

Measuring and Modeling the Free Content Web*

Abdulrahman Alabduljabbar¹, Runyu Ma², Ahmed Abusnaina³, Rhongho Jang⁴, Songqing Chen², DaeHun Nyang⁵, and David Mohaisen¹

¹University of Central Florida, ²George Mason University, ³Meta Inc, ⁴Wayne State University, ⁵Ewha Womans University

Abstract

Free content websites that provide free books, music, games, movies, etc., have existed on the Internet for many years. While it is a common belief that such websites might be different from premium websites providing the same content types, an analysis that supports this belief is lacking in the literature. In particular, it is unclear if those websites are as safe as their premium counterparts. In this paper, we set out to investigate, by analysis and quantification, the similarities and differences between free content and premium websites, including their risk profiles. To conduct this analysis, we assembled a list of 834 free content websites offering books, games, movies, music, and software, and 728 premium websites offering content of the same type. We then contribute domain-, content-, and risk-level analysis, examining and contrasting the websites' domain names, creation times, SSL certificates, HTTP requests, page size, average load time, and content type. For risk analysis, we consider and examine the maliciousness of these websites at the website- and component-level. Among other interesting findings, we show that free content websites tend to be vastly distributed across the TLDs and exhibit more dynamics with an upward trend for newly registered domains. Moreover, the free content websites are 4.5 times

*This research was supported by Global Research Laboratory (GRL) Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2016K1A1A2912757). A. Alabduljabbar is supported in part by the Saudi Arabian Cultural Mission (SACM). R. Ma and S. Chen are partly supported by a Commonwealth Cyber Initiative grant, and NSF grant CNS-2007153. (*corresponding authors: David Mohaisen; mohaisen@ucf.edu.*) [†]Equal Contributors

more likely to utilize an expired certificate, 19 times more likely to be malicious at the website level, and 2.64 times more likely to be malicious at the component level. Encouraged by the clear differences between the two types of websites, we explore the automation and generalization of the risk modeling of the free content risky websites, showing that a simple machine learning-based technique can produce 86.81% accuracy in identifying them.

Keywords: Free Content Websites; Web Security; Web Mining

1. Introduction

Websites are categorized into two broad categories based on their monetization options: free content and premium websites. While the free content websites provide content free of charge and are typically sustained by proceeds of advertisements and user donations [1, 2, 3, 4], the premium websites offer services through fees, e.g., subscriptions or pay-as-you-use models [5]. Premium websites ensure a very high level of quality of service as a result of well-designed websites that are well-maintained through dedicated engineering and operational efforts [6]. In contrast, free content websites are believed to lack such a high expectation for the quality of service and are often user-driven [7].

The lax expectations for functional and security qualities, user-driven content, and the extensive utilization of third-party advertisements on free content platforms introduce various risks [8, 9, 10, 11]. For example, advertisements on these websites can be exploited for data and information leakage or even the distribution and execution of malicious scripts on the user device [12, 13]. Moreover, the lack of strict maintenance operation rules in free content websites allows for various risks: web frameworks used in free content websites are rarely updated, allowing for the exploitation of old unpatched vulnerabilities and exposing their users to various levels of risk [14].

However, untested hypotheses and widely unverified beliefs aside, are free content websites different from premium websites delivering the same type of content? Do free content websites differ in their structure, content, and security

properties from premium websites? Do these websites come with a hidden cost to users, outweighing the perceived benefits, i.e., being free? To answer these questions, we proceed with a systematic analysis of a carefully assembled dataset that curates 834 free content websites and 728 premium websites. Our study combines both domain- and content-level analysis, coupled with security analysis across various dimensions. For the domain-level analysis, we examine the domain name system features, creation time, and SSL (Secure Sockets Layer) features as measures of intent. For the content-level analysis, we examine the HTTP (Hypertext Transfer Protocol) request, page size, loading time, and content type, all measuring website complexity. For security analysis, we examine both the website- and component-level detection and vulnerability using two major off-the-shelf tools, *VirusTotal API* [15] and *Sucuri API* [16].

Our analysis concludes that there are significant, fundamental, and intrinsic differences between free content and premium websites delivering the same type of content. Among other interesting findings, we report that free content websites are exclusively vastly distributed across TLDs (Top-level Domains), although using common SLDs (Second-level Domains). Moreover, they frequently change their domains, are likely to evade blacklisting, and are more often associated with invalid SSL certificates. Content-wise, free content websites tend to require significantly fewer HTTP requests for smaller requested page sizes, although at a penalty of significant load time due to extensively employing redirection with more script objects. Risk-wise, we found that free content websites are 19 and 2.64 times more likely to be malicious than premium websites at the page level (38% vs. 2%) and file level (45% vs. 17%), respectively.

We leverage our insights from those analyses to generalize and extrapolate by modeling free content websites' risk. To this end, we defined risk using pure performance metrics. Moreover, we were able to group the risky websites with very high accuracy (more than 86%).

Contributions and Findings. This paper delivers in-depth comparative analyses of the free and premium websites of the same content types across various

dimensions: domains, content, and security. Enabled by a feature-rich analysis, we build a machine learning-based approach to score the risk of free content websites with high accuracy. In the following, we elaborate on our contributions.

1. **Free Content Websites Curation (§3).** We assembled a list of more than 1,500 free content and premium websites offering the same type of content. The websites are obtained from the top search results of Google, DuckDuckGo, and Bing search engines. The websites are then crawled to obtain their content, including scripts, images, HTML, CSS, etc.
2. **Domain-level Analysis (§4.1).** To examine the domain-level features of free content websites, we analyze three aspects: their TLD (Top-level Domain), SSL certificates, and creation date. As a result, we found a significant increase in the number of free content websites, in contrast to a decrease in newly created premium websites. Moreover, we observe more frequent domain name dynamics in free content websites than in premium websites. Almost one-third of the free content websites operated using an invalid or unmatched SSL certificate.
3. **Content-level Analysis (§4.2).** To examine the content-level features, we analyze three aspects: the HTTP requests, page size, and average load time. Among other findings, we observe that the premium websites contain significantly more images, and their average size is three times the size of free content websites. Interestingly, however, we found the load time appears comparable due to various intrinsic design choices, including the utilization of scripts and redirection to deliver advertisements, which are more prevalent in free content websites.
4. **Free Content Websites Risk Analysis (§5.1).** We leverage two popular off-the-shelf tools, *VirusTotal* and *Sucuri*, to assess the security risks associated with free content websites. Our analysis shows that free content websites are significantly more likely to be associated with maliciousness

than premium websites. However, the discovery of premium websites detected as malicious is quite interesting and calls for further exploration.

5. **Risk Modeling (§5.2).** Both the performance and security metrics analysis highlight significant differences between free content and premium websites. Moreover, their risk profiles are vastly different from one another. Motivated by the differences in their features, we build a simple machine learning algorithm that utilizes easy-to-obtain domain- and content-level features to predict the risk of a website. We report a promising accuracy of 86.81% for modeling the risk of free content websites.

