WEB APPLICATION SCANNING FOR MALWARE ATTACK DETECTION WITH PROVIDE APPROPRIATE INCIDENT REPORT BY USING HYBRID METHOD

AINA NABILA BINTI ABDUL RAZAK

FSKTM 2019 25

WEB APPLICATION SCANNING FOR MALWARE ATTACK DETECTION WITH PROVIDE APPROPRIATE INCIDENT REPORT BY USING HYBRID METHOD

By

AINA NABILA BINTI ABDUL RAZAK

Thesis submitted to the Faculty of Computer Science and Information Technology, University Putra Malaysia, in fulfillment of the requirements for the Master of Information Security

JUNE 2019

COPYRIGHT PAGE

DEDICATIONS

"This sweet dedication goes to respected lecturers, thoughtful friends and supportive family"

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Information Security

WEB APPLICATION SCANNING FOR MALWARE ATTACK DETECTION WITH PROVIDE APPRORIATE INCIDENT REPORT BY USING HYBRID METHOD

By

AINA NABILA BINTI ABDUL RAZAK

JUNE 2019

Supervisor: Dr. Noor Afiza Mohd Ariffin
Faculty: Faculty of Computer Science and Information Technology

Nowadays, antivirus software is one of the ways to measure the increasing number of malware not only on the computer but also on the information system as well as the software that needs to be protected from any attacks. The malware detection process becomes a challenge because the attacker has a new technique to penetrate it. Most anti-virus software uses unmatched signatures to prevent the increase in the number of malware variants.

Signature is a unique confirmation for binary files. It is created by binary file analyzer using static analysis method. In addition, the next analysis is known as the dynamic analysis that requires behavior and action during execution to identify whether it can be malware or not. Both methods have their own advantages and disadvantages.

This project proposes a static and dynamic analysis method of combining to produce a method known as hybrid.

It will analyze as well as classify files vulnerable to unknown malware. Additionally, in order to create this method, it is necessary to use a machine learning where a malware program is used as a data set. Feature vectors have been selected by analyzing binary code and dynamic

behavior. The hybrid method uses the advantages of static and dynamic analysis and impact rather than it will improve the classification results. Therefore, expecting this approach is able to detect time and accuracy taken for each method to detect malware detection attack which lead to results.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia Sebagai memenuhi keperluan untuk Ijazah Sarjana Keselamatan Maklumat

WEB APPLICATION SCANNING FOR MALWARE ATTACK DETECTION WITH PROVIDE APPROPRIATE INCIDENT REPORT BY USING HYBRID METHOD

Oleh

AINA NABILA BINTI ABDUL RAZAK

JUN 2019

Penyelia: Dr. Noor Afiza Mohd Ariffin

Fakulti: Fakulti Sains Komputer dan Teknologi Maklumat

Kini, perisian antivirus adalah salah satu cara bagi menyukat pertambahan bilangan "malware" bukan sahaja hanya pada komputer malah kepada sistem maklumat juga perisian yang perlu dilindungi daripada sebarang serangannya. Proses pengesanan "malware" menjadi cabaran kerana penyerang mempunyai pelbahai teknik baru untuk menembusinya. Kebanyakan perisian antivirus menggunakan tandatangan yang tidak sesuai untuk menghalang peningkatan bilangan varian "malware".

Tandatangan adalah pengesahan unik untuk fail binari. Ia tercipta dengan penganalisis biner file menggunakan kaedah analisis statik. Selain itu, analisis seterusnya dikenali sebagai dinamik analisis yang memerlukan tingkah laku dan tindakan semasa pelaksanaan bagi mengenal pasti sama ada boleh terjadinya malware atau tidak. Kedua-dua kaedah mempunyai kelebihan dan kelemahan sendiri.

Projek ini mencadangkan kaedah analisis statik dan dinamik bergabung bagi menghasilkan kaedah yang dikenali sebagai "hybrid".

Ia akan menganalisis sekaligus mengklasifikasikan fail boleh terdedah kepada “malware” ataupan fail yang tidak diketahui. Selain itu, bagi mencipta kaedah ini, ia perlulah menggunakan mesin pembelajaran dimana program malware digunakan sebagai set data.

