Poster: Measuring and Assessing the Risks of Free Content Websites

Abdulrahman Alabduljabbar*, Runyu Ma†, Sultan Alshamrani*,
Rhongho Jang‡, Songqing Chen†, and David Mohaisen*
*University of Central Florida, †George Mason University, ‡Wayne State University
{jabbar, salshamrani}@knights.ucf.edu, {rma5, sqchen}@gmu.edu, r.jang@wayne.edu, mohaisen@ucf.edu

***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 from the literature. In particular, it is unclear if those websites are as safe as their premium counterparts. In this work, we set out to investigate the risks associated with free content websites. In particular, we examine the maliciousness of these websites at the website- and content-level. Our findings uncover that the free content websites are 4.5 times more likely to use 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 (content) level.**

I. Introduction

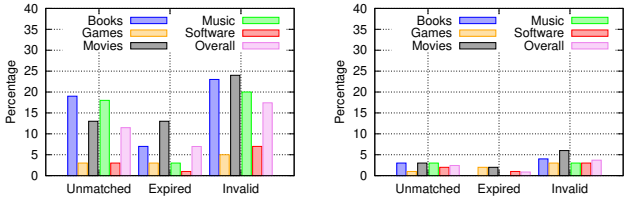
Online services and websites are categorized into two broad categories based on their monetization options: free content and premium websites. While the free content websites provide content free of charge, and are typically sustained by proceeds of advertising and user donations [2], the premium websites offer services through fees, e.g., subscriptions or pay-as-you-use operation models. Due to their monetization model, the extensive utilization of third-party advertisements on free content platforms introduce various risks. For instance, advertisements on these websites can be exploited by malicious actors for data and information leakage, or even the distribution of malicious scripts on the user device [4].

In this work, we explore and assess the risks associated with free content websites. In particular, we examine both the website- and component-level detection and vulnerability using two major off-the-shelf tools, *VirusTotal* [7] and *Sucuri* [6]. Our analysis concludes that there are significant security concerns associated with free content websites. We report that free content websites are more often associated with invalid SSL certificates, and 2.64 to 19 times more likely to be malicious, depending on the analysis level.

Contributions and Findings. We assembled a list of more than 1,500 free content and premium websites offering the same type of content. The websites are then crawled to obtain their content for further analysis. Further, we leverage *VirusTotal* and *Sucuri* APIs to assess the security risks associated with the 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 the premium websites detected as malicious is quite interesting and calls for further exploration.

TABLE I: An overview of the collected dataset.

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



(a) Free content websites (b) Premium websites

Fig. 1: The SSL certificate analysis results.

II. Dataset Overview

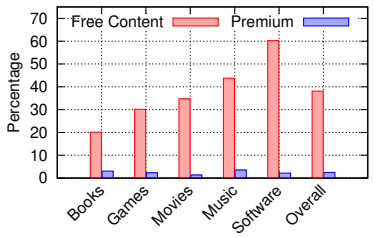
In this work, we compiled a list of 1,562 free content and premium websites. When selecting the websites, we consider two factors: (i) the most popular websites, (ii) websites that appear in the top results by Google, DuckDuckGo, and Bing search engines. Each website was manually examined and labeled to either premium or free content. The websites are then categorized into five groups based on the content they provide: books, games, movies, music, or software.

To understand the risks associated with free content websites, we crawled each website's contents (i.e., files) using PyWebCopy [5]. The obtained files are then used for the risk analysis and modeling, as they reflect the behavior of the provided service. Table I shows the distribution of the collected dataset. Notice that the average files crawled from premium websites are significantly larger than the average files obtained from the crawling of the free content websites.

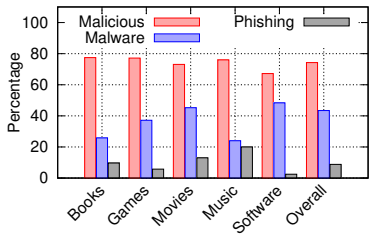
III. Risk Assessment Analysis

To assess the risk of websites, we leverage two public APIs: *VirusTotal* [7] and *Sucuri* [6] for harmful behavior analysis, two public website behavior and heuristic analysis tools.

