BEHAVIORAL MALWARE DETECTION BY DATA MINING

By

Allan Ninyesiga

SEP15/COMP/0644U

Option-Computer Security

Supervisor

Dr.John Ngubiri

.....

UTAMU

A Proposal submitted to the Gradute School for a dissertation in partial fulfilment of the requirement for the award of MSc Computing of Uganda Technology and Management University.

August, 2016

1 INTRODUCTION

1.1 Background

In a modern age, the use of computers, smart phones and internet has become popular in our daily lives. Most businesses, education are online that is using computers and internet. This has provided a convenient environment for people and organizations as well. Due to a huge use of Internet, integrity of our systems and information is more important. This calls for maintaining the consistency, accuracy, and trustworthiness of data over its entire life cycle. Data should not be accessed in transit, and steps must be taken to ensure that data cannot be altered by unauthorized people. Computers bring together people. In the world out of at least 7 billion people a quarter use computers and internet (Internet World Stats, 2016). The development of software has also increased. And for these software to work, they have to be installed on the computers. The combination of software, platforms, and hardware is what keeps people together. They are developed for good that is to perform a certain functionality. However, some software are developed to harm or to damage (a virus, a worm). Which are commonly known as malicious programs or malware. A single code can cause havoc billions people using computers. For example the I LOVE YOU worm developed by Philippines in May, 2000. It managed to wreak havoc on computer systems all over the world, causing damages totaling in at an estimate of $10 billion (Robert,K, 2016). Systems are mostly developed for functionality not security. Many attacks are quiet and not detectable by user. Mechanisms are required to detect these attacks. Lack of these mechanisms makes ICT a high risk of people's privacy and valuable data.

1.2 Problem Statement

Malware has become a serious problem in the field of computing. Although detection techniques like signature-based have tried hard to detect them, they have failed to detect new and unknown malware, Behavior based methods have tried to detect the unknown malware but their results yield to a lot of false positives.

1.3 Justification

The reasons to why this problem is to be solved include;

1. A lot of traditional techniques to detect malware exist but most of them are signature-based which fail to detect new and unknown malware.
2. Many of the research conducted on malware detection in general but a little research exists in behavior of malware detection, so this research participates in enhancing and enriching this research area.
3. Most device operating systems in the world are dominated by Microsoft's Windows operating system; this gives an indicator those windows platforms, and its Portable Executable (PE) files will be the target of attacks by attackers

1.4 Objectives

1.4.1 General Objectives

The main objective of this research is to improve general security of software using data mining techniques based on anomaly detection that can detect known, new and unseen malware from windows Portable Executable (PE) file.

1.4.2 Specific objectives

1. To collect dataset to perform experiments.
2. Finding and selecting the features and feature type that appropriate for the mining process.
3. To develop a classifier using data mining classifications that help to differentiate the malicious from benign programs.
4. To detect the behavior of both known and known malware based on API call analysis using data mining techniques and reduce false positives.

1.5 Expected Outcome

We intend to come up with a malware detection technique to detect the behavior of known, new and unknown malware and reduce the false positive detection ratio.

2 LITERATURE REVIEW

2.1 Malware

2.1.1 Malware Definition

In simple terms, malware is software i.e. a computer program used to perform malicious actions (perform harm than good). According to Imtithal and Ali, (2013), the term malware is a combination of words malicious and software and can be used to indicate any unwanted software. It is a malicious software which is used with the intention of breaching a computer systems security policy with respect to confidentiality, integrity and availability (Landage and Wankhade, 2013). Attackers or cyber criminals usually install malware on computers or devices to gain control over them or gain access of what they contain. Once malware installed, these attackers can use them to spy on user's activities online, steal user's passwords and files or use attacked system to attack others.

2.1.2 Malware Categories

Malware can be divided into different categories, which include; virus, worm, spyware, Trojan, adware, logic bomb among others.

