Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Survey on Different Kinds of Malware and their Detection

Mansi Goyal
CSE& Kurukshetra University
Haryana, India

Ankita Sharma
CSE& Kurukshetra University
Haryana, India

***Abstract**—As computer technology is becoming necessity in our day to day life in various aspects like education, communication, banking etc., computer system's security becomes the main concern nowadays. Malware is getting its roots strong due to this emerging growth of high speed internet, so detection and removal of these malware in an effective manner is very essential. Malware detectors are the tools to protect against malware. The goal of this research is to study the concept of malware, its kinds and several detection techniques for their recognition. The main objective is to compare the detection techniques to check their effectiveness. The metrics used to compare the performance of these techniques have also been studied.*

***Keywords**— Malware, Detection Techniques, Signatures, Anomaly, Behaviour.*

I. INTRODUCTION

In today's world, sharing of information through email is getting popularity. All the business information and private credentials are shared through email only. Messages are exchanged using SMTP i.e. simple mail transfer protocol. There are two types of agents involved in it: mail delivery agents and mail transfer agents [6]. Cybercriminals exploits the benefits of using email to enter target networks. Its usability in offices, either physical or virtual, had proved to be an effective method to introduce attacks. A severe threat to security today is malicious executable, specifically new, unseen malicious executable frequently received as email attachments. These new malicious executable are produced in a large proportion every year and pose a serious security threat. The message that has been intentionally shaped to cause harm at the server or the client side is known as a malicious email message. This message might include a virus, or it may be due to the message being shaped. Recent research indicated that organizations got at least 20 billion malicious emails each quarter.

A. MALWARE

Malware may be defined as malicious software that enters our system deprived of the permission of user of the system. Malware poses a great threat in computing environment these days. It continues to grow in capacity and evolve in complexity. After entering in the system, the malicious software examines for vulnerabilities of operating system and exploits them to accomplish unintended activities on the system which will slow down the effectiveness of the system. [2] Malicious code is defined as "the code which has been added, altered, or detached from software intentionally so as to cause harm or disrupt the intended functions of the system."

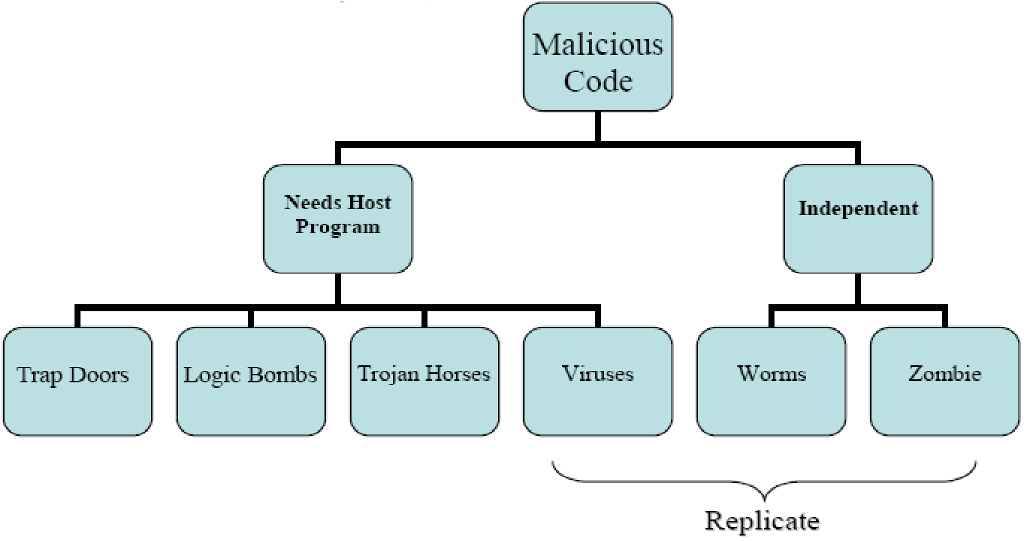

Fig 1: Kinds of Malicious Code

B. CATEGORIES OF MALWARE

Malware code can be classified in to various classes:

1) Viruses

A small program having malicious intent and the ability to create copies of itself is termed as virus. Virus code attaches itself with other programs and then replicates [3]. It works by inserting virus code in an executable file. Virus code gets executed when the file is run. Metamorphic viruses evolve into new variants by changing itself. A program in which virus has inserted itself is denoted as the host for the virus. A virus requires an existing host program so as to cause damage.

