Malware and Malware Detection Techniques: A Survey

Jyoti Landage
ME, Dept of Comp Engg
Sinhgad College of Engg, Vadgaon, Pune

Prof. M. P. Wankhade
Professor,Dept of Comp Engg.
Sinhgad College of Engg, Vadgaon, Pune

Abstract

Now a day's malicious program is a serious threat. It is developed to damage the computer system and some of them are spread over the connected system in the network or internet connection. Researchers are taking great efforts to produce anti-malware system with effective malware detection methods to protect computer system. Two basic approaches have been proposed for it i.e. signature-based and heuristic-based detection. These approaches detect known malware accurately but cannot detect the new, unknown malware.

Recently different researchers have proposed malware detection system using data mining and machine learning methods to detect known as well as unknown malwares. In this paper, a detailed analysis has been conducted on the current state of malware infection and work done to improve the malware detection systems.

Keywords: anti-malware system, data mining, heuristic-based, malware, malware detection system, signature-based.

1. Introduction

Now a day the use of internet is the most integral part of modern life. The internet browser downloads different types of computer software. One drawback of the widespread use of internet is that many computer systems are vulnerable to attacks and get infected with malwares. There are different names for malware for example malicious code, malicious program or malicious executable. Malware is malicious software which is used with the intention of breaching a computer system's security policy with respect to confidentiality, integrity and availability of data. It can add, change or remove any program from the system to intentionally harm the system's required functions. Malware comes in different forms such as virus, Trojan horse, spyware, scareware, adware or trapdoor etc. A more recent annual report on the Internet security threat-2013 from Symantec says "Threat to online security have grown and evolved considerably in 2012, In particular, social media and mobile devices have come under increasing attack in 2012"[1].

Malware detection system is a system used to determine whether a program has malicious intent or not. Detection system includes two tasks - analysis and detection [4]. Malware detector is used as a tool to defense against the malware. The qualities of such detectors are determined by the techniques it uses. It takes two inputs first is signature or behavioral parameters of a given code and second is the program under inspection, it can employ its detection technique to decide if the program is malware or benign.

The main purpose of this review paper is to investigate the different forms of malware, malware analysis and detection techniques, their comparative study to understand their strength and limitations. Thus one can have clear idea of current state of the art and try to come up with better solution. The paper is organized as follows: section 2 describes the malware classification; section 3 describes the malware analysis techniques and their comparison, whereas in section 4 we illustrate the detection process of malware by explaining malware detection techniques. In section 5 the survey of existing work on advanced detection method has been presented. Section 6 concludes the paper with remarkable comments.

2. Malware Classification

The classification of malware is a difficult process. Software that allows unauthorized control of a system is obviously malicious. Malware comes in various forms and categories. These are usually classified according to their propagation method and their actions that are performed on the infected machine using the designed malicious program.

The following list presents the common types of malware.

i. Virus: A malicious program propagates from one program to another or from one computer to another by inserting their code into other program.

ii. Worm: It is a self-replicating program which spread from one computer to another by transmitting copy of itself via a network without user authorization [14].

iii. Trojan horse: Trojans mask themselves by appearing to be something legitimate. Trojans typically destroy data or attempt to extract confidential information including financial data & passwords [5]

iv. Spyware: Spyware is any software installed on system without user's knowledge. It is a collective term for software which monitors and gathers personal information about the user and sends that information back to the attacker so the attacker can use the stolen information in some disreputable way. It generally enters a system when free or trial software is downloaded and installed on the system without the user's knowledge, changes the setting of your browser or adds abominable browser toolbars. [15, 17].

v. Scareware: Scareware is a malware masquerading as free or trial anti-virus software or some other free online scam. It can be installed by the user when downloading bogus security software, opening attachments or by visiting a malicious website. After installation it collects all information stored on your computer (financial details, personal info) which could be sold to other cyber criminals.

vi. Adware: Adware is advertising supported software that automatically plays, displays, or downloads advertisements to a computer after malicious software is installed or application is used. This piece of code is generally set into free downloaded software. The most common source of adware programs are free games, peer-to-peer clients like KaZaa, BearShare etc.[6]

vii. Botnet: A botnet is remotely controlled autonomous software. It is usually a zombie program (worms, Trojans) under common control for any network infrastructure [6].

3. Malware analysis Technique

Malware analysis is necessary to develop effective malware detection technique. It is the process of analyzing the purpose and functionality of a malware, so the goal of malware analysis is to understand how a specific piece of malware works so that defense can be built to protect the organization's network. There are three types of malware analysis which achieve the same goal of explaining, how malware works, their effects on the system but the tools, time and skills required to perform the analysis are very different.