Vektor ciri telah dipilih dengan menganalisis kod binari serta tingkah laku yang dinamik. “Hybrid” method menggunakan kelebihan daripada analisis statik dan dinamik dan impak daripada itu akan memperbaiki hasil klasifikasi. Oleh itu, diharapkan pendekatan ini dapat mengesan masa dan ketepatan yang diambil untuk setiap kaedah untuk mengesan serangan pengesanan “malware” yang membawa kepada keputusan.

ACKNOWLEDGEMENT

First of all, I feel grateful and thanks Allah SWT for His mercy and bless. During this period, I learned a lot of new things especially relating to Security in Information Technology. This thanksgiving and appreciation are also directed to my supervisor, Dr. Noor Afiza Mohd Ariffin who always gives new ideas and knowledge provided for improvement so that the project is successful. Next my appreciation goes to my parent En Abdul Razak B. Mat Jidin and Puan Zarimah bt Jaafar, who always give endless moral and encouragement support both physically and mentally. Not forgetting, thanks to my friends who are not tired of exchanging opinions and comments that can improve my project. May Allah bless you all.

APPROVAL FORM

This thesis was submitted to the Senate of University Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Master Information Security. The members of the Supervisory Committee were as follows:

DR. NOOR AFIZA MOHD ARIFFIN

Faculty of Computer Science and Information Technology

University Putra Malaysia

(Supervisor)

Date: June 2019

DECLARATION FORM

Declaration by graduate student

I hereby confirm that:

- This thesis is my original work
- Quotations, illustrations and citations have been duly referenced
- This thesis has not been submitted previously or concurrently for any other degree at any other institutions
- Intellectual property from the thesis and copyright of thesis are fully-owned by University Putra Malaysia (UPM)
- Written permission must be obtained from supervisor and Deputy Vice-Chancellor (Research and Innovation) before thesis is published in book form
- There is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity was upheld as according to Rule 59 in Rules 2003 (Revision 2013-2014). The thesis has undergone plagiarism detection software

Signature: _____ Date: _____

Name and Matric No.: AINA NABILA BINTI ABDUL RAZAK (GS 50602)

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.0 Introduction

As for introduction, this section will briefly explain about research background, highlight the problem statement and research objective also define the research scope.

1.1 Background Research

In this background research, there is a brief description about related field of study for this project. Background research is mainly to collect information to have understanding in-depth about related subject. For this project, understanding about malware especially using both method either static or dynamic also the machine language skills are need to have overall overview about this project. Besides, study about malware detection system is also important in order to know how malware analysis is being done.

The Internet is one of the important sources used in everyday life not only simplifies everyday processes such as online banking payments, connect with other users such as Facebook and so on. Therefore, internet users are likely to face security threats due to malware attacks. Malware refers to malicious programs or files and will have an adverse

effect on computer users [1]. Among malware are computer viruses, worms, trojan and spyware. Malware inflicts on various objectives such as stealing, encrypting or removing sensitive data, altering or retrieving the rights of core computing functionality as well as monitoring the activities of computer users without their knowledge and consent. Because of this, the number of everyday and existing emerging malware in this extremely high range will evolve in their structure as well as difficult to detect.

1.1.1 Malware

Malicious software also known as Malware is any malicious code in software that can be used to compromise computer operations, collect sensitive information, do illegitimate action on data, gain access to private computer resources, host or networks. It will damage computer programs without user consent. Malware are able to exploit resource from various system platforms. Malware comes with different types of threats such as virus, worms, Trojans, rootkits and so on. Malware considered as high-level issues because it had potential to attack the security goals which are confidentiality, integrity and availability.

In the era of Internet of Things (IoT) technology, potentially malware to attack any software, computer, server or network is from various ways. Each types of malware have its own technique to damaging computers and data, so it comes with different malware challenge and removal method to prevent it. By represent the malware analysis the organization will determine potential new threat and defend themselves against malware attacks.

Antivirus is one of malware analysis tool in order to distinguish malicious from benign code. Anti-virus use signature based to classify the malware for identify unknown malware programs. Every signature has their unique identity of binary file which will compare them to a database of malware programs. There consist of method to created signature either by using static, dynamic or combination of both knows as hybrid as will have stored in signature database. [1] There consist 4 most common evasive technique used by malware design to avoid detection and analysis:

1. Environmental awareness

This will give permission for malware samples to detect underlying runtime environment of the system that will infect. This technique will allow malware to search differences between virtualized that compare to bare metal environment.

