Mobile Keylogger Detection By Using Machine Learning Technique

[1]S.vinothkumar, [2]S.Aruna sankaralingam
[1]M-Tech (Software Engineering) PG Scholar, [2]Assistant Professor (Software Engineering).
[1,2]Dept of Software Engineering ,SRM University ,Chennai, Tamilnadu , India
[1]vinoth.raman90@gmail.com, [2]aruna.s@ktr.srmuniv.ac.in

Abstract - Keylogger, a highly specialized tool designed to record every keystroke made on the machine to giving the attacker the ability to steal large amounts of sensitive information silently. The primary objective of this project is to detect keylogger applications and prevent data loss and sensitive information leakage. In This project aims to identify the set of permissions and storage level owned by each of the applications and hence differentiate applications with proper permissions and keylogger applications that can abuse permissions .This technique of detecting keyloggers is completely Black-box its based on behavioral characteristics common to all keyloggers and it does not rely on the internal structure of the keylogger. The paper intends to develop a machine learning-based keylogger detection system on mobile phones to detect malware applications.

Keywords - Keylogger, Black-box, malware, machine learning, Smartphone, spyware.

I.INTRODUCTION

Keyloggers are used in IT organizations to troubleshoot technical problems with computers and business networks. The Keyloggers can also be used by a family or business people to monitor the network usage of people without their direct knowledge. Mobile Keylogger is installed onto a mobile phone to monitor the activities of the mobile phone user. For example spyware ,flexispy that can monitor activities on the mobile device and then send them to a remote server. Passwords and banking information have been stolen using keyloggers. Keyloggers are sometimes part of malware packages downloaded onto the mobile phone without the owners' knowledge. It intercepts all of the communication with out user's knowledge. This can be used for legitimate purposes that the mobile phone users are using their phones for inappropriate activities. This system exploits machine learning techniques to distinguish between normal and malware applications.

The first real mobile phone virus was found in the world and could replicate on its own. On June 15 2004, Finnish anti-virus firm F-Secure and Russian rival Kaspersky released details about a piece of mobile phone malware that used Bluetooth to try and spread to other Symbian series60-based mobile phones. the usage of Smartphone is increasing.Mobile malwares are increased because the mobile devices are vulnerable to viruses, spams and malwares, and the usage of the mobile devices are increased. Smartphones are also required to be made safe and secure as we do for personal computers. The overall Smartphone security market is segmented on the basis of mobile operating systems, ownership, and features. The market by ownership is segmented into individual and business segment. In the early 1990s, very few PC users had the concept of the threat posed by malware. Anti-virus software was not widely deployed at the time, and much of the earliest malware was not designed to maliciously exploit other systems. The PC malware has increased dramatically over the past 15–20 years. Today's varieties focus on fraud, identity theft and distributed denial of service (DDoS) attacks they have become stealthier, finding new and innovative ways to conceal themselves on PCs and spread undetected. Despite billions of dollars invested in R&D to combat malware, it is estimated that approximately 30–50 percent goes undetected by anti-virus software. Kindsight Security Labs has found that 15–20 percent of home networks consistently show infections, including Trojans, botnets, spambots and keyloggers.The first virus capable of infecting mobile devices without needing a PC to transmit itself was discovered in the summer of 2004, the Cabir worm began affecting phones running the Symbian operating system.

The first confirmed malware affecting Apple iPhones appeared in late 2009. However, because such malware could be transmitted to and infect only 'jailbroken' iPhones. Malware for Google Android devices began to appear in Chinese app markets in late 2010. However, much like the original Cabir worm, these programs were more 'proofs-of-concept' than sophisticated attacks. That quickly changed in 2011 when a number of vendors and other observers detected a notable rise in malware communications—more specifically, the command-and-control (C&C) protocols used by malware to call home with stolen information—coming from Android devices. Lookout Mobile Security reported that upwards of one million people were affected by Android malware in the first half of 2011.

The number of infections continued to climb over the second half of the year, with Kindsight Security Labs reporting a 400 percent increase in Android infections per week from early September to late November 2011. In the Mobile Threats Report, Juniper Networks are noted a whopping 3,325 percents to be increase in malware specifically to targeting the Android

platform—from just 400 in June to more than 13,300 by the end of the year. one of the biggest malware-related concerns is identity theft, a serious crime in which personal informations including banking and credit card details is stolen and used without permission to commit fraud or a number of other crimes—often with devastating financial consequences. Because today's mobile malware is not quite as sophisticated as PC-based malware and it tends to focus on a lower level of identity theft: the stealing of contact lists and address books from mobile devices in order to send unsolicited SMS and email messages under the guise of the device owner .Its Not only an inconvenience to the people receiving these spam messages, it can also cost device owners money by racking up fees for data usage or the sending/receiving of premium SMS messages. This approach may represent the beginning of an SMS spam market that could eventually rival the traditional email spam used in wireless networks.