3.1. Static analysis

It is also called as code analysis. It is the process of analyzing the program by examining it i.e. software code of malware is observed to gain the knowledge of how malware's functions work. In this technique reverse engineering is performed by using disassemble tool, decompile tool, debugger, source code analyzer tools such as IDA Pro and Ollydbg in order to understand structure of malware [9]. Before program is executed, static information is found in the executable including header data and the sequence of bytes is used to determine whether it is malicious. Disassembly technique is one of the techniques of static analysis. With static analysis executable file is disassembled using disassemble tools like XXD, Hexdump, NetWide command, to get the assembly language program file. From this file the opcode is extracted as a feature to statically analyze the application behavior to detect the malware.

3.2. Dynamic analysis

It is also called as behavioral analysis. Analysis of infected file during its execution is known as dynamic analysis [2]. Infected files are analyzed in simulated environment like a virtual machine, simulator, emulator, sandbox etc [6]. After that malware researchers use SysAnalyzer, Process Explorer, ProcMon, RegShot, and other tools to identify the general behavior of file [9]. In dynamic analysis the file is detected after executing it in real environment, during execution of file its system interaction, its behavior and effect on the machine are monitored. The advantage of dynamic analysis is that it accurately analyzes the known as well as unknown, new malware. It's easy to detect unknown malware also it can analyze the obfuscated, polymorphic malware by observing their behavior but this analysis technique is more time consuming. It requires as much time as to prepare the environment for malware analysis such as virtual machine environment or sandboxes.

3.3. Hybrid Analysis

This technique is proposed to overcome the limitations of static and dynamic analysis techniques. It firstly analyses the signature specification of any malware code & then combines it with the other behavioral parameters for enhancement of complete malware analysis. Due to this approach hybrid analysis overcomes the limitations of both static and dynamic analysis [6].

Static and dynamic analyses are differentiated in following Table 1 in the form of its advantages and disadvantages.

Table 1. Comparison of static and dynamic analysis

Sr.No.	Static analysis	Dynamic Analysis
1.	Fast & safe	Time Consuming & vulnerable
2.	Good in analyzing the multipath malware (Global view)	Difficult to analyze the multipath malware
3.	Can't analyze the obfuscated & polymorphic malware	Can analyze the obfuscated and polymorphic malware
4.	Can't detect new, unknown malware	Detect known as well as unknown malware
5.	Low level of false positive (accuracy is high)	High level of false positive (accuracy is low)

4. Malware Detection Technique

Malware detection techniques are used to detect the malware and prevent the computer system from being infected, protecting it from potential information loss and system compromise. They can be categorized into signature-based detection, behavior-based detection and specification-based detection.

4.1 Signature-based detection

It is also called as Misuse detection. It maintains the database of signature and detects malware by comparing pattern against the database. General flow of signature-based malware detection and analysis is explained in detail in [15]. Most of the antivirus tools are based on the signature-based detection techniques. These signatures are created by examining the disassembled code of malware binary. Disassembled code is analyzed and features are extracted. These features are used in constructing the signature of particular malware family. A library of known code signatures is updated and refreshed constantly by the antivirus software vendor so this technique can detect the known instances of malware accurately. The main advantages of this technique is that it can detect known instances of malware accurately, less amount of resources are required to detect the malware and it mainly focus on signature of attack. The major drawback is that it can't detect the new, unknown instances of malware as no signature is available for such type of malware.

4.2 Heuristic-based detection

It is also called as behavior or anomaly-based detection. The main purpose is to analyze the behavior of known or unknown malwares. Behavioral parameter includes various factors such as source or destination address of malware, types of attachments, and other countable statistical features. It usually occurs in two phase: Training phase and detection phase. During training phase the behavior of system is observed in the absence of attack and machine learning technique is used to create a profile of such normal behavior. In detection phase this profile is compared against the current behavior and differences are flagged as potential attacks [13]. Figure 1 show the behavior detector which basically consists of following components.

A. Data collection: This component collects the dynamic and static information.
B. Interpretation: Converts the raw information collected by data collection module into intermediate representations.
C. Matching Algorithm: It is used to compare the representation with the behavior signature.

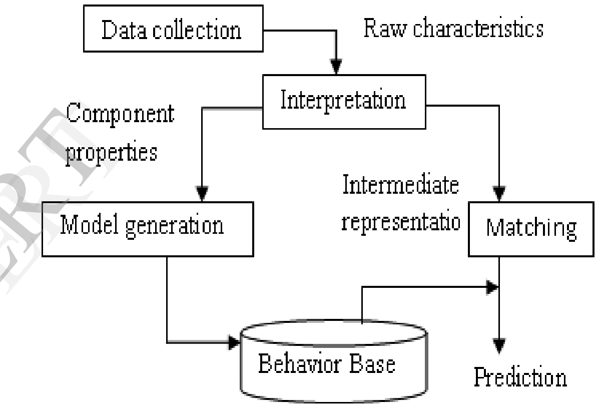

Figure 1. Behavior detector [17]

The advantage of this technique is that it can detect known as well as new, unknown instances of malware and it focuses on the behavior of system to detect unknown attack. The disadvantage of this technique is that it needs to update the data describing the system behavior and the statistics in normal profile but it tends to be large. It need more resources like CPU time, memory and disk space and level of false positive is high.