2. Confusing automated tools

This tools purpose will allow malware to avoid and prevent detection from signature-based antivirus software. It will be changing the malware domain on daily basis to increase the difficulty of blocking traffic associated with malware.

3. Time based evasion

This type behaviour used by malware to exploit certain action taken by the user that will include opening a new window and once user click some links to activate only after the system is reboots and running at specific date and time. This purpose will check from time to time either the system is infected by machine hardcoded into the executable and allow Black POS in periods while remain dormant at rest of time.

4. Obfuscating internal data

Malware might use any number of tricks to run code that cannot be detected using the analysis system. This technique will replace API names with hash value and used it to ignore certain process from being communicate server and effective encrypts the traffic. It will increase the difficulty for system to identify ROM's malicious nature.

1.1.2 Static and Dynamic Method

Malware analysis could be retrieved either by using static or dynamic method. Static analysis act to examine the malware without actually running it while dynamic analysis will execute malware in a controlled and monitor the environment to observe detailed particular process of malware detection that will analyze the whole process behavior of malware. Each technique comes with different element to categorize either as basic or advanced. There consist of some advantage and limitation based on the method of malware analysis.

[2] Static malware analysis uses a signature-based approach for example that involves file fingerprints, virus scanning, reverse engineering the binary, file obfuscation, analyze memory artifact, packer detection and debugging. Signature based is identify the presence of malware that infect by match at least one-byte signature also known as blacklist. It is ineffective against the sophisticated malware programs and codes. Static analysis fails at different code by using obfuscation technique used by virus coders also polymorphic and metamorphic malware but there is advantage from binary code information that contains very useful information about malicious behavior of program in term of code sequence and parameters [3].

Dynamic malware analysis uses a behavior-based approach for malware detection that will analyze the suspicious activity. It involves the API call, by intrusion detection traces, any changes of registry, calls for network and system also the memory write. It is effective against all types of malware because it will execute the sample of malware. However dynamic also have some limitation for obfuscation techniques and polymorphic malware but it is necessary complement compare to static approach [4]. Hybrid is the combination of Static and dynamic technique. This project will use integrated static and dynamic method which known as hybrid to analyze the malware attack detection.

