Into the Dark: Unveiling Internal Site Search Abused for Black Hat SEO

Yunyi Zhang[†‡], Mingxuan Liu[§], Baojun Liu[‡§], Yiming Zhang[‡], Haixin Duan[‡§], Min Zhang[†], Hui Jiang[‡¶✉], Yanzhe Li[¶], Fan Shi[†✉]

[†]National University of Defense Technology, [‡]Tsinghua University
[§]Zhongguancun Laboratory, [¶]Baidu Inc

Abstract

Internal site **S**earch **A**buse **P**romotion (*ISAP*) is a prevalent Black Hat Search Engine Optimization (SEO) technique, which exploits the reputation of abused internal search websites with minimal effort. However, ISAP is underappreciated and not systematically understood by the security community. To shed light on ISAP risks, we established a collaboration with Baidu, a leading search engine in China. The key challenge of efficiently detecting ISAP risks stems from the sheer volume of daily search traffic, which involves billions of URLs. To address these efficiency bottlenecks, we introduced a first-of-its-kind lightweight detector utilizing a funnel-like approach, tailored to the unique characteristics of ISAP. This approach allows us to single out 3,222,864 ISAP URLs from 10,209 abused websites from Baidu's traffic data. We found that the businesses most likely to fall prey to this practice are porn and gambling, with two emerging areas: self-promotion for SEO and promotion for anonymous servers. By analyzing Baidu's search logs, we discovered that these malicious websites had reached millions of users in just 4 days. We further evaluated this threat on Google and Bing, thereby confirming the widespread presence of ISAP across various search engines. Moreover, we responsibly disclosed the issue to affected search engines and websites, and actively helped them fix it. In summary, our findings highlight the widespread impact and prevalence of ISAP, emphasizing the urgent need for the security community to prioritize and address such risks.

1 Introduction

Internal Site Search (ISS) is an essential service for efficient resource location. Figure 1 illustrates the ISS feature on Washington State University's official website. While ISS shares similarities with search engines, its primary purpose is to access the website's internal resources. Due to its convenience and crucial functionality for resource retrieval, ISS has been adopted by 47.23% of Alexa Top 1M websites [25].

✉ Corresponding author.

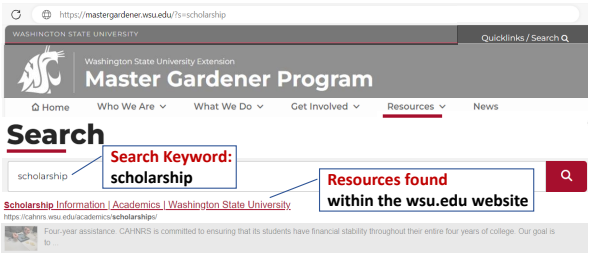
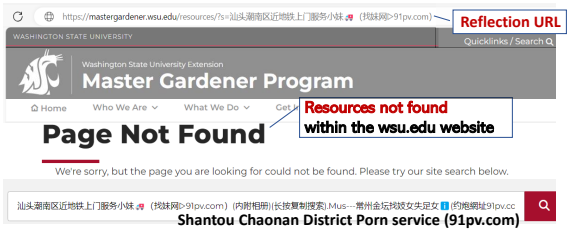


Figure 1: Example of internal site search on `wsu.edu`.

However, ISS has unfortunately become an exploitation target for Black Hat Search Engine Optimization (SEO) activities [9, 21, 57]. Figure 2 shows an abuse example of the ISS function in Washington State University (WSU). Such abuse stems from ISS's improper handling of non-existent resources. Specifically, ISS generates new URLs and webpages containing search keywords even with empty search results, as shown in Figure 2a. Black hat SEOers embed promotion targets (e.g., illegal websites, phone numbers) into search keywords. Thus, they obtain a large number of searching URLs and webpages, which are defined as *reflection URLs and webpages*. These reflection URLs, once indexed in search engines, become a vehicle for illegal promotional content. Leveraging the reputation of the websites, such as the well-known university WSU, the promotional content has more potential to be displayed to users through search engines (Figure 2b). In this paper, we term this black hat SEO technique as **I**nternal site **S**earch **A**buse **P**romotion (*ISAP*).

The impact of ISAP is significant. First, instead of compromising other websites [61] or deploying complex strategies or infrastructure [11, 14, 32, 53], it enables cost-effective black hat SEO by exploiting high-credibility websites, catalyzing promotion abuse. Second, users risk being misled whenever they encounter connections between reputable sites (e.g., educational and government sites) and malicious promotions in search results, which even appear on the first page. Worse still, ISAP operates covertly, successfully evading detection by the abused websites.

Research gap. Despite being mentioned in an existing



(a) ISAP example on `wsu.edu`.

(b) Search result of ISAP on `wsu.edu` in Google.

Figure 2: Internal site search abuse promotion of Washington State University.

study [14], ISAP is not systematically understood and effectively detected. Its low cost and fast promotion enable promotional content on the top page of search results within two hours. Unfortunately, existing black hat SEO detection approaches are unsuitable and ineffective against real-world ISAP. Existing methods are scenario-specific, such as wildcard-DNS-based detection [14] and cloaking-based detection [54, 56], or computationally intensive, like semantic-based detection [15, 31, 40, 46] and website-tampering-based detection [61]. This research gap hinders ISAP detection and mitigation, allowing it to remain a severe threat.

Our work. To bridge the research gap in understanding and detecting ISAP, this paper endeavors to devise an efficient and lightweight detection system capable of identifying ISAP in real-world search engines. Moreover, we aim to comprehensively reveal the ISAP ecosystem and assess its impact, utilizing active and passive detection methods.

In this study, we confront the overlooked security threat of ISAP by collaborating with Baidu [5], China's largest search engine vendor, to carry out the first-ever detection and measurement of ISAP in a practical search dataset. We initially collected the ground-truth dataset of ISAP and conducted empirical analysis to unearth key characteristics of ISAP. Drawing insights from these characteristics, we devised a lightweight ISAP detection approach with minimal data requirements, crafted explicitly for search engines, with inspiration from the funnel-like idea. This method is designed to circumvent the efficiency bottlenecks, thus facilitating efficient detection of high-volume traffic (at the billion level). Our detection system has been deployed in Baidu, and daily checks are performed. Furthermore, for defensive purposes, we introduced a method to assist website administrators in proactively assessing website susceptibility to ISAP threats.

Our findings. Our method empowers our collaborators to process billions of daily traffic in a mere span of 2 hours. From May 1 to September 19, 2023, we unveiled ISAP in over one billion URLs, uncovering **3,222,864** ISAP URLs from **10,209** abused websites, which were dispersed among **4,458** distribution websites. Our findings testify to the extensive use of ISAP by SEO practitioners, with the adult industry accounting for a staggering 77.44%. To add to this, we identified two emerging businesses: *self-promotion within the SEO industry* and *promotion of anonymous servers*.

In addition, we discovered that black hat SEO practitioners strategically exploit various types of websites, including popular ones (35.3%) and non-popular ones (64.67%). They proliferate a large number of ISAP URLs on individual popular domains. Conversely, for non-popular domains, black hat SEO practitioners aggregate multiple domains' promotional abilities to achieve their desired promotion scale, typically characterized by a relatively low number of URLs for each domain. Different distribution websites employ varied promotional strategies, including sustained, gradual, or explosive rapid distribution. Using search logs provided by our collaborators, we assessed the Page Views (PV) of identified ISAP URLs. The results reveal a significant impact, with ISAPs reaching at least **6 million** users through search engines within 4 days.

For defensive evaluation, we discovered that 11.76% of Tranco's top 10K domains (e.g., `bbc.com`), 24.59% of EDU domains (e.g., `berkeley.edu`), and 14.10% of GOV domains (e.g., `nasa.gov`) can be exploited, indicating the severity of ISAP. Furthermore, we confirmed the prevalence of ISAP in other search engines utilizing detected ISAP URLs (Section 6.2). Last but not least, we have responsibly disclosed our findings to Baidu, Google, and Bing. Currently, Baidu has deployed our detection system, and Bing indicated they had implemented a fix for ISAP threat based on our disclosure. Additionally, we have disclosed the threat and are helping website administrators fix it. To date, we have received acknowledgment from 37 organizations.

Contributions. This paper makes the following contributions:

A novel, efficient detection method. We designed and implemented a lightweight ISAP detection method. This method has greatly enabled our partner search engine's efficiency in processing billions of daily data within two hours, thereby significantly improving ISAP detection and mitigation.

A comprehensive understanding of the ISAP ecosystem. We conducted a systematic study and analysis of ISAP for the first time. Not only did we uncover the promotion strategies of ISAP from a search engine perspective, but we have also demonstrated the risk existing in a large number of well-known websites in the wild through proactive evaluation.

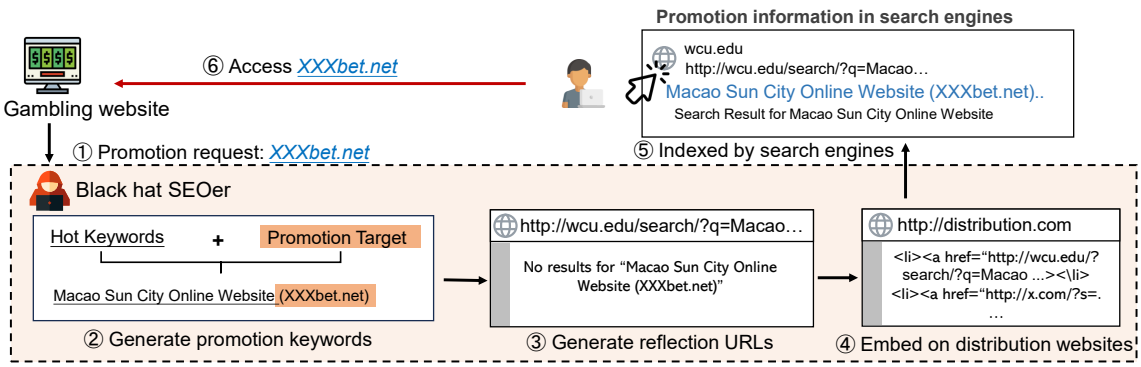


Figure 3: Promotion process of ISAP.

