

BACHELOR'S THESIS COMPUTING SCIENCE



RADBOUD UNIVERSITY NIJMEGEN

Detecting Brandjacking-Based Malvertising

Author:
Chris Musteata
s1035303

First supervisor/assessor:
dr. Güneş Acar

Second assessor:
dr. ir. Bart Mennink

July 8, 2024

Abstract

Online advertising is the cornerstone of revenue generation for brands across the world. Buying products is a process influenced by the brands associated with them and the ads we see around us. Sometimes, seeing the name of a company is enough motivation to buy their services. This leaves space for individuals to misuse a brand's trust with malicious intentions, by brandjacking. This study aims to investigate the use of web crawling techniques to detect instances of brandjacking-based malvertising. The focus is on developing a Playwright-based web crawler capable of identifying malicious advertisements by analyzing landing URLs, redirection chains, and ad content. Results indicate that the web crawler successfully identifies instances of malvertising and more specifically brandjacking. This study offers insights into enhancing the real-time detection of malicious advertisements and offers a basis for future research to improve detection accuracy and coverage.

Contents

1	Introduction	2
1.1	Problem Motivation	3
1.2	Knowledge Gap	4
1.3	Problem Solution	4
2	Preliminaries	5
3	Related Work	7
3.1	Introduction	7
3.2	Key Studies	7
4	Methodology	10
4.1	Tool Design	11
4.2	Data Collection	11
4.3	Data Analysis	13
5	Results	19
5.1	Summary of results	19
5.2	Malicious results	21
5.3	Non-malicious results	21
5.4	Domain statistics	22
6	Discussion	23
6.1	Malicious cases	25
6.2	Efficiency of detection methods	31
6.3	Limitations	32
6.4	Research Ethics	33
6.5	Future work	33
7	Conclusions	34
A	Appendix	39

Chapter 1

Introduction

In the evolving landscape of digital advertising, malvertising has emerged as a significant threat to online security and user trust. Malvertising, short for malicious advertising, involves integrating malicious code within online advertisements, which can lead to malware infections, data breaches, and other harmful consequences when users interact with these ads [1].

A particularly deceptive form of malvertising is brandjacking-based malvertising, where cybercriminals assume the identity of well-known brands to create ads that appear legitimate, thereby exploiting the trust that users place in these brands. This tactic not only harms users but also damages the reputation of the targeted brands. An example of such an advertisement can be seen below in Figure 1.1. This ad aims to fool unsuspecting users by impersonating “Uber Eats”, a platform used for food delivery. Upon clicking the link, the user is greeted by a message telling them that they can redeem a 29\$ coupon (Figure 1.2). They are prompted to enter their phone number/e-mail and their password to log in. While the true intent of the person who created the page is unknown to us, our tool found this ad to be a case of brandjacking.

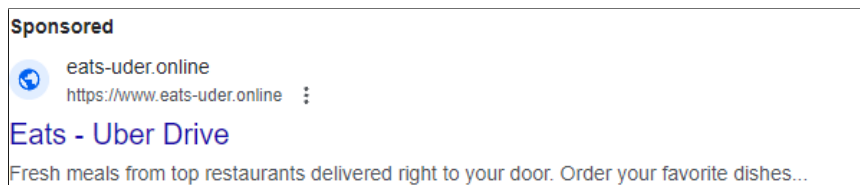


Figure 1.1: Example of ad impersonating “Uber Eats”

The research question this paper aims to answer is: *How can web crawling, alongside detection and analysis tools, be used to identify instances of brandjacking-based malvertising?*

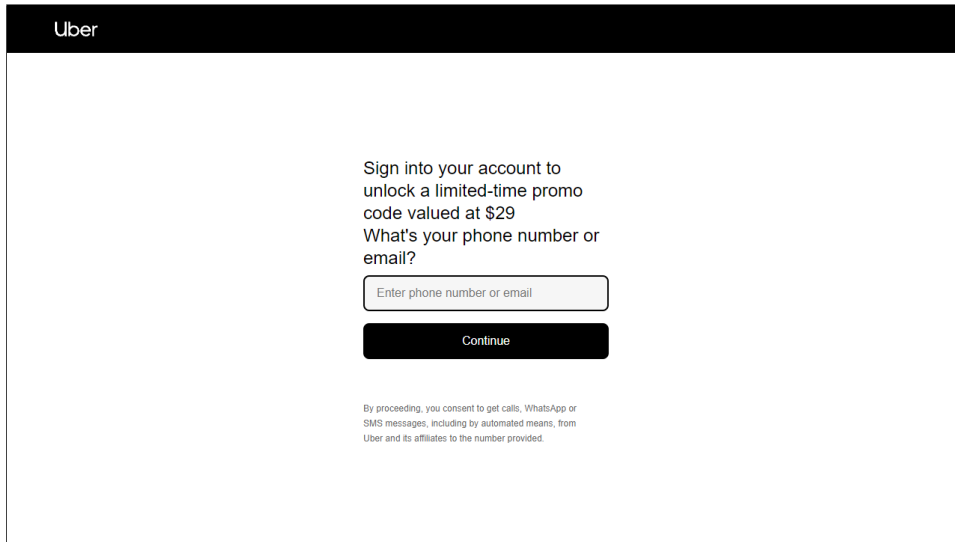


Figure 1.2: Landing page for `eats-uder.online`

1.1 Problem Motivation

Between 2022 and 2023, the Internet advertising industry generated an all-time high revenue of \$225 billion. This represents an increase of over 7% compared to the previous year [2]. This ecosystem’s importance makes it a prime target for cybercriminals seeking to exploit unsuspecting users through tactics such as brandjacking-based malvertising [3]. Brandjacking involves the unauthorized use of reputable brands’ identities to deceive users into interacting with malicious advertisements, potentially leading to malware infections, phishing attacks, and identity theft.

Cases of brandjacking can be traced back as far as the early 2000s [3]. However, despite advertiser networks’ efforts to combat malvertising, detection methods often struggle to keep pace with cybercriminals’ evolving tactics. This is evidenced by reports of such ads throughout 2023 [4, 5] and as recently as January 2024 [6]. A 2022 study from Semrush found that approximately 45% of users click on a search result within 5 seconds, and nearly 74% do so within 15 seconds [7]. This rapid browsing behavior, combined with cybercriminals’ persistent efforts to avoid detection, leaves users vulnerable to interacting with deceptive advertisements posing as legitimate brand promotions. This shows the urgent need for more advanced and proactive detection methods to protect users from these quickly evolving threats.

1.2 Knowledge Gap

While both industry and academia are working to combat malvertising [1], the ongoing threat of brandjacking-based malvertising highlights critical gaps in the current understanding and capabilities. Traditional detection methods such as blocklist-based filtering are often reactive and fail to keep up with the dynamic tactics used by cybercriminals [8]. DataProt reported that in 2023, a new phishing site was created every 11 seconds [9]. This makes it nearly impossible for blocklists to counter these attacks effectively.

Existing research often focuses on the broader picture of malvertising detection [1, 10, 11, 12]. However, this does not thoroughly address the specific challenges associated with brandjacking. The seemingly random surges in malvertising incidents underscore the recurring nature of this threat [4, 5, 6]. Despite awareness of the issue, there is a notable lack of solutions that address the detection of brandjacking-based malvertising. Particularly, there is a need for solutions that integrate real-time web crawling to improve detection accuracy and response times.

1.3 Problem Solution

The proposed tool makes automated Google searches, creating a dataset of advertisements targeted at a predefined list of terms. Information about each ad (screenshots, video recordings, HAR files) is stored and analyzed. Landing domains and corresponding advertisements are then labeled as malicious or non-malicious using four different detection platforms. Finally, our tool showcases statistics about the presence of malicious ads in different contexts, e.g., country-based, term-based, domain-based.

Structure Chapter 2 offers background information about search-based advertising and malvertising. Chapter 3 gives an overview of existing methods and related academic work. Chapter 4 presents the tool’s design and the methods used for data collection and analysis. Chapter 5 reports on the findings of our crawler. Chapter 6 analyzes the results, presents interesting findings, and discusses limitations. Chapter 7 concludes this paper.

Chapter 2

Preliminaries

Search-based advertising is one of the components of digital marketing, primarily involving ads that appear on the main pages of search engines like Google, Bing, and Yahoo. When users input search queries, these search engines display relevant ads alongside the organic search results [13]. The placement of these advertisements is determined by a process known as real-time bidding, or RTB for short.

In RTB, advertisers bid for ad space in real time, competing for visibility based on the relevance of their ads to the search queries. Advertisers leverage various user data, including demographics, previous search behavior, and browsing history, to create targeted ads that are more likely to resonate with potential customers [14]. This targeted advertising ensures that users see ads that are relevant to their interests and needs, increasing the effectiveness of the advertising campaign. The entire process, from the user's search to the display of targeted ads, happens within milliseconds, making it a highly dynamic and efficient system.

Malvertising disrupts this advertising ecosystem by embedding malicious code within seemingly legitimate ads. These ads can exploit vulnerabilities in the ad delivery process, leading to malware infections, data theft, or redirection to phishing websites when users click on the ads [10]. Since search engines display ads mixed in with organic results, malvertising poses a significant risk to users who might unknowingly interact with harmful ads.

The large presence of advertising networks, as well as the automated nature of ad placement through RTB, provides malvertisers with opportunities to infiltrate said networks. By exploiting the trust users place in well-known brands, malvertisers can deploy their content more effectively, making it a significant threat for digital advertising.

Countering malvertisements is not straightforward. To evade detection and maximize the impact of their campaigns, malvertisers employ various techniques. One common method is cloaking, where the content presented differs from one user to another. This helps malicious websites go undetected by ad network reviewers. This tactic involves targeting specific demographics or user profiles who receive the malicious content, while others are served the “clean” versions [15]. By doing so, malvertisers can pass initial reviews and gain entry into legitimate ad networks.

Additionally, malvertisers frequently change their tactics and infrastructures to stay ahead of detection efforts. This includes rotating domains [16], using short-lived campaigns, and leveraging ad fraud techniques such as click fraud and impression fraud [17, 18]. These methods make it challenging for ad networks to identify and block malicious ads with the necessary speed.

Chapter 3

Related Work

In this chapter, the focus is on the existing body of research surrounding the detection and mitigation of malicious advertising. By examining the methodologies, findings, and limitations of these works, the aim is to get an understanding of the current research within the broader context of the field and highlight the unique contributions of our approach.

