Dissertation Proposal

Demystifying the Network Characteristics of the Free Content Hosting Infrastructure

Mohammed Alqadhi

Date: October 20, 2023

Department of Computer Science
University of Central Florida
Orlando, FL 32816

Doctoral Committee:
Dr. David Mohaisen (Chair)
Dr. Cliff Zou
Dr. Xueqiang Wang
Dr. Sung Choi Yoo

Mohammed Alqadhi

Department of Electrical and Computer Engineering, University of Central Florida (UCF)
4000 Central Florida Blvd., HPA1-111, Orlando, FL 32816-2362 USA

Education

PH.D. IN COMPUTER SCIENCE (2020 – CURRENT)
University of Central Florida

M.SC. IN SOFTWARE ENGINEERING (2018 – 2020)
M.SC. IN INFORMATION ASSURANCE AND CYBERSECURITY (2018 – 2020)
Florida Institute of Technology
CGPA: 4.00

B.SC. IN SOFTWARE ENGINEERING (2011 – 2015)
King Saud University
CGPA: 3.5

Peer-Reviewed Publications

1. **Mohammed Alqadhi**, Abdulrahman Alabduljabbar, Kyle Thomas, Saeed Salem, DaeHun Nyang, and David Mohaisen, *Do Content Management Systems Impact the Security of Free Content Websites?*. The 11th International Conference on Computational Data and Social Networks (CSoNet), 2022

2. **Mohammed Alqadhi**, Mohammed Alkinoon, Jie Lin, Ahmed Abdalaal, and David Mohaisen, *Entangled Clouds: Measuring the Hosting Infrastructure of the Free Contents Web*. The ACM Cloud Computing Security Workshop (CCSW), in conjunction with the ACM Conference on Computer and Communications Security (CCS), 2023.

3. **Mohammed Alqadhi**, Ali Alkinoon, Saeed Salem, and David Mohaisen. *Understanding the Country-Level Security of Free Content Websites and their Hosting Infrastructure*. The 10th IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2023

4. **Mohammed Alqadhi**, and David Mohaisen. *The Infrastructure Utilization of Free Contents Websites Reveal their Security Characteristics: A Correlation Analysis*. *Submitted to* The 12th International Conference on Computational Data and Social Networks (CSoNet), 2023

Contents

Abstract

Free content websites are a significant component of the Web. They also constitute a significant threat dimension due to their user-base which may be affected by their malicious contents. As such, the security of free content websites is critical, although the existing work is limited to high-level security attributes. In particular, although recent studies reported a significant difference in the security of free content websites compared to premium content websites, they did not investigate the root causes of such characteristics in free content websites. Therefore, in this dissertation, we address this gap by exploring the latent aspects of the severity of free content websites' security, focusing mainly on the affinities of the infrastructure associated with such websites.

First, we investigate the potential causes of vulnerabilities in free content websites to address risks by exploring the reused codes represented by the content management system (CMS). We assemble 1,562 websites from free and premium content websites to identify their CMS and maliciousness. We used frequency analysis in both aggregate and per category (books, games, movies, music, and software), utilizing unpatched vulnerabilities, total vulnerabilities, malicious count, and percentiles to uncover trends and affinities of usage CMS's and their contribution to those websites. Moreover, despite the significant number of websites based on custom code, the use of CMS's is pervasive, with varying trends across types and categories. We find that even a small number of unpatched vulnerabilities in popular CMS's could be a potential cause of significant maliciousness.

Second, we statically explore the distribution of free content websites globally by analyzing their hosting network scale, cloud service provider, and country-level distribution, combined and per the content category, by contrasting these measurements to the characteristics of premium content websites. We further contrast the distribution of these websites with general websites sampled from the Alexa top-1M websites and explore their security attribute using various security indicators. We found that free and premium content websites are hosted mainly in medium-scale networks, a scale that is associated with a high concentration of malicious websites. Moreover, free content websites' cloud and country-level distribution are shown to be heavy-tailed, although with unique patterns compared to premium content websites.

Finally, we explore the geographical distribution and affinities of the free and premium content websites by studying their colocation in different countries and contrasting that with general websites. We focus on the distribution of malicious websites and their correlation with the National Cyber Security Index (NCSI), which measures a country's cyber security maturity and its ability to deter the hosting of such malicious websites. We found that most investigated websites are hosted in the United States. Interestingly, countries with higher malicious website rates have relatively low NCSI, mainly due to a lower score in privacy policy development. Ultimately, our research aims to identify regional vulnerabilities in the host country of free content websites and guide policy improvements at the country level to mitigate possible cyber threats.

1 Introduction

Websites have become an essential part of daily life. They provide countless services to their users. Some websites deliver one or more types of content, while others perform a specific task. Websites that deliver content to their users can be classified according to the fee they require to access their content. If website content is accessed for a fee, the websites are classified as Premium Content Websites. On the other hand, Free Content Websites provide content to the user without explicit cost. Recent studies show that a hidden cost is associated with the use of free content websites [10, 11, 13]. Some of the hidden costs are related to user privacy. Another cost is the risk exposure to which the user is exposed by using a vulnerable or malicious free content websites. In this dissertation, we will investigate the hidden causes for the lax free content websites' security, informing decision-makers to form higher security standards that will protect the Internet. We are particularly interested in the most popular types of content according to Alabduljabbar *et al*. [12], which are (books, games, music, movies, and software) content websites.

One of the potential causes of the security lack in free content websites is their Content Management System (CMS). The vulnerabilities associated with CMSs can decrease the security level of free content websites or premium content websites. The vulnerable CMS can spread and multiply widely because multiple websites use it. Therefore, we investigate the frequency distribution of free and premium content websites over CMSs. Looking to find the highly used CMS and check how many vulnerabilities have been discovered for this CMS. How severe are these vulnerabilities? Are there differences between free and premium content websites in terms of their CMSs? free content websites are shown to be the most vulnerable and malicious environment compared to other parts. Therefore, studying the distribution of free content websites compared to premium content websites on the Internet is very important to determine the root causes of malicious and vulnerable environments. Knowing the distribution on the different scales of the network, the Cloud Service Provider (CSP), and the geographical locations that contain this malicious environment is a step toward identifying and eliminating the risks that could threaten the Internet.

Exploring the enumeration of network vulnerability gives a readiness for the worst-case scenario. In addition, it will allow content providers to make their websites more resilient and robust. Additionally, it will help to perform a better enumeration that will be significant. Moreover, it will provide some prevalence affinity characterization of vulnerabilities with certain other variables, such as types of websites that may be excluded. However, it will help to understand which infrastructures are mainly used for certain types of malicious behavior to control them. Understanding the distribution of infrastructure on various network scales is a fundamental aspect of network management. By mapping the infrastructure, one can devise and implement effective containment and isolation techniques. If the infrastructure resides predominantly within a specific, smaller segment of the network, this offers distinct advantages. This allows for quick isolation, which in turn minimizes the potential disruption to the broader network. In such a scenario, the unaffected parts of the network, which host benign content, can continue to function without interruption.

In contrast, when infrastructure is distributed more broadly or diffusely across larger network scales or multiple networks, challenges arise. Here, the development of a targeted containment strategy becomes complex. Such a widespread distribution could potentially lead to larger disruptions during containment efforts, highlighting the importance of nuanced network profiling in the development of efficient strategies. Understanding the distribution of malicious content across CSPs is vital to effective risk mitigation. Large CSP networks make blocking all nodes costly, whereas smaller networks with concentrated malicious nodes can be feasibly blocked. If a CSP has few malicious nodes within a vast network, the strategic approach involves investigating the provider's country of operation to leverage local regulations for appropriate actions.

Assessing the geographical distribution of the free content websites helps point out malicious website hotspots. Users benefit by understanding cross-border vulnerabilities. If victimized by free content websites from another country, knowledge of that nation's cyber laws can inform legal recourse. This also shapes defense strategies against CSPs with many malicious websites. Existing cyber security agreements between countries can influence strategies. A country's concentration of malicious free content websites could inversely correlate with its maturity in cyber security policy, suggesting potential avenues for improvement.

Given the fast increase in the number of data breaches on free content websites. We find it important to focus on the need to investigate the root causes of the security lax on free content websites compared to other types of websites. Taking into account the popularity and global spread of free content websites.

1.1 Statement of Research

In this dissertation, we take a step forward in investigating the security of free content websites. Proposing four comprehensive studies that explore the affinities of the different hosting patterns and infrastructures utilized by free content websites and their malicious behavior. We further elaborate on every study.

Do Content Management Systems Impact the Security of Free Content Websites? (§ 3). Website vulnerabilities, particularly within free content platforms, can put countless users at risk. These vulnerabilities, once introduced, can rapidly proliferate across the Internet. Often, these free platforms are attractive targets for malevolent activities, leveraging newly discovered weaknesses. A key observation is the prevalent use of Content Management Systems (CMSs) among such sites, leading to a situation where a single CMS vulnerability can compromise multiple websites, amplifying the risk for users. To go deeper into this issue, we conducted a methodical study of more than 1,500 free and premium content websites. Our analysis focused on the CMSs they employ and their related security attributes. Through the evaluation of metrics such as unpatched and total vulnerabilities, along with malicious activity counts, we were able to discern patterns across various content genres such as books, games, music, movies, and software.

Measuring the Hosting Infrastructure of the Free Contents Web (§ 4). Recent research underscores the widespread of free content websites relative to premium. Multiple websites often share a single hosting infrastructure, meaning that the security of one can impact the others. If a significant portion of nodes within a network are malicious, the entire network may be considered suspect. Assessing the security of networks that host free content sites aids in developing more effective containment approaches. By segmenting the issue, we can discern the prevalent structures of these networks. Recognizing high concentrations of malicious sites can guide preventive measures, but care must be taken not to penalize benign sites. A proactive approach involves regulating the hosting infrastructure, especially focusing on cloud service providers (CSPs) and specific geographic distributions linked with malicious activity. Enhanced governance over these CSPs and regions can further mitigate risks. To this end, we conducted a detailed analysis of the global distribution patterns of free content websites, comparing their hosting, Cloud Service Providers (CSP), and country-specific distributions to premium sites. By benchmarking against Alexa's top-1M general sites, we delved into their security features. Our findings reveal that both content types predominantly use medium-sized networks, often associated with high malicious risk. Furthermore, free content sites showed distinct distribution trends. This research provides a clearer understanding of the ecosystem of free content websites through a quantitatively focused assessment.

Understanding the Country-Level Security of Free Content Websites and their Hosting Infrastructure (§ 5). Free content websites are popular and diverse around the globe. To control the spread of malicious free content websites on the Internet, collaboration is required between hosting providers in terms of geographical location. Hosting countries can emphasize the security standard in hosting free content websites and regulate the spread of malicious websites. Improving cyber security policies is required to ensure the existence of collaboration between nations. The National Cyber Security Index (NCSI) provides an index score that measures the different aspects of security development in particular nations. Indicating the maturity of cyber security policies, digital development, and the level of protection of digital services per country, linking that with the hosting pattern of malicious and benign free content websites would reveal the weakness that needs to be improved around the globe. Thus, we examined the distribution of malicious sites among 1,562 content websites relative to the NCSI, a metric reflecting a country's cybersecurity maturity. The majority of these sites, encompassing various content categories, were compared with a subset of Alexa's top million websites. By identifying the predominant hosting nations and their associated percentages of malicious sites, our research highlights regional hosting vulnerabilities and underscores the need for strategic policy enhancements to counteract cyber threats.

2 Related Work

Several prior works studied the security, privacy, and modeling of free content websites [10–13,48, 68], the role of the infrastructure, for example, using CMS in free content websites [14], detecting bots in game content [49], malicious activities including malware classification and vulnerability assessment [15, 16, 23, 41, 58, 60, 79], understanding and improving the security of the top-used websites and network-level characteristics [22, 25, 26, 28, 30, 36, 46, 47, 50, 52, 54, 57, 66, 69, 73, 81, 82], statistically analyzing domain-specific security breaches in web services (e.g., of health providers) and associated network characterization [17–20], and investigating the security of the infrastructure and its role in securing web services [43,44,61,67,80]. Numerous works [24,32,42, 53] analyzed various security features of the web infrastructure.

Moreover, several works explored the regional analysis for domain-specific websites, such as governments and universities. The key difference in this study is that we investigate the security of free content websites across different countries, utilizing various new features of their modeling and contrasting them to the general web population of websites. This, as a result, supplements the other efforts focused on understanding the security, privacy policies, accessibility, or performance of such websites [24, 70, 72, 75–78, 83–85]. Given the multitude of studies and space constraints, we focus on a select group of highly relevant studies concerning this work and findings.

2.1 Websites Security Analysis

free content website's security and privacy have been a significant concern. Alabduljabbar *et al*. [12] investigated the security of free content websites by analyzing SSL certificates and examining the certificate issuer, validity, and signature. They also studied the validity of these certificates by identifying if they were genuine or fake, the coverage term, validity, and website security. They found that 36% of the free content websites use invalid, expired, or fake SSL certificates. Another study performed component and website-level analyzes to understand vulnerabilities using two main standard tools, VirusTotal and Sucuri [11], linking free content websites to significant threats. Mindful of the implicit security cost, another work has looked into the interplay between privacy policies and the quality of those websites. Namely, the prior work examined user comprehension of risks linked to service use through privacy policy understanding [13]. The researchers investigated free content websites' privacy policies and their expressiveness utilizing TLDR [10], a natural language processing (NLP) pipeline for privacy policy analysis [40]. They also examined the uniqueness of the policy for each free content website compared to premium content websites. Among other interesting findings, they concluded that free content websites' privacy policies are unclear in stating their data collection practices do not provide useful information compared to premium content websites' policies, and utilize mostly predefined privacy policy templates, which may not state the actual data tracking, storage, and sharing practices of users data.

Free hosting infrastructure and its security are also studied. Roy *et al.* [68] examined the problem of phishing attacks hosted on free web hosting domains (FHDs), which can evade detection and take-down by anti-phishing entities. A large-scale analysis of 8.8k FHD URLs shared on Twitter and Facebook found that these attacks remain active for 1.5 times longer than regular phishing URLs, have 1.7 times lower coverage from blocklists, and take 3.8 times longer to be detected by security tools compared to regular phishing attacks. Several works examined the security of the top-used website by Alexa. Kontaxis *et al.* [46] performed a study on the security of cross-domain policies in Rich Internet Applications (RIA) such as Microsoft Silverlight and Adobe Flash. These tools are used widely with cross-domain policies that might be maliciously used to threaten user privacy. The authors performed their study on Alexa's top 100K websites and Fortune 500 companies' websites at the country level. They found more than 6500 vulnerable websites and were exposed to security attacks.

Several works investigated the relationship between the quality of service, usability, and security of websites based on their content. Figueras-Martín [33] analyzed the Freenet darknet's website connectivity, relationships, and content. The results showed a significant general availability of websites on Freenet, significant nodes within the network connectivity structure, and predominant illegal content. Li *et al.* [50] investigated websites' malicious advertising activities, using 90,000 leading websites by identifying how attackers reach advertising networks. They further unveiled the role of malicious nodes in malicious web advertising by studying their different characteristics and interaction information.

2.2 Measuring the Hosting Infrastructure

The security of the website's infrastructure is essential to secure the network since almost half of the used websites often reside within specific content management systems [14, 59]. Analyzing network association and security of websites has been explored by Noroozian *et al.* [62], who performed a longitudinal study of broadband CSP security efforts in mitigating IoT malware, namely Mirai. They investigated the Mirai infection rate across 342 global CSPs and found that 55% of the difference in the infection rate is due to the number of subscribers to that CSP.

Wickramasinghe *et al.* [82] examined the hosting patterns of malicious domains by analyzing the hosting types of IP addresses employed by malicious websites. They found that more than 95% of the malicious websites are hosted on regular hosting IPs. In contrast, 97.1% of these malicious websites are co-hosted with unrelated benign websites, and the top 5 hosting providers for malicious domain hosts were Cloudflare, Amazon, Google, OVH, and Microsoft. The study suggests that more should be done by hosting providers to safeguard their shared hosting infrastructures.

Fryer *et al.* [34] investigated the problem of malicious web pages and the solutions the hosting providers could implement. Liao *et al.* [51] examined the problem of long-tail SEO (search engine optimization) spam on cloud web hosting services and identified 3,186 abusive cloud directories for long-tailed SEO spam by analyzing 15,774 cloud directories across 10 major providers. They unveil monetization strategies spammers use and their evasion techniques, such as obfuscation via link shorteners and client-side JavaScript when a platform does not support server-side scripting.

Tajalizadehkhoob *et al*. [74] studied the distribution of web security features and software patching practices in shared hosting providers to understand their influence on website compromise. Wang *et al*. [80] examined the consolidation of DNS and web hosting providers, an increasing trend that may have wide-reaching implications for the security, reliability, and accessibility of the Internet. They found that Amazon and Cloudflare are responsible for exclusively hosting the name servers for over 40% of domains and only five organizations (Cloudflare, Amazon, Akamai, Fastly, and Google) host about 62% of the Tranco top 10K index pages and most external page resources for these sites.

2.3 Country-Level Security of Free Content Websites

Another study investigated the security of the top-used website lists provided by Alexa. Raponi and Di Pietro [66] performed a detailed analysis of Alexa's top 200 websites with domains registered in certain European countries and analyzed the password recovery management mechanisms adopted by each website. They found more than 54% of the websites in France, 36% in Italy, 47% in Spain, and 33% in the United Kingdom to be vulnerable in December 2017, with almost no difference a year later, highlighting minimal progress in adapting the General Data Protection Regulation (GDPR) standard. Verkijika *et al*. [77] investigated the public values delivery of web-based platforms in Sub-Saharan Africa by analyzing 279 e-government websites from 31 countries, revealing a lack of various features associated with accessibility, citizen engagement, trust development, responsiveness, dialogue, and quality of service among the surveyed websites.

Shafqat *et al*. [72] conducted a comparative analysis of 20 countries' National Cyber Security Strategies (NCSS) using different metrics, such as perception of cyber threats, organization overviews, critical sectors and infrastructure, incident response capabilities, etc. Their results show that while all countries have defensive measures to protect their cyberspace from threats, there is variation in the approaches they use. The study concludes with recommendations nations may use to design their NCSS documents with global best practices for improved cyber resilience.

Vaughan *et al*. [75] examined the coverage of websites from four countries—United States, China, Singapore, and Taiwan—in four major search engines—Google, Yahoo!, MSN, and Yahoo! China—and found that the United States-based sites had higher coverage rates than those from other countries. Moreover, they found that the Chinese sites had the lowest average coverage rate. They also found that the language factor did not explain this difference in representation, although the visibility, as measured by the number of links to a site, did affect its chance of being indexed. Yahoo! China provided better coverage for Chinese and surrounding region sites than the global Yahoo! search engine.

Velasquez *et al*. [76] investigated the accessibility, resources, and staff availability provided by 1,517 public library websites in Australia, Canada, and the United States. The research aimed to extend the definition of physical library branches into their digital counterparts. To assess this, 18 criteria were used to determine if they were present on each website, and descriptive statistics revealed that Canadian and United States libraries met more criteria than Australian libraries. However, many similarities between all three countries' websites were found overall. Bangera *et al*. [24] presents the results of an extensive study of web hosting, with a particular focus on differences between ads and regular content. A virtual private network (VPN) service was used to collect data from top country-specific websites in 52 countries, and the findings show that ads employ more servers for broader load distribution. At the same time, replication is local for ads and global for regular content.

While there is an overlap between the prior work and ours in the analyzed modalities, our work stands out in utilizing those modalities to understand the ecosystem of free content websites with the premium and general websites to their geographical distribution and co-location in countries with varying security policy standings.

3 Do Content Management Systems Impact the Security of Free Content Websites?

3.1 Summary of Completed Work

Users often face vulnerabilities when accessing free content websites. This study explores the causes of these vulnerabilities by analyzing more than 1,500 websites, free and premium, to identify their content management system (CMS) and malicious traits. Through frequency analysis across various content categories, we examine unpatched vulnerabilities, total vulnerabilities, and malicious counts. Our findings highlight the prevalent use of CMSs, with differing trends across content types. In particular, even a few unpatched vulnerabilities in popular CMSs can lead to significant malicious activity.

3.2 Introduction

Today, free content websites are an essential part of the Internet, providing ample resources to users in free books, movies, software, and games. The security of free content websites has always been a focal point of debate and studies. The main questions around the study of free content websites have been their security and privacy: what are the costs associated with using those websites? Those costs have been studied by contrasting free content websites with premium websites–websites that provide similar content but charge fees–across multiple dimensions, e.g., vulnerabilities in code, infrastructure utilization, and the richness of the privacy policies [8,19,20,65,77].

For instance, it was found that there is a higher level of maliciousness in free content websites than in premium websites, which makes free content websites unsafe for the users [11]. Digital certificates used by those websites are shown to be problematic [12]. Their privacy policies are shown to be limited in covering essential policy elements [13]. Despite the importance of the literature, it falls short in determining the root cause for the lack of security and privacy in free content websites. The contrast provided in the literature highlights that free content websites are a source of lurking risks and vulnerabilities that could expose users and their data to significant security costs. However, there is a lack of a study that looks into various potential contributors to the vulnerability to better understand a mitigation strategy for those risks.

To address this gap, we revisit the security analysis of free content websites. The critical insight we utilize for our analysis is that the security of any website is best understood by understanding the codebase of its content. In essence, we also hypothesize that many of the vulnerabilities associated with those websites could be caused by a repeated software design pattern in their codebase, as is the case with other web technologies. We find that we can understand the repeated patterns by studying the utilization of third-party content management systems (CMS's), which are heavily used in today's websites.

