UNIVERSITI PUTRA MALAYSIA

A SCORE BASED MALWARE CLASSIFICATION APPROACH FOR MOBILE FORENSIC ANALYSIS

RAMYAA A/P GOBI

FSKTM 2019 40

A SCORE BASED MALWARE CLASSIFICATION APPROACH FOR MOBILE FORENSIC ANALYSIS

By

RAMYAA A/P GOBI

Thesis Submitted to the School of Graduate Studies,

University Putra Malaysia, in Fulfilment of the

Requirements for Degree of Master of Information Security

June 2019

DEDICATIONS

I am dedicating this thesis to my parents who have supported me throughout the research. I am also taking this opportunity to thank my family members and friends for their continuous love and motivation.

Abstract of thesis presented to the Senate of University Putra Malaysia in fulfilment of the requirement for the degree of Master of Information Security

A SCORE BASED MALWARE CLASSIFICATION APPROACH FOR MOBILE FORENSIC ANALYSIS

By

RAMYAA A/P GOBI

June 2019

Chair: Assoc. Prof. Dr. Zurina Mohd Hanapi

Faculty: Faculty of Computer Science and Information Technology

The rapid growth of Android as one of the leading Operating System (OS) for mobile devices drives the need of effective security measures to ensure the users have a safer platform to use. Boolean based features used for application permissions degrades the precision, recall, F-1 score and accuracy of malware detection. The reason for this is that Boolean based features classify the benign and malware applications based on true or false rule which is done based on the binary 0 for benign and 1 for malware. FAMOUS (Forensic Analysis of MObile devices Using Scoring of application permissions) which incorporates Effective Maliciousness Score of Permission (EMSP), a score based representation for permissions which replaces the Boolean representation for permissions have produced better result for the accuracy, precision,

recall and F1-score over the Boolean based feature from existing works. FAMOUS is tested on the crawled datasets that are collected from multiple public archives such as Cantagio dump, AndroMalShare, Derbin project and Andrototal. This crawled datasets are then labelled by the result captured from Virus Total engines. Thus, FAMOUS did not use any standard dataset for its analysis. In his research, we will implement the EMSP, a score based triage and test it over Android Malware Dataset (AMD) and Android PRAGuard dataset to ensure reliable result obtained for the Accuracy, Precision, Recall and F1-Score through Machine Learning classifiers. Total of five classifiers have been used to train and test the datasets which consist of Random Forest, Decision Tree, Naive Bayes, K-nearest neighbours, and Support Vector Machine. EMSP will be implemented using Python programming language on Windows system. The performance metrics evaluated for the research are precision, recall, F-1 score and accuracy. The accuracy obtained varies for different classifiers for AMD and Android PRAGuard dataset. The best result obtained for Random Forest classifier when using AMD.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia

Sebagai memenuhi keperluan untuk ijazah Sarjana Keselamatan Maklumat

A SCORE BASED MALWARE CLASSIFICATION APPROACH FOR MOBILE FORENSIC ANALYSIS

Oleh

RAMYAA A/P GOBI

Jun 2019

Pengerusi: Assoc. Prof. Dr. Zurina Mohd Hanapi

Fakulti: Fakulti Sains Komputer Dan Teknologi Maklumat

Pertumbuhan pesat Android sebagai salah satu Sistem Operasi yang terkemuka untuk peranti mudah alih memacu keperluan langkah keselamatan yang berkesan untuk memastikan pengguna mempunyai platform yang lebih selamat untuk digunakan. Ciri berasaskan *Boolean* yang digunakan untuk kebenaran aplikasi merendahkan ketelitian, keimbasan, skor F-1 dan ketepatan pengesanan malware. Sebabnya ialah ciri berdasarkan *Boolean* mengklasifikasikan aplikasi baik dan aplikasi negatif berasaskan peraturan benar atau palsu yang dilakukan berdasarkan binari 0 untuk aplikasi baik dan 1 untuk aplikasi negatif. (*Forensic Analysis of MObile devices Using Scoring of application permissions*) yang menggabungkan *Effective Maliciousness Score of Permission (EMSP),* perwakilan berdasarkan nilai untuk kebenaran yang