3.1 Introduction

The field of detecting and mitigating malicious advertising has seen significant research and development in recent years, particularly through the use of web scraper tools. Many studies have employed web scraping technologies to collect and analyze data from various online sources [11, 12, 19], similar to the approach taken in this paper. However, the web crawler tool developed in this study is unique in its design and application by focusing on Google's ad network, rather than on individual websites. This choice stems from the abundance of existing research on the latter category [10, 11, 12, 19]. In turn, this allows the current study to contribute new insights and strategies to the field.

3.2 Key Studies

Zhou et al. delved into the tactics used by cybercriminals to spread malicious ads online, namely obfuscation and frequent modifications to avoid detection [1]. Zhou et al. developed a framework that combined static and dynamic analysis to examine ads and the behavior of associated landing pages. The methodology involved crawling the top 90,000 Alexa websites, reconstructing ad redirection chains, identifying ad-delivery paths, annotating nodes with various attributes, and using a machine-learning approach to generate detection rules for malvertising. Their resulting tool, MadTracer, achieved a great detection rate, while also maintaining the number of false

positives very low. This study showcased the sophisticated and ever-evolving nature of malvertising and underscored the need for a multifaceted approach to combat these threats.

Zarras et al. went further into the shady world of malicious advertising by aiming to uncover the tricks and tactics used by cybercriminals. A dataset was first built by performing a large-scale web crawl to collect over 600,000 real-world advertisements [10]. The websites crawled were compiled from two different data feeds, one provided by an antivirus company and the second one being Alexa’s top 1M sites. These websites were visited using a browser automated via Selenium [20], capturing the content of the ads and the HTTP traffic. The researchers also designed an oracle, consisting of three components, that would automatically classify the recorded advertisements. This study showed that most of the time, publishers and advertisers work on the basis of trust. This means that publishers rarely employ additional detection measures to filter out malicious ads. This work also showed that some ad exchanges serve more malicious ads than others due to insufficient detection systems and the ad arbitration process allows malicious ads to more easily infiltrate ad exchanges.

Moti et al. investigated the prevalence and nature of tracking and malvertising on websites aimed at children. They find that a significant number of these websites host tracking technologies that collect data without explicit consent, thus often violating privacy regulations designed to protect minors [11]. The results were collected by crawling a list of 2,000 child-directed websites. Information was collected about the advertisements themselves, as well as the advertisers (using the “Why this ad?” section). The data was analyzed using multilingual language models, to classify any ads that may be problematic for children. The study revealed that 27% websites contained ads that should not be displayed without explicit parental consent.

Masri and Aldwairi explored a system designed to identify malicious advertisements by combining three online malware detection tools: VirusTotal, URLVoid, and TrendMicro [12]. Their methodology involved selecting websites from two different data feeds (Alexa’s top 1M websites and a blacklist) and then using Selenium for web browser automation. The tool extracted advertisement URLs and then submitted these URLs to the three aforementioned platforms. Each ad was classified based on the results returned. This study found URLVoid to be a more reliable tool than the other two. Nonetheless, it showed that no tool is the “best” and that the only way to ensure higher accuracy is by combining functionalities from different platforms.

Subramani et al. studied the growing issue of push-based advertisements as a significant vector for delivering malicious content. Their 2020 paper presented PushAdMiner, a system for automated collection and analysis of web-based push notifications (WPNs) [21]. They extended existing Chromium browser crawlers to also track Service Workers and WPNs in detail. The crawler had the capability of visiting websites and granting notification permissions. The resulting WPNs and their landing pages were collected and analyzed. Subramani et al. found a total of 5,143 WPN ads and classified more than 50% of them to be malicious. The study was the first to systematically capture and analyze this malvertising attack vector. The high ineffectiveness of traditional ad-blockers and URL filters coupled with the large amount of data collected, showed the urgent need for better detection methods.

The 2024 paper by Nettersheim et al. investigated the effectiveness of various internet services in identifying and categorizing ad malware [19]. The researchers extended their tool, Katti, to crawl websites and fetch HTTP requests related to online advertisements. They then queried these requests against filtered DNS providers and VirusTotal and compared the responses from different Internet services to evaluate the definition of ad malware. By leveraging the aforementioned services, they examined how these services label suspicious content. Their findings revealed significant discrepancies in how different services classify ad malware, with some services flagging a higher proportion of domains than others. The study showed the need for standardized definitions and more transparent criteria in ad malware detection, as well as the importance of considering the entire URL structure in future research.

While these studies make significant contributions to the field, there remain significant gaps in the existing literature that the current research aims to address. By focusing on the development of a unique web crawler tool and original detection methodologies, this study seeks to provide new insights and strategies for combating malicious brandjacking-based advertising.

Building upon the foundation of the related work discussed in this chapter, the next chapter will delve into the methodology used in this study. It will provide a detailed description of the web crawler tool developed for detecting malicious advertisements, including its architecture, data collection process, and analysis techniques.

Chapter 4

Methodology

This chapter outlines the systematic approach used to develop, implement, and evaluate the web crawler which was designed to detect brandjacking-based malvertising. The methodology contains the design and development of the web crawler, the process for collecting data, as well as the techniques used to analyze this data. Each step is detailed in order to provide a clear understanding of the methods used and to ensure the tool's reliability and effectiveness in identifying malicious ads. An overview of the tool's functionality can be seen in Figure 4.1.

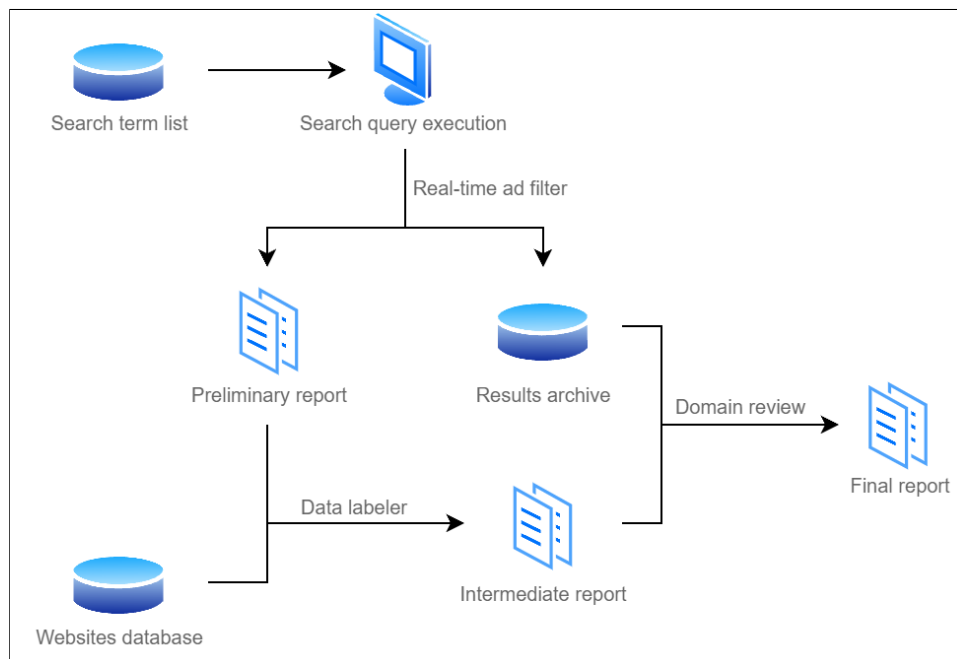


Figure 4.1: Overview of tool functionality

4.1 Tool Design

The web crawler was developed using Playwright [22], a powerful framework for web automation that can interact with web pages in a similar way that a human would. Its functionality is closely related to other tools such as Selenium [20] or Cypress [23].

Setup and configuration

The tool was configured to run on Chromium-based browsers. Different settings were fine-tuned in order to more accurately simulate a real user interaction. For example, the crawler runs with headless mode disabled and pauses between command execution. This was done to minimize the risk of being flagged as a bot and ensure that the results are representative of genuine user experiences.

Search query execution

The crawler was programmed to input predefined search terms related to popular software and websites from a dictionary of 50 terms. This list was compiled from two data feeds, namely the Tranco [24] and Kantar [25, 26] lists. Each term in the dictionary is paired with an allowlist of known legitimate domains, which are later used to analyze the ads. This information is available in appendix A.

VPN setup

One of the main criteria that Google bases ads on is the location of the user. In order to get a better view and maximize the chances of encountering brand-jacking ads, different locations should be explored. This is done using the Mullvad VPN service [27]. Each run is made across 10 different countries and the results are recorded independently. This list consists of the USA (**us**), Canada (**ca**), Australia (**au**), United Kingdom (**gb**), Netherlands (**nl**), Sweden (**se**), Romania (**ro**), Brazil (**br**), South Africa (**za**), Thailand (**th**).

4.2 Data Collection

The data collection process involves gathering advertisements from search engine results based on predefined search terms. This section details the procedures used to capture relevant data, including the identification of ads, the collection of information (i.e., screenshots, video recordings, HAR files), and the methods used to ensure the accuracy and relevance of collected data. The aim is to build a comprehensive dataset for subsequent analysis.

Advertisement filter

The tool scrolls the search results page up to the chosen point and renders all advertisements. It then treats each ad individually and assesses its relevance. This filter minimizes the number of flagged ads by assessing how closely each ad matches the search term or its associated known domain. Since the point of the crawler is to identify brand-jacking attempts, any ads that do not resemble the search term are deemed irrelevant. For example, if the crawler is analyzing advertisements for “Amazon”, it will only inspect ads which contain the search term or its associated known domain.

Main page collection

When an ad of interest is found, the crawler captures screenshots of crucial information, including the ad itself (Figure A.1) and the “Why this ad?” section. This section contains information about the advertiser (Figure A.2) and it also offers an insight into what criteria the search engine used for displaying the ad (Figure A.3). During this step, the advertiser ID is also recorded, accessible through the “See more ads” button. This ID can be used later to view all ads currently being run by this company or individual (Figure A.4).

Accessing the advertisement

A request listener is also enabled at this point in the process. This listener has the purpose of storing the redirection chain for any navigation requests the browser makes. Some ads may decide in real time that the crawler is not within the target population and redirect it to a legitimate landing page. The redirect chain helps capture any suspicious websites along the way. The next step consists of clicking the ad. Then, the crawler screenshots the landing page of the website (Figure 4.2). It also saves the landing URL for later inspection.