II.RELATED WORK

We first review related work on behavior-based malware detection techniques have been proposed in the desktop environments. Next, we review on keylogger detection based on machine learning techniques .Unprivileged black-box approach for accurate detection of the most common user space keyloggers is presented in [1] and the behavior of a keylogger is modeled by surgically correlating the input with the output. They used the Pearson product correlation co-efficient (PCC) as the detection algorithm. In their approach the keylogger may rely on aggressive buffering and keylogger may trigger the keylogging activity only in occurrence of particular events. Malware is analyzed with respect to its behavior irrespective of the similarities in file contents in [10].

In This framework is proposed to overcome the computer security problems, such as denial-of-service attacks, identity theft, or distribution of spam and phishing contents. It uses learning algorithms such as clustering and classification, for analysis of malware behavior. Bu using their approach, the run-time as well as memory requirements can be reduced .But large number of malware samples to be collected to implement it practically and should monitor the behavior of the malware in the sandbox environment. In [9], a learning based method for the detection of malicious android applications is proposed. Support Vector Machine is used to classify the samples into normal and malicious applications. Main advantage of this system is that it is able to run in time linear in the number of nodes and can therefore process graphs with thousands of nodes such as the function call graphs of Android applications. This method outperforms these approaches and detects 89% of the malware at a false-positive rate of only 1%|corresponding to only one false alarm when installing 100 applications on a Smartphone .In [8] presented a robust and lightweight approach for detecting Android malware based on different classifiers. Rather than following a heuristic basedapproach for determining the feature vector of the classifiers, we have statically analyzed a large corpus of Android malwares belonging to different families and a large benign set belonging to different categories. [7] uses Approximate flow graph matching algorithm that uses the decompilation technique of structuring.

In This system proposed malware classification using either the edit distance between structured control flow graphs or the estimation of isomorphism between control flow graphs. It develops an effective and efficient system to solve the polymorphic malware problem. Kernel based behavior analysis system [6] achieves collecting log data that only contains data of target activities. The Log data is analyzed by signature-based pattern matching. Kernel-based behavior analysis can be applied for security inspections for Android application markets. It detects malicious behaviors of the unknown applications. Low error rate of a false negative and a false positive is achieved by carefully described signatures.[2] characterizes two types of human behavior they are operational behavior and mobile behavior .In the order to observe and uncover the propagation mechanisms of mobile viruses .It can be extended to incorporate additional characteristics of human mobility and operations. Naive Bayesian model [3] has been applied in non-DTN settings. The extension of the behavioral characterization of proximity malware for strategic malware detection evasion with game theory. It presents two techniques dogmatic filtering and adaptive look-ahead. [4] Presented a tool that allows testers to define and execute models of attackers, aiming at finding software vulnerabilities. The tool VERA Executes the Right Attacks are allows testers to define attacker models by means of extended finite state machines (EFSM). [5] Presented an effective approach to alleviate malware targeting the android platform .Problem based on Bayesian classification models obtained from static code analysis. The results presented in the paper showed better detection rates than were achieved by popular signature-based antivirus software. Investigate the classifier performance with larger sample sets.

Table 1: Comparison of Different Methodologies

S.NO	Paper Title	Authors	Journal/Conference	Year	Methods & Technique Used	Conclusion & Future Enhancement
1	Unprivileged Black-Box Detection of User-Space Keyloggers	Stefano Ortolani, Cristiano Giuffrida, and Bruno Crispo,Senior Member, IEEE	IEEE Transactions on dependable and secure computing, Vol.10,No.1	2013	Pearson product moment correlation coefficient.	Presented an unprivileged black-box approach for accurate detection of the most common keyloggers.
2	Modeling and Restraining	Chao Gao and Jiming Liu, Fellow,	IEEE Transactions on Mobile computing Vol. 12,	2013	Autonomy-oriented computing	Characterizes two types of human behavior, i.e.,