2) Worms

The software which makes its copies by executing its own code regardless of any other program is known as worm. It sends copies of itself to other systems in the network unnoticeably without the authorization of the user. Worms consume the bandwidth of the network to halt it. Worms spread through network linkages and have the aim to infect most of the computer systems associated with the network. Worms do not require the existence of any file unlike virus. It could encrypt files, delete files or send junk email.

3) Spyware

The software which monitors and collects private information of the user is termed as spyware. It usually gathers data like the key pressed by user, email address, pages regularly visited, credit card number etc [4]. This can happen when users download free or trial software. The users are monitored by scouts in this type of malware therefore their account numbers, passwords and every second personal element become exposed.

4) Adware

Adware also known as advertising-supported software presents, plays, or downloads advertisements in a computer system automatically once malicious software is installed or application is used. The code of this malware is normally embedded into free software. Its main objective is to monitor the activities of the user of the system. Free games and peer-to-peer clients are some examples of the common adware programs.

5) Trojans

A Trojan horse is a malware injected by its designer in an application or system. The application or system is accomplishing some illegal action but seems to carry out certain valuable function. Remote hijackers use this malware to launch their attack and they use our system for their own purpose. They could acquire our passwords, monitors what is happening on our system or corrupt the system files.

6) Botnet

Botnet is a type of malware that takes the control of our system distantly and sends spam or spyware. Mostly, botnets are like zombie and work under the command of the party who runs it. Bot doesn’t wait for a command from the third party by sitting around the infected machine. On the other hand, it looks for the communication having like occurrences of bots awaiting instructions. Simple and hierarchical are the two types of botnets.

C. MALWARE DETECTION TECHNIQUES

Malware detection is referred as classifying the code in to two classes: genuinely benign and malicious. Malware detection technique should have the capability of handling obfuscated malware efficiently for robust malware detection. The main idea behind code obfuscation is that it modifies only malware syntax but preserves its proposed behaviour [1]. A system that tries to recognize malware using signatures and other heuristics factors is known as malware detector.

Four methods employed for malware detection are:

1) Signature-Based Detection

Signatures are typically a sequence of bytes which the various antivirus scanners seek inside the malware code to state that the program scanned is malicious in nature. By observing the disassembled code of malware binary, these signatures are designed. Several debuggers and disassemblers are present for disassembling the portable executable. After that features are extracted by analyzing the disassembled code. These features are then used in the creation of the signature of specific category of malware.

2) Behavior-Based Detection

Behavior based detection focuses on the actions performed by the malware instead of the binary sequence. The programs having identical behavior but different syntax are recognized. Consequently several illustrations of malware are recognized by a single behavior. The main aim is to analyse the behaviour of well-known or unfamiliar malwares. Several elements such as source and destination address of malwares, attachments types and other statistical features constitute the behavioral parameters. These forms of detection mechanisms are helpful in discovering those malwares that continuously generates novel mutants as system resources and services are used by them in a similar way.

3) Specification-based Detection

Specification-based detection is derived from anomaly based detection. Specification-based detection estimates the necessities of an application or a system rather than estimating their execution. There is a training phase in specification-based method in which we try to acquire information about the legal behavior of a program or system which is being examined[4].The chief drawback of specification based technique is that it is very challenging to correctly specify the behavior of the system or program.