SSL Certificate Analysis. In today's world, secure HTTP is a necessity, as it implements an encryption mechanism to protect the transferred content. Analyzing free content websites, we found that 36% of the websites have invalid HTTPS, compared to only 7% of the premium websites. Moreover, we found



(a) Websites labeled as malicious.



(b) Malicious free content websites.

Fig. 2: The potential maliciousness of free content and premium websites.

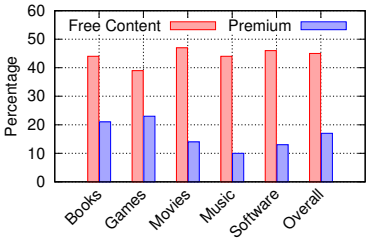
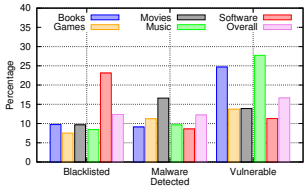
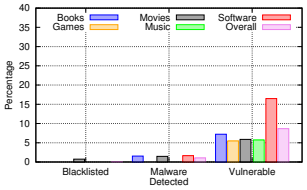


Fig. 3: The malicious file detected by *VirusTotal*: free content vs. premium.



(a) Free content websites



(b) Premium websites

Fig. 4: Sucuri API: Assessing the websites' maliciousness.

that 26% still allow HTTP (insecure) access, whereas 0% of the premium websites do. Further, we investigate the validity of the SSL certificate for both the 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 1 shows that, in total, 36% of the free content websites have issues with their certificates, compared to a total of only 7% of premium websites.

Malicious URLs Detection and Annotation. Using *VirusTotal* API, we extracted the malicious characteristics associated with the website URL, shown in Figure 2. Figure 2a 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 (≈74%) were labeled as malicious (Figure 2b); i.e., 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. Analyzing the scripts and website files is critical, as recent study [3] has shown that such content can be exploited, leading to information and data leakage, in addition to abusing the resources of the end-user device. We leverage *VirusTotal* API to understand the risks of free content websites. Figure 3 shows the percentage of malicious files detected in the free content and premium websites. Notice that 45% of the free content websites contain files that have been labeled as malicious, in comparison with 17% of its premium counterparts.

Websites' Vulnerability and Blacklisting. We leveraged Sucuri API [6] to obtain information of domains activities for free content and premium websites. 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 4. 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. In terms of blacklisting, Figure 4a shows that 12% of the free content websites were blacklisted, with *"Software"* free content websites have a significantly higher percentage (23.12%) in comparison with other categories.

IV. Conclusion and Future Direction

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 work "scratches the surface" of the risks associated with free contents websites, highlighting important problems and calling for further actions. While providing free contents to their users, free content websites are 19 times more likely to be associated with malicious behavior, questioning the safety of their provided services.

Future Direction: We will examine various features of free content websites, including top-level Domain distribution and HTTP requests variables. We will further investigate using these features to predict risk associated with these websites.

Acknowledgement. This work is supported by NRF grant 2016K1A1A2912757. The full version of this work is in [1]

References

[1] A. Alabduljabbar, R. Ma, A. Abusnaina, S. Alshamrani, R. Jang, S. Chen, and D. Mohaisen, "No free lunch: Measuring and modeling the free content websites in the wild," Uni. of Central Florida, Tech. Rep., 2022.

[2] M. Carvajal, J. A. García-Avilés, and J. L. González, "Crowdfunding and non-profit media: The emergence of new models for public interest journalism," *Journalism practice*, vol. 6, no. 5-6, pp. 638–647, 2012.

[3] A. Cohen, N. Nissim, and Y. Elovici, "Maljpeg: Machine learning based solution for the detection of malicious JPEG images," *IEEE Access*, vol. 8, pp. 19 997–20 011, 2020.

[4] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, "Knowing your enemy: understanding and detecting malicious web advertising," in *ACM conference on Computer and communications security*, 2012, pp. 674–686.

[5] PyWebCopy, "Pywebcopy: Tool for scraping and saving webpages and websites with python," January 2022. [Online]. Available: https://pypi.org/project/pywebcopy/

[6] Sucuri, "website security check & malware scanner," January 2022. [Online]. Available: https://sucuri.net/

[7] VirusTotal, "Analyze suspicious files and URLs to detect types of malware, automatically," January 2022. [Online]. Available: https://www.virustotal.com/