Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.930 – 935

REVIEW ARTICLE

Techniques in Detection and Analyzing Malware Executables: A Review

Milan Jain[1], Punam Bajaj[2]

[1,2]Department of CSE, Chandigarh Engineering College, Mohali, Punjab, India
[1] milan.jain2@gmail.com, [2] cecm.cse.punb@gmail.com

Abstract- Today computer field has gained a lot of importance in our day to day life to deal with many aspects like education, entertainment purpose etc. System security is warned by weapons named as malicious software to fulfill malicious intention of its authors. Malicious software known as malware is one of the common problem faced by the internet today. The key to detect these threats are also available like AV Scanners, Intrusion Detection System, and Firewalls etc. In this paper we discussed various data mining techniques, several anti-virus systems are there for detecting the malware i.e. malicious code written manually but these approaches are very expensive and oftentimes ineffective. Therefore, there is a requirement to present a data-mining framework that can detect new, malicious executables precisely and systematically. This survey paper highlights the techniques used in analyzing and detecting malware data.

Keywords- Malicious Code Detection; Data Mining; Security; Malware; prediction

I. INTRODUCTION

As the application of computer and Internet is gaining popularity, as it is the easiest method to find the malicious code gives chances to malicious code occurrence or malicious programs, including computer viruses [1]. These programs include: spyware or malware, Trojans, viruses etc.. Spyware is not visible to the user and can be hard to find. Spyware like keyloggers, is installed by the user of a corporate sector, or public system in order to check out the users who are using. Spyware term means software that monitors a user's computation, the functions of spyware can enhance simple monitoring [9]. Malicious code can be any data, like user's personal information and Internet surfing activities, logins of user, and bank account information. Some spyware can modify overall computer settings or options, resulting in very slow Internet connection paces, unauthenticated modifications in browser settings options. The goal of this paper is to explore the limits of static, dynamic and hybrid analysis for the detection of malicious code.

Malware is a malicious data that replicates over the systems which are in connection with network. Malicious code (or malware) is defined as software that fulfils the harmful intention of an attacker. This scenario is increasing at a large scale due to advanced computing technology and communicating entities of network. Malware is the entity in which current features are easily added to enhance its dark side effects in the form of different types of attacks. These malwares are very dangerous and harmful along with their all side effects on the infected machines like disabling malware detectors or AV Scanners which are installed for the security reasons. According to statistical data, 70-80% of the spyware comes from known sites [11]. Malicious code activities and security solutions is not yet leveling. In fact, attacks have also been increased against Web browsers and malicious code variants [8].

II. **DATA MINING**

Data mining task is the automatic or semi-automatic analysis of large quantities of data from unknown interesting patterns. It has database methodologies like spatial indices. These patterns have detailed type of the input data, and useful for further analysis or, for instance, in machine learning and predictive analysis. For example, the data mining step might finds out the number of groups in the particular data, further used for obtaining very precise predicted consequences. Data collection, data preparation and result interpretation and reporting are not subset of the data mining step, related to overall KDD process as additional steps.

As Information Technology is growing tremendously, the fast growth of data raised the capability of the processing data manually. This is more essential to guide people for extracting the general knowledge from the bulk of data. For implementing this, data mining technique is discovered recently and became research direction for scientists and engineers. It is able to find out the unusual connecting behavior of network in real-time to discover the trace of type of worm virus, specifically the precaution is to the new worm virus to make administrator to follow corresponding measure to avoid tremendous loss of data.

III. MALICIOUS EXECUTABLES

Malware is any software that are being used for disturbing normal system functions, collect important information, or can access to private computer systems also called malicious software and can be in the form of coding, scripts, active material etc. 'Malware' term is used for a variety of forms of hostile or intrusive software [7]. Malwares are dangerous as they have too many limitations on the disordered machines like disabling malware detectors or AV scanners which installed for security purposes.

As we know malware is categorized into three generations that includes payload, enabling vulnerability & propagation mechanism. In the First Generation malwares share properties of virus which are replicated or propagated by some human activities, such as emails and file sharing whereas in the Second Generation malware carries properties of worms which are hybrid in nature involving some features of virus and Trojans which are not replicated by any human actions. Hence in the Third Generation malwares are in the geographical region or are specific in organization. These malwares usually attacks security technologies and products [8].

Malware can be broadly classified into following categories.