In this work, we contribute to the state-of-the-art by analyzing and contrasting the security of free content websites through the lenses of CMS analysis using 1,562 websites. We annotate the websites with their malicious attributes and systematically evaluate the role of CMS as a contributing factor. We find that a significant number of the websites (≈44%) use CMS's, which comes with vulnerabilities and contributes to maliciousness. We find that the use pattern of CMS's is unique across different types of websites and categories. The top-used CMS's have common aspects, such as unpatched vulnerabilities, which help explain the maliciousness of websites using them.

3.3 Dataset and Data Annotation

Websites. We compiled a dataset that contains 1,562 websites, with 834 free and 728 premium, which has been used in previous work [11–13]. In selecting those websites, we consider their popularity while maintaining a balance per the subcategory of a website. To determine the popularity of a website, we used the results of the Bing, DuckDuckGo, and Google search engines as a proxy, where highly ranked websites are considered popular. To balance the dataset, we undertook a manual verification approach to vet each website across the subcategory (see below). That is, we classified the websites into five categories based on the content they predominantly serve: software, music, movies, games, or books. The following are the free and premium content websites count per category: books (154 free, 195 premium), games (80 free, 113 premium), movies (331 free, 152 premium), music (83 free, 86 premium), and software (186 free, 182 premium).

Dataset annotation. For our analysis, we supplement the data set in various ways. We focus mainly on information reflecting the exposure to user risk [12]. We determine whether a website is malicious or benign using the VirusTotal API [6]. VirusTotal is a framework that offers cyber threat detection, which helps us analyze, detect, and correlate threats while reducing the required effort through automation. Specifically, the API allowed us to identify malicious IP addresses, domains, or URLs associated with the websites we use for augmentation.

CMS's. Since this work aims to understand the role of software (CMS, in particular) used across websites and its contribution to threat exposure, we follow a two-step approach: (1) website crawling and (2) manual inspection and annotation. First, we crawl each of the websites and inspect its elements to find the source folder for the website. From the source folder, we list the source and content for each website to identify the CMS used to develop this website. This approach requires us to build a database of the different available CMS's to allow annotation automation through regular expression matching. We cross-validate our annotation utilizing existing online tools used for CMS detection. We use CMS-detector and w3techs, two popular tools, to extract the CMS's used for the list of websites. For automation, we build a wrapper that prepares the query with the website, retrieves the response of the CMS used from the corresponding tool, and compares it to the manually identified set in the previous step. Among the CMS's identified, WordPress is the most popular, followed by Drupal, Django, Next.js, Laravel, CodeIgniter, and DataLife. In total, we find 77 unique CMS's used across the different websites, not including custom-coded websites.

Vulnerabilities. Our dataset's final augmentation and annotation are the vulnerability count and patching patterns. For each CMS, we crawl the results available in various portals concerning the current version of the CMS to identify the associated vulnerability. Namely, we crawl such information from cvedetails, snyk.io, openbugbounty, and wordfence. Finally, to determine whether a vulnerability is patched or not (thus counting the number of unpatched vulnerabilities), we query cybersecurity-help. [1,3–5,27,35,63].

3.4 Analysis Methods

The key motivation behind our analysis is to understand the potential contribution of CMS's to the (in)security of free content websites, which has been established already in the prior work, as highlighted in section 2. To achieve this goal, we pursue two directions. The first is a holistic analysis geared toward understanding the distribution of various features associated with free content and premium websites (combined). The second is a fine-grained analysis that considers the per-category analysis of vulnerabilities. In essence, our study utilizes frequency analysis of various features to understand trends and affinities and a holistic view of vulnerabilities. The features are:

❶ **CMS.** This feature signifies the industry name of the content management system utilized by the free content website, premium content website, or both.

❷ **Count.** Signifies the number of websites that use a given CMS for their operation. Particularly, we assume each given website utilizes only one CMS, which has been the case in our analysis.

❸ **Percent.** This feature signifies the normalization of the count feature by the total number of studied websites. We use the percentage to understand a relative order of the CMS's contribution that is easier to interpret.

❹ **Malicious count (MC).** This is calculated per CMS. It highlights the total number of websites utilizing the given CMS deemed malicious. For our maliciousness check, we utilize the output of VirusTotal, where a website is deemed malicious if at least one scanner has flagged it as malicious.

❺ **Malicious percentage per CMS (MPCP).** This feature signifies the normalized MC by the count feature. It highlights the significance (as a percentage) of the malicious websites for the given CMS. It highlights the actual relative contribution of the CMS to the maliciousness of websites taking into account their relative representation in our dataset.

❻ **Malicious percentage (MP).** This feature signifies the MC feature normalized by the total MC value (i.e., the overall number of malicious websites) by capturing a given CMS's relative contribution to the total number of malicious websites. It signifies the contribution irrespective of the representation of that CMS in our dataset. The gap between MPCP and MP signifies whether a given CMS is more secure in the abstract or not.

❼ **Total vulnerabilities (TV).** This feature signifies the total number of vulnerabilities associated with the given CMS.

❽ **Unpatched vulnerabilities (UV).** This feature signifies the total number of unpatched vulnerabilities associated with the given CMS.

❾ **Correlation analysis.** This feature identifies the relationship between the CMS's and the maliciousness of the sites. For that, we use the Pearson correlation, defined as $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$. Here, X is a random variable associated with free content/premium content type (malicious vs. benign), and Y is a random variable capturing the CMS's associated with the given type.

3.5 Results and Discussion

3.5.1 Overall Analysis Results and Discussion

First, we explore the distribution of the various features outlined earlier in a holistic manner, considering the free vs. premium labels of the websites. The results are shown in Table 1, and we make the following observations.

(1) The total number of malicious websites is 525 out of 1,561 websites, corresponding to 33.63% of them. This number is surprisingly large, especially in contrast to general website maliciousness levels, which are estimated at 1%[1].

(2) In terms of vulnerability, the maliciousness of those websites corresponds to 2,760 vulnerabilities in the CMS's the websites employ. Among them, 145 are unpatched at the time of our scanning. While small as a percentage (only 5.25%), we note that some of those unpatched vulnerabilities are associated with the most popular CMS's our dataset. For example, one unpatched vulnerability is associated with WordPress, which is used by more than 24% of websites and is associated with 32% of the total number of malicious websites. This supports our hypothesis on the role of CMS's as an amplification avenue of vulnerabilities and associated impact, where a single vulnerability could be utilized to contaminate a large number of websites and recruit them into malicious endeavors.

(3) We observe that a majority of the websites (883, or 56.6%) in our dataset use custom code, with 30.5% (or 269) of them being malicious. Custom-coded websites made up 51.2% of the total malicious websites. In contrast, while the websites that used CMS's represented 43.4% of all websites, they had 48.8% of all malicious websites, which corresponds to 37.8% maliciousness among those that utilize CMS's.

[1] https://ophtek.com/what-are-malicious-websites/

(4) Our estimate of the role of CMS's serves only a lower bound, as we do not consider the potential for shared codes among custom websites (i.e., websites that do not use a standard CMS). Those websites might be reusing cross-website codes, which could amplify the vulnerabilities.

(5) We observe a range of percentages of vulnerabilities and maliciousness across the website groups utilizing different CMS's, where that percentage sometimes exceeds 40% (well above the average) even with major CMS's; e.g., WordPress (44.3%), Next.js (53.85%), and Shopify (70%). These results show a significant trend in the maliciousness of the websites based on their platforms.

3.5.2 Category-based Analysis Results and Discussion

The main results provided in section 3.5.1 are profound, although they do not look into the individual categories and how they differ (if at all). To help answer this question, we conduct the same analysis we had in section 3.5.1, but per category; for books, games, movies, music, and software. Our analysis provides a contrast against the mean (§3.5.1) and group (free vs premium).

General Observations. Before delving into the specific analysis of each category, we make the following broad observations. **(1)** We notice that while the use of CMS's is common among both the free content and premium websites, the usage follows different patterns: whereby the number of CMS's utilized by the free content websites is small, it is prominent in the case of premium websites, with a more significant heavy-tailed distribution (i.e., a significant number of the CMS's have a minimal representation in terms of the websites that utilize them). This is very well-captured in the "others" row in every table, where we combine the CMS's with 1-2 websites. We observe that "others" in the case of premium websites is significantly more than that in the free content websites part of the table. **(2)** Across the different websites, we observe inconsistent patterns concerning the division between custom code and CMS: where it is significantly greater in the case of free content websites vs premium for books (75% vs 43%), movies (86% vs 51%), music (65% vs 51%), and software (59% vs 30%), the pattern does not hold for games (47% vs 56%).

❶ **Books.** The results in Table 2 show that there are 153 free content websites and 195 premium websites. With 348 websites, 149 use a CMS, and 199 use custom code. Under this category, 46 (30.8%) of free content and 53 (27.2%) of premium websites are malicious. In total, 99 (28.5%) of the books' websites are malicious. This result shows that slightly more free content websites are malicious. In contrast, both types of websites have a malicious percentage that is less than the average (33.6%) per Table 1. Interestingly, the free content websites have a 39.5% chance of being malicious vs. a 30.6% chance for the premium.

It is natural to ask whether the ranking of the CMS's persists in both the free content and the premium websites. While the top CMS is the same in both cases, others in the top 4 for the free content (ordered) are Drupal, Django, vBulletin vs. Shopify, Drupal, and Magento for premium. Shopify is the most malicious CMS (percentage-wise) with 70%. It is used only in the premium books category, in contrast to the top (count-wise) malicious CMS (WordPress) used in both.

Table 1: Distribution of the combined free and premium websites across different CMS's. Studied distribution characteristics are for each CMS: the percentage among all websites (percent), the count, malicious count (MC), malicious per CMS websites count (MPCP), the malicious percentage among the total websites (MP), the total number of identified vulnerabilities with the given CMS (TV), and the total number of unpatched vulnerabilities for the given CMS (UV).

CMS	Count	Percent	MC	MPCP	MP	TV	UV
Custom code	883	56.57	269	30.46	17.23	–	–
WordPress	379	24.28	168	44.33	10.76	8	1
Zendesk	26	1.67	11	42.31	0.70	2	2
Drupal	25	1.60	3	12.00	0.19	228	0
Adobe EM	22	1.41	1	4.55	0.06	93	0
Shopify	20	1.28	14	70.00	0.90	0	0
Magento	18	1.15	5	27.78	0.32	210	3
Next.js	13	0.83	7	53.85	0.45	9	0
Laravel	9	0.58	2	22.22	0.13	9	1
vBulletin	9	0.58	3	33.33	0.19	0	0
HubSpot CMS	8	0.51	5	62.50	0.32	3	0
Bigcommerce	6	0.38	0	0.00	0.00	20	0
Django Framework	6	0.38	1	16.67	0.06	1	0
Salesforce C360	6	0.38	0	0.00	0.00	1	0
Gatsby	5	0.32	2	40.00	0.13	1	0
IPS Community	5	0.32	2	40.00	0.13	3	0
Joomla	5	0.32	1	20.00	0.06	83	2
Oracle CX	5	0.32	0	0.00	0.00	25	0
Salesforce Cloud	5	0.32	5	100	0.32	56	0
Sitecore CMS	5	0.32	2	40.00	0.13	19	1
Others	101	6.47	24	23.76	1.54	1,989	135
Total	1,561	100	525	–	33.63	2,760	145

❷ **Games.** Similarly, as shown in Table 3, there are 80 free content and 113 premium websites for games. With 193 websites in total, 91 of them are shown to use CMS, while 102 used custom code. Among the free content games websites,43 (53.75%) are shown to be malicious in comparison to 35 (31%) of the premium games websites. Put together, the total number of malicious games websites were 78 (roughly 40%). From these results, we make several observations: (1) significantly more free content websites are malicious, (2) both types of websites have a malicious percentage that is close to or significantly higher than the average (33.6%) per Table 1, and (3) the free websites have a 50% chance of being malicious when using a CMS compared to about 40% in the case of the premium websites.

We notice that the top CMS is observed to be the same in both cases. However, others in the top 4 for free content (ordered) are DataLife Engine, vBulletin, Discuz! vs. Magento, Zendesk, and Bigcommerce for premium game websites. DataLife Engine is also shown to be the most malicious CMS (percentage-wise) at 100%, and is only used in the free games category, in contrast to the top

Table 2: Distribution of free vs. premium **books content websites** across different CMS's. Studied distribution characteristics are for each CMS; Keys are as in Table 1.

Free Content Websites					
CMS	Count	Percent	MC	MPCP	MP
Custom code	115	75.16	31	26.96	20.26
WordPress	22	14.38	10	45.45	6.54
Drupal	3	1.96	0	0.00	0.00
Django Framework	2	1.31	1	50.00	0.65
vBulletin	2	1.31	1	50.00	0.65
Others	9	5.88	3	33.33	1.96
Total	153	100	46	–	30.07
Premium Websites					
Custom code	84	43.08	19	22.62	9.74
WordPress	46	23.59	12	26.09	6.15
Shopify	10	5.13	7	70.00	3.59
Drupal	7	3.59	0	0.00	0.00
Magento	6	3.08	2	33.33	1.03
Others	42	21.53	13	30.95	6.67
Total	195	100	53	–	27.18

Table 3: Distribution of free vs. premium **games content websites** across different CMS's. Studied distribution characteristics are for each CMS; Keys are as in Table 1.

Free Content Websites					
CMS	Count	Percent	MC	MPCP	MP
Custom code	38	47.50	22	57.89	27.50
WordPress	34	13.75	18	52.94	22.50
DataLife Engine	2	2.50	2	100	2.50
vBulletin	2	2.50	0	0.00	0.00
Discuz!	1	1.25	0	0.00	0.00
Others	3	3.75	1	33.33	1.25
Total	80	100	43	–	53.75
Premium Websites					
Custom code	64	56.64	15	23.44	13.27
WordPress	22	19.47	8	36.36	7.08
Magento	4	3.54	2	50.00	1.77
Zendesk	4	3.54	1	25.00	0.88
Bigcommerce	2	1.77	0	0.00	0.00
Others	17	15.04	9	52.94	7.96
Total	113	100	35	–	30.97

(count-wise) malicious CMS (WordPress) used in both categories.

❸ **Movies.** As shown in Table 4, we found that 331 free content websites and 152 premium content websites serve movies. Among the 483 websites, 120 used a CMS, while 363 used a custom code.

Table 4: Distribution of free vs. premium **movies content websites** across various CMS's. Studied distribution characteristics are for each CMS; Keys are as in Table 1.

Free Content Websites					
CMS	Count	Percent	MC	MPCP	MP
Custom code	285	86.10	105	36.84	31.72
Wordpress	34	10.27	17	50.00	5.14
Django Framework	2	0.60	0	0.00	0.00
Laravel	2	0.60	1	50.00	0.30
DataLife Engine	1	0.30	1	100	0.30
Others	7	2.11	4	57.14	1.21
Total	331	100	128	–	38.67
Premium Websites					
Custom code	78	51.32	6	7.69	3.95
WordPress	18	11.84	5	27.78	3.29
Zendesk	11	7.24	5	45.45	3.29
Adobe EM	6	3.95	0	0.00	0.00
Drupal	4	2.63	2	50.00	1.32
Others	35	23.03	5	14.29	3.29
Total	152	100	23	–	15.13

On the other hand, 128 (38.7%) are shown to be malicious in the free content category vs. 23 (15.1%) for the premium websites category. This result is somewhat expected, given the general association of free movie websites with malicious content distribution. Overall, we found that 151 (30.26%) of the websites in the movies category are malicious. As such, we make the following observations: (1) significantly more free content websites are malicious (23% gap), and (2) both types of websites are slightly less likely to be malicious than the average (33.6%).

We also explore the trend in top CMS's, which are shown to be the same in both categories. Others in the top 4 for the free content (ordered) are Django Framework, Laravel, DataLife Engine vs. Zendesk, Adobe Experience Manager, and Drupal for premium. We notice a draw between WordPress and Laravel for the most malicious CMS (as a percentage) for the free websites with 50%. We also notice that Drupal is the most malicious CMS used only in premium movies with a percentage of 50%. On the other hand, the top (count-wise) malicious CMS is (WordPress) which is used in the free and premium movies websites.

❹ **Music.** As illustrated in Table 5 there are 83 free content and 86 premium content websites, out of which 98 websites use custom code and 71 use a CMS, giving us a total of 169 websites. Overall, 32 (38.6%) of free content were reported as malicious sites compared to 15 (17.44%) of premium content sites. This result conveys the following: (1) Free music websites are significantly more like to be malicious compared to premium music sites. (2) It is noticeable that music websites are slightly less likely to be malicious than the average. Namely, the free websites have twice the chance of being malicious compared to the premium websites (40% vs. 20%).

Table 5: Distribution of free vs. premium **music content websites** across different CMS's; Keys are as in Table 1.

Free Content Websites					
CMS	Count	Percent	MC	MPCP	MP
Custom code	54	65.06	24	44.44	28.92
WordPress	18	21.69	5	27.78	6.02
Drupal	2	2.41	0	0.00	0.00
MediaWiki	2	2.41	0	0.00	0.00
Shopify	2	2.41	1	50.00	1.20
Others	5	6.02	2	40.00	2.41
Total	83	100	32	–	38.55
Premium Websites					
Custom code	44	51.16	7	15.91	8.14
WordPress	19	22.09	2	10.53	2.33
Zendesk	4	4.65	2	50.00	2.33
Gatsby	3	3.49	1	33.33	1.16
Oracle CX	2	2.33	0	0.00	0.00
Others	14	16.28	3	21.43	3.49
Total	86	100	15	–	17.44

Similarly, the most utilized CMS under the music category is WordPress in the free and premium content alike, followed by (ordered) Drupal, MediaWiki, and Shopify in free content, and Zendesk, Gatsby, and Oracle CX Commerce in the premium websites. Where the most malicious CMS (percentage-wise) is Shopify in free music, the results show that Zendesk is the most malicious CMS in premium music with a similar percentage for both CMSs (50%).

❺ **Software.** For The last category, results in Table 6 show that 247 websites use CMS's compared to 121 websites that use custom code, for a total of 368 websites; 186 for free content and 182 are premium. Overall, we found 116 (62.4%) of the free content websites are malicious vs. 34 (18.7%) of the premium websites, which adds up to a total of 150 (40.76%) malicious software websites. The results illustrate a significant difference between the free and the premium malicious websites with enormous trends in software-free content. On the other hand, both types have a more malicious percentage than the average (33.63%). Interestingly, the free software websites have more than 60% chance of being malicious against the 20% chance for the premium websites. Unique to this category, the top code base is not the custom code but WordPress, which deviates from the last four categories. Moreover, the other most used CMS's (ordered) are Contentteller, IPS Community Suite, and Jimdo, among free websites, vs. Adobe Experience Manager, Drupal, and Next.js for the premium. Nevertheless, Contentteller is shown to be the most malicious CMS with a chance of 100% maliciousness and has been used only in the free software category. In contrast to the top (count-wise) malicious CMS (WordPress) used in both free and premium software.

Table 6: Distribution of free vs. premium **software content websites** across different CMS's. Studied distribution characteristics are for each CMS; Keys are as in Table 1.

Free Content Websites					
CMS	Count	Percent	MC	MPCP	MP
WordPress	111	59.68	81	72.97	43.55
Custom code	69	37.10	33	47.83	17.74
Contentteller	1	0.54	0	0.00	0.00
IPS Community	1	0.54	1	100	0.54
Jimdo	1	0.54	0	0.00	0.00
Others	3	1.61	1	33.33	0.54
Total	186	100	116	–	62.37
Premium Websites					
WordPress	55	30.22	10	18.18	5.49
Custom code	52	28.57	7	13.46	3.85
Adobe EM	14	7.69	0	0.00	0.00
Drupal	7	3.85	0	0.00	0.00
Next.js	5	2.75	4	80.00	2.20
Others	49	26.92	13	26.53	7.14
Total	182	100	34	–	18.68

3.5.3 Putting it Together: Discussion

❶ **Correlation Heatmap.** The correlation heatmap is shown in Figure 1. Most malicious sites are free content websites based on the correlation heat map. We also find that the free software websites are the most malicious, shown in Table 6 with (62.37%) malicious percent. In contrast, the relation between the maliciousness of a website and the premium category is relatively weak. We also find that the premium category uses more CMS's than the free websites. Ultimately, free content websites using custom code are the most malicious. The second most are the sites using WordPress, which are likely to be malicious for both the free and premium categories. Premium websites using Zendesk and Shopify are the most malicious among the other premium websites.

❷ **Further Discussion.** Based on the previous results, we infer that CMS websites have a higher malicious percentage than custom code websites. Table 1 shows that 30.46% of the custom code websites are malicious against 38.55% of the websites that use CMS's. We also find that the free content sites have the highest vulnerabilities and maliciousness compared to premium websites in the per-category comparison. We also notice that websites that use specific CMS's, such as WordPress, Shopify, Next.js, Gatsby, and Sitecore, have a high chance of being malicious. It will be hard to generalize these results among the CMS's that only occur once or twice, even if they have a 50% chance of being malicious.