4) Anomaly-based detection

The anomaly detection technique is based on the idea of a baseline of the network behavior. The baseline specifies the accepted network behavior, which is stated by the network administrators. Any behavior that does not correspond with the predefined or accepted model of behavior causes events in an anomaly detection engine. It is a two-step approach that involves first training a system with data to create certain conception of normality and then use the established profile on actual data to report deviations. This approach gives anomaly-based IDSs the power to detect new attacks which are new and for which signatures are not present.

D. METRICS FOR DETECTION

The outcome of detection function can be categorized as:

1) False positive

It results when virus is truly not present in a non-infected file but a virus scanner erroneously detects it. A false positive arises when the signature used to spot the virus also appears in legal or non-infected software and is not solely for that virus [3].

2) False negative

It results when virus is actually present in an infected file but a virus scanner fails to detect it. Due to the non-availability of the virus signature and its novel nature, antivirus scanners proved unsuccessful in detection of that virus. The reason for the failure can also be attributed to the dynamic and complex configuration settings for that virus.

3) Hit ratio

It results when a malware detector accurately detects a virus in an infected file by matching the malware signature with the signature already present.

E. USE OF DATA MINING FOR MALWARE DETECTION

Data mining techniques are useful in detecting patterns in huge amount of data, for example byte code, and for detecting future occurrences in similar data by using these patterns. Data mining framework detect new malicious executable with the help of classifiers. A classifier is defined as a set of rules or detection model. It is generated by a data mining algorithm that is trained over a particular set of training data. One of the main difficulties faced by the virus community is to develop techniques for detecting new malicious programs that have not been examined yet. About ten to twelve malicious programs are generated daily and most of them cannot be correctly detected until signatures have been created for them. During that time period, systems that were protected with signature-based algorithms are susceptible to attacks. Nowadays, data mining has turn out to be the emphasis of many malware researchers to detect unidentified malware and to categorize malware from benign files. Features form the input data for the detection systems, and they can also be used as classification patterns in malware detection systems.

An approach in data mining for malware detection generally consists of using statistical techniques of classification. The model is constructed in each classification algorithm using machine learning for representing the benign and malicious classes. A labelled training set is needed in this method to form the class models during the supervised learning process [8]. The various existing statistical algorithms for classification includes Naive Bayes (NB) Algorithm, Sequential Minimal Optimization (SMO) Algorithm, k−Nearest Neighbour (KNN) Algorithm, Back propagation Neural Network Algorithm, J48 decision tree and Logistic Regression.

II. RELATED STUDY

A. PREVAILING ANALYSIS AND DETECTION TECHNIQUES

[1] Kirti Mathur in April, 2013 highlighted the prevailing analysis and detection techniques that are used for obfuscated malicious code. The major threat to computer system security is various malware that do the malicious actions. AV Scanners, Intrusion Detection System, and Firewalls are the various solutions that are used to detect these threats. Traditionally, all of these solutions for detection of malware detect their presence in our system by using malware signatures. But malware authors employ some obfuscation methods for which these methods proved unsuccessful.

B. THREE DATA MINING ALGORITHMS TO PRODUCE NEW CLASSIFIERS

[2] Milan Jain in August, 2014 proposed three data mining algorithms to produce new classifiers with separate features. The three algorithms are RIPPER, Naïve – Bayes and a Multi Naïve Bayes Classifier. The author also compared these three algorithms. Three phases comprising it is root kit data collection, pre-processing of data, then its classification and evaluation of performance. With the growth in high-speed Internet connections, malware are spreading very rapidly. Therefore, it is very essential to detect and delete benign malware in an effective way.

C. DEFINITION, TYPES, PROPAGATION OF MALWARE, AND THEIR DETECTING TECHNIQUES

[3] Mohsen Damshenas, Ali Dehghantanha, Ramlan Mahmoud in 2013 closely looked into the concept of malware, to know the definition, types, propagation of malware, and their detecting techniques so as to enhance the method of protection and security. The security experts practice all promising techniques, strategies and methods to halt and eliminate the threats whereas the malware authors exploit new types of malwares that bypass employed security features.