2. Related Work

The work most related to our contribution in this paper falls under two broad branches: website analysis and malicious web content analysis. In the following, we provide an overview of some of the efforts in both directions.

2.1. Websites Analysis

Websites are continuously evolving in content and usage, paralleled by an increase in the complexity and richness of their components. However, with such an evolution, various security risks emerge due to the interplay between such components [17, 18, 19, 20]. One of the vastly unexplored security aspects in the literature has been the validity of websites’ certificate [21]. To address this issue, Chung *et al.* [21] proposed an in-depth analysis of certificates in the web PKI (Public Key Infrastructure), showing that the vast majority of certificates in the web PKI are invalid. Their study also investigated the source of the invalid certificates, concluding that they were generated mainly by end-user devices, with periodic regeneration of new self-signed certificates.

Libert *et al.* [22] evaluated the privacy-compromising practices employed by a million popular websites, e.g., data leakage. They concluded that roughly nine out of ten websites shared user data with third-party services without user consent. Using a similar dataset, Lavrenovs *et al.* [23] conducted a comprehensive

assessment of the security of Alexa top-million websites, showing that 29.1% of HTTPS requests have incorrect TLS (Transport Layer Security) configurations, and the HTTP Strict Transport Security (HSTS) policy is implemented in only 17.5% of the websites. These findings are alarming and demonstrate the worrisome state of the security policies followed by such popular websites.

Exploring environments to evaluate the security flaw in web applications, Alsmadi *et al.* [24] designed a component-based testing mechanism for various invalid inputs and used this mechanism to investigate websites' behavior, including security, due to such inputs. Since the invalid input is a consistent part of the attack surface, the security of the online services and web applications is strengthened by eliminating those inputs (i.e., rejecting invalid inputs). To do so, they proposed several methods for detecting invalid inputs, uncovering many SQL injection vulnerabilities.

2.2. Malicious Web Content

Recent studies have shown that adversaries are capable of embedding malicious codes within *JavaScript*, *GIF*, or *Redirection* components of the websites [25, 26, 27, 28, 29]. The security (and safety) of end-users depend significantly on detecting and preventing such malicious content, which has also been studied. To do so, researchers have leveraged various features of web applications, including URL (Uniform Resource Locator) domain components, webpage content, HTTP headers, and loaded scripts, and used them to detect malicious web applications [30, 31]. It has also been shown that a promising feature set is the HTTP header information [22], where McGahagan *et al.* [32] leveraged 672 of those features to build a system for malicious website detection. To examine the feasibility of using components and content (i.e., files and scripts) as features for detection, the authors conducted a comprehensive evaluation of different webpage content features. These 17 engineered new features can improve malicious websites' detection performance.

One crucial yet unexplored aspect of websites is the interplay between advertisements deployed on them and their associated maliciousness. Li *et al.* [33]

investigated various malicious online advertising and marketing methods, e.g., malware propagation [34, 35, 36, 37, 38], click fraud, etc. Their study used a large-scale dataset of ads-related web traces, showing malicious advertisement practices in hundreds of high-ranked websites. To examine the effectiveness of malicious advertisement detection, Masri *et al.* [39] evaluated three tools, *VirusTotal*, *URLVoid*, and *TrendMicro*, showing *URLVoid* to provide the best performance.

Another prominent threat that has been explored is the distribution of malicious content on free download portals [40]. Such portals can be maliciously utilized for distributing harmful software to end-user devices. Rivera *et al.* [41] conducted a systematic analysis of PUP (Potentially Unwanted Programs) and malware obtained using free download portals, showing that, on average, 8% to 26% of the downloaded content are either PUP or malicious.

Machine learning algorithms have also been widely used for effectively detecting malicious websites [42]. However, they are impaired by two key challenges, feature selection and evasion. To address the feature selection problem, Singh and Goyal argued for coupling the feature selection with overhead performance and accuracy in their analysis [43]. Detection evasion, the other issue, is often associated with intrinsic features, including the usage of redirection and hidden iFrames. In this domain, Liu and Lee [44] proposed an effective Convolutional Neural Network-based malicious content detection based on a screenshot of a webpage.

This Work. In this work, we explore and assess the maliciousness of free content websites in contrast with premium websites. Our findings show worrisome increasing trends in the portion of malicious content within free content websites. To proactively address this concern, we model the risks associated with these websites through easy-to-obtain performance features and identify up to 86.81% of the risky websites verified against ground truth.

Table 1: An overview of the collected dataset. The collected URLs are associated with five different categories and belong to free content and premium websites. Overall, 1,562 websites were crawled for the purpose of this study.

Category	Free Content Websites			Premium Websites		
	URLs	Files	Avg. Files	URLs	Files	Avg. Files
Books	154	7,073	45.93	195	17,840	91.49
Games	80	6,439	80.49	113	11,314	100.12
Movies	331	9,821	29.67	152	10,738	70.64
Music	83	6,059	73.00	86	7,225	84.01
Software	186	11,561	62.16	182	18,742	102.98
Overall	834	40,953	49.10	728	65,859	90.47

3. Dataset Overview

In the following, we highlight the approach we followed in creating our dataset, including initial selection and associated criteria, manual annotation, crawling, and augmentation.

Websites Selection. We compiled a list of 1,562 free content and premium websites for conducting our measurements. In selecting the websites, various constraints for representation. In particular, the following criteria are utilized in selecting our websites:

(1) Popularity: Each website has to be among the most popular websites on the web. Given that those websites may not necessarily in the most popular websites, we use search engines’ results as a proxy for estimating their popularity. A website is considered popular if it appears in the top results by at least one of the used search engines: Google, DuckDuckGo, and Bing.

(2) Balanced Representation: In composing our overall dataset, we ensure that our dataset is balanced per category. To that end, we expand our queries until we achieve close-to-balanced representation across categories for both the free and premium websites.

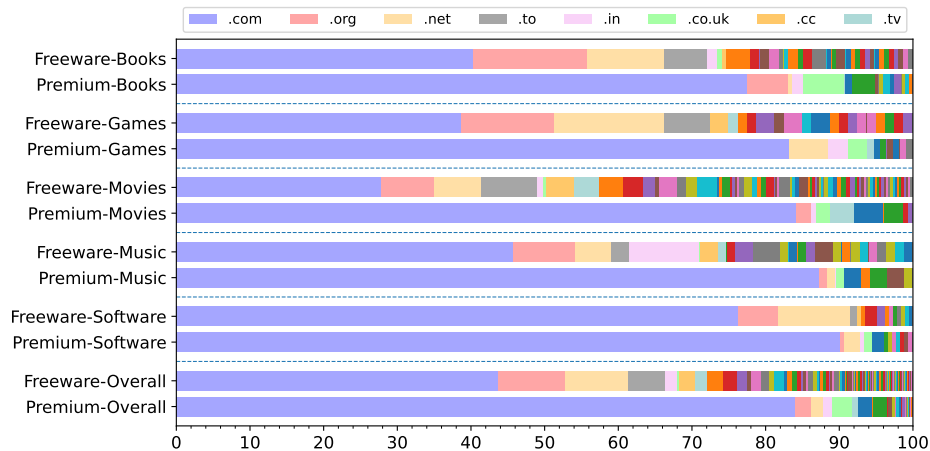


Figure 1: TLD distribution of free vs. premium websites. The free content websites are more distributed among the TLDs, in contrast to the premium websites.