One outcome of this study is the assessment of the risk of CMS's that are generally associated with malicious websites. Taking five as a threshold of occurrence with a malicious percent

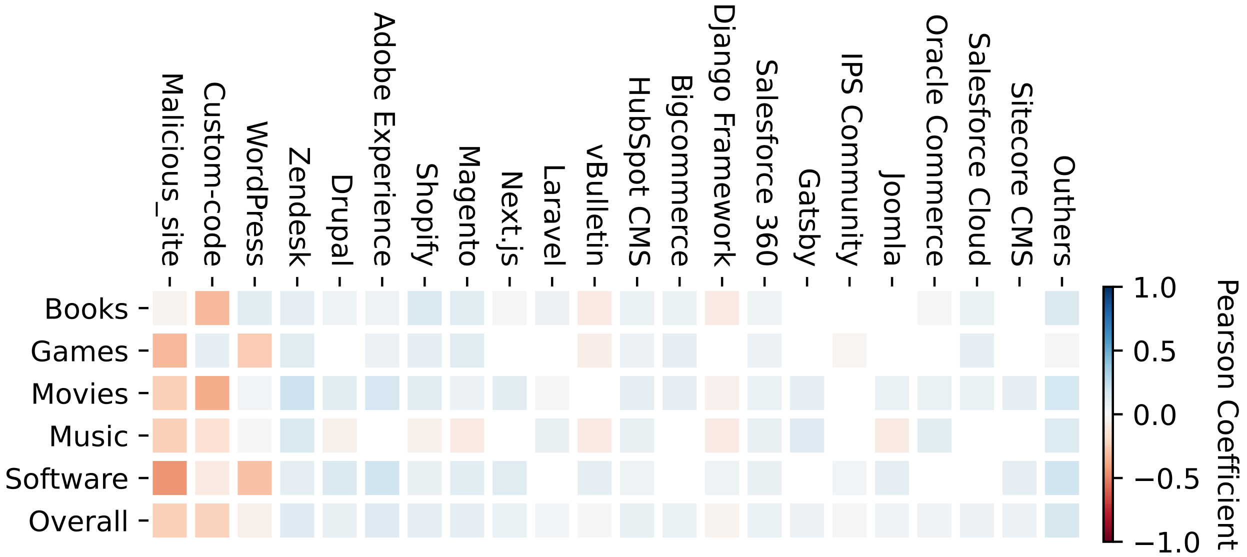


Figure 1: Correlation Heatmap: Pearson correlation coefficient shows the linear relation between two features. This case shows the most contributor CMS's based on Table 1 and the five content categories (Books, Games, Movies, Music, and Software). It depicts the relation between the maliciousness of the websites and the different categories. The yellow color means that this feature strongly relates to free categories. In contrast, the blue color reflects the strong relation with the premium content. The white color represents the weak relation.

higher than 30%, Shopify is considered high-risk, with 20 websites and 70% of them being malicious. When applying the same criterion to other CMS's, we find that Next.js (53.85%, 13 times), WordPress (44.33%, 379 times), and Zendesk (42.31%, 26 times) are high-risk. We highlight the possibility of reducing the attack surface of websites by not using a high-risk CMS or by fixing the CMS to restrict these vulnerabilities.

We noticed that the CMS with a lower malicious percentage has the highest number of total vulnerabilities but the lowest number of unpatched vulnerabilities. We highlight a pattern to use in practice: those with lower unpatched vulnerabilities are likely CMS's that provide good maintenance and apply the latest security standards. One can recommend reducing the risk of websites using CMS's using the same insights. It has been argued that this could be accomplished by having ongoing monitoring and management of the free content websites [64].

3.6 Summary and Work to be Completed

Our study shows various analyses to uncover specific risks associated with those websites in contrast to premium. It highlights the significant challenges with free content websites regarding increased vulnerabilities to maliciousness. Although well-established that free content websites are more likely to be malicious, we tie this likelihood to their utilization of CMS's, in aggregate and at a per-category analysis. Recognizing this problem and the potential role CMS's play in websites security, it is essential to generalize this insight to a more significant number of websites, contrast those trends to other general websites (besides the free content vs. premium), and conduct measurements over time to capture the security dynamics.

4 Measuring the Hosting Infrastructure of the Free Contents Web.

4.1 Summary of Completed Work

In our research, we analyzed the network distribution and security characteristics of free websites compared to premium websites and a set of benchmarks from Alexa's top million websites. A significant observation was that free and premium content websites are predominantly hosted on medium-scale networks, which have a notable association with malicious web activity. The distinct distribution patterns observed for free websites provide insights into potential strategies to mitigate security threats, highlighting the potential for targeted isolation based on these patterns.

4.2 Introduction

Websites are broadly classified into two types based on whether they require payments: free and premium content websites [13]. As the name indicates, free content websites do not require any payment for their service while providing content such as books, music, movies, software, and games. On the other hand, premium content websites ask for a premium to access the same type of content. Both types of web content and associated websites are prevalent, although free content websites are rising for their appeal because of the convenience. However, this appeal resulting in wide use makes them a target for security and privacy threats [11]. Free content websites are an essential part of the Internet, and their wide use signifies their associated risks. This risk is further seen in the various analyses showing that the free content websites have poor privacy policies that do not protect the users' rights and data [10]. Moreover, free content websites contain some of the most vulnerable and malicious contents compared to their premium counterparts or the general web (e.g., Alexa's top million websites) [9, 11–14, 21, 47].

Shortcomings of the Literature. Despite several studies on understanding free content websites' ecosystem and security, the literature did not explore the infrastructure employed by free content websites and how they differ from premium content websites. To gain an understanding of the interplay between free content websites and Internet infrastructure, it would be essential to (1) study the contributing networks to the phenomenon of free content websites, and whether there is an affinity between the networks size and the security of free content websites, (2) analyze the affinity between Cloud Service Providers (CSPs), their attributes, and the security of free content websites, (3) understand the distribution of the free content websites, in contrast to the premium content websites, spatially at the country level. In this work, we explore these dimensions by modeling free content websites, contrasted to premium content websites and the general web population, to improve our understanding of their ecosystem and associated risks.

Our Approach and its Rationale. Our study is multi-faceted, and each of the studied dimensions carries importance and rationale. The study of the network characteristics of free websites in contrast to premium websites is critical to model a part of the Internet, given its prevalence, and to shed light on more profound insights into the root causes, or even correlations, of Internet vulnerabilities in situ. Moreover, understanding the distribution of free websites over the different network scales, CSPs, and geographical locations containing the most malicious content is a step toward identifying and eliminating the risks that pose threats to Internet users. Finally, understanding the deployment of free websites at the country level will highlight the effectiveness of security policies and regulations in encountering potentially malicious web content and their hosting.

We employ network analysis methods to understand the hosting patterns of free websites by studying the size of networks they mostly reside within. To achieve this goal, we divided the networks into four sizes: small, medium, large, and very large. We identify each type based on the subnet mask assigned to the IP address with each domain. Each IP is assigned a subnet mask that represents the number of possible addresses reserved by that hosting provider, which indicates the number of possible associated neighboring publicly addressable hosts to the studied websites.

Knowing the distribution of free content websites over network scales will help guide practical defenses. Since free content websites were reported to be more malicious than other websites types of websites [13, 14], such a knowledge of the most used network scales will simplify the implementation of a containment and isolation technique within the malicious infrastructure at the network level. For example, if the scale of the network that hosts the potentially malicious free content websites is small, isolating the malicious hosts from a whole network (prefix) could be the most effective approach without risking interrupting communication with many benign hosts. Alternatively, if the network with the potentially malicious free content websites is extensive, such a decision cannot be made without interrupting the communication, favoring the individual hosts filtering for containing these free content websites.

Understanding the distribution of malicious free content websites over networks can also guide containment strategies under limited resources. For example, identifying which networks are hosting a majority of the malicious free content websites can help prioritize those networks in a containment effort. Similarly, CSP-level profiling of such free content websites can help determine the appropriate risk prevention procedures without affecting other benign host associations.

By investigating the geographical locations of these CSPs, we can better understand what strategies can be done to ensure network security based on the country's rules, regulations, and Cybersecurity policies. Understanding the distribution of free content websites over countries is necessary to identify the concentration of malicious websites and infrastructures. As a result, it will help the users of these services protect themselves from being victimized by a vulnerability that cannot be controlled or governed by the law of the same country. For example, if a user uses a free content website that resides in a different country and the website victimizes the user, the user would know if the law of the free content website country is strict and elaborated so that the user can perform legal action against that free website. They can legally remove any acquisitions, definitions, misinformation, or malicious content from the free content website. Also, it will help to determine the necessary actions against the CSPs that contain the most malicious websites.

Contributions. Starting with a dataset that consists of 1,562 free and premium content websites (combined), used in the previous works [10, 13, 14] and augmented across various dimensions, and a sample of Alexa's million websites [37], we deliver the following contributions. (1) **Hosting and Network-level Analysis.** We systematically and comprehensively measure, analyze, and contrast the hosting patterns of free and premium content websites, and their distribution among network scales (§4.4.1). Moreover, we examine free and premium content websites distribution among scales by taking into account the type of content they serve (§4.4.2). We measure the concentration of malicious websites in both cases and use various metrics of security across the different network scales. (2) **Hosting Networks Spatial Analysis.** We identify the malicious free and premium content websites and examine their affinities with various infrastructure attributes. Besides the network scale, we enumerate the hosting countries and highlight heavy-tailed and highly focused hosting patterns (§4.4.3). (3) **Cloud Service Providers (CSPs) Analysis.** We enumerate the commonly used CSPs for hosting free content websites and determine which CSPs and study the affinities in hosting as well malicious free content websites across CSPs, in isolation and in contrast with both premium content websites and general websites (§4.4.4).

4.3 Methodology

In this section, we review our research questions (§4.3.1), the dataset and the utilized data collection methods for augmentation (§4.3.2), and the various analysis dimensions (§4.3.7).

4.3.1 Research Questions

The main goal of this work is a systematic understanding of the hosting patterns of free content websites, their utilization of Internet infrastructure, and their contrast to premium content websites and the general website populations. Moreover, we take on the task of identifying any specific patterns concerning infrastructure utilization by considering the specific content types (e.g., books, games, movies, music, or software). To this end, our pursuit raises several questions we attempt to answer. **RQ1** What are the hosting patterns of the malicious free content websites? **RQ2** Is there any network pattern associated with malicious free content websites? **RQ3** Is there an affinity between the distribution of free content websites, premium content websites, and their hosting patterns at the country scale? **RQ4** How do the hosting patterns of free and premium content websites compare to those of the general websites? **RQ5** What are the main distribution characteristics of free content websites with respect to the major cloud service providers?

4.3.2 Dataset and Data Collection

Our effort to understand our research questions relies on several datasets: 1. a primary dataset of free content websites, premium content websites, and their associated annotations, 2. two complementary dataset for augmenting the analysis of the primary dataset in terms of security (maliciousness detection) and network scale enumeration, and 3. a dataset for general websites population to aid in our contrast analysis against free content websites, premium content websites, and their utilization of infrastructure. In the following, we review those datasets and how we obtained them.

4.3.3 Free and Premium Content Websites Dataset

The main free and premium content websites dataset utilized in this study consists of 1,562 websites obtained from Alabduljabbar *et al.* and utilized in their previous works [11–14]. The criteria followed for whether to include a website in our list are inherited from the prior work: (1) popularity, (2) language, and (3) activity. The *popularity* of a website is assessed by the ranking of the website in major search engines as a returned result for a keyword search. The returned websites are then examined with respect to the *language* they use, and only websites that use English as the primary language are retained for further analysis. Finally, the *activity* is determined by examining whether the website returned by the search engine for the keyword is online (live) at the time of our analysis. At the time of the work of Alabduljabbar *et al.*, all of the websites are online. Moreover, we note that Abduljabbar's criteria are more elaborate, including ensuring a balanced dataset across the different categories of content the websites provide.

In their original work, Alabduljabbar *et al.* use three search engines, Google, DuckDuckGo, and Bing, as a proxy for the popularity estimation of websites. The determination of whether a website is an free content website or premium content website is based on a manual inspection. Also, each website is manually labeled with a category based on its contents: books, games, movies, music, or software. The categories and other indicative keywords (e.g., free, premium, paid, pay-per-use, etc.) are used as the search keywords in the respective search engine.

Upon filtering the websites according to the above criteria, we query all the domain names to extract their associated IP addresses. As a result, we verified that 1,509 websites are online (representing 96.6% of the total websites). Of those online websites, 788 were free, and 721 were premium content websites. Specifically, we classified the online websites into five categories based on their content type. Each category holds the distribution over the free and premium type as follows (free vs. premium): books (144 vs. 191), games (87 vs. 111), movies (310 vs. 152), music (80 vs. 86), and software (176 vs. 181).

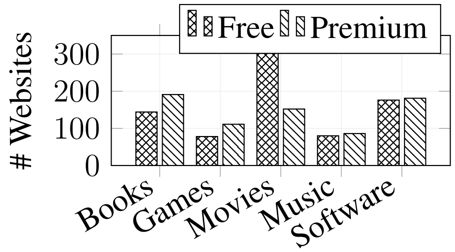


Figure 2: Per-category distribution of the Free content websites vs. Premium content websites.

4.3.4 Malicious Websites Annotation

A primary goal of this work is to examine the latent variables that may help explain the lax security in free content websites, particularly concerning their usage of Internet infrastructure. To start, we used VirusTotal [6] to identify if a website (at the time of the analysis) was malicious or benign. VirusTotal, an online tool that integrates more than 70 combined scanning engines, is used to determine whether a domain name (URL), IP address, or binary—which could be identified by a unique identifier; *i.e.*, hash value of its contents—is malicious or benign. VirusTotal allowed us to identify malicious IP addresses, domains, or URLs associated with the websites we used in this study. We augmented the collected data based on the output of VirusTotal. Since VirusTotal returns many detection results, we consider an entity—website or IP—to be malicious if at least one of the returned scan results from VirusTotal for that entity is marked as malicious.

4.3.5 Network Scale Enumeration

Another goal of this work is to understand the scale of the network infrastructure associated with free and premium content websites, which leads us to define network scale. To identify the scale of the network for each website in our dataset, we use the associated IP address with that domain as an analysis feature. Then, we used two major APIs, *ipdata* [2] and *IPSHU* [7], to extract intelligent information related to the given IP address, such as domain name, subnet mask, cloud service provider, and geographical location, for further augmenting our dataset with scale information. The subnet mask for each IP address is extracted to identify the network scale for each website. Then, each website is classified based on the network scale using the CIDR (Classless Inter-Domain Routing) notation as follows: (1) **small network:** any network that is between (/25 and /32), (2) **medium networks:** any network that is between (/16 and /24), (3) **large networks:** any network between (/8 and /15), and (4) **very large networks:** any network that is /7 and below. The sizes above correspond to the value range of (20 to 27), (28 to 216), (217 to 224), and (225 and more). The characteristics of the network scales are in Table 7.

Table 7: Network scales and their characteristics. The network size is represented by each slash bit of the CIDR notation, where the decimal number after the slash character represents the number of bits in the network prefix of the IP address. The maximum slash bit is 32 (IPv4). x represents the number of bits and y represents the number of addresses.

Scale	Bits in CIDR	# Addresses
Small (SN)	$/24 < x \leq /32$	$2^8 > y \geq 2^0$
Medium (MN)	$/16 < x \leq /24$	$2^{16} > y \geq 2^8$
Large (LN)	$/8 < x \leq /16$	$2^{24} > y \geq 2^{16}$
Very Large (VLN)	$/0 < x \leq /8$	$2^{32} > y \geq 2^{24}$

4.3.6 General Websites Sample

A benchmark dataset representative of the web ecosystem is needed further to understand the infrastructure utilization for free and premium content websites. To this end, an unbiased random sample of 2,400 from Alexa's top million websites dataset [37] was generated and used. To ensure that the measured characteristics of the websites represent the larger population, we fixed the margin of error and the confidence interval to 2% and 95%, respectively, to produce a sample size of 2,400. By definition, the change in the sample size is insignificant as the population grows. Therefore, the number of samples is kept at 2,400. We call this dataset "general" for short in the subsequent implementation.

As in the preprocessing and augmentation procedures of free and premium content websites dataset, the general websites were examined, whether online or offline (*i.e.*,*activity*). As a result, only 2,057 websites are online, corresponding to 85.7% compared to 96.5% for the final dataset of free and premium content websites. We then extracted each sample's CSPs, countries, and subnet mask information using the *ipdata* API. Moreover, VirusTotal was used to identify the malicious websites and their concentration among the different network scales for comparison.

4.3.7 Analysis Dimensions

In this study, we use the statistical analysis approach to recognize the patterns and statistical differences between free content websites, premium content websites, and general websites across different analysis dimensions. This study uses eight major dimensions: the network scale, CSP, country, maliciousness (count and percentage), maliciousness per feature, count, and percentage. In the following, we define each of those dimensions. The workflow of this analysis is in Figure 3.

Network Scale. The network scale analysis dimension is based on the network scale feature, defined in §4.3.5. This feature signifies the network size where the studied websites reside. Based on the annotation in §4.3.5, this feature has four valid values: small, medium, large, and very large.

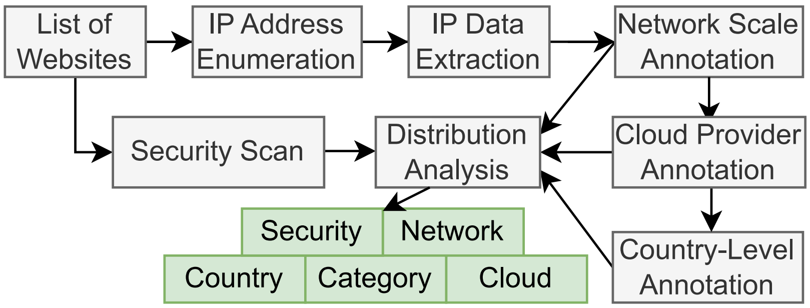


Figure 3: The data enumeration and feature extraction leading to the final distribution of websites.

Cloud Service Provider (CSP). Signifies the name of the cloud service provider where free content websites, premium content websites, and the general websites reside. Based on the eventual analysis that we present later in 4.4, this feature has 298 valid values (service providers' names).

Country. This feature signifies the country's name where the infrastructure (drive based on the IP allocation) of free content websites, premium content websites, and general websites reside. Our eventual analysis shows this feature has 41 valid values for different countries.

Malicious Count (MC). This feature signifies the number of malicious websites residing in a specific infrastructure entity based on the studied feature scale (CSP, country, or network size). The maliciousness of a website is determined by the VirusTotal scan results, as highlighted in §4.3.4.

Malicious Percentage (MP). This feature signifies the *normalized* malicious websites count for the studied feature (i.e., country, CSP, network scale) over the individual sample's total malicious count (i.e., total malicious websites in free content websites, premium content websites, both, or general websites). In essence, this feature highlights the contribution of one specific studied infrastructure entity among all other entities to the maliciousness ascribed to the entity type (country, CSP, or network scale). Namely, $MP = MC/$(Total # Malicious Websites).

Malicious Per Feature Percentage (MPFP). This feature signifies the normalized number of malicious websites over the number of websites that reside in the given infrastructure entity (country, CSP, or network scale). This feature describes the contribution of the studied entity to the malicious websites population, considering their respective size in our dataset. Compared to the MC dimension, which characterizes the contribution of a given entity to the total maliciousness indicated in our analysis, MPFP normalizes this quantity by the total number of websites potentially residing in the given entity to address the fact that different entities could have vastly different scales. In contrast to the MC feature indication, the MPFP means that an entity large in scale could contribute very little to the maliciousness once this scale is considered. Namely, this feature is formulated as $MPFP = MC/$(# Malicious Websites under One Dimension).

Count. This feature signifies the number of websites that reside within the assigned entity type: network scale, hosting CSP, or hosting country.

Percentage. This feature signifies the count of websites—free content websites, premium content websites, or both, and general websites—that reside in a given entity type normalized by the total number of the studied websites for that given website type. This feature is used to understand the variance of the distribution of the websites among the studied features.

4.4 Analysis Results

This section presents the findings of our distribution analysis pipeline applied to the extracted dataset. The trends and patterns of free content websites distribution across various network scales, CSPs, and countries are described and compared with premium content websites. The distribution of free and premium content websites on the top million most frequently used websites is examined. Finally, an overall analysis that compares the outcomes of each distribution feature analysis is provided, followed by a detailed analysis of categorical outcomes for the different content types: books, games, movies, music, and software.

4.4.1 General Network Scale Analysis

The distribution analysis performed over the network scale holds several vital insights that can be summarized as follows: 1. Most websites reside in the medium-scale networks; 81.24% of the total number of the studied free and premium websites combined with general websites. 2. The premium content websites use large networks, shown in the results with a more secure percentage of websites than medium networks. 3. The free content websites in medium networks are the riskiest type of websites where almost 90% of free content websites are in medium networks and 40% of them are classified as malicious. 4. In our per-category analysis, books, movies, and software websites are shown to use large networks more than games and music websites. They are also reported as being less malicious than games and music websites except for the free software category, where most software websites resided in medium networks and reported the highest MP. This final result is expected, as one natural to recruit victim devices by convincing users to install unauthenticated free software on their devices—thus affecting the eventual security label. 5. All of the category websites (books, games, movies, music, and software) free and premium combined primarily reside in medium networks where the average premium websites concentration in medium networks varies between ≈75% to ≈85%, compared ≈84% and over 97% in free websites, where the game category has the highest concentration in both websites. 6. Most of the CSPs are equally distributed over medium- and large-scale networks. 7. The network hosting of free and premium content websites is dominant mainly in the United States, where ≈58% of the websites reside. 8. Large-scale networks are mostly in the US since ≈71% of these websites reside.

