An Approach for Malware Behavior Identification and Classification

Mohamad Fadli Zolkipli
School of Computer Science
Universiti Sains Malaysia, USM
Penang, Malaysia
e-mail: fadli@ump.edu.my

Aman Jantan
School of Computer Science
Universiti Sains Malaysia, USM
Penang, Malaysia
e-mail: aman@cs.usm.my

***Abstract*— Malware is one of the major security threats that can break computer operation. However, commercial anti-virus or anti-spyware that used signature-based matching to detects malware cannot solve that kind of threats. Nowadays malware writers try to avoid detection by using several techniques such as polymorphic, metamorphic and also hiding technique. In order to overcome that issue, we proposed a new framework for malware behavior identification and classification that apply dynamic approach. This framework consists of two major processes such as behavior identification and malware classification. These two major processes will integrate together as interrelated process in our proposed framework. Result from this study is a new framework that able to identify and classify malware based on it behaviors.**

Keywords-computer security; malware; behavior analysis; malware classification

I. INTRODUCTION

Nowadays numbers of computer security threat that cause by malware attack have extremely increased. It is consist of seventeen categories such as viruses, worms, Trojan horse, spyware and also other malicious software that causes a billion of losses to the computer operation worldwide. Although all types of malware have their specific objective, the main purpose is to break the computer operation. G Data Security Labs[1] was shown the higher growth of the malware attack in the graph in Figure 1.

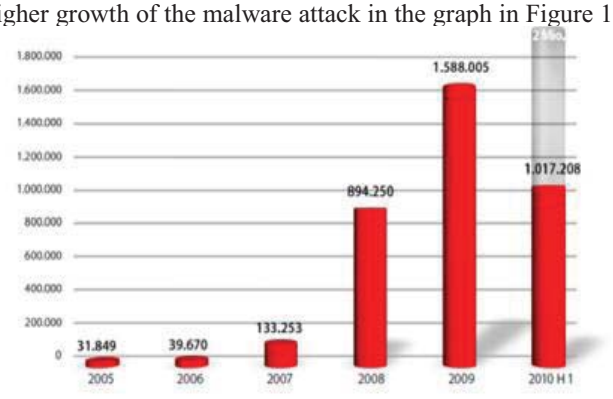

Figure 1. The number of malware growth since 2005 until 2010.[1]

Signature-based matching technique was commercially applied by intrusion detection system, anti-virus or anti-spyware product in the market. Although this technique is very popular and reliable for host-based security tool, there are some limitations on this technique need to be solved. Basically, signature-based technique cannot provide the opportunity to learn and understand the threat because the detection process is only based on the string matching without knowing the goals and behaviors of the malware. Understanding malware goals and behaviors is the best practice because that information is very useful in implementing prevention mechanism on the computer systems as it is different based on times, organizations types and methods of attacks.

However, because of technology advancement many malware writers try to use better hiding techniques to avoid detection [2]. The hiding technique is created by combining the previous behavior in order to attack and at the same time to avoid signature-based detection. Other common techniques that commonly used are polymorphic and metamorphic. Further research need to be done in order to find possible solution on avoiding current technique of malware attacks.

In this paper, we proposed a new framework for malware behavior identification and classification. The proposed framework has three main objectives. First objective is to develop architecture for secure environment for behavior-based malware analysis. Secondly is to implement comprehensive approach to conduct malware analysis. The third objective is to classified malware into new possible group using adapted AI technique. This is truly host-based framework that was designed for the window environment that has high potential of malware attacks.

The rest of the paper is structured as follows. Section II provides state of the art of the study that discusses background and related work. Section III explains the proposed framework which is used to conduct malware behavior analysis and classification. The method of evaluation is described in section IV and the conclusion is provided in section V.

II. STATE OF THE ART

Malware is a program with malicious intent designed to damage the computer on which it executes or the network over which it communicates [3]. Although all types of

978-1-61284-840-2/11/$26.00 ©2011 IEEE

malware have their specific objective, the main purpose is to break the computer operation. Because of that, the security control needs to be implemented [4] in order to protect all code and data against modification, replacement or sub-versioned.

