# Collecting Fake Shopping Website URLs via Product Image Search

Masaki Yoshii
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
y.masaki@nict.go.jp

Shintaro Kawamura
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
kawashin@nict.go.jp

Nagisa Nakamura
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
n-nagisa@nict.go.jp

Hidekazu Tanaka
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
tanaka.hidekazu@nict.go.jp

Shingo Yasuda
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
s-yasuda@nict.go.jp

Daisuke Inoue
*Cybersecurity Nexus*
*National Institute of Information and Communications Technology (NICT)*
Tokyo, Japan
dai@nict.go.jp

***Abstract*—One type of phishing site is the fake shopping website, which imitates legitimate e-commerce platforms to steal money, credit card details, and personal information. These sites often reuse product images from legitimate stores without permission and lure users through compromised legitimate sites and SEO (Searach Engine Optimization) poisoning. As part of the WarpDrive project, NICT has been conducting passive observation and analysis of web-based attacks by collecting web access logs via the Tachikoma Security Agent distributed to users. However, because malicious websites appear and disappear frequently, proactive URL collection is essential for effective trend analysis. In this study, we focused on fake shopping websites and developed a method to collect URLs by using product images as seeds. Leveraging the Google Cloud Vision API, our approach retrieves the URLs of websites hosting similar images and captures both pre- and post-redirect URLs, enabling more effective acquisition of malicious site information.**

***Keywords*—Observing Web-based Attacks, Web Security, Cybersecurity, Malicious Website**

## I. INTRODUCTION

Phishing sites targeting users in Japan often exploit legitimate websites, including not only shopping sites but also those of educational institutions, judicial organizations, and tourism services. Attackers gain unauthorized access to these sites, tamper with their content, or insert redirect scripts. Such compromised legitimate sites, hereafter referred to as stepping-stone sites, are used to lure users to phishing pages. In addition, SEO poisoning, which manipulates web search results, increases the likelihood that these stepping-stone sites leading to phishing pages appear at the top of search results.

In response to this situation, the National Institute of Information and Communications Technology (NICT)* has been promoting a user participatory cybersecurity project in Japan called WarpDrive (Web-based Attack Response with Practical and Deployable Research InitiatiVE) [1]. The project aims to understand the actual state of web-based attacks and to improve countermeasure technologies. Under WarpDrive, NICT distributes two free applications to participating users: the Tachikoma** Security Agent (SA) for PCs and the Tachikoma SA Mobile for Android smartphones. These applications observe, analyze, and alert users about web access activities, enabling continuous monitoring and analysis of phishing and other malicious websites, including fake shopping sites [2][3][4].

The WarpDrive project collects web access logs based on participants' browsing activities. Phishing sites often modify or delete content and change URLs; for fake shopping sites, 13.7% of domains change within 17 days [5]. As phishing behaviors vary over time, proactive data collection is needed. Because product images are frequently reused on fake sites, using them as search queries is an effective way to collect related URLs.

In this study, we aimed to proactively collect a large number of malicious site URLs and improve analysis performance. To meet this objective, we defined the following requirements:

*Requirement 1*: The method must enable semi-automated collection of fake shopping site URLs by using product images as seeds.

*Requirement 2*: The collected URLs must allow investigation of the actual characteristics of fake shopping sites.

In this paper, "semi-automated" refers to actively collecting URLs and performing maliciousness classification based on seed images. To satisfy Requirement 1, product images from suspected fake shopping sites collected by Tachikoma SA and Tachikoma Mobile are used as seeds. We retrieve the URLs of sites hosting identical or similar images and their redirect destinations, then evaluate maliciousness, the number of top-level domains (TLDs), and the number of domains.

The structure of this paper is as follows. Section 2 describes the new method for collecting fake shopping site URLs using product images. Section 3 presents an evaluation of the new method. Section 4 addresses research ethics. Section 5 concludes this paper.

## II. COLLECTING FAKE SHOPPING SITE URLS USING PRODUCT IMAGE SEARCH

### A. Product Images on Fake Shopping Websites

According to Reference [5], two methods have been identified for loading product images onto fake shopping websites:

* https://www.nict.go.jp/

**© Masamune Shirow, Production I.G / Kodansha, Ghost in the Shell: SAC_2045 Production Committee

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

*1)* Specifying the image URL of a legitimate shopping site directly in the “src” attribute of the <img> tag.