Upon initially selecting the unique websites for inclusion in our dataset, we proceed by manually examining and labeling each of them as either premium or free content websites. Each of the websites is then categorized, also manually, into one of five groups based on the type of content the website mainly provides: books, games, movies, music, or software.

Websites Crawling. To understand the risks associated with free content websites, we crawled each website’s content (i.e., files) using PyWebCopy [45], a python package for cloning websites and downloading their associated files. The obtained files are then used for the risk analysis and modeling, as they reflect the behavior of the provided services. Our dataset is then augmented with various attributes categorized into two broad groups, the domain-level attributes (TLD, domain creation information, SSL certificate information) and content-level attributes (HTTP request information, page size, load time, and content type).

High-level Characteristics. Table 1 shows the distribution of the collected dataset. Notice that the average files crawled from premium websites are significantly larger than the average files for free content websites.

4. Websites Analyses

In order to understand the fundamental differences between free content and premium websites, we conduct two types of analyses: domain-level analysis and content-level analysis. Domains are the gateways to websites, and they are rich in information that can be utilized to understand their intent. Supplementing the domain-level features with content-level features improves the visibility into the websites intent. In the following, we provide our analysis results based on both of those features groups.

4.1. Domain-level Analyses

The domain-level analysis provides us with a high-level and interesting view and understanding of the website as an infrastructure across the owner information, creation date, and the used TLD. We pursue such an analysis to contrast the associated domains of free and premium websites.

Top-level Domains Analysis. The TLD is the highest level domains in the hierarchical domain name system, followed by the SLD (Second-level Domain); in `example.com`, `example` is the SLD, and `com` is the TLD. Recently, the number of TLDs has grown significantly with the introduction of the new generic TLDs (gTLDs), although `.com`, `.net`, `.org`, and `.edu` remain the most prominent [46]. In this work, we investigate the distribution of free content and premium websites among the TLDs, shown in Figure 1.

We found that `.com` is the most prominent TLD domain, with 44% and 84% of free content and premium websites using `.com`, respectively. However, interesting, we found that the total number of unique TLDs used by the premium websites in our dataset to be only 24, while this number is 98 domains in the free content websites. We note that this widespread distribution could be triggered by the mechanisms employed for malicious website blocking implemented by major browsers and systems. For instance, Chrome and Firefox rely on user reports when using safe browsing service [47] to collect and block malicious websites. To evade blocking, free content websites change their domain

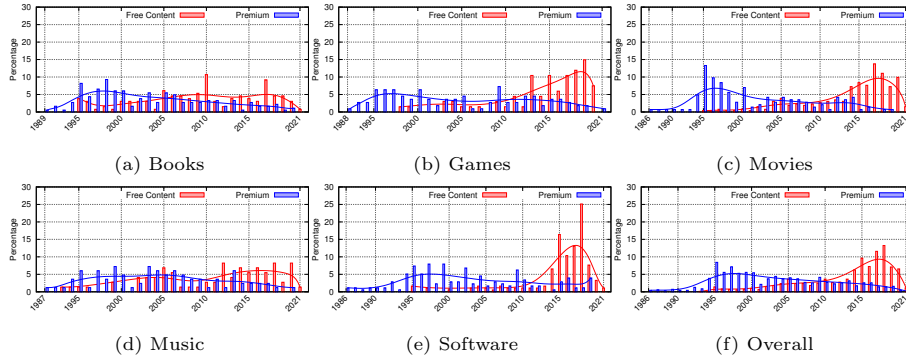


Figure 2: The domain creation year comparison between free and premium website. By comparing the trend across the various content types, we observe the significant upwards trend of free content domain creation compared to premium websites.

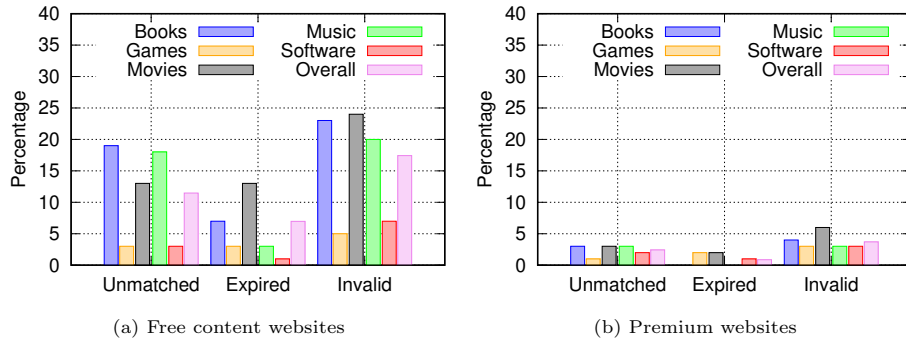


Figure 3: The SSL certificate analysis results. We observe that almost 36% of the free content websites have problematic SSL certificates compared to 7% in premium websites.

name periodically. However, free content operators maintain the same SLD and migrate their websites to other TLDs to retain the existing users and some of them change their TLD to evade blocking.

Domain Name Creation. We examine the website creation dates, where we observe an increasing trend in the number of newly created free content websites, in contrast to the declining number of newly created premium websites, as shown in Figure 2. This growing trend, particularly in the period of 2015–2021, motivates us to examine and understand the risks associated with using online free content websites. To further support that, we found from the TLDs analysis

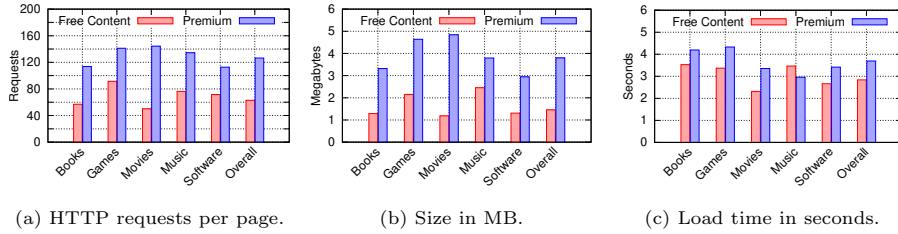


Figure 4: Page-related comparison between the free content and premium websites (average statistics). Despite having different page sizes, the free content and premium websites average comparable page load times, indicating other reasons than size that affect time.

that free content websites tend to change their domain name periodically to avoid content blocking or blacklisting.

