Journal of Advanced Computer Applications
vol-01, issue-01, December, 2024

ISSN No.: XXXX-XXXX
Pages 69-73

JOURNAL OF ADVANCED COMPUTER APPLICATIONS

ISSN: XXXX-XXXX (Online)
Vol: 01 **Issue:** 01
Contents available at: https://www.swamivivekanandauniversity.ac.in/jaca/

A Malware Analysis Based on Malicious URLs

Dipam Mishra[1*]

[1] *Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore-700121, WB, INDIA*
Dipam.svu@gmail.com

ABSTRACT

Malicious distribution channels on the internet function as platforms for disseminating harmful data, playing a pivotal role in transmitting various cyber threats like spam, malware, adware, spoofing, inappropriate content, and other detrimental resources. This activity exposes individuals to potential exploitation, leading to information disclosure, unauthorized access, financial losses, and even extortion. Cyber attackers frequently exploit web applications to clandestinely access and monitor sensitive data. They lure victims through email, social media platforms, or web searches, redirecting them to malicious URLs, thereby compromising their security and privacy. Conventional methods employed to counter these threats, such as blacklisting, signature matching, and pattern matching, are becoming more intricate due to the growing number of signatures, patterns, and features. Adapting to the dynamic technological landscape necessitates continuous innovation to sustain these protective techniques. In this article, I propose a novel approach to classify malicious URLs. Our method involves comparing benign web-based URLs with established classifiers. By assessing the benign URLs while considering other malicious parameters, we can more efficiently identify and categorize potential threats. The objective is to enhance overall security and protection for web users..

Keyword: Malware, URL, Website, SGD, WHOIS

I. INTRODUCTION

The increasing importance of the World Wide Web (WWW) creates an avenue for malware to compromise both individuals and organizations. This compromise often occurs through the embedding of malicious web URLs containing scripts, exploits, and executable files, allowing for remote access. A uniform resource locator (URL) serves as a unique identifier for resources on the Internet. Attackers seek to manipulate protocols and alter the URL structure to deceive targeted users, redirecting them to fraudulent web pages. This tactic exposes users to malicious unintentional downloads, phishing, social engineering, adware, and spam, resulting in financial losses, sensitive information disclosure, and data extortion.

* Authors for Correspondence

Malicious web URLs can be categorized into four types: malware, spoofing, phishing, and defacement. Malware refers to malicious software that is planted on victims' devices to gain unauthorized access, often associated with system files capable of contaminating government or corporate websites and cloud systems. Spoofing involves compromising victims' personal information, such as usernames, passwords, and credit/debit card details. Victims, thinking they are interacting with a legitimate website, unknowingly communicate with replicated websites featuring similar designs and functionalities hosted on attacker servers. Phishing is accomplished through deceptive emails containing fraudulent web links that compromise end-user information.

Defacement entails altering the legitimate content and structure of a trusted website by injecting malicious scripts. Attackers gain unauthorized access to replace the genuinely hosted website with their malicious version, leading to phishing, code injection, and cross-site scripting. To detect and identify these malware-based malicious web URLs, security researchers have developed various methods such as defence strategies include static analysis, dynamic analysis, mongrel analysis, blacklisting- grounded analysis, and heuristic-grounded analysis to fight malware- grounded vicious web links. stationary analysis involves crawling and examining web links using fine- statistical features (1, 2). Dynamic analysis utilizes tools like the ditz sandbox and Yara autographs to examine networks, relating suspicious malware scripts (3). mongrel analysis combines static and dynamic styles, reducing elusion chances by continuously crawling and covering the gist of suspicious scripts within web URLs. In blacklisting- grounded approaches, URLs are scrutinized against a predefined list of vicious web URLs. Heuristic- grounded styles use known patterns and autographs to overlook suspicious web URLs. still, blacklisting can be finessed by modifying the URL structure, and heuristic- grounded blacklisting fails for web URLs unknown to the scanning module. As per current trends, the discovery of malware- grounded vicious web URLs relies on a combination of autographs, pattern matching, and behavioral analysis (4). Advanced machine literacy or deep literacy ways are employed for this purpose. My exploration utilizes advanced classifiers to classify vicious web URLs grounded on features and actions. From static and dynamic actions, features are insulated and uprooted. These features also suffer analysis by multiple classifiers similar as arbitrary timber, decision tree, redundant tree, Gaussian Naive Bayes, neighbors, SGD, and AdaBoost. The thing is to descry and dissect malware through vicious web URLs, considering delicacy, macro average, weighted normal, perfection, recall, F1- score, and support.

