International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

A Study on Malicious Web Page Detection

Syeda Farheen Sultana & Dr. Sameena Banu
[1]M. Tech Student KBNCE, Kalaburagi
[2] CSE Dept., KBNCE, Kalaburagi

Abstract:

Web security is a challenging issue due to emerging trends in the web attacks. Malicious websites steal the valuable information of the visitors and infect their system for further attacks. Various methodologies are proposed to detect the malicious websites based on features like web contents, HTML codes, session information, and dynamic behaviors. This paper classifies the detection methods in three categories- static, dynamic and hybrid approaches. The limitations in these methods are discussed. This paper also describes the difficulties in classification methods, reliability of test data sources, limitation of various features, and their collection methods. The detailed analysis carried out in this paper provides a new road map for the research in this area.

Keywords

Malicious Websites, Detection, Machine learning algorithm, Classification methods

1. INTRODUCTION

The web attacks are the challenging issues of the web community. When the user visits the malicious web site the attack is initiated through various features (lexical, domain, path, web content and hyperlink etc.) [1]. To prevent the user against accessing the malicious websites, several automated analysis and detection methods have been proposed. The attackers lure the visitor to access malicious web sites and they steal crucial information from the client machine or install the spyware for further exploits. Dynamic HTML gives attackers a new and powerful technique to compromise the security of computer systems. A malicious dynamic HTML code is usually embedded in a normal webpage. The malicious webpage infects the victim when a user browses it. Furthermore, such DHTML code can disguise itself easily through obfuscation or transformation, which makes the detection even harder. So, detecting and preventing the user from these attacks are significant task. A huge number of attacks have been observed in last few years. Malicious attack detection and prevention system plays an immense role against these attacks by protecting the system's critical information. The internet security softwares and fire walls are not enough to provide full protection to the system. Hence efficient detection systems are essential for web security.

The malicious web site detection methods are classified into three types as shown in the figure 1. They are static approaches, dynamic approaches and hybrid approaches [2]. The various features, algorithms and datasets are discussed. The limitations in various classification methods are also analyzed. In this paper we propose a research roadmap for improvement to overcome the limitation of the existing techniques in detecting malicious websites.

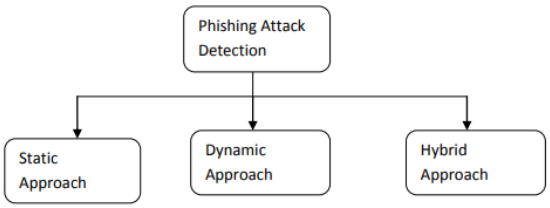

Figure 1 Types of Malicious URL detection methods

2. LITERATURE SURVEY

2.1 Static Approach

The static attacks depend on the source code and features such as URL Structure, host-based information, webpage content and characterization of malicious payload. Most of the detection techniques are signature based approaches which compares the URL with the black listed profiles [3]. But the phishers keep changing their websites in short time span, so these methods are not efficient. The malicious web sites are the corner stone of criminal activities, as a result there has been broad interest in developing systems to prevent the end user from visiting such sites. Justinma et al [14] proposed an approach for classifying URLs automatically as either malicious or benign based on supervised learning across both lexical and host-based features. This approach is complementary to black listing which cannot predict the status of unknown URLs.

2.2 Dynamic Approaches

DaHuang et al [8] identify the malicious URLs based on dynamically extracted lexical patterns from URLs (ex: *PayPal.*.cgi.*/login.php"). They developed a new method to mine their URL patterns, which are not assembled using any pre-defined items and thus cannot be mined using any existing frequent pattern mining methods. It can provide new flexibility and

capability on capturing malicious URLs algorithmically generated by malicious programs.

Hossian Shahriar and Mohammed zulkernine [13] proposed a tool Phish Tester to test the trustworthiness of the website based on the behaviour of the web application. They use finite state machine to represent the behaviour of the web application. They classify the genuine and malicious behaviour of the website using state traversal. In [6], a fast pre-filtering technique combining URL structure, host-based information and page content is proposed and is demonstrated to significantly reduce the execution load of a dynamic analysis technique. The general limitation of considering only page content is the high risk of obfuscated content (e.g., multilevel obfuscation of JavaScript code).

