International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET231029

A Novel Approach for Detection of Malicious Websites using Machine Learning Techniques

Dr. Md. Sirajuddin[1], B. Bhavani [2] ,Y. Akshaya[3] , P. Reethika[4] ,T. Sriram Reddy [5]

[1]Professor, Head of the Department,Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur (Dt), Andhra Pradesh, India

[2, 3, 4, 5]B. Tech Students, Department of Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur(Dt), Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 02 March 2023
Published: 15 March 2023

Publication Issue

Volume 10, Issue 2
March-April-2023

Page Number

68-74

ABSTRACT

When an unsuspecting victim visits a malicious website, it infects her machine to steal valuable information, redirects her to malicious targets, or compromises her system to launch future attacks. While current approaches have. There are still open issues in effectively and efficiently addressing: filtering of web pages from the wild, coverage of a wide range of malicious characteristics to capture the big picture, continuous evolution of web page features, systematic combination of features, semantic implications of feature values on characterizing web pages, ease and cost of flexibility and scalability of analysis and detection technology. In this position paper, we highlight our ongoing efforts towards effective and efficient analysis and detection of malicious websites, with a particular emphasis on broader feature space and attack-payloads, technique flexibility with changes in malicious characteristics and web pages, and, most importantly, technique usability in defending users against malicious websites.

Keywords-Malicious Websites, Detection, Efficiency, Effec- Tiveness

I. INTRODUCTION

Attackers trick an unsuspecting victim into visiting malicious websites, where they steal vital credentials or install malware on the victim's machine to use as a springboard for future exploits. When a victim visits a malicious website, the attack is launched, and the attack payload is executed if evidence of exploitable vulnerabilities (e.g., browser components or browser extensions) is found. Several automated analysis and detection techniques have been proposed to protect Web users from malicious websites. However, given the alarming prevalence of malicious websites and the ever-changing techniques in crafting attack payloads combined with emerging threats, current approaches to tackling the problem have common and specific limitations in effectively and efficiently: characterizing the malicious payloads using a more complete feature set; incorporating inevitable evolution of web page features; systematic methods of selecting and composing web page features; ensuring the feasibility of web page features; ensuring the feasibility of web page features.

COVID-19 has surprised the world with significant impact on the lives of billions of people [17]. This pandemic has accelerated the global digitization [18]. In present digital era, all the transactions are performed digitally across the globe [18]. This digitization has opened up many opportunities for intruders to launch their attacks to gain unauthorized access to the institutional assets or people's personal information. In this digital era, cyber security has become prominent aspect to consider. This paper enlightens on how an attacker uses malicious websites to steal personal information of users.

II. MALICIOUS WEBSITES : THE SIZE OF THE PROBLEM

Approaches proposed to combat the effects of malicious websites fall into two complementary categories: static and dynamic analysis. To perform analysis and construct characterizations of malicious payloads, the former rely primarily on the source code and some static features such as URL structure, host-based information, and web page content. The latter are concerned with capturing "behaviours" that emerge when the page is rendered in a controlled environment. A strategy shared by both approaches is that they extract features of some kind for further analysis in order to obtain patterns of malicious payloads, which are then used to train a classification algorithm using machine learning techniques. A popular protection method is black-listing known malicious URLs and IP addresses gathered through manual reporting, honeypots, and custom analysis techniques. While blacklisting is simple to set up and use, it is only effective if malicious websites are thoroughly identified and the blacklist is updated on a regular basis. In practise, this is impossible because: new websites are too new to be blacklisted, even if they are malicious, some websites may escape blacklisting due to incorrect analysis (e.g., "cloaking"), and attackers may frequently change where the malicious websites are hosted. As a result, URLs and IP addresses may also change.

Lexical aspects of URLs (e.g., URL length, domain name length, query length, path length) and host-based information (e.g., WHOIS information, DNS records) have been shown to be effective in economically characterising malicious web pages in [2], [7], and partially in [1]. The main assumption in such approaches is that malicious URL tokens and host-based values differ from those of benign URLs. The speed with which such approaches extract features without executing the URL is their strength. However, if we look at the WHOIS information of recently registered websites by low-reputation registrars, such websites are more likely to be classified as malicious due to low reputation scores. In practise, there is a high likelihood of false positives.. False negatives may also occur, as old and well-known registrars may host malicious domains. Another source of false negatives could be websites that use free hosting services, which are already compromised but have benign-looking URL and host information. Above all, the feature space considered is far too small to capture the most lethal attacks such as malicious code injection, drive-by downloads, and social engineering tricks.