2 Background and Threat Model of ISAP

In this section, we introduce ISS and elucidate the abuse of ISAP. Subsequently, we delineate this adversarial promotion technique for black hat SEO and illustrate its threat model.

2.1 Internal Site Search

Internal site search (ISS) is a website function that enables users to quickly find specific content within the site. Similar to search engines, ISS allows users to enter keywords and send queries in a search box, then retrieve relevant results from web pages, documents, or products. For example, users can efficiently find information about "scholarships" on a university website using ISS, as shown in Figure 1. This feature improves user experience by providing a quick and efficient way to find specific information within a website. It is particularly valuable for websites with extensive content or complex navigation. ISS plays a crucial role in information retrieval and enhances overall website usability. Its adoption is widespread; 47.23% of Alexa Top 1M websites utilize internal site search functionality [25].

2.2 Internal Site Search Abuse Promotion

While internal site search improves user convenience, flawed implementation can invite exploitation by miscreants, particularly for promotional abuse. This abuse typically occurs when internal site searches for non-existent resources are improperly implemented, making them vulnerable to Internal site Search Abuse Promotion (ISAP). Despite similarities with search engines, ISS's implementation standards are less rigorous, especially regarding non-existent resources. Specifically, some websites send search queries through `GET` method, which results in incorporating the non-existent search keywords into the returned resources, including 1) the returned URL, thereby generating a new URL, denoted as the "reflection URL" or "ISAP URL"; 2) the returned webpage's content (e.g., title, search bar, body), denoted as the "reflection webpage". Given the strong correlation between reflection URLs/webpages and search keywords, they serve as affordable promotional platforms. Furthermore, since these domains may belong to reputable websites, black hat SEO practitioners (black hat SEOers) can exploit the trust these sites have established with search engines for effective promotion.

Figure 3 depicts the promotion process of ISAP. As the initial step (①), miscreants often leverage promotional services provided by black hat SEO practitioners to advertise their illegal or malicious businesses (e.g., the illicit gambling website `XXXbet.net`), referred to as *promotion targets* in this work. When dealing with promotion requests, black hat SEOers first generate promotion keywords for specific targets (②). In the underground industry, it is not typical for users to search for promotional targets directly. To increase search keyword exposure, black hat SEOers combine a hot keyword with the target, taking advantage of the high popularity of the hot keyword. For instance, "Sun City", a renowned gambling venue in Macao, could be used as the hot keyword for a specific gambling website (`XXXbet.net`). In step ③, the website with ISAP risks is abused to generate a multitude of candidate reflection URLs, where the search keywords as URL query parameters (i.e., `?search=...Macao Sun City Website(XXXbet.net)...`).

Next, it is necessary for black hat SEOers to enable search engines to index these reflection URLs. The most direct and fastest way is through active indexing platforms like Google's interception tool [20]. However, these platforms have verification processes that require the submitter to own the site and verify the page's content [3, 6, 7, 17]. Black hat SEOers, however, do not own the domain name of reflection URLs, and their content is associated with illegal and harmful activities. Thus, submitting directly through this method is challenging. Hence, black hat SEOers distribute reflection URLs through distribution websites (④). The distribution website refers to a site indexed by search engines where adversaries can embed external links, such as spider pools [14] or prominent blog sites [39, 48]. By embedding reflection URLs on these sites, search engine crawlers are guided to crawl and index them. Once indexed, these reflection URLs leverage the reputation of well-known websites, resulting in higher search engine

ranking (⑤). Consequently, when users search for content related to the hot keywords, these promotional reflection URLs are returned to users and may appear on the top page of search results. Finally (⑥), users gain information about the promotion target and are exposed to malicious resources (like illicit gambling sites).

2.3 Threat Model of ISAP

Adversary's Goal. We assume that the attacker's primary goal is to effectively inject promotional content into search engine results. By displaying injected content in user search results, the visibility of these promotional content is significantly increased to reach a broader user base.

Adversary's Capacity. To effectively implement ISAP, black hat SEOers require two key conditions. First, they need numerous vulnerable websites with ISS function as promotional sources. Active scanning of prominent websites enables them to identify susceptible targets. Our active analysis in Section 6.2 enables us to identify ISAP risks before the actual abuse. The results reveal that a significant number of websites within the Tranco Top 10k [41], as well as education- and government-related websites, are exploitable. Second, reflection URLs can be quickly indexed by search engines. Constructing distribution websites for indexing is not a major challenge, as discussed in prior work [11, 14, 39, 48], it can be done through various methods such as spider pools, blog sites, link farms, or using self-built distribution websites.

Assumption of Victim Website. ISAP exploits the ISS of victim websites to distribute promotional content. Victim websites should fulfill two requirements: 1) have an internal search function, and 2) when no search results are found, display a "no resources found" notification webpage that includes the search keywords in the URL and webpage.

3 Dataset and Empirical Study

In this section, we present our dataset collection, including the URL snapshot dataset, ISAP ground-truth dataset, and website list. Then, we introduce the empirical study results for the ground-truth dataset.

3.1 Datasets

We collaborate with the security department of Baidu, collecting URL snapshots preserved across diverse functional units. Furthermore, we employ a database of ISAP search patterns, meticulously collected by Baidu's security team, to construct a robust ground-truth data set for our empirical study, which provides insights into the ISAP detector's design in Section 4. Additionally, we used the Tranco Top 10K and our collected education and government website lists for active evaluation in Section 6.

URL Snapshot Dataset is derived from integrating various Baidu business data maintained by the Baidu Security Department, including search engine, advertising service, and terminal security monitoring service. This dataset comprises focused and prioritized data from key business areas, as identified by the Security Department. This diverse coverage allows us to access real data from different facets of the search engine's operations. It is worth noting that snapshot data serves as the initial gateway for search engines to index websites, including those submitted by users or discovered by search engine crawlers during network resource explorations. For the purpose of efficient data storage, our collaborator only retains the first layer of URL data and its corresponding HTML content without performing deep crawling on the embedded links (i.e., external links [14]). Consequently, in the context of ISAP detection and analysis, snapshot data provides information on distribution websites, where the external links are more likely to be ISAP URLs. On the contrary, for benign cases, snapshot data consists of normal advertising websites that may contain legitimate promotional external links, such as product advertisements.

This snapshot dataset is stored daily, and historical data for the past 4 months is retained. The storage format for the snapshots consists of <url, html_source_code> pairs. Therefore, we got the URLs for each day from this dataset along with the corresponding crawled page results, i.e., their HyperText Markup Language (HTML) data. In the end, we collected snapshot data from May 1 to September 19, 2023. However, data from 17 days were lost due to storage issues. In total, we obtained over 7 billion URLs and their HTML within 125 days, with an average of 60 million URLs daily.

Ground-truth Dataset. To the best of our knowledge, there is currently no publicly available dataset specifically dedicated to ISAP URLs. Hence, in our study, we bootstrap our research by creating a ground-truth dataset. Over a long-term observation period, our collaborators have diligently maintained an abuse database, including identified ISAP URLs and search patterns (e.g., /search/q) of abused websites, encompassing thousands of popular websites, as shown in Appendix A. Most of these abused sites have been reported by users' complaints and subsequently verified through a manual process. While the coverage of this database may be limited, the data it contains have undergone meticulous manual verification, ensuring a high level of accuracy for our ground-truth dataset. Under the premise of ensuring no ethical issues, we will open-source part of our labeled data [1] to aid the security community in understanding the ISAP threat.

1) URL ground-truth dataset. From August 13 to 31, 2023, we selected 7 sampling days at three-day intervals to collect reflection URLs from the aforementioned abuse database. On each sampling day, we randomly selected 3,000 reflection URLs. Two researchers manually verified the promotional

[1] http://isap-check.com/#dataset

intent of search keywords in reflection URLs to confirm ISAP. For benign ground-truth, we randomly sampled 3,000 search URLs in the same 7 sampling days from the snapshot dataset. Two security researchers conducted additional manual verification to ensure the accuracy and reliability of the normal ground-truth dataset, by removing duplicates and pre-identified ISAP URLs. As a result, a cumulative collection of 20,999 reflection URLs and 18,349 normal URLs were gathered as our ground-truth dataset. In addition, to facilitate a comprehensive analysis of the collected URL data, we also gathered the distribution websites that embedded these abused and benign URLs from the snapshot data.

2) Promotion keyword ground-truth dataset. We manually labeled promotion keywords to identify different promotion business categories. In consultation with our collaborating security team, we identified 5 distinct categories, including gambling, adult content (porn), SEO, anonymous servers, and others. To simplify analysis and discussion, we consolidated multiple low-frequency categories into a single category labeled "others", like software development, fake invoice, and loan service. These industries within the "others" category are small in scale; thus, combining them exerts minimal impact on subsequent analysis. Notably, we enforced a minimum requirement of 2,000 keywords per category to ensure an adequate dataset size for training. We randomly chose promotional keywords in reflection URLs from the same 7 sampling days and manually categorized them by semantic meaning. As a result, we obtained a total of 26,643 labeled promotion keywords, distributed as follows: 7,039 entries for gambling, 7,028 for adult content, 7,339 for SEO, 2,128 for anonymous servers, and 3,109 for other categories.

