Automatic Feature Extraction, Categorization and Detection of Malicious Code in Android Applications

Muhammad Zuhair Qadir  
Department of Computer Sciences  
School of Science and Engineering  
Lahore University of Management and Sciences  
Lahore, Pakistan  
mzuhairqadir@gmail.com

Atif Nisar Jilani  
Department of Computer Sciences  
School of Science and Engineering  
Lahore University of Management and Sciences  
Lahore, Pakistan  
jilani.ansar@gmail.com

Hassam Ullah Sheikh  
School of Computer Science  
University of Manchester  
Manchester, UK  
sheikhh@cs.man.ac.uk

Abstract—Since Android has become a popular software platform for mobile devices recently; they offer almost the same functionality as personal computers. Malwares have also become a big concern. As the number of new Android applications tends to be rapidly increased in the near future, there is a need for automatic malware detection quickly and efficiently. In this paper, we define a simple static analysis approach to first extract the features of the Android application based on intents and categories the application into a known major category and later on mapping it with the permissions requested by the application and also comparing it with the most obvious intents of category. As a result, getting to know which apps are using features which they are not supposed to use or they don’t need.

Keywords — Feature Extraction, Malicious code, Android, Categorization, Automated approach.

I. INTRODUCTION

Android has seen a tremendous growth since its introduction in the market. The numbers of applications are growing with extraordinary volume. It has a wide variety of applications targeting various needs and use cases. Since its launch the privilege to the developers who can upload any application draws attention to the malicious developers whose intents are to deceive and take advantage of this privilege and without much inspection it is getting hard to investigate a valid and benign application which can filter out the malicious application from the pool of millions of apps. Malicious code is being injected into mobile applications and threatens the privacy of user’s personal data and device integrity also it can lead to breaches of user data and violate application security policies. In 2011, malware attacks are increased by 155 percent across all platforms [1] in particular, Android is the platform with the highest malware growth rate by the end of 2011.

Static detection techniques (also called signature matching) have high detection rates and consume fewer resources [8]. Static analysis of these applications can result in faster detection of malicious apps and as it involves automatic application code lookup and detecting required content without running it or testing it. Static analysis of Android applications is important because quality and reliability are keys to success on the Android market [2].

Because of diversity of applications available in different markets, there is a need to categories the application first.

We perform static analysis on Android applications in order to determine the feature set of the application based on its functionality. Our program also looks for the category in which the app should fall.

II. RELATED WORK

There is lot of study on dynamic and static approaches for android malware detection analysis. Static being faster as it does not require runtime cost and exploiting the code on runtime.

Confidentiality and authorization are two key goals which are addressed in program analysis techniques. SCanDroid extracts security specifications from the manifest of an app and checks whether data flows through the app are consistent with the stated specifications [3].

Table 4 [4] gives the crisp comparison of the tools being made to address various malicious code detection and analysis with various techniques.

They are various tools written on static analysis among them the most famous one used in AndroGuard [5] it’s an open source project to statically detect Android malwares, it basically reverse engineer APK from byte code (assembly source code) to readable format and afterward it visualizes
your application with Gephi which is a control flow graph of the method. Check if an android application is present in a database or not, it maintains the list of pre-defined malware's. Also it compares the control flow graph of both the application to check the similarity between that malware and the given application.

There are various papers which works on the permission analysis approach, one [7] of them detects whether an app is

