Journal of Forensic Sciences
And
Criminal Investigation
ISSN: 2476-1311

Juniper PUBLISHERS
key to the Researchers

Review Article
Volume - 7 Issue - 3 February 2018
DOI: 10.19080/JFSCI.2018.07.555712

J Forensic Sci & Criminal Inves

Malicious Website Detection: A Review

Abdulghani Ali Ahmed*, Nik QuosthoniSunaidi

Faculty of Computer Systems & Software Engineering University, Pahang, Malaysia

Submission: January 25, 2018; **Published:** February 01, 2018

***Corresponding author:** Abdulghani Ali Ahmed, Nik Quosthoni Sunaidi, Faculty of Computer Systems & Software Engineering University, Pahang, Malaysia Email: abdulghani@ump.edu.my

Abstract

Malicious websites represent a critical threat to computer users as hackers use them to take control of many computers simultaneously. Malicious websites distribute all types of malware which may take over the device of their victim and perform several kinds of cybercrimes. Many solutions have been developed to address the detection and prevention of malicious websites in real time. Although these strategies are helpful, they are still prone to several problems that may critically affect the efficiency and capability of identifying and preventing online cybercrimes, which continuously change and evolve. This paper reviews the existing methods and techniques of malicious websites detection. This paper also analyzees the features of the reviewed methods and critically discusses their limitations.

Keywords : Malicious websites; Drive by; Phishing; Detection

Introduction

As the usage of internet continue to grow exponentially, the growth of malicious website is also seen in the environment. These malicious websites are created for unlawful purposes, for example, spam-advertising, malware propagating, and financial fraud through phishing and malvertising. As an example, on the aspect of malvertising activities, its growth has been recorded at a hundred and thirty-two percent in 2016 [1]. It is also reported by Symantec in 2016 [2] that seventy- six percent of serve scanned were fund to have vulnerabilities which could mean that it harbors malicious intention or codes. This would include luring unsuspecting uses to a spoof sites and obtain key information by impersonating valid sites [3]. These malicious websites not only steal or damage the info from users, but also let the hackers to regulate the computers. It becomes a platform helping diverse Internet crimes. Because of this, detecting the malicious sites to stay away from the damage is of a significant priority. What's more, these websites often appear to be genuine websites. Sometimes it will request an unsuspecting user to install software that the machine seems to need. Another one is, a video website might request for a codec installation which in turns compromise the machine itself.

There have been many works in malicious websites detections efforts. The pioneering effort begins with the blacklist approach [4]. A blacklist is a list containing IP address information, website name or URLs of known malicious websites such as where it is listed at sites such as phishtank.com and vxvault.net. These sites provide credible validation of whether the site is malicious or not since it is based on real feedback from those who discovered it and impacted by it. Though the credibility is high the speed of which it is updated is quite slow and since it is updated after and impact has potentially been made it is reactive in nature. Furthermore, it needs an extended time period from finding, verifying and upgrading a fresh malicious website to the blacklist. In the on the other hand, users are threatened by the new harmful websites. Moreover, the price and enough time required by making a harmful website are low. Any new malicious website can't be detected by blacklist-based system given that they have different IP addresses and domain names. Beyond the reactive nature of malware detection, there are also significant works done in the proactive approaches of malware websites detection which includes blacklist [5], honey clients [6], machine learning techniques [7] and webpage content [8].

Related Works

In the pursuit of malicious website detection, a list of efforts would include leveraging capabilities that would include approaches such as blacklist, honey clients, machine learning, web page content and an integrated approach.

Blacklist Approach

A regular strategy for supporting users avoid malicious websites is by using the blacklist via known bad websites. Microsoft Internet Explorer, Google Chrome and Mozilla Firefox warn users when they make an effort to visit a web page present on a blacklist. This might also include leveraging information from known exterior websites such as vxvault.net and phishing.com for example. These techniques are suffering from a few shortcomings like the need to upgrade it periodically, sluggish to reveal new harmful websites and poor overall coverage of these destructive websites. However the information in the blacklist sites are valuable when used as an initial exposure for training platform [9] and predictions of potential malicious domain in the environment [10].

Honey clients

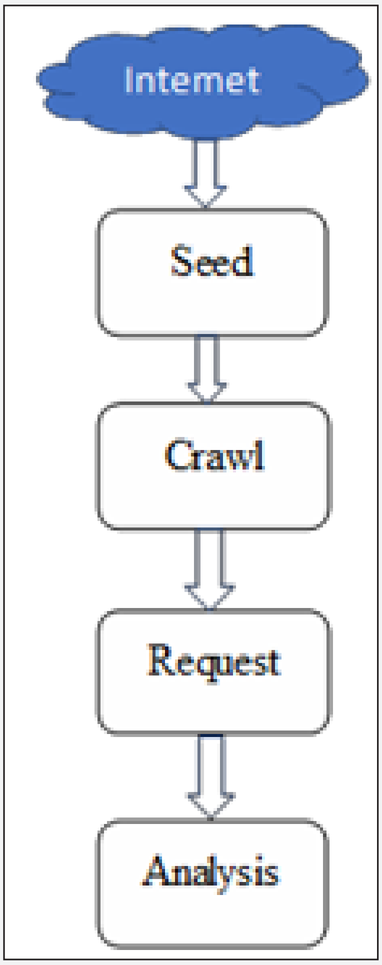

Figure 1: General Overview of Honey client.