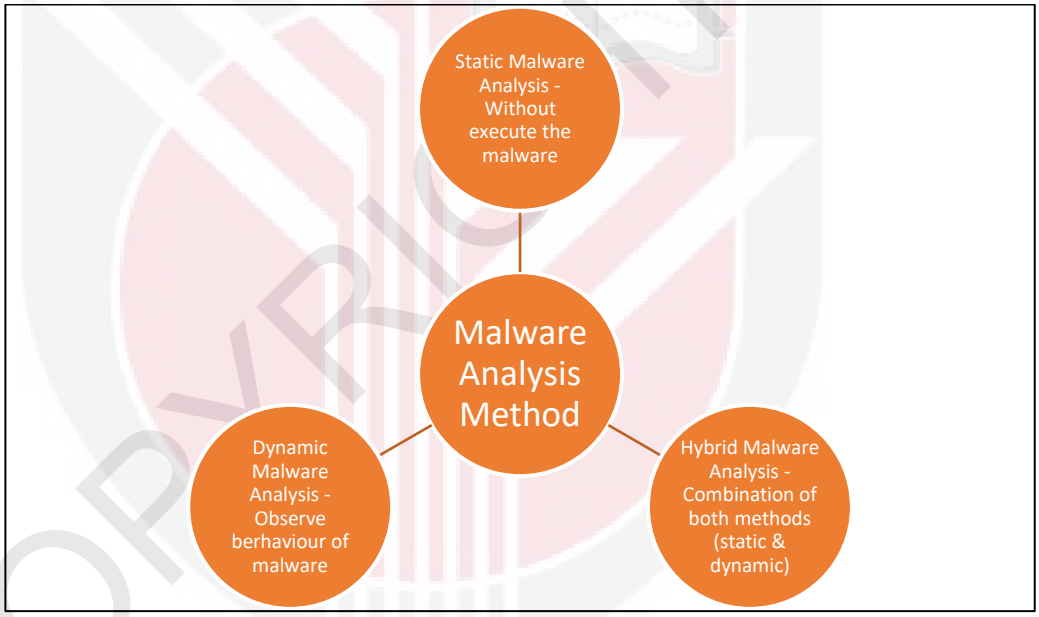

Figure 1: Malware Analysis Method

1.1.3 Hybrid Malware Analysis Method

Hybrid malware analysis method will collect information about malware from static and dynamic analysis. By implement this method, it will reduce the limitation from both method either static or dynamic analysis. Therefore, improving the ability to detect the program intend to properly [5]. It will be observed by the code analysis by checking

the signature of malware and running in the virtual environment to observe actual behavior.

1.1.4 Machine Learning

Machine learning is method that give system or computer to learn without being explicit programmed by discover the algorithm and formalize the principles that underlie the data such as unseen samples. In the area of malware detection analysis, it hidden property that could be malware or benign of mathematically formalized set of principle called as model. It has variety approach that have different capacity and different task they suit best.

Machine learning have two major methods which are supervised and unsupervised learning that have differentiate between the labelling and the aim is to classify the files under the binary (malware or not malware). Machine learning will define a problem by collecting the data and processing it to the algorithm. After collecting malware samples available, it will extract features of information for built dataset. Based on that, machine learning will see the discrepancies between the dataset [6]. In this project static and dynamic feature are integrated each other. The integrated feature is used for training and classification of data. Vector machines that include the decision tree and random forest is some of technique for malware classification. Gradient boosting machine learning will cover for this area of project. The similarities between static and dynamic sample malware samples will be attributed to Cuckoo Sandbox.

1.2 Problem Statement

Most of anti-virus software that be a tool to prevent malware uses a signature-based detection which inefficient if involves the rapid increase in the number and variants of malware. Signature has a unique identification for a binary file. It will have created using static analysis methods. Dynamic use the behavior-based approach to identify whether it is malware or not. Both methods have its own advantage and disadvantages [7]. To make it more reliable, the web-based malware detection must to be develop because it has availability of higher speeds and bandwidth.

The previous web application to detect malware attack always comes with either static or dynamic method only and not comes with the summary report. It will be difficult to see again the history tracking for each training data that already tested using any vector. Besides, it had difficulty to produce web application with time and accuracy to malware detection which lead to results.

1.3 Research Objective

The objective of this project is generally to propose web application scanning that able to detect the malware attack using integrated static and dynamic methods known as hybrid technique by analyze and classify an unknown executable file that also included the implementation of machine learning into the development process. Besides it will provide appropriate incident report for future references such as to know the level of risk of malware attack to the file or system. Other than that, this project proposes analysis

performance based on time and accuracy taken for each method to detect malware detection attack which lead to results.

1.4 Research Scope

As for this project research scope requirement, operating system that involve is Windows XP for purpose to Cuckoo Sandbox. It also works with Windows 7 with disable the function of User Access Control for analysis purpose. Besides, we are used Ubuntu that was installed on VMware Workstation 15 act as a server for virtualization product that makes it possible to a single physical server into multiple virtual machines. Virtual machine is used as the secure environment. Cuckoo is used to run and analysis the files of malware files and generate result based on the behavior of malware while in execution. The log file contains API calls made during execution, registry modifications and the information such as heap memory address and process address.

Besides, the dataset for testing malware detection will be collected from virustotal.com. It is free and can be used to hunt down malware samples based on static, dynamic and relational properties. The search parameters can be combined in order to identify either the files match highly to the criteria by clustering and filtering noise and focusing on threats that are relevant to do an investigation [8].

1.5 Research Schedule

This duration for this project is about10 months which start in September 2018 and expected to finish in June 2019. Generally, this project has six activities which are project implementation plan, knowledge gathering, experimentation design, implementation and development, testing and evaluation and lastly report write up. Each

project activities have their own milestones that need to be achieved. The details for project activities and time take can be refer in Gantt chart in Appendix section.

1.6 Thesis Structure

The structure of this thesis consists of six chapters including Introduction, Literature Review, Methodology, Project Implementation, Result and Discussion and last chapter is Conclusion.