menggantikan kerja yang sedia ada yang menggunakan perwakilan Boolean untuk mendapatkan kebenaran telah menghasilkan hasil yang lebih baik untuk terperinci, keimbasan, skor F-1 dan ketepatan pengesanan malware ke atas ciri berasaskan Boolean daripada kerja sedia ada. *FAMOUS* diuji pada kumpulan data yang dikumpulkan yang dikumpulkan dari pelbagai arkib awam seperti Cantagio dump, AndroMalShare, Derbin dan Andrototal. Data-data yang dikemas kini kemudian dilabelkan dengan hasil yang ditangkap daripada enjin Virus Total. Oleh itu, FAMOUS tidak menggunakan sebarang dataset standard untuk analisisnya. Dalam kajian ini, kami akan melaksanakan *EMSP*, *triage* berasaskan skor dan mengujinya melalui dataset Android Malware Dataset (AMD) dan Android PRAGuard untuk memastikan hasil yang boleh dipercayai untuk ketelitian, keimbasan, skor F-1 dan ketepatan melalui pengeluar *Machine Learning*. Jumlah 5 pengelasan telah digunakan untuk melatih dan menguji dataset yang terdiri daripada *Random Forest, Decision Tree, Naive Bayes, K-nearest neighbours*, dan *Support Vector Machine*. *EMSP* akan dilaksanakan menggunakan bahasa pengaturcaraan Python pada sistem Windows. Metrik prestasi yang dinilai untuk penyelidikan adalah ketelitian, keimbasan, skor F-1 dan ketepatan. Ketepatan yang diperolehi berbeza untuk klasifikasi yang berbeza untuk dataset AMD dan Android PRAGuard. Hasil terbaik yang diperoleh untuk pengelas *Random Forest* apabila menggunakan AMD.

ACKNOWLEDGEMENTS

Primarily, I would like to thank my supervisor, Associate Professor Dr Zurina Mohd Hanapi for her guidance and support in helping me to complete my research as well as to prepare this thesis. She have always been a great support by providing proper guidance throughout the period of this research.

Secondly, I would also like to thank my parents for their continuous love and trust. They have been a great pillar of support to me since day one of my Master journey. Special thanks to my father for supporting me financially to help me achieve my dreams.

I would also like to thank my friends for being there in time of need. Learning together with fantastic course mates make a lot of difference. I am thanking each and every one of them for being keen to share their knowledge and expertise with me.

Last but not least, thank you to all FSKTM lecturers and staffs for their support and care to all FSKTM student. Thank you to them for providing a comfortable, nice, efficient environment and facilities for students like me so that we can focus on our study.

APPROVAL

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Information Security. The members of the Supervisory Committee were as follows:

Signature:_____

Zurina Mohd. Hanapi, PhD

Associate Professor

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

(Supervisor)

Date:_____

DECLARATION

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustration and citation have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, report, lecturer notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012.

Signature: _____ Date: _____

Name and Matric No.: **Ramyaa A/P Gobi GS49776**

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BSP	Benign Score of Permission
DT	Decision Tree
FES	Feature Extraction and Scoring
EMSP	Effective Maliciousness Score of Permission
kNN	k-Nearest Neighbors
MSP	Malicious Score of Permission
NB	Naive Bayes
RF	Random Forest
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Background

The popularity of smartphones have grown exponentially in the past years (Sanz et al., 2013). The drive to this popularity is the technology change which has bring all the necessary facilities in a mobile phone in the form of applications. Smartphones are one of the powerful small computers that have the capability of accompanying us everywhere.