Dataset versus Benchmark.. As demonstrated in Table 8, a concentration of malicious websites is observed in the medium-sized networks for both the combined free content websites/premium

Table 8: free content websites+premium content websites compared to the general websites distribution across different network scales using the website count (#), percentage (%), malicious count (MC), malicious count per feature percentage (MPFP), and percentage of malicious websites among all websites (MP) for each scale.

Free and Premium Content Websites					
Scale	#	%	MC	MPFP	MP
Small	38	2.52	6	15.79	0.40
Medium	1271	84.23	348	27.38	23.06
Large	199	13.19	23	11.56	1.52
Very Large	1	0.07	0	0.00	0.00
Total	1509	100	377	24.98	24.98
General Websites					
Small	0	0	0	0	0
Medium	1626	79.05	80	4.92	3.89
Large	430	20.90	12	2.79	0.58
Very Large	1	0.05	0	0.00	0.00
Total	2057	100	92	4.47	4.47

content websites and general datasets, with MP of 23.06% and 3.89%, respectively. Specifically, 27.38% of medium-network websites in the free and premium content websites dataset exhibit malicious behaviors per network scale count (MPFP). In comparison, the general dataset has a considerably lower rate of 4.92% per the MPFP feature. This significant difference supports our hypothesis that a higher proportion of malicious websites are hosted on the free and premium content websites datasets. Consequently, it is essential to consider the network scale and the extent of malicious activity when effectively managing network security risks. These findings emphasize the importance of considering these factors when comparing and analyzing different datasets. Failure to do so may lead to inadequate strategies for addressing network security threats, potentially compromising the safety and integrity of online systems.

Free versus Premium Websites. Per Table 9, the majority of free and premium content websites reside in medium networks, accounting for ≈89.1% and ≈78.9%, respectively. We observe that the MPFP for free content is nearly double that of premium content websites', with ≈40.5% compared to ≈22.2%. The highest MP in both free and premium content websites is observed in medium networks, with ≈37.7% for free content websites and ≈19.8% for premium content websites.

These findings highlight a need for implementing measures against websites in medium networks hosting malicious content. Moreover, ≈20% of premium content websites are hosted in large networks, which may contribute to their enhanced security, as fewer free content websites are in these networks. These results support our hypothesis that malicious websites in free and premium content websites tend to exhibit similar hosting patterns. A deeper investigation of these patterns is necessary to address potential vulnerabilities and enhance the security of those systems.

4.4.2 Per-Category Network Scale Analysis

In this section, we review the results and findings of our measurements through a per-category analysis, considering websites associated with books, games, movies, music, and software.

Book Websites. Table 9 reveals significant trends across different network scales hosting free and premium content websites. 85% of the free content websites and 80.1% of the premium content websites are hosted in medium networks, which, combined, account for ≈82.4% of both types of websites. The MPFP for the free content websites is ≈30% and ≈27.8% for the premium content websites. Interestingly, ≈31.7% of free content websites were found malicious compared to ≈30% of premium content websites, pointing to a substantial issue with the maliciousness of book websites in medium-scale networks. In these networks, ≈27% of free content websites and ≈24% of premium content websites contribute to the malicious websites' total MP. It is worth noting that ≈17.3% of premium content websites reside in large networks compared to only 10% of free content websites. Additionally, ≈28.6% of small networks in free content websites were identified as malicious compared to 0% in premium content websites. While the difference is not substantial for large networks, it is quite significant for small networks, potentially relating to the difference in the total MP between free and premium content websites offering book content.

Games Websites. Per 9, a significant concentration of game websites is in medium-scale networks, with ≈97.4% of free content websites and ≈85.6% of premium content websites. In total, ≈90.5% of both websites are situated in medium networks. This suggests that organizations providing gaming content prefer to use medium networks, possibly to ensure maximum network speed for users worldwide. Additionally, ≈12.6% of premium content websites use large networks compared to ≈1.3% of free content websites. This supports our finding that large networks enhance the security of premium content websites, as it contributes to only 1.8% MP while the total MP of premium content websites is ≈31.5%, in contrast to the 64.1% MP of free content websites. This highlights the increased risk associated with free gaming websites compared to premium gaming

Table 9: An overview of the distribution per category (Free content websites vs. Premium content websites, books, games) across different network scales.

Free vs. Premium

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	26	3.30	6	23.08	0.76
Medium	702	89.09	297	42.31	37.69
Large	60	7.61	16	26.67	2.03
Very Large	0	0.00	0	0.00	0.00
Total	788	100	319	40.48	40.48
Premium Content Websites					
Small	12	1.66	1.00	8.33	0.14
Medium	569	78.92	143	25.13	19.83
Large	139	19.28	16	11.51	2.22
Very Large	1	0.14	0	0.00	0.00
Total	721	100	160	22.19	22.19

Books

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	7	5.00	2	28.57	1.39
Medium	123	85.00	39	31.71	27.08
Large	14	10.00	2	14.29	1.39
Very Large	0	0.00	0	0.00	0.00
Total	144	100	43	29.86	29.86
Premium Content Websites					
Small	4	2.09	0	0.00	0.00
Medium	153	80.10	46	30.07	24.08
Large	33	17.28	7	21.21	3.66
Very Large	1	0.52	0	0.00	0.00
Total	191	100	53	27.75	27.75

Games

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	1	1.28	0	0.00	0.00
Medium	76	97.44	50	65.79	64.10
Large	1	1.28	0	0.00	0.00
Very Large	0	0.00	0	0.00	0.00
Total	78	100	50	64.10	64.10
Premium Content Websites					
Small	2	1.80	0	0.00	0.00
Medium	95	85.59	33	34.74	29.73
Large	14	12.61	2	14.29	1.80
Very Large	0	0.00	0	0.00	0.00
Total	111	100	35	31.53	31.53

Table 10: The distribution per category (movies, music, and software) of Free content websites vs. Premium content websites across different network scales.

Movies

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	5	1.61	0	0.00	0.00
Medium	284	91.61	75	26.41	24.19
Large	21	6.77	7	33.33	2.26
Very Large	0	0.00	0	0.00	0.00
Total	310	100.00	82	26.45	26.45
Premium Content Websites					
Small	2	1.32	1	50.00	0.66
Medium	115	75.66	21	18.26	13.82
Large	35	23.03	1	2.86	0.66
Very Large	0	0.00	0	0.00	0.00
Total	152	100.00	23	15.13	15.13

Music

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	1	1.25	0	0.00	0.00
Medium	72	90.00	31	43.06	38.75
Large	7	8.75	0	0.00	0.00
Very Large	0	0.00	0	0.00	0.00
Total	80	100	31	38.75	38.75
Premium Content Websites					
Small	1	1.16	0	0.00	0.00
Medium	66	76.74	12	18.18	13.95
Large	19	22.09	3	15.79	3.49
Very Large	0	0.00	0	0.00	0.00
Total	86	100	15	17.44	17.44

Software

Free Content Websites					
Network	#	%	MC	MPFP	MP
Small	12	6.82	4	33.33	2.27
Medium	147	83.52	102	69.39	57.95
Large	17	9.66	7	41.18	3.98
Very Large	0	0.00	0	0.00	0.00
Total	176	100	113	64.20	64.20
Premium Content Websites					
Small	3	1.10	0	0.00	0.00
Medium	140	77.35	31	22.14	17.13
Large	38	20.99	3	7.89	1.66
Very Large	0	0.00	0	0.00	0.00
Total	181	100	34	18.78	18.78

and both game websites.

Movie Websites. In the realm of movie websites, as seen in the games category, most free and premium content websites are situated within medium-scale networks. Table 10 reveals that 91.61% of free content websites and 75.66% of premium content websites fall into this category, meaning that nearly 9 out of 10 free content websites and 3 out of 4 premium content websites reside in medium networks. Large networks are particularly appealing to premium websites, hosting 23.03% of them while only accommodating 6.77% of free websites. Intriguingly, 26.41% of free websites within medium networks have been identified as malicious, accounting for 24.19% of the total 26.45% MP. In contrast, premium content websites exhibit a lower MP of 15.13%, with 13.82% contributed by websites on medium networks. Notably, 18.26% of every four premium content websites may be identified as malicious. Small and very large networks are relatively scarce for both free and premium content websites. A striking disparity emerges when examining small-scale networks, as 50% of premium content websites found malicious compared to 0% of free websites, although the overall count remains minimal. These insights imply that the movie category lags behind the book and games categories in terms of security. Movie websites frequently utilize cross-domain video players, potentially escalating the number of reported security threats and the susceptibility to malicious attacks. In addition, these findings emphasize the prevalence of medium-scale networks in hosting movie websites and the heightened MP among free websites relative to their premium counterparts. Furthermore, the allure of large networks for premium content providers could be attributed to enhanced security or superior performance.

Music Websites. Table 10 shows the distribution of music websites across various network scales for the movies category. Both free and premium content websites are primarily concentrated in medium-scale networks, accounting for more than 90% and 75% of their respective distributions. Large-scale networks host ≈8.8% of free content websites and ≈22.1% of premium content websites, highlighting a preference for premium content websites in these networks. Moreover, a significant disparity exists between free and premium content websites regarding MP, with close to 40% of free content websites classified as malicious compared to only ≈17% of premium content websites. Medium-scale networks contribute significantly to this difference, as 43% of free content websites in these networks are malicious, in contrast to ≈18% of premium content websites. This results in a total MP contribution of 100% for free content websites and 80% for premium content websites in medium-scale networks. Large-scale networks play a role in premium content websites' security, demonstrating a lower MP of ≈3.5%.

Interestingly, no malicious free content websites are found in large-scale networks, even though the overall MP of free content websites is more than double that of premium content websites. This suggests that large networks are safer for free content websites. Common among the movie and music websites is their use of the same content across multiple domains. This may increase the potential for malicious activities targeting these websites.

Software Websites. Table 10 shows the results, where free websites have a high MPFP of ≈70% in medium networks—≈84% of the websites reside. This is considered the highest malicious concentration among all categories, which was expected since software applications are the most likely to require access to system commands when a user installs any software from free content websites, making them highly exposed to malicious activities. In comparison, premium content websites—where ≈77% are in the medium networks—have a relatively high MPFP of ≈22%.

premium content websites use more large networks, with ≈21% compared to ≈9.7% of free content websites. Free content websites show significant use of small networks, with ≈6.8% of the websites, where ≈33.3% of them are identified as malicious, compared to a 0% MPFP in premium content websites. This signifies the risk of using free software websites, where they show a high MPFP in large networks at ≈41.2%, making the total MP ≈64.2% versus ≈18.8% in premium content websites. The implication of these results highlights the severity of using software websites, especially those in the free content websites category that reside in medium networks, as they exhibit a high MPFP of ≈69.4%.

4.4.3 Networks' Spatial Analysis

While abstract network-level distribution analysis sheds light on networks structure of the free and premium content websites, a provider- and country-level annotation may provide insight into the interdependencies in this ecosystem. We start this annotation of the cloud service providers (CSPs) and country of provider while providing the networks' spatial analysis. Table 11 illustrates the CSPs distribution for free and premium content websites across various network scales. We observe that most websites (≈84%) are hosted in medium networks, with large networks hosting only ≈13% of websites, while only ≈2.5% are in small networks. A negligible (¡0.1%) percent of websites reside in very large networks. Moreover, free and premium content websites are shown to be dispersed across multiple CSPs, with the highest concentrations being in Cloudflare (≈27%)

Table 11: The distribution (count; #) of the CSPs and hosting countries (Alpha-3) of free and premium content websites across the small (SN), medium (MN), large (LN), and very large (VLN) network scales.

CSPs.

Networks Distribution Over CSPs					
CSP	#	SN	MN	LN	VLN
Cloudflare	410	0	410	0	0
Amazon	240	0	121	119	0
Liquid	72	0	72	0	0
Trellian	42	0	42	0	0
Google	41	0	28	13	0
LeaseWeb	37	0	37	0	0
Sp-Team	35	0	35	0	0
Akamai	33	0	33	0	0
Fastly	26	0	26	0	0
Microsoft	21	0	2	19	0
Others	552	38	465	48	1
Total	1509	38	1271	199	1
%	100	2.52	84.23	13.19	0.07

Countries.

Networks Distribution Over CSPs					
Country	#	SN	MN	LN	VLN
USA	884	24	718	141	1
BEL	99	0	99	0	0
NLD	95	0	95	0	0
DEU	89	4	78	7	0
AUS	48	0	46	2	0
FRA	35	1	28	6	0
CHN	33	1	26	6	0
GBR	31	6	18	7	0
CAN	24	0	18	6	0
IRL	22	0	12	10	0
Others	149	2	133	14	0
Total	1509	38	1271	199	1
%	100	2.52	84.23	13.19	0.07

and Amazon (≈16%), which predominantly operate in medium and large networks (per their IP allocation). Liquid (4.8%), Trellian (2.8%), and Google (2.7%) also host a significant number of websites, primarily in medium networks.

The other CSPs host fewer than 3% of the websites and are only associated with medium networks. The "Others" category represents ≈36.6% of the websites, dispersed across various network scales with a notable concentration in the medium networks at ≈30.8%. Further examination of the CSPs distribution in different countries is essential, as there may be regional variations that could impact their overall security.

Countries of Networks. Table 11 shows our results characterizing the distribution of the hosting countries for most free and premium websites across the various network scales. A significant fraction of these websites (≈84.2%) are hosted in medium networks, with the US hosting the majority (≈56.5%). Moreover, we found that the US hosts 58.7% of websites distributed across small, medium, large, and very large networks. Belgium and the Netherlands host a considerable number of websites, primarily in the medium networks, with ≈6.6% and 6.3% of the websites.

This analysis reveals a diverse hosting pattern across countries and networks, where most websites are hosted in medium networks. These findings emphasize the importance of examining the maturity of cybersecurity policies, as most malicious websites seem to rely on medium-scale networks for their operation. By focusing on these networks and understanding scale-level practices, it may be possible to combat malicious websites' proliferation effectively.

CSPs Over Countries. Table 12 demonstrates the distribution of the most commonly used CSPs across the top contributing countries. Among these CSPs, Cloudflare leads with a total of 410

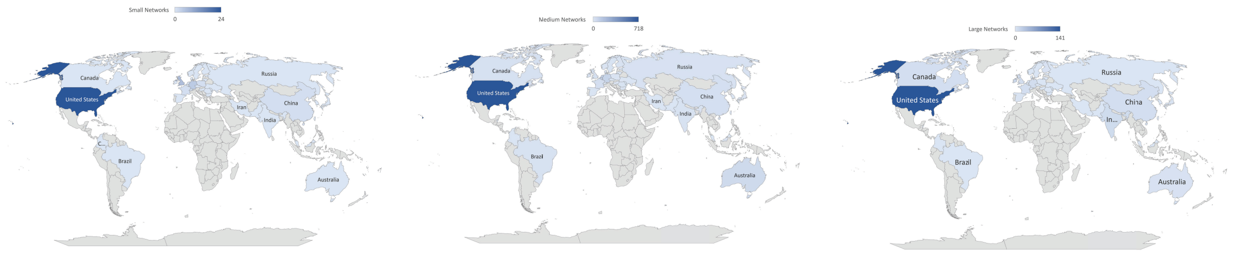
Table 12: The country-level distribution of the most commonly used CSPs. The features are the count per provider (#), and the count per country. The names are coded using Alpha-3.

Cloud Service Providers Distribution Over Countries												
CSP	#	USA	BEL	NLD	DEU	AUS	FRA	CHN	GBR	CAN	IRL	Others
Cloudflare	410	296	98	1	0	0	13	0	1	0	0	1
Amazon	240	191	0	0	0	3	1	0	5	1	20	19
Liquid	72	72	0	0	0	0	0	0	0	0	0	0
Trellian	42	0	0		0	42	0	0	0	0	0	0
Google	41	34	1	4	0	0	0	0	0	0	0	2
LeaseWeb	37	2	0	34	1	0	0	0	0	0	0	0
Sp-Team	35	0	0	0	35	0	0	0	0	0	0	0
Akamai	33	1	0	28	2	0	0	1	0	0	1	0
Fastly	26	12	0	1	1	2	0	0	0	0	0	10
Microsoft	21	16	0	0	0	0	0	0	2	1	1	1
Others	552	260	0	27	50	1	21	32	23	22	0	116
Total	1509	884	99	95	89	48	35	33	31	24	22	149
%	100	58.58	6.56	6.30	5.90	3.18	2.32	2.19	2.05	1.59	1.46	9.87

websites, primarily used in the United States (296 websites) and Belgium (98 websites). Amazon follows, with 240 websites, with the majority (191 websites, 79.6%) being hosted in the United States. Liquid Web and Trellian account for 72 and 42 websites, respectively, with Liquid Web exclusively used in the United States (72 websites) and Trellian exclusively utilized in Australia (42 websites).

Additionally, Google hosts 41 websites, with 34 in the United States, while LeaseWeb serves 37 websites, with a majority (34 websites) in the Netherlands. SP-Team has 35 websites, all in Germany. Akamai hosts 33 websites with 28 in the Netherlands, and Fastly accounts for 26 websites, primarily in the United States (12 websites). Microsoft hosts 21 websites, with 16 located in the United States. The "Others" category in Table 12 has 552 websites, whereas the US hosts 260 websites among them. In total, 1,509 websites were analyzed, and the US has the highest percentage of websites hosted at ≈58.6%, followed by Belgium, the Netherlands, and Germany.

Network Distribution Heatmaps. We generated heatmaps that display the distribution of network



(a) Small **(b)** Medium **(c)** Large

Figure 4: The spatial country-level distribution of small, medium, and large networks hosting free and premium content websites. The very large networks were only concentrated in the US.

scales among countries based on the data from Table 11. Figure 4a highlights the distribution of small-scale networks (SN column), revealing the United States as the primary host. In Figure 4b, the distribution of free and premium content websites within the medium networks (MN column) shows the United States, Belgium, and the Netherlands as the top hosting countries. Figure 4c illustrates the distribution of free and premium content websites within large-scale networks (LN column), where the United States, Germany, and France emerge as leading hosting countries.

The visualizations provide answers to **RQ1**, **RQ2** by giving insights into network scales' geographical distribution and emphasizing medium-scale networks' dominance in hosting free and premium websites, compatible with the derived results that reveal hosting patterns and networks' potential impact on websites' security and reliability. Finally, Medium-scale networks are identified as less secure or reliable than large-scale networks. Studying different types of medium-scale networks is necessary to better understand the most severe network. Implementing robust defensive measures against websites in medium networks, particularly those offering free content to users. Identifying the primary locations of malicious websites is crucial to improving online security.

4.4.4 Cloud Service Providers Analysis

As highlighted in §4.4.3, a CSP-level analysis provides better insight into the context of the free and premium content websites, especially for malicious websites. To this end, we dive deeper into this analysis by understanding the affinities between different categories of websites and the major cloud providers across our assessment metrics.

The distribution of free and premium content websites over different CSPs conveys different aspects as follows: 1. Most free content websites, premium content websites, and general websites use Cloudflare. Moreover, Cloudflare is reported to have the highest concentration of malicious websites among all CSPs for all three types of websites. 2. Amazon, although one of the largest providers, is the provider with the least concentration of malicious websites. Although one cannot point out conclusively a reason behind this behavior, one possible explanation is the measures taken by Amazon to curtail security risks in shared infrastructure, compared to more lax providers. 3. For the per-category websites analysis, free content websites mostly use Cloudflare, which is used by premium content websites only in the game category. In contrast, Amazon is the most used hosting provider for the rest of the categories in premium content websites. 4. Providers with the highest concentration of malicious websites reside in the US and Belgium. This can be attributed to using providers such as Cloudflare, which is mostly bound to those countries. 5. In general, there is a strong affinity between the state of a given website (malicious or benign) and the provider utilized by such a website.

Free and Premium Websites Comparison. The hosting pattern follows a heavy-tailed distribution. For both, we observed that the top eight providers (Cloudflare, Amazon, Liquid Web, LeaseWeb, SP-Team, Akamai International, Fastly, and Microsoft) host 63.42% of the websites, while the remaining websites are distributed across 290 providers, as partially shown in Table 14. It is worth noting that 80.59% of the malicious websites are hosted in these top CSPs.

The top five providers in terms of MPFP are Cloudflare, Liquid Web, LeaseWeb, SP-Team, and Trellian. Interestingly, Amazon, the second top provider in terms of hosting, has a relatively low MPFP (≈12.9%) and a smaller MP compared to other top providers. Fastly, which hosts 26 websites, has no malicious content websites. The "Others" category encompasses 552 websites (≈36.6%) and has an MPFP of ≈16.9% and an MP of 6.16%. The analysis highlights the varying security levels of different CSPs.

Although the most popular hosting providers among free and premium content websites, Cloudflare and Amazon, differ significantly regarding their MPFP and MP. Cloudflare, the top provider, contributes ≈68.5% of MPFP and ≈18.6% of the total MP of the reaming ≈31.7%, being also the top provider in terms of both features. On the other hand, the second provider in terms of MP is Liquid Web which contributes only ≈2.1% of the total malicious websites. This clear distinction in security levels suggests that a low MP may be related to a low percentage of hosted websites. At the same time, the significant difference in MPFP indicates the security of individual CSPs.

Benchmark Websites. Upon closer examination of the tables, it becomes evident that the data presented in Table 13 in comparison with Table 14 we can provide a unified insight based on the information from Table 14 and either Table 13.