*2)* Specifying the URL of an image proxy server in the “src” attribute to indirectly fetch the image.

Therefore, performing image searches using product images obtained from fake shopping websites is expected to find numerous mirror sites and stepping-stone sites.

Figure 1 shows an example of search results obtained by querying with product images collected from websites suspected by Tachikoma SA to be fake shopping sites. For ethical reasons, the product images and site names are blurred. In the search results, more than 70 websites were found to use exactly the same product image. The URLs of these sites included TLDs such as “.sn” (Senegal) and “.cl” (Chile). In several cases, accessing these sites resulted in redirection to completely different domains. Some of the sites appearing in the search results were found to be compromised legitimate sites that redirected users to malicious sites, including fake shopping websites. Therefore, collecting the URLs of websites that appear in image search results can help identify sites where redirection occurs, including stepping-stone sites.

## *B. Product Image Search Method*

Figure 2 shows an overview of the method for collecting fake shopping site URLs using product images. The procedure is as follows:

*1)* URL information of malicious websites actually accessed by Tachikoma SA users is shared.

*2)* URLs suspected to belong to fake shopping sites are provided from the BlockList maintained by Tachikoma SA.

*3)* On the STARDUST platform [6], the provided URLs are actually visited and product images are manually collected. The collected images are hereafter referred to as seed images.

*4)* The seed images are submitted to Google Cloud via the Google Cloud Vision API [7] (hereafter, Google Vision API).

*5)* Using the “Web entities and pages” feature of the Google Vision API, URLs of sites hosting visually similar images are extracted.

*6)* The URL extraction results are received from Google Cloud.

*7)* The collected URLs and the curl command are used to retrieve redirect destination URLs.

*8)* The redirect destination URLs are scanned by the antivirus (AV) engines hosted by VirusTotal [8] via the API (hereafter, VT-API). We obtain the number of anti-virus engines that flag a URL as malicious (hereafter, the score).

*9)* On the STARDUST platform, URLs judged as malicious are visited and visually inspected to extract characteristic features.

Fig. 1 Example of Search Results Based on Product Images from Suspected Fake Shopping Websites. (In this example, product images of automobile parts are used for the search.)

Previous work has reported cases where redirection does not occur unless the Referrer header from a search engine is present [5]. Although methods that emulate human web interaction, such as using Selenium, could be employed to trigger such redirects, we opted to collect redirect destination URLs using the curl command in this study to avoid imposing excessive load on the search engines. Because the redirection conditions may vary depending on the stepping-stone site, differences in performance are expected between manual URL collection and collection using the curl command. Therefore, this aspect is evaluated in Section 3. Each function from steps (4) to (8) was implemented individually using Python 3.8.10.

The overview and rationale for adopting the STARDUST platform, Google Vision API, and VT-API are described separately in Sections 2-C to 2-E.

## *C. STARDUST*

The STARDUST platform [6], developed by NICT, is a cyberattack attraction system that enables observation of human-like attack behaviors. While tools such as Windows Sandbox could also be used to collect product images from fake shopping sites, they pose risks of executing malicious programs [9]. Therefore, STARDUST was adopted for its ability to safely collect product images and analysis data even when malicious programs are executed.

## *D. Google Vision API*

Using Google’s standard image search can detect dozens of URLs per query, but automated searching violates Google’s terms of service [10], and no dedicated API is available for image-based web searches. To address this, the

