**Machine Learning Approach for Android Malware Detection**

Materials of the Bauman MSTU Postgraduate Conference

External postgraduate student at the Department of Information Security

Andrey Artyushkin

Moscow, 2017

## Introduction

In 2012, Novosibirsk Sobolev Institute for Mathematics published the article *On the History of Cryptography in Russia*. Natalia Tokareva, the author, highlights some key events related to Russian cryptography since the loss in the Cold War. This paper covers the establishment of Information Security and Cryptography programs at civil universities as well as the foundation of Kaspersky Lab [1]. I had the chance to achieve a diploma in this specialty at Bauman MSTU and work for several years as a malware analyst at Kaspersky Lab. My name is Andrey Artyushkin. Now, I continue my career as a malware analyst at Positive Technologies. My talk will focus on the general aspects of malware detection and defense strategies. I will also present some results of my recent research on infected Android applications.

## History of cryptography in Russia

The history of Information Security and information attacks reaches back to ancient times. By the 13th century, Russians were using several different successive types of cryptography [2]. There is also evidence of cryptography being used by Sergius of Radonezh and Russian princes in preparation for a Holy War against the Mongol-Tatar [2].

Cryptographic activities of the Ministry of Foreign Affairs headed by K.V. Nesselrode played an important role in the outcome of the 1812 Patriotic War and Napoleon's setback in Russia [1].

"We were very much helped by the fact that we always knew the intentions of your emperor from his own dispatch ...

"I find it very strange that you managed to read it. Someone probably gave you a key.

"Far from it! I give you my word of honor that none of this happened. We have just successfully decrypted the message." From the conversation that took place after the war between Alexander I and Marshal McDonald of Napoleon's army [1].

During the World War I, Russian army was caught in the German trap due to some reasons. One of the main reasons was unsatisfactory level of cryptographic service.

"In August 1914, during the East-Prussian operation, two Russian armies (headed by Samsonov and Rennenkampf), performing until the end of the mobilization, had to pull back a part of the German forces, thereby disrupting Germany's attack against France. However, another scenario took place. It turned out that the army of P. K. Rennenkampf had already received a new cipher, and the old one had been destroyed, while Samsonov's army were still using the old cipher. Therefore, radio communication between these two armies were conducted in cleartext. Germans took advantage of this situation. Moreover, Samsonov’s army did not possess any stockpiled telegraph wire. This forced the command headquarters and reconnaissance units to use even phones of local residents for communication. At the same time, the orders on the timely withdrawal of armies to certain milestones (sent by the commander of the front) simply did not reach Samsonov. His army was surrounded and fought heroically, left without any support. Rennenkampf’s army should have come for rescue, but did not come: according to historians, it was an actual betrayal. As a result, the army of Samsonov was destroyed. Losses amounted to tens of thousands of killed, wounded, and prisoned ..." [3]

"Russian army used their devices so carelessly, as if they did not expect that the same receivers are at the disposal of Austrians, who could easily tune to the appropriate wavelength. Austrians were more provident when using their radio facilities, mainly, for eavesdropping. Sometimes, they succeeded in deciphering by guessing, while other situations required direct requests on the radio during the broadcast. Russians willingly helped the 'insiders' whom they regarded as their colleagues." M. Ronge, the Chief of the Austrian Intelligence Bureau [1].

Soviet cryptographic service has achieved considerable success during the Great Patriotic War and in preparation for it. The equipment for secret telephone communication, radio communication, and text encoders were created and put into operation. Vladimir Aleksandrovich Kotelnikov, being an outstanding scientist and Moscow Bauman State Technical University graduate, played the key role in the organization of a reliable secret communication of the highest level during the Great Patriotic War. Germans failed to decipher messages encrypted via Soviet equipment. "Radio reconnaissance played a special role at the very beginning of the war, but later it did not lose its significance. However, we have never been able to intercept and decipher radiograms sent by your General Headquarters, front staff and armies." Chief of the Supreme High Command of the German Armed Forces, Lieutenant-General A. Jodl. [1]

"Good performance of cryptographers helped to win numerous battles." Deputy Supreme Commander G.K. Zhukov. [4]

"This secret communication line was, as the saying goes, was sent-from-heaven. It helped us so many times, was so stable under the most difficult conditions, that we should pay tribute to our equipment and our signalmen who followed close behind every man using this communication." Marshal I. S. Konev.