First, Cloudflare and Amazon emerge as the most popular CSPs across all providers. Cloudflare hosts the highest number of websites and has the highest MC count. Per Table 14, Cloudflare hosts ≈27.2% of the websites, while in the other table, it hosts ≈16.4%. Amazon, ranking second, hosts ≈15.9% of websites in the first table and around ≈10.9% in the other table. Liquid Web consistently has the second-highest MPFP. In the first table, it ranks third and hosts ≈4.8% of websites, while in the other table, it ranks sixth and hosts ≈1.9% of websites. Fastly is unique because it hosts no MC websites in the first table. However, in the benchmark results, Fastly has an MPFP of ≈4.2% and an MP of only 0.05%. The "Others" category represents a considerable portion of websites in all tables, ranging from ≈36.6% to ≈53.5% in the other two tables.

In conclusion, Cloudflare and Amazon are the most popular CSPs among free and premium content websites, with Cloudflare having the highest number of MC websites and the highest MPFP. Liquid Web ranks second in terms of MPFP. Fastly stands out as a provider without any MC websites in its hosting, as shown in the first table.

Free Websites. Analyzing the distribution of free content websites across various CSPs as shown in Table 13, we observe that Cloudflare dominates the market by hosting ≈33.8% of free content websites, with ≈64.3% of its hosted websites identified as malicious, resulting in 21.7% MP. Liquid Web and Amazon are the second and third most popular CSPs, respectively, hosting 8.5% and ≈6.9% of free content websites. Liquid Web has an MP of ≈4%, while Amazon's MP stands at 1.9%. CSPs like Trellian, LeaseWeb, and Sp-Team each host ≈5% of the free content websites

Table 13: An overview of the distribution of the (top-1M, free content websites, and premium content websites) across different cloud service providers.

General Websites

CSP	#	%	MC	MPFP	MP
Cloudflare	337	16.38	14	4.15	0.68
Amazon	224	10.86	5	2.33	0.24
Google	95	4.62	0	0	0
OVH	72	3.50	4	5.56	0.19
Hetzner Online	56	2.72	3	5.36	0.15
Microsoft	42	2.04	1	2.38	0.05
Liquid Web	39	1.90	9	23.08	0.44
Automattic	36	1.75	0	0	0
Alibaba	29	1.41	1	3.45	0.05
Digitalocean	27	1.31	2	7.41	0.10
Others	1100	53.48	53	4.82	2.58
Total	2057	100	92	4.47	4.47

free content websites

CSP	#	%	MC	MPFP	MP
Cloudflare	266	33.76	171	64.29	21.70
Liquid Web	67	8.50	32	47.76	4.06
Amazon	54	6.85	15	27.78	1.90
Trellian	42	5.33	10	23.81	1.27
LeaseWeb	36	4.57	10	27.78	1.27
Sp-Team	35	4.44	10	28.57	1.27
Bodis	17	2.16	4	23.53	0.51
SEDO GmbH	13	1.65	2	15.38	0.25
OVH	11	1.40	4	36.36	0.51
Google	11	1.40	4	36.36	0.51
Others	236	29.95	57	24.15	7.23
Total	788	100	319	40.48	40.48

premium content websites

CSP	#	%	MC	MPFP	MP
Amazon	186	25.80	16	8.60	2.22
Cloudflare	144	19.97	110	76.39	15.26
Akamai	32	4.44	2	6.25	0.28
Google	30	4.16	3	10	0.42
Fastly	23		0	0	0
Microsoft	18	2.50	2	11.11	0.28
Sp-Shopify	12	1.66	10	83.33	1.39
Ebay	8	1.11	0	0	0
Wal-Mart	8	1.11	1	12.50	0.14
Ovh	7	0.97	1	14.29	0.14
Others	253	35.09	15	5.93	2.08
Total	721	100	160	22.19	22.19

and exhibit similar MC and MP values. Notably, the "Others" category, encompassing a variety of CSPs, hosts ≈30% of free content websites and presents an MP of ≈7.2%. With a total of 788 free content websites, 319 (≈40.5%) were malicious.

Premium Websites. In Table 13, Amazon emerges as the most prominent host, accommodating 25.8% of the total premium websites. Among the websites hosted by Amazon, 8.6% are malicious, resulting in an overall MP of ≈2.2%. Cloudflare ranks as the second-largest host with ≈20% of premium websites, with a higher proportion (≈76.4%) of malicious websites, leading to an MP of ≈15.3%. Other notable CSPs include Akamai, Google, Fastly, and Microsoft, hosting around 2% to 4% of premium content websites. Regarding malicious content, Google and Microsoft exhibit MPs of ≈0.4% and ≈0.3%, respectively, while Akamai has a lower MP of ≈0.3%. Fastly and Ebay host ≈3.2% and ≈1.1% of premium content websites, respectively, but neither has any malicious content. Interestingly, Sp-Shopify hosts only ≈1.7% of premium content websites but has a high proportion (≈83.3%) of malicious websites, resulting in an MP of ≈1.4%. Wal-Mart and OVH each host about 1% of premium content websites and have MPs of ≈0.1%. Lastly, the "Others" category, which includes a variety of CSPs, hosts ≈35.1% of the total premium content websites. With only ≈6% of its hosted websites classified as malicious, the category exhibits an MP of ≈2.1%. The table shows 721 premium content websites, with 160 (≈22.2%) being malicious.

Free versus Premium Websites. Upon comparing the distribution of free and premium content websites across different CSPs using Table 13, several insights are drawn. First, we found that Cloudflare is the most prominent hosting cloud for free content websites, hosting ≈33.8% of the total free content websites, while Amazon is the most prominent host for premium content websites, with 25.8%. Interestingly, the MP of Cloudflare is higher for premium content websites (≈15.3%) compared to free content websites (≈21.7%), indicating that Cloudflare hosts a larger proportion of malicious premium content websites than free content websites. Conversely, Amazon has a higher MP for free content websites (1.9%) than premium content websites (≈2.2%), suggesting that it hosts proportionally more malicious free content websites than premium content websites. Google has a relatively low MP for both free content websites (≈0.5%) and premium content websites (≈0.4%), implying that it hosts a smaller proportion of malicious websites than other CSPs. The total number of websites is higher for free content websites (788) than for premium content websites (721), with 41% and ≈22.2% being malicious, respectively, indicating that free content websites have a higher overall prevalence of malicious content than premium content websites—some CSPs, e.g., Liquid Web, Trellian, LeaseWeb, and Sp-Team, host only free content websites. In contrast, others like Akamai, Fastly, Microsoft, Sp-Shopify, eBay, and Wal-Mart only host premium content websites, suggesting that different CSPs may have different preferences when hosting free content websites or premium content websites or affinities in those types of websites for selecting a specific provider.

4.4.5 Per-Category Cloud Service Providers Analysis

In the following, we outline the per-category results of the CSP analysis, for specific trends.

Book Websites. Table 14 shows the distribution of free and premium content websites for books across CSPs. In free content websites, Cloudflare hosts the majority, with 39 websites accounting for ≈27% of the total. Amazon follows it with 11 websites (≈7.6%), Liquid Web with ten websites (≈7%), Trellian and Sp-Team with 6 and 5 websites, respectively, and others collectively hosting 73 websites (≈50.7%). For premium content websites, Amazon tops the list with 41 websites (≈21.5%), closely followed by Cloudflare with 40 websites (≈21%). Other CSPs in this category include Google, Sp-Shopify, Fastly, and others, with varying counts and percentages. Regarding the MC, Cloudflare dominates in both free and premium content websites with 28 and 32 instances, respectively. The MPFP is highest for Cloudflare among free content websites (≈71.8%) and Sp-Shopify among premium content websites (≈75%). The MP is fairly distributed between Cloudflare (≈19.4% for free websites and ≈16.8% for premium websites) and other CSPs.

Games Websites. Table 14 presents the distribution of free and premium Games Websites across different CSPs. For free websites, Cloudflare is the dominant CSP, hosting 42 websites, which account for 53.85% of the total. Other CSPs in this category include Mivocloud with 5 websites (≈6.4%), LeaseWeb and Liquid Web with 3 websites each (≈3.9% each), Amazon with two websites (≈2.6%), and others collectively hosting 23 websites (≈30%). For premium content websites, Cloudflare is the leading CSP, hosting 37 websites (≈33.3%). Amazon follows it with 22 websites (≈20%), Akamai with 11 websites (≈10%), Fastly with 5 websites (4.50%), Google with 4 websites (3.6%), others hosting 32 websites (≈28.8%). Considering the MC aspect, Cloudflare has the highest count for both free and premium websites, with 39 and 29 instances, respectively. The MPFP for Liquid Web is highest among free websites at 100%, while Cloudflare leads among premium websites at ≈78.4%. The MP is distributed between various CSPs, with Cloudflare accounting for 50% in free websites and ≈26.1% in premium websites.

Table 14: An overview of the distribution per category (combined, books, and games) across different cloud service providers.

Combined

CSP	#	%	MC	MPFP	MP
Cloudflare	410	27.17	281	68.54	18.62
Amazon	240	15.90	31	12.92	2.05
Liquid Web	72	4.77	32	44.44	2.12
Trellian	42	2.78	10	23.81	0.66
Google	41	2.72	7	17.07	0.46
LeaseWeb	37	2.45	10	27.03	0.66
Sp-Team	35	2.32	10	28.57	0.66
Akamai	33	2.19	2	6.06	0.13
Fastly	26	1.72	0	0	0
Microsoft	21	1.39	3	14.29	0.20
Ovh	18	1.19	5	27.78	0.33
Bodis	17	1.13	4	23.53	0.27
Linode	13	0.86	5	38.46	0.33
SEDO GmbH	13	0.86	2	15.38	0.13
Others	491	32.54	77	15.68	5.10
Total	1509	100	479	31.74	31.74

Books

Free Content Websites					
CSP	#	%	MC	MPFP	MP
Cloudflare	39	27.08	28	71.79	19.44
Amazon	11	7.64	1	9.09	0.69
Liquid Web	10	6.94	3	30	2.08
Trellian	6	4.17	0	0	0
Sp-Team	5	3.47	0	0	0
Others	73	50.69	6	8.22	4.17
Total	144	100	43	29.86	29.86
Premium Content Websites					
Amazon	41	21.47	3	7.32	1.57
Cloudflare	40	20.94	32	80	16.75
Google	9	4.71	1	11.11	0.52
Sp-Shopify	8	4.19	6	75	3.14
Fastly	5	2.62	0	0	0
Others	88	46.07	64	72.73	33.51
Total	191	100	53	27.75	27.75

Games

Free Content Websites					
CSP	#	%	MC	MPFP	MP
Cloudflare	42	53.85	39	92.86	50
Mivocloud	5	6.41	0	0	0
LeaseWeb	3	3.85	2	66.67	2.56
Liquid Web	3	3.85	3	100	3.85
Amazon	2	2.56	1	50	1.28
Others	23	29.49	5	21.74	6.41
Total	78	100	50	64.10	64.10
Premium Content Websites					
Cloudflare	37	33.33	29	78.38	26.13
Amazon	22	19.82	3	13.64	2.70
Akamai	11	9.91	0	0	0
Fastly	5	4.50	0	0	0
Google	4	3.60	0	0	0
Others	32	28.83	3	9.38	2.70
Total	111	100	35	31.53	31.53

Movie Websites. Table 15 shows the distribution of free and premium Movies Content Websites across different CSPs. For free content websites, Cloudflare is the leading CSP, followed by Liquid Web (≈11.6%), Trellian (≈9.7%), Sp-Team with (≈7.7%), Amazon (≈6.1%), and all others hosting 118 websites (≈38.1%). Regarding premium websites, Amazon leads with 56 websites (≈36.8%), followed by Cloudflare Inc with 18 websites (≈11.8%), Akamai with nine websites (≈5.9%), Google with seven websites (≈4.6%), Fastly with five websites (≈3.3%), and others hosting 57 websites (37.5%). In terms of the MC, the largest count in free content websites is observed with Cloudflare (17) and in premium content websites with Cloudflare Inc (15). The MPFP highlights Sp-Team as the highest in free content websites with 37.5%, while Cloudflare Inc tops premium content websites with ≈83.3%. The MP is distributed among various CSPs: Cloudflare has ≈5.5% in free content websites, and Cloudflare Inc has ≈9.9% in premium content websites.

Music Websites. Table 15 presents the distribution of free and premium Music Content Websites across different CSPs. For free content websites, Cloudflare is the dominant CSP, hosting 22 websites (27.5% of the total). Sp-Team follows with six websites (7.5%), then Google with four websites (5%), Amazon with three websites (≈3.8%), Liquid Web with two websites (2.5%), and the Others category with 43 websites (≈53.8%). In the case of premium content websites, Amazon leads with 30 websites (≈35%), followed by Cloudflare with 12 websites (≈14%), Fastly and Google each with 4 websites (≈4.7%), Apple with 2 websites (≈2.3%), and Others with 34 websites (≈39.5%). Regarding MC, Cloudflare has the highest count for both free content websites (16) and premium content websites (9). In terms of the MPFP, Liquid Web has the highest percentage in free content websites with 100%, while Cloudflare takes the lead in premium content websites with 75%. The MP is distributed among various CSPs: Cloudflare accounts for 20% in free content websites, and in premium content websites, Cloudflare accounts for ≈10.5%.

Table 15: An overview of the distribution per category (movies, music, and software) across different cloud service providers.

Movies

Free Content Websites					
CSP	#	%	MC	MPFP	MP
Cloudflare	83	26.77	17	20.48	5.48
Liquid Web	36	11.61	11	30.56	3.55
Trellian	30	9.68	8	26.67	2.58
Sp-Team	24	7.74	9	37.50	2.90
Amazon	19	6.13	5	26.32	1.61
Others	118	38.06	32	27.12	10.32
Total	310	100	82	26.45	26.45
Premium Content Websites					
Amazon	56	36.84	1	1.79	0.66
Cloudflare Inc	18	11.84	15	83.33	9.87
Akamai	9	5.92	1	11.11	0.66
Google	7	4.61	2	28.57	1.32
Fastly	5	3.29	0	0	0
Others	57	37.50	4	7.02	2.63
Total	152	100	23	15.13	15.13

Music

Free Content Websites					
CSP	#	%	MC	MPFP	MP
Cloudflare	22	27.50	16	72.73	20
Sp-Team	6	7.50	1	16.67	1.25
Google	4	5	1	25	1.25
Amazon	3	3.75	1	33.33	1.25
Liquid Web	2	2.50	2	100	2.50
Others	43	53.75	10	23.26	12.50
Total	80	100	31	38.75	38.75
Premium Content Websites					
Amazon	30	34.88	3	10	3.49
Cloudflare	12	13.95	9	75	10.47
Fastly	4	4.65	0	0	0
Google	4	4.65	0	0	0
Apple	2	2.33	0	0	0
Others	34	39.53	3	8.82	3.49
Total	86	100	15	17.44	17.44

Software

Free Content Websites					
CSP	#	%	MC	MPFP	MP
Cloudflare	80	45.45	71	88.75	40.34
Amazon	19	10.80	7	36.84	3.98
Liquid Web	16	9.09	13	81.25	7.39
LeaseWeb	11	6.25	3	27.27	1.70
Voxility LLP	4	2.27	0	0	0
Others	46	26.14	19	41.30	10.80
Total	176	100	113	64.20	64.20
Premium Content Websites					
Amazon	37	20.44	6	16.22	3.31
Cloudflare	37	20.44	25	67.57	13.81
Microsoft	9	4.97	0	0	0
Akamai	6	3.31	0	0	0
Google	6	3.31	0	0	0
Others	86	47.51	3	3.49	1.66
Total	181	100	34	18.78	18.78

Software Websites. Table 15 shows the distribution of free and premium software websites across various CSPs. In the case of free content websites, Cloudflare is the leading CSP with 80 websites (≈45.5%), followed by Amazon with 19 websites (10.8%), Liquid Web with 16 websites (≈9.1%), LeaseWeb with 11 websites (≈6.3%), Voxility LLP with 4 websites (≈2.3%), and Others with 46 websites (≈26.1%). On the other hand, for premium content websites, Amazon and Cloudflare are the most prominent CSPs, each hosting 37 websites (≈20.4%), followed by Microsoft with 9 websites (≈5%), Akamai and Google each with 6 websites (≈3.3%), and Others with 86 websites (≈47.5%). In terms of MC, Cloudflare has the highest count for free content websites (71) and the second-highest for premium content websites (25). For free content websites, the highest MPFP is found in Cloudflare (≈88.8%), while for premium content websites, it is also in Cloudflare with 67.57%. Regarding MP, Cloudflare has the highest percentage in free content websites at ≈40.3% and the second-highest in premium content websites at ≈13.8%.

4.5 Discussion

4.5.1 Main Takeaways

To summarize our findings, the results of the network-scale distribution show interesting trends between the benchmark dataset and the free and premium websites dataset as follows.

Network Distribution. 1. free and premium content websites are predominantly concentrated in medium-scale networks across all categories. 2. Malicious websites show a strong correlation with medium-scale networks. Premium content websites use more large-scale networks than free content websites, indicating a relationship between reliable networks and residents in large networks. 3. Using large networks requires higher security standards than medium or small ones. 4. The ranking of category websites based on their average MP is (1) games (47.82%), (2) software (41.49%), (3) books (28.81%), (4) music (28.1%), and (5) movies (20.79%), with an overall average MP of 31.34%. 5. The same ranking applies to the average distribution over medium-scale networks, providing answers for**RQ1** and **RQ2**. 6. A similar conclusion can be applied to distributing malicious content websites across different network scales. Isolating this network scale might be ineffective, as numerous legitimate websites also reside within the same scale. 7. It is essential to break the medium-scale into different types and study the distribution characteristics of free and premium content websites on these different network scales to find the most severe network pattern.

Spatial Distribution. The main takeaways are as follows. 1. The results of the spatial analysis convey answers to **RQ3**. We found the most contributing hosting countries: small and large networks were primarily used in the US, while the other countries mainly used medium networks. 2. We found that more than half of the top CSPs (58.58%) reside in the US, while the rest are heavy–tailed distributed over the top hosting countries. 3. Examining the distribution of hosting CSPs and determining the primary locations of malicious websites is crucial to identifying underlying issues and implementing appropriate preventative measures.

free content websites & premium content websites versus the Benchmark. Comparing the combined datasets of free and premium content websites against the benchmark datasets provides an understanding of the studied **RQ4**. The takeaways of this study are as follows. 1. Where Where we notice a similarity in the network scale distribution results, there are significant differences in the results of the top hosting CSPs distribution. 2. The CSPS distribution analysis reveals that Cloudflare and Amazon are the most popular CSPs for hosting free content websites, premium content websites, and benchmark websites. 3. While the free and premium content websites reported being higher malicious than the top 1 million websites, where in detail, the higher rate of the malicious websites in the combined datasets is due to the free content websites and mostly those residing in the top used CSPs. 4. We observed that the free and premium content websites are heavy-tailed distributed over the top hosting CSPs. The case is the same for the top 1 million websites, where in free and premium content websites, the most benign websites associated with hosting CSPs are mostly found in premium content websites. 5. Noteworthy, Liquid Web was found to host the most malicious websites with the highest MPFP in the benchmark, with a significant difference between the first and the second CSP 23.08%, 7.41% respectively, while it is considered the second-highest MPFP CSP in the combined results.

CSPs Distribution. The results of provide sufficient answers for **RQ5** the CSPs distribution uncover the affinities of free and premium content websites hosting patterns. 1. There is a higher prevalence of malicious content among free content websites than premium content websites. Some CSPs are unique to either free content websites or premium content websites, indicating that different CSPs may have preferences or specializations when hosting these types of content. 2. Amazon generally has a lower MPFP than other heavily used CSPs in free and premium content websites, making it a good example of secure CSPs. 3. Cloudflare hosts the highest number of malicious content websites and exhibits the highest MPFP among free and premium content websites. 4. Liquid Web ranked second in terms of MPFP, while Fastly did not host any malicious content websites in one of the tables. 5. Furthermore, a similar conclusion can be made that the most malicious websites hosting CSPs cannot be isolated due to the high overlap between the malicious and benign websites on these CSPs. 6. Thus, the distribution of the most contributed countries needs to be investigated, especially for countries with the most malicious websites hosting CSPs: The US, Belgium, Netherlands, Germany, and Australia.

4.5.2 Limitations of the Study

The limitations of this study are related to the used datasets. First, the dataset is a snapshot of free and premium content websites at a particular time, which may not represent all malicious and benign websites associated with premium content websites and free content websites. Second, errors might have been introduced during the manual annotation of data. Moreover, in this study, only VirusTotal was used to scan the malicious websites, limited by the representation of scanners. Moreover, since free content websites hosting patterns change rapidly and depend on CSPS' policies and regulations, the results may not be applicable in other contexts or periods. Additionally, the study briefly discussed other factors affecting website security, such as privacy policies, due to space constraints, and further research is needed to address them. To address the distribution of services among different companies, the study proposed a solution that combines all the various entities under the umbrella of one major company.

For instance, Amazon CSPs provide their services on a regional basis, such as Amazon Data Services Canada, Amazon Data Services France, and Amazon Prod. Consequently, all these CSPs were aggregated into one entity known as "Amazon" for further analysis regarding their service distribution. Nonetheless, it is imperative that further investigations need to be conducted to ensure security concerning each distributed service. Finally, although correlation analysis was performed to understand how different hosting patterns interact for free content website and premium content website classification tasks, more advanced machine-learning models should be applied to achieve better accuracy rates when classifying malicious domains.