D. SEVERAL MALWARE DETECTION METHODS

[4] Vinod P. focused on several malware detection methods like signature based detection methods, reverse engineering of obfuscated code, for detecting malicious codes. Malwares are malicious software's. They are intended to harm computer systems devoid of the knowledge of the owner of the system. Software's that came from trustworthy vendors also contain malicious code which disrupts the system and discloses private information to remote servers. Malware's consist of computer viruses, spyware, ad-ware, Trojans etc. [5] Nwokedi Idika and Aditya P. Mathur in February 2007 had examined 45 malware detection methods and provided a chance to compare them with one another which would help in the process of decision making involved in the development of a secure application. The survey also provided a comprehensive bibliography to assist the researchers in malware detection. Malware detectors are the chief tools that provide protection against malware. The technique used by such malware detectors determines their effectiveness. Therefore it is very important to study malware detection techniques and recognize their merits and demerits.

E. CLASSIFICATION OF MALICIOUS EMAILS USING NAÏVE BAYES CLASSIFIER

[6] B.V.R.R.Nagarjuna in July 2013 explained how the malicious emails are classified and how these are deleted and how to know the contents of messages. The author used Bayesian spam filtering, Email filtering and J48 process for classification which overcomes the difficulties arose in linear C-Support Vector Machine (C-SVM). This machine had given the correct results when compared to the existing one. There are three steps to examine first one is to detect the malicious email, next step is to apply classifier to classify according to the emails received and send to trash automatically and delete directly. [7] Ion Androutsopoulos, John Koutsias, Konstantinos, Constantine D. Spyropoulos and V. Chandrinos in Aug, 2000 proposed a method in which a Naïve Bayesian classifier is automatically trained to identify spam messages. The author tested this method on a large set of personal e-mail messages, which are available in encrypted form publicly. The author presented proper cost-sensitive measures. The author also examined the influence of training data size, lemmatization, size of attribute set, stop lists and the issues that had not been explored till then. Lastly, the Naive Bayesian classifier is compared to a filter makes use of keyword patterns to find its effectiveness.

F. DATA MINING APPLICATIONS AND SEVERAL CLASSIFICATION ALGORITHMS

[8] Vishnu Kumar Goyal in April, 2014 had worked with diverse data mining applications and several classification algorithms. The algorithms had been applied on different dataset to find out the effectiveness of the algorithms. Classification is an important technique of data mining. It has wide applications in classifying the various kinds of data used in almost every field of human life. The author analysed the five main classification algorithms: Decision Tree (DT), Decision Stump (DS), k-nearest neighbourhood (KNN), Naive Bayes (NB) and Rule Induction (RI) and compared their performance. The results were verified on five datasets namely Golf, Iris, Weighting, Deals and Labor using Rapid Miner Studio.

Table 1: Comparison of various malware detection techniques

Technique	Approach	Based on	Effective in	Pros	Cons	Accuracy
Signature Based Detection	Actively compare and match current behaviour against a large collection of signatures**.**	Predefined rules for known attacks	Detection of attacks having a fixed behaviour pattern.	Signature can be used as a standalone system. Works well for known signatures.	Signature extraction and distribution is a complex task. The signature generation involves manual intervention. The growing size of signature repository.	Lower false alarm rates Low false positives
Behavior Based Detection	identifies the action performed by Malware. Single	Signatures of malicious	Detection of mutants of malware	May detect a wide range of novel attacks	Usage patterns may change often.	Higher false alarm

	behaviour signature can identify various samples of malware.	behavior		Can be cheap to deploy and monitor	Post-facto, attack already occurred Easy to evade once known	rates Low false positives
Anomaly Based Detection	Detects behaviors that fall outside the predefined or accepted model of behavior.	Notion of normality	Detection of new unknown attacks	Can detect potentially a wide range of novel attacks.	May miss known attacks High overhead	High false alarm rates High false positives
Specification Based Detection	Leverage some specification or rule set of what is valid behavior. Programs violating the specification are considered malicious.	Statistical machine learning	Detection of attacks having fixed specification.	Approximates the requirements of application or system. Address the high false alarm rate problem.	it is very difficult to accurately specify the behaviour of the system or program.	Low false alarm rate High false positives