A. *Viruses*

Virus is a program with many harmful effects on internet that repeats itself. When we execute the code, these programs are automatically executed. The actual virus modifies itself by changing into new variants. These programs can spread from any one computer to other computer via network or malicious data such as in USB devices. The main target is binary executable file in MSDOS and hard disk, general purpose script.

B. *Worms*

Worms are repeatedly occuring programs and has network to transmit different sets of data to other systems without user authorization and copies are created itself. Worms can affect the overall network by using the bandwidth or frequency bands. There is no need to use any other file for supporting it. It can delete files, encrypt files when crypto viral attack or junk email attacks on it.

C. *Spyware*

Spyware is the word for software of system that can monitor and collects the personal information related to client like the web pages recently visited, emails, ATM number etc. It comes automatically while downloading of any free or trial software.

D. *Adware*

Adware or advertisement of supported software are automatically played, displayed and downloaded to the system after malicious software is downloaded or any application can be used. Adware is piece of code which is embedded into software. The situation is, many programmers does not like these advertisements software but checking out the internet’s user activities.

E. *Trojans*

Trojan horses follows the behavior of the authenticated program like login cell and hijacked user password for gaining complete control of system remotely. Some other activities can include damage system resources like computer files and any data, denies particular services.

F. *Botnet*

A botnet is software that is collection of autonomous software robots which are remotely controlled. It is usually a Zombie program which is controlled by secure network or common control. These are used for sending spam spyware remotely. Botnets need not to wait for the instruction from the third party inspite of that, it finds for the communication with similar sets of bots that are waiting for instructions. There is a structure called hierarchical in which both master is connected to hundreds of bots that are further connected to many bots.

IV. VARIANTS OF MALWARE

In malware detection, we put our major focus on the different variants produced or which are already in existence related to the security threats that pose to the system, the detailed are following as:-

A. *Polymorphic Malware*

If a virus is programmed that look different every times when it is replicated, but also keep the original code intact. The virus of this type is referred as polymorphic virus. Polymorphic code is a common method which is implemented in malware and uses a polymorphic generator to mutate the code thus keeping the original code intact [8]. A typical implementation of a code is to encrypt malware and it also includes the encryptor/decryptor within the code. It have mutation engines which are specially designed. Just by changing the order of instructions, the Mutation Engine generates a new decryption routine. Hence, the body cannot be scanned as it is encrypted, and thus mutation engine has the capacity of generating too many different routines which are decrypted.

B. Metamorphic Malware

If a body of virus changes from one instance to another then it is known as metamorphic virus. It is also known as body-polymorphic malware. The nature of the malware helps malicious executables to mutate due to which they spread across the network and thus make signature based detection completely ineffective. Various Techniques employed in metamorphism are as follows: Disassembly-Depermutation / Shrinking, Expansion Permutation, Assembling and Other transformations.

Two techniques/methods used for malware detection:

1) *Behavior-Based Techniques*

Here the main aim of this technique is to know the behavior of known or benign malwares. The parameters of behavior-based technique include variety of factors such as source/ destination address, types of attachments and other countable statistical features of malwares.

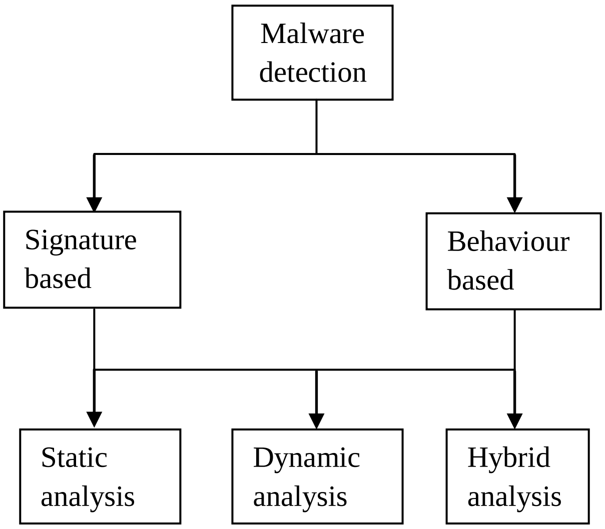

Fig1. Malware detection and analysis

2) *Signature-Based Techniques*

The signature-based technique includes most of the antivirus tools which are used for detection. These signatures are created by examining the code which is disassembled. Various disassemblers and debuggers are known which help in disassembling the portable executables [8]. Thus disassembled code is analyzed and features are extracted. Hence these features play a significant role in constructing the signature of particular malware family.