Website List. The URLs collected from Baidu could represent the actual abuse status of ISAP (already exploited). The abuse of ISAP on internal search websites is highly covert and invisible to website administrators. Therefore, to assess websites for potential ISAP risks in the real world, we also collected two distinct categories of website lists, including popular domains from Tranco and high-profile domains (including education and government apex domains). These websites have a high reputation in search engines and are more likely to be targeted by adversaries.

• Popular domain list. Tranco [41] is an aggregated top domain list that collects multiple existing top lists as its input, including Alexa [4], Majestic [1], Quantcast [43], and Cisco Umbrella [12]. Thus, it possesses commendable credibility and resistance to manipulation. We chose the top 10K domains of Tranco to evaluate popular websites.

• Education and government apex domain. In order to create a more comprehensive domain list, we collected education domains (referred to as EDU) and government domains (referred to as GOV) from 2 data sources, including 1) popular domain lists (Tranco [41] and SecRank [58]); 2) passive DNS data from 114 DNS[1]. We specifically extracted domains with the extensions .gov, .gov.[ccTLD], .edu, and .edu.[ccTLD] to represent government and education domain names, respectively. After eliminating duplicates and inaccessible websites, we collected 22,641 education and 24,947 government domains.

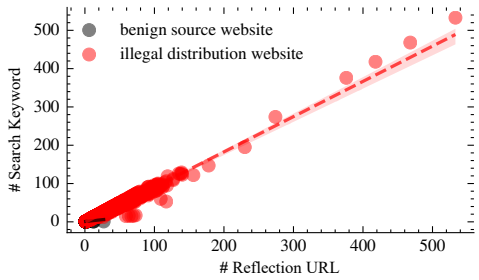


Figure 4: Distribution of the number of search keywords and URLs embedded on each distribution website.

3.2 Empirical Study of ISAP

Given the limited understanding of ISAP within the security community, we initially undertook an extensive analysis of real-world ISAP ground-truth data encountered by search engines. This analysis enables us to identify the following 3 key observations of ISAP, which played a crucial role in informing the ISAP detector's design in Section 4.

Key Observation 1: Illegal distribution websites for ISAP exhibit a substantial number of external reflection URLs, which are embedded with limited distinct promotion keywords. To boost promotional effectiveness, black hat SEO practitioners leverage their distribution websites to spread multiple reflection URLs for the same promotion target, each incorporating different promotion keywords. This approach maximizes promotion efficiency, while a single distribution website may also deploy multiple sets of reflection URLs to promote various targets. Figure 4 presents the statistical results from the snapshot dataset, revealing that illegal distribution websites exhibit significantly higher reflection URLs than benign source websites. Analytical results indicate that only 2.84% of illegal distribution websites contain fewer than 4 external links, while 99.95% of benign websites have no more than 3. We manually inspected 200 randomly-sampled illegal distribution webpages with fewer than four external links. The result shows that these webpages are caused by the truncation of HTML pages with storage errors in Baidu's database. The results of our manual confirmation and URL Screener evaluation (Section 4.4) show that this data error on ISAP detection is manageable. Errors occur equally on illegal and legitimate sites, resulting in fewer observed external links. We implemented strict threshold selection to mitigate this issue, as described in Section 4.4. Moreover, although truncation results in missing a small percentage of corrupt illegal distribution URLs (only 2.84%), many untruncated distribu-

[1] https://www.114dns.com/

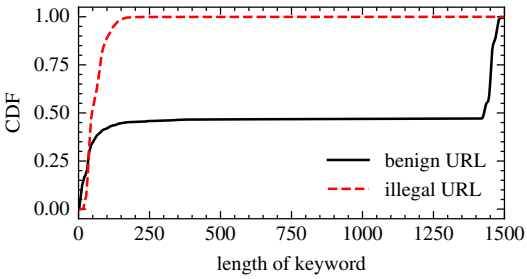


Figure 5: Distribution of search keywords' length.

tion URLs still operate under the domains of these corrupted URLs. Thus, given our dataset's size, these truncated webpages do not compromise the diversity of our detection results. Besides, most data points cluster along a line of slope 1, indicating that each reflection URL contains unique promotion keywords. Conversely, benign source websites show minimal variation in keywords despite the presence of some reflection URLs, since each URL contains the same keywords.

Key Observation 2: To ensure the clarity of promotional information, the length of promotion keywords within reflection URLs typically falls within a specific range. When generating reflection URLs, black hat SEO practitioners combine hot keywords from their keyword databases and promotion targets to construct promotion keywords, i.e., the entire search keyword. This process results in promotion keywords with consistent patterns, including similar lengths. However, for effective user comprehension, promotion keywords must also convey sufficient semantic information. As a result, unlike randomly varied lengths seen in regular search keywords, the length of promotion keywords falls within a specific range, which is neither too short nor too long. Figure 5 demonstrates that 90% of the search keywords in illegal URLs are in the range of 14 to 108. The average length of search keywords within illicit URLs comprises 59.27 characters. This may be attributed to attackers' requirements to embed sufficient information in the keywords, including not only promotion targets but also attractive keywords.

Key Observation 3: While the promotion keywords within illegal reflection URLs often change, the promotion target remains constant. Black hat SEO practitioners of ISAP employ search keywords embedded in reflection URLs for promotional purposes. These keywords are carefully crafted to include explicit promotion targets, such as websites or contact information, which are consistently reiterated. Consequently, when generating reflection URLs, practitioners ensure that the promotion target remains unchanged while hot keywords are changed.

4 ISAP Detector

In this section, we illustrate the ISAP detector. We begin by presenting the overview and each part of our detector, followed by a detailed description of its implementation. Finally, we evaluate its effectiveness.

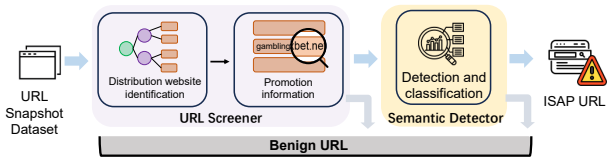


Figure 6: Workflow of ISAP Detector.

4.1 Overview

Given the extensive volume of URLs, it remains a challenge to effectively detect ISAP for search engines. While the semantic features of ISAP are clearly discernible, directly applying natural language processing (NLP) models poses challenges due to the immense volume of daily URL traffic. For instance, the efficiency of DMOS [61] is notable, as it processed 50 million webpages over a period of five months. However, this efficiency falls significantly short when compared to the billions of webpages handled daily by search engines. To address this efficiency challenge, we propose a lightweight detection methodology that incorporates a progressive data reduction strategy, resembling a cascade filter. Our detection methodology is derived from the 3 key observations of the ground-truth dataset in Section 3.2, with distinctive promotional characteristics found in ISAP URLs, including distribution website, promotion target, and promotional content features.

Architecture. Figure 6 depicts the workflow of the ISAP detector, including two primary units, URL Screener and Semantic Detector. First, URL Screener (Section 4.2) could significantly reduce the data volume by eliminating legitimate URLs, based on the distribution website and promotion target. Subsequently, for the remaining URLs requiring inspection (in a relatively small quantity), Semantic Detector (Section 4.3) performs multi-classification based on their semantic features, enabling precise identification of ISAP URLs. Notably, ISAP Detector processes all URLs in the daily snapshot database.

4.2 URL Screener

The primary objective of the URL Screener is to effectively minimize the number of legitimate URLs, enabling real-time detection of the ISAP detector. Drawing from our analysis of ISAP cases, we develop two sets of indicators, including distribution website identification and promotion information.

Distribution website identification. According to *Key Observation 1*, black hat SEO practitioners generate a significant quantity of reflection URLs and embed them across multiple distribution websites as their external URLs. Hence, the distribution websites are the entries in the ISAP URL indexing and spreading process, which is regarded as the initial gateway for detecting ISAP URLs. Specifically, we identify distribution

websites from the snapshot dataset based on the condition of external URLs embedded on these websites, including the number of URLs and search keywords. To avoid excluding ISAP URLs, we choose a conservative filtering threshold. In the end, we consider websites exceeding 3 external URLs and 3 unique keywords as potential distribution sites based on empirical analysis (Section 3.2). Such strict threshold selection is reasonable given the adversaries' objectives and strategies for malicious promotion. Deploying a small number of reflection URLs on a distribution site is not cost-effective for attackers. This not only increases the cost of deploying the distribution website but also makes it difficult to distribute large numbers of reflection URLs in a short period of time.

Promotion information. ISAP promotional information resides within search keywords, requiring comprehensive and easily understandable semantic content. However, the time and computational costs associated with deep learning models render them unsuitable for direct application in this scenario. To address this, we propose a more efficient approach that leverages character features within promotion keywords, i.e., character lengths. First, based on *Key Observation 2*, we note that search keywords are typically longer because they need to include both promotional goals and hot keywords. Therefore, leveraging the insights from Figure 5, we employ a filtering threshold of length 5. Any search keyword shorter than 5 characters is excluded, regardless of whether they are in Chinese, English, or other languages. Although adversaries may use shorter messages to bypass detection, this may weaken their promotional content. Thus, shorter search terms tend to be less likely to serve as ISAP promotion keywords. Furthermore, based on *Key Observation 3*, we observed that while promotion keywords vary, the promotion targets remain consistent. Accordingly, we propose a character co-occurrence-based approach to identify the promotion targets. To extract promotion targets in reflection URLs, we calculate the co-occurrence frequency of each string, i.e., the ratio of a string appearing in total search keywords. Simultaneously, we build a stop word list to filter out false positives, which is comprised of domain names from the collected website list (in Section 3.1) and common words, such as "Facebook" and "Telegram". After that, we identify the candidate promotion targets detected on the distribution website. Moreover, we analyze these promotion targets in Section 6.1.