Virus: This is a type of malware that propagates by inserting a copy of itself into and becoming part of another program. It spreads from one computer to another while leaving infections as it travels. Viruses usually run with user involvement, and cannot do anything alone it must be executed by the carrier program which has been infected (Zahra et al., 2013).

Worm: Worms are similar to viruses in that they replicate functional copies of themselves and can cause the type of damage. Worms are independent meaning they don't need for a host program to start lifecycle (Imtithal and Ali, 2013)

Spyware: is secretly installed on a user computer for the purpose of collecting information about users without their knowledge (Zahra et al., 2013, Imtithal and Ali, 2013). It is a software which monitors and gather user personal information and sends that information back to the attacker to use the stolen information in a notorious way.

Trojan: Sometimes called a Trojan horse. It is a malware that appears legitimate and useful, but in fact can compromise computer or device security and cause much damage. It actually steals information or corrupts data (Imtithal and Ali, 2013).

Adware: Adware is an advertising software package that automatically play advertisements to the user without need. Adware objective is to gain financial profit for their author (Imtithal and Ali, 2013). They are not harmful but they interrupt users thinking by always put themselves in form of a pop-up window.

Logic bomb: Logic bomb is a piece of code intentionally inserted into a software system with the aim of setting off a malicious function when specified conditions are met.

2.1.3 Malware characterization

According to Imtithal and Ali, (2013), Malware can be characterized by the ability of propagation, replication, corruption of computer and self-replication. Zahra et al., (2013), show that malware can be characterized by their concealment strategies. Using these strategies, malware developers try to make malware to evade anti-malware strategies. These include;

Obfuscation: Malware developers use actions such as adding garbage commands, unnecessary jumps etc. to prevent signature based detection methods from detecting their malware.

Polymorphic strategy: in this strategy, malware usually encrypts its self by an encryption algorithm. And a different decryption key is used in any infection. Unlimited number of encryption algorithms can also be used to avoid detection. Syntaxes of mal-code mutate in each time of infection without change in semantics (Imtithal and Ali, 2013).

Metamorphic strategy: In this case malicious software change themselves in such a way that the new instance has no any resemblance to the original ones. This aspect makes the too complex type of malware.

Remote execution malware: (Imtithal and Ali, 2013), hackers normally by using the infrastructure of internet achieve their intention remotely by using malicious software to perform remote malicious functions.

2.2 Malware detection techniques

To fight against malware, Organizations have developed techniques which help in weakening the malware from attacking systems. The most commonly used techniques include signature based and behavior based detection techniques. We describe these techniques and how perform in their malware detection role.

2.2.1 Signature Based Detection

This is the most common method in malware detection. Signature based methods use the patterns extracted from various malwares to identify them and are more efficient and faster than any other methods (Zahra et al., 2013).The signatures are often extracted with special sensitivity for being unique, this makes those detection methods that use this signature have small error rate. Using signature-based techniques in identifying maliciousness of a file, scanner software is used to evaluate its information to a vocabulary of virus signatures in a database to know whether a signature exists (Imtithal et al., 2013). The technique maintains the database of signature and detects malware by comparing pattern against the database (Landage and Wankhade, 2013, Osaghe, 2015).

2.2.2 Behavior Based Detection

These malware detection techniques observe behavior of a program to conclude whether it is malicious or not. A behavior based detector is used to conclude whether a program is malicious by inspecting what it does rather than what it says (Zahra et al., 2013). Behavior-based techniques are not susceptible to the shortcomings of signature-based techniques since they observe what an executable file does (Zahra et al., 2013). They detect computer malicious software by monitoring system activities and classifying it as either normal or anomalous (Imtithal et al., 2013). Programs with the same behavior are collected and using a single behavior signature various samples of malware can be identified.

2.3 Strengths and weaknesses

2.3.1 Signature based detection

Strengths

This technique can detect the known instances of malware accurately with less amount of resources required to detect the malware, and it mainly focus on signature of attack (Landage and Wankhade, 2013, Osaghe, 2015).