Video recordings

Throughout the entire search query execution, the crawler is also set up to record videos of each page [28]. Hence, after finalizing, recordings of the main page as well as each landing page are saved for future reference. These recordings are useful for analyzing unforeseen edge cases and serve as evidence if the landing page becomes unavailable in the future.

HAR files

To facilitate a more detailed analysis of the HTTP traffic and redirections, the crawler also stores the HTTP archive of the web browser. In case a

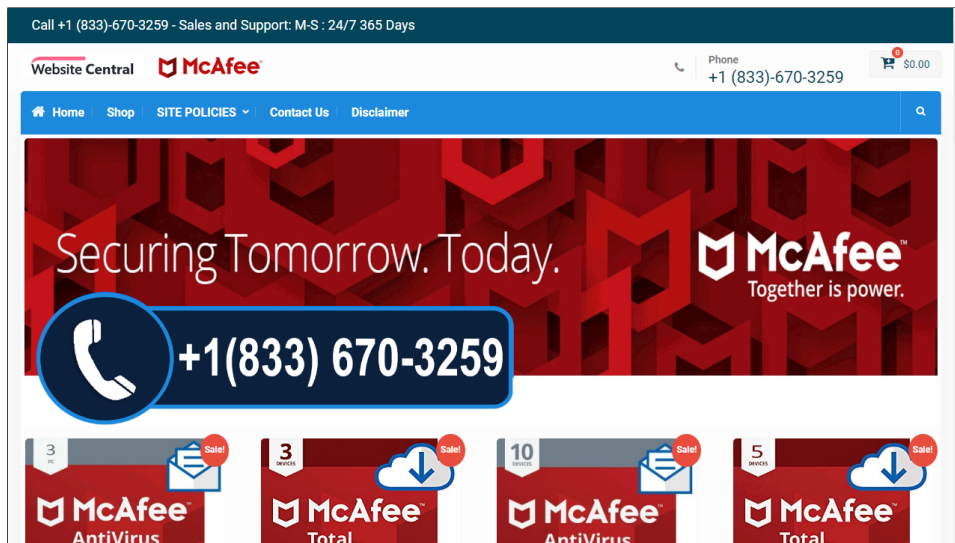


Figure 4.2: Landing page screenshot captured by the crawler

search term is interesting (i.e., it has advertisement campaigns running), the crawler will record the HAR file of the entire interaction. This makes it possible to re-create the environment in which the ads were found, as well as further analyze any suspicious requests and pages.

4.3 Data Analysis

The data analysis section outlines the techniques used to examine the collected ads to determine their legitimacy and detect instances of brandjacking-based malvertising. This includes validating URLs as well as analyzing landing pages and redirect chains. Due to the nature of brandjacking-based malvertisement, a final review may be required in order to ensure the validity of claims.

Result evaluation

Upon finalizing an entire run, i.e., all search terms were queried, the results are automatically evaluated by the crawler. This step is done in order to present only relevant information and reduce the time and effort spent on domain review. Different information is compared with the allowlist and raises corresponding warnings. Firstly, the landing URL of each ad is matched with the known domain(s) using regex and a warning is raised in case of a mismatch. Secondly, the redirect chain is also compared with the allowlist of domains and redirects. If an unknown website is detected among the requests, the corresponding warning is raised. If needed, this system is easily expandable to add more automatic checks.

Warning modes Both warning systems have a **strict** and a **relaxed** mode. In the **strict** mode, the landing URL of each ad is matched only with the known domain of the specific term. Similarly, every redirect of each ad is matched with the known domain of the specific term and with the redirect allowlist. In the **relaxed** version, both landing URLs and redirects are matched with all the known domains available in the crawler. This latter version of the warning system results in fewer flagged ads, while not affecting the number of detected malicious cases. The reasoning here is that landing on a website within the known domain list will not be malicious, except in the highly unlikely event that one of the known domains gets hijacked.

Preliminary report

The results of a run are displayed in the form of a report. This report contains relevant information for each ad, grouped on search terms. It can also be seen how many advertisements were found per term, as well as which warnings were raised. For each warning corresponding information is also displayed, i.e., if the URL warning is raised, then the landing URL is also displayed in the report. This feature works similarly for the redirection chain. The advertiser ID is displayed to facilitate further investigations in case any warnings are raised. An excerpt of a summary can be seen below for better visualization of the results. The discrepancy between the total number of ads found and the number of ads displayed comes from the ad filter. Since 4 advertisements are missing that means that they were not relevant to the search term, as explained in the “Advertisement filter” section.

```
avast — Ads found: 9
  Ad 1: 0/2 warnings

  Ad 3: 0/2 warnings

  Ad 4: 1/2 warnings
    Redirect chain: {'mcafeeinc.demdex.net', 'www.mcafee.com', ...}
    Ads transparency ID: AR07041635852870483969?origin=ata

  Ad 6: 0/2 warnings

  Ad 7: 2/2 warnings
    Landing URL: https://blitzhandel24.co.uk/avast/[...]
    Redirect chain: {'monitor.clickcease.com', 'blitzhandel24.co.uk', ...}
    Ads transparency ID: AR03282036780072697857?origin=ata
```


Data labeling

In the preliminary report, the number of flagged domains is quite high. For example, the largest part of flagged domains consists of outlet stores. Terms such as “samsung” or “nike” lead to a significant number of flagged advertisements due to various outlet stores selling these products. An example of such an ad can be seen in Figure 4.3. The crawler flags these ads correctly, as they do not land on the known domain. However, this does not immediately mean the ad is malicious.

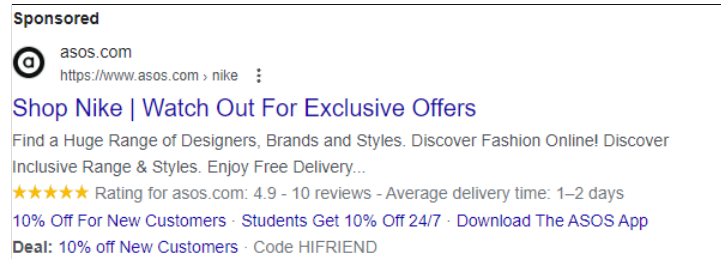


Figure 4.3: Ad of an outlet store for search term “nike”

The second largest contributor to flagged domains is related but legitimate software. Sometimes, related software may target keywords included in the tool’s search term list. This leads to the ad filter deeming these advertisements as relevant. Then, these ads are flagged as suspicious due to the difference between the landing URL and the known domain. An example of such a related ad can be seen in Figure 4.4.

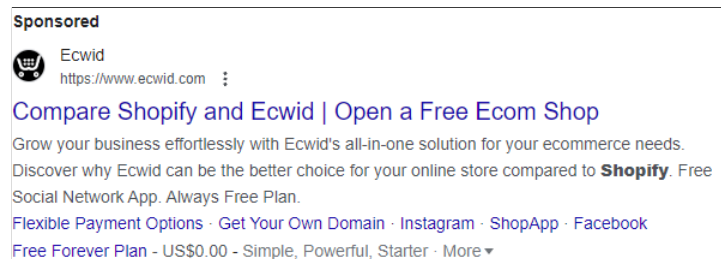


Figure 4.4: Ad of related software for keyword “shopify”

Labeler A separate data labeler is used to further categorize the flagged domains and advertisements present in the preliminary report. This labeler makes use of a database of websites, where each website is part of a category (presented later in the “**Flag System**” subsection). All preliminary reports run through the labeler, which analyzes the results of each ad individually. If the landing URL warning is present in the preliminary report, then the labeler compares this URL with its database. Flags are assigned to the ad based on which category the landing URL falls under. If the landing URL

is not sufficient for a verdict, then the labeler also compares the advertiser ID with its database and flags the ad accordingly. If an automatic verdict cannot be given, for example, due to the landing URLs and/or advertiser IDs not being categorized, then the user is notified that a domain review is required. This notification consists of a message that points the user to the ad (and its associated file) which needs reviewing.

Database The database of websites previously mentioned is compiled with the help of the labeler. The process is circular as the labeler results are used to extend the database, which in turn improves the performance of said labeler. During labeling, the occurrences of each domain are counted. If this number is larger than or equal to the cut-off value (chosen for our tool to be 3) a notification is sent. This notification tells the user that a domain requires categorization. The user can then manually review this domain and distribute it in one of the categories. The process of domain review will be explained more in-depth in the next subsection. The aforementioned cut-off value was chosen to be 3 as it offers the “greatest balance” between manually reviewing domains and manually reviewing ads. The definition of “greatest balance” will be given in chapter 6. All domains that needed review were grouped under one of the nine categories: *outlet/retail*, *related software*, *blogs/forums*, *course platforms*, *courier services*, *search engines*, *unrelated*, *app download platforms*, and *brandjacking*. The *unrelated* category refers to ads that were flagged due to multiple interpretations of a keyword. For example, “UPS” can refer to “United Parcel Service” (our target) and also to “Uninterrupted Power Supply” (unrelated). Similarly, “facebook” can return ads for contact pages of other websites (Figure 4.5) and “booking” is widely used by platforms similar to `booking.com` without the intention of brand-jacking (Figure 4.6).

