Visitor analytic service providers.

No.	Provider	# SEOer	# Domain	% Domain
1	hm.baidu.com	2,335	31,076	37.87%
2	51la.com	1,745	20,977	25.56%
3	51yes.com	279	2,497	3.04%
4	cnzz.com	432	781	0.95%
5	others	-	26,730	32.58%
Total	-	**4,791**	**82,061**	**100.00%**

and cnzz.com. Among those, Baidu is the most popular provider adopted by blackhat SEOers.

Comparison. We compared the clustering results between the external link-based method and the ID-based method. The former captures a group of domains that promote the same target websites while the latter identifies the domains dominated by the same blackhat SEOer. In order to improve the effectiveness of promotion, a website owner may purchase an SEO service from multiple blackhat SEOers to form a larger-scale campaign. Table 8 lists the top 5 campaigns with the most participating domains and the number of statistics ID it used. Since most of the objects we detect are blackhat SEO pages targeting Chinese users, they mainly use analytic providers in China. We also find a small portion of blackhat SEO pages using Google Analytics.

Table 8: Top 5 SEO Campaigns for promotion.

	Target	# Domain	# IP	# Country	# ID
Camp. 1	7497999.com	1,758	35	2	2
Camp. 2	hg18217.com	1,440	1,026	2	22
Camp. 3	yifageen68.cn	1,294	19	2	4
Camp. 4	jiuwuzhizun8.com	823	14	3	1
Camp. 5	187bet.com	661	335	4	3

5.4 Real-World Deployment

During our study, we deployed the SCDS at the gateway of our campus network (with more than 30K users). We collected the domains from DNS query records through the campus network, crawled the domains and checked whether their webpages contain the mingled semantic content (*i.e.*, semantic confusion). We deployed our system from April 5, 2020 to January 10, 2021, obtained 11,547,471 unique domain names, and detected 23,093 domains with semantic confusion (including 21,097 blackhat SEO and 1,996 gray pages).

6 Practical Issues

Impact on Search Engines. As illustrated in Section 3, for recursive detection, we extended blackhat SEO pages with semantic confusion by utilizing Google. In

Table 9: Number of SEO URLs appearing in search results of SEO domains.

Domain	Category	Google	Baidu
tcppower.com	Gambling	12	20
tongyunhr.com	Gambling	20	5
skfjr.com	Gambling	0	20
flyco360.com	Gambling	20	20
jrzgxw.com.cn	Porn	20	0

this step, we first selected 1,000 blackhat SEO domains (including 800 gambling and 200 pornographic domains), and searched them using Google. We then obtained 5,373 valid search results and confirmed that 5,191 blackhat SEO URLs are indexed by Google. From this, we can see that these blackhat SEO pages have an actual effect on search engine results. To compare the impact among different search engines, we randomly selected 5 SEO domains and searched them in Baidu (the most widely used search engine in China), collecting the top 20 results. We counted the number of blackhat SEO domains in the search results, as shown in Table 9. It can be seen that these sites pollute search engine results.

Popular Website Templates. In order to enhance the cheating effect, blackhat SEO sites crawl HTML pages from authoritative sites and mix blackhat SEO content with them. To cheat search engines, blackhat SEO pages may keep part of the original title information in the new pages, *e.g.*, making a new title as a combination of Blackhat SEO keywords and original title information. We extracted all the "original title information" parts and clustered them into groups. In particular, we found that there are 771 blackhat SEO pages built based on the official webpage of "Chinese Academic of Science" and 164 pages built based on the official webpage of "Tencent," which operates the most popular Instant Messaging tools in China. Moreover, we noticed that the blackhat SEOers widely use open-source webpage templates to construct their pages. The bottom of these pages usually contain words such as "`Powered by [template provider]`", with a link to the template supply website. During the link extraction process, we found that `dedecms.com`, `dede58.com`, and `adminbuy.cn` are the most commonly used open-source template supply websites for constructing blackhat SEO webpages, enabling 1,703, 1,562, and 1,387 domains, respectively.

7 Discussion

Responsible Disclosure and Feedback. We reported the detection results to QiAnXin [10] and our network administration team every day. Our work can help in the following ways: (1) help search engines to identify webpages with blackhat SEO content, (2) help security practitioners to be aware of the trend of all kinds of underground economies with emerging content, (3) help security enterprises to classify webpages more accurately, especially those in "gray" host

[10] QiAnXin is a leading Cybersecurity company in China (`https://en.qianxin.com`).

hybrid content. In the future, we will collaborate with security practitioners to deploy SCDS in real-world network environments to further investigate blackhat SEO webpages at a large scale.

Language Dependency. Our training procedure relies on the training data, and the detection result is dependent on the quality of the collected training data. Due to dataset limits, our detection mainly focused on blackhat SEO pages that promote illegal gambling/pornography content on Chinese webpages. However, we believe the identified phenomenon also exists in such SEO pages in other languages. Moreover, we demonstrated the effectiveness of our detection. If datasets belong to other languages could be obtained, our method will effectively detect the corresponding blackhat SEO pages.