SSL Certificate Analysis. HTTP transfers website content, e.g., HTML, from the web server to the user browser. However, this protocol is not secure, and the transferred data can be exposed to unauthorized access. Therefore, most websites have moved to use the secure version of HTTP (HTTPS), which implements an encryption mechanism to protect the transferred content. Our analysis found that 36% of the free content websites have invalid HTTPS compared to only 7% of the premium websites. Moreover, we found that 26% of free content websites still allow HTTP (insecure) access, whereas 0% of the premium websites allow HTTP access. SSL certificate is a digital authentication method that authenticates the identity of a website and provides HTTPS with an encrypted connection between a server and a client machine. The SSL certificate is a critical component of a website to secure user data and protect them against, e.g., phishing.

In this work, we investigate the validity of the SSL certificate for both free content and premium websites. In particular, we study three aspects: (i) unmatched hostname in the certificate, (ii) expired certificate, and (iii) invalid/-fabricated certificate. Figure 3 shows that, in total, 36% of the free content websites have issues with their certificates (i.e., 11.5% unmatched name, 7% expired, and 17.5% invalid certificate), compared to a total of only 7% of the

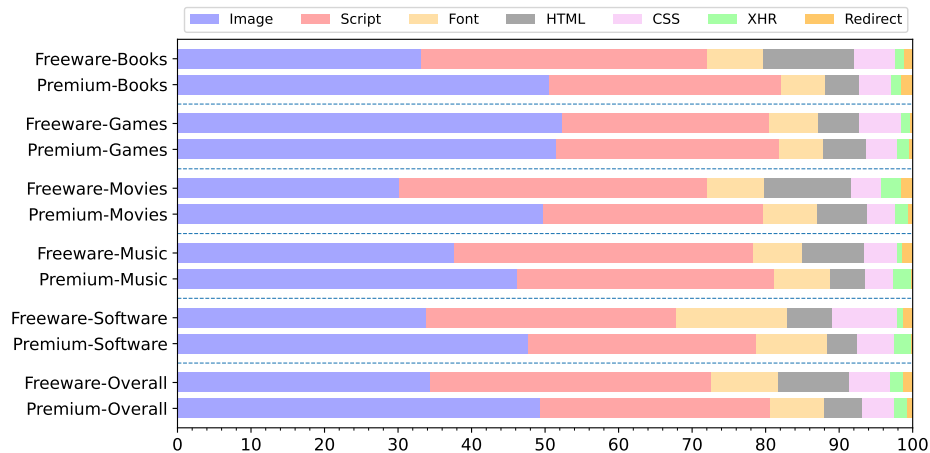


Figure 5: Content-type comparison between the free content and the premium websites. We observe some differences in the website file types, notably in the *images* type.

premium websites with problems in their associated SSL certificates. This is more noticeable in the “*Movies*”, “*Books*”, and “*Music*” categories. As shown, free content websites are more likely to have issues with their SSL certificate. This may be attributed to the fact that free content operators are not renewing the SSL certificate, unwilling to increase their operational cost. Nonetheless, this practice leads to potential risks regarding user information and data privacy.

Takeaways: Through domain-level analyses, we found that (i) the free content websites are newer, and their growth has been increasing significantly in recent years, whereas the premium websites’ growth is decreasing, with fewer websites introduced every year, (ii) the free content websites are more distributed across the TLDs as they change their domain to avoid malicious website blocking mechanisms, (iii) the free content websites are more likely to have invalid or expired SSL certificate. These findings complement our analysis concerning the safety of using free content websites and the risks associated with them.

4.2. Content-level Analyses

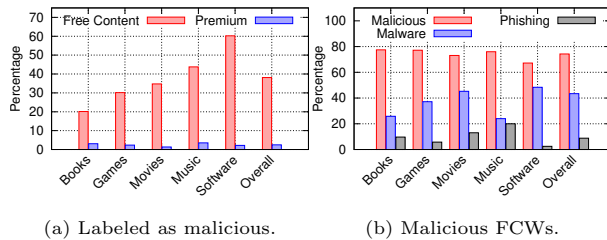
To gain insight into the content-level features of free content and premium websites, we analyze the extracted files in both types of websites. In this analysis, we focus on four features: the number of HTTP requests, page size, page load time, and content type.

HTTP Requests. HTTP requests are made by clients to request access to resources on servers (e.g., HTML files, CSS, images), and their numbers per page are an indication of the complexity of the requested page. Figure 4a shows the average number of HTTP requests made for free content and premium websites. We observe that a client would initiate almost twice the number of requests to access a premium website compared to accessing a free content website. This is quite anticipated, given that the premium website pages are larger in size. However, we observe that the average page size in premium websites is 3x the free content websites, whereas the number of HTTP requests is only 2x more, indicating that visiting a free content page requires more HTTP requests for the same amount of data. That could be a result of redirection, where each redirection triggers one or more independent HTTP requests and consumes more time for loading, as shown later.

Average page size. According to the page weight report by HTTP Archive [48], the average page size of the top one million websites is around 2.07 MB. How far is the size of the average page that belongs to either category? To answer this question, we examine the average page size of the free content and premium websites, with the results reported in Figure 4b. We observe that the free content websites follow the normal distribution of the page sizes reported by the HTTP Archive [48], while the premium websites have an average homepage size of 3.9MB, three times the average size of a free content page. A potential explanation might be that the free content websites rely on redirecting users to other websites' content or advertisement websites, as we demonstrate later, instead of including and presenting content in the free content page body.

Average page load time. We define page load time as the time it takes the page to be loaded fully and measured to understand additional aspects of the website’s complexity. Figure 4c shows the average page load time, calculated using the SolarWinds Pingdom API (Application Programming Interface) [49], for both the free content and premium websites. While the average size of the premium websites is three times the average free content page size, we notice that the average load time is comparable across them, indicating aspects beyond the size that affect the load time, i.e., degraded performance and extensive usage of redirection.

Content type. The page size does not seem to fully explain the complexity and loading time of websites, which calls for a deeper analysis of the content of the website. The content type is another statistical feature of the website’s content at the component level (i.e., files). These components include *Image* (*GIF*, *PNG*, *JPEG*), *JavaScript*, *Text*, *HTML*, *CSS*, *XHR*, and *Redirection*. We found that *Image* is the most common component, followed by *JavaScript*, whereas the *Redirection* content is the least common among these components. Figure 5 shows the average distribution (%) of the different components in the free content and premium websites. Overall, premium websites have 15% more images than free content websites. However, we notice the extensive usage of *Redirection* in the free content websites, as it is often a method to deliver advertisements and mislead the filtering algorithm. We found that the (rounded) ratio of the redirection in free content compared to premium pages to be 6 (software), 7 (music), 3 (movie), 1 (games), 1 (books). Overall, free content websites redirect twice as much as premium sites, and have twice the HTML, 1.5 times the CSS, and 1.23 times the JavaScript.