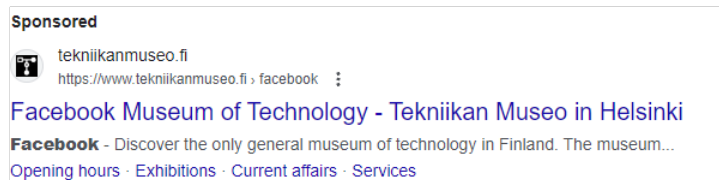


Figure 4.5: Unrelated ad pointing to a Facebook contact page



Figure 4.6: Unrelated ad due for keyword “ups”

Flag system This system categorizes the landing URL of each advertisement based on the previously presented database. Corresponding flags are then assigned to the ad, which allows for better visualization of the results. The flag system is comprised of the following:

F - non-malicious	S - malicious
O - outlet/retail store	E - search engine
R - related software	P - app download platform
C - course platforms	M - brandjacking
B - blogs/forums	X - manually reviewed
I - courier services	
U - unrelated	
L - legitimate	

An excerpt of the intermediate report returned by the labeler can be seen below. The advertisements initially recorded by the crawler are now categorized. Ad 4 was deemed legitimate as the advertiser ID is associated with the “McAfee” antivirus company. Ad 7 was deemed as non-malicious as `blitzhandel24` is a known store for online products.

```
avast - Ads found: 9
  Ad 1: 0/2 warnings

  Ad 3: 0/2 warnings

  Ad 4: 1/2 warnings FL
    Redirect chain: {'mcafeeinc.demdex.net', 'www.mcafee.com', ...}
    Ads transparency ID: AR07041635852870483969?origin=ata

  Ad 6: 0/2 warnings

  Ad 7: 2/2 warnings FO
    Landing URL: https://blitzhandel24.co.uk/avast/[...]
    Redirect chain: {'monitor.clickcease.com', 'blitzhandel24.co.uk', ...}
    Ads transparency ID: AR03282036780072697857?origin=ata
```

Backwards compatibility The crawler requires a significant amount of time to collect results. During the data collection process, a run needed on average one hour per country. This means that any modifications to the real-time detection would require an additional ten hours of running the crawler to gather new results. Hence, the data labeler is built as an independent tool, separated from the crawler. This makes it possible to label new results and re-label old results with minimal effort. It is especially useful in the scenario where one of the non-malicious domains becomes a “threat” and a new computation of the results is needed.

Domain review

As a final step in the data analysis phase, the domain review process aims to validate and improve the accuracy of the automated detection methods. Domain review may be needed either for frequently occurring domains or for uncategorized advertisements. Both processes work similarly, based on a “threat level”.

The effective top-level domain (eTLD) and, if applicable, the immediate next level (eTLD+1) are checked on three different platforms. These are `VirusTotal` [29], `URLVoid` [30], and `IPQS` [31]. The “threat level” of the domain increases by 1 for each platform that reports it as malicious. If the age of the domain (more accurately displayed on `VirusTotal`) is less than one year, the “threat level” of the domain is increased by 0,5. A final verification is done on `scammer.info`, a crowd-sourced scam alert forum [32]. If the domain or the displayed phone number is reported on the website, then the “threat level” also increases by 1. With 4 different sources of information, the chances of malicious ads going unnoticed are further reduced.

The “threat level” threshold chosen for review in this paper is 1. The reader has the freedom to modify these values as they see fit and compare the results. This paper will report on findings based only on the aforementioned numbers. If the final “threat level” of an advertisement surpasses the threshold, then it is labeled as *malicious*. Otherwise, it falls under the *non-malicious* category.

Chapter 5

Results

A continuous process of data collection happened during the creation of the crawler, from February 2023 until May 2023. The results of these pilot crawls were used to improve the tool and expand the list of known domains (presented in appendix A). We will analyze the results collected by the crawler throughout June 2023, across 64 runs.

5.1 Summary of results

country	runs	total ads	relevant	legitimate	non-mal	malicious	inc
us	8	1,676	1,325	739	451	111	24
gb	6	1,807	1,140	539	491	45	65
au	7	1,203	826	440	298	39	49
ca	7	1,852	1,269	680	428	124	37
nl	7	1,638	1,294	655	583	50	6
ro	7	1,685	1,139	494	606	31	8
se	7	1,644	1,270	603	609	42	16
za	5	765	558	288	227	17	26
br	5	1,129	819	383	391	29	16
th	5	568	427	237	132	25	33
TOTAL	64	13,967	10,067	5,058	4,216	513	280
	runs	total ads	relevant	legitimate	non-mal	malicious	inc

Table 5.1: Crawler results per country. Column “**non-mal**” stands for non-malicious ads, and column “**inc**” stands for inconclusive results.

In Table 5.1 the results recorded for each country can be seen. Across 3,200 queries (64 runs of 50 queries each), the crawler has seen a total of 13,967 advertisements. Out of these, 10,067 were deemed relevant by the ad filter which was described in chapter 4. The column labeled “**legitimate**” represents the number of legitimate ads. These are ads that did not raise any warnings during the crawling process, i.e., landed on the known domain and had no suspicious redirects. Column “**non-mal**” presents the number of non-malicious ads recorded. These are ads for domains that were tagged as non-malicious by the detection tools. The number of interesting cases, that our tool reported as malicious can be seen in the column “**malicious**”. The last column “**inc**” refers to the advertisements with inconclusive results, namely when the crawler ran into unforeseen issues in its process. This can be, for example, pop-ups requiring user input which the crawler was not developed to handle, due to low reproducibility rate.

The percentages of each country can be better visualized in Table 5.2. These results were computed based on the number of relevant ads collected by the crawler. Approximately half of the advertisements were deemed legitimate by the tool. Non-malicious ads made up $\approx 42\%$ of the relevant ads. Out of 10,067 recorded advertisements 5% were labelled as malicious. The inconclusive rate was below 3%.

country	legitimate%	non-mal%	malicious%	inc%
us	55,77	34,04	8,38	1,81
gb	47,28	43,07	3,95	5,70
au	53,27	36,08	4,72	5,93
ca	53,59	33,73	9,77	2,92
nl	50,62	45,05	3,86	0,46
ro	43,37	53,20	2,72	0,70
se	47,48	47,95	3,31	1,26
za	51,61	40,68	3,05	4,66
br	46,76	47,74	3,54	1,95
th	55,50	30,91	5,85	7,73
TOTAL	50,24	41,88	5,10	2,78
	legitimate%	non-mal%	malicious%	inc%

Table 5.2: Ratio of results per country

5.2 Malicious results

Table 5.3 presents the number of malicious domains per country, as well as the number of ads found for these domains. Across all runs, 49 unique malicious domains were identified. Out of these, 22 domains had cases of brandjacking with various severity, 20 were malicious search engines and 7 were platforms for app downloads. The full list of malicious domains alongside the evaluation results can be found in Table A.1 and Table A.2.

country	malicious domains	malicious ads	% out of relevant ads
us	24	111	8,38
gb	9	45	3,95
au	5	39	4,72
ca	24	124	9,77
nl	6	50	3,86
ro	2	31	2,72
se	8	42	3,31
za	7	17	3,05
br	8	29	3,54
th	5	25	5,85
<hr/>			
TOTAL	49	513	5,10
	malicious domains	malicious ads	% out of relevant ads

Table 5.3: Breakdown of malicious results per country

The top 5 search terms targeted by malicious ads were, in order, “mcafee”, “avast”, “microsoft”, “anydesk”, and “teamviewer”.

5.3 Non-malicious results

In Table 5.4 the categorization of advertisements labeled as non-malicious can be seen. Each entry corresponds to one of the categories described in chapter 4. Here it can be seen that 44% of the captured ads are from outlet/retail stores selling a variety of products. Ads about related software contribute for $\approx 24\%$. These ads refer to legitimate software that targets keywords found in the search term list. Around 10% of the advertisements (column “**unrel**”) were flagged as relevant due to containing the search term.

category	domains	ads	%
outlet	212	1,855	44,00
related	83	987	23,41
courses	10	222	5,27
blogs	26	320	7,59
courier	14	136	3,23
app platforms	8	88	2,09
search engines	21	194	4,60
unrelated	41	414	9,82

Table 5.4: Breakdown of non-malicious cases

However, due to multiple interpretations of a search term (“United Parcel Services” vs. “Uninterrupted Power Supply”) these proved to be unrelated. The remaining 22,77% of non-malicious cases consisted of various platforms offering courses, blogs/forums, courier services, app downloads, or search engine services.

5.4 Domain statistics

From the dataset, a list of all the visited domains was compiled along with the occurrences of each domain. These values were used to decide the cut-off value for reviewing domains, as mentioned in chapter 4. In total, the crawler accessed 854 unique domains over 4,729 ads (malicious and non-malicious). The top 11 domains were accessed more than 50 times.

# of visits	domains	ads
1	293	293
≤ 2	425	557
≤ 3	511	815
≤ 5	641	1,383
≤ 10	769	2,351

Table 5.5

Chapter 6

Discussion

In this chapter, the previously presented results will be further analyzed. Some of the malicious advertisements flagged by the designed tool will be presented, along with the supporting evidence. Limitations of the crawler will be discussed as well as possible improvements that can be made.

Geographic distribution The highest percentages of malicious ads were recorded in the United States and Canada. This can be attributed to the following factors. Firstly, these countries together have a large number of internet users and substantial online economic activity [33], making them lucrative targets for cybercriminals. Moreover, it is interesting to note that the US and Canada have more sophisticated advertising ecosystems [34], where ads can be highly targeted based on user behavior and demographics. This precision in targeting may increase the effectiveness of malicious ads, as they can be made to appear more legitimate to specific users or target a more vulnerable audience.

Target audience The prevalence of malicious ads targeting antivirus software (McAfee, Avast), Microsoft, and remote desktop software (AnyDesk, Teamviewer) may be motivated by the demographics of their user base. These software products are used more frequently by the elderly population, who are more susceptible to phishing scams and less familiar with the latest cybersecurity practices [35]. This makes them attractive targets for those looking to exploit a broad, vulnerable audience. As previously mentioned, brandjackers leverage well-known and trusted brands to increase the credibility of their phishing attempts. By mimicking these brands, they can deceive users into downloading malicious software or providing sensitive information, making said software prime targets for such attacks.

Domain age An interesting observation from the results is the influence of domain age on the likelihood of a domain being malicious. Most ma-

malicious domains identified had a domain age of less than one year, which is consistent with the tactics used by cybercriminals who frequently register new domains to avoid detection. However, there were exceptions where some malicious domains were older than one year old, and some even older than 5 years. These findings coincide with the 2014 study by Bilge et al. [16]. Older malicious domains may be indicative of more advanced attackers who can maintain control over a domain for extended periods without detection. It should also be noted that websites standing for more than 5 years were tagged as legitimate by VirusTotal, URLVoid, and IPQS. The only reports of maliciousness were from users on scammer.info. The persistence of such domains highlights the need for continuous monitoring and updating of detection methods to account for both new and long-standing threats.