Another proposal is to use machine learning techniques to classify web pages based on features such as text content, HTML, native JavaScript functions and objects, ActiveX objects, and iFrame size [8], [9]. Such approaches can also quickly extract content features to train classifiers for web pages, and they are especially effective at detecting spam and phishing scams. [1] proposes and demonstrates a fast pre-filtering technique that combines URL structure, host-based information, and page content to significantly reduce the execution load of a dynamic analysis technique. The general limitation of focusing solely on page content is the high risk of obfuscated content (for example, multilevel obfuscation of JavaScript code) and ignoring web pages that must be executed. Honey-clients are systems that visit a web page in a dedicated sandbox environment (e.g., virtual machine) to mimic a human visitor. When a page is rendered, the execution dynamics are captured and analysed to infer evidence for attack payloads. Honey-clients are classified into three types: low-interaction (use simulated browser and minimal OS features), high-interaction (use real browser and full OS features), and hybrid (combine qualities of both). Low-interaction honey-clients, such as HoneyC, are typically limited to monitoring the traces of activities during the interaction against pre-defined signatures; as a result, they are unable to detect zero-day exploits due to the static nature of the reference signatures.On the contrary, high- interaction honey-clients such as Capture HPC and HoneyMonkey [16] check integrity changes in system states which requires monitoring file system, registry entries, processes, network connection, and physical resources like memory and CPU consumption anomalies. The advantage of honey-clients, specially high-interaction ones is the deep insight they provide as to the internal details of attack payloads embedded in malicious websites. However, they are resource-intensive as they need to load and execute individual pages under analysis and modern web pages are usually stuffed with rich client-side code and multimedia taking longer analysis time per a web page. Moreover, not all web pages are likely to launch attacks upon visiting. There are web pages which demand user interaction or wait for time-bombs to take action, which makes honey-clients

inflexible for such websites. From evasion perspective, IP addresses of honey-clients are likely to be black-listed by malicious servers, their virtual machines be detected through advanced fingerprint identification techniques and they may also be victims of CAPTCHA verification that necessarily involve human visitors.

Given the common and specific limitations of existing approaches, a reliable method to analyse a web page and alert web users prior to visiting a potentially malicious website is still far from available to protect web users from attacks. Furthermore, one unresolved issue is the fine-grained characterization of a rich set of features relevant to the ever-changing malicious payloads and multiple artefacts of malicious websites. Following are research questions with an emphasis on effectiveness and efficiency issues that challenge existing approaches to common challenges, web page feature issues, and analysis and detection techniques.

Common Challenges.

The vast majority of existing approaches to analysing and detecting malicious websites are based on a prominent attack as the foundation of their techniques. An attacker, on the other hand, can create virtually any variation of an existing or newly devised attack and embed it in web pages. As a result, the majority of the techniques not only ignore a different type of attack than the one for which they were designed, but they are also likely to miss the fine-grained features that characterise a malicious web page because the techniques are based on a limited set of features. As a result, the techniques fall short of providing valuable information about a comprehensive snapshot of attack payloads on which the analysis and detection techniques could be based.

No matter how effective and efficient a security technique is, it is only as strong as its resistance to potential attacker countermeasures. Techniques for analysing and detecting malicious websites are no exception. Honey-clients, for example, are vulnerable to fingerprinting by definition. Even methods that rely on page content are vulnerable to evasion due to sophisticated obfuscation and cloaking.

Feature Completeness: The number and semantics of feature values in the state-of-the-art are insufficient to capture a comprehensive snapshot of malicious web page characteristics.

Feature Type: While identifying feature types is useful in developing effective and efficient feature extraction and selection schemes, none of the existing approaches incorporate or assume recurring features (e.g., WHOIS information, URL tokens) in comparison to features that change frequently (e.g., page content).

Feature Selection, Composition and Values: Given the different magnitude of contribution by feature values in characterizing malicious web pages, there is no systematic way to prioritize and define an affective and efficient method for computing, selecting, and combining features of different type and semantics in a manner that enhances existing feature extraction techniques. Moreover, feature selection, value computation and, weight assigning are subjective across approaches.