4.5.3 Recommendations

Consistent with the literature [10–14], our work found free content websites are more malicious and vulnerable than premium content websites in all comparison aspects. Similar to other works that studied the security of the top-used websites [22, 46, 50, 52, 66], our contrast analysis shows that the top million used websites are, in fact, less malicious than free content websites which implies that they might be more vulnerable to any security breaches even if they are benign free content websites. Furthermore, previous work analyzed the security factors of the CSPs and the networks [31, 34, 38, 39, 43–45, 51, 55, 56, 61, 62, 67, 71, 74, 80], and suggested different techniques that can be used to protect the different network applications with affinities to certain CSPs, which is echoed in our work that points out CSPs that are used to host malicious free content websites.

The study results suggest that network administrators should implement more stringent security measures to protect their networks from malicious activities. Specifically, organizations should focus on isolating medium-scale networks, as they are frequently associated with malicious content websites. Additionally, investigating the CSPs used by either free content websites or premium content websites can help identify which CSPs host more malicious websites than benign ones, and legal action should be taken if necessary. Further security annotation needs to be investigated by combining more than VirusTotal scanners, such as Google Safe Browsing, PhishTank, or other scanners, to get the most accurate malicious classification.

Further research is necessary to understand the relationship between hosting patterns and malicious content websites. For instance, future studies could analyze other factors, such as website age or domain registration date, that may influence website classification. Moreover, analyzing the dynamic code of free content websites can help evaluate the severity of their vulnerabilities. This can improve classification accuracy and our understanding of these websites' functions.

In addition to the previous recommendations, researchers should explore alternative methods for detecting malicious activity with medium-scale networks to improve overall security across all internet domains. While the study results suggest that most websites reside within medium-scale networks, more research is necessary to accurately determine which size of the medium-scale network poses the most significant security risk. To achieve this, we suggest dividing medium networks into different levels and assessing the security of websites within each level.

Such analysis enables organizations to target their security defenses towards networks with a heightened vulnerability only, leading to better risk management. Finally, it is essential to investigate how attackers exploit free services hosted by trusted providers to detect and prevent attacks before they occur, minimizing their duration. Understanding the attackers' methods to exploit vulnerabilities within services, organizations can proactively take steps to improve their defenses, preventing attacks from happening in the first place or reducing their impact if they do occur.

4.6 Summary and Work to be Completed

Our results show that free and premium websites are concentrated in medium networks, similar to malicious websites, implying that isolating this type of network alone may not be an effective solution. Furthermore, we identified Cloudflare 68.87%, Liquid Web 44.44%, LeaseWeb 29.41%, SP-Team 28.57%, and Trellian 23.81% as the most used CSPs with high overlap between malicious and benign websites located within these CSPs, indicating a need for further investigation into their distribution and potential weaknesses in security protocols or policies within those countries.

Further future work is needed to investigate if there are any changes to the free and premium content websites distribution over time and whether these changes have a specific seasonal or periodic pattern. Identify the most effective strategies for containing and limiting the spread of malicious free content websites, and how these strategies can differ based on the network scale, CSP, or the hosting country. Nevertheless, comparing the distribution and the hosting patterns of free content websites to other types of cyber threats, such as phishing, scams, or ransomware attacks, is needed to recognize if there are any commonalities or differences in their distribution.

This study highlights the need for continued efforts to improve the security of free content websites. In the future, it would be interesting to investigate the vulnerability enumeration of free content websites to increase the users' awareness by identifying vulnerable points within free content website's infrastructure before attackers can exploit them.

5 Understanding the Country-Level Security of Free Content Websites and their Hosting Infrastructure.

5.1 Summary of Completed Work

This work examines the distribution of free and premium content websites around the world in relation to malicious websites. Finding correlations between malicious website hosting and the National Cyber Security Index (NCSI). Despite the United States being a dominant host, it has a low NCSI, largely due to its privacy policy criteria. Similar trends have been observed in other countries. By mapping free and premium content websites by country and noting the prevalence of malicious websites, this research aims to underscore regional hosting vulnerabilities and inspire policy shifts to mitigate cyber threats.

5.2 Introduction

Free content websites provide free content to their users, including books, games, music, movies, and software. The same type of content can be provided to the user at a cost in premium content websites. In previous work [13], free content website privacy policies were shown to be less elaborate, unlike premium content websites, where free content websites reuse policies depending on their hosting providers. These hosting providers usually reside in one or more countries where users can access free and premium content websites.

It is important to assess the country-level (geo-distribution) security features of free and premium content websites to understand their ecosystem and provide the appropriate security recommendations. Moreover, as the characteristics of those websites may differ based on the type of content they provide, a per-category analysis is paramount for deeper contrasts. Finally, as the target of analysis at the country level, it is essential to understand the security of such websites and country security matureness—e.g., measured by the National Cyber Security Index (NCSI).

Why Study the Geographical Distribution. By investigating the geo-distribution of free content websites, we can design better strategies to ensure network security based on the country's regulations and cyber security policies. Understanding the distribution of free content websites over countries is necessary to identify the concentration of malicious websites and infrastructures. As a result, it will help users of these services protect themselves from being victimized by a vulnerability that cannot be controlled or governed by the law of the same country. For example, if a network user uses a free content website that resides in a different country and the website victimizes the user, the user would know if the law of the free content website country is strict and elaborated so that the user can perform legal action against that free content website. They can legally remove any acquisitions, definitions, misinformation, or malicious content from the free content website.

Also, it will help to determine the necessary actions against the hosting providers that contain the most malicious websites.

Why Study the Cyber Security Policy. To understand the factors that improve security at the country level, we investigate the maturity of cyber security policies of the countries where most free content websites are hosted. Security agreements could exist between countries with the least malicious websites, indicating the effectiveness of such policies and agreements. Investigating the maturity of the cyber security policies of the countries would reveal if the countries with the highest concentration of malicious free websites are those with the least mature cyber security policies. Knowing the NCSI for each country will help us find if there is any correlation between the actual security landscape measured by the prevalence of those websites and the security policies.

Our results show certain country-level distribution patterns for free and premium content websites, where most of the malicious websites are heavily concentrated in some countries. Similar conclusions are derived from the results of the general websites with some differences in concentration where the NCSI average shows a unique pattern for the top hosting countries, where we notice lower scores in some digital development aspects in the countries that contribute to hosting a higher rate of the malicious free content websites, where we notice the significant difference in the distribution of the category websites analysis results between every category and between free and premium content websites in general.

Contributions. Using 1,562 free and premium content websites [10, 13, 14], we contribute the following. 1. *Malicious free content websites Measurements.* We identify the malicious free and premium content websites and analyze their connections to various infrastructure entities and characteristics. More specifically, we revealed the countries contributing most to (malicious) free content website hosting. 2. *Comparative Analysis.* A thorough comparison was made between free and premium content websites in terms of their utilization of infrastructure and security features in every hosting country. Furthermore, the top hosting countries for Alexa's one million websites were compared to the free and premium content websites and maliciousness within each hosting entity. This yielded a precise, inclusive, and contextualized description of free content websites when put next to premium content websites. 3. *Per-category Analysis.* We performed a comprehensive analysis of the most contributed hosting countries for the different content free and premium content websites categories. We give a detailed comparison in every content type between free and premium content websites, describing the affinities for every studied category. 4. *Security Policy Impact.* We investigated the correlation analysis that sheds light on the average NCSI score of the top free and premium content websites hosting countries to identify the role of the privacy policy development on the percentage of the malicious hosted free content websites in that country.

5.3 Methodology

5.3.1 Research Questions

This work aims to understand how free content websites are hosted for their geographic locations and the divergence from premium content websites and general website populations. To accomplish our goal, we endeavor to address several questions. **RQ1**. Where are malicious websites mainly concentrated, and what correlations exist regarding their geographical locations? **RQ2**. Are there any similarities or differences between free content websites, premium content websites, and the general websites in their geographical distribution regarding the malicious websites? **RQ3**. What are the main geographical distribution characteristics of the different content category websites? **RQ4**. Are there any inconsistencies between top countries hosting malicious websites and the national cyber security index (NCSI) of those countries?

5.3.2 Dataset and Data Collection

We use several datasets: 1. A primary dataset consisting of free content websites, premium content websites, and their corresponding annotations. 2. A dataset for the general website population to facilitate our contrast evaluation between free content websites, premium content websites, and their utilization of infrastructure. 3. The results of network features extraction using ipdata [2] and IPSHU [7] in order to obtain the country-level annotation for every website in the scanned dataset. 4. The results of malicious annotation websites by scanning all websites datasets using VirusTotal [6]. In the following, we will review those datasets and how we obtained them.

5.3.3 Free and Premium Content Websites Dataset

Our study employs a dataset of 1,562 websites compiled as per the criteria established in the prior works [11–14]. When determining whether to add a website to this list, the primary considerations are its level of popularity, the language used on the site, and how active it is. To assess *popularity*, upon entering a keyword, the ranking of each website on major search engines is assessed. Websites that utilize English as their primary language are kept for additional examination. Meanwhile, *activity* is determined by verifying that the website returned from the search engine is online (active) at the time of the evaluation.

Similarly, three search engines—Google, DuckDuckGo, and Bing—are employed to estimate website popularity, where the average rank of the returned website is taken in making this estimation. Manual inspection is employed in determining whether a website is an free content website or premium content website. Moreover, each webpage is assigned a category based on the content it features—books, games, movies, music, and software. Finally, the classification of websites and other relevant keywords (e.g., free, premium, paid, etc.) are used for searching in the related search engine. Upon obtaining the filtered websites from the previous steps, we initiated a query to their domain names to acquire their hosts' IP addresses. We found that 1,509 websites were available, which represents 96.6% of the total websites we queried. Among these active websites, 788 were free content websites, and 721 were premium content websites. Based on the content type, we categorize the free and premium websites into five categories: books (144 free and 191 premium), games (78 free and 111 premium), movies (310 free and 152 premium), music (80 free and 86 premium) and software (176 free and 181 premium).

5.3.4 General Websites Sample

While measuring and characterizing the free and premium content websites in isolation may shed light on their characteristics, contrasting them with general websites sample can put them into perspective. To this end, we collect a benchmark dataset that is representative of the general websites population. To obtain an unbiased random sample, 2,400 websites were drawn from Alexa's Top One Million website dataset [37]. The sample size is selected to be comparable to the total number of websites in the free and premium content websites while ensuring a small error and high confidence when considering the mean estimation of the overall population of the websites. In particular, we used a margin of error of 2% and a confidence interval of 95%, resulting in a sample size of 2,400. We note that the size of the population (1 million in this case) has an insignificant effect on the size. We refer to this dataset as the "general" for simplicity in presenting.

As in the preprocessing and augmentation of free and premium content websites, we consider whether a general website is active or not—i.e., online or offline at the time of the data acquisition. We found that only 2,057 websites were active, which accounts for only 85.7% compared to 96.5% for the final dataset of free and premium content websites. We obtained each sample's IP address and hosting countries with the *ipdata* API in the free and premium content websites.

5.3.5 Malicious Websites Annotation

The main goal of this work is to assess the regional concentration of malicious free content websites in comparison to the premium content websites and the general websites. To begin, we took advantage of VirusTotal [6], a service that combines more than 70 scanning engines and can be used to classify whether a domain name (URL), IP address, or binary (file)—identified by its unique hash value—is malicious or benign. Upon passing a file to VirusTotal, it returns a list of antivirus scanners and their associated detections. Our datasets of the different types of websites are further enhanced using the annotation provided by VirusTotal, where we consider a website to be malicious if at least one of the returned results by VirusTotal is malicious and benign otherwise.

5.3.6 The National Cyber Security Index

The national cyber security index (NCSI) [29] is provided by the e-governance academy and identifies a rating of countries based on 12 metrics: 1. cyber security policy of that country, 2. identified and analyzed security threats, 3. education and professional development, 4. contribution to global cyber security, 5. protection of digital services, 6. protection of essential services, 7. electronic identification and digital signature, 8. protection of personal data, 9. cyber incident response, 10. cyber crisis management, 11. cyber crimes fighting, and 12. military cyber operation

We use NCSI to understand the role of cyber security policy and its association with favorable security outcomes, such as the lack of malicious websites in a given country. We hypothesize that countries with a high NCSI would have a low percentage of such malicious websites, and we examine this hypothesis through correlation analysis. Our justification is that countries with the most malicious free content websites have less mature cyber security policies or may not be aware of the latest cyber threats. This is consistent with the rationale of developing NCSI based on objective qualities of the cyber security posture at the country level, as those countries with a lower rating in NCSI might not be digitally well developed to analyze the recent threats and implement the analysis outcomes into defenses for taking down such websites, or may not protect digital services by applying a high-quality standard and providing a competent supervisory authority that tracks private or public digital services on both free and premium content websites.

To sum up, low-ranked countries may: 1. have less mature cyber security policies, 2. lack of awareness of the cyber security threats, 3. not using recent technology to identify and analyze the risk, 4. provide less protection of digital services, and 5. provide low protection of personal data.

5.3.7 Analysis Dimensions

We conduct analyses to detect patterns and disparities between free content websites, premium content websites, and general websites across six categories: the country of origin, the number of websites (count), the proportion of the studied websites per each country (percentage), the total maliciousness contribution (malicious count and malicious percentage), and the maliciousness of the hosted websites per feature count and percentage. Each studied dimension is defined below, and the workflow of our analysis pipeline is shown in Figure 5.

Country. The country where the IP-based infrastructure of free content websites, premium content websites, and general websites is located. Our investigation discovered that this attribute has 41 different values that correspond to as many countries.

Count. The total number of websites hosted on an IP address that is located in that particular country and associated with the given class of websites.

Percentage. The number of websites (free content websites, premium content websites, general) in a country normalized by the total number of investigated websites for that type of the studied

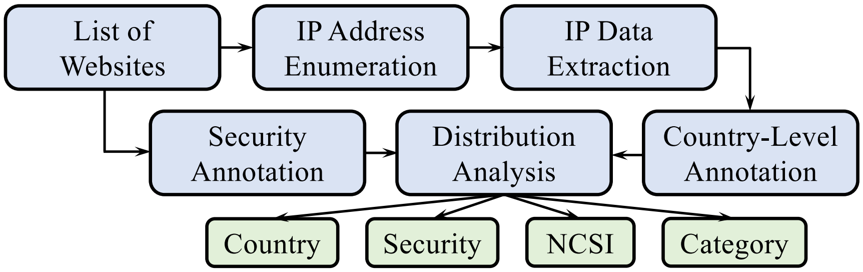


Figure 5: Feature extraction and data enumeration workflow, along with comprehensive distribution analysis leading to the free content websites country-level analysis results.

class of websites. This feature is employed to comprehend any differences in website distribution.

Malicious Count (MC). A count of the discovered malicious websites by VirusTotal as described in section 5.3.5 that are hosted within a specific country.

Malicious Per Country Percentage (MPCP). The proportion of malicious websites relative to the number of websites in a given country. This feature emphasizes the contribution of the studied entity to the total number of malicious websites by considering their size within our dataset. As opposed to MC, which gauges an entity's overall contribution to the total malicious hosting contribution according to our analysis, MPCP normalizes this value by considering how many potentially malicious websites reside within a particular country, acknowledging that countries may vary significantly in their size (scale).

Malicious Percentage (MP). This feature indicates the ratio of MC to all websites in the country under analysis for that particular sample, meaning the total number of malicious websites in free content websites, premium content websites, both, or general websites. Unlike MC's indication, MPCP implies that even large entities may contribute little to the total number of malicious websites when their size is considered.

The National Cyber Security Index. Provided by NCSI, we studied the following features that determine the high malicious percentage and the weaknesses in NCSI scores: 1. **country**, which is the name of the country reported being one of the most hosting free content website/premium content website countries, 2. **count**, the number of websites found in the given country, 3. **MPCP**, the percentage of the malicious websites, as described earlier, discovered per each country, 4. **MP**, normalizes the number of malicious websites in every country over the total number of websites, 5. **NCSI**, which is the National Cyber Security Index, defined earlier, 6. **DDL**, index signifies the Digital Development Level, 7. **CSPD**, index signifies the Cyber Security Policy Development, 8. **CTAI**, index signifies the Cyber Threat Analysis and Information, 9. **PDS**, index signifies the Protection of Digital Services, 10. **PPD**, index signifies the Protection of Personal Data, and 11. **Average**, which is the average score for each of the previous features.

5.4 Analysis Results

We provide the results of our findings through an analysis pipeline used to examine the distribution of free and premium content websites. We describe and compare trends in free and premium content websites distribution across countries, and their comparison to the concentration of the general websites. We will compare the results from each feature analysis we performed, followed by an analysis that takes into account such characteristics with regard to the type of content being hosted on the given website – i.e., content category analysis (books, games, movies, music, and software). Finally, we provide the NCSI scores analysis, which may reveal some hosting affinities or correlations between the hosting countries and the hosted free and premium content websites.

5.4.1 Country-Level Distribution Analysis

General Insights. The results of the distribution analysis of free and premium content websites over different countries reflect the following insights followed by the analysis tables results: 1. More than half of the websites reside in the United States, and 33.6% of the free and premium content websites are malicious, as it appears in Table 17. Implies that employing security measures on websites in the United States would improve the security of websites by almost 20% for websites with free and premium content. 2. Preventing free websites from being deployed in Belgium will contribute to changing the classification of Belgium from the second top in hosting malicious websites to exclusively hosting benign websites. 3. Compared to locations worldwide, free and premium websites in all categories are primarily hosted in the US rather than in other countries.

Free and Premium Websites. Table 17 shows the distribution of free and premium content websites, along with their respective MC, MPCP, and MP. The United States leads with 58.5% of the total websites, followed by Belgium at 6.6%, the Netherlands at 6.3%, and Germany at 5.9%. Australia, the United Kingdom, France, China, Canada, and Ireland contribute to the overall distribution, as shown in Figure 6.

In terms of MC, the United States has the highest count of malicious websites (297), followed by Belgium (67) and the Netherlands (19). Moreover, the United States, Belgium, and France exhibit significantly higher MPCP for free content websites, while the other countries maintain relatively lower percentages across all categories. Overall, out of 1,509 websites, 479 were malicious. These results shed light on the prevalence of malicious content in different countries, offering valuable insights into potential areas of focus for cyber security efforts.

Benchmark Websites. Table 16 shows the distribution of general websites in countries for free and premium content websites. The United States leads, hosting 45.8% of the websites, followed by Germany (7.1%), France (4.9%), China (3%), the Netherlands (3%), and Canada (2.5%). The United Kingdom, Australia, Ireland, and Belgium have a smaller presence. In terms of MC, the United States has the highest count (44). The total MP for the Top One Million websites is 4.5%.

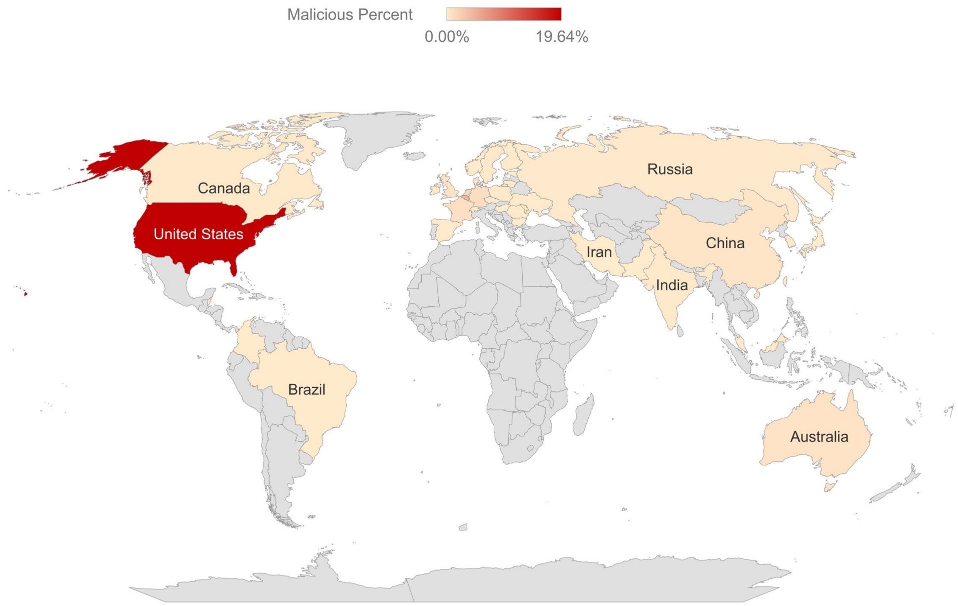


Figure 6: This heatmap illustrates the distribution of malicious websites across various countries.

Other countries account for 29.5% of the websites and have an MC of 121.

In table 17, the United States also leads in hosting free and premium websites, with ≈59%. Belgium, the Netherlands, and Germany follow, but with much smaller percentages. The distribution of malicious content is more pronounced in the United States, with a considerably higher MP of 19.6% compared to the Top One Million websites. MPCP values differ between countries.

In summary, the United States is the dominant country for both general websites, free content websites, and premium content websites. However, there is a notable difference in the distribution of malicious content, with a higher prevalence in the latter category. This insight highlights the potential need for increased cyber security measures for free and premium content websites, especially in the United States. The "Others" category, while not an individual country, still contributes significantly to the total distribution of websites and malicious content.

Free Websites. Table 16 displays the distribution of free content websites in the top hosting countries. The United States leads the list, hosting 50.6% of the free content websites, followed by Belgium (11.2%), Germany (9.4%), the Netherlands (7%), and Australia (5.3%). France, the United Kingdom, Russia, Canada, and Romania have smaller percentages of free content websites. In terms of MC, the United States has the highest count (171), while Belgium has the highest MPCP at 70.5%. The overall MP for free content websites is 40.5%.