Design choices In chapter 4 we mentioned finding the “greatest balance” between reviewing domains and reviewing advertisements. This balance decided the cut-off value for labeling domains into the existing databases. The results presented in the previous chapter will be used to motivate the choice for a cut-off point of 3 occurrences. The data labeler was created with the aim of reducing the need for review. The higher the frequency of a domain, the more advertisements are covered with only one domain check. On the other hand, a lower frequency of a domain means less ad coverage. The following cases were treated:

1. Review all (854) domains \Rightarrow automatically label all advertisements.
2. Review the domains visited at least 2 times (561) \Rightarrow label the remaining 293 ads.
3. Review the domains visited at least 3 times (429) \Rightarrow label the remaining 557 ads.
4. Review the domains visited at least 5 times (261) \Rightarrow label the remaining 1,143 ads.

Cases 1 and 2 would require an equal amount of 854 reviews. Case 3 would require 986 reviews (132 reviews over cases 1/2), while case 4 already surpasses the 1,000 mark (550 reviews over cases 1/2). Going for a higher cut-off value would mean an even larger amount of ads to review, with exponential growth. Between cases 1/2 and case 3, the latter was picked. Due to how information is structured during collection, review for advertisements has more data easily available and thus is faster. Reviewing a domain with accompanying screenshots, videos, and HAR files first requires finding an instance of that domain within the results and accessing the corresponding folder. For advertisements, this information is already available in the folder where the ad is flagged. The difference in required effort is what motivated the choice of setting the cut-off value to 3.

6.1 Malicious cases

A total of 49 domains out of the 854 visited domains were tagged as malicious due to high “threat levels”. The malicious domains were further split into three different categories, namely cases of brandjacking, search engines, and platforms for app downloads. Each of these will be presented in the following sections.

6.1.1 Brandjacking cases

Across the dataset of malicious domains, 22 were marked as cases of brandjacking. The largest part of this category (17 out of 22) consists of websites that follow a pattern of “tech support scams” [36]. This includes but is not limited to, claiming that the user’s computer is infected, using another brand’s logo to increase trustworthiness, and displaying phone numbers to call for those in need of customer support. These advertisements were mainly targeted at the search terms “mcafee”, “avast” and “microsoft”. The websites sold licenses for said software or advertised offering customer support.

Example 1 - windowstechies.com

This website offers a troubleshooting tool for “common PC issues”. It targets a vast amount of keywords, as can be seen in Table A.1. The targeted terms have individual advertisements (Figure 6.1) and lead to different versions of the website (Figure 6.2). The domain has been active for 12 years at the time of writing and the 3 detection tools (VT, URLVoid, IPQS) mark it as legitimate. The reports of maliciousness come from users on `scammer.info`.

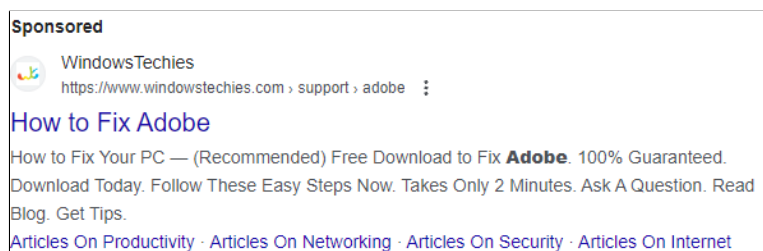


Figure 6.1: Example of `windowstechies.com` ad targeted to keyword “adobe”

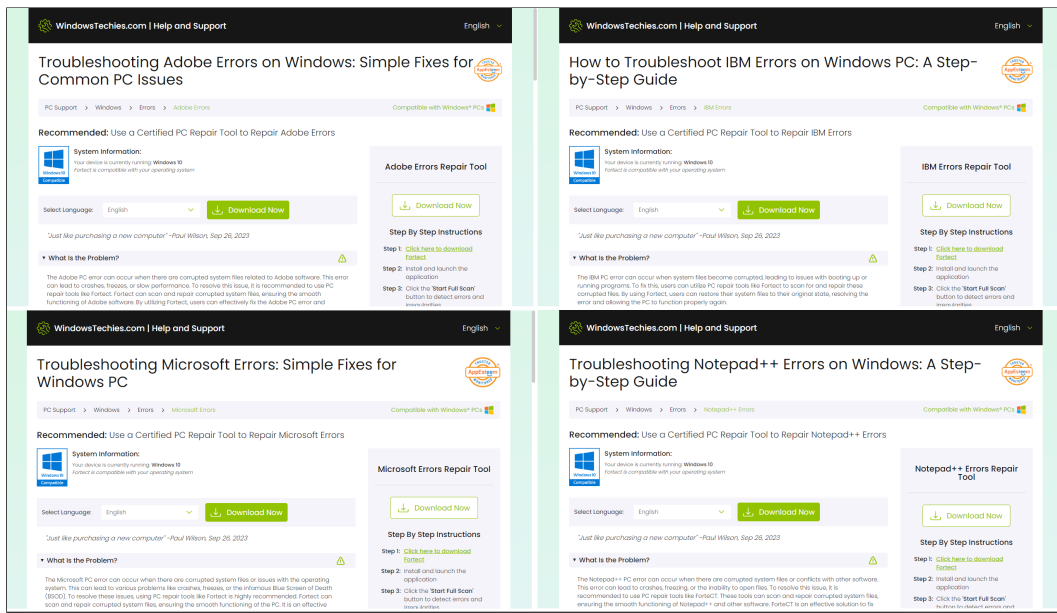


Figure 6.2: Different landing pages for windowstechies.com

Example 2 - deal.risecenter.shop

This website advertises licenses for the McAfee antivirus software. It follows patterns common to tech support scams as previously mentioned. It is flagged as malicious on VirusTotal and IPQS, as well as reported for scam calls on scammer.info. The usage of the McAfee brand with no (known) affiliation is a case of brandjacking. An extremely similar case is the website deal.websitecentral.shop, which has an identical layout. The only difference is the displayed phone number and the website's logo.

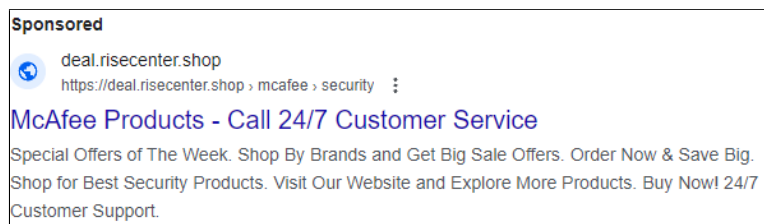


Figure 6.3: Advertisement of deal.risecenter.com for keyword "mcafee"

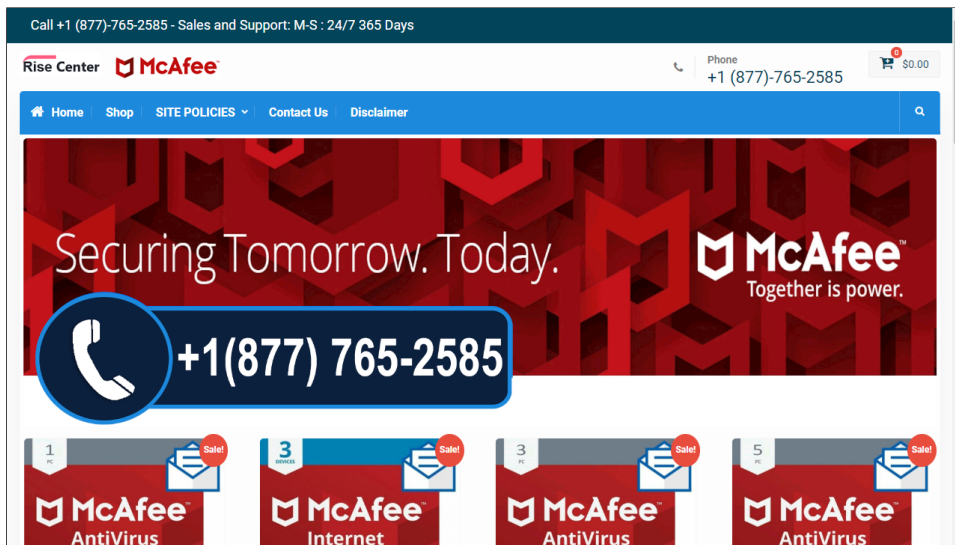


Figure 6.4: Landing page for deal.risecenter.com

Example 3 - aolsolution.info

This is a website that also falls under the tech support scam category, this time offering customer support. It is flagged as malicious by VirusTotal and also reported by users on scammer.info. It impersonates the antivirus software Avast.

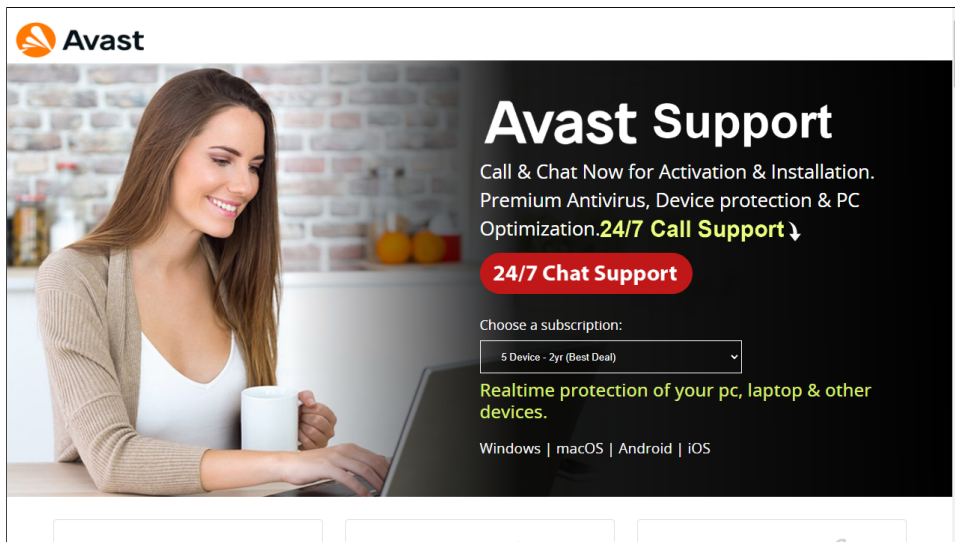


Figure 6.5: Landing page for aolsolution.info

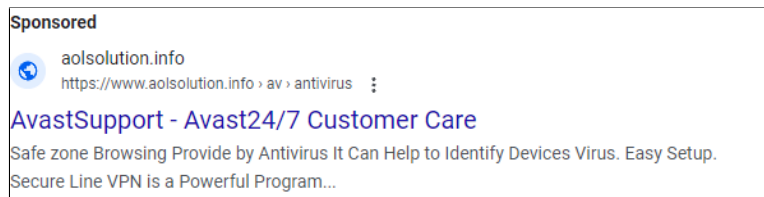


Figure 6.6: Advertisement of aolsolution.info for keyword “avast”

Example 4 - office-staples.org

This website offers “Microsoft Windows” licenses. It does, however, not follow the previously seen patterns. The landing page does not display a phone number to call and does not claim any direct affiliation with Microsoft. Nonetheless, it was flagged by VirusTotal and IPQS.