Feature Evolution: Existing web page features may change (become obsolete) and new features may emerge as web pages evolve due to inevitable changes such as change in: page content, functionality, protocols, configuration, browser components, browser extensions, and usage policies. The major bottleneck with most of the existing techniques is lack of flexibility to (semi)automatically and quickly revise and accordingly upgrade the extraction, analysis, and detection techniques, which rely on evolving features. Analysis and Detection Strategies.

Although machine learning techniques are potentially effec- tive in the analysis and detection of malicious websites, there are yet unaddressed questions. First of all, different approaches use partial snapshot of the web page to eval- uate different machine learning techniques (e.g. classifier builders). Secondly, the dataset used even for the same feature set is so diverse. Third, the relationship between the web page features used and the machine learning techniques applied is not well-traced experimentally and not in large- scale context except in few recent works such as [1]. As a result, the answer to the question ”which machine learning technique is effective and efficient for malicious website detection and why?” is still subjective.

Concerning training of models to build classifiers, whether training is conducted after combining all the distinct feature classes into a single thick feature vector or it is sepa- rately conducted for each feature class, remains a question partially-answered. Recently, in [1], it is suggested to com- bine the models learned for each feature set (URL tokens, host-based information and page content in this case) to perform a collaboration-oriented classification. However, the detail of the composition technique and why it is effective needs further investigation.

III. THE ROADMAP

In response to the problems and questions posed in the previous section, we present our proposal to address the general problem of effectively and efficiently detecting malicious websites and the specific questions about the com- mon challenges shared by current approaches, issues about web page features, and analysis and detection techniques. The core mission in our proposal is not only to detect known malicious websites in the wild but also to uncover not yet known malicious websites with optimized resource consumption.

We envision a more comprehensive, more effective and more efficient approach that takes into account : common challenges, web page feature-related issues and analysis and detection techniques. To tackle the common challenge of focusing on a prominent attack, we propose an aggregation of different types of attack payloads and investigating the relationship among each attack type to look for patterns of convergence towards capturing the big snapshot of an ideal malicious web page.

To achieve a more complete set of features, based on the existing feature set in the state of the art, we are currently building a richer set of features and a feature extraction en- gine to enhance both feature quantity and feature granularity. Rather than considering restricted set of features, limited to capturing part of the big picture, we are incorporating all the relevant features of web page identity, URL tokens, web page content and web page execution trace. In line with this, we are also investigating the effective methods to combine such features to achieve a more complete characterization of attack payloads. To make the features more useful, we are also refactoring some of them to have a more fine-grained feature values. For example, current approaches consider only presence of remote links on a page. Part of our proposal is to split links into local and remote. In addition, identifying the final targets of links (e.g., other pages, executable files, PDF documents, images) is useful to capture fine-grained values of features.

Another investigation we are undertaking is to draw a solid line between stable features and those changing frequently. Such an identification is a basis for estimating the resource consumption of analyzing a single web page based on which an optimization strategy could be devised. In this regard, one of the possible ways forward is to identify features that are based on stable feature sources and monitor their resource consumption during extraction from the web page. For example, WHOIS information features are relatively stable as opposed to sequence and number of functions called when the browser invokes a plugin.

By devising an algorithm that identifies the most visible contributors in the feature scores and validating these score values using historical profile of feature extraction in addi- tion to domain knowledge, we can filter only the features for which acceptable values are extracted. To this end, we plan to apply state-of-the-art feature selection techniques that minimize redundancy and maximize relevance to get the best candidate feature set with respect to improving effectiveness and efficiency of analysis and detection.

With an ultimate goal of identifying methods to map set of features to set of algorithms, we can incorporate features that are subject to change and as a result define a frame of reference for training on the fly using live feeds of URLs from the real world. In this respect, online learning algorithms [7], [17] instead of the classical ones are demonstrated to be more effective and much faster with URL tokens and host-based features. We plan to extend the usage of online learning techniques by introducing page content and execution trace features. Building up on the suggestions in [1] about combining models after separate training, another interesting line of investigation we are planning is to experimentally verify the merits and demerits of training each feature class versus training for the union of the feature classes with respect to efficiency and detection accuracy. Another equally important issue to investigate concerning training is the frequency with which the models are updated since the set and values of features extracted from a web page analyzed at time t1 may not be the same at time t2, and the time frame t2-t1 has implications both on the accuracy and efficiency of the analysis and detection.