Chapter 1 briefly describes the introduction of background studies of subjects related to this project. This session contains problem statements, research objectives, research scopes, expectations of research results and thesis structures. The research objective is derived from the problem statement which has been studied through anchor paper and is expected to be achieved at the end of the project goal as stated in research objective. The scope of research and research schedule is to highlight the scope of this project and ensure that the project is on the right track according to the set schedule.

Chapter 2 is a list of literature review for this project. Literature review is a source of research article and journal that being used to give more understanding and gain more idea to implement about related topic. This chapter is important as it is to ensure this project is possible to be done and to avoid any duplication of previous work. By this chapter, table of differentiate can be classify based on all the requirement gather from various research article and journal.

Chapter 3 is the chapter that explained methodology that being used to develop this project. Methodology is process flow for every project as it to ensure the project is properly plan and can be execute smoothly. In this chapter also explained the framework that being used in this project.

Chapter 4 is project implementation and development. In this chapter, the approach used is being explained in detail. Besides, the design and the functionalities of the approach also are highlighted. This chapter provides overall and process flow chart for this project. All the detail according to the implementation of the approach also can be found in this chapter.

Chapter 5 is an explaining and display the result. In this chapter all the result and finding related to this project will be provided. An evaluation of the result and the discussion are explained in this chapter.

Chapter 6 is the chapter of conclusion for this project. It consists the advantage and limitation for this project. Besides, there are also suggested future enhancements that can be highlight for next project. This chapter also concludes the whole project, result and the achievement while doing this project.

REFERENCES

[1] Difference Between Static Malware Analysis and Dynamic Malware Analysis | Difference Between | Static Malware Analysis vs Dynamic Malware Analysis. (2018).

[2] Difference Between Static Malware Analysis and Dynamic Malware Analysis | Difference Between | Static Malware Analysis vs Dynamic Malware Analysis. (2018).

[3] A. Moser, C. Kruegel, and E. Kirda. Limits of static analysis for malware detection, in Computer Security Applications Conference, 2007. ACSAC 2007, pp. 421-430, Dec 2007.

[4] I. You and K. Yim. Malware obfuscation techniques: A brief survey. In International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA), pp. 297-300, Nov 2010.

[5] Difference Between Static Malware Analysis and Dynamic Malware Analysis | Difference Between | Static Malware Analysis vs Dynamic Malware Analysis. (2018).

[6] P.V.Shijo, A.Salim,et al. (2015) Integrated static and dynamic analysis for malware detection, pp. 804-811. https://doi.org/10.1016/j.procs.2015.02.149

[7] Christodorescu, M., Jha, S., Maughan, D., Song, D., Wang, C.: Malware Detection. In: Advances in Information Security. Verlag New York, Inc. Secaucus, NJ, USA: Springer (2007).

[8] Weka 3: Data Mining open source Software. Accessed 2014. www.cs.waikato.ac.nz/ml/weka/.

[9] B.Tewfik,B. Zakaria A. Nemrat, B. Chafika,et al (2016) Scalable malware classification based on integrated static and dynamic feature pp. 113-124. DOI: 10.1007/978-3-319-51064-4_10

[10] Y. Zhenlong, L. Yongqiang, A.Izzat,Z.Mohammad, et al (2016) Droiddetector: Android malware characterization and detection using deep learning. pp. 114-123. https://doi:10.1109/TST.2016.7399288

[11] N. Awang, D. Yusof, S. Ariffin, et al (2015) Manual Malware Analysis Using Static Method pp. 324-328.

[12] Mohammad, A. Mamoun, A.Izzat,Z.Mohammad,et al. (2016) The Malware Detection Challenge of Accuracy **DOI:** 10.1109/OSSCOM.2016.7863676

[13] V.Deepti, S.P.Choudary., et al, (2015) A Simple Method for Detection of Metamorphic Malware using Dynamic Analysis and Text Mining pp.265-270. DOI: 10.1016/j.procs.2015.06.031

[14] Cao, Ying., et al, (2013) A Malware Behavior Capturing System Implemented at Virtual Machine Monitor Layer **DOI:** 10.1109/CIS.2012.126

[15] The Cuckoo sandbox. Accessed 2014. http://www.cuckoosandbox.org/