Small error rate: This small error rate is the main reason that most common antiviruses use this technique (Zahra et al., 2013).

Weaknesses

Signature based detection techniques are unable to detect unknown malware variants (Zahra et al., 2013, Imtithal et al., 2013, Landage and Wankhade, 2013, Ryo, Daiki &Shigeki, 2013, Zarni, & Win, 2013).

They also require high amount of manpower, time, and money to extract unique signatures (Zahra et al., 2013).

Another weakness is the inability to comfort the malwares that mutate their codes in each infection such as polymorphic and metamorphic malware (Zahra et al., 2013).

Signature based techniques are also susceptible to evasion (Mujumdar, Masiwal, & Meshram, 2013).

2.3.2 Behavior based detection

Strengths

Behavior based malware detection techniques has the ability to detect the type of malwares that signature base techniques are unable to detect such as unknown and polymorphic malware variants (Zahra et al., 2013, Landage and Wankhade, 2013, Zarni, & Win, 2013)

Weaknesses

The drawback is they are susceptible to high false positives (Zarni, & Win, 2013, Ashwini, Gayatri, & Meshram, 2013) and high amount of scanning time (Zahra et al., 2013, Landage and Wankhade, 2013).

The technique needs to update the data describing the system behavior and the statistics in normal profile but it tends to be large which requires more resources like CPU time, memory and disk space (Landage and Wankhade, 2013).

2.4 Data mining

Data mining which is sometimes called knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. According to Mehedy, Latifur, & Bhavani,(2012), Data mining is the process of posing various queries and extracting useful and often previously unknown and unexpected information, patterns, and trends from large quantities of data. The goals of data mining may include detecting abnormal patterns, predicting the future based on the past etc. Data mining can be divided into different tasks (Muazzam, 2008, Fadel, 2013).

Predictive data mining, which consists of predicting unknown values based upon given items.

Descriptive data mining, which consists of patterns describing the data.

Classification: this deals with discovery of a predictive learning function that classifies a data item into one of more predefined classes (Fadel, 2013).

Regression: this involves the discovery of a predictive learning function that maps a data item to a real value prediction variable. Clustering: involves identifying a finite set of clusters or categories to describe the data.

Summarization: this is a descriptive task which involves methods for finding a solid description for a class or subclass of data (Fadel, 2013).

2.5 Data mining for malware detection

Many authors have researched about malware detection using different techniques such as signature-based detection but they have not been able to detect new malware variants. Others have tried to use behavior based detection methods but they have a problem of false positives and low accuracy. Recent researchers have shown using data mining approaches the behavior of malware can easily detected basing on features such as windows API calls which is the interest of our research. In this section we provide related work about Data mining techniques and API call feature techniques.

Hamid, Mehdi, &Ahmad, (2014) present a data mining approach to predict executable behavior using API that provides sequences captured of a running process. The approach is divided into four steps which include; collecting executable files and executing them, Registering interactions between execution files and the system, Extraction of API functions, selecting them as characteristics and choosing the number of iterations as the value of characteristics, and finally using different classifications for detection. Experimental results show the method is effective on detecting polymorphic and metamorphic malware with the accuracy of 93.5% with a detection rate of 95%. That accuracy is not perfect.

Abhay Pratap Singh, (2013) presents a data mining approach which improves the malware detection ratio with a high precision. The proposed approach consists of five steps which include Data collection, data processing, Apply Ida Pro, generating .ASM file, converting .csv file, them applying csv file to Weka tool for classification. Results show that the proposed methodology can detect the malware along with obsification when compared with existing antivirus scanner.

Chun-I Fan, et al.., (2015) used hooking methods in order to trace dynamic signatures that malware tries to hide. Then use data mining techniques to compare the behavioral differences between malware and benign for malware identification. The approach use API call as dynamic features for analysis. A tracing module is developed in virtual environment to trace the behaviors that malware attempts to conceal. Then data mining tools are adopted to analyze and build a description model used to identify malware and benign programs.The results show the method can achieve a high detection rate with low

complexity by having a detection rate of 95% with only 80 attributes.