Gary Page. In this study, we explicitly identified the gray pages that have not been well examined before. Limited by dataset, we only focused on gray pages related to gambling/pornography. If there are more types of training data, more gray pages related to different underground businesses could be detected.

Evasion. If blackhat SEOers had been aware of our detection mechanism, they may explore the evasion by reducing the content related to the underground economy in HTML pages or adding various normal topics. However, such strategies would also significantly reduce the effectiveness of promotion, which is also a goal of our detection system.

Ethical Considerations. Since we deployed the SCDS on the gateway of our campus network, there could a concern about personal information and privacy leakage. Here we only extracted domain names and excluded any other information. Therefore, there is no risk of revealing a user's personal and private information to a third party in our study.

8 Related Work

Blackhat SEO. To understand the blackhat SEO activities, significant efforts have been spent to explore its ecosystem and practice. Wang *et al.* [27] infiltrated an SEO botnet and showed that it is quite effective in poisoning trending search terms, given its small size. Liao *et al.* [18] characterized the long-tail SEO that promotes longer and more specific keyword phrases targeting niche markets. To thwart the threat of blackhat SEO, search engines have developed many defense mechanisms. John *et al.* [16] leveraged URL signatures to identify SEO pages from a dataset of URLs provided by search engines. Lu *et al.* [20] detected search poisoning by inspecting the redirection chains unfolded when visiting a search result. The countermeasures of search engines were effective to some extent, but challenges still remain [17,26,16]. Complementary to the existing studies, our work focuses on an emerging trend of blackhat SEO, which leverages semantic confusion to disguise the content of underground economy with the text of normal topics.

Underground Economy. Prior studies have conducted in-depth analysis of different types of the underground economy, including scam [25], email spam [7], promotion infection [19], illegal commodity transaction [22] and social media

spam [23]. In particular, to evade detection, the practitioners of the underground economy usually use jargon to disguise their activities. Yang *et al.* [28] and Yuan *et al.* [29] proposed different methodologies to discover such jargon or black keywords.

9 Conclusion

In this paper, we conducted the first systematic investigation of a recent trend in blackhat SEO, *semantic confusion*, by which blackhat SEOers can promote underground business by disguising their content with legitimate topics. To address this emerging threat, we developed an effective detection system that can identify such semantic-confusion-based blackhat SEO webpages with high precision. Further, we performed a comprehensive measurement study to characterize the ecosystem of blackhat SEO. Finally, we deployed our system in a gateway for real-world evaluation, showing that semantic confusion has been prevalent in blackhat SEO. Our study will help the security community to pay more serious attention to blackhat SEO detection for fighting cybercrime. In the future, we will collaborate with security practitioners to enlarge the detection scale of `SCDS`.

References

1. Beautiful soup documentation - beautiful soup 4.9.0 documentation. `https://www.crummy.com/software/BeautifulSoup/bs4/doc/` (2020)
2. Github - fxsjy/jieba. `https://github.com/fxsjy/jieba` (2020)
3. Keras: the Python deep learning API. `https://keras.io/` (2020)
4. SeleniumHQ Browser Automation. `https://www.selenium.dev/` (2020)
5. Textcnn - pytorch and keras — kaggle. `https://www.kaggle.com/mlwhiz/textcnn-pytorch-and-keras` (2020)
6. Chung, Y.j., Toyoda, M., Kitsuregawa, M.: A Study of Link Farm Distribution and Evolution Using a Time Series of Web Snapshots. In: International Workshop on Adversarial Information Retrieval on the Web (2009)
7. Cormack, G.V.: Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval **1**(4), 335–455 (2007)
8. Du, K., Yang, H., Li, Z., Duan, H., Zhang, K.: The Ever-Changing Labyrinth: A Large-Scale Analysis of Wildcard DNS Powered Blackhat SEO. In: USENIX Security (2016)
9. Enge, E., Spencer, S., Fishkin, R., Stricchiola, J.: The Art of SEO. O'Reilly Media, Inc. (2012)
10. Farsight: `https://www.farsightsecurity.com/` (2020)
11. Fishkin, R.: Indexation for SEO: Real Numbers in 5 Easy Steps. `https://moz.com/blog/indexation-for-seo-real-numbers-in-5-easy-steps` (2010)
12. Google: Search Engine Optimization Starter Guide. `http://static.googleusercontent.com/media/www.google.com/en//webmasters/docs/search-engine-optimization-starter-guide.pdf` (2008)
13. Google: Search Engine Optimization (SEO) Starter Guide. `https://support.google.com/webmasters/answer/7451184?hl=en` (2020)