In order to utilize the possibilities that smartphones offers, smartphone users are required to install applications. There are tons of mobile applications, which can be categorized from basic applications such as calculator to critical applications such as banking applications. Smartphones are the most targeted for cyber-attack due to the sensitivity of the data stored and transmitted over.

Among the operating system available for smartphones, Android has the exponential growth which holds around 85% of the market share based on the International Data Corporation's forecast report (Kumar et al., 2018).

One of the major factor for the tremendous growth of Android as an operating system is that it comes with the capability of installing applications from many application markets to broaden the features that is already available in Android smartphone. The default application market for Android is widely known as Google Play Store while

there are many third party markets such as Samsung Galaxy app, Slide MF, Amazon App Store and many more that increases the usage and popularity of Android operating system for smartphones. Along with the popularity, the level of threat has also increased over the years. One of the top most treat is installation of malicious applications (Zhou et al., 2012).

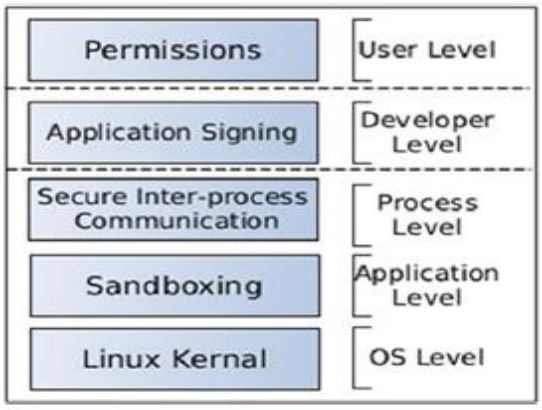

Figure 1.1 Android multi-layered security measures
(Source: Kumar et al., 2018)

Android platform is plot with many different layers of security measures as shown in Figure 1.1 Among the other layers in security mechanisms as shown in Figure 1.1, the weakest link is known to be the user level which has the outline of the application permission. Each activity that is carried out by the user on their smartphone requires permissions during the installation of an application. Apparently, the smartphone users are not aware of the risk in granting permission to proceed with the application installation without knowing the malicious intention of the application. Attackers takes advantage of this weak spot to intrude on the users' device intentionally. Not only Android applications have this threat but even iOS applications are also being targeted.

Detection of Android malicious application is one of the active research area that is being researched by many researches all around the world with approaches starts from

pattern matching up to machine learning (Faruki et al., 2013; Zheng et al., 2013; Sato et al., 2013; Huang et al., 2013; Sanz et al., 2013; Wu et al., 2012; Alam et al., 2013; Chan et al., 2014; Chuang et al., 2015). The tremendous growth of malware attacks on mobile created a space where mobile devices have been used for detecting the malwares with the use of machine learning (Shabtai et al., 2010). Ensemble of multiple classifiers used to categorize benign apps has been studied in a recent work (Wang et al., 2017).

Android classifier feature set can be categorised into three categories known as static, dynamic and hybrid respectively. To extract the static feature, it is not necessary to execute the application. Example of static feature are permission, intent, API calls ad meta-data. Meanwhile, to extract dynamic feature set, the applications are required to be executed and the features extracted come from the log activities which triggers when the application is executed. Example of dynamic feature set are network activity, memory analysis, OS interaction, and process trace. Hybrid feature set are combination of both static and dynamic features.

The existing works have used static feature set analysis, dynamic feature set analysis and hybrid feature set analysis. Most of research work discussed in Chapter 2, have used machine learning and deep learning techniques to obtain the accuracy, precision, recall and the F1-score for android permission analysis.

This research adopts the static feature set focusing on application permission to construct a machine learning based model to detect malicious applications.