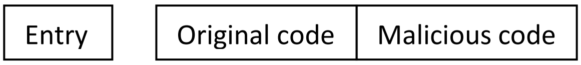

Fig2. Basic kind of virus

V. MALWARE ANALYSIS TECHNIQUE

Various techniques used for signature and behavior are categorized into static, dynamic and hybrid analysis which are as follows.

A. *Static Analysis*

It is the method of detecting executable code without the actual execution of file. In this type of analysis we draw out the low level information from codes and it involves various tools which can be used for working such as program debuggers, analyzers and disassemblers. Static analysis has the advantage of showing the code structure of the program. The main benefit of static over dynamic analyses is that it is free from the overhead of execution time. So they can easily detect multipath malware and are fast and safe. They may fail in analysis of unknown malware.

B. Dynamic Analysis

It is also referred to as behavioral analysis, which involves the malware which is executed and its behavior is monitored, interaction with the system, and the effects on the machine [8]. As we know malware writers make use of machine tools so that they can hide its function resulting in invisible working environment. So here it becomes difficult to detect multipath malware and easy to analyze benign malware. Hence they are neither fast nor safe.

C. Hybrid Analysis

Hybrid analysis is a combine approach of both static and dynamic analysis. It firstly detects the signature of any malware data which are in proper specification then combines with the other behavioral parameters for enhancement so that it results in complete analysis. This approach is a good scope as it overcomes the drawbacks of both static and dynamic analysis.

VI. EXISTING ALGORITHM

In this work, we are using three data mining algorithms to produce new classifiers with separate features: RIPPER, Naive Bayes and a Multi Navie Bayes and the comparison between these three algorithms.

A. RIPPER

The first algorithm used, RIPPER [7], is an inductive rule based learning algorithm which produces a model for detecting malware (malicious code) composed of resource rules. RIPPER uses LibBFD particulars as features. RIPPER is a rule-based learner that is building a set of rules that determines the classes while reducing the rate of ambiguity. The error or ambiguity is evaluated by taking the set of trained instances misclassified by the rules.

B. Naive Bayes

The next classifier is a naive Bayes which calculates the likelihood that a program is having a malicious code that are present in the program. This approach uses both strings and byte sequence data to compute a probability of a binary's malicious code having some features. In Naive Bayes method, each executable's features is used as a text document and classified on that basis. The key supposition in this technique is that the binaries contain same features such as signatures, machine instructions, etc.

C. Multi-Naive Bayes

The next data mining algorithm is Multi-Naive Bayes. This algorithm was importantly a collection of Naïve Bayes algorithms that supported on an overall classification for this instance. Naive Bayes algorithm everytimes do the classification the examples in the test set as malicious or benign and this counted as a single choice. Then these choices are merged by the Multi- Naive Bayes algorithm to output a final categorization for all the Naive Bayes. This method was needed because even using a machine with one gigabytes of RAM , as the size of the binary information was very huge to get into memory. To overcome this problem we categorized them into tiny pieces that would easily get into the memory and therefore trained a Navie Bayes classifier.

VII. LITERATURE SURVEY

In the paper [11] explained that malicious program is one of the main problem faced by the internet. Basically software is composed to disorder the operation, collecting information useful for unauthorized access and other targeted behavior. Malwares from their early patterns that were needed for replication have now progress into more advanced form. Finally we are explaining trends in malware designs and current attacking models and also the current mitigation strategies.

In the paper [8] the computer technology plays a great role in our day to day life to deal with various aspects like education, banking, communication, entertainment etc. Computer system's security is warned by weapons named as malicious programs to fulfill malicious intention of its authors. But these approaches are also avoided due to some obfuscated techniques employed by malware writers. This survey paper highlights the techniques used in analyzing and detection malware executables.

In the paper [1] explained various data mining techniques for security applications. These requisition include but are not limited to malicious executable detection by mining it binary executables, network intrusion detecting by mining network traffic, anomaly detecting, and data stream mining process. They summarize their acquirement and present works at the University of Texas at Dallas on intrusion detection, and cyber-security research.