Honey clients approach is considered as an effective technique in malicious website detection. Honey clients are security tools used to recognize websites that host harmful content. People to web sites are attacked whenever a website's malicious content exploits vulnerabilities within the browser or the browser's plug-ins. Honey clients works by crawling on the internet and scowering for malicious websites and analyze drive-by-download attacks [11]. These honey clients can be categorized by low interaction or a high interaction. Low interaction honey clients uses an emulator to do the crawling of the internet. Thus the risk of itself being infected is low [12]. High-interaction honey clients on the other hand are powered by the basic principle of monitoring system condition changes throughout a website visit using real web browsers installed on the host's system. Websites are categorized as destructive if the web browser accesses or tries to modify supervised security-sensitive listing and files [13] (Figure 1).

Machine Learning

In the machine learning space, the works on detecting malicious websites focuses on URLs detectors. In the machine learning technique, it makes use of a collection of URLs as training data, and predicated on the statistical properties, find out a prediction function to classify an URL as malicious or benign. Thus, giving them the capability to generalize to brand-new URLs unlike blacklisting strategies. The primary requirement of training a machine learning model certainly is the existence of training data. In the context of malicious URL detection, it would definitely correspond to a collection of large numbers of URLs. By extracting decent feature representations of URLs, working out a prediction model on training data of both malicious and benign URLs will occur. Key features would include lexical features which are obtained from the URL string. The more common lexical features used include statistical properties of the URL string, its length and length of each components (Hostname, Primary domain etc) [14].

In the machine learning approaches, these features are extracted via static and dynamic features. In static approach, the URL is analyzed without the need to execute it. This method is considered safer than dynamic analysis due to the fact that in dynamic analysis mode, monitoring the behavior of the systems calls for abnormal behaviors which usually are potential victims [15], to consider any anomaly that have intrinsic risks, and so are difficult to implement and generalize. There are a multitude of machine learning algorithms in literature which can be directly utilized in the context of Malicious URL Detection such as Support Vector Machine [14-17], Logic Regression [17-19], Naïve Bayes [8,18-20] and Decision Tree respectively [17,21-22].

Page Content Analysis

Page content based approach is the more detail analysis in comparison to the URL based approach. This would require the most processing and analysis time as considerable information need to be extracted from a particular website. Should the URL-based features neglect to identify a malicious URL, according to [18] this technique is a far more thorough analysis of the content-based features which can help in early threat detection. This approach derives key features such as HTML and JavaScript [8] proposed a malicious webpage detection based on dynamic HTML and Java Script using decision tree structured algorithm [23] on the other hand, proposed CANTINA which uses machine learning approach and leveraging HTML Document Object Model, search engines and third-party tools [24-27].

Critical Analysis

(Table 1)

Table 1: In this paper, we looked into the features anddrawbacks of each method of malicious website detection.

Name	Feature
Blacklist	Uses precompiled list of known malicious Sites. The accuracy and validity is both high and based on communal feedback.
	Drawback
	Resource constraint which required periodic updates and hackers actively evade blacklists by making minor changes to the original URL [28]
Honey Client	**Feature**
	Proactively crawls the internet and detects malicious website in a low interaction or high interaction mode.
	Drawback
	Prone to evasion by malicious site owners [6].
Machine Learning	**Feature**
	Uses already existing information from the URL and develops a learning model to classify whether a site is malicious or benign. Classification algorithm can include Support Vector Machine, Decision Trees, etc.
	Drawback
	Finding correct training data is a challenge due to the generous number of instances and features [29]
Page Content	**Feature**
	Inspects the page content and does matching calculations through comparisons with valid pages and a set of specified base rules.
	Drawback
	Requires sizable time when querying, as
	Example in the Google [3].

Conclusion

In this review, we looked at multiple approaches that has been developed in detecting malicious websites. Malicious URL detection plays a crucial role for most cyber security applications, and the detection efforts are plays a crucial part of it. In this review, we carried out a review on Malicious URL detection using methods such blacklisting, honey customers, machine learning and web page content analysis methods. In this review, we categorized most, if not absolutely all, its features, related works connected and how it operates. Finally, we highlighted its features and drawbacks. We're able to see still even more opportunities specifically in the device learning space in improving the Malicious Website detection agenda.