1.2 Problem Statement

Forensic Analysis of MObile devices Using Scoring of application permissions (FAMOUS) which incorporates Effective Maliciousness Score of Permission (EMSP), a score based representation for permissions that replaces the existing work which uses Boolean representation for permissions. FAMOUS use crawled collection of dataset that was trained and tested on Machine Learning classifiers. However, standard dataset have not been tested on FAMOUS (i.e. Android Malware Dataset (AMD) & Android PRAGuard Dataset). The impact of using crawled dataset is that the dataset is labelled from the result captured from Virus Total engine. The rule set for the labelling is that the dataset is labelled as malware when at least one engine detects the dataset as malicious. Virus Total is a tool that operates from human intelligence where the community of the Virus Total vote the dataset as malware or benign thus the result obtained from crawled dataset is still arguable.

1.3 Objective

The objective of this research is to implement the EMSP score based triage over Android Malware Dataset (AMD) and Android PRAGuard dataset to ensure reliable result obtained for the Accuracy, Precision, Recall and F1-Score.

1.4 Scope

The scope of this research is limited to Android Operating System which runs on smartphones. This project is tested on AMD and Android PRAGuard Dataset. The development of the proposed method is programed on Windows system using Python programming language.

1.5 Thesis Structure

The rest of this thesis is organized as follows; Chapter Two presents the literature review explained the background of crawled dataset and standard dataset. Chapter Three elaborates the methodology of this research, which consist of the EMSP architecture, the experimentation and tools used to implement this research work. Chapter Four describes the details on the results obtained from the conducted experiments. The results focuses on the accuracy, precision, recall and F1-score obtained from the machine learning classifiers. Finally, Chapter Five conclude the overall research work.

REFERENCES

Alam, M.S. and Vuong, S.T. 2013. Random forest classification for detecting android Malware from *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, IEEE* 663–669.

Arshad, S., Shah, M., Wahid, A., Mehmood, A., Song, H., & Yu, H. 2018. SAMADroid: A Novel 3-Level Hybrid Malware Detection Model for Android Operating System from *IEEE Access* 6: 4321-4339.

Chan, P.P. and Song, W.-K. 2014. Static detection of android malware by using permissions and api calls from *2014 International Conference on Machine Learning and Cybernetics, Vol. 1, IEEE* 82–87.

Clemens, J. 2015. Automatic classification of object code using machine learning from *Digit. Investig* 14 S156–S162.

Dini, G., Martinelli, F., Matteucci, I., Petrocchi, M., Saracino, A., & Sgandurra, D. 2018. Risk analysis of Android applications: A user-centric solution from *Future Generation Computer Systems* 80: 505-518.

Fang, Z., Han, W. and Li, Y. 2014. Permission based android security: Issues and countermeasures from *Comput. Secur* 43 205–218.

Faruki, P., Ganmoor, V., Laxmi, V., Gaur, M.S. and Bharmal, A. 2013. AndroSimilar: robust statistical feature signature for android malware detection from *Proceedings of the 6th International Conference on Security of Information and Networks, ACM* 152–159.

Felt, A.P., Chin, E., Hanna, S., Song, D. and Wagner, D. 2011. Android permissions

demystified from *Proceedings of the 18th ACM Conference on Computer and Communications Security, ACM* 627–638.

Geneiatakis, D., Fovino, I.N., Kounelis, I. and Stirparo, P. 2015. A Permission verification approach for android mobile applications from *Comput. Secur* 49 192–205.

Huang, C.-Y., Tsai, Y.-T. and Hsu, C.-H. 2013. Performance evaluation on permission based detection for android malware from Advances in Intelligent Systems and Applications from *Springer* 2:111–120.

Idrees, F., Rajarajan, M., Conti, M., Chen, T.M. and Rahulamathavan, Y. 2017. Pindroid: A novel android malware detection system using ensemble learning methods from *Comput. Secur* 68 36–46.

Kim, T., Kang, B., Rho, M., Sezer, S., & Im, E. 2019. A Multimodal Deep Learning Method for Android Malware Detection Using Various Features *from IEEE Transactions On Information Forensics And Security* 773-788.