Haixu Xi and Hongjin Zhu, (2016) design a kind of data mining technology in order to solve the various problems of anti-virus software. The technology is designed on the basis of the new features of malicious programs intelligent detection rules, and this method use the Windows platform executable file format as the main characteristics, extraction of executable file, then analyzes and gets the new characteristics of malicious programs. To extract the detection rules, data mining technology is used in order to out the hidden malicious program rules, and improves the accuracy.

Guanghui, Jianmin, & Chao, (2016) propose a classification technique using dynamic analysis based on behavior profile. API calls and other essential information of running malware are captured when the malware is running, then their multilayer dependency chain is established according to dependency relationship of these function calls. To identify the degree of similarity between malware variants, the similarity comparison algorithm is used.

Hyun-il Lim, (2016) propose an approach to detecting malicious behaviors of software by analyzing dynamic API function calls. API functions in Microsoft Windows operating systems are classified and an approach to representing malicious behavior of software is proposed. Malicious behaviors are abstracted as a set of k-grams and can be identified by calculating similarity between the sets of k-grams and a sequence of API function calls, thus increase the efficiency and the tolerance of the analysis. Malicious behavior of software is represented as a form of behavior automata that is generated from API flow graph of a program, and then automation is traced with a sequence of API function calls that are extracted during program execution in order to recognize the malicious behavior of software. The behavior automation is abstracted as a set of k-grams to increase the efficiency and the tolerance of analysis. To identify malicious behavior, the similarity between the set of k-grams and a sequence of API function calls is calculated.

Zahra, et al.., (2013) presents the use of features such as API calls, OpCodes, N-Grams etc. that can be used in methods to detect the behavior of malware. The author show that using API calls as a feature has some advantages which are; they help in detecting polymorphic and unknown malware, outperforms other classification approaches in both detection ratio and accuracy, obfuscated malware variants can easily be detected, and

help to detect malware before execution. This show that API calls is a good feature for malware detection.

Youngjoon, Eunjin , & Huy, (2015) proposed an approach for dynamic analysis of malware by adopting DNA sequence alignment algorithms. The algorithms are used to extract common API call sequence patterns of malicious function from different categories of malware. Hooking process monitors are used to track the program's API call sequences when a new program needs to be traced. The system then compares the extracted API call sequences with API call sequence of API-based malware detection system database (APIMDS). If there is a match, APIMDS alerts the administrator.

Mojtaba, Zeinab, &Sattar (2013) show API calls are important features being able to describe the behavior of programs; therefore, they are appropriate features which can be used by a classifier in order to categorize executable files based on their behavior

3 METHODOLOGY

In this section we give an explanation on what we are going to do.

3.1 Data collection

We intend to infect some devices, or get malicious data containing malware programs and run them on devices. Basing on their API calls we record the behavior information of the programs. Then we record the API data or information. We do the same to devices that are not infected by malware and collect other dataset.

3.2 Data processing

We focus on extraction of API call as features to distinguish between benign and malware file. Hackers can use packer techniques to hide content of their malware, and also benign executables may use these techniques to protect applications against cracking. We intend to unpack some of the files (PE file) in the data set that are either compressed or packed. Then each binary executable is disassembled, and API call information extracted. Relevant API calls are then selected. The collected data is divided into two subsequent phases which are training and testing. In the training phase the classifier will learn to classify the file as either benign or malicious, while the testing phase is used for evaluating the performance.

3.3 Analysis

Basing on two datasets, we do analysis by applying data mining techniques. And we apply a classifier basing on the analysis of the two different kinds of datasets to determine whether the program is malicious or benign.

4 Conclusion

Most of the malware detection techniques such as signature-based have tried to help in detecting of malicious activities. Although they try to do so, malware programmers have tried to come up with new and unknown malware programs that are hard to be detected by those techniques. Some behavioral detection techniques try to detect more than signature behavior methods but they also have a false positive problem. This calls for data mining techniques to detect the behavior of malware by detecting new and unknown malware, reduce the false positive detection rate and improve the accuracy.