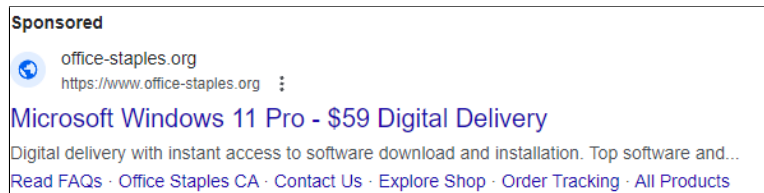


Figure 6.7: Advertisement of office-staples.org for keyword “microsoft”

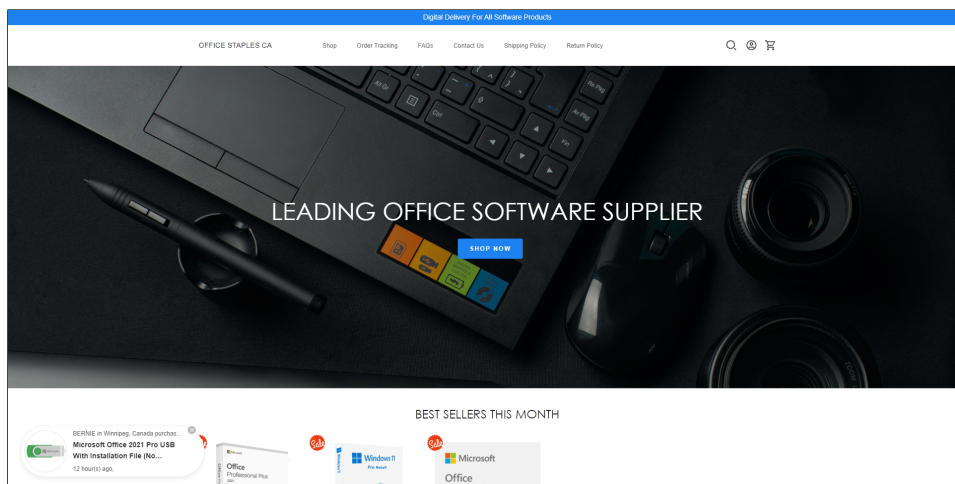


Figure 6.8: Landing page for office-staples.org

Example 5 - auth.onuberconnect.com

This website is a redirection from other advertised links, `eats-uder.online` and `couponcave.online`. These two websites can also be navigated to manually, but lead to a different landing page than the one captured by the crawler (Figure 6.10). Upon manually accessing the URL where the crawler landed, we were redirected to the main “Yahoo!” search page. This may be an indication of the advertisers using ad cloaking. The website `auth.onuberconnect.com` is flagged only by IPQS and is designed to collect login details of users, supposedly offering a coupon. Due to the inability to access the URL manually, we could not further look into the intent of the page for processing this data.

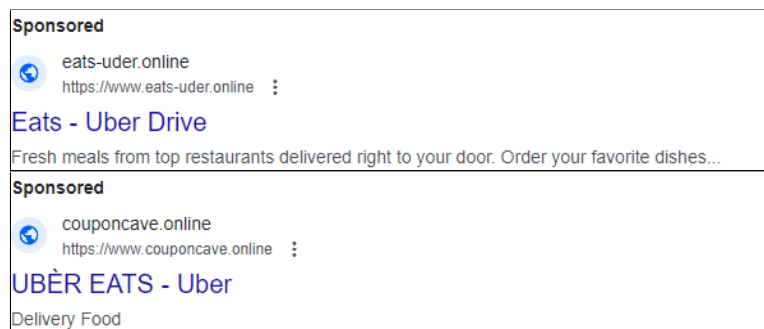


Figure 6.9: Two different ads leading to `auth.onuberconnect.com` for keyword “uber”

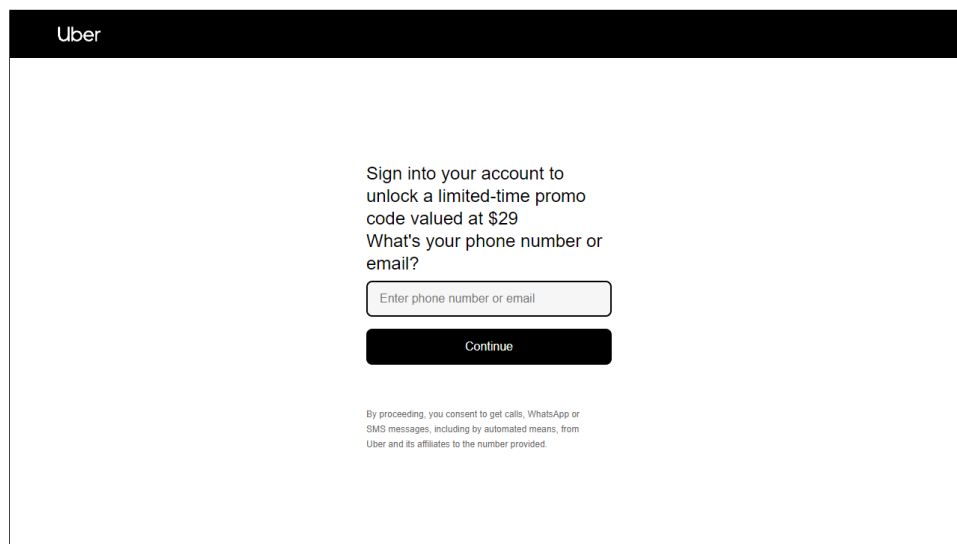


Figure 6.10: Landing page for `auth.onuberconnect.com`

6.1.2 Malicious search engines

During the crawling process, a total of 41 search engines were encountered in advertisements. Almost half of these were tagged as malicious by VirusTotal and/or URLVoid. The full details can be seen in Table A.2. Such malicious search engines may manipulate search results to promote phishing sites, ad fraud schemes, and other deceptive content. In Figure 6.11, the landing of a malicious search engine can be seen. The associated ad targeted the keyword “ebay”. It can be seen that the legitimate “ebay” page cannot be seen anywhere in the top four results.

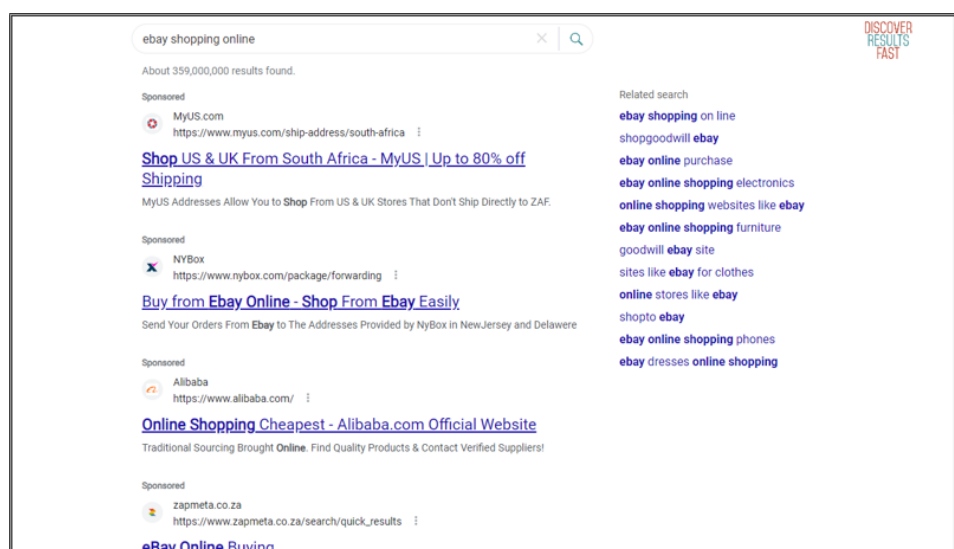


Figure 6.11: Landing page for discoverresultsfast.com, a malicious search engine

6.1.3 Malicious app download platforms

Platforms advertising software downloads were also encountered during crawling. Out of 15 domains, 7 of them were tagged as malicious by VirusTotal, URLVoid or IPQS. These platforms were reported for distributing adware. The details are available in Table A.2. In Figure 6.12 an example of such a platform can be seen. The corresponding ad was targeted at the keyword “capcut”, but advertisements for other terms such as “facebook” were also captured by the crawler.

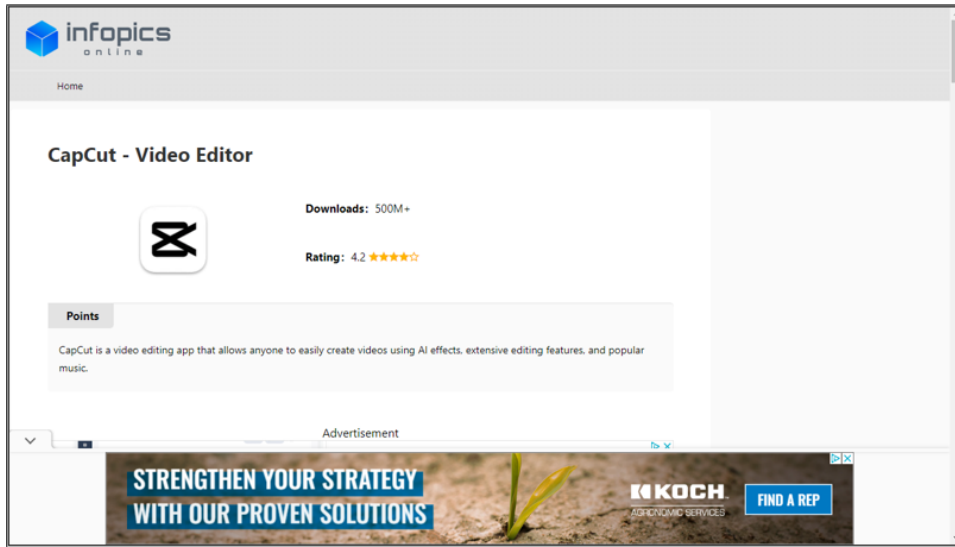


Figure 6.12: Landing page for `infopics.online`, a malicious app platform

6.2 Efficiency of detection methods

As presented throughout the paper, 4 platforms were used to assess whether a domain is malicious or not. Individual cases can be visualized in Table A.1 and Table A.2. A more compact version can be seen below in Table 6.1.