14. ICANN: Data Protection/Privacy Issues. https://www.icann.org/dataprotectionprivacy (2018)
15. Invernizzi, L., Thomas, K., Kapravelos, A., Comanescu, O., Picod, J.M., Bursztein, E.: Cloak of Visibility: Detecting When Machines Browse a Different Web. In: IEEE S&P (2016)
16. John, J.P., Yu, F., Xie, Y., Krishnamurthy, A., Abadi, M.: deSEO: Combating Search-Result Poisoning. In: USENIX Security (2011)
17. Leontiadis, N., Moore, T., Christin, N.: A Nearly Four-Year Longitudinal Study of Search-Engine Poisoning. In: ACM CCS (2014)
18. Liao, X., Liu, C., McCoy, D., Shi, E., Hao, S., Beyah, R.A.: Characterizing Long-tail SEO Spam on Cloud Web Hosting Services. In: WWW (2016)
19. Liao, X., Yuan, K., Wang, X., Pei, Z., Yang, H., Chen, J., Duan, H., Du, K., Alowaisheq, E., Alrwais, S., Xing, L., Beyah, R.: Seeking Nonsense, Looking for Trouble: Efficient Promotional-Infection Detection through Semantic Inconsistency Search. In: IEEE S&P (2016)
20. Lu, L., Perdisci, R., Lee, W.: Surf: detecting and measuring search poisoning. In: ACM CCS (2011)
21. MaxMind: https://www.maxmind.com/en/home (2020)
22. Motoyama, M., McCoy, D., Levchenko, K., Savage, S., Voelker, G.M.: An analysis of underground forums. In: ACM IMC (2011)
23. Nilizadeh, S., Labrèche, F., Sedighian, A., Zand, A., Fernandez, J., Kruegel, C., Stringhini, G., Vigna, G.: Poised: Spotting twitter spam off the beaten paths. In: ACM CCS (2017)
24. SEOmoz: Google Algorithm Change History. https://moz.com/google-algorithm-change (2016)
25. Tu, H., Doupé, A., Zhao, Z., Ahn, G.J.: Users really do answer telephone scams. In: USENIX Security (2019)
26. Wang, D.Y., Der, M., Karami, M., Saul, L., McCoy, D., Savage, S., Voelker, G.M.: Search+Seizure: The Effectiveness of Interventions on SEO Campaigns. In: ACM IMC (2014)
27. Wang, D.Y., Savage, S., Voelker, G.M.: Juice: A Longitudinal Study of an SEO Botnet. In: NDSS (2013)
28. Yang, H., Ma, X., Du, K., Li, Z., Duan, H., Su, X., Liu, G., Geng, Z., Wu, J.: How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In: IEEE S&P (2017)
29. Yuan, K., Lu, H., Liao, X., Wang, X.: Reading Thieves' Cant: Automatically Identifying and Understanding Dark Jargons from Cybercrime Marketplaces. In: USENIX Security (2018)
30. Zhang, Q., Wang, D.Y., Voelker, G.M.: DSpin: Detecting Automatically Spun Content on the Web. In: NDSS (2014)

Appendix

A Practices of Semantic Confusion

We further analyzed the method of embedding mingled semantics in SEO pages and identified three main categories: (1) Modify only the <title> tag, and keep all other parts the same. This is because search engines often pay more attention to the <title> tag and the text in the title has a higher probability

Fig. 12: Gambling software development in GitHub.

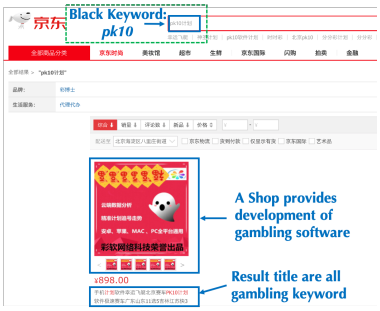

Fig. 13: Gambling shop in JingDong (`jd.com`).

to be indexed. Also, blackhat SEOers do not want to be detected due to the modification of the pages or the mixed illegal content. (2) Embed the same promotion keywords into different paragraphs repeatedly. Appropriate repeats are helpful for search engines to extract keywords and give them a higher rank. These two methods mainly target search engines. (3) Replace the total paragraph with promotion content. This category mainly targets visitors and aims to attract them immediately upon arrival at the webpage. The replaced content is short-lived because it is easily noticed by search engines and webmasters.

B Keyword Promotion in Other Platforms

GitHub. In our study, we found that some blackhat SEOers are promoting GitHub repositories of gambling software development services. Specifically, when we searched "`github.com+[gambling keywords]`" in Google, the results show many GitHub repositories introducing gambling software, and the descriptions of these repositories include the developer's contact information (*e.g.*, phone numbers and IM IDs). Another practice is to place a large number of gambling keywords in a repository's introduction through which search engines can index them. For example, Figure 12 shows the search results of a gambling keyword "真人视讯" (a dark jargon that means "Live Dealer Casino Games" in underground gambling business) with "`github.com`" in Google. The top results are mostly GitHub repositories that promote gambling development services.

E-Commerce. We also found that the keywords promoted by blackhat SEO pages were not only used in search engines, but they also appeared on E-Commerce websites. For example, when we searched the most frequent keyword, `pk10` on `jd.com` (a well-known E-Commerce website in China), there were shops that provide illegal gambling software development services, as shown in Figure 13. Therefore, we recommend that E-Commerce websites should also pay attention to the identification and purification of such activities related to the underground business in search results.