References

[1] Jyoti, L., and Wankhade, M.P.(2013). Malware and Malware Detection Techniques. A Survey. International Journal of Engineering Research and Technology (IJERT), vol. 2, no. 12, pp. 61-68 .

[2] Imtithal, A. S., and Ali, M. A. A.(2013). A Survey on Malware and Malware Detection Systems. International Journal of Computer Applications,vol. 67, no. 16, pp. 25-31.

[3] Zahra,B., Hashem,H., Seyed,M. H. F., and Ali. H.(2013). A Survey on Heuristic Malware Detection Techniques. 5th Conference on Information and Knowledge Technology (IKT), vol.13, (13), pp. 113 -120.

[4] Osaghae, E.O. (2015). Improved Signature-Based Antivirus System, International Journal of Computer Science and Information Technology Research, Vol. 3, No. 4, pp. 250-254.

[5] Ryo, S., Daiki, C. and Shigeki G.(2013) Detecting Android Malware by Analyzing Manifest Files, Proceedings of the Asia-Pacific Advanced Network, Vol. 36, No. 14, pp. 23-3.

[6] Zarni, A., and Win, Z.(2013) Permission-Based Android Malware Detection, International Journal of Scientific & Technology Research, Vol. 2, No. 3, pp. 228-234.

[7] Ashwini, M., Gayatri , M., and Meshram, B. (2013). Analysis of Signature-Based and Behavior-Based Anti-Malware Approaches, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 2, No. 6, pp. 2037-2039.

[8] Mehedy, M., Latifur, K, and Bhavani, T.(2012). Data mining tools for malware detection, New York, CRC Press.

[9] Muazzam, S.(2008). Data mining methods for malware detection, (Doctoral dissertation), University of Central Florida Libraries.

[10] Fadel, O. S.(2013).Spyware detection using data mining for windows portable executable files, (Master's thesis).

[11] Hamid ,R. R., Mehdi, S., and Ahmad,K.(2014). A novel data mining method for malware detection. Journal of Theoretical and Applied Information Technology, vol. 70, no. 1, pp. 43-51.

[12] Abhay,P.S.(2013). Improving the malware detection ratio using data mining techniques. 2nd International Conference on Science, Technology and Management,pp. 852-857.

[13] Chun-I,F.,Han-Wei. H.,Chun-Han, C., and Yi-Fan, T.(2015). Malware Detection System Based on API Log Data Mining .

[14] Haixu, X., and Hongjin, Z.(2016). Data Mining Methods for New Feature of Malicious Program. International Journal of Hybrid Information Technology, vol. 9, no. 3, pp. 171-178 .

[15] Guanghui ,L., Jianmin, P., and Chao, D.(2016). A Behavior-Based Malware Variant Classification Techniqu. International Journal of Information and Education Technology vol. 6, no. 4, pp. 291-295.

[16] Hyun-il,L.(2016). Detecting Malicious Behaviors of Software through Analysis of API Sequence k-grams. Computer Science and Information Technology,vol. 4, no. 3, pp. 85-91.

[17] Youngjoon,K.,Eunjin, K., and Huy, K. K.(2015). A novel approach to detect malware based on API Call Sequence analysis. International Journal of Distributed Sensor Networks, http://dx.doi.org/10.1155/2015/659101.

[18] Mojtaba, E., Zeinab, K., and Sattar, H.(2013). HDM-Analyser, a hybrid analysis approach based on data mining techniques for malware detection. J Comput Virol Hack Tech, vol. 9,pp. 77-93.

[19] Robert, K.(2016, March 8). 15 Most Dangerous Malware Of All Time. Retrieved from http://zoorepairs.com.au/computer-tips/list-of-most-dangerous-malware-of-all-time/

[20] Internet World Stats.(2016). Retrieved from http://www.internetworldstats.com/stats.htm