4.3 Semantic Detector

Leveraging URL Screener, we reduce the number of URLs to be processed by at least 4 orders of magnitude for later efficient real-time detection. Given the similarity between legitimate advertising distribution websites and illicit distribution websites used by black hat SEO [61], it is necessary to verify based on the semantic information of search keywords. To effectively identify ISAP URLs, we develop a detection model based on BERT [13] to learn the semantic distinctions between black hat SEO promotion keywords and regular search keywords.

Table 1: Representative promotion keywords of each category.

Category		Promotion Keywords*
Gambling		Colors Millennium Fortune Website [URL: gd***.cc]
Adult Content		Anime Beauty Characters [BA**.CC] Meet the Live Streaming
BlackHat SEO		SEO Channel[Open:**SEO.cc]
Anonymous Server		AWS Cloud: Open Account Discount with Free Records [TG Telegram: @AK***3]
Other	**Software Development**	Dating IM chat development TG teleg-{ram:@FF***6}
	Fake Invoice	Training fee invoice/c**.htm [Wechat: fp****8]
	Illicit Exam Understudy	Professional Exam understudy: Safe and Reliable [URL AK***9.COM]
	Loan Service	Local loan company, Wechat_k12****7
	Unknown	Hamster Maze [url:gd***-cc]

*: The underlined part is the promotion target, namely the contact information

Furthermore, to understand the ISAP ecosystem comprehensively, it is essential to determine the promotion businesses associated with ISAP URLs. Accordingly, we develop a BERT-based multi-classifier to classify their businesses. We categorize ISAP's promotional activities into five main business categories based on our empirical study (refer to Table 1), including gambling, adult content, SEO, anonymous servers, and others.

4.4 Implementation and Evaluation

In this section, we describe the implementation and evaluation of our ISAP Detector.

Implementation. We implemented our semantic model in a local experimental environment, which is deployed on a Linux server with an Intel(R) Xeon(R) Gold 6139 CPU and 376GB of memory, without GPU. We utilized the pre-trained model "bert-base-chinese" [22] as our base model, which is motivated by the fact that our search engine collaborator primarily caters to Chinese users. In the pre-processing stage, we calculated the text's embedding vector with the default "BertTokenizer". Moreover, we randomly divided the training, test, and validation set 8:1:1. In the end, we trained two models: 1) a binary classifier to detect promotional semantics for ISAP (trained on URL ground-truth dataset); 2) and a five-category classifier for business types (trained on labeled promotion keyword ground-truth dataset). Evaluating the binary classifier on the test and validation sets yielded accuracy of 95.02% and 95.22% for promotional URL detection, respectively. Regarding the five-category classifier, the accuracy on the test and validation sets are 99.47% and 99.40%, respectively.

Evaluation. URL Screener significantly reduces the load an-

alyzed by the semantic detector, allowing the ISAP detector to be scalable enough for search engines. Thus, we first analyzed the individual effect of *URL Screener* on actual traffic. The URL snapshot dataset we use processes an average of 60 million URLs daily. After using the URL Screener, the data volume is reduced to about 10K per day. This amount is already suitable for applying deep learning techniques for efficient semantic analysis. Then, to check whether URL Screener is incorrectly filtering out ISAP URLs, we randomly selected 1,000 URLs each day from the filtered URLs in 14 days from August 17 to August 30, 2023, for a total of 14K URLs. After spendign a day evaluating these URLs manually, two researchers found no ISAP URLs, indicating that the URL Screener is accurate.

Moreover, to assess the effectiveness of our ISAP Detector in real traffic, we conducted manual sampling reviews of the detection results. Over a period of 7 consecutive days, we randomly selected 200 URLs daily, including 100 legitimate URLs and 100 ISAP URLs, and manually checked for false positives and false negatives. For this task, we enlisted 7 volunteers who underwent training to understand the distinctions between promotion keywords and regular search keywords. Based on 1,400 detection samples, our method exhibited a 2.14% false-negative rate and a 1.86% false-positive rate. Therefore, based on the data filtered by URL Screener, we estimate that approximately 200 URLs per day are false positives. These error rates become almost negligible when considering the entire dataset (60 million URLs per day). Moreover, manual checks show false positives often involve adult or gambling sites, which require regulatory oversight in special regions. For instance, pornography sites often place numerous search links with illegal semantic search keywords (e.g., sexual video names) on their pages to attract search engines. Our collaborators received no user complaints after removing the detected ISAP URLs from search results, indicating that these false positives will not affect normal websites. In addition, utilizing the identified promotion targets, we picked out the missed reflection links to reduce the false negatives. For the multi-classification task, we manually evaluated 500 randomly selected samples from each category, resulting in an accuracy of 98.86%. Most misclassifications occurred in the "others" category, likely due to its diverse sub-categories, impacting the model's accuracy.

5 ISAP Website Finder

Notably, ISAP presents a significant challenge to website owners due to its covert and imperceptible nature. From search engines (Section 4), we can detect ISAP that has been abused. However, website owners struggle to identify and prevent this risk proactively. Thus, we design and implement *ISAP Website Finder* to prevent and avoid risks before their sites are exploited and make exploitation for adversaries much harder. Subsequently, we provide a website [1] to offer ISAP risk testing services for website owners.

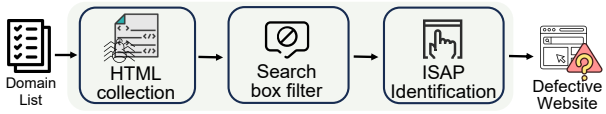


Figure 7: Workflow of ISAP Website Finder.

5.1 Methodology

Challenge. The identification of ISS is challenging due to the lack of a standardized paradigm across various implementations. Kats et al. [25] noted that default prompt words like "search" are commonly used in search boxes for ISS, leading to the keyword matching algorithm. However, this method is limited, including time-consuming, incomplete keyword lists, and difficulties in multilingual settings. Thus, accurately identifying search boxes remains a challenge.

Search box implementation analysis. We found that user input plays a crucial role in the search function, thus serving as the primary motivation for our method. To identify search box candidates, we extract tags for user input from webpages. However, a single tag can serve multiple functions. For example, the <*input*> tag can be used to create a checkbox by setting the attribute <@*type=checkbox*>. While a website may contain numerous <*input*> elements, our focus is solely on the search box. As a result, we leverage tag attributes to reduce noise from non-target tags.

Our observation reveals two distinct types of search box implementations: explicit and implicit. Explicit search boxes are visibly displayed on the webpage, providing a text input field for user entries. On the other hand, implicit search boxes are initially invisible and only appear dynamically when triggered by user clicks, making them undetectable in the original page state. To address this, we employ simulated clicks to activate the dynamic rendering process.

Workflow. Based on the analysis above, we designed an automated ISAP website finder, as shown in Figure 7. It identifies input tags, followed by simulated search operations to verify whether websites can be abused by ISAP.

Step 1: HTML collection. For each given target domain, we crawl its HTML source code, taking into consideration the issue of dynamic rendering of page content.

Step 2: Search box filter. Our approach involves two steps to identify genuine search boxes. First, we identify search boxes with explicit search entity declarations and gather their search URLs following the guidelines[2]. These schemes define the structural attributes associated with ISS. Conse-

[1] http://isap-check.com

[2] `schema.org`, a platform dedicated to creating, maintaining, and promoting structured schemes [44].

Table 2: HTML tag filtering rules.

Attribute	Condition
tag	<input>, <textarea>
tag type	hidden, submit, button, checkbox, radio, email, password, image, tel, reset, file, color, range, date, datetime, datetime-local, month, week

Table 3: Results of manual validation for ISAP Website Finder.

	Predicted to be not at risk			Predicted to be at risk.			Total
	Top10K	GOV	EDU	Top10K	GOV	EDU	
# domain	96	93	96	98	99	100	582
False positive	-	-	-	4	1	0	4
False negative	12	6	15	-	-	-	33

quently, we can directly extract search URLs from these websites, such as `http://example.com/?s={search_keyword}`. For websites lacking explicit declarations, we employ a two-fold approach to handle both explicit and implicit search boxes. We find that clicking on a tag can trigger webpage changes (for implicit search boxes) or result in no changes (for explicit search boxes). By analyzing the post-click webpage, we can identify both types of search boxes. To do this, we leverage the presence of the term "search" within tag attributes (such as `@class`, `@id`) of buttons associated with search functions. This language-independent characteristic allows us to identify potential search function buttons on the webpage and simulate a click on them. We then extract all input-related tags from the new webpages after clicking, including `<input>` and `<textarea>`. Next, we establish our filtering rules based on the analysis of `<input>` types from Tranco Top 10K websites (Table 2). This enables us to obtain a set of candidate search boxes efficiently.

Step 3: ISAP identification. For candidate search boxes, we simulate search operations with specific search keywords (e.g., "isap-test"), then record the redirection URLs and response pages. Subsequently, we use two criteria to verify ISAP issues: 1) existence of search keywords in redirection URLs; 2) containment of search keywords in webpage titles. We concentrate on the title because search engines usually emphasize it (see Figure 2), which enhances promotion. We ignore keyword occurrences in other parts of the webpage. Websites that satisfy these two criteria are deemed affected.

5.2 Implementation and Evaluation

Implementation. We used Pyppeteer [42], a Python library, to implement our ISAP Website Finder. It provides an API to control and automate headless Chromium browsers for web scraping and other automation tasks. It helps us access dynamically rendered webpage content. To enhance evaluation efficiency, we adopted asynchronous multiprocessing, where each process reuses browser resources to minimize unnecessary resource consumption. As a result, we achieved an average evaluation time of 5 seconds per website.