4.3 Specification-based detection

It is derivative of behavior-based detection that tries to overcome the typical high false alarm rate associated with it. Specification based detection relies on program specifications that describe the intended behavior of security critical programs [13]. It involves monitoring program executions and detecting deviation of their behavior from the specification, rather than detecting the occurrence of specific attack patterns. This technique is similar to anomaly detection but the difference is that instead of relying on machine learning techniques, it will be based on manually developed specifications that capture legitimate system behavior [13]. The advantage of this technique is that it can detect known and unknown instances of malware and level of false positive is low but level of false negative is high and not as effective as behavior based detection in detecting new attacks; especially in network probing and denial of service attacks. Development of detailed specification is time consuming.

From last decade data mining has been the main focus of many malware researcher for detecting the new, unknown malwares; they have added data mining as a fourth proposed malware detection technique. In 2001 Schultz [7] first introduced the idea of applying the data mining and machine learning method for the detection of new, unknown malware based on their respective binary codes. Then different studies have been conducted for detection of different malwares. Data mining helps in analyzing the data, with automated statistical analysis techniques, by identifying meaningful patterns or correlations. The results from this analysis can be summarized into useful information and can be used for prediction. Machine learning algorithms are used for detecting patterns or relations in data, which are further used to develop a classifier [10]. The common method of applying the data mining technique for malware detection is to start with generating a feature sets. These feature sets include instruction sequence, API/System call sequence, hexadecimal byte code sequence (n-gram) etc. The numbers of extracted features are very high so various text categorization techniques are applied to select consistent features and generate the training and test feature sets. Then classification algorithms are applied on the consistent training feature set to generate and train the classifier and test feature set is examined by using these trained classifiers. The performance of each classifier is evaluated by identifying the rate of False Positive, False Negative, True Positive, True Negative and calculate the TPR, FPR, Recall, precision and F1-measure. The survey of various feature selection technique & classification technique used for data mining is presented in [16]. The advantage of data mining based detection is that detection rate is high as compared to signature-based detection method [7]. It detects the known as well as unknown, new instances of malware.

5. Survey of Existing work

In the previous section we have presented the malware analysis and detection technique as base of this paper work. Following table shows comparison of the studied literature survey papers of various techniques that are used to detect the malwares using data mining and machine learning method.

Table 2. Comparison of Studied papers

Study	Feature representation and extraction	Feature selection	Classifiers	Conclusion
Detecting unknown malicious code by applying classification techniques on OpCode patterns (2012) [3]	Opcode n-gram	TF-IDF	SVM, LR,RF,ANN, DT,NB and their boosted version BNB and BDT	Evaluated number of experiments & found that setting of 2-gram, TF, using 300 features selected by DF measure outperformed. The performance of decision tree & boosted decision tree was very well as compared to NB & boosted NB
Detecting scareware by Mining Variable Length Instruction Sequences (2011) [12]	Opcode n-gram	TF-IDF, CPD	Jrip, SMO, DT, IBk, NB, Random forest	This paper presents the static analysis method based on data mining which extends the general heuristic detection technique using a variable length instruction sequence mining approach for the purpose of scareware detection
Accurate Adware Detection using Opcode Sequence Extraction (2011) [10]	Opcode n-gram	TF-IDF, CPD	ZeroR ,Naïve Bayes, SMO, IBk, J48, JRip	Detects adware using data mining & ML method. KNN and SVM were effective when the data was noisy, KNN's performance is superior incrementally when new training samples are introduced, JRip and J48 algorithms are expensive in term of time consumption to train and generate the model but it is easy to analyze the rules and trees generated to differentiate the non-malicious and malicious files.
Detection of Spyware by Mining Executable Files(2010) [11]	Byte sequence, n-gram	CFBE and FBFE	ZeroR, Naïve bayes, SMO, J48, Random forest, JRip	Detects spyware by using data mining &ML method. Feature set generated by CFBE selection method generally produced better results with regard to accuracy than feature sets generated by FBFE method.