It has been decades since then. Nowadays, issues related to cryptography and information security affect everyone.

## The current state of the information security world

We have entered the age of total computerization. Information technology is used in almost every sphere of life. People use computers at both home and workplace. Furthermore, almost everyone has a smartphone in their pocket these days. We entrust control of our personal and work-related data to computers, communicate through social networks and email, use online banking, and make purchases via the internet. Everyone is dependent on information technology to run business. The larger the business, the stronger this dependency.

This mass transfer to electronic format has encouraged a rise in cybercriminal activities. At the dawn of computer era, attackers committed pranks and small crimes at worst; today, extortion, theft, and fraud are extremely widespread. Mentions of serious [5] crimes such as sabotage and espionage started appearing rather recently. The first officially registered act of malware sabotage was an attempt to disrupt a uranium enrichment plant in Iran [6]. By the way, this incident was discovered by Sergey Ulasen, a Belarusian security researcher [6]. As the internet of things becomes a reality and industrial processes are computerized, we can expect new attack vectors to arise. There was no well-known computing platform invulnerable to malware thus far.

In general, hardware for most devices is manufactured in China while software is made by the United States. Licensed applications from Western vendors (or their cracked versions) are usually installed on these devices. Average users spend most of their time surfing the web and reading email messages with browsers or editing documents in various office applications. Other user activities depend on their nature and professional interests. In outline, a significant portion of users pursues their professional tasks when working on computer, and others use their PC mainly for gaming and watching movies.

Anyway, everyone storing sensitive information on their computer, using internet banking services, and chatting online want to protect themselves against data leakage. Users rely on hardware and operating system vendors, application developers, hackers, who provide cracked copies of applications, and all the owners of web services. In case users care about the security of sensitive information and install an antivirus software, they also rely on antivirus developers. Trust is the basis of modern society: a huge amount of different people trusts each other.

## Evolution of malicious android-based applications

Matters of trust arise in relation to both application sources from which these applications are installed and antivirus software used for scanning them. We conducted a research on malware distributed through Google market and would like to share the results of that study.

Android is a modern mobile platform, which together with the associated web services creates a coherent software ecosystem. An important part of this system is Google Market, a service allowing you to download and install applications. In most cases, Android users install applications to their mobile devices via Google Market. Developers can create a Google Market account to publish their own applications. It is no wonder that malware developers also use this service to distribute malware.

By the time the first version of Android was released, malware for all other popular platforms had already existed. Android was no exception. At first, the simplest malware was created for this platform: ransomware Trojans and fraudulent applications. Shortly after, came the spyware stealing passwords and encryption-based ransomware. Then, experts discovered a number of Android vulnerabilities that allowed attackers to create exploits and rootkits for mobile devices. Today, information security researchers saw an increase in both the variety of types of Android malware and botnet networks of malware-infected mobile devices. In general, evolution of Android-based malware seems to be repeating history with the PC malware. In some instances, malicious code from a mobile device can infect a PC, and vice versa.

Most often, the attackers' targets are mobile contact lists, social network account credentials, email accounts, money from the subscriber's account, and bank card data. Malware developers access accounts and contact lists to send spam messages and infect user devices. In some cases, infected devices are used as proxy servers, or as bots for DDoS attacks.

## Sampling and testing of machine learning techniques

Most mobile devices have Google Play Market installed. Although, there are other application repositories available. Some of these repositories were created by respective international companies, others—by relatively unknown developers. Google Market is the most popular among other online stores of Android-based applications. A series of automatic security checks is performed before publishing an application to Google Market. The aim of our study was to determine the probability of getting malware from Google Play Market and reveal ways of detecting malicious code in Android applications.