| Solution     | Aim                          | Flow Analysis         | Classification Policy                                      | Evaluation Scale |
|--------------|------------------------------|-----------------------|-----------------------------------------------------------|------------------|
| SCanDroid    | Enforcement of confidentiality, integrity | Data, string          | Constraints on permission logics                          | N/A              |
| CHEX         | Discovery of exposed component API | Data                 | Component exported to public without restrictions        | 5,486 apps       |
| RiskRanker   | Detection of abnormal code/behavior patterns | Data, control        | Multiple malware behavior signatures                      | 118, 318 apps    |
| Woodpecker   | Firmware permission          | Data, control         | N/A                                                       | 8 phone images, 13 permissions |
| AndroidLeaks | Confidentiality              | Data                  | Sensitive data used by risky APIs                        | 24,350 apps      |
| SCANDAL      | Confidentiality              | Data                  | Sensitive data used by risky APIs                        | 90 apps & 8 malware |
| Stowaway     | Detection of overprivileged apps | String, Intent control flow | Compare required and requested permissions               | 940 apps         |
| ComDroid     | Detection of apps communication vulnerabilities | Intent control flow   | Implicit Intent with weak or no permission                 | 100 apps         |
| PiOS         | Confidentiality              | Data                  | Sensitive data used by risky APIs                        | 1,407 apps       |
| UID          | Identification of unauthorized calls | Data, event specific control | Trigger-operation dependence for privileged function calls | 708 apps & 482 malware |

TABLE I. Comparison

System API calls and performs dynamic detection using machine learning on android market.

Basically it does static analysis on android application output of the readelf tool which extracts their system calls and then they are compared with the pre-defined list of malicious applications from benign ones based on the combinations of system calls used in the executable [8].

III. APPROACH

We propose and implemented an approach to detect malicious applications statically. Android applications can interact with other applications, and with the system, through a well-defined API. A number of components can make up an application. In particular, Android defines activities, services, content providers, and broadcast receivers.

Activities, services, and broadcast receivers are activated by intents, i.e., asynchronous messages exchanged between individual components to request an action. Activity and service intents specify actions to be performed. Conversely, broadcast receiver intents define the received event and are delivered to the interested broadcast receivers. Our algorithm consists of three main steps.

Step I: We get the set of APKs to analyze and transform to decompressed files and then into byte code using APKTool [6].

Step II: Then we extract features of the application present in the byte code.
Step III: Categorize the application to known major categories.  
Step IV: Relate features to category and point out features which are not needed for the application and can be considered as malicious.

In this four step process, APKTool does the conversion part to get the readable format code of the original APKs. While in the second step, feature extraction or more appropriately, code tagging is done through our custom tool written in C++ which parse the code and gets all the intents of the code and tags it with the detected features which will be used later.

Intents provide an easy way to detect what an application actually trying to do and how it is utilizing to the resources of the device. This step is crucial and essential part of the algorithm making an intelligent system to separate benign apps and malicious apps. Our tool is flexible as takes a file full of intents which should be extracted and tagged in the code. Making a useful plugin for various alterations of this algorithm and also targeting specific type of malicious codes which exploits certain features only.

In the last phase, we use tagged code and categorized APK to achieve this categorization.

### IV. IMPLEMENTATION

We used APK tools for extracting code out of APK into smali code [10] which is actually sort of byte code along with Dalvik Opcodes [11] and having its own syntax. APK tools is a very powerful tool available at android’s developer website for various apk reverse engineering tasks. We wrote a batch script which takes the application name , and after wards with the help of APK tool, it extracts all the folder that are compressed within APK into Dalvik code which is not human readable for that purpose batch in the next step converts it into samli code which is byte level human readable code.

Custom C++ tool was written to tag features parsing whole bunch of smali files obtained from previous step. In the tool, a file is passed as parameter having all features to be tagged with the code line number and count of that feature in a specific smali byte code file.