Kumar, A., Kuppusamy, K.S. and Aghila, G. 2018. FAMOUS: Forensic Analysis of MObile devices Using Scoring of application permissions from *Future Generation Computer Systems* 83 158–172.

Maiorca, D., Ariu, D., Corona, I., Aresu, M., & Giacinto, G. 2015. Stealth attacks: An extended insight into the obfuscation effects on Android malware from *Computers & Security* 51: 16-31.

Marturana, F. and Tacconi, S. 2013. A machine learning-based triage methodology for automated categorization of digital media from *Digit. Investig*. 10 (2) 193–204.

Milosevic, N., Dehghantanha, A. and Choo, K.K.R. 2017. Machine learning aided Android malware classification from *Comput. Electr. Eng*.

Moonsamy, V., Rong, J. and Liu, S. 2014. Mining permission patterns for contrasting

clean and malicious android applications from *Future Gener. Comput. Syst*. 36 122–132.

Rehman, Z., Khan, S., Muhammad, K., Lee, J., Lv, Z., & Baik, S. et al. 2018. Machine learning-assisted signature and heuristic-based detection of malwares in Android devices from *Computers & Electrical Engin*eering 69: 828-841.

Samra, A.A.A., Yim, K. and Ghanem, O.A. 2013. Analysis of clustering technique in android malware detection from *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, (IMIS), IEEE* 729–733.

Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Bringas, P.G. and Álvarez, G. 2013. Mama: manifest analysis for malware detection in android from *Cybern. Syst* 469–488.

Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Bringas, P.G. and Álvarez, G. 2013. Puma: Permission usage to detect malware in android from *International Joint Conference CISIS12-ICEUTE´ 12-SOCO´ 12 Special Sessions, Springer* 289–298.

Sato, R., Chiba, D. and Goto, S. 2013. Detecting android malware by analyzing manifest Files from *Proceedings of the Asia-Pacific Advanced Network* 36: 23–31.

Shabtai, A., Fledel, Y. and Elovici, Y. 2010. Automated static code analysis for classifying android applications using machine learning from *2010 International Conference on Computational Intelligence and Security, (CIS), IEEE* 329–333.

Sharma, K., & Gupta, B. 2018. Mitigation and risk factor analysis of android applications from *Computers & Electrical Engineering* 71: 416-430.

Talha, K.A., Alper, D.I. and Aydin, C. 2015. Apk auditor: Permission-based android malware detection system from *Digit. Investig*. 13 1–14.

Wang, W., Li, Y., Wang, X., Liu, J. and Zhang, X. 2017. Detecting android malicious apps and categorizing benign apps with ensemble of classifiers *from Future Gener. Comput. Syst.*

Wang, X., Wang, W., He, Y., Liu, J., Han, Z. and Zhang, X. 2017. Characterizing android apps behavior for effective detection of malapps at large scale from *Future Gener. Comput. Syst.* 75 30–45.

Wu, D.-J., Mao, C.-H., Wei, T.-E., Lee, H.-M. and Wu, K.-P. 2012. Droidmat: Android malware detection through manifest and api calls tracing from *2012 Seventh Asia Joint Conference on Information Security, (Asia JCIS), IEEE* 62–69.

Yerima, S.Y., Sezer, S. , McWilliams, G. and Muttik, I. 2013. A new android malware detection approach using bayesian classification from *2013 IEEE 27th International Conference on Advanced Information Networking and Applications, (AINA), IEEE* 121–128.

Yousefi-Azar, M., Hamey, L., Varadharajan, V., & Chen, S. 2018. Malytics: A Malware Detection Scheme from *IEEE Access* 6: 49418-49431.

Zheng, M., Sun, M. and Lui, J.C. 2013. Droid analytics: a signature based analytic system to collect, extract, analyze and associate android malware from *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, IEEE* 163–171.

Zhou, Y. & Jiang, X. 2012. Dissecting android malware: Characterization and evolution from *2012 IEEE Symposium on Security and Privacy, (SP), IEEE* 95–109.