Platform	Total flagged	Brand-jacking	Search engines	App platforms
VirusTotal	37	11	20	6
URLVoid	23	4	15	4
IPQS	10	8	0	2
<code>scammer.info</code>	14	14	0	0

Table 6.1: Summarised results of the used detection tools

VirusTotal VirusTotal identified the highest number of total cases, with a strong detection rate on malicious search engine-related threats. This is likely due to its use of multiple antivirus engines and URL scanning tools, making it a robust option for broad detection purposes.

URLVoid URLVoid performed well in identifying search engine-related threats, but was less effective in detecting brand-jacking and app download platform threats. This suggests that URLVoid is particularly strong in analyzing URLs and leveraging multiple blocklists.

IPQS IPQS showed great results in detecting brand-jacking cases, but did not flag any search engine-related threats. This indicates that IPQS is more specialized in fraud prevention and IP reputation scoring, particularly for brand protection, but may lack URL scanning capabilities for broader contexts, unlike the previous two tools.

scammer.info `scammer.info` is entirely focused on brand-jacking, reflecting its crowd-based approach to reporting scams and fraudulent activities. The absence of detections in other categories confirms a narrow, but highly effective focus on brand-related threats. It was also effective in flagging domains older than one year that went undetected by the previous tools.

Each detection tool comes with its strengths and weaknesses. However, the inconsistency between tools seen in Table A.1 shows that none of them are highly capable of identifying brandjacking-based malvertising.

6.3 Limitations

One of the main limitations of this study is that it targeted advertisements from a single ad network, namely Google. While Google plays a large part in the online advertising ecosystem, focusing solely on its network means that the findings may not be representative of the broader landscape of online advertisements. This narrow scope overlooks possible malvertising cases present on other ad networks, such as Facebook or Bing. These platforms have different layouts for displaying pages and advertisements, which can lead to the crawler not functioning as intended.

Additionally, the list of search terms used in this study was limited to a set of 50 popular software and websites. Although these terms were chosen based on high relevance (Kantar BrandZ list) and large traffic (Tranco list), they do not encompass the full range of potential targets for malvertising. Consequently, many malicious campaigns that target less common or emerging search terms might have been missed. Thus, the presence of brandjacking-based malvertising could be higher than reported in this study.

Another limitation to be kept in mind is the previously mentioned technique of ad cloaking. Although there was an instance where the web crawler successfully passed cloaking and landed on a malicious page (Section 6.1.1, example 5), this does not indicate the crawler’s capability of always passing cloaking. It is likely that numerous other malicious ads employing cloaking evaded detection, resulting in an underestimation of the true presence of brandjacking-based malvertising.

Despite having a low need for review ($\approx 5\%$), this process can be further improved. For example, reviewing domains on `VirusTotal`, `URLVoid`, and `IPQS` can be automated with the usage of APIs. Such a method was presented in the paper of Masri and Aldwairi [12]. This would lead to a more streamlined and quicker process of review.

Similarly, domain categorization as presented in chapter 4 can also be enhanced. This may be done, for example, by implementing machine learning algorithms to automatically label domains [37]. Similarly, existing databases or classifiers may be found and used so that the amount of reviewed domains is lowered [38].

Although no extensive research has been done with regards to the validity of `scammer.info`, the platform has been used as a source of information in various other research papers [39, 40, 41, 42]. Due to its presence in numerous “tech support scam” related studies, we deem it a trustworthy database for our tool as well.

6.4 Research Ethics

Our crawler requires clicking on advertisements and accessing the associated landing pages. This means that advertisers may be subject to small costs due to our clicks, as they may have to pay publishers and ad networks for the ad services. In order to assess the impact of our research, we checked the number of occurrences for each visited domain. The most frequent domain in the dataset was visited 117 times. The average number of visits per domain was 5. According to 2024 statistics, the CPM (cost per 1,000 impressions) for Google Search Ads is around \$38,40 [43]. The highest incurred cost for the most visited domain would be \$4,45 and the average cost would be around \$0,19. Thus, we believe our crawler did not significantly impact advertisers and falls in line with other research and associated costs [21, 44].

6.5 Future work

Future research should focus on the previously mentioned limitations. Expanding the range of ad networks and search terms to provide a more comprehensive understanding of the scope and strategies of malvertising campaigns should be a priority. Our crawler is set up to use a fresh profile for each search term. This means that the ad network cannot create a behavioral profile and serve us more personalized ads. Future work can experiment with more persistent profiles and investigate how results change.

Chapter 7

Conclusions

This paper has explored the development and implementation of a web crawler using Playwright to detect brandjacking-based malvertising. Within the dataset of 10,067 ads, we successfully identified 49 malicious domains across 513 malicious advertisements. Out of these, 22 domains were cases of brandjacking. This shows that brandjacking is a present threat in the online advertising ecosystem and it should not be dismissed.

Additionally, the results of our research reveal gaps in Google's current system to verify the advertisements displayed on its search page. Despite Google's extensive resources, the presence of malicious ads indicates a possible failure in properly screening ad content and the legitimacy of advertisers. This poses a great risk to users, who rely on the perceived security and reliability of Google's advertising platform.

Our study also evaluated the effectiveness of four detection tools: `VirusTotal`, `URLVoid`, `IPQS`, and `scammer.info`. While each tool demonstrated strengths in certain scenarios, none proved to be an all-around solution for detecting all instances of brandjacking-based malvertising. The varying performance across different categories shows the need for a better approach to detecting brandjacking-based malvertising.

In summary, this paper highlights the need for improved detection methods and stricter verification processes to address the still present issue of brandjacking-based malvertising. The development of better, more focused tools that integrate the strengths of existing solutions is essential for ensuring the security of online advertising platforms and reducing the associated effects of brandjacking.

Bibliography

- [1] L. Zhou, Z. Kehuan, X. Yinglian, Y. Fang, and W. XiaoFeng, “Knowing your enemy: understanding and detecting malicious web advertising,” in *Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12*, p. 674–686, Association for Computing Machinery, 2012.
- [2] IAB, “IAB/PwC Internet Advertising Revenue Report 2024,” 2024. Available at: <https://www.iab.com/insights/internet-advertising-revenue-report-2024/>.
- [3] M. Milam, “The rise of brandjacking against major brands,” *Computer Fraud & Security*, vol. 2008, no. 10, pp. 10–13, 2008.
- [4] J. Segura, “Malvertising via brand impersonation is back again,” 2023. <https://www.malwarebytes.com/blog/threat-intelligence/2023/05/malvertising-its-a-jungle-out-there>.
- [5] A. Savčín, “Avast researchers detect a September surge in malvertising,” 2023. <https://blog.avast.com/avast-threat-report-q3-2023-malvertising>.
- [6] B. Krebs, “Using Google Search to Find Software Can Be Risky,” 2024. <https://krebsonsecurity.com/2024/01/using-google-search-to-find-software-can-be-risky/>.
- [7] M. Tober, “Zero-Clicks Study,” 2022. <https://www.semrush.com/blog/zero-clicks-study/>.
- [8] E. Nowroozi, Abhishek, M. Mohammadi, and M. Conti, “An Adversarial Attack Analysis on Malicious Advertisement URL Detection Framework,” *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1332–1344, 2023.
- [9] N. Cveticanin, “Phishing Statistics & How to Avoid Taking the Bait,” 2023. Available at: <https://dataprot.net/statistics/phishing-statistics/>.

- [10] A. Zarras, A. Kapravelos, G. Stringhini, T. Holz, C. Kruegel, and G. Vigna, “The Dark Alleys of Madison Avenue: Understanding Malicious Advertisements,” in *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC ’14, p. 373–380, Association for Computing Machinery, 2014.
- [11] Z. Moti, A. Senol, H. Bostani, F. Zuiderveen Borgesius, V. Moonsamy, A. Mathur, and G. Acar, “Targeted and Troublesome: Tracking and Advertising on Children’s Websites,” in *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 121–121, IEEE Computer Society, 2024.
- [12] R. Masri and M. Aldwairi, “Automated malicious advertisement detection using VirusTotal, URLVoid, and TrendMicro,” in *2017 8th International Conference on Information and Communication Systems (ICICS)*, pp. 336–341, 2017.
- [13] D. Lewandowski, “Users’ understanding of search engine advertisements,” *Journal of Information Science Theory and Practice*, vol. 5, no. 4, pp. 6–25, 2017.
- [14] S. Yuan, J. Wang, and X. Zhao, “Real-time bidding for online advertising: measurement and analysis,” in *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD ’13, 2013.
- [15] D. Y. Wang, S. Savage, and G. M. Voelker, “Cloak and dagger: dynamics of web search cloaking,” in *Proceedings of the 18th ACM conference on Computer and communications security*, pp. 477–490, 2011.
- [16] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel, “Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains,” *ACM Transactions on Information and System Security*, vol. 16, no. 4, 2014.
- [17] K. C. Wilbur and Y. Zhu, “Click fraud,” *Marketing Science*, vol. 28, no. 2, pp. 293–308, 2009.
- [18] K. Springborn and P. Barford, “Impression Fraud in On-line Advertising via Pay-Per-View Networks,” in *22nd USENIX Security Symposium (USENIX Security 13)*, pp. 211–226, 2013.
- [19] F. Nettersheim, S. Arlt, and M. Rademacher, “Dismantling Common Internet Services for Ad-Malware Detection,” 2024.
- [20] “Selenium,” 2024. <https://www.selenium.dev/>.