Signature-based matching technique is one of the most popular approaches to malware detection [3]. This technique was commercially applied by anti-virus or anti-spyware product in the market. The main limitation of signature-based technique is on detecting new malware. This technique that used unique bytes string always fails to detect previously unseen malware[5] that normally known as zero-day attacks [6]. The problem will happen when [6] computers must be infected before a new virus pattern can be captured, creating signature and stored for future use. [7] concluded that the signature-based technology suffers from two main shortcomings which is an inability to detect zero-day attacks and an inability to cope with an exponential growth in new malware.

Wagener et al. [8] proposed a flexible and automated approach to extract malware behavior by observing all the system function calls performed in a virtualized execution environment. Similarities and distances between malware behaviors are computed which allows classifying malware behaviors. The main features of this approach reside in coupling a sequence alignment method to compute similarities and leverage the Hellinger distance to compute associated distances. The classification process proposed by this work is using phylogenetic tree. However, this technique still has a limitation due to the wrongly classified a few malware behavior.

Zhou and Inge applied machine learning algorithm on their malware detection technique [9]. That technique was using adaptive data compression in order to counter the limitation of signature-based technique in current commercial anti-malware tool. Zhou and Inge identified two limitations of signature-based technique. First, not all malicious programs have bit patterns that are evidence of their malicious nature and also not recorded in the virus dictionary. Second, obfuscated malware that take many forms of bit patterns will not working on signature-based technique. The proposed technique used adaptive data compression model and prediction by partial matching as learning engine to build two compression models. This technique works on unstructured input, that is, raw executables, with an underlying statistical compression model.

Bergeron et al. proposed a new approach for the static detection of malware code in executable programs [10]. This approach carried out directly on binary code using semantic analysis based on behavior of unknown malware. The reason for targeting binary executables is that the source code of those programs that need to detect malicious code is often not available. The primary objective of the research is to elaborate practical methods and tools with theoretical foundations for the static detection. The experiment was done in three steps such as generating an intermediate representation by using IDA32 Pro, analyzing the control and data flows and doing static verification by using their own prototype tool.

Ulrich et al. presented an approach to improve the efficiency of dynamic malware analysis systems [11]. It is to overcome the huge number of new malicious files currently appears. It is due to mutations of only a few malware programs. The proposed system avoids analyzing malware binaries that simply constitute mutated instances of already analyzed polymorphic malware. It can drastically reduce the amount of time required for analyzing a set of malware programs. The limitation of this approach is due to the changes of the behavior after the analysis process that cause by the limitation of dynamic analysis.

Bayer et al. was used a scalable clustering approach to identify and group malware samples that exhibit similar behavior [10]. This approach also performs dynamic analysis to obtain the execution traces of malware programs using automated tools. The execution traces are generalized into behavioral profiles, which characterize the activity of a program in more abstract terms. Then the profiles serve as input to an efficient clustering algorithm that allows handling sample sets larger than previous approaches in term of malware behaviors. Bayer et al. also stated that it is not sufficient while automating the analysis of a single malware sample in a first step because the analyst will facing a thousands of reports every day that need to be examined.

Tabish et al. was proposed malware detection that applied data mining which is based on the analysis of bytelevel file content [12]. This technique also designed to provide protection against first day launched malware. This non-signature based technique has the potential to detect previously unknown and new launch malware. It does not memorize specific byte-sequences or string that appearing in the actual file content. Standard data mining algorithm was used to classify the file content of every block as normal or potentially malicious by categorize it as benign or malware. The proposed technique was tasted using six different file types such as doc, exe, jpg, mp3, .pdf and zip. Six different types of malware that consist of backdoor, Trojan, virus, worm, constructor and miscellaneous was used as dataset.

III. PROPOSED FRAMEWORK

A. Preliminary Study

This research was proposed to overcome the limitation of signature-based technique in learning and understanding the malware threats. It is because the detection process is only based on the string matching without knowing the goals and behaviors of the malware. Another issues that must be faces in this research is a technique that used by malware writers in creating new malware. Nowadays, new techniques that normally used such as polymorphic and metamorphic were increased the number of malware dramatically.