References

1. Arghire I (2017) Malvertising Jumped 132% in 2016.
2. Wood P (2016) Internet Security Threat Report 21: 79.
3. Ahmed AA, NA Abdullah (2016) Real time detection of phishing websites in Information Technology, Electronics and Mobile Communication Conference (IEMCON), 7th Annual.
4. M Justin (2011) Learning to detect malicious URLs. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3): 30.
5. Eshete B (2013) Effective analysis, characterization, and detection of malicious web pages. In Proceedings of the 22nd International Conference on World Wide Web ACM pp. 355-360.
6. Qassrawi MT, H Zhang (2011) Detecting Malicious Web Servers with Honeyclients. JNW 6(1): 145-152.
7. Damodaram, RD (2012) Experimental Study on Meta Heuristic Optimization Algorithms for Fake Website Detection. International Association of Scientific Innovation and Research (IASIR) pp. 43-53.
8. Hou YT (2010) Malicious web content detection by machine learning. Expert Systems with Applications 37(1): 55-60.
9. Likarish PF (2011) Early detection of malicious web content with applied machine learning. The University of Iowa.
10. Felegyhazi M, Kreibich C, V Paxson (2010) On the Potential of Proactive Domain Blacklisting LEET.
11. Akiyama M (2010) Design and implementation of high interaction client honeypot for drive-by-download attacks. IEICE transactions on communications 93(5): 1131-1139.
12. ChibA D (2012) Detecting malicious websites by learning IP address features. IEEE/IPSJ 12th International Symposium on Applications and the Internet, SAINT in Applications and the Internet (SAINT), IEEE/IPSJ 12th International Symposium on Applications and the Internet pp. 29-39.
13. Mansoori M, I Welch, Q Fu YALIH (2014) yet another low interaction honeyclient. In Proceedings of the Twelfth Australasian Information Security Conference 149: 7-15.
14. Kolari P, T Finin, A Joshi (2006) SVMs for the Blogosphere: Blog Identification and Splog Detection. in AAAI Spring Symposium Computational Approaches to Analyzing Weblogs.
15. Canfora G (2014) Detection of malicious webpages using system calls sequences in International Conference on Availability, Reliability, and Security.
16. Ahmed, Abdulghani Ali, Aman Jantan, Tat-Chee Wan (2016) "Filtration model for the detection of malicious traffic in large-scale networks." Computer Communications 82: 59-70.

17. Ma j (2009) Beyond blacklists: learning to detect malicious web sites from suspicious URLs in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

18. Canali D (2011) Prophiler: a fast filter for the large-scale detection of malicious web pages in Proceedings of the 20th international conference on World Wide Web. ACM.

19. Ahmed Abdulghani Ali, Chua Xue Li (2018) "Analyzing data remnant remains on user devices to determine probative artifacts in cloud environment." Journal of forensic sciences 63(1): 112-121.

20. Ahmed Abdulghani Ali (2017) "Investigation Approach for Network Attack Intention Recognition." International Journal of Digital Crime and Forensics (IJDCF) 9(1): 17-38.

21. Seifert C, I Welch, P Komisarczuk (2008) Identification of malicious web pages with static heuristics. IEEE. Telecommunication Networks and Applications Conference.

22. Ahmed Abdulghani Ali, Nurul Amirah Abdullah (2016) "Real time detection of phishing websites." In Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual, pp. 1-6. IEEE.

23. Xiang G (2011) Cantina+: A feature-rich machine learning framework for detecting phishing web sites. ACM Transactions on Information and System Security (TISSEC) 14(2): 21.

24. Dunlop M, Groat S, Shelly D (2010) Goldphish: Using images for content-based phishing analysis IEEE Internet Monitoring and Protection (ICIMP), 5th International Conference.

25. Ahmed Abdulghani Ali, Chua Xue Li (2018) "Analyzing data remnant remains on user devices to determine probative artifacts in cloud environment." Journal of forensic sciences 63(1): 112-121.

26. Ahmed Abdulghani Ali, Ali Safa Sadiq, Mohamad Fadli Zolkipli (2016) "Traceback model for identifying sources of distributed attacks in real time." Security and Communication Networks 9(13): 2173-2185.

27. Ahmed Abdulghani Ali, Aman Jantan, Tat-Chee Wan (2011) "SLA-based complementary approach for network intrusion detection." Computer Communications 34(14): 1738-1749.

28. Prakash P (2010) Phishnet: predictive blacklisting to detect phishing attacks IEEE INFOCOM 2010 Proceedings IEEE.

29. SahoO D, C Liu, SC Hoi (2017) Malicious URL Detection using Machine Learning: A Survey.

DOI : 10.19080/JFSCI.2018.07.555712