(a) Labeled as malicious.

(b) Malicious FCWs.

Figure 6: The potential maliciousness of free content and premium websites.

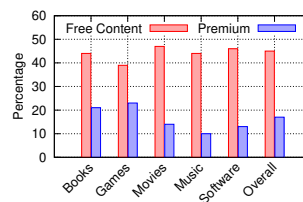


Figure 7: The malicious files detected by *VirusTotal*: free content vs. premium.

Takeaways: Our content-level analyses shed light on the main differences between free content and premium websites. We found the following. (i) The premium websites have almost twice the number of requests as free content websites and three times the average size of free content websites pages. (ii) Nevertheless, the average homepage load time is comparable for free content and premium websites. (iii) Content type-wise, the free content websites have a higher portion of *redirection* components, as they are a primary method to deliver advertisements.

5. Maliciousness Analyses

The analysis we conducted so far considered the performance and non-security characteristics of free content and premium websites, which highlight clear differences that contribute to both direct and indirect costs. One of the most important and obvious metrics to measure the cost of free content websites is by understanding their security and associated risk. In this section, we conduct this analysis, focusing on indicators of threat, such as maliciousness of URLs, files, and associated vulnerabilities (§5.1). Towards automating the discovery of such risks, we also report the results of a machine learning-based tool that shows the risk boundaries of websites based on features obtained from the risk analysis (§5.2).

5.1. Risk Assessment

The study of the maliciousness and vulnerabilities of both services websites, by shedding examining how they potentially affect users experience, safety, and security, is important. Motivated by that, we define the risk of a website using several metrics, namely: (i) containing malware, (ii) running malicious scripts, (iii) exploiting user device’s resources, or (iv) containing vulnerabilities, outdated software versions, or unpatched frameworks.

To assess the risk of each type of website without reinventing the wheel, we leverage two public APIs: *VirusTotal* [15] and *Sucuri* [16] for harmful behavior analysis. *VirusTotal* is an online service that aggregates the scanning results of more than 70 scanning engines and can be used for scanning files and URLs alike. On the other hand, *Sucuri* is a service that tests websites against several known malware, viruses, blacklisting lists, vulnerabilities, outdated frameworks, and malicious code.

Malicious URLs Detection and Annotation. Using *VirusTotal* API, we extracted malicious activities associated with the website URL, shown in Figure 6. We notice that there is a noticeable discrepancy between free content and premium websites in terms of maliciousness. In particular, Figure 6a shows that 38% of the free content websites are considered malicious by *VirusTotal*, compared to only 2% of the premium websites. A significant number of those detected websites ($\approx 74\%$) were labeled as malicious (Figure 6b), a website created to promote scams, attacks, and frauds. We also notice that a significant portion of the free content URLs is detected as malicious, ranging from 20% (“*Books*” websites) to 60% (“*Software*” websites). In contrast, premium websites have a very low detection rate, ranging from 1% to 4% only.

Malicious File Detection and Formats Analysis. In order to understand the behavior of a given service (i.e., content providers), it is essential to analyze the behavioral characteristics of the executable scripts hosted by the service. These scripts are forwarded to the end-user as files, including images, JavaScript codes, HTML, among other formats, and are often rendered or executed on

Table 2: The distribution of malicious files for different file formats in free content and premium websites. We observe that a large portion of “.gif” files are labeled as malicious in both cases, although almost twice as much (percentage) in free content.

	Category	.gif	.html	.png	.js	.php	.woff	.jpg	.eot	.woff2	.svg	.ttf	.log	.css
Free Content	Books	28%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Games	7%	13%	0%	1%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	Movies	40%	6%	1%	2%	0%	1%	0%	1%	0%	1%	1%	1%	0%
	Music	26%	6%	0%	1%	4%	3%	0%	3%	3%	0%	2%	0%	0%
	Software	11%	30%	4%	1%	1%	4%	0%	2%	4%	2%	3%	1%	0%
	Overall	26%	11%	2%	1%	2%	2%	0%	2%	4%	1%	2%	1%	0%
Premium	Books	19%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Games	21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Movies	9%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Music	21%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Software	5%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Overall	15%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

the user’s device. Analyzing the scripts and website files is critical, as recent studies [50, 51, 52, 53] have shown that such content can be exploited, leading to information and data leakage, in addition to abusing the resources of the end-user device. In order to understand the risks of free content websites, we leverage *VirusTotal* API for malicious file identification. In contrast, Figure 7 shows the percentage of malicious files detected by the *VirusTotal* API in the free content and premium websites. While the number of URLs that pertain to the premium websites and are labeled as malicious is only 2%, the number of their files labeled as malicious was 17%.

We notice that the trend persists overall, although magnified: 45% of the free content websites had files that have been labeled as malicious (compared to 17% in premium). To better understand this observation, we investigate the distribution of the format of the malicious files, where the comparative results are shown in Table 2.

Based on Table 2, we report that the majority of malicious files have ‘.gif’ and ‘.html’ formats. This is a result of either (i) the ‘.gif’ and ‘.html’ files

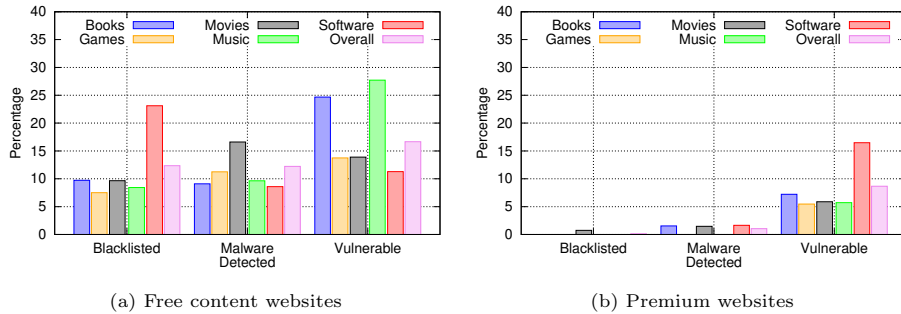


Figure 8: Assessing the maliciousness of the free content and premium websites. We show the percentage of the websites labeled as blacklisted, malware, and vulnerable.

containing malicious embedded scripts, or (ii) the VirusTotal engines considering the ‘.gif’ files as malicious content in general (i.e., potential false positives). It is worth noting that we manually inspected the ‘.gif’ files, and found that the majority of the malicious-labeled ‘.gif’ files are advertisement-related content.

Websites Vulnerability and Blacklisting. In order to analyze the potential exploitable vulnerabilities and blacklisting, we leveraged Sucuri API [16] to obtain information of domains activities for both types of services. As a result, we found that 12% of the free content websites were detected as *containing malware*, compared to only 1% of their premium counterparts, as shown in Figure 8. Moreover, we found the free “*Movie*” websites have the highest percentage of malware detection (16.67%), as shown in Figure 8a.