Evaluation. To evaluate the accuracy of our method, we manually inspected a subset of websites due to the lack of website datasets with search boxes and those at risk of ISAP. We randomly selected 200 websites from each of the three lists, including 100 at-risk and 100 not-at-risk, totaling 600 websites. We recruited four volunteers to inspect these websites, assigning each 200 sites—150 unique to them and 50 for cross-validation. Before the inspection, we explained the testing objectives and ISAP risk criteria. We also instructed volunteers to thoroughly inspect each page, noting that the search box might be less visible, located at the bottom of the page, or indicated only by a small magnifying glass icon.

After excluding websites with status changes after their initial assessment, we manually inspected 582 websites, including 285 classified as no-risk and 297 assessed as at risk (Table 3). Our analysis revealed a false positive rate of 1.68% and a false negative rate of 11.58%. These results demonstrate the lower limit of the impact of ISAP. Upon re-evaluating the initially missed domain names, we found that most were overlooked due to wrong HTTP responses during evaluation. False positives are mainly those sites that do not implement internal site search independently, but call the search engine interface directly. These websites use the advanced search syntax (e.g., `site:[target site] [search keyword]`) of search engines to retrieve content indexed in specific sites, rather than searching their own resource libraries. In future work, we will identify such websites to reduce false positives.

6 Understanding the ISAP Abuse

In this section, we analyze ISAP's characteristics based on data from May to September 2023. Additionally, we evaluate the prevalence of ISAP across multiple search engines, including Google and Bing. Moreover, to understand the potential risk of ISAP, we actively assess 3 high-profile website lists.

6.1 Characterizing ISAP in the wild

Scope and Magnitude. Our detector performs efficiently in our collaborators' search engine, processing billions of daily traffic within 2 hours. Through analyzing 125 days of URL snapshots, from May 1 to September 19, 2023, we identified **3,222,864** reflection URLs from **10,209** websites, including 3,607 from the Tranco Top 1M. Some abused websites belong to well-known vendors like `store.google.com` and `support.microsoft.com`, as well as 228 education websites (e.g., `science.mit.edu`) and 162 government websites (e.g., `apod.nasa.gov`). In addition, we found **4,458** distribution websites to spread reflection URLs.

Promotion business. Promotional activity and strategies vary

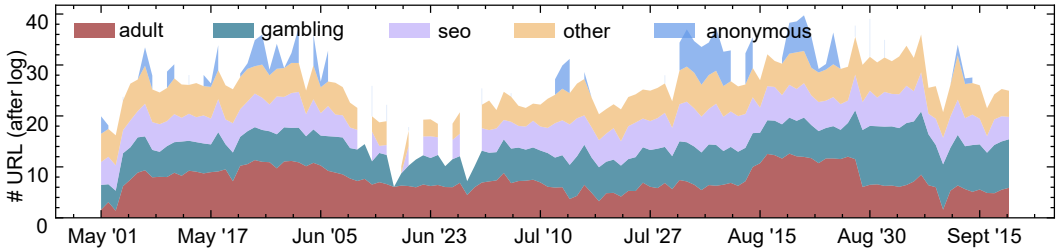


Figure 8: The number of identified promotion URLs per day.

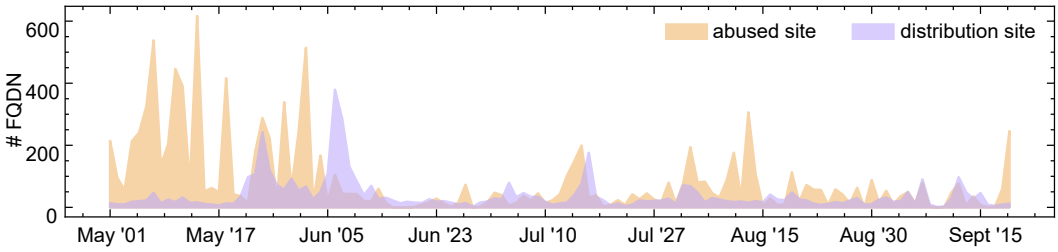


Figure 9: The number of first observed abused and distribution websites per day.

across different businesses. In Figure 8, we observed various businesses being promoted, with diverse daily promotional activities. Adult content and gambling dominate, accounting for 77.44% and 20.41% respectively, consistently showing high daily promotion volumes. Interestingly, we discovered promotion for black hat SEO in ISAP, as shown in Table 1. Additionally, we identified a novel promotion approach involving anonymous servers, as the example in Table 1. Despite lacking sustainability, anonymous server promotion exhibits a burst surge within a short period.

Promotion activity. ISAP is widely and persistently utilized by black hat SEOers. Figure 8 presents the daily ISAP URL distribution of different businesses. It reveals a significant number of abused reflection URLs daily, with peak activity reaching up to 300K. Furthermore, Figure 9 displays the daily trend of newly observed abused and distribution websites. The result indicates a continuous influx of new distribution and abused websites involved in ISAP activities.

Abused website. In total, we identified 10,209 websites abused by various illicit businesses. We observed that 909 websites were abused for more than 30 days. For example, `virtualsc.org` was abused for 100 days. Notably, some websites experienced prolonged abuse, with their promotional activities evolving over time. For example, the website `www.bible.com` was abused for 106 days, and initially promoted adult content but later shifted to gambling. We also conducted a detailed analysis of the top 10 most frequently abused websites, as illustrated in Figure 10. We tracked the daily count of reflection URLs, variations in promotional activities, and the number of promotion targets and distribution websites for these websites. These abused sites are widely propagated across numerous distribution sites, potentially appearing on over 200 different websites in a single day. However, the promotion targets are concentrated, typically involving fewer than five targets per day.

We observed differences in miscreants' strategies for abusing popular and non-popular websites, which may stem from variations in the websites' reputation and influence. Popular domains (in the Tranco Top 1M), known for their high credibility and influence, are maximally exploited by adversaries once an abusable instance is identified. Adversaries generate a large number of reflection URLs on a single site. For example, we observed 154,969 abused reflection URLs under `www.unhcr.org` (ranked 3,969 in Tranco). In contrast, the abuse of less popular sites tends to follow a strategy of accumulating small gains; that is, only a limited number of reflection URLs are created on each affected site, but when aggregated across many such sites, they enable large-scale malicious promotion.

In domain utilization and selection, strategic considerations come into play. Subdomains are also the abused target. `timegenie.com` tops the list with 41 subdomains abused, followed by `facebook.com` with 23 subdomains. This occurs because these websites use subdomains to serve users in different regions. We further examined the categories of abused websites using the classification service of Fortiguard [47]. We found that *Business* websites experienced the highest degree of abuse, comprising a notable 38.19% of the total (Table 4). *Shopping* websites closely followed, while education and government websites were also significantly affected.

Distribution website. Out of 4,458 observed distribution websites, each with an average active period of 3 days, we identified 2 distinct distribution patterns. The first pattern is characterized by stable single-business websites. These web-

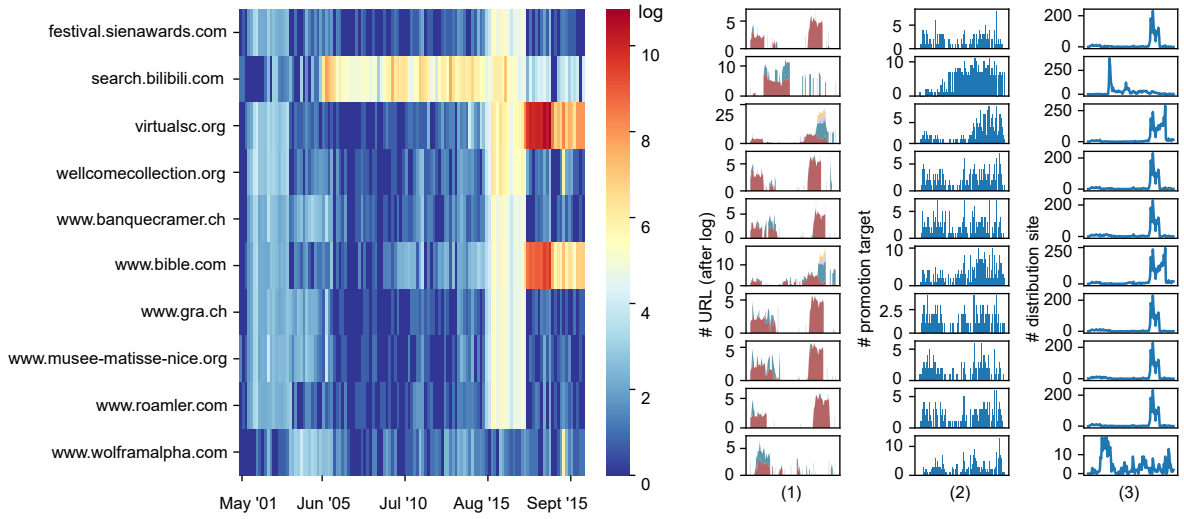


Figure 10: Information of persistently top 10 abused websites. (1) presents the number of promotion URLs (after log) and their category per day for each website. (2) is the number of daily promotion targets for each website. (3) displays the number of distribution websites per day.

Table 4: Category classification results of abused websites.