	Mobile Virus Propagation	IEEE	No. 3,		(AOC) strategy.	operational behavior and mobile behavior, to observe the propagation mechanisms of mobile viruses. Extended to incorporate additional characteristics of human mobility and operations.
3	Behavioral Malware Detection in Delay Tolerant Networks	Wei Peng, Student Member, IEEE, {Feng Li, Xukai Zou}, Member, IEEE, and Jie Wu, Fellow, IEEE	IEEE Transaction on Parallel and Distributed Systems.	2013	Presents two techniques dogmatic filtering and adaptive look-ahead.	Naive Bayesian model has been applied in non-DTN settings. The extension of the behavioral characterization of proximity malware for strategic malware detection evasion with game theory..
4	VERA: A flexible model-based vulnerability testing tool	Abian Blome, Mart´ın Ochoa, Keqin Li, Michele Peroli, Mohammad Torabi Dashti	IEEE Sixth International Conference on Software Testing, Verification and Validation	2013	The tool VERA Executes the Right Attacks, which allows testers to define attacker models by means of extended finite state machines (EFSM).	Presented a tool that allows testers to define and execute models of attackers, aiming at finding software vulnerabilities.
5	A New Android Malware Detection Approach Using Bayesian Classification	Suleiman Y. Yerima, Sakir Sezer, Gavin McWilliams, Igor Muttik	IEEE 27th International Conference on Advanced Information Networking and Applications	2013	Novel application of Bayesian classification is Applied.	The results presented in the paper showed better detection rates than were achieved by popular signature-based antivirus software. Investigate the classifier performance with larger sample sets
6	Kernel-based Behavior Analysis for Android Malware Detection	Takamasa Isohara, Keisuke Takemori and Ayumu Kubota	Seventh International Conference on Computational Intelligence and Security	2011	Uses Jailbreak Techniques.	The system achieves collecting log data that only contains data of target activities. Log data is analyzed by signature-based

						pattern matching. Kernel-based behavior analysis can be applied for security inspections for Android application markets.
7	Malwise—An Effective and Efficient Classification System for Packed and Polymorphic Malware	Silvio Cesare, Student Member, IEEE, Yang Xiang, Senior Member, IEEE, and Wanlei Zhou, Senior Member, IEEE	IEEE Transaction on Computers, Vol. 62, No. 6,	2013	Uses Approximate flowgraph matching algorithm that uses the decompilation technique of structuring.	Proposed malware classification using either the edit distance between structured control flow graphs, or the estimation of isomorphism between control flow graphs. Develop an effective and efficient system to solve the polymorphic malware problem
8	Droid APIMiner: Mining API-Level Features for Robust Malware Detection in Android	Yousra Aafer, Wenliang Du, and Heng Yin	9th International Conference on Security and Privacy in Communication Networks, Sydney, Australia	2013	Uses Malware detection techniques that rely on static analysis.	To predict whether an app is benign or malicious, the classifiers rely on the semantic information within the bytecode of the applications ranging from critical API calls, package level information and some dangerous parameters invoked. Reduce the false positives and negatives through analyzing the samples
9	Structural Detection of Android Malware using Embedded Call Graphs	Hugo Gascon, Fabian Yamaguchi, Daniel Arp, Konrad Rieck	Proc. of 6th ACM CCS Workshop on Artificial Intelligence and Security (AISEC)	2013	Uses Support vector machine technique.	Automatically identify Android malware with a detection rate of 89% with 1% false positives, corresponding to one false alarm in 100 installed applications on a smartphone. Adapting the method to other platforms
10	Automatic Analysis of	Konrad Rieck,	published in the Journal of Computer	2011	Uses learning algorithms.	A framework is proposed to

	Malware Behavior using Machine Learning	Philipp Trinius, Carsten Willems, and Thorsten Holz,	Security, IOS Press			overcome the problems of computer security, such as denial-of-service attacks, identity theft, or distribution of spam and phishing contents and enhance the current state-of-the-art.
11	Permission-Based Android Malware Detection	Zarni Aung, Win Zaw	International Jounal Of Scientific & Technology Research Vol 2,Issue 3.	2013	K-Means Algorithm-A machine learning technique	A framework is proposed for classifying Android applications using machine-learning techniques whether they are malware or normal applications

III. PROPOSED METHODOLOGY

We focused on detecting the mobile keylogger using the support Vector Machine(SVM) algorithm, a machine learning technique. Our aim is to prevent the keylogger from stealing the sensitive information. The design of our system consists of three different components, a component for the mobile Applications and its permissions Gathering, Permissions Analyze, and Keyloggers Detector. Bu using these components the applications and permission for each application which are installed in a mobile phone are analyzed and the malicious applications are detected based on the learning model. The overview design of proposed system architecture is shown in Figure 1.