2.3 Hybrid approaches

Honggeunkim et al [12] proposed a method to detect malicious web pages. The proposed method introduces a pattern-based static analysis for detecting web attacks efficiently. A high-interaction client honey pot performs the static analysis before carrying out execution-based dynamic analysis. The static analysis classifies sample web pages into two groups, the first one assumed to be attacks and the second one assumed to be without attacks. Then dynamic analysis is performed using sequential visitation algorithm for the filtered pages to ensure the malicious attack. This method reduces cost of identifying the malicious web pages. Birhanu Eshete et al [5] proposed an approach called BINSPECT to defend against malicious web pages fall into two major blocks, i.e., static analysis and dynamic analysis techniques. This approach applies supervised learning techniques in detecting malicious web pages pertinent to drive-by-download, phishing, injection, and malware distribution by introducing new features (URL features, Page-Source features like HTML and JavaScript, and Social-Reputation features) that can effectively discriminate malicious and benign web pages. Experimental evaluation of BINSPECT in large scale achieved above 97% accuracy with low false signals. Igor Santos et al [10], developed OPEM, a hybrid unknown malware detector which combines the frequency of occurrence of operational codes (statically obtained) with the information of the execution trace of an executable (dynamically obtained). They proved that this hybrid approach enhances the performance of both the above approaches.

3. ISSUES IN STATED APPROACHES

3.1 Issues in Static approaches

The commonly used protection technique is blacklisting of known malicious URLs and IP address collected through manual reporting, data sources, honey part and custom analysis techniques. This approach uses various lexical features of URL. This light weight approach is easy to deploy and use. This approach is effective only when one can exhaustively analyze the malicious web site and the update the black list regularly. The drawback is the inability to find the new websites even if they are malicious. A huge number of false positives are reported due to incorrect analysis. So, these approaches should be improved. Another drawback of this method is that it can be slow due to time consuming verification process. Nowadays, the weapon of choice in combat against malicious software is signature-based anti-virus scanners that match a pre-generated set of signatures against the files of a user. These signatures are created in a way so that they only match malicious software. This approach has at least two major drawbacks. First, the signatures are commonly created by human analysts. This, often, is a tedious and error-prone task. Second, the usage of signatures inherently prevents the detection of unknown threats for which no signatures exist. Thus, whenever a new threat is detected, it needs to be analyzed, and signatures need to be created for this threat. After the central signature database has been updated, the new information needs to be deployed to all clients that rely on that database. The signatures are created by human analysts, unfortunately there is room for error.

3.2 Issues in dynamic approaches

The behaviour based model is dynamically detecting the malicious attack in web page. They also have some limitation. The Finite State Machine (FSM) [13] model uses the various states of the malicious behaviour and they detect the malicious website based on their state traversals. But this approach only detects the attacks based on predefined states(behaviour). This method is not capable of detecting random inputs and new behaviors. Malicious URL are detected by dynamically mining the lexical patterns [8] of the URL. The complete pattern set algorithm and greedy selection algorithms are used for this purpose. As the size of data set increases, the algorithms running time also increases drastically. So, the existing pattern selection algorithms are not delivered a desirable performance, so a better pattern selection algorithm is needed. Bottracer is a tool to detect bot like malware on end systems through detecting the bot start-up, preparation, and attack behaviour during execution. This tool implements a prototype of Bot-tracer based on VMware and Windows XP Professional. But if a bot first detects user activities before it launches itself, the current BotTracer would fail to detect such

bots. They also fail to detect time bomb bots. SpyProxy [4] analyzes the behaviour of the website through execution. It is not able to detect the dynamically changing malicious content. However, highly interactive web pages resemble general-purpose programs whose execution paths depend on non-deterministic factors such as randomness, time, unique system properties, or user input [9]. An attacker could use non-determinism to evade detection. For example, a malicious script could flip a coin to decide whether to carry out an attack, this simple scheme defeats Spy Proxy 50% of the time. Guanghui Liang at al [11] developed a classification technique to detect malicious websites. A dynamic analysis is used to capture API calls and other running information of the malware. Finally, a similarity comparison algorithm is used to diagnose the degree of similarity between malware variants. This method is not capable of identifying anti-detection malware.

3.3 Issues in Hybrid approaches

Though static and dynamic approaches yield high performance, they took long time to identify the malicious web pages and tend to miss some attacks like time bomb [12]. This approach contains two phases static analysis and dynamic detection, so the model is complex and difficult to adopt new changes. This model requires lot of training before deploying it in real time. For example, Cujo a hybrid system for detection and prevention of JavaScript attacks, the detection procedure is repeated for 10 times to report the result. The hybrid approaches increase the performance overhead cost. So, the hybrid approaches are effective in detecting attacks. But increase in detection time is the major drawback of this approach.