Website Category	# Website	%
Business	3,893	38.19%
Shopping	1,425	13.98%
Information Technology	753	7.39%
Education	540	5.30%
News and Media	361	3.54%
Government	307	3.01%
Health	292	2.86%
Entertainment	268	2.63%
Travel	186	1.82%
Pornography	180	1.77%
Other	1,989	19.51%

sites have prolonged activity periods, with the longest being 75 days. For example, `www.0852028.com` remained active for 75 days, exclusively promoting gambling content at an average rate of 78 reflection URLs per day. The second pattern is the explosive distribution of websites. These websites have shorter activity periods but distribute a large number of ISAP URLs. For instance, `www.pr779.cn` was active for only 4 days but distributed over 50,000 ISAP URLs in a single day.

Promotion target. Through the string co-occurrence algorithm, we identified 294 promotion targets, with details shown in Table 5, including 205 domains and other contact information. We observed different promotional tendencies among businesses. Adult content and gambling tend to use domains as their promotion targets, whereas other services like SEO and anonymous servers prefer Telegram. Table 6 displays top 10 promotion targets. Promotions targets of the anonymous server are more concentrated, where only 25 distinct

Table 5: Identification results of promotion targets.

Category	Number	# ISAP URL	# Abused site	# Distribution site
domain	205	3,163,724	7,121	4,125
telegram	47	34,267	5,704	181
wechat	25	18,565	1,243	226
telephone	17	6,308	2,693	116

promotion targets were identified in total. For 205 domains, we actively crawled them to assess their promotion activity timeliness. Note that we define the target's active time as the interval between its discovery and our evaluation. Previous studies indicated that illegal activities often have short lifespans [33]. However, we found that promoted websites exhibit much longer, with 67.14% remaining active for over a month. Notably, 23 domains were active for more than 158 days, from May 1 to October 5, 2023. This implies that these websites have effectively evaded existing detection mechanisms, sustaining their operations over time. Enhancing governance over ISAP proves to be a measure to counteract these websites and improve current detection approaches.

User-side impact. We solely measured the link clicks (i.e., Page View, PV) to assess ISAP's impact on users, thereby avoiding the privacy risks associated with analyzing data related to user privacy. In collaboration with Baidu, we quantified PV values [19] for ISAP and distribution URLs. Our estimation indicates that ISAP URLs generated approximately 6 million PVs, and distribution websites accumulated over 2 million PVs within a 4-day period (September 30 to October 3, 2023) based on real search log data. Nearly 6 million page views indicate that ISAP has received significant exposure and interest, suggesting that thousands of users may be lured to illegal services. This underscores the significant and severe

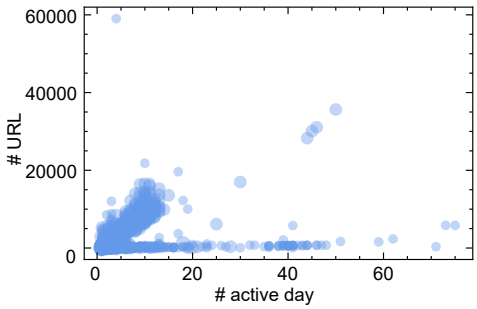


Figure 11: Distribution of reflection URLs and active days for distribution websites.

Table 6: Information of top 10 promotion targets.

Target	Category	# URL	# Day [1]	# Abused site	Business	Active[2]
B**6.CC	domain	1,880,829	16	1,011	adult	Yes
W**7.CC	domain	602,059	25	419	gambling	Yes
z***7.cc	domain	502,127	21	933	adult	No
**v.cc	domain	34,673	31	5	adult	Yes
y**8.me	domain	31,147	13	933	adult	Yes
@A***3	telegram	13,269	34	4,347	anonymous	-
me**ao	wechat	11,188	76	747	seo	-
@A***7	telegram	10,889	39	4,090	seo	-
5**9.com	domain	9,804	42	372	adult	Yes
e**9.com	domain	7,263	33	1,121	adult	Yes

[1]: The number of days present in promotion URLs.
[2]: The activity status of each target was checked on October 15, 2023. For domain names, we visit domains and determine if they are still available. For other target types, we cannot determine their status as we lack an account on that platform.

impact of ISAP.

Summary. The ISAP is extensively used by black hat SEO practitioners to promote various illegal businesses, even including new businesses (e.g., anonymous servers). Black hat SEO practitioners strategically select websites (popular or non-popular) to generate numerous reflection URLs and distribute them according to their business requirements to achieve their promotional goals. Furthermore, the identified ISAPs have had a significant impact on users.

6.2 Investigating ISAP on Google and Bing

Using the ISAP detector, we identified numerous ISAP URLs from a single search engine, Baidu. To demonstrate the prevalence of ISAP across various search engines, we evaluated Google [18] and Bing [37], which are representative search engines. Due to the absence of a URL database, we estimated the ISAP scale and user reachability on these platforms by simulating user search behaviors.

To estimate the scale of ISAP across various search engines, we sampled 182 detected promotion targets as search keywords within these engines and counted the number of results returned. We considered only ISAP URLs on the first page of search results to avoid search engine bias. For Google, we used the exact match mode. In addition, we used attractive keywords to assess user exposure to promotional content. Specifically, we randomly selected ten attractive keywords from the detected promotional keywords of each promotion type. These attractive keywords were used as search queries, and we counted the number of ISAP URLs on the first page of the search results.

ISAP scale on Google and Bing. In Google, we utilized exact match mode to search for promotion targets. We found that search results on the first page for **98 (53.84%)** promotion targets yielded 801 ISAP URLs. Furthermore, we obtained that Google returned up to 32,663,275 URLs in search results for these 98 promotion targets. Such a substantial quantity of ISAP URLs identified through exact matching provides strong evidence of the prevalence of ISAP URLs in Google. In Bing, 387 ISAP URLs appeared on the first-page results for **75 (41.21%)** targets, with 2,097,161,500 search results in total. Since Bing offers no exact search mode, 2 billion ISAP URLs obtained under broad search can approximately simulate the abuse status. Therefore, it is reasonable to infer that a substantial number of ISAP URLs permeate Bing.

ISAP exposure. In Google, **27 (54%)** keywords displayed ISAP URLs on the first page of search results. Remarkably, all first-page results of 5 keywords were exclusively ISAP-promoted content. In Bing, **9 (18%)** keywords revealed ISAP URLs on the first page. We observed that Google focuses on entire search keywords, while Bing returns content related to specific strings within the keywords. For instance, for the search keyword "Anguilla Facebook bidding", Google returns ISAP promotional content, while Bing predominantly returns tutorials related to Facebook bidding, which bear no relation to "Anguilla". This discrepancy may arise from each search engine's unique data processing and algorithms.

Summary. ISAP is prevalent in search engines, injecting significant amounts of promotional content. This negatively impacts data quality, engine performance, and user experience. The prevalence of promotional content on the first page obstructs users from finding desired information and heightens the risk of visiting malicious sites. Moreover, we conducted responsible vulnerability disclosure to the affected search engines and websites (detailed in Section 7.1).

6.3 Evaluating Potential ISAP in the wild

To evaluate the potential impact scope of ISAP in the real world, we employed the method (Section 5) to assess the 3 high-profile domain lists collected in Section 3. Table 7 lists our evaluation results. Notably, some websites were active during the collection phase but were not during the evaluation stage. Additionally, since Tranco integrates multiple lists, including PDNS-based ones, some domains might not offer web services. These factors result in the number of Eva-domain being smaller than Apex. Moreover, we checked each domain three times to mitigate issues brought by network fluctuations.

A substantial number of domains—whether in the Tranco Top 10K, EDU, or GOV—possess the potential to be abused by ISAP in the wild. As listed in Table 7, we checked 57,410 high-profile domains and identified **9,233 (18.19%)** domains

Table 7: Evaluation results of potential ISAP.

Source	# Apex	# Eva-domain[1]	# Vul-domain (%) [2]
Top 10K	10,000	6,647	**782 (11.76%)**
EDU	22,641	21,273	**5,230 (24.59%)**
GOV	24,947	23,008	**3,245 (14.10%)**
Total	57,410	50,762	**9,233 (18.19%)**

[1]: Eva-domain is the domains we evaluated. Others remained inaccessible despite multiple attempts, preventing evaluation.
[2]: Vul-domain represents the domains that are potentially exploited by ISAP. % is the fraction of the Vul-domain and Eva-domain (i.e., Vul-domain / Eva-domain).

subjected to ISAP risks. Among the Top 10K domains, we identified 782 domains that can be exploited by ISAP, like *bbc.com* and *aol.com*.

Moreover, within the EDU, affected domains are most prevalent, with up to **24.59%** being susceptible, thereby offering abundant resources for black hat SEOers. Although numerous domains within the GOV generate URLs containing search keywords, the actual number of affected domains is less than in the EDU. This can be attributed to the manner in which these websites are designed and implemented. Our observation revealed that the page titles of certain government sites remain constant, often merely reflecting the names of the institutions they represent. Such behavior hampers the promotional efforts undertaken by black hat SEOers.

Summary. Many high-profile domains in the wild are at risk of ISAP due to their implementation flaws, furnishing black hat SEO practitioners with an abundance of free resources.

7 Discussion

7.1 Responsible Disclosure

To reduce the impact of ISAP, we responsibly disclosed this threat to affected search engines and websites, and received positive responses.

• *For search engines*, we disclosed ISAP risks and identified ISAP URLs for affected search engines, including Baidu, Bing, and Google. Baidu operates our system daily on the offline snapshot database to detect ISAP. Furthermore, Baidu also removed detected ISAP URLs and associated distribution sites from search results, greatly mitigating the impact of ISAP. Specifically, when removing ISAP URLs, Baidu takes care to maintain the reputation of the website. Moreover, Bing indicated they had implemented a fix for ISAP, while we are discussing mitigation strategies with Google.