Study	Feature representation and extraction	Feature selection	Classifiers	Conclusion
Using Multi-Feature and Classifier Ensembles to Improve Malware Detection(2010) [18]	Content (DLL/API function call from PE) and Behavior based feature	Information gain	K-nearest neighbor, Naïve bayes, SVM, decision trees	Improved the accuracy of ML using classifier ensemble to replace individual classifier. Introduced ensemble learning algo. Like bagging, boosting, voting etc. proposed new ensemble learning method SVM-AR which outperforms nearly all popular ensemble learning algo.
Data Mining Methods for Detection of New Malicious Executables (2001) [7]	PE, String, Byte Sequence (n-gram)	NA	Ripper, Naïve Bayes, multi-naïve bayes	The highest accuracy and detection rate of multi-naïve bayes algo with 97.76%, over double the signature-based method.

6. Conclusion

Malware is a critical threat to user's computer system in terms of stealing confidential information, corrupting or disabling security system. This survey paper presents some existing technologies used by security researchers to tackle these threats. It explains static, dynamic and hybrid malware analysis techniques, their comparative study, existing traditional malware detection techniques and their advantages-disadvantages. According to their comparative study we are going to use advanced malware detection technique i.e. data mining and machine learning method to overcome the drawbacks of existing malware detection techniques.

References

[1] Symantec Corporation, Internet security threat report-2013, Volume 18

[2] Ammar Ahmed E. Elhadi, Mohd Aizaini Maarof and Ahmed Hamza Osman, Malware Detection Based on Hybrid Signature Behaviour Application Programming Interface Call Graph, American Journal of Applied Sciences 9 (3): 283-288, 2012, ISSN 1546-9239, 2012, Science Publications

[3] Asaf Shabtai, Robert Moskovitch, Clint Feher, Shlomi Dolev and Yuval Elovici, Detecting unknown malicious code by applying classification techniques on OpCode patterns, Security Informatics 2012, 1:1, http://www.securityinformatics.com/content/1/1/1.

[4] Imtithal A. Saeed, Ali Selamat, Ali M. A. Abuagoub, A Survey on Malware and Malware Detection Systems, International Journal of Computer Applications (0975 – 8887) Volume 67– No.16, April 2013

[5] Jonathan joseph bloun, adaptive rule-based malware detection employing learning classifier systems, Thesis-Master of science in computer science, Missouri University of science and technology, 2011.

[6] 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, ISSN: 2277 128X, Volume 3, Issue 4, April 2013.

[7] Matthew G. Schultz, Eleazar Eskin, Erez Zadok, and Salvatore J. Stolfo, Data Mining Methods for Detection of

New Malicious Executables, in Proceedings of the Symposium on Security and Privacy, 2001, pp. 3849.

[8] Muazzam Ahmed Siddiqui, Data Mining Methods For Malware Detection, Thesis, B.E. NED University of Engineering and Technology, M.S. University of Central Florida, 2008

[9] Pham Van Hung, An approach to fast malware classification with machine learning technique, Keio University,5322 Endo Fujisawa Kanagawa 252-0882 JAPAN, 2011

[10] Raja Khurram Shahzad, Niklas Lavesson, Henric Johnson, Accurate Adware Detection using Opcode Sequence Extraction, in Proc. of the 6th International Conference on Availability, Reliability and Security (ARES11),Prague, Czech Republic. IEEE, 2011, pp. 189195.

[11] R. K. Shahzad, S. I. Haider, and N. Lavesson, Detection of spyware by mining executable files, in Proceedings of the 5th International Conference on Availability, Reliability, and Security. IEEE Computer Society, 2010, pp. 295302.

[12] R. K. Shahzad and N. Lavesson, Detecting scareware by mining variable length instruction sequences,in Proc. of the 10th Annual Information Security South Africa Conference (ISSA11), Johannesburg, South Africa. IEEE, August 2011, pp. 18.

[13] Robiah Y, Siti Rahayu S., Mohd Zaki M, Shahrin S., Faizal M. A., Marliza R.,A New Generic Taxonomy on Hybrid Malware Detection Technique, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 5, No. 1, 2009.

[14] Robin Sharp, An Introduction to Malware, Spring 2012. Retrieved on April, 10, 2013

[15] Ronghua Tian, An Integrated Malware Detection and Classification System, Changchun University of Science and Technology, Thesis, August, 2011

[16]Sunita Beniwal, Jitender Arora, Classification and Feature Selection Techniques in Data Mining, International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181

[17] Vinod P. V.Laxmi,M.S.Gaur: Survey on Malware Detection Methods, 3rd Hackers‟ Workshop on Computer and Internet Security, Department of Computer Science and Engineering, Prabhu Goel Research Centre for Computer & Internet Security,IIT, Kanpur, pp-74-79, March,2009.

[18] Yi-Bin Lu, Shu-Chang Din, Chao-Fu Zheng, and Bai-Jian Gao,Using Multi-Feature and Classifier Ensembles to Improve Malware Detection, JOURNAL OF C.C.I.T., VOL.39, NO.2, NOV., 2010.