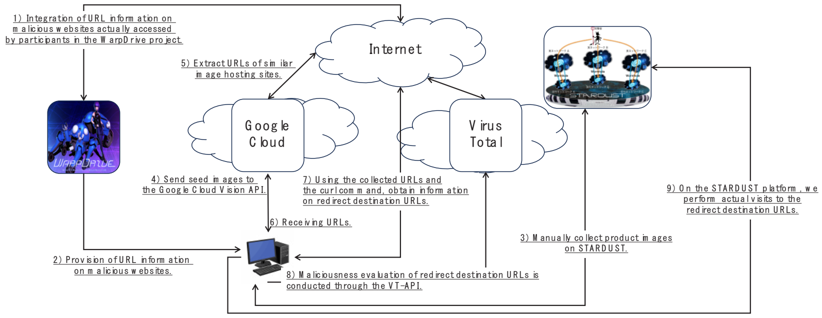
1) Integration of URL information on malicious websites actually accessed by participants in the WarpDrive project.
Internet
5) Extract URLs of similar image hosting sites.
WarpDrive
Google Cloud
Virus Total
STARDUST
4) Send seed images to the Google Cloud Vision API.
7) Using the collected URLs and the curl command, obtain information on redirect destination URLs.
9) On the STARDUST platform, we perform actual visits to the redirect destination URLs.
6) Receiving URLs.
3) Manually collect product images on STARDUST.
2) Provision of URL information on malicious websites.
8) Maliciousness evaluation of redirect destination URLs is conducted through the VT-API.

Fig. 2 Method for Collecting URLs Using Product Images from Fake Shopping Websites.

Google Vision API [7] was used. This API, provided through the Google Cloud Platform, applies machine learning to recognize objects and faces and to extract URLs of sites containing identical or similar images, although it returns only about 10 URLs per request.

### E. *VirusTotal API*

The VT-API [8] is an interface that enables automated file uploads, URL scans, and retrieval of scan results from VT. VT uses multiple AV engines from various security vendors to reduce false detections; therefore, it was adopted in this study. The number of AV engines that label a URL as "malicious" is used as the score.

## III. Evaluation of the Image-based Method

### A. *Evaluation Metrics*

Because the purpose of our method is to collect malicious URLs, it is necessary to evaluate both the collected URLs and their maliciousness classification. Therefore, an evaluation was conducted from the following perspectives. As a prerequisite, 426 non-duplicate product images collected from one suspected fake shopping website were obtained through Tachikoma SA and used in the image-based method. Duplicates were identified by checking for identical file sizes and through visual inspection.

***Evaluation 1: Number of Collected URLs***

We evaluated the number of URLs collected via the Google Vision API; the number of redirect destination URLs manually collected using these URLs; and the number automatically collected with the curl command when setting the Google, Yahoo, and Bing Referrer headers. Each Referrer type was evaluated separately because previous research [5] has shown that redirection behavior varies depending on the Referrer source. In the evaluation, if the redirect destination URL was identical to the URL obtained by the Google Vision API, or if no redirect occurs, it was excluded from the count.

***Evaluation 2: TLD Analysis of Stepping-Stone Sites***

We evaluated the number of TLDs used by stepping-stone sites and quantitatively analyzed both country code and generic TLD distributions. The analysis used the results obtained from executing the curl command with the Google Referrer setting.

***Evaluation 3: Frequent Domains of Stepping-Stone Sites***

We quantitatively evaluated the occurrence of company names, brand names, and commercial domains to analyze what kinds of sites are being exploited as stepping-stone sites. The analysis also used the results of the curl command with the Google Referrer setting.

***Evaluation 4: VT-API Scan Results***

We evaluated the number of URLs whose redirect destinations were detected as malicious, defined as those with an AV score of three or higher. For comparison, the numbers of URLs with AV scores of one or higher and two or higher were also obtained. Both manually collected URLs and URLs collected automatically using the curl command with the Google Referrer setting were evaluated.

***Evaluation 5: TLD Analysis of URLs Scanned by VT-API***

We evaluated the distribution of TLDs among the URLs scanned via the VT-API and quantitatively analyzed their trends.

### B. *Evaluation 1: Number of Collected Stepping-Stone URLs*

Figure 3 shows the number of collected URLs. The Google Vision API was used to obtain 1,959 URLs. Manual access to these URLs yielded 1,050 redirect destinations. With the curl command, 509, 506, and 28 URLs were collected when setting the Google, Yahoo, and Bing Referrers, respectively. On average, about 5 URLs were collected per product image (426 images).