We also scanned the websites for vulnerabilities and found that the free “*Books*” and “*Music*” websites have the highest vulnerabilities overall. Despite the low reporting rate in the premium websites, 17% of “*Software*” were labeled as vulnerable, a higher portion than in free content websites (12%), which is quite surprising. According to *Sucuri* reports, a high percentage of the legitimate “*Software*” websites vulnerabilities are due to outdated framework versions, which is common in “*Software*” services websites.

In terms of blacklisting, Figure 8a shows that 12% of the free content websites were blacklisted by the *Sucuri* scanning engines, including Google, McAfee,

Yandex, Norton, ESET, and AVAST engines. We observe that the “*Software*” free content websites have a significantly higher percentage of blacklisted URLs (23.12%) compared to other categories, which all had at most 12% blacklisting rate. One reason for this behavior is the fact that these websites are changing their domain names frequently using a different TLD.

Takeaways: To assess the risks associated with free content websites, we leveraged *VirusTotal* and *Sucuri* APIs for analyzing the maliciousness of domain and files of both service types. Our analyses show worrisome trends among free content websites, including (i) free content websites are more likely to be associated with maliciousness at a domain-level (38% of the free content websites), and (ii) they are more likely to be associated with maliciousness at the file-level (45% of them). These trends are not limited to maliciousness, which led to high blacklisting, but include exploitable vulnerabilities that can expose visitors to leakage attacks. Our analysis also unveils that 17% of the free content websites are vulnerable.

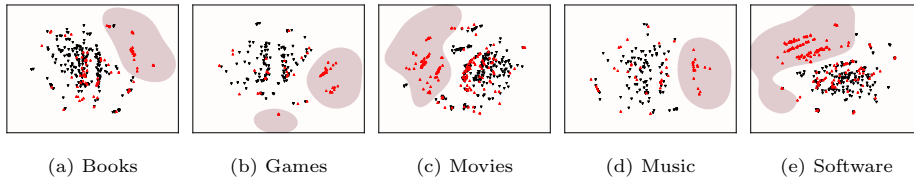


Figure 9: The decision boundary of the risk-free and risky websites. A risky website is a website with potential malicious intention. Notice that this malicious behaviour can be characterized (i.e., determined) using a support vector machine.

5.2. Risk Modeling

The insights that we have provided thus far are intriguing, although we are left with a key question: how much of these insights can be generalized across sites of the same population and type for risk assessment? To answer this question, we report on our effort to identify *risky* websites using simple machine learning algorithms.

Table 3: The description of the website’s characterization features. The features are extracted from three sources, (i) The website’s content, (ii) The website’s public information, (iii) The website’s SSL certificate information. We include the characteristics extracted from VirusTotal and Sucuri APIs for risk characterization and potential detection (D). Org: the origin of the feature, *[c]*: categorical feature, *[b]*: boolean feature (T/F), *[n]*: numerical feature, *[p]*: percentage feature.

#	§	Type	Description	#	§	Type	Description
1	§4.1	[c]	TLD name used by the website	14	§4.2	[p]	% of Redirect content in the website
2	§4.1	[b]	Domain not matched	15	§5.1	[b]	Domain detected by VirusToal API
3	§4.1	[b]	Expired SSL certificate	16	§5.1	[b]	Website is detected as malicious
4	§4.1	[b]	SSL certificate cannot be verified	17	§5.1	[b]	Website containing malware
5	§4.2	[n]	Average number of HTTP requests	18	§5.1	[b]	Website detected as phishing
6	§4.2	[n]	Average content size	19	§5.1	[b]	Files detected as malicious
7	§4.2	[n]	Page load time	20	§5.1	[b]	URL detected as malicious
8	§4.2	[p]	% of images in the website	21	§5.1	[b]	Blacklisted by scanning engines
9	§4.2	[p]	% of script files in the website	22	§5.1	[b]	Vulnerability in the website
10	§4.2	[p]	% of Fonts content in the website	23	§5.2	[n]	Website’s IP address lifetime
11	§4.2	[p]	% of HTML files in the website	24	§5.2	[b]	Using/used Cloudflare as a CDN
12	§4.2	[p]	% of CSS files in the website	25	§5.2	[b]	Using/used Akamai as a CDN
13	§4.2	[p]	% of XHR content in the website				

Risky Websites. A website in our analysis is considered risky if it is associated with any of the following:

- (1) *Malicious Domain.* Websites that are associated with URLs responsible for malicious activities are considered risky [54].
- (2) *Malicious Files.* Upon visiting a website, multiple scripts are executed on the host. As such, we consider any website with malicious files [55], regardless of its *VirusTotal* label, as a risky website.
- (3) *Blacklisted URLs.* Blacklisting can occur due to (i) massive user reporting, or (ii) previous maliciousness by the website (e.g., scam attacks). As such, we consider all blacklisted websites as risky [54].
- (4) *Vulnerable Websites.* Websites that are identified as vulnerable by *Sucuri* are considered risky, for the potential exploitability.

We note that the risk modeling is not limited to free content websites. We also consider any free content and premium website with one or more of the

aforementioned aspects as a risky website and otherwise a risk-free website.

Website Features. To model the risks associated with each service, and as common in the relevant literature [34, 54, 56, 57], we leverage the aforementioned extracted features as a representation. In particular, Table 3 shows the superset of potential features that we use to represent each online service, including SSL certificate, page size, load time, TLD, and website content features. Additionally, we include three more features extracted using *SecurityTrails* [58]: (i) the lifetime of a service IP address, (ii) whether a website is using or previously used Cloudflare as a Content Delivery Network (CDN), and (iii) whether a website is using or previously used Akamai Tech as a CDN.

Hold-out Data. The data obtained by *VirusTotal* and *Sucuri* (#15–#25) in Table 3 is held out, and is only used to model validation. This allows us to utilize easy-to-obtain website quality metrics that do not require access to third-party information to model the website risk. We envision that our lightweight modeling, in contrast to third-party risk data, would be more practical, since the third-party labels are determined based on reporting and expensive analyses accumulated over a period of time. Solely relying on third-party tools, such as *VirusTotal* to identify risks would exclude a significant number of websites, including those newly created for free content.

Risk Boundaries. Considering the aforementioned features, we visualize the boundaries between *risky* and *risk-free* websites, shown in Figure 9. In particular, we use the t-distributed stochastic neighbor embedding (t-SNE) visualization technique [59] to plot the features of the websites. Then, using a support vector machine, we estimate the risk boundaries, shown in the red-shaded area in Figure 9. Based on the validation, we find that the riskiest websites are clustered together, as they share different website features. Our modeling is capable of identifying risky websites with an accuracy of 86.81%, despite some limitations (e.g., potential false positives among our sampled websites).