For categorization we wrote another custom tool in C++ to categorize the application, we analyze the group of intents and manifest file for permissions to get most of the intents and features of the app helping us to understand the type of the application whether it is a game, utility, image tool, sound recorder etc. An isolated and unprivileged application has very limited functionality. Therefore, smartphone platforms allow access to individual sensitive resources (e.g., address book, GPS) using permissions. Permission is a form of capability. In order to categorize we have maintained a list of major permissions by analyzing various applications before writing our tool. On the basis of these permissions, we compare the permission presented in manifest file with our list, and also we have set of rules that if an application has access to certain set of permissions then we can say that it may belongs to a certain category as we mentioned in our Table III. If the application is from Google’s application market i.e. play then we also take into consideration the category [23] assigned in the Google app store.

| No | List of Some Targeted Intents      | Set of Some Permissions                                                                 |
|----|------------------------------------|----------------------------------------------------------------------------------------|
| 1  | android.hardware.Camera.PictureCallback | android.permission.WRITE_SMS  
anroid.permission.SEND_SMS  
anroid.permission.CALL_PHONE  
anroid.permission.READ_SMS |
| 2  | android.telephony.SmsMessage       | android.permissionINTERNET  
anroid.permission.READ_PHONE_STATE                                                          |
| 3  | android.telephony.SmsManager       | android.permissiongroup.LOCATION  
anroid.permission.READ_CONTACTS  
anroid.permission.READ_SOCIAL_STREAM  
anroid.permissiongroup.ACCOUNTS  
anroid.permissionINTERNET |
| 4  | android.telephony.CellLocation     | android.permissiongroup.BATTERY_STATS  
anroid.permissiongroup.SYSTEM_TOOLS  
anroid.permissiongroup.BLUETOOTH_ADMIN  
anroid.permissiongroup.KILL_BACKGROUND_PROCESSES |
| 5  | android.location.LocationManager   | android.permissiongroup.STORAGE  
anroid.permissiongroup.READ_EXTERNAL_STORAGES  
anroid.permissiongroup.CAMERA  
anroid.permissiongroup.RECORD_AUDIO  
anroid.permissiongroup.MODIFY_AUDIO_SETTINGS  
anroid.permissiongroup.BATTERY_STATS  
anroid.permissiongroup.INTENT  
anroid.permissiongroup.READ_SMS  
anroid.permissiongroup.WRITE_SMS  
anroid.permissiongroup.SEND_SMS  
anroid.permissiongroup.CALL_PHONE  
anroid.permissiongroup.READ_SMS |
| 6  | android.permissiongroup.INTERNAL  
anroid.permissiongroup.STORAGE  
anroid.permissiongroup.READ_EXTERNAL_STORAGES  
anroid.permissiongroup.CAMERA  
anroid.permissiongroup.RECORD_AUDIO  
anroid.permissiongroup.MODIFY_AUDIO_SETTINGS  
anroid.permissiongroup.BATTERY_STATS  
anroid.permissiongroup.INTENT  
anroid.permissiongroup.READ_SMS  
anroid.permissiongroup.WRITE_SMS  
anroid.permissiongroup.SEND_SMS  
anroid.permissiongroup.CALL_PHONE  
anroid.permissiongroup.READ_SMS |
| 7  | android.appwidget.action.APPWIDGET_UPDATES  
anroid.appwidget.action.APPWIDGET_CONFIGURATIONS  
anroid.appwidget.action.APPWIDGET_DRAWABLE |
| 8  | android.permissiongroup.LOCATION  
anroid.permissiongroup.INTENT |

**TABLE III. Categories**

In the last phase, we use tagged code and categorized APK and point out the features malicious for the application and as a result classifying whether the concerned APK should be considered as malicious or not.
V. FUTURE WORK

We had another approach still to implement to improve the results, in which we will use machine learning approach to further classify the detected malicious to pinpoint the category of malware. So two level filtering will further decrease the false positives and give more accurate results.

Basically the behavior of smart phones is rather protected by the user permissions, also there are numerous permission-protected [20] method calls that are not part of the public Android API, but are in classes that are resident on the phone, we will then examine all the obtained smali files to find method calls used by application and each method call is then compared to the list of all method calls that we have in our list of permission-protected Android API calls to build an association. That association set is then compared to the permission set that is declared in the application’s AndroidManifest.xml file, in this way we can determine whether the application has extra permissions, lacks permissions, or has exactly the permission set that it requires based on its functionality.