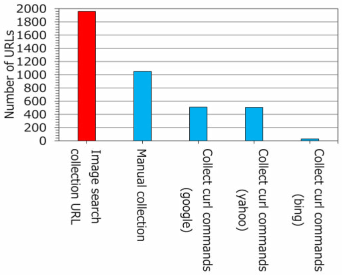

Fig. 3 Evaluation Results of Collected URL Count.

Manual collection produced more results than did curl, likely because some stepping-stone sites block requests without full browser Referrers. Although automation with Selenium could improve efficiency, it cannot be used due to Google's terms of service. The number of collected URLs differed by Referrer type (Google > Yahoo > Bing), suggesting that attackers mainly target users of search engines popular in Japan.

### C. *Evaluation 2: TLD Analysis of Stepping-Stone Sites and Evaluation 3: Frequent Domain Analysis of Stepping-Stone Sites*

Figure 4 shows the top 20 TLDs used by stepping-stone sites. The most frequent was ".in" (India, 159 cases), followed by ".br" (Brazil, 139) and ".com" (generic TLD, 88).

Figure 5 presents the top 20 frequent domains, with partial masking (*) for ethical reasons. The leading domains

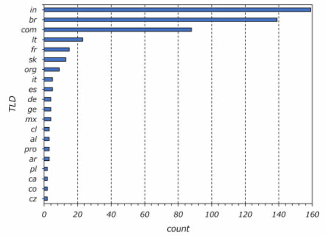

Fig. 4 Evaluation of the Number of TLDs in Redirect Sites.

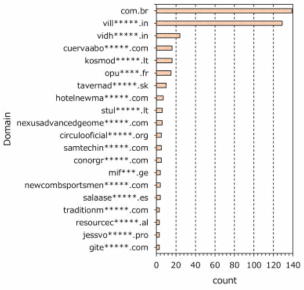

Fig. 5 Number of Corporate, Brand, and Commercial Domains.

were “.com.br” (Brazilian commercial, 139), “vill*****.in” (Indian tourism, 129), and “vidh*****.in” (Indian legal/educational, 24).

Many compromised tourism and hotel sites appear related to increased post-COVID-19 inbound demand and insufficient security measures [11].

### D. Evaluation 4: Scan Results Using the VT-API

Figure 6 shows the VT-API scan results. Among manually collected redirect URLs, 669 were flagged as “malicious” (score ≥ 1), 619 as ≥2, and 596 as ≥3. For URLs collected using curl with the Google Referrer, 596, 509, and 490 were respectively flagged at the same thresholds. Thus, 57% of manually collected URLs (596 of 1,050) and 82% of automatically collected URLs (490 of 509) had a score ≥ 3. Manual collection captured more malicious URLs because it included user-driven redirects, while curl mainly retrieved URLs embedded in JavaScript. However, repaired stepping-stone sites sometimes redirected browsers to legitimate pages, increasing non-malicious detection.

The high malicious ratio (≈82%) in curl results indicates that many redirects are intentionally embedded in compromised site scripts.

### E. Evaluation 5: TLD Analysis of URLs Scanned by the VT-API

Figure 7 shows the TLD types and counts among manually collected redirect URLs with a “malicious” score of ≥3. The most frequent TLD was “.com” (generic TLD, 521 cases), followed by “.shop” (23) and “.click” (20). Prior studies [5] have also reported frequent use of “.shop” and “.click” in suspected fake shopping sites. In total, 14 different TLDs were observed.

Figure 8 shows the TLDs of redirect URLs collected using the curl command with a score of three or higher. The most frequent was “.com” (477 cases), followed by “.click” (13); only these two TLDs appeared. TLDs such as “.com” and “.shop” are inexpensive or free to obtain, which likely explains their frequent use by attackers. The prevalence of “.click” suggests intentional use to track or trigger user redirections. Similar trends were observed for searches via “bing.com” and “yahoo.co.jp.”

These results indicate that attackers may adjust redirect conditions based on TLDs, while the limited variety of TLDs used highlights an area for further analysis.

## IV. Research Ethics

In this study, product images were collected from websites suspected to be fake shopping sites. To minimize the load on target sites, the collection was performed manually. For URL collection using the curl command, the same consideration was applied, and the command was executed only once per site. No data other than redirect URLs were collected.