III. CONCLUSION

We have studied about the concept of malware and its types as well as various techniques used for detecting it. From the comparison of various techniques, it has been found that for the attacks having a fixed behavior pattern signature based methods works well. But they are not effective in detecting novel attacks for which no signatures are present. To detect new and unspecified attacks, anomaly based method gives better performance. In the case of polymorphic and metamorphic malware which change their signatures and creates their new variants every time, behavior based approach is most effective.

REFERENCES

[1] Kirti Mathur. "A Survey on Techniques in Detection and Analyzing Malware Executable". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[2] Milan Jain. "Malicious Detection Using Multiple Classification Algorithms & Their Comparison Using Different Clustering Techniques". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[3] Mohsen Damshenas, Ali Dehghantanha, Ramlan Mahmoud. "A survey on malware propagation, analysis and detection". International Journal of Cyber-Security and Digital Forensics (IJCSDF), 2013.

[4] Vinod P. "Survey on Malware Detection Methods". Department of Computer Engineering, Malaviya National Institute of Technology.

[5] Nwokedi Idika and Aditya P. Mathur. "A Survey of Malware Detection Techniques". Research supported by Committee on Institutional Cooperation, February 2007.

[6] B.V.R.R.Nagarjuna, V. Sujatha. "An Innovative Approach for Detecting Targeted malicious E-mail". International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 7, July 2013.

[7] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, and Constantine D. Spyropoulos. " An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages". In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Aug 2000.

[8] Vishnu Kumar Goyal, Dept. of Computer Engineering. "A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio". International Journal of Innovative Research in Science & Engineering, April 2014.

[9] Gerard Wagener,Radu State,Alexandre Dulaunoy,"Malware Behavior Analysis",Springer-Verlag France 2007.

[10] Jonathan A.P. Marpaung, Mangal Sain and Hoon-Jae Lee: Survey on malware evasion techniques: state of the art and challenges, International Conference of Advanced Communication Technology, pp 19-22, 2012.

[11] Rizwan Rehmani , G.C. Hazarika and Gunadeep Chetia : Malware Threats and Mitigation Strategies: A Survey, Journalof Theoretical and Applied Information Technology, Vol. 29 No.2, 2011.

[12] F. Adelstein, M. Stillerman, and D. Kozen. Malicious code detection for open firmware. In Proceedings of the 18th Annual Computer Security Applications Conference, 2002.

[13] Anshul Goyal and Rajni Mehta (2012). "Performance Comparison of Naïve Bayes and J48 Classification Algorithms".

[14] Matthew G. Schultz, EleazarEskin, ErezZadok and Salvatore J. Stolfo "Data Mining Methods for Detection of New Malicious Executables".

[15] M. Hall and E. Frank. Combining naive Bayes and decision tables. In Proc 21st Florida Artificial Intelligence Research Society Conference, Miami, Florida. AAAI Press, 2008.

[16] J. ZicoKolter and Marcus A. Maloof, “Learning to Detect and Classify Malicious Executables in the Wild,” Journal of Machine Learning Research, 7, 2006, p 2721-2744.

[17] Ross Thomas, Dmitry Samosseiko, SophosLabs Canada. “The game goes on: an analysis of modern spam Techniques”. Virus bulletin conference, October 2006.

[18] UsukhbayarBaldangombo, NyamjavJambaljav and Shi-Jinn Horng, “a static malware detection system using data mining methods”.