In the paper [14] addressed that the Early Detection, Alert and Response (eDare) system is aimed at purifying Web traffic propagating via the premises of Network Service Providers (NSP) from malicious code. To achieve this goal, the system employs powerful network traffic scanners capable of cleaning traffic from known malicious code. The remaining traffic is monitored and Machine Learning (ML) algorithms are invoked in an attempt to pinpoint unknown malicious code exhibiting suspicious morphological patterns. Decision trees, Neural Networks and Bayesian Networks are used for static code analysis in order to determine whether a suspicious executable file actually inhabits malicious code.

In the paper [2]The amount of spyware increases rapidly over the internet and it is difficult for the average user to check if a software application hosts spyware. Moreover, algorithms performance is much better than the current state-of-the-art EULA analysis method. Based on it, we show a novel tool that is helpful in preventing the spyware installation.

VIII. CONCLUSION

In this paper, we have presented a comprehensive survey on the static, dynamic and hybrid malware analysis techniques and detection techniques. Today the Internet is becoming popular changing our ways of thinking and standard of living is as useful, as easy to destroy as well. The clear growth in E-commerce, today's Open source nature of malware, the growing penetration of the Internet in respect to insecure connected PCs, are among the main driving factors of the scene. Organizations should adopt a multilayered Web defense strategy that can protect their users and networks from increasingly sophisticated threats and malware is posing a threat to users computer systems in terms of stealing personal and private information, corrupting or disabling our security systems This paper has surveyed on data mining techniques for security applications. Different studies have been done till now that indicates that data mining techniques perform well for finding malicious code. Finally, there is requirement of research in the hope of stimulating further research in this thriving area.

REFERENCES

[1] Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen,"*Data Mining for Security Applications* ",2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing.

[2] Boldt, M. ; Dept. of Syst. & Software Eng., Blekinge Inst. of Technol., Ronneby ; Jacobsson, A. ; Lavesson, N. ; Davidsson, P., "*Automated Spyware Detection Using End User License Agreements"* Information Security and Assurance, 2008. ISA 2008. International Conference on 24-26 April 2008; 978-0-7695-3126-7.

[3] D.Michie, D.J.Spiegelhalter, and C.C.TaylorD. Machine learning of rules and trees. In Machine Learning, Neural and Statistical Classification. Ellis Horwood, 1994.

[4] Guillermo Suarez-Tangle, "*Evolution, Detection and Analysis of Malware for Smart Devices"* IEEE communications surveys & tutorials, accepted for publication, pp.1-27, 2013.

[5] H. Dediu, "When will tablets outsell traditional pcs?"March2012,http://www.asymco.com/2012/03/02/ when-will-the-tablet-market-be-larger-than-the-pc-market/.

[6] Johannes Kinder, "*Detecting Malicious Code by Model Checking*" pure.rhul.ac.uk/portal/files/17566588/mcodedimva05.pdf.

[7] Matthew G. Schultz "*Data Mining Methods for Detection of New Malicious Executables*", academiccommons.columbia.edu/download/.../binaryeval-ieeesp01.pdf.

[8] Kirti Mathur ,Saroj Hiranwal, "*A Survey on Techniques in Detection and Analyzing Malware Executables* "International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 4, April 2013 ISSN: 2277 128X.

[9] Parisa Bahraminikoo "*Utilization Data Mining to Detect Spyware*", IOSR Journal of Computer Engineering (IOSRJCE), Volume 4, Issue 3, pp.01-04, 2012.

[10] Peter Miller. Hexdump. Online publication, 2000 http://www.pcug.org.au/millerp/hexdump.html.

[11] Rizwan rehman, Dr. g.c. hazarika, Gunadeep chetia, "*malware threats and mitigation strategies: A survey"*; Journal of Theoretical and Applied Information Technology; 31st July 2011. Vol. 29 No.2, ISSN: 1992-8645.

[12] Robert Moskovitch "*Detecting unknown malicious code by applying classification techniques on OpCode patterns*" Springer-Verlag "http://link.springer.com/article/10.1186%2F2190-8532-1-1" 2012.\.

[13] Wildlist Organization. Virus descriptions of viruses in the wild. Online publication, 2000. http://www.fsecure.com/virus-info/wild.html.

[14] Yuval Elovici, Asaf Shabtai, Robert Moskovitch, Gil Tahan, and Chanan Glezer" *Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic"*.