Takeaways: We address the need for lightweight risk modeling of free content websites using a representation of 17 generic and file-related features. Our modeling is shown to be effective, producing an accuracy of 86.81%.

6. Conclusion and Future Work

Free content websites are an interesting element of the makeup of the web today, and their characteristics are not rigorously analyzed nor understood in contrast to other websites that offer the same content. This paper provides the first look into a comparative analysis of such websites across various domain- and content-level dimensions, as well as their risk profiles. Our curated datasets offer valuable resources for exploring this uncharted space, and our findings shed light on the fundamental differences between free content websites in contrast to premium websites.

We believe that our analysis in this paper only “scratches the surface” of this important problem and calls for further explorations and actions. For instance, our domain- and content-level analyses have been only focused on easy-to-obtain metadata features and did not consider the in-depth features, e.g., linguistic, network topology information, regional information, deep content type, and organization attributes (e.g., in the case of SSL certificates; signing authorities, and hosting infrastructure). All of these dimensions could shed more light on the characteristics of such websites and constitute our future work. Finally, we notice that our analysis utilizes a single snapshot of those websites, and we did not consider the temporal dimension of their characteristics, which would be a very interesting yet challenging aspect to explore.

References

- [1] F. Hecker, Setting up shop: The business of open-source software, *IEEE Softw.* 16 (1) (1999) 45–51.

- [2] K. Greenhill, C. Wiebrands, No library required: the free and easy backwaters of online content sharing, VALA 2012: eM-powering eFutures.
- [3] M. Carvajal, J. A. García-Avilés, J. L. González, Crowdfunding and non-profit media: The emergence of new models for public interest journalism, *Journalism practice* 6 (5-6) (2012) 638–647.
- [4] R. Snijder, The profits of free books: an experiment to measure the impact of open access publishing, *Learn. Publ.* 23 (4) (2010) 293–301.
- [5] J. R. Mayer, J. C. Mitchell, Third-party web tracking: Policy and technology, in: *IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA, IEEE Computer Society, 2012*, pp. 413–427.
URL <https://doi.org/10.1109/SP.2012.47>
- [6] C.-J. Liang, H.-J. Chen, A study of the impacts of website quality on customer relationship performance, *Total Quality Management* 20 (9) (2009) 971–988.
- [7] T.-C. Lin, J. S.-C. Hsu, H.-C. Chen, Customer willingness to pay for online music: The role of free mentality., *Journal of Electronic Commerce Research* 14 (4).
- [8] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, J. Forné, Online advertising: Analysis of privacy threats and protection approaches, *Comput. Commun.* 100 (2017) 32–51.
URL <https://doi.org/10.1016/j.comcom.2016.12.016>
- [9] A. Alabduljabbar, R. Ma, S. Alshamrani, R. Jang, S. Chen, D. Mohaisen, Poster: Measuring and assessing the risks of free content websites, in: *Network and Distributed System Security Symposium, (NDSS), 2022*.
- [10] A. Alabduljabbar, R. Ma, S. Choi, R. Jang, S. Chen, D. Mohaisen, Understanding the security of free content websites by analyzing their SSL

certificates: A comparative study, in: The 1st Workshop on Cybersecurity and Social Sciences, CySSS, 2022, pp. 19–25.

- [11] A. Alabduljabbar, D. Mohaisen, Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation, in: Companion of The Web Conference, ACM, 2022, pp. 860–867.
- [12] Z. Li, K. Zhang, Y. Xie, F. Yu, X. Wang, Knowing your enemy: understanding and detecting malicious web advertising, in: Proceedings of the 2012 ACM conference on Computer and communications security, 2012, pp. 674–686.
- [13] W. D. Groef, D. Devriese, N. Nikiforakis, F. Piessens, Flowfox: a web browser with flexible and precise information flow control, in: the ACM Conference on Computer and Communications Security, CCS, ACM, 2012, pp. 748–759.
URL <https://doi.org/10.1145/2382196.2382275>
- [14] A. Nappa, R. Johnson, L. Bilge, J. Caballero, T. Dumitras, The attack of the clones: A study of the impact of shared code on vulnerability patching, in: IEEE Symposium on Security and Privacy, SP, 2015, pp. 692–708.
URL <https://doi.org/10.1109/SP.2015.48>
- [15] VirusTotal, Analyze suspicious files and URLs to detect types of malware, automatically (May 2022).
URL <https://www.virustotal.com/>
- [16] Sucuri, website security check & malware scanner (May 2022).
URL <https://sucuri.net/>
- [17] D. Perdices, J. E. L. de Vergara, I. González, L. de Pedro, Web browsing privacy in the deep learning era: Beyond vpns and encryption, *Comput. Networks* 220 (2023) 109471. doi:10.1016/j.comnet.2022.109471.
URL <https://doi.org/10.1016/j.comnet.2022.109471>

- [18] H. Zou, J. Su, Z. Wei, S. Chen, B. Zhao, An efficient cross-domain few-shot website fingerprinting attack with brownian distance covariance, *Comput. Networks* 219 (2022) 109461. doi:10.1016/j.comnet.2022.109461. URL <https://doi.org/10.1016/j.comnet.2022.109461>
- [19] G. Sun, Z. Zhang, Y. Cheng, T. Chai, Adaptive segmented webpage text based malicious website detection, *Comput. Networks* 216 (2022) 109236. doi:10.1016/j.comnet.2022.109236. URL <https://doi.org/10.1016/j.comnet.2022.109236>
- [20] A. Faroughi, A. Morichetta, L. Vassio, F. V. D. de Figueiredo, M. Mellia, R. Javidan, Towards website domain name classification using graph based semi-supervised learning, *Comput. Networks* 188 (2021) 107865. doi:10.1016/j.comnet.2021.107865. URL <https://doi.org/10.1016/j.comnet.2021.107865>
- [21] T. Chung, Y. Liu, D. R. Choffnes, D. Levin, B. M. Maggs, A. Mislove, C. Wilson, Measuring and Applying Invalid SSL Certificates: The Silent Majority, in: *ACM Internet Measurement Conference, IMC, 2016*, pp. 527–541.
- [22] T. Libert, Exposing the hidden web: An analysis of third-party HTTP requests on 1 million websites, *CoRR* abs/1511.00619.
- [23] A. Lavrenovs, F. J. R. Melon, HTTP security headers analysis of top one million websites, in: *International Conference on Cyber Conflict, CyCon, 2018*, pp. 345–370.
- [24] I. Alsmadi, F. Mira, Website security analysis: variation of detection methods and decisions, in: *2018 21st Saudi Computer Society National Computer Conference (NCC), 2018*, pp. 1–5.
- [25] G. Tan, P. Zhang, Q. Liu, X. Liu, C. Zhu, F. Dou, Adaptive malicious URL detection: Learning in the presence of concept drifts, in: *17th IEEE International Conference On Trust, Security And Privacy In Computing And*

Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018, IEEE, 2018, pp. 737–743.