## V. Conclusion

In this paper, we describe a semi-automated method for collecting stepping-stone URLs and redirect destination URLs of websites suspected to be fake shopping sites, using product images as seeds. To achieve this, two requirements were defined (Section I). As a result, 1,959 image search URLs and 490 redirect URLs with a “malicious” score of ≥3 were obtained through the VT-API, satisfying Requirement 1.

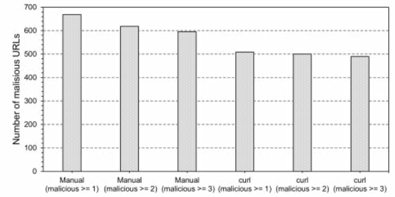

Fig. 6 VT-API Scan Results.

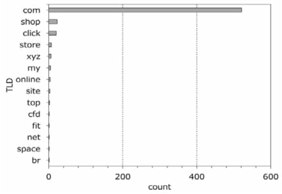

Fig. 7 Evaluation of the TLD Count in Maliciously Identified URLs Collected Manually.

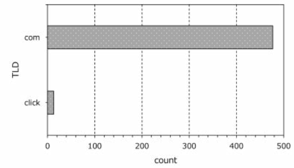

Fig. 8 Evaluation of the TLD Count in Maliciously Identified URLs Collected via Curl Commands.

Future work includes verifying the adequacy of Requirement 2, improving the accuracy of redirect URL acquisition, and conducting data analysis from a cognitive science perspective.

## References

[1] “*WarpDrive*,” https://warpdrive-project.jp/ (accessed 7 Oct. 2025).

[2] M. Hasegawa, A. Saino, A. Fujita, K. Takada, R. Tanabe, C. H. Ganan, M. van Eeten, and K. Yoshioka, “*POSTER: Do You Sell This? Utilizing Product Searches to Find SEO-driven Fake Shopping Sites*,” Network and Distributed System Security Symposium (NDSS2025), Feb. 2025.

[3] D. Miyashita, S. Kobayashi, and T. Yamauchi, “*Investigation Towards Detecting Landing Websites for Fake Japanese Shopping Websites*,” 13th International Conference on Emerging Internet, Data and Web Technologies (EIDWT2025), Lecture Notes on Data Engineering and Communications Technologies, Vol. 243, pp. 107-119, Apr. 2025.

[4] T. Yamauchi, R. Orito, K. Ebisu, and M. Sato, “*Detecting Unintended Redirects to Malicious Websites on Android Devices Based on URL-Switching Interval*,” IEEE Access, Vol. 12, pp. 153285-153294, Oct. 2024. DOI: 10.1109/ACCESS.2024.3478748.

[5] H. Kodera, T. Koide, D. Chiba, K. Aoki, and M. Akiyama, “*Understanding Attacks with Fake Shopping Websites*,” IPSJ Journal, Vol. 62, No. 9, pp. 1523-1535, Sep. 2021 (in Japanese). DOI: 10.20729/00212758.

[6] N. Kanaya, E. Suzuki, O. Nakamura, Y. Umemura, and S. Sato, “*STARDUST: Large-scale Infrastructure for Luring Cyber Adversaries*,” Journal of NICT, Vol. 70, No. 2, pp. 15-27, 2024.

[7] “*Cloud Vision API*,” https://cloud.google.com/vision?hl=en (accessed 7 Oct. 2025).

[8] “*VirusTotal*,” https://www.virustotal.com/gui/home/url (accessed 7 Oct. 2025).

[9] NISC, “Alert: Cyberattacks by MirrorFace (Provisional Translation),” https://www.nisc.go.jp/eng/pdf/Alert_MirrorFace.pdf (accessed 7 Oct. 2025).

[10] Google, “*Google Privacy & Terms*,” https://policies.google.com/terms?hl=-en-US (accessed 7 Oct. 2025).

[11] L. Florido-Benítez, “*The Role of Cybersecurity as a Preventive Measure in Digital Tourism and Travel, A Systematic Literature Review*,” Discover Computing, Vol. 28, No. 28, Apr. 2025. DOI: 10.1007/s10791-025-09523-3.