Furthermore, to make more refine categorization of applications we aim to implement the scheme as implemented in LACTA [14]. LACTA finds certain keywords and does code analysis of the application and based on learning certain keywords and function names it categorizes the application. This will add two level filter on categorization step of our algorithm making it fine-tuned and more effective.

VI. CONCLUSION

We have devised a simple approach for automatic static analysis which is capable of source code tagging with its prominent features and application categorization which helps in identifying irrelevant features which should not be present in the app. It is quick and efficient and relies on intents present in the source code. It can be made more efficient using machine learning techniques to train on apps first then predicting the malicious code snippets but it has a tradeoff with performance.

We have also proposed additional filters in the future work that could be added to our approach to increase the viability and accuracy of our system.

REFERENCES

[1] Juniper Networks: 2011 Mobile Threats Report (February 2012)
[2] Static analysis of Android programs
É Payet, F Spoto - Automated Deduction–CADE-23, 2011 – Springer
[3] SCanDroid: Automated security certification of Android applications
AP Fuchs, A Chaudhuri, JS Foster - Univ. of Maryland, 2009 – Citeseer
[4] A Static Assurance Analysis of Android Applications
KO Elish, DD Yao, BG Ryder, X Jiang - people.cs.vt.edu
[5] APKTool for reverse engineering android applications
https://code.google.com/p/android-apktool/
[6] Android permissions demystified
AP Felt, E Chin, S Hanna, D Song, 2011 - dl.acm.org
[7] Androguard - Reverse engineering, Malware and goodware
code.google.com/p/androguard/
[8] T.K. Ho, “Random Decision Forest,” Proc. International Conference on Document Analysis and Recognition ,1995, pp. 278-282
[9] Schmidt, A.D., Bye, R., Schmidt, H.G., Clausen, J., Kiraz, O., Yuksel, K.A., Camtepe, S.A., Albayrak, S.: Static analysis of executables for collaborative malware detection on android. In: Proceedings of IEEE International Conference on Communications.(2009).
[10] Smali code
https://code.google.com/p/smali/
[11] DalvikOpcodes
http://pallergabor.uw.hu/androidblog/dalvik_opcodes.html
[12] I. Burguera, U.Z., Nadjim-Tehrani, S.: Crowdroid: Behavior- Based Malware Detection System for Android. In: LACTA[14], ACM(October 2011)
[13] Permission-Based Android Malware Detection
[14] LACTA: An Enhanced Automatic Software Categorization on the Native Code of Android Applications
[15] Google Inc. The Android Developer's Guide.
http://developer.android.com/guide/, 2011.
[16] The Android Open Source Project. Bytecode for the Dalvik VM.
http://source.android.com/tech/dalvik/dalvik-bytecode.html, 2007.
[17] Android Developers Reference.
http://developer.android.com/reference/.
[18] The dex2jarproject.Available: http://code.google.com/p/dex2jar/
[19] A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, Y. Weiss: Andromaly: a behavioral malware detection framework for android devices. Journal of Intelligent Information Systems 38(1) (January 2011) 161-190
[20] Analysis of Android Applications' Permissions
R Johnson, Z Wang, C Gagnon... - Software Security and …, 2012 - ieexplorer.ieee.org
[21] A framework for static detection of privacy leaks in android applications C Mann, A Starostin - Proceedings of the 27th Annual ACM Symposium ..., 2012 - dl.acm.org
[22] A Malicious Application Detection Framework using Automatic Feature Extraction Tool on Android Market
D Kim, J Kim, S Kim - pscentre.org 2012
[23] On the automatic categorization of android applications
B Sanz, I Santos, C Laorden… - … (CCNC), 2012 IEEE, 2012 - ieexplorer.ieee.org
[24] Exploiting common Intent vulnerabilities in Android Applications
[25] A System Call-Centric Analysis and Stimulation Technique to Automatically Reconstruct Android Malware Behaviors
A Reina, A Fattori, I Cavallaro - 2013 - security.