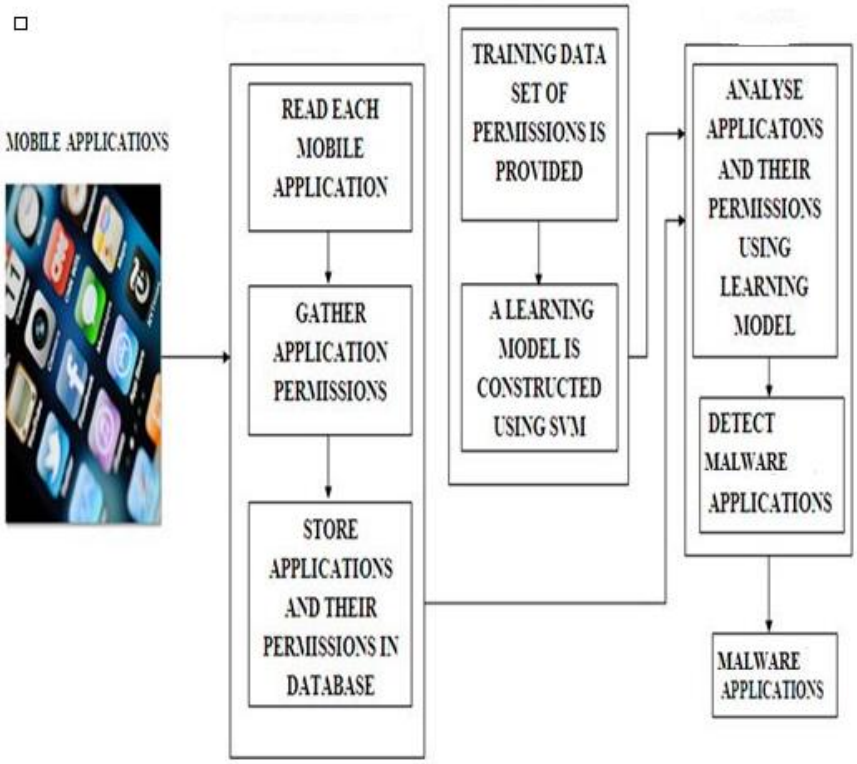

Figure 1

The mobile devices are the targets for malicious applications because of the vulnerable nature of mobile devices than personal computers. Cyber criminals make use of the malware—malicious software to exploit mobile devices such as Smartphones and tablets. The mobile users can email, use on line banking, purchase goods, and use social networking websites. Because of the money transactions through the mobile phones, the attackers are attracted to steal the information and it may used for malicious activities. The most common way for mobile users to get their devices infected with malware is by unknowingly downloading malicious apps. Cyber crooks design their evil little programs to look like authentic games or other useful apps and put them

online on forums and even open app markets, like the Android Market. If the users are not careful when downloading an app, there's a chance the device will turn into a spying tool. By taking these issues into consideration, we expect that mobile devices should have proper mobile security application to detect the keylogger.

The proposed system intends to develop a machine learning-based malware detection system on mobile phones to detect malware applications. This system enhances security and privacy of mobile phone users. It monitors various permission based features and events obtained from the android applications and analyses these features by using machine learning classifiers to classify whether the application is a normal application or malware. In the paper is organized as follows to we review the related literature in Section 2and Section 3 presents an overview of system. Section 4 describes the nature of work. Section 5 describes our implementation of the machine learning algorithm that we use to detect malicious behavior from normal behavior. Section 6 concludes the paper.

IV.NATURE OF WORK

Permission Gathering

The applications and permission for each application are listed using Package Manager API, then they are stored into the sqlite database. Package Manager API is a class for retrieving various kinds of information related to the application packages that are currently installed on the device. This information is stored in a sqllite database. Sqlite database is a relational database management system contained in a small C programming library. It implements a self-contained server less, zero configuration transactional SQL database engine.

Permission analyzer

Permission analyzer uses **Support Vector Machine (SVM)** algorithm to construct a learning model with training data set. The training data set includes permissions and their protection levels. Support Vector Machine (SVM) is a machine learning algorithm which that analyze data and recognize patterns.

Keylogger detector

Keylogger detector analyzes the mobile applications and their permissions using learning model. It detects keylogger applications and prompt users to disable keylogger applications with permissions that can lead to critical security risks. The key advantage of our approach is all types of keyloggers can be detected with in a less computation time.

V. IMPLEMENTATION

Support Vector Machine (SVM)

History and basic concept of SVM

SVM was first introduced in 1992 which is related to statistical learning theory. SVM becomes popular because of its success in handwritten digit recognition. It is now regarded as one of the key area in machine learning. Many commercial analyzers contain SVM. In our approach for machine learning WEKA tool is used.