IV. III.Malicious URL

The malicious URL appears to be a normal URL. A simple URL will cause significant harm to your digital device. These malicious links are regarded as one of the most serious threats to the modern digital world. They can also carry out these attacks by sending these viruses or any links via email[5]. According to reports, the Malicious URL link was created to promote scams, attacks, and frauds. If we visit an affected URL, the malware or trojan will automatically download and take control of the device. Spam and phishing are the most common malicious URL scams. What are the different types of frauds used by criminals who attempt to hack and steal personal information?

We will obtain these URLs by receiving them in an email from an unknown or stranger.through message. These malicious URLs are thought to be attack launchpads. By visiting these URLs, fraudsters can steal personal information or automatically install malware, viruses, or trojans without the user's knowledge. Threats to cybersecurity are regarded as such[8].

Since social websites or applications such as Facebook, Instagram, Twitter, Telegram, or WhatsApp are widely used by nearly 90% of people worldwide, data thieves are now simply using these platforms to commit crimes by creating spam websites, advertisements, or pictures, and we select those spams.

V. Malicious URL Detection

Methods for Detecting Malicious URLs in General:

Some general patterns can be used to detect malicious sites, which can be a noticeable pattern when it comes to user infections. The most common targets are gambling, gaming, porn, and video streaming websites. When we visit these websites, if we click on a link without our permission, 2-3 windows will open in the browser, or we will be asked to download new software or browser extensions. They only rely on two things: traffic or ad clicks. The malicious hacker will fully exploit the flaws in the plugins to infect the ads or popups that infect the user, end-user[10-15].In this paper, we use Machine Learning to identify malicious websites.

VI. Machine Learning

Machine Learning is doing their own thing according to their experiences. It is a type of computer algorithm that works automatically based on their past experiences and mistakes. AI (Artificial Intelligence) subsets Machine Learning[8-10]. There are numerous algorithms in machine learning that can be used to train data. Machine learning is widely used in almost all fields to obtain results from computers. To apply this Machine Learning process in any field to obtain the We must feed the required algorithm and data to the system in advance and define the analysis rules for pattern recognition. Following this, the system executes the following tasks:

- Finding the data, retrieving it, and summarising it;
- Making predictions based on the analysis data; • Calculating the probabilities of specific results;
- Adapting to certain developments autonomously; and
- Finding the data, retrieving and summarising the data;
- Making predictions based on the analysis data; • Calculating the probabilities of specific results; • Adapting to certain developments autonomously; and • Optimizing the process based on the recognised pattern.

VII. CONCLUSION

Current approaches in automated analysis and detection of malicious websites have concrete limitations in considering the holistic snapshot of an ideal web page and resilience to possible evasion. Moreover, there are important questions that are yet to be addressed through effective and efficient techniques of feature type identification, maintaining feature evolution, feature composition, and feature value computation. In fact, the analysis and detection techniques which mostly rely on machine learning algorithms also need enhancements in terms of dealing with evolving features, different feature types, and evasion attempts by attackers.

In this position paper, we proposed a holistic approach to effectively and efficiently analyze and detect malicious websites. In particular, we are currently working on enhancement of feature set and feature extraction technique for characterizing malicious payloads by combining page identity, URL tokens, page content, and execution trace so as to capture a complete snapshot of a malicious web page. Extending proposals by past research, we are also investigating effective incorporation of feature evolution, feature types and systematic composition of feature classes to improve analysis and detection of malicious websites. Im- proving the effectiveness and efficiency of training strategies (e.g., training frequency) using online learning techniques and large-scale experimental validation of our approach, using industry and research benchmarks, with live feed of real-life websites is also within the scope of our future plan.

VIII. REFERENCES

[1]. D. Canali, M. Cova, G. Vigna, and C. Kruegel, “Prophiler: a fast filter for the large-scale detection of malicious web pages,” in Proceedings of the 20th international conference on World wide web, 2011, pp. 197–206.