4. CHALLENGES IN THE DETECTION METHODS

Most of the existing methods to detect malicious web site are based on their core techniques for a well-known attack. But the attacker invents changes in the existing approach and introduces new techniques to be embedded in webpage. The existing approach rely on the fixed set of features but the attacker makes changes in the existing features and introduces new features. As a result, the detection methods are not able to detect the new attacks. So, the analysis and detection techniques need to be improved. The various techniques like signature based, features based and behaviour based approaches to detect malicious website and contents are facing these limitations due to sophisticated invasions. Due to the limitations the various existing features are not sufficient to detect malicious websites. For example, existing approaches are not able to detect malicious websites based on the domain name because the attacker frequently changes the domain. Apart from that none of the feature collection techniques can collect the emerging features. The existing detection methods suffer a lot from the true and false negatives. So, there is need a for new approach to overcome all these limitations.

The performance is a major problem. Most of the detection methods affects the performance of the system. The hybrid approaches consume more time due to their analysis and detection phases. Most of our real-time applications like financial management, health care etc., are time critical applications. So, time efficiency needs to be addressed. The emerging features, limitation of the detection method and performance are the major challenges in detecting malicious web sites. Hence these issues need to be considered while designing a new technique.

5. CONCLUSION

Most of the existing approaches to detect malicious websites have concrete limitations due to the emerging techniques in malicious attacks. The analysis and detection techniques rely on machine learning algorithms also need to be improved in terms of dealing evolving features, different feature types, and evasion attempts by attackers. Various static, dynamic and hybrid approaches to detect malicious websites are analyzed. The static approaches alone are not sufficient to detect the emerging threats of web application. The dynamic and hybrid approaches opt for the present scenario. The dynamic and hybrid approaches consume more time for detection. The limitations in classification techniques and issues in data sources are explained.

6. REFERENCES

1. Aaron Blum, Brad Wardman and Thamar Solorio. “Lexical Feature Based Phishing URL Detection Using Online Learning”, 3rd Workshop on Artificial Intelligence and Security, Chicago, Illinois, USA, pp. 54-60, October 8, 2010.

2. Aishwarya Vishwakarma and Niket Bhargava "A combination approach for Web Page Classification using Page Rank and Feature Selection Technique”, International Journal of Computer Theory and Engineering, Vol.2, No.6, pp. 897-900, December 2010.

3. Anh Le, Athina Markopoulos and Michalis Faloutsos, "PhishDef: URL Names Say It All", in Proceedings of IEEE INFOCOM, pp.191-195, 2011.

4. 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, pp. 3:1–3:16, 2007.

5. Birhanu Eshete, Adolfo Villaorita, and Komminist Weldemariam." BINSPECT: Holistic Analysis and Detection of Malicious Web Pages", In Proceedings of Security and Privacy in Communication Networks (SECURECOMM), Padua, Italy, September 2012.

6. 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, Hyderabad, India, pp. 197–206, 2011.

7. D. K. McGrath and M. Gupta, "Behind phishing: An examination of phishermodi operandi," in Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats(LEET), San Francisco, USA, 2008.

8. Da Huang, Kai Hu and Jian Pei "Malicious URL Detection by Dynamically Mining Patterns without Pre-Defined Elements", International Journal of Word Wide Web. Springer US. Vol26 Issue 1, 10th August 2013. DOI10.1007/s11280-013-0250-4.

9. Dinaburg, A., Royal, P., Sharif, M. and Lee, W," Ether: Malware Analysis via Hardware Virtualization Extensions", In Proceedings of the 15th ACM Conference on Computer and Communications Security, CCS'08, Alexandria, pp. 51-62, October 2008.

10. Igor Santos, Jaime Devesa, Felix Brezo, Jaview Nieves, and Pablo G. Bringas," OPEM: A staticdynamic approach for machine-learning-based malware detection", In Proceedings of InternationalJoint Conference CISIS'12-ICEUTE´ 12-SOCO´ 12 Special Sessions. Springer Berlin Heidelberg, German, pp. 271-280, 2013.

11. Guanghui Liang, Jianmin Pang, and Chao Dai, "A Behavior-Based Malware Variant Classification Technique", International Journal of Information and Education Technology, Vol.6 (4), pp. 291-295, April 2016.

12. HongGeun Kim, Dong-Jin Kim, Seongje Cho, Moonju Park, and Minkyu Park. "Efficient Detection of Malicious Webpages using High-Interaction Cline Honeypots", Journal of Information science and engineering, Vol.28, Issue 5, pp.911-924, May 2012.

13. Hossain Shahriar and Mohammad Zulkernine, "Trustworthiness testing of phishing websites: A behaviour model-based approach", International Journal of Future Generation Computer System. Volume 28, Issue 8, pp.1258- 1271. October 2012).

14. Justin Ma, Lawrence K. Saul, Stefan Savage and Geoffrey M. Volker, "Identifying Suspicious URLs: An Application of Large-Scale Online Learning", In Proceedings of the 26th Annual International Conference on Machine Learning. ACM New York, USA 2009. DOI:10.1145/1553374.1553462.