Properties of SVM

Support Vector Machine (SVM) is Flexible in choosing a similarity function. It provides sparseness of solution when dealing with large data sets.

Strengths and weaknesses of SVM

Training is relatively easy by using SVM. It scales relatively high dimensional data. In The tradeoff between classifier complexity and error can be controlled. Non traditional data like strings and trees can be used as input to SVM. But it has to select a good kernel function.

WEKA

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms written in Java and runs on almost all platforms. It was developed at the University of Waikato in New Zealand. In the algorithms can either be applied directly to a dataset or called from our own Java code. Weka is free software available under GNU General Public License (GPL). So it is possible to analyze how the algorithms work and to modify them.

Main Features of WEKA

It consists of 49 data preprocessing tools and 76 classification/regression algorithms, 8 clustering algorithms, 15 attribute/subset evaluators + 10 search algorithms for feature selection..

VI.CONCLUSION AND FUTURE ENHANCEMENT

In this paper we proposed a system to obtained and analyze the mobile phone applications with their permissions using Support Vector Machine (SVM). With this machine learning algorithm, the system is capable enough to differentiate the normal and malicious applications. Our approach utilize the technique of detecting keyloggers is completely based on behavioral characteristics common to all keyloggers and it does not rely on the internal structure of the keylogger. As the future enhancement the memory usage, control flow & resource usage can be added as the feature vector to identify the keylogger.

REFERENCES

[1] Stefano Ortolani, Cristiano Giuffrida, and Bruno Crispo, "Unprivileged Black-Box Detection of User-Space Keyloggers" IEEE Transactions on dependable and secure computing, Vol.10,No.1, January/february 2013.

[2] Chao Gao and Jiming Liu, "Modeling and Restraining Mobile Virus Propagation "IEEE Transactions on Mobile computing Vol. 12, No. 3, March 2013.

[3] Wei Peng, {Feng Li, Xukai Zou}, and Jie, Wu, " Behavioral Malware Detection in Delay Tolerant Networks" IEEE Transaction on Parallel and Distributed Systems,2013.

[4] Abian Blome, Mart´ın Ochoa, Keqin Li, Michele Peroli, Mohammad Torabi Dashti," VERA: A flexible model-based vulnerability testing tool" IEEE Sixth International Conference on Software Testing, Verification and Validation, March 2013

[5] Suleiman Y. Yerima, Sakir Sezer, Gavin McWilliams Igor Muttik, " A New Android Malware Detection Approach Using Bayesian Classification" 27th International Conference on Advanced Information Networking and Applications, '13 Proceedings of the 2013 IEEE, Pages 121-128, 2013

[6] Takamasa Isohara, Keisuke Takemori and Ayumu Kubota, "Kernel-based Behavior Analysis for Android Malware Detection" Seventh International Conference on Computational Intelligence and Security,IEEE, Page(s):1011 – 1015, 2011

[7] Silvio Cesare, Yang Xiang, Wanlei Zhou, " Malwise—An Effective and Efficient Classification System for Packed and Polymorphic Malware" IEEE Transaction on Computers, Vol. 62, No. 6, June 2013

[8] Yousra Aafer, Wenliang Du, and Heng Yin, " Droid APIMiner: Mining API-Level Features for Robust Malware Detection in Android" 9th International Conference on Security and Privacy in Communication Networks, Sydney, Australia, September 2013.

[9] Hugo Gascon, Fabian Yamaguchi, Daniel Arp,Konrad Rieck, "Structural Detection of Android Malware using Embedded Call Graphs" Proc. of 6th ACM CCS Workshop on Artificial Intelligence and Security (AISEC), Pages 45-54, November 2013

[10] Konrad Rieck, Philipp Trinius, Carsten Willems, and Thorsten Holz, "Automatic Analysis of Malware Behavior using Machine Learning" Published in the Journal of Computer Security, IOS Press, Volume 19 Issue 4, Pages639-668,December2011.

[11] Zarni Aung, Win Za, "Permission-Based Android Malware Detection" International Jounal Of Scientific & Technology Research Vol 2,Issue 3, March 2013.

[12] I. Burguera, U.Z., Nadijm-Tehrani, S.Crowdroid, "Behavior- Based Malware Detection System for Android" In: SPSM'11, ACM ,October 2011

[13] J. Wright, M. E. Dawson Jr. and M. Omar, " Cyber security And Mobile Threats: The Need For Antivirus Applications For Smart Phones" JISTP - Volume 5, Issue 14 (2012), pp. 40-60