This research also will overcome the limitation of static approaches on malware behavior analysis. Firstly, it describe about an approach to identifying malware behavior by using hybrid technique on analysing malware behavior on secure environment. Next we will describe about optimizing the malware classification using artificial intelligent technique. Knowledge based will be used in order to make malware classification.

The data used for this research consists of current malware samples that spread through computer network. In order to get as much as possible malware samples, this task can be done using a number of security tools such as malware collector, honeyport and intrusion detection system. Two malware collector tools were selected in this research such as HoneyClients and Amun. Both of tools were selected in order to avoid failure and to maximize the varieties of malware collection. Those tools were continuously running on network within two or three months.

B. *Secure Environment*

The behaviors analysis provides detailed information about malware that suitable for learning and understanding malware samples. The samples are run in a Windows virtual machine environment and their behavior is identified during program execution. Virtual machine is the suitable solution that can be used to create secure environment in order to analyze malware behavior. It is to avoid damage to the real operating system and computer components if the malware executed.

In order to do an analysis, behavior from a piece of malware should be extract in a controlled environment. We decided to improve secure environment architecture that use virtual machine tools to simulate the Windows environment in our proposed framework. Virtual machine operating system is the suitable solution on doing malware behavior analysis because malware often pose strong threat to the computer system. It provides a tightly controlled set of resources because untrusted process cannot run out of the virtual machine [13, 14]. However, there are some malware samples that try to prevent against malware analysis that used virtual environment tool.

C. *Proposed Work*

Figure 2 shows the proposed framework of this research. It consists of two major processes such as behavior identification and malware classification. It also involved the secure environment and database platform to fulfil the need of this framework.

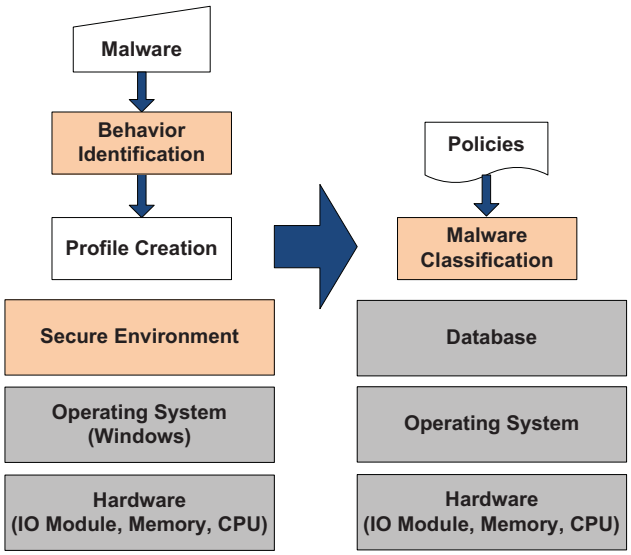

Figure 2. Framework for Malware Behavior Identification and Classification.

1) Behavior Identification: Malware behaviors analysis is a process to understand the types and characteristics of malicious software. The behavior describes the purpose and the function of the malware attacks. Understanding malware behaviors is very important and critical tasks in order to use it for defination and classification of the malware. This task will be done by using dynamic analysis. Two main processes in this task are run time analysis and resource monitoring. New behavior definion process will be added to filter and generate possible malware behavior.

2) Malware Classification: The defined malware behavior from each sample will be used in this process. Normally some of malware sample will share the same characteristic and behavior because it was produced from polymorphic and metamorphic technique. Related classification techniques will be applied for identifying the shared behavior of each malware family. However, the classification process will be optimized by using any Artificial Intelligent technique. Currently more than seven thin types of malware were exist. The other way of classifying malware well be proposed in order to limit the number of malware group.

IV. EVALUATION

First objective of this research is to develop architecture for secure environment for behavior-based malware analysis. Quantitative measurement will be applied in this stage. It is used to present the successful rate in conducting malware behavior using proposed secure environment architecture. The variables such as number of malware and percentage of each malware groups will be used.