- [21] K. Subramani, X. Yuan, O. Setayeshfar, P. Vadrevu, K. Lee, and R. Perdisci, “When Push Comes to Ads: Measuring the Rise of (Malicious) Push Advertising,” in *Proceedings of the ACM Internet Measurement Conference, IMC ’20*, p. 724–737, Association for Computing Machinery, 2020.
- [22] “Fast and reliable end-to-end testing for modern web apps,” 2024. <https://playwright.dev/>.
- [23] “Testing Frameworks for JavaScript,” 2024. <https://www.cypress.io/>.
- [24] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, “Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation,” in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, 2019.
- [25] Kantar, “Kantar BrandZ Most Valuable Global Brands 2023,” 2024. Available at: <https://www.kantar.com/campaigns/brandz/global>.
- [26] Y. Hsiao, M. Yang, Y. Lo, C. Feng, W. Hsu, and R. Chen, “A Study on Competitive Trend of Global Top 100 Brands,” *Journal of Economics, Business and Management*, vol. 10, no. 4, 2022.
- [27] “Mullvad VPN - Free the Internet,” 2024. <https://mullvad.net/>.
- [28] “Videos | Playwright,” 2024. <https://playwright.dev/python/docs/videos>.
- [29] “VirusTotal - Home,” 2024. <https://www.virustotal.com/gui/home/url>.
- [30] “Check if a Website is Malicious/Scam or Safe/Legit | URLVoid,” 2024. <https://www.urlvoid.com/>.
- [31] “Malicious URL Scanner,” 2024. <https://www.ipqualityscore.com/threat-feeds/malicious-url-scanner/>.
- [32] “Scammer Info - Scambait Forums and Scam Number Database,” 2024. <https://scammer.info/>.
- [33] K. S., “Digital 2024: Global Overview Report,” 2024. <https://datareportal.com/reports/digital-2024-global-overview-report>.
- [34] D. S. Evans, “The Online Advertising Industry: Economics, Evolution, and Privacy,” *Journal of Economic Perspectives*, vol. 23, no. 3, p. 37–60, 2009.
- [35] M. Jiang, H. Tsai, S. Cotten, N. Rifon, R. Larose, and S. Alhabash, “Generational Differences in Online Safety Perceptions, Knowledge, and Practices,” *Educational Gerontology*, vol. 42, 2016.

- [36] N. Miramirkhani, O. Starov, and N. Nikiforakis, “Dial One for Scam: A Large-Scale Analysis of Technical Support Scams,” in *Proceedings 2017 Network and Distributed System Security Symposium*, 2017.
- [37] M. Cova, C. Kruegel, and G. Vigna, “Detection and analysis of drive-by-download attacks and malicious JavaScript code,” in *Proceedings of the 19th international conference on World wide web*, p. 281–290, 2010.
- [38] X. Song, Y. Zhu, Z. Xuemei, and X. Chen, “Hierarchical contaminated web page classification based on meta tag denoising disposal,” *Security and Communication Networks*, vol. 2021, pp. 1–11, 11 2021.
- [39] M. Berney, J. Ondrus, and A. Holzer, “Navigating the Shadows of Cyber Vigilantism: A Preliminary Analysis of Social Dynamics and Activities of Scambaiting,” in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA ’24, 2024.
- [40] J. Liu, P. Pun, P. Vadrevu, and R. Perdisci, “Understanding, Measuring, and Detecting Modern Technical Support Scams,” in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 18–38, 2023.
- [41] Y.-C. Chen, J.-L. Chen, and Y.-W. Ma, “AI@TSS- Intelligent technical support scam detection system,” *Journal of Information Security and Applications*, vol. 61, p. 102921, 2021.
- [42] E. Tsai, A. Singhal, and A. Prakash, “Terms of Deception: Exposing Obscured Financial Obligations in Online Agreements with Deep Learning,” 2024.
- [43] “What is Average CPM Google Ads & How to Reduce It,” 2024. <https://nestscale.com/blog/cpm-google-ads.html>.
- [44] P. Vadrevu and R. Perdisci, “What you see is not what you get: Discovering and tracking social engineering attack campaigns,” in *Proceedings of the Internet Measurement Conference*, pp. 308–321, 2019.

Appendix A

Appendix

The list of terms and associated known domains can be found below. In some places, information was omitted for better readability. This was marked by [...]. In the tool itself, the space was filled with the possible top-level domains of the platforms. For example .co.uk for websites in the United Kingdom, .ro for websites in Romania, .ca for websites in Canada, etc.

```
dictionary = {
    "7zip"      : "7-zip\.org",
    "winrar"    : "win-rar\.com",
    "amazon"    : "amazon(\.com| [...])",
    "apple"     : "apple\.com",
    "samsung"   : "samsung\.com",
    "microsoft" : "microsoft\.com",
    "bol"       : "bol\.com",
    "ebay"      : "ebay(\.com| [...])",
    "ibm"       : "ibm\.com",
    "ikea"      : "ikea\.com",
    "facebook"  : "facebook\.com",
    "nike"      : "nike(\.com|\.com\.br)",
    "adidas"    : "adidas(\.com| [...])",
    "oracle"    : "oracle\.com",
    "VLC"       : "videolan\.org",
    "paypal"    : "paypal\.com",
    "tiktok"    : "tiktok\.com",
    "temu"      : "temu\.com",
    "taobao"    : "taobao\.com",
    "aliexpress": "(aliexpress(\.com|\.us))|(aliexpressromania\.com)",
    "shein"     : "(shein(\.com|\.co\.uk|\.se))|(shoturlcl\.com)",
    "zalando"   : "zalando(-lounge|)(\.com| [...])",
    "vinted"    : "vinted(\.com| [...])",
    "pinterest" : "pinterest\.com",
```

```
"adobe"      : "adobe\.com",
"capcut"     : "capcut\.com",
"audacity"   : "audacityteam\.org",
"blender3d" : "blender\.org",
"virtualbox": "virtualbox\.org",
"OBS"        : "obsproject\.com",
"notepad++"  : "notepad-plus-plus\.org",
"ups"        : "ups\.com",
"dhl"        : "dhl(express|ecommerce|parcel|)(\.com| [...])",
"disney"     : "disney(plus|store|)(\.com| [...])",
"netflix"    : "netflix\.com",
"hbo"        : "(hbomax\.com)|(hbogo\.co\.th)",
"dropbox"    : "dropbox\.com",
"trivago"    : "trivago(\.com [...])",
"slack"      : "slack\.com",
"cisco"      : "cisco\.com",
"imdb"       : "imdb\.com",
"shopify"    : "shopify\.com",
"discord"    : "discord\.com",
"booking"    : "booking\.com",
"mcafee"     : "mcafee\.com",
"avast"      : "avast\.com",
"eset"       : "eset(\.com|\.ro)",
"teamviewer" : "teamviewer\.com",
"anydesk"    : "anydesk\.com",
"uber"       : "uber(eats|carshare|)\.com"
}
```

Verification results of websites flagged as malicious:


Domain	VT	URLVoid	IPQS	scammer.info	term(s)
windowstechies.com				✓	audacity, notepad++, adobe, ups, dropbox, oracle, teamviewer, microsoft, virtualbox, ibm, slack
deal.websitecentral.shop	✓			✓	mcafee
deal.risecenter.shop	✓		✓	✓	mcafee
newlanecart.xyz			✓	✓	mcafee
al-bazaar.xyz	✓			✓	avast, mcafee
office-staples.org	✓		✓		microsoft
fortect.com	✓				microsoft
apexaibricks.com	✓	✓	✓	✓	mcafee
securemypcsoftware.com		✓	✓	✓	mcafee
auth.onuberconnect.com			✓		uber
sbcomexp.com.br	✓				dhl, ups
orac-server.com			✓		oracle
naspeo.com				✓	mcafee
softcartllc.com	✓			✓	mcafee
aolsolution.info	✓			✓	avast
247techies.tech				✓	microsoft
247techiesau.tech				✓	microsoft
insoftassist.com				✓	mcafee
peodeal.com			✓		mcafee
tomatomovies.com	✓	✓			hbo
envioparaexterior.com.br	✓	✓			dhl
raycerlx.com				✓	mcafee

Table A.1: Results for brandjacking malvertisements

Domain	VT	URLVoid	IPQS
SEARCH ENGINES			
searchresultsdelivery.com	✓	✓	
informationvine.com	✓		
searchinfotoday.com	✓	✓	
discoverresultsfast.com	✓	✓	
findresultsnow.com	✓		
greatselections.co	✓	✓	
encontrerapidinho.com	✓		
newsearchtoday.co	✓	✓	
smartshopsearch.com	✓		
search.nation.online	✓	✓	
readytodistribute.com	✓	✓	
findbestresults.co	✓	✓	
frequentsearches.com	✓	✓	
allshoppinghub.com	✓	✓	
discovertoday.co	✓	✓	
answerroot.com	✓	✓	
allinfosearch.com	✓		
allshoppinghub.com	✓	✓	
givemeanswers.net	✓	✓	
quicklyseek.com	✓	✓	
APP PLATFORMS			
pcapp.store	✓	✓	
softonic.com	✓		
softonic.com.br	✓		
infopics.online	✓		✓
rocketdrivers.com	✓	✓	
tunefab.com		✓	
appreview.cc	✓	✓	

Table A.2: Results for malicious search engines and app platforms



Sponsored



 deal.websitecentral.shop
https://deal.websitecentral.shop › mcafee › security

McAfee Total Protection | Call 24/7 Customer Service

Special Offers of The Week. Shop By Brands and Get Big Sale Offers. Order Now & Save Big.

Figure A.1: Ad example

 **About this advertiser** 



 Advertiser identity verified by Google 

Advertiser
Nabh Joshi

Location
India

[See more ads](#) this advertiser has shown using Google

Figure A.2: Information about advertiser

 **Why you're seeing this ad** 

The following information was used to show you this ad.

- Your current search terms
- Google's estimation of your approximate current location
- Google's estimation of your areas of interest, based on your activity
- Google's estimation of relevant locations

[Learn how personalized ads work](#)

Figure A.3: Search engine criteria

Google Ads Transparency Center

Any time ▾ Shown anywhere ▾ All formats ▾

🔍 All topics ▾ Search by advertiser or website name

Nabh Joshi

Legal name: Nabh Joshi

Based in: India

👤 Advertiser has verified their identity

1 ad

Sponsored

deal.websitecentral.shop
deal.websitecentral.shop/mcafee/security

McAfee Customer Service - Call 24/7 Customer Service

Special Offers of The Week. Shop By Brands and Get
Big Sale Offers. Order Now & Save Big. Shop for...

Nabh Joshi

Figure A.4: Excerpt of “See more ads” page