- [26] R. Masri, M. Aldwairi, Automated malicious advertisement detection using virustotal, urlvoid, and trendmicro, in: International Conference on Information and Communication Systems, ICICS, 2017, pp. 336–341.
- [27] K. Patil, Isolating malicious content scripts of browser extensions, *Int. J. Inf. Priv. Secur. Integr.* 3 (1) (2017) 18–37.
- [28] V. R. L. Shen, C. Wei, T. T. Juang, Javascript malware detection using A high-level fuzzy petri net, in: 2018 International Conference on Machine Learning and Cybernetics, ICMLC 2018, Chengdu, China, July 15-18, 2018, IEEE, 2018, pp. 511–514.
- [29] R. Wang, Y. Zhu, J. Tan, B. Zhou, Detection of malicious web pages based on hybrid analysis, *J. Inf. Secur. Appl.* 35 (2017) 68–74.
- [30] C. Johnson, B. Khadka, R. B. Basnet, T. Doleck, Towards detecting and classifying malicious urls using deep learning, *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.* 11 (4) (2020) 31–48.
- [31] A. Desai, J. Jatakia, R. Naik, N. Raul, Malicious web content detection using machine leaning, in: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE, 2017, pp. 1432–1436.
- [32] J. M. IV, D. Bhansali, M. Gratian, M. Cukier, A comprehensive evaluation of HTTP header features for detecting malicious websites, in: European Dependable Computing Conference, EDCC, 2019, pp. 75–82.
- [33] Z. Li, K. Zhang, Y. Xie, F. Yu, X. Wang, Knowing your enemy: understanding and detecting malicious web advertising, in: ACM Conference on Computer and Communications Security, CCS, 2012, pp. 674–686.

- [34] A. Mohaisen, O. Alrawi, Unveiling zeus: automated classification of malware samples, in: WWW, 2013, pp. 829–832.
URL <https://doi.org/10.1145/2487788.2488056>
- [35] H. Kang, J. Jang, A. Mohaisen, H. K. Kim, Detecting and classifying android malware using static analysis along with creator information, *Int. J. Distributed Sens. Networks* 11.
URL <https://doi.org/10.1155/2015/479174>
- [36] H. Kang, J.-w. Jang, A. Mohaisen, H. K. Kim, Detecting and classifying android malware using static analysis along with creator information, *International Journal of Distributed Sensor Networks* 11 (6) (2015) 479174.
- [37] A. Mohaisen, O. Alrawi, M. Mohaisen, AMAL: high-fidelity, behavior-based automated malware analysis and classification, *Comput. Secur.* 52 (2015) 251–266.
URL <https://doi.org/10.1016/j.cose.2015.04.001>
- [38] H. Alasmary, A. Khormali, A. Anwar, J. Park, J. Choi, A. Abusnaina, A. Awad, D. Nyang, A. Mohaisen, Analyzing and detecting emerging internet of things malware: A graph-based approach, *IEEE Internet Things J.* 6 (5).
- [39] R. Masri, M. Aldwairi, Automated malicious advertisement detection using virustotal, urlvoid, and trendmicro, in: 2017 8th International Conference on Information and Communication Systems (ICICS), 2017, pp. 336–341.
- [40] A. Geniola, M. Antikainen, T. Aura, A large-scale analysis of download portals and freeware installers, in: *Secure IT Systems - 22nd Nordic Conference, NordSec 2017, Tartu, Estonia, November 8-10, 2017, Proceedings*, Vol. 10674 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 209–225.
- [41] R. Rivera, P. Kotzias, A. Sudhodanan, J. Caballero, Costly freeware: a

- systematic analysis of abuse in download portals, *IET Inf. Secur.* 13 (1) (2019) 27–35.
- [42] A. S. Manjeri, R. Kaushik, M. Ajay, P. C. Nair, A machine learning approach for detecting malicious websites using url features, in: 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 555–561.
- [43] A. K. Singh, N. Goyal, A comparison of machine learning attributes for detecting malicious websites, in: 11th International Conference on Communication Systems & Networks, COMSNETS 2019, Bengaluru, India, January 7-11, 2019, IEEE, 2019, pp. 352–358.
- [44] D. Liu, J. Lee, CNN based malicious website detection by invalidating multiple web spams, *IEEE Access* 8 (2020) 97258–97266.
- [45] PyWebCopy, Pywebcopy: Tool for scraping and saving webpages and websites with python (May 2022).
URL <https://pypi.org/project/pywebcopy/>
- [46] D. state, Domain Tools, Stats, News, Forum and Directory (May 2022).
URL <https://www.domainstate.com/registrar-tld-breakup.html>
- [47] Google, User report-based google safe browsing for chrome and firefox (May 2022).
URL https://safebrowsing.google.com/safebrowsing/report_general/
- [48] H. I. Archive, Top 1,000,000: Page weight report (May 2022).
URL <https://httparchive.org/reports/page-weight?lens=top1m>
- [49] Pingdom, Website Performance and Availability Monitoring (May 2022).
URL <https://www.pingdom.com/>
- [50] A. Cohen, N. Nissim, Y. Elovici, Maljpeg: Machine learning based solution for the detection of malicious JPEG images, *IEEE Access* 8 (2020) 19997–20011.

- [51] L. South, D. Saffo, M. A. Borkin, Detecting and defending against seizure-inducing gifs in social media, in: CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021, ACM, 2021, pp. 273:1–273:17.
- [52] W. Yost, C. Jaiswal, Malfire: Malware firewall for malicious content detection and protection, in: 8th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017, New York City, NY, USA, October 19-21, 2017, IEEE, 2017, pp. 428–433.
- [53] D. Jung, S. Lee, I. Euom, Imagedetox: Method for the neutralization of malicious code hidden in image files, *Symmetry* 12 (10) (2020) 1621.
- [54] A. G. West, A. Mohaisen, Metadata-driven threat classification of network endpoints appearing in malware, in: Proc. of DIMVA, 2014.
- [55] A. E. Kosba, A. Mohaisen, A. G. West, T. Tonn, H. K. Kim, ADAM: Automated Detection and Attribution of Malicious Webpages, in: 15th International Workshop on Information Security Applications, WISA, 2014, pp. 3–16.
- [56] A. Mohaisen, O. Alrawi, AMAL: high-fidelity, behavior-based automated malware analysis and classification, in: Proc. of WISA, 2014.
- [57] A. Mohaisen, Towards automatic and lightweight detection and classification of malicious web contents, in: Proc. of IEEE HotWeb, 2015.
- [58] SecurityTrails, Explore complete current and historical data for any internet assets. IP & DNS history (May 2022).
URL <https://securitytrails.com/>
- [59] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., *Journal of machine learning research* 9 (11).