The data reveals that the United States is the dominant hosting country for free content websites, with over half of the websites hosted there. However, Belgium has the highest proportion of malicious content, as indicated by the MPCP value. This information highlights the need for increased security measures for free content websites, particularly in countries with higher con-

Table 16: An overview of the distribution of the (top-1M, free content websites, and premium content websites) across different countries. The names are coded using Alpha-3, where GBR stands for the United Kingdom, which here includes Northern Ireland.

General Websites

Country	#	%	MC	MPCP	MP
USA	941	45.75	44	4.68	2.14
DEU	146	7.10	3	2.05	0.15
FRA	101	4.91	6	5.94	0.29
CHN	62	3.01	2	3.23	0.10
NLD	62	3.01	2	3.23	0.10
CAN	51	2.48	2	3.92	0.10
GBR	40	1.94	0	0	0
AUS	24	1.17	3	12.50	0.15
IRL	19	0.92	1	5.26	0.05
BEL	4	0.19	0	0	0
Otr.	607	29.51	121	19.93	5.88
Total	2057	100	92	4.47	4.47

free content websites

Country	#	%	MC	MPCP	MP
USA	399	50.63	171	42.86	21.70
BEL	88	11.17	62	70.45	7.87
DEU	74	9.39	16	21.62	2.03
NLD	55	6.98	17	30.91	2.16
AUS	42	5.33	10	23.81	1.27
FRA	20	2.54	13	65	1.65
GBR	17	2.16	9	52.94	1.14
RUS	13	1.65	2	15.38	0.25
CAN	8	1.02	0	0	0
ROU	7	0.89	3	42.86	0.38
Otr.	65	8.25	16	24.62	2.03
Total	788	100	319	40.48	40.48

premium content websites

Country	#	%	MC	MPCP	MP
USA	485	67.27	126	25.98	17.48
NLD	40	5.55	2	5	0.28
CHN	28	3.88	6	21.43	0.83
GBR	22	3.05	3	13.64	0.42
IRL	21	2.91	1	4.76	0.14
CAN	16	2.22	3	18.75	0.42
IND	16	2.22	2	12.50	0.28
DEU	15	2.08	1	6.67	0.14
FRA	15	2.08	2	13.33	0.28
BEL	11	1.53	5	45.45	0.69
Otr.	52	7.21	9	17.31	1.25
Total	721	100	160	22.19	22.19

centrations of malicious content. The “Others” category, which collectively represents 8.3% of the free content websites, also contributes significantly to the total distribution of malicious content, with an MC of 16 and an MP of 2%.

Premium Websites. Table 16 shows the distribution of premium content websites in the top hosting countries. The United States dominates the list, hosting 67.3% of the premium content websites. The Netherlands, China, the United Kingdom, and Ireland follow with smaller percentages of 5. 6%, 3. 9%, 3. 1% and 2. 9%. Canada, India, Germany, France, and Belgium also host a minor portion of premium content websites. In terms of MC, the United States has the highest count at 126, but Belgium has the highest MPCP at 45.5%.

The overall MP for premium content websites is 22.2%. The data highlights the significant concentration of premium content websites in the United States, but it also shows that Belgium has a higher proportion of malicious content in its premium content websites, as evidenced by its MPCP value. This suggests that security measures should be strengthened for premium content websites, especially in countries with a higher concentration of malicious content. The “Others” category, which collectively represents 7.2% of the premium content websites, also contributes significantly to the total distribution of malicious content, with an MC of 9 and an MP of 1.3%.

Free Websites versus Premium Websites. A comparison between the distribution of free and premium content websites across the top hosting countries reveals several insights, as shown in Table 16. The United States is the top hosting country for both types of websites, with 50.6% of free content websites and 67.3% of premium content websites. However, the distribution of free content websites is more spread across various countries, with Belgium (11.2%), Germany (9.4%), and the Netherlands (7%) hosting a substantial percentage of free content websites. On the other hand, the distribution of premium content websites is more concentrated in the United States, with other countries like the Netherlands, China, and the United Kingdom hosting smaller percentages of 5.6%, 3.9%, and 3.1%, respectively.

Free content websites have an overall MC of 319 compared to premium content websites with 160. The MP of free websites is significantly higher at 40.5% compared to premium ones, which have an MP of 22.2%. This suggests that free content websites are more likely to contain malicious content than premium content websites. Belgium has the highest MPCP for both free content websites (70.5%) and premium content websites (45.5%). This suggests that despite hosting a smaller percentage of websites, Belgium has a high proportion of malicious content.

Overall, the distribution of free content websites is more dispersed across various countries than premium content websites, primarily concentrated in the United States. The malicious content rates in both types of websites highlight the need for improved security measures, particularly in countries with a high concentration of malicious content.

5.4.2 Per-Category Country-level Analysis

Consistent with the prior work that initiated this line of work, we explore the geographical distribution of the free and premium content websites across the various analysis dimensions by considering their category type, emphasizing the type of content such websites serve. Namely, the contents are divided into books, games, movies, music, and software websites.

Books Websites. Table 17 shows the distribution of free and premium book websites across countries. For free websites, the United States has the highest share, accounting for roughly 58% of the total, followed by Germany at roughly 8%, Belgium at roughly 6%, and Australia at roughly 4%. The United States also has the highest MC with 32 instances, while Belgium has the highest MPCP at roughly two-thirds of the total. Furthermore, the United States leads in MP with 22.2%.

For premium content websites, the United States maintains dominance with roughly 62%, followed by China, Canada, the Netherlands, and the United Kingdom. Note that the United States has the highest MC with 34 instances, while Canada and the United Kingdom have the highest MPCP, around 38%. Once again, the United States has the highest MP, at roughly 18%. The United States dominates both types of websites. Belgium has a significantly high MPCP in free content websites. Furthermore, China, Canada, and the United Kingdom have considerable MPCP values in premium content websites.

In summary, there is a significant and clear difference in malicious content distribution in free and premium content websites, with slightly higher MP values found in free content websites than premium content websites.

Games Websites. Table 17 summarizes the distribution for the category of games among countries. For free websites, the United States has almost half, followed by Belgium at almost 13%, then Moldova at 6%, and the Netherlands at 5%. The United States has the highest MC with 31 instances and the highest MPCP at almost 80%. Notably, Belgium stands with the highest MPCP of 100% and a MP of almost 13%. Moreover, the United States leads in MP with almost 40%.

As for premium content websites, the United States dominates once again with more than 60%,

Table 17: An overview of the distribution per category (free content websites vs. Premium content websites, books, games) across different countries.

Overall

Country	#	%	MC	MPCP	MP
USA	884	58.47	297	33.60	19.64
BEL	99	6.55	67	67.68	4.43
NLD	95	6.28	19	20	1.26
DEU	89	5.89	17	19.10	1.12
AUS	48	3.17	10	20.83	0.66
GBR	39	2.58	12	30.77	0.08
FRA	35	2.31	15	42.86	0.99
CHN	33	2.18	7	21.21	0.46
CAN	24	1.59	3	12.50	0.20
IRL	22	1.46	1	4.55	0.07
IND	18	1.19	3	16.67	0.20
RUS	15	0.99	2	13.33	0.13
FIN	12	0.8	1	8.33	0.07
SGP	10	0.66	1	1	0.07
Otr.	86	5.7	24	27.91	1.59
Total	1509	100	479	31.75	31.75

Books

Free Content Websites					
Country	#	%	MC	MPCP	MP
USA	84	58.33	32	38.10	22.22
DEU	11	7.64	0	0	0
BEL	9	6.25	6	66.67	4.17
AUS	6	4.17	0	0	0
FRA	4	2.78	2	50	1.39
Otr.	30	20.83	3	10	2.08
Total	144	9.54	43	29.86	29.86
Premium Content Websites					
USA	118	61.78	34	28.81	17.80
CHN	13	6.81	4	30.77	2.09
CAN	8	4.19	3	37.50	1.57
NLD	8	4.19	1	12.50	0.52
GBR	8	4.19	3	37.50	1.57
Otr.	36	18.85	8	22.22	4.19
Total	191	100	53	27.75	27.75

Games

Free Content Websites					
Country	#	%	MC	MPCP	MP
USA	39	50	31	79.49	39.74
BEL	10	12.82	10	100	12.82
MDA	5	6.41	0	0	0
NLD	4	5.13	3	75	3.85
ROU	3	3.85	0	0	0
Otr.	17	21.79	6	35.29	7.69
Total	78	100	50	64.10	64.10
Premium Content Websites					
USA	67	60.36	31	46.27	27.93
NLD	13	11.71	0	0	0
GBR	5	4.50	0	0	0
CHN	4	3.60	0	0	0
FRA	4	3.60	0	0	0
Otr.	18	16.22	4	22.22	3.60
Total	111	100	35	31.53	31.53

followed by the Netherlands, the United Kingdom, China, and France. In terms of MP, the United States leads with almost 28%, while other countries such as the United Kingdom, China, and France—surprisingly—have no reported malicious instances.

In summary, the United States leads in both free and premium game contents websites, with a higher percentage of malicious content in free content websites. Belgium also has a significant presence in free content websites, with a striking 100% MPCP rate. Other countries have a lesser contribution, and others have no malicious content reported in either free or premium websites.

Movies Websites. Table 18 shows the distribution of movie websites in the dimension studied in different countries. In the case of free websites, the United States has the largest share with approximately 47%, followed by Germany at 15%, Australia at roughly 10%, and the Netherlands at approximately 9%. The United States has the highest MC of 34 and an MPCP of 23%.

For MPCP, we found that Belgium has the highest MPCP, at 31.8%. Moreover, the United States is shown to have the highest MP at 11%. Similarly, for premium content websites, we found that the United States dominates with more than 77%, followed by the Netherlands, China, and Ireland, each of which has only around 4%–5%. The United States again has the highest MC with 20 instances and an MPCP of 17%. China and the Netherlands exhibit similar MPCP values distribution with around 14% and 13%, respectively. Moreover, the United States leads in MP at around only 13%, while most other countries have a smaller number of malicious instances. When comparing free and premium content websites, it is evident that the United States has a more significant share of both types of websites.

In summary, the United States has a higher MP in premium content websites compared to free content websites, while the MC and MPCP are lower in premium content websites. Moreover, the highest MPCP value for a country in free content websites is observed in Belgium, compared to

Table 18: An overview of the distribution per category (movies, music, and software) across different countries.

Movies

Free Content Websites					
Country	#	%	MC	MPCP	MP
USA	146	47.10	34	23.29	10.97
DEU	46	14.84	12	26.09	3.87
AUS	30	9.68	8	26.67	2.58
NLD	29	9.35	8	27.59	2.58
BEL	22	7.10	7	31.82	2.26
Otr.	37	11.94	13	35.14	4.19
Total	310	100	82	26.45	26.45
Premium Content Websites					
USA	118	77.63	20	16.95	13.16
NLD	8	5.26	1	12.50	0.66
CHN	7	4.61	1	14.29	0.66
IRL	6	3.95	0	0	0
AUS	2	1.32	0	0	0
Otr.	11	7.24	1	9.09	0.66
Total	152	100	23	15.13	15.13

Music

Free Content Websites					
Country	#	%	MC	MPCP	MP
USA	43	53.75	19	44.19	23.75
DEU	9	11.25	3	33.33	3.75
BEL	5	6.25	4	80	5
NLD	5	6.25	1	20	1.25
CAN	3	3.75	0	0	0
Otr.	15	18.75	4	26.67	5
Total	80	100	31	38.75	38.75
Premium Content Websites					
USA	58	67.44	13	22.41	15.12
FIN	4	4.65	0	0	0
IRL	4	4.65	0	0	0
NLD	4	4.65	0	0	0
GBR	3	3.49	0	0	0
Otr.	13	15.12	2	15.38	2.33
Total	86	100	15	17.44	17.44

Software

Free Content Websites					
Country	#	%	MC	MPCP	MP
USA	87	49.43	55	63.22	31.25
BEL	42	23.86	35	83.33	19.89
NLD	13	7.39	5	38.46	2.84
GBR	7	3.98	6	85.71	3.41
DEU	6	3.41	1	16.67	0.57
Otr.	21	11.93	11	52.38	6.25
Total	176	100	113	64.20	64.20
Premium Content Websites					
USA	124	68.51	28	22.58	15.47
DEU	9	4.97	1	11.11	0.55
BEL	7	3.87	3	42.86	1.66
NLD	7	3.87	0	0	0
FRA	6	3.31	1	16.67	0.55
Otr.	28	15.47	1	3.57	0.55
Total	181	100	34	18.78	18.78

more evenly distributed values in premium content websites among countries like China and the Netherlands. This suggests that there may be a difference in the distribution of malicious content between free and premium content websites for this category.

Music Websites. Table 18 shows the results of the music websites category distribution across the different countries. From a distribution standpoint, the United States still leads in free and premium content websites, accounting for more than 53% and 67% of each category, respectively. Germany, Belgium, and the Netherlands also have a considerable presence in the free content websites. Analyzing malicious content, free content websites exhibit a higher MP in countries such as the United States (≈24%) and Belgium (5%), with the "Others" category showing a collective MP of 5%. In contrast, premium content websites have lower malicious content rates in most countries, with the United States having an MP of ≈15% and the collective "Others" category at 2.3%. Finland, Ireland, the Netherlands, and the United Kingdom have no reported malicious content in their premium content websites.

In summary, the United States is the primary contributor to both free and premium music content websites, with higher MP observed in the free content websites compared to the premium content websites. Other countries such as Germany, Belgium, and the Netherlands also contribute significantly to music content distribution, displaying varying patterns of malicious content between free and premium content websites.

Software Websites. Table 18 shows the distribution of the free and premium websites across countries, again highlighting a lead of the United States at almost 50% and 69% in the free and premium websites, respectively. Moreover, Belgium and the Netherlands also have a notable presence in both categories. On the other hand, and for malicious content, the free websites have a higher MP overall, particularly in the United States (≈31.3%) and Belgium (≈20%). In contrast, the premium websites have lower malicious content rates across all countries. For instance, the United States has an MP of ≈16%, followed by Belgium (≈2%) in premium content websites.

In summary, the United States is a major contributor to both free and premium content websites, with a higher percentage of malicious content observed in free content websites. Other countries, such as Belgium and the Netherlands, also contribute significantly with software content, with varying levels of maliciousness across free and premium content websites.

5.4.3 National Cyber Security Index

NCSI measures the country-level cyber security maturity, and we use this dimension of analysis to understand if there is any trend in the availability of free content malicious websites in a given country and their association with such an index.

Table 19 lists the relationship between the MPCP of the leading countries in hosting free and premium content websites and their scores on different NCSI criteria. The results reveal that the countries hosting websites marked as malicious, as indicated with a high MPCP and MP, have a varying range of NCSI scores, indicating the limitations in some aspects of the scoring criteria to capture this essential feature (security) of those websites at the country level. For instance, the United States, which has a high MPCP, scored only 20 in the cyber threat analysis and information sharing (CTAI) and protection of digital services (PDS) criteria. On the other hand, a country like Belgium, with 4.4% of MP and 67.68 of MPCP, had a DDL of 75.3% and a CTAI of 80%. With 20% of the hosted websites in it being malicious, the Netherlands had 83.5% in DDL, and 57% in CSPD. Noteworthy, Germany, with 19.1% MPCP and 1.1% of the total MP, scored 90.9% in NCSI, which is a relatively higher rate than the other countries. However, we observe that the same country also had a DDL score of 81.4%.

In Australia, the United Kingdom, and Canada, 20.8%, 30.8%, and 12.5% MPCP, contributed only 0.9% of the total MP, but scored 20% in PDS, 78.7% DDL in Australia, 81.6% in the United Kingdom and 77.1% in Canada. However, the CSPD score of Canada and the United Kingdom is 71%. The same trend applies to other countries, where the results show an average of 75.33% of the countries that host free and premium websites, averaging 31. 8% of MPCP and 2. 9% of MP while scoring an average of 62% in PDS and 72% in CTAI and CSPD, and 77.5% in DDL.

This insight supports the previous hypothesis that some countries may need to improve their cyber security measures to combat cyber threats effectively as the scores may not be indicative of the level of security in certain categories–such as free content websites security. Moreover, we observed that the highest malicious free and premium content websites hosting is in countries performing 20% in CSPD, CTAI, and PDS, highlighting the importance of these criteria in measuring the country-level security matureness regarding the studied threat.

Table 19: The distribution of free and premium content websites across different countries associated with NCSI scores. Studied distribution characteristics for each country: the count, MPCP, MP, and the NCSI ranking scores.

CN	#	MPCP	MP	NCSI	DDL	CSPD	CTAI	PDS
USA	884	33.60	19.64	64.94	82	100	20	20
BEL	99	67.68	4.43	93.51	75.34	100	80	100
NLD	95	20	1.26	83.12	83.48	57	100	80
DEU	89	19.10	1.12	90.91	81.43	100	100	100
AUS	48	20.83	0.66	66.23	78.68	100	100	20
GBR	39	30.77	0.08	77.92	81.55	71	100	20
FRA	35	42.86	0.99	84.42	78.59	86	80	80
CHN	33	21.21	0.46	51.95	60.81	14	20	80
CAN	24	12.50	0.20	70.13	77.09	71	100	20
IRL	22	4.55	0.07	70.13	76.23	71	20	100
AVG	137	31.75	2.89	75.33	77.52	77	72	62

5.5 Discussion

Overall Takeaway. The results of the country-level analysis convey answers to **RQ1** and **RQ2**. We found that the majority of the investigated websites are located in the United States, with 33.6% of free and premium content websites, compared to 45.8% of the general websites. At the same time, a significant number of the studied websites were identified as malicious. Overall, the free content websites, premium content websites, and general websites had a heavy-tailed distribution over the top hosting countries.

Surprisingly, the vast majority of the malicious websites in all three types of websites are mostly hosted in the United States, where the MPCP shows a very high percentage in the case of free content websites for most of the studied countries. In contrast, the highest MPCP in premium content websites mostly concentrated around the top hosting countries. Interestingly, the case is different on the general websites, where the highest MPCP appears in the eighth of the top hosting countries, indicating the severity of the free content websites in comparison to the other types of websites where the MP in free content websites 40.5%, 22.2% in premium content websites, and only 4.5% in the general websites.

Per-category Analysis Takeaway. The summary of the category websites analysis holds answers to **RQ3** by showing the distribution of free and premium content websites over countries. Again, we found that the United States dominates both free and premium content websites across the top hosting countries and various content categories. However, the distribution of free content websites is more spread across various countries, such as Belgium, Germany, and the Netherlands. For MC, MP, and MPCP, the free content websites generally have higher malicious content rates than premium content websites, with Belgium exhibiting the highest MPCP for both types of websites.

We conclude that the United States has the highest MC, while Belgium has a relatively high proportion of malicious content despite hosting a smaller percentage of websites. The results highlight the need for improved security measures, particularly in countries with high concentrations of malicious content on free and premium content websites. Although such measures are desired, at a minimum, this study directly points out the insufficiency of the accepted standard for characterizing the security maturity of a country in light of a specific domain and calls for revising such a standard for hosting capabilities and associated security. On the positive side, Germany, which is considered the second dominant hosting country in most free content websites except in the game and software categories, had one of the lowest malicious hosting scores, captured well in the associated measures.

NCSI Analysis Takeaway. The derived results indicate an answer to **RQ4**, where we found that the most malicious websites are concentrated in countries that have gotten a lower score, at least in one of the following aspects: DDL, CSPD, CTAI, and PDS. This finding supports the examined hypothesis that the weakness in these cyber security aspects could contribute to the high concentration of malicious content on these countries' websites. Therefore, we recommend prioritizing the development of these aspects to improve the overall cyber security measures of these countries, reduce the number of malicious websites, and increase the network's security.

The distribution of free and premium content websites over countries identified the most contributed hosting countries; however, due to the high overlap between malicious and benign websites within these countries, it is essential to investigate other factors which could cause this weakness. It was also found that the most malicious contributed countries have gotten a lower score in at least one aspect, such as Cyber Security Policy Development (CSPD), Cyber Threat Analysis and Information (CTAI), or Protection of Digital Services (PDS).

Contrast with the Literature. The results of the NCSI analysis show compatibility with Alabduljabbar *et al.*'s [13] work where we found indicators that the score of the policy development is relatively low in the countries with a high MPCP, in general. Their findings were drawn from examining the privacy policies, where our results are drawn by tracking the security development indicators of the country and the security state of such websites hosted in a country. As such, our study provides other means of supporting the findings of this prior work.

We also found prior work investigated the security of websites and their geographical distribution at the country level and showed variations between the vulnerable websites per country [21, 24, 59, 70, 72, 75–78, 83–85]. However, most of these studies focus on websites from e-government, universities, and libraries. As such, the distribution of malicious free content websites has not been discussed in prior work in contrast with the studied in-depth dimensions, although our findings are consistent with some of such literature. Both our findings and previous work conclude the need for more regulations on website hosting by considering security as an essential criterion, while our work additionally substantiates this need with an evidence-based study that highlights the performance of existing measures and the gaps that call for further improvements.