• *For vulnerable websites,* for one hand, we organized online training and provided remediation guidance for vulnerable university and government websites, with the help of the national Computer Emergency Response Team and the China Internet Network Information Center. Until now, Xiamen University has fixed this issue, and other organizations have plans to deploy fixes based on our recommendations. In addition, we disclosed the information about vulnerable corporate websites to their Security Response Center (SRC). For example, Tencent and Yahoo acknowledged the threat and awarded us a bug bounty.

Up to now, we have received 37 acknowledgment letters and are still helping vulnerable entities to fix this threat.

7.2 Limitations

Despite our best efforts to understand and assess the ISAP threat, some limitations remain. First, singular data sources confine our understanding of ISAP, despite its vast user base of over a billion users. To address this limitation, we also examined ISAP for other search engines in Section 6.2. While this assessment is preliminary and approximate, it still provides an overview of ISAP and confirms the pervasive nature of ISAP threats. Second, our methodology still has limitations. For the ISAP detector, other languages in ISAP may impact the performance of the NLP model. We tested pretrained models supporting multiple languages but observed no significant performance improvements. For ISAP website finder, although we have made struggle efforts to account for various implementations of internal site search functions, unpredictable factors may still exist. Methodological limitations lead us to find the low-bound of ISAP threats in the wild. However, our findings are sufficient to highlight the risk and offer valuable defensive insights to the security community. Malicious promoters may change their strategies to evade our detector. However, it is important to note that our method employs a strict threshold, and any attempts to circumvent it will either destroy promotion information or increase promotion costs. According to the analysis of detection results provided by our collaborators, no significant evasion cases have been observed thus far. Lastly, we only evaluated high-profile apex domains, while many of their subdomains remain susceptible to abuse. Therefore, our active assessment can provide a conservative estimate for potentially abused domains. We will public our proactive evaluation methodology in Section 5, empowering website operators to assess their own websites regularly.

7.3 Ethics

In accordance with established ethical guidelines, specifically the Belmont Report [16] and the Menlo Report [26], our study adhered strictly to ethical principles in 3 key areas: ISAP analysis, data collection, and vulnerability disclosure. For detected ISAP targets, e.g., illegal websites, we did not interactively analyze their content (like webpages), avoiding the ethical aspects of studying and interacting with potentially illegal or unethical content, such as adult or gambling sites. Regarding data collection, our collected data did not contain any user-specific information, only including domain names, URLs, and HTML content. Given potential ethical concerns, all data processing and analysis were conducted on secure internal

servers within Baidu, with limited access granted to interns. The handling of Baidu's data was supervised by their legal committee. For the evaluation of high-profile domains, we conducted only a few specific legal test requests to minimize any potential impact. Furthermore, we took precautions to prevent search engines from indexing the newly generated search URLs, thereby avoiding the pollution of search engines.

7.4 Lessons Learned

Based on our findings, we propose the following recommendations for mitigating ISAP. Search engines should proactively detect and mitigate ISAPs by implementing technical measures, leading to improved search results and enhanced user experience. Website operators should assess their internal search functionality for ISAP risks. Our proactive evaluating tool can help websites assess their websites for ISAP threats, while the access will be under specific conditions to prevent potential misuse. They can also utilize tools like Google Search Central [2] to analyze indexed data and promptly report abused URLs to search engines for remediation. For affected websites, we advise mitigating ISAP risks in two aspects. First, avoid generating abusable reflection URLs, e.g., using `POST` for parameter transmission and excluding searching keywords in result page URLs and titles. Second, prevent indexing of non-existent resource URLs, by configuring "noindex" in the response for search keywords with empty search results [45].

8 Related Works of Black Hat SEO

Previous studies have established the effectiveness of black hat SEO in disseminating fraudulent, and illicit content [29, 35, 36, 49]. Moore et al. [38] and Nektarios et al. [27, 28] revealed the illicit promotion activities targeting emerging popular keywords to drive substantial user traffic for their promotions. Several studies examined various promotional tactics employed in black hat SEO, including distributing spam messages in forums [39]. Wang et al. [51] investigated the effectiveness of black hat SEO botnets in manipulating popular search terms. John et al. [23] uncovered a method involving the compromise of legitimate website servers to generate and inject numerous counterfeit promotional pages associated with popular search terms.

As understanding of black hat SEO techniques deepens, adversaries continue to enhance their promotional technologies. Adversaries are diversifying their resources and network infrastructure. Liao et al. [30] confirmed the utilization of cloud hosting platforms for malicious promotion. In addition to exploiting search functions, miscreants also misuse automatic search term recommendation services [53] and nearby business search services [52] for malicious promotion. Moreover, adversaries employ obfuscation techniques in their promotional content, including homoglyph attacks to deceive search engines [8, 64], phonetic confusion poisoning [24], and cloaking techniques to hide malicious promotional content [50].

Various detection methodologies have been proposed based on the characteristics of known SEO attacks, falling into two main categories: network behavior and text content analysis. Network behavior-based methods often rely on the characteristics and connectivity of the infrastructure used in black hat SEO activities. Du et al. [14] identified wildcard domain names exploited by spider pools, and Wang et al. [23] detected promotional links hosted on compromised websites based on fingerprint features in URL dimensions. As search redirection is a crucial black hat SEO technique, some studies proposed detection methods based on link relationships [10, 34, 55, 63]. In addition, some work studied detection methods based on the cloaking phenomena [54, 56] and search engine visibility assessment [62]. Another detection method is based on textual content [15, 40, 46]. Liao et al. [31] recommended detecting based on the semantic differences between malicious keywords and top-level domains, whereas Yang et al. [60] introduced detection using the semantic discrepancies between page topic keywords and search keywords. Additionally, Yang et al. [61] proposed a label-embedded mechanism with natural language techniques (NLP) to detect SEO defacement. To address pages with a mix of normal and malicious content, Yang et al. [59] proposed a method combining multiple NLP models' detection capabilities.

Our research focuses on an overlooked malicious promotion technique, i.e., Internal site Search Abuse Promotion (ISAP). While previous studies observed ISAP URLs through external links in the spider pool [14], we fill the gap of effective detection methods and systematic understanding for ISAP. In collaboration with a renowned search engine vendor, we develop a lightweight detection method and provide the first comprehensive understanding of the ISAP ecosystem based on longitude monitoring results.

9 Conclusion

In this paper, we conducted the first systematic study on ISAP, assisting our collaborators in effectively mitigating the ISAP threats. Based on our empirical analysis of the real-world ISAP cases, we proposed a lightweight detection method. We undertook a four-month detection of ISAP, processing over one billion URLs with high efficiency. In the end, we identified 3,222,864 ISAP URLs, abusing 10,209 websites. Moreover, we revealed the abusive strategies towards various websites and the promotion patterns of distribution websites. In addition, we developed an automated ISAP website finder to assist website operators in proactively evaluating ISAP risks. Utilizing this tool, we assessed three high-profile domain lists and identified 9,233 affected domains, indicating the severity of ISAP. Finally, we responsibly disclosed the ISAP threat to vulnerable search engines and websites with ISS functions, and are helping them to fix it.

Acknowledgments

We sincerely thank all anonymous reviewers and our shepherd for their valuable comments on improving the paper. This work is in part supported by the National Key Research and Development Program of China (No. 2021YFB3100500), the National Natural Science Foundation of China (62102218), CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (CCF-Tencent RAGR20230116). Haixin Duan is supported by the Taishan Scholars Program. Most of this work was done by Yunyi Zhang during the joint Ph.D. program between National University of Defense Technology and Tsinghua University.

References

[1] Majestic Million CSV now free for all, daily. https://blog.majestic.com/development/majestic-million-csv-daily/, 2012.

[2] Google Search Central Blog. Prevent portions of your site from being abused by spamr. https://developers.google.com/search/blog/2021/05/prevent-portions-of-site-from-spam, 2023.

[3] 360 Search Engine. Link Submission. https://info.so.360.cn/site_submit.html, 2023.

[4] Alexa Internet, Inc. Global Top Sites. https://web.archive.org/web/20081216072512/http://www.alexa.com:80/site/ds/top_sites, 2008.

[5] Baidu. Baidu Search Enginge. https://www.baidu.com/, 2023.

[6] Baidu Search Engine. Baidu Link submission. https://ziyuan.baidu.com/linksubmit/url, 2023.

[7] Bing Search Engine. Anonymous URL Submission Tool Being Retired. https://blogs.bing.com/webmaster/september-2018/Anonymous-URL-Submission-Tool-Being-Retired, 2018.

[8] Nicholas Boucher, Luca Pajola, Ilia Shumailov, Ross J. Anderson, and Mauro Conti. Boosting big brother: Attacking search engines with encodings. *CoRR*, 2023.

[9] Bryan Henderson. User-generated spam from internal search. https://support.google.com/webmasters/thread/105827120/user-generated-spam-from-internal-search?hl=en, 2021.

[10] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. Know your neighbors: web spam detection using the web topology. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[11] Young-joo Chung, Masashi Toyoda, and Masaru Kitsuregawa. A study of link farm distribution and evolution using a time series of web snapshots. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, 2009.

[12] Cisco. Cisco Umbrella. https://umbrella.cisco.com/, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[14] Kun Du, Hao Yang, Zhou Li, Hai-Xin Duan, and Kehuan Zhang. The ever-changing labyrinth: A large-scale analysis of wildcard DNS powered blackhat SEO. In *25th USENIX Security Symposium*, 2016.

[15] Dennis Fetterly, Mark S. Manasse, and Marc Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In Sihem Amer-Yahia and Luis Gravano, editors, *Proceedings of the Seventh International Workshop on the Web and Databases*, 2004.

[16] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*. Department of Health, Education and Welfare, 1979.