[2]. J. Ma, S. L. K., S. Stefan, and V. G. M., “Beyond black- lists: learning to detect malicious web sites from suspicious urls,” in Proceedings of the 15th international conference on Knowledge discovery and data mining, 2009, pp. 1245–1254.

[3]. M. Qassrawi and H. Zhang, “Detecting malicious web servers with honeyclients,” Journal of Networks, vol. 6, no. 1, 2011.

[4]. A. Dewald, T. Holz, and F. C. Freiling, “Adsandbox: sandboxing javascript to fight malicious websites,” in ACM Sym- posium on Applied Computing, 2010, pp. 1859–1864.

[5]. G. Aggarwal, B. E., J. C., and D. Boneh, “An analysis of private browsing modes in modern browsers.” in Proceedings of the 19th USENIX conference on Security, 2010, pp. 6–6.

[6]. B. S., K. S.T., M. P., and W. M., “Vex: vetting browser extensions for security vulnerabilities,” in Proceedings of the 19th USENIX conference on Security, 2010, pp. 22–22.

[7]. M. Justin, S. L. K., S. Stefan, and V. G. M., “Identifying suspicious urls: an application of large-scale online learning,” in Proceedings of the 26th Annual International Conference on Machine Learning, ser. ICML ’09, 2009, pp. 681–688.

[8]. H. Yung-Tsung, C. Yimeng, C. Tsuhan, L. Chi-Sung, and C. Chia-Mei, “Malicious web content detection by machine learning,” Expert Syst. Appl., vol. 37, pp. 55–60, 2010.

[9]. C. Seifert, I. Welch, and P. Komisarczuk, “Identification of malicious web pages with static heuristics,” in Proceedings of the Australasian Telecommunication Networks and Appli- cations Conference, 2008.

[10]. C. Marco, K. Christopher, and V. Giovanni, “Detection and analysis of drive-by-download attacks and malicious javascript code,” in Proceedings of the 19th international conference on World wide web, 2010, pp. 281–290.

[11]. M. Alexander, B. Tanya, D. Damien, G. S. D., and L. H. M., “Spyproxy: execution-based detection of malicious web content,” in Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium, 2007, pp. 3:1–3:16.

[12]. S. Ford, M. Cova, C. Kruegel, and G. Vigna, “Analyzing and Detecting Malicious Flash Advertisements,” in Proceedings of the Annual Computer Security Applications Conference, 2009.

[13]. A. Ikinci, T. Holz, and F. Freiling, “Monkey-spider: Detecting malicious websites with low-interaction honeyclients,” in Proceedings of Sicherheit, Schutz und ZuverlŁssigkeit, 2008, pp. 407–421.

[14]. B.-I. K., C.-T. I., and H.-C. J., “Suspicious malicious web site detection with strength analysis of a javascript obfuscation,” in International Journal of Advanced Science and Technology, 2011, pp. 19–32.

[15]. R. K., K. T., and D. A., “Cujo: efficient detection and prevention of drive-by-download attacks.” in Proceedings of the 26th Annual Computer Security Applications Conference, 2010, pp. 31–39.

[16]. Y.-M. W., X. J. Doug B., C. V. Roussi R., and S. T. K. Shuo C., “Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities.” in Proceedings of the Network and Distributed System Security Symposium, 2006.

[17]. Rama Krishna, S., Sirajuddin, M. (2022). A Role of Emerging Technologies in the Design of Novel Framework for COVID-19 Data Analysis and Decision Support System. In: Nayak, J., Naik, B., Abraham, A. (eds) Understanding COVID-19: The Role of Computational Intelligence. Studies in Computational Intelligence, vol 963. Springer, Cham. https://doi.org/10.1007/978-3-030-74761-9_14

[18]. Mohammad Sirajuddin, G.Joel Sunny Deol, et al., “Convergence of Blockchain and Computational Intelligence for Industry 4. 0: A Survey”, Computational Intelligence of Blockchain Systems, Nova Science Publisher, ISBN: 978-1-68507-891-1, https://doi.org/10.52305/NNYE5750

Cite this article as :

Dr. Md. Sirajuddin, B. Bhavani, Y. Akshaya, P. Reethika, T. Sriram Reddy, "A Novel Approach for Detection of Malicious Websites using Machine Learning Techniques ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 2, pp. 68-74, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRSET231029

Journal URL : https://ijsrset.com/IJSRSET231029