Limitations and Recommendations. One of the unexpected results is that malicious free content websites are highly distributed over the hosting countries. This indicates a need to improve to cyber security policies and agreements across those countries to protect users. Moreover, while some of those regulations might be in existence—as indicated by the discrepancy between the NCSI and our measurements, the higher MPCP discovered rate may be due to a lack of *enforcement* of such regulations and policies, calling for tracking the enforcement as an equally important aspect of the matureness of the cyber security policy at the country level. Given the broad usage of the websites class we studied, it is important to note that our results and findings highlight how cybercrime may transcend nations, making it difficult to contain without a coordinated international collaboration and dialogue, which should be embodied in the nation-level security matureness scoring.

One potential explanation for the United States is the lead in some of the measurements we conducted is that the collection of the websites (free and premium content websites) took place from hosts located in the United States, which biases the returned results to only those relevant in the United States—e.g., Google considers a combination of factors to determine the results, including user's location, language, search history, and the relevance of website. We note that such bias would be at the level of the contents, and not unclear whether the infrastructure—the main studied aspect in this study—is taken into account when returning search results.

While we did not consider the root cause for the maliciousness of those websites—as that is an important yet orthogonal pursuit, it would be interesting to explore that in the future. One potential factor contributing to our results concerning the distribution of malicious websites across countries is perhaps the difference in access restrictions, data privacy laws, or other digital security measures across different regions or nations, which could lead to varying levels of risk when accessing content from those areas (and, by the same token, security assurance).

As a primary recommendation of this work, and based on the key findings, there is a need to develop better cyber security policies and regulations to reduce the risk exposure for users who access these websites. Moreover, while this work provides an overview of the country-level distribution patterns associated with free content websites, premium content websites, and their association, much work remains to be done to identify the correlation between the maliciousness of a website and its regional environments.

5.6 Summary and Work to be Completed

We examined the distribution of free and premium content websites in different countries, showing that malicious free content websites are highly concentrated in some countries and highlighting the need for more mature cyber security policies to ensure security. We also examined the discrepancy between the NCSI scores of hosting countries and malicious free content website averages. The findings presented here can help inform strategies to better protect users against vulnerabilities beyond their control. Our study is not without limitations, including the need to revisit data collection, annotation, and website types, which can all be continuously improved to provide better coverage, representation, and data balance. In the future, it would also be important to understand through measurements the different types of maliciousness of websites, and their severity, which may impact the weight of the different policy scores and their relevance. Identifying the weaknesses exploited in different categories or regions may help shed light on more precise policy.

6 Concluding Remarks

The security of free content websites is very important because they are popular websites that are accessed by many users. The impact of malicious free content websites can spread widely and affect many users around the world. Thus, in this dissertation, we explored three topics that could impact the security of free content websites that could reflect the security of the Internet. At first, we examined the content management systems of free content websites and found that custom-coded websites are dangerous considering the entropy of the code used and the various possibilities to transform benign websites into malicious ones by targeting their source code. Second, we classified the free content websites based on their favorite network scale and found that they predominantly reside within medium-scale networks with a high association of malicious and benign websites at the same time. This led to testing the security of the CSPs for free content websites, where we found they were heavily distributed over the top ten CSPs with a high concentration of malicious websites. On the other hand, the rules and regulations for the CSPs are related to the hosting countries. In the third work, we explored the distribution of free content websites over the hosting countries. Highlight hidden correlations between hosting a high number of malicious free content websites with the National Cyber Security Index scores. The top hosting countries tend to have vague security policies that provide no collaborations with other nations, creating a golden opportunity for more malicious free content websites to be hosted in their countries. However, this work comprehensively investigated the affinities of hosting infrastructures that host free content websites. Future work is recommended to reveal hidden correlations for free content websites with network scales, CSPs, or hosting countries to reveal the direct relation for hosting malicious free content websites in order to overcome these causes and ensure the security of users.

References

[1] —. What CMS is That? Use CMS Detector and Find Out, 2022. Last access March 25, 2022.

[2] —. Reliable IP ddress Data, 2022. Last access December 14, 2022.

[3] —. Find and automatically fix vulnerabilities in your code, open source dependencies, containers, and infrastructure as code, 2022. Last access March 15, 2022.

[4] —. W3Techs - World Wide Web Technology Surveys, 2022. Last access April 28, 2022.

[5] —. The ultimate security vulnerability datasource, 2022. Last access March 15, 2022.

[6] —. Analyze suspicious files and URLs to detect types of malware automatically, 2022. Last access December 14, 2022.

[7] —. IP Address Lookup Tools, 2023. Last access January 19, 2023.

[8] S. A. Adepoju, I. O. Oyefolahan, M. B. Abdullahi, A. A. Mohammed, and M. O. Ibiyo. A Human-Centered Usability Evaluation of University Websites Using SNECAAS Model. In *Handbook of Research on the Role of Human Factors in IT Project Management*, pages 173–185. IGI Global, 2020.

[9] D. Akhawe, A. Barth, P. E. Lam, J. C. Mitchell, and D. Song. Towards a Formal Foundation of Web Security. In *Proceedings of the 23rd IEEE Computer Security Foundations Symposium, CSF*, pages 290–304, 2010.

[10] A. Alabduljabbar, A. Abusnaina, Ülkü Meteriz-Yıldıran, and D. Mohaisen. TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. In *ACM WPES*, pages 103–118, 2021.

[11] A. Alabduljabbar, R. Ma, S. Alshamrani, R. Jang, S. Chen, and D. Mohaisen. Poster: Measuring and assessing the risks of free content websites. In *NDSS*, 2022.

[12] A. Alabduljabbar, R. Ma, S. Choi, R. Jang, S. Chen, and D. Mohaisen. Understanding the Security of Free Content Websites by Analyzing their SSL Certificates: A Comparative Study. In *CySSS@AsiaCCS*, pages 19–25, 2022.

[13] A. Alabduljabbar and D. Mohaisen. Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation. In *Companion of The Web Conference, WWW*, pages 860–867, 2022.

[14] M. Alaqdhi, A. Alabduljabbar, K. Thomas, S. Salem, D. Nyang, and D. Mohaisen. Do Content Management Systems Impact the Security of Free Content Websites? A Correlation Analysis. In *CSoNet*, 2022.

[15] H. Alasmary, A. Anwar, A. Abusnaina, A. Alabduljabbar, M. Abuhamad, A. Wang, D. Nyang, A. Awad, and D. Mohaisen. ShellCore: Automating Malicious IoT Software Detection Using Shell Commands Representation. *IEEE Internet Things J.*, pages 2485–2496, 2022.

[16] H. Alasmary, A. Khormali, A. Anwar, J. Park, J. Choi, A. Abusnaina, A. Awad, D. Nyang, and A. Mohaisen. Analyzing and Detecting Emerging Internet of Things Malware: A Graph-Based Approach. *IEEE Internet Things J.*, pages 8977–8988, 2019.

[17] M. Alkinoon, A. Alabduljabbar, H. Althebeiti, R. Jang, D. Nyang, and D. Mohaisen. Understanding the Security and Performance of the Web Presence of Hospitals: A Measurement

Study. In *32nd International Conference on Computer Communications and Networks, ICCCN, Honolulu, HI, USA*, pages 1–10, 2023.

[18] M. Alkinoon, A. Alabduljabbar, H. Althebeiti, R. Jang, D. Nyang, and D. Mohaisen. Understanding the Security and Performance of the Web Presence of Hospitals: A Measurement Study. *CoRR*, 2023.

[19] M. Alkinoon, S. J. Choi, and D. Mohaisen. Measuring Healthcare Data Breaches. In *Proceedings of the 22nd International Conference on Information Security Applications, WISA*, pages 265–277, 2021.

[20] M. Alkinoon, M. Omar, M. Mohaisen, and D. Mohaisen. Security Breaches in the Healthcare Domain: A Spatiotemporal Analysis. In *Proceedings of the 10th International Conference on Computational Data and Social Networks (CSoNet)*, pages 171–183. Springer, 2021.

[21] O. Alrawi and A. Mohaisen. Chains of Distrust: Towards Understanding Certificates Used for Signing Malicious Applications. In *Proceedings of the 25th International Conference on World Wide Web,(WWW)*, pages 451–456, 2016.

[22] I. Alsmadi and F. Mira. Website security analysis: variation of detection methods and decisions. In *Proceedings of the 21st IEEE/Saudi Computer Society National Computer Conference (NCC)*, 2018.

[23] H. Althebeiti and D. Mohaisen. Enriching Vulnerability Reports Through Automated and Augmented Description Summarization. *CoRR*, 2022.

[24] P. Bangera and S. Gorinsky. Ads versus regular contents: Dissecting the web hosting ecosystem. In *Proceedings of Networking Conference, IFIP Networking and Workshops, Stockholm, Sweden, IEEE*, pages 1–9, 2017.

[25] S. Calzavara, A. Rabitti, and M. Bugliesi. Content Security Problems?: Evaluating the Effectiveness of Content Security Policy in the Wild. In *ACM CCS*, pages 1365–1375, 2016.

[26] S. Calzavara, A. Rabitti, and M. Bugliesi. Semantics-Based Analysis of Content Security Policy Deployment. *ACM Trans. Web*, 12(2):10:1–10:36, 2018.

[27] cybersecurity help. Vulnerability Database , 2022. Last access March 15, 2022.

[28] D. G. Dobolyi and A. Abbasi. PhishMonger: A free and open source public archive of realworld phishing websites. In *Proceedings of IEEE Conference on Intelligence and Security Informatics, ISI*, pages 31–36, 2016.

[29] e Governance Academy. National Cyber Security Index, 2023. Last access February 8, 2023.

[30] S. Englehardt and A. Narayanan. Online Tracking: A 1-million-site Measurement and Analysis. In *ACM CCS*, pages 1388–1401, 2016.

[31] B. Everman and Z. Zong. GreenWeb: Hosting High-Load Websites Using Low-Power Servers. In *Proceedings of the Ninth International Green and Sustainable Computing Conference, IEEE*, pages 1–6, 2018.

[32] D. Fett, R. Küsters, and G. Schmitz. The Web SSO Standard OpenID Connect: In-depth Formal Security Analysis and Security Guidelines. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium, CSF*, pages 189–202, 2017.

[33] E. Figueras-Martín, R. Magán-Carrión, and J. Boubeta-Puig. Drawing the web structure and content analysis beyond the Tor darknet: Freenet as a case of study. *J. Inf. Secur. Appl.*, 68(8):103229, 2022.

[34] H. Fryer, S. StallaBourdillon, and T. Chown. Malicious web pages: What if hosting providers could actually do something.. *Comput. Law Secur. Rev.*, 31(4):490–505, 2015.

[35] R. Gall. WordPress 5.9.2 Security Update Fixes XSS and Prototype Pollution Vulnerabilities, 2022. Last access March 18, 2022.

[36] S. M. Ghaffarian and H. R. Shahriari. Software Vulnerability Analysis and Discovery Using Machine-Learning and Data-Mining Techniques: A Survey. *ACM Comput. Surv.*, 50(4):56:1–56:36, 2017.

[37] S. Ghodke. Top 1 Million Websites, 2022. Last access December 8, 2022.

[38] Marie Vasek and Matthew Weeden and Tyler Moore. Measuring the impact of sharing abuse data with web hosting providers. In *Proceedings of the Workshop on Information Sharing and Collaborative Security, ACM*, pages 71–80, 2016.

[39] T. T. Huynh, T. D. Nguyen, N. T. H. Nguyen, and H. Tan. Privacy-Preserving for Web Hosting. In *Industrial Networks and Intelligent Systems - 6th EAI International Conference*, volume 334 of *Proceedings of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer*, pages 314–323, 2020.

[40] S. Jayanthi and M. S. Sasikala. XGraphticsCLUS: Web mining hyperlinks and content of terrorism websites for homeland security. *International Journal of Advanced Networking and Applications*, 2(6):941–949, 2011.

[41] H. Kang, J. Jang, A. Mohaisen, and H. K. Kim. Detecting and Classifying Android Malware Using Static Analysis along with Creator Information. *Int. J. Distributed Sens. Networks*, pages 479174:1–479174:9, 2015.

[42] R. P. Kasturi, Y. Sun, R. Duan, O. Alrawi, E. Asdar, V. Zhu, Y. Kwon, and B. Saltaformaggio. TARDIS: Rolling Back The Clock On CMS-Targeting Cyber Attacks. In *Proceedings of the IEEE Symposium on Security and Privacy, SP*, pages 1156–1171, 2020.

[43] S. Khare and A. Badholia. Analysis of Cloud and Self-Web-Hosting Services Based on Security Parameters. *Int. J. Inf. Syst. Model. Des.*, 13(6):1–14, 2022.

[44] J. Kohout and T. Pevný. Automatic discovery of web servers hosting similar applications. In *IFIP International Symposium on Integrated Network Management, IEEE*, pages 1310–1315, 2015.

[45] S. Kondakci. A concise cost analysis of Internet malware. *Comput. Secur.*, 28(7):648–659, 2009.

[46] G. Kontaxis, D. Antoniades, I. Polakis, and E. P. Markatos. An empirical study on the security of cross-domain policies in rich internet applications. In *Proceedings of the Fourth European Workshop on System Security, EuroSec*, 2011.

[47] A. E. Kosba, A. Mohaisen, A. G. West, T. Tonn, and H. K. Kim. ADAM: Automated Detection and Attribution of Malicious Webpages. In *Proceedings of the 15th International Workshop on Information Security Applications, WISA*, pages 3–16, 2014.

[48] D. Lee, K. Nam, I. Han, and K. Cho. From free to fee: Monetizing digital content through expected utility-based recommender systems. *Inf. Manag.*, 59(6):103681, 2022.

[49] E. Lee, J. Woo, H. Kim, A. Mohaisen, and H. K. Kim. You are a Game Bot!: Uncovering Game Bots in MMORPGs via Self-similarity in the Wild. In *23rd Annual Network and Distributed System Security Symposium, NDSS, San Diego, California, USA*, 2016.

[50] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In *Proceedings of the ACM Conference on Computer and Communications Security, CCS*, pages 674–686, 2012.

[51] X. Liao, C. Liu, D. McCoy, E. Shi, S. Hao, and R. A. Beyah. Characterizing Long-tail SEO Spam on Cloud Web Hosting Services. In *Proceedings of the 25th International Conference on World Wide Web, ACM*, pages 321–332, 2016.

[52] T. Libert. Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites. *CoRR*, 2015.

[53] E. Mannes and C. Maziero. Naming Content on the Network Layer: A Security Analysis of the Information-Centric Network Model. *ACM Comput. Surv.*, 52(3):44:1–44:28, 2019.

[54] S. Matic, G. Tyson, and G. Stringhini. PYTHIA: a Framework for the Automated Analysis of Web Hosting Environments. In *World Wide Web Conference*, pages 3072–3078, 2019.

[55] S. A. Mirheidari, S. Arshad, S. Khoshkdahan, and R. Jalili. Two novel server-side attacks against log file in Shared Web Hosting servers. In *Proceedings of The 7th International Conference for Internet Technology and Secured Transactions, ICITST IEEE*, pages 318–323, 2012.

[56] S. A. Mirheidari, S. Arshad, S. Khoshkdahan, and R. Jalili. A Comprehensive Approach to Abusing Locality in Shared Web Hosting Servers. *CoRR*, abs/1811.00922, 2018.

[57] A. Mohaisen. Towards Automatic and Lightweight Detection and Classification of Malicious Web Contents. In *IEEE Hot Topics in Web Systems and Technologies*, pages 67–72, 2015.

[58] A. Mohaisen and O. Alrawi. Unveiling Zeus: automated classification of malware samples. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, Companion Volume*, pages 829–832, 2013.

[59] A. Mohaisen, O. Alrawi, and M. Mohaisen. AMAL: High-fidelity, behavior-based automated malware analysis and classification. *Comput. Secur.*, 52:251–266, 2015.

[60] A. Mohaisen, A. G. West, A. Mankin, and O. Alrawi. Chatter: Classifying malware families using system event ordering. In *IEEE Conference on Communications and Network Security, CNS, San Francisco, CA, USA*, pages 283–291, 2014.

[61] V. L. Nguyen, P. Lin, and R. Hwang. Preventing the attempts of abusing cheap-hosting Webservers for monetization attacks. *CoRR*, abs/1903.05470, 2019.

[62] A. Noroozian, E. Rodríguez, E. Lastdrager, T. Kasama, M. van Eeten, and C. Gañán. Can ISPs Help Mitigate IoT Malware? A Longitudinal Study of Broadband ISP Security Efforts. In *Proceedings of the IEEE European Symposium on Security and Privacy, EuroS&P*, pages 337–352, 2021.

[63] openbugbounty. The complete list of bug bounty and security vulnerability disclosure programs lauhched and operated by open bug bounty community, 2022. Last access March 15, 2022.

[64] A. Ostroushko. Restricting Access to Websites as an New Procedure of Government Coercion. *Financial Law and Management*, pages 167–173, 2015.

[65] X. Pan, Y. Cao, S. Liu, Y. Zhou, Y. Chen, and T. Zhou. Cspautogen: Black-box enforcement of content security policy upon real-world websites. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 653–665, 2016.

[66] S. Raponi and R. D. Pietro. A Longitudinal Study on Web-Sites Password Management (in)Security: Evidence and Remedies. *IEEE Access*, pages 52075–52090, 2020.

[67] S. R. Rizvi, B. D. Killough, A. Cherry, and S. Gowda. Lessons Learned and Cost Analysis of Hosting a Full Stack Open Data Cube (ODC) Application on the Amazon Web Services (AWS). In *Proceedings of International Geoscience and Remote Sensing Symposium, IEEE*, pages 8643–8646, 2018.

[68] S. S. Roy, U. Karanjit, and S. Nilizadeh. A Large-Scale Analysis of Phishing Websites Hosted on Free Web Hosting Domains. *CoRR*, abs/2212.02563, 2022.

[69] N. Samarasinghe, A. Adhikari, M. Mannan, and A. M. Youssef. Et tu, Brute? Privacy Analysis of Government Websites and Mobile Apps. In *ACM Web Conference*, 2022.

[70] L. Sandoval-Guzman and H. Petrie. Using Freedom of Information requests to understand usability problems with e-government websites. In *HCI*, 2017.

[71] M. Schulz and M. Pieper. Web Compliance Management: Barrier-Free Websites Just by Simply Pressing the Button? Accessibility and the Use of Content-Management-Systems. In *Proceedings of The Universal Access in Ambient Intelligence Environments, 9th ERCIM Workshop on User Interfaces for All, Königswinter, Springer*, pages 419–426, 2006.

[72] N. Shafqat and A. Masood. Comparative analysis of various national cyber security strategies. *International Journal of Computer Science and Information Security*, 14(1):129–136, 2016.

[73] H. Shimamoto, N. Yanai, S. Okamura, J. P. Cruz, S. Ou, and T. Okubo. Towards Further Formal Foundation of Web Security: Expression of Temporal Logic in Alloy and Its Application to a Security Model With Cache. *IEEE Access*, 7:74941–74960, 2019.

[74] S. Tajalizadehkhoob, T. van Goethem, M. Korczynski, A. Noroozian, R. Böhme, T. Moore, W. Joosen, and M. van Eeten. Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting. In *Proceedings of the SIGSAC Conference on Computer and Communications Security, ACM*, pages 553–567, 2017.

[75] L. Vaughan and Y. Zhang. Equal Representation by Search Engines? A Comparison of Websites across Countries and Domains. *J. Comput. Mediat. Commun.*, pages 888–909, 2007.

[76] D. L. Velasquez and N. Evans. Public library Websites as electronic branches: a multi-country quantitative evaluation. *Inf. Res.*, 23(1), 2018.

[77] S. F. Verkijika and L. de Wet. Quality assessment of e-government websites in Sub-Saharan Africa: A public values perspective. *Electron. J. Inf. Syst. Dev. Ctries.*, 84(2):12015, 2018.

[78] S. Wakeling, D. Kingsley, H. R. Jamali, M. A. Kennan, and M. Sarrafzadeh. Free for all, or free-for-all? A content analysis of Australian university open access policies. *Inf. Res.*, 27(2), 2022.

[79] A. Wang, A. Mohaisen, W. Chang, and S. Chen. Capturing DDoS Attack Dynamics Behind the Scenes. In *Detection of Intrusions and Malware, and Vulnerability Assessment - 12th International Conference, DIMVA, Milan, Italy*, pages 205–215. Springer, 2015.

[80] S. Wang, K. MacMillan, B. Schaffner, N. Feamster, and M. Chetty. A First Look at the Consolidation of DNS and Web Hosting Providers. *CoRR*, abs/2110.15345, 2021.

[81] R. Wash, E. J. Rader, R. Berman, and Z. Wellmer. Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites. In *Symposium on Usable Privacy and Security, SOUPS*, 2016.

[82] N. Wickramasinghe, M. Nabeel, K. Thilakaratne, C. Keppitiyagama, and K. D. Zoysa. Uncovering IP Address Hosting Types Behind Malicious Websites. *CoRR*, abs/2111.00142, 2021.

[83] M. A. Zare and A. Z. Ravasan. A Framework for Assessing Governmental Websites Quality: The Case of Iranian Free Economic Zones Websites. *Int. J. E Serv. Mob. Appl.*, 6(1):44–65, 2014.

[84] J. J. Zhao and S. Y. Zhao. Opportunities and threats: A security assessment of state e-government websites. *Gov. Inf. Q.*, 27(1):49–56, 2010.

[85] Şevval Seray Macakoğlu, S. Peker, and İhsan Tolga Medeni. Accessibility, usability, and security evaluation of universities' prospective student web pages: a comparative study of Europe, North America, and Oceania. *Universal Access in the Information Society*, pages 1–13, 2022.