[17] Google. Google Link Submission. https://www.google.com/webmasters/tools/submit-url, 2023.

[18] Google. Google Search Engine. https://www.google.com/, 2023.

[19] Google. How Site Search metrics are calculated. https://support.google.com/analytics/answer/1032321, 2023.

[20] Google. URL Inspection Tool. https://support.google.com/webmasters/answer/9012289?hl=en, 2023.

[21] Guangsuan Technology. What are Google Search Traces. https://www.guangsuan.com/post/%e4%bb%80%e4%b9%88%e6%98%af%e8%b0%b7%e6%ad%8c%e6%90%9c%e7%b4%a2%e7%95%99%e7%97%95/, 2024.

[22] Hugging Face. bert-base-chinese. https://huggingface.co/bert-base-chinese, 2023.

[23] John P. John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martín Abadi. deseo: Combating search-result poisoning. In *20th USENIX Security Symposium*, 2011.

[24] Matthew Joslin, Neng Li, Shuang Hao, Minhui Xue, and Haojin Zhu. Measuring and analyzing search engine poisoning of linguistic collisions. In *2019 IEEE Symposium on Security and Privacy*, 2019.

[25] Daniel Kats, David Luz Silva, and Johann Roturier. Who knows i like jelly beans? an investigation into search privacy. *Proceedings on Privacy Enhancing Technologies*, 2022.

[26] Erin Kenneally and David Dittrich. The menlo report: Ethical principles guiding information and communication technology research. *Available at SSRN 2445102*, 2012.

[27] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *20th USENIX Security Symposium*, 2011.

[28] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[29] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Click trajectories: End-to-end analysis of the spam value chain. In *32nd IEEE Symposium on Security and Privacy*, 2011.

[30] Xiaojing Liao, Chang Liu, Damon McCoy, Elaine Shi, Shuang Hao, and Raheem A. Beyah. Characterizing long-tail SEO spam on cloud web hosting services. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[31] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhongyu Pei, Hao Yang, Jianjun Chen, Hai-Xin Duan, Kun Du, Eihal Alowaisheq, Sumayah A. Alrwais, Luyi Xing, and Raheem A. Beyah. Seeking nonsense, looking for trouble: Efficient promotional-infection detection through semantic inconsistency search. In *IEEE Symposium on Security and Privacy*, 2016.

[32] Zilong Lin, Zhengyi Li, Xiaojing Liao, XiaoFeng Wang, and Xiaozhong Liu. MAWSEO: adversarial wiki search poisoning for illicit online promotion. In *Proceedings of 2024 IEEE Symposium on Security and Privacy*, Oakland S&P '24, 2024.

[33] Mingxuan Liu, Yiming Zhang, Baojun Liu, Zhou Li, Haixin Duan, and Donghong Sun. Detecting and characterizing SMS spearphishing attacks. In *ACSAC '21: Annual Computer Security Applications Conference*, 2021.

[34] Long Lu, Roberto Perdisci, and Wenke Lee. SURF: detecting and measuring search poisoning. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, 2011.

[35] Xiulin Ma. Research on black hat seo behaviour measurement. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018.

[36] Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M. Voelker, Stefan Savage, and Kirill Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21th USENIX Security Symposium*, 2012.

[37] Microsoft. Bing Search Engine. https://www.bing.com/, 2023.

[38] Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. Fashion crimes: trending-term exploitation on the web. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, 2011.

[39] Yuan Niu, Hao Chen, Francis Hsu, Yi-Min Wang, and Ming Ma. A quantitative study of forum spamming using context-based analysis. In *Proceedings of the Network and Distributed System Security Symposium*, 2007.

[40] Alexandros Ntoulas, Marc Najork, Mark S. Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, 2006.

[41] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *26th Annual Network and Distributed System Security Symposium*, 2019.

[42] pyppeteer. Pyppeteer. https://github.com/pyppeteer/pyppeteer, 2023.

[43] Quantcast. Ranking Top Websites by Demographics | Quantcast. https://www.quantcast.com/blog/ranking-websites-by-demographics/, 2023.

[44] Schema.org. SearchAction. https://schema.org/SearchAction, 2023.

[45] Search Console Help. Pharm Spam Search Query Crawl and Indexed by Google. https://support.google.com/webmasters/thread/103486956?hl=en, 2024.

[46] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking web spam with HTML style similarities. *ACM Trans. Web*, 2008.

[47] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. Misshapes, mistakes, misfits: An analysis of domain classification services. In *Proceedings of the ACM Internet Measurement Conference*, 2020.

[48] Tom van Goethem, Najmeh Miramirkhani, Wouter Joosen, and Nick Nikiforakis. Purchased fame: Exploring the ecosystem of private blog networks. In Steven D. Galbraith, Giovanni Russello, Willy Susilo, Dieter Gollmann, Engin Kirda, and Zhenkai Liang, editors, *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019.

[49] David Y. Wang, Matthew F. Der, Mohammad Karami, Lawrence K. Saul, Damon McCoy, Stefan Savage, and Geoffrey M. Voelker. Search + seizure: The effectiveness of interventions on SEO campaigns. In *Proceedings of the 2014 Internet Measurement Conference*, 2014.

[50] David Y. Wang, Stefan Savage, and Geoffrey M. Voelker. Cloak and dagger: dynamics of web search cloaking. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, 2011.

[51] David Y. Wang, Stefan Savage, and Geoffrey M. Voelker. Juice: A longitudinal study of an SEO botnet. In *20th Annual Network and Distributed System Security Symposium*, 2013.

[52] Peng Wang, Zilong Lin, Xiaojing Liao, and XiaoFeng Wang. Demystifying local business search poisoning for illicit drug promotion. In *29th Annual Network and Distributed System Security Symposium*, 2022.

[53] Peng Wang, Xianghang Mi, Xiaojing Liao, XiaoFeng Wang, Kan Yuan, Feng Qian, and Raheem A. Beyah. Game of missuggestions: Semantic analysis of search-autocomplete manipulations. In *25th Annual Network and Distributed System Security Symposium*, 2018.

[54] Baoning Wu and Brian D. Davison. Cloaking and redirection: A preliminary study. In *AIRWeb 2005, First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[55] Baoning Wu and Brian D. Davison. Identifying link farm spam pages. In *Proceedings of the 14th international conference on World Wide Web*, 2005.

[56] Baoning Wu and Brian D. Davison. Detecting semantic cloaking on the web. In *Proceedings of the 15th international conference on World Wide Web*, 2006.

[57] Xiaojinniu. How to dominate the search results in Google at zero cost? https://www.youtube.com/watch?v=8TlOQevghjk, 2024.

[58] Qinge Xie, Shujun Tang, Xiaofeng Zheng, Qingran Lin, Baojun Liu, Haixin Duan, and Frank Li. Building an open, robust, and stable voting-based domain top list. In *31st USENIX Security Symposium*, 2022.

[59] Hao Yang, Kun Du, Yubao Zhang, Shuai Hao, Haining Wang, Jia Zhang, and Haixin Duan. Mingling of clear and muddy water: Understanding and detecting semantic confusion in blackhat SEO. In *Computer Security - ESORICS 2021 - 26th European Symposium on Research in Computer Security*, 2021.

[60] Hao Yang, Xiulin Ma, Kun Du, Zhou Li, Hai-Xin Duan, XiaoDong Su, Guang Liu, Zhifeng Geng, and Jianping Wu. How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In *2017 IEEE Symposium on Security and Privacy*, 2017.

[61] Ronghai Yang, Xianbo Wang, Cheng Chi, Dawei Wang, Jiawei He, Siming Pang, and Wing Cheong Lau. Scalable detection of promotional website defacements in black hat SEO campaigns. In *30th USENIX Security Symposium*, 2021.

[62] Jialong Zhang, Xin Hu, Jiyong Jang, Ting Wang, Guofei Gu, and Marc Ph. Stoecklin. Hunting for invisibility: Characterizing and detecting malicious web infrastructures through server visibility analysis. In *35th Annual IEEE International Conference on Computer Communications*, 2016.

[63] Jialong Zhang, Chao Yang, Zhaoyan Xu, and Guofei Gu. Poisonamplifier: A guided approach of discovering compromised websites through reversing search poisoning attacks. In *Research in Attacks, Intrusions, and Defenses - 15th International Symposium*, 2012.

[64] Qing Zhang, David Y. Wang, and Geoffrey M. Voelker. Dspin: Detecting automatically spun content on the web. In *21st Annual Network and Distributed System Security Symposium*, 2014.

A ISAP URL Examples

In this section, Table 8 provides several examples of ISAP URLs and embedded search keywords.

Table 8: ISAP URL examples.

Distribution Website	Category	Reflection URL	Search Keyword	Promotion Target
http://54***ie.top	Adult	https://www.bhliquors.com/catalogsearch/result/?q={search keyword}	Yunnan Dali home massage appointment phone number {WeChat 152****8840} provides first-class door-to-door service UteCW	152****8840
		https://www.ncbi.nlm.nih.gov/medgen/?term={search keyword}	How to find special escort services in Wuxi {WeChat phone number 132****9532} provides first-class door-to-door service eZXzl	132****9532
		https://store.google.com/br/search?q={search keyword}	Xiamen door-to-door (one-stop door-to-door service) {WeChat phone number 132****9532} provides first-class service ikYQHy	
http://3g.08***28.com	Gambling	http://www.sbaad.no/?s={search keyword}	World Expo International Platform official website homepage (900***.net)	900***.net
		http://app.pureflix.com/videos/search?q={search keyword}	Xinbao Sports APP official website (visit 900***.net)	
		http://blog.cpanel.com/?s={search keyword}	Macau Lisboa official website login registration (input 52***.net)	52***.net