Second objective is to improve an approach for behavior analysis that can produce possible malware behaviors. It also the main contribution for this research where a

comprehensive approach on conducting malware analysis will be presented. In order to evaluate the capability of the proposed approach to analyze malware behavior, quantitative measurements are applied. The possible malware behavior that generated from behavior analysis is identified as the quantitative variable. That variable will be compared to the previous static and dynamic approaches.

Third objective of this research is to optimize malware classification using artificial intelligent technique. It will contribute to the new classification technique that applied artificial intelligent components. Evaluation measure for this research contribution will applied qualitative measurements. The new malware class is identified as qualitative variable. The new malware class will be compared to the previous malware group that classified based on types.

On the other hand, in order to evaluate the ability of the new proposed framework, both quantitative and qualitative measurements are applied. It will refer back to the ability of behavior identification and malware classification that measured before. It is because all two processes are the main component in this framework. However, comparison with another framework will be done by comparing the component and purpose of the framework.

V. CONCLUSION

In this paper, we have identified problem and reviewed the existing malware detection techniques. From the analysis, we have proposed a new approach for malware behavior identification and classification. Output from this framework can be used by system administrator to plan and implement prevention mechanism in order to minimize future malware threat. This research will be continuing by implementing the proposed approach that can integrate all two major processes of this framework.

ACKNOWLEDGMENT

The authors would like to thank the members in the Computer Security and Forensic Lab and also Security Research Group for their helpful discussions and suggestions. This work was supported by Short-term Grant No.304/PKOMP/639021 and No.1001/PKOMP/822126, School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia.

REFERENCES

[1] B. Ralf and B. Sabrina, *Malware Report: Half-year report January-June 2010.* G Data Security Labs, 2010(http://www.gdatasoftware.co.uk).

[2] Christodorescu, M., et al. *Semantics-aware malware detection*. in *Security and Privacy, 2005 IEEE Symposium on*. 2005.

[3] M. D. Preda, et al., *A Semantics-Based Approach to Malware Detection.* ACM Transactions on Programming Languages and Systems, 2008. **30**(No. 5, Article 25).

[4] Langweg, H. *Framework for malware resistance metrics*. in *PROCEEDINGS - QoPŠ06*. 2006: ACM.

[5] Yanfang, Y., J. Qingshan, and Z. Weiwei. *Associative classification and post-processing techniques used for malware detection*. in *Anti-counterfeiting, Security and Identification, 2008. ASID 2008. 2nd International Conference on*. 2008.

[6] Zhou, R., et al. *Application of CLIPS Expert System to Malware Detection System*. in *Computational Intelligence and Security, 2008. CIS '08. International Conference on*. 2008.

[7] Syed Bilal, M., T. Ajay Kumar, and F. Muddassar, *IMAD: in-execution malware analysis and detection*, in *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*. 2009, ACM: Montreal, Qu\&\#233;bec, Canada.

[8] Gérard Wagener, R.S.a.A.D., *Malware behaviour analysis.* Journal in Computer Virology, 2008. Volume 4: p. 279-287.

[9] Yan, Z. and W.M. Inge, *Malware detection using adaptive data compression*, in *Proceedings of the 1st ACM workshop on Workshop on AISec*. 2008, ACM: Alexandria, Virginia, USA.

[10] J. Bergeron, M.D., J. Desharnais, M. M. Erhioui, Y. Lavoie and N. Tawbi, *Static detection of malicious code in executable programs.* Int. J. of Req. Eng, 2001.

[11] Ulrich, B., K. Engin, and K. Christopher, Improving the efficiency of dynamic malware analysis, in Proceedings of the 2010 ACM Symposium on Applied Computing. 2010, ACM: Sierre, Switzerland.

[12] Tabish, M., Z. Shafiq, and M. Farooq, Malware detection using statistical analysis of byte-level file content, in CSI-KDDŠ09. 2009, ACM. p. 23–31.

[13] Hengli, Z., et al. *Malicious Executables Classification Based on Behavioral Factor Analysis*. in *e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E '10. International Conference on*. 2010.

[14] Vasudevan, A. *MalTRAK: Tracking and Eliminating Unknown Malware*. in *Computer Security Applications Conference, 2008. ACSAC 2008. Annual*. 2008