II. METHODOLOGY

The objective of my proposed system for detecting malicious URLs is to analyse, classify, and identify web-based URLs as either malicious or benign.

1.1 Data Processor

In order to identify vicious URLs and classify them as either vicious or benign, the original step involves crawling the web to prize URLs and regularize them. latterly, indexing is necessary to organize and classify them into orders similar as benign, malware, vandalization, and phishing. The coming phase involves rooting features through verbal scanning of the web strings. Eventually, the pre-processed data is divided into two subsets, with 80 allocated for training to fit the model and 20 for testing to assess the fitted model. Selection features are deduced from crawled web- grounded URLs to effectively classify and identify them as either malware or benign. The crucial point groups employed for detecting vicious web URLs include verbal features, which encompass special characters similar as ('@', '?','-',' = ', '.','#',',',',','$', '!'' *',',',',//') that are set up in a URL.

Features based on reputation evaluate a website's reputation and offer indexing. Examples of these web indexes include Alexa, Google, Yahoo, and Baidu. Features based on entropy assess the level of randomness or uncertainty in web URLs. A higher entropy indicates a greater degree of randomness, and this metric is employed to distinguish between normal and abnormal websites.

Features based on content are obtained by either crawling or downloading the entire web page. Malicious websites typically contain sensitive content. The tokenized approach is employed to count sensitive content, focusing on areas such as 'login,' 'sign in,' 'signup,' 'confirm,' 'account,' and 'secure.'

Features based on the host are derived from host attributes, revealing details about the location, identity, technology, and impact level of malicious servers. The specified host attributes include standardizations for censorship, protocols, and regulations.
Features based on the domain encompass 'Who Is' (WHOIS) domain information, covering aspects such as domain name, domain expiration, domain registry, server section, and 'domain length.

1.2 Training Module

During the training phase, the web URLs data undergoes feature extraction to accurately identify whether they are malicious or benign. Subsequently, the extracted data is labelled and subjected to various classification algorithms such as decision trees, extra trees, K-nearest neighbors, Gaussian Naive Bayes, random forests, AdaBoost, and stochastic gradient descent classifiers. The model's performance in the testing phase is based on these algorithms. In the training module, feature extraction involves multiple levels of processing.

Initially, the data is organized based on lexical features, followed by matching against the standard URL structure. The data is then further categorized into normal and abnormal URLs. Following the categorization of abnormal URLs, the URLs are arranged to ascertain the presence of a secure communication channel using either the Hypertext Transfer Protocol (HTTP) or the secure variant, HTTP Secure (HTTPS).

Following the sorting based on the security channel (HTTP or HTTPS), the next step involves comparing with legitimate URL shortening services. Subsequently, after sorting based on the security channel (HTTP or HTTPS), the standard IP address structure that is currently in use is considered. Ultimately, a correlated heat map with appropriate labels is generated by passing the data through multiple layers of extraction processes. This data is then trained using multiple classifiers to achieve the desired results.

1.2 Testing Module

The testing phase relies on the results obtained from training. Initially, the testing web URLs undergo the feature extraction process. Subsequently, these extracted features are analysed by the classifiers to ascertain whether the URL is malicious or benign.

III. CONCLUSION

In this research, I have formulated a strategy employing various classifiers to identify URLs compromised by benign content, malware, phishing, and defacement. The process involves crawling web URLs and extracting both lexical and content-based features to proficiently distinguish between benign and malicious entities. In subsequent stages, I intend to acquire datasets, integrate them with existing systems, and create a user interface for a web application. This implementation can be employed by information security systems or serve as a foundation for further research and development by other researchers.

REFERENCES

1. Saul JKL, Savage S, Voelker MG (2009) Beyond blacklists learning to discover vicious web spots from suspicious URLs. In Proceedings of the 15th ACM SIGKDD transnational conference on Knowledge discovery and data mining. Paris, pp 1245 – 1254. https//doi.org/10.1145/1557019.1557153
2. Wang S, Chen Z, Yan Q, Ji K, Peng L, Yang B, Conti M (2020) Deep and broad URL point mining for android malware discovery. Inf Sci 513600 – 613. https//doi.org/10.1016/j.ins.2019.11.008
3. Kim S, Kim J, Nam S, Kim D (2018) WebMon ML- and YARA- grounded vicious webpage discovery. Comput Netw 137(4) 119 – 131. https//doi.org/10.1016/j.comnet.2018.03.006