More than 20,000 of the most popular applications were downloaded from Google Play Market using the unofficial Google API. The most frequently downloaded applications from different categories were selected: chats, cartographic applications, games, etc. In addition, search by keywords was carried out (for example, the names of professions, science fields, cities, and so on. It was intended to use these 20,000 APK files as a representative collection of secure applications. These standards are used to solve the problems of machine learning. Our research was focused on recognition of malicious APK files using various machine-learning techniques; therefore, reference samples were selected with great care. It was necessary to exclude all files that cause even the slightest suspicion about their security from the collection of applications downloaded from Google Market. To accomplish this task, PT MultiScanner was used.

PT MultiScanner is a multithreaded system for identifying malicious content, which allows you to scan both individual files and large collections of files and provides detection verdicts from antivirus software. PT MultiScanner was used to scan file storages, email services, and network traffic via a given set of protocols. The framework of this study involved using PT MultiScanner in paranoid mode in order to identify APK files within the collection, which at least one antivirus recognized as a potential threat. It is not necessary that a file detected by a single antivirus is actually malware. However, since we needed a collection of secure applications, all suspicious files were filtered in this way.

It should be noted that 1,500 files recognized by at least one antivirus as a threat were detected in the collection of 20,000 most popular applications from Google Market. After additional examination, some of these suspicious files were classified as false positives. For instance, our experts detected Trojans disguised as tools for working a mobile phone's flashlight. These software flashlights provide a single button in GUI that is responsible for turning the light on and off. At the same time, these Trojans require an extensive set of privileges including access to phone contacts, camera, and sending SMS messages. Now these applications have been removed from Google Market.

600 files from VirusTotal, in each of which at least 10 antivirus scanners found threats, were taken as standards of malicious applications. The collection of malicious files was selectively checked manually, and no false positives were found. Therefore, the collection of malicious files from VirusTotal was found suitable for use as standards of unsafe applications when testing machine-learning techniques.

The next step required us to identify features that can be used to detect malicious Android applications. According to foreign authors, there are many signs that can detect malicious APK files, but the best for this purpose are the privileges listed in the Android application manifest [8]. In this paper, such privileges were used as one of the signs. We summarized the number of privileges that were rather more often found in the collection of pure standards than in the collection of malicious samples. Some of these privileges are listed in Table 1.

Table 1. Privilege statistics from the application manifest

|  |  |  |
| --- | --- | --- |
| **Privilege** | **The number of malicious samples with such a privilege in the manifest** | **The total number of samples with such a privilege in the manifest** |
| android.permission.INTERNET | 2262.0 | 20981.0 |
| android.permission.ACCESS\_NETWORK\_STATE | 1969.0 | 19652.0 |
| android.permission.WRITE\_EXTERNAL\_STORAGE | 1929.0 | 15162.0 |
| android.permission.WAKE\_LOCK | 505.0 | 10647.0 |
| android.permission.ACCESS\_WIFI\_STATE | 1593.0 | 9492.0 |
| android.permission.READ\_PHONE\_STATE | 2042.0 | 9096.0 |
| android.permission.VIBRATE | 492.0 | 7220.0 |
| android.permission.GET\_ACCOUNTS | 193.0 | 5674.0 |
| android.permission.ACCESS\_COARSE\_LOCATION | 938.0 | 5578.0 |
| android.permission.READ\_EXTERNAL\_STORAGE | 647.0 | 5567.0 |
| android.permission.ACCESS\_FINE\_LOCATION | 726.0 | 5366.0 |
| android.permission.RECEIVE\_BOOT\_COMPLETED | 687.0 | 5003.0 |
| android.permission.CAMERA | 49.0 | 3458.0 |
| android.permission.GET\_TASKS | 641.0 | 2648.0 |
| android.permission.SEND\_SMS | 1768.0 | 2392.0 |
| android.permission.RECEIVE\_SMS | 1613.0 | 2312.0 |
| android.permission.READ\_SMS | 1611.0 | 2193.0 |

The second metric was the number of advertising blocks in the APK file. It was assumed that advertisers are more likely to collaborate with developers of secure applications than with those who make viruses. During cooperation with suppliers of advertising modules, the developer, as a rule, passes a number of checks. Thus, to classify Android applications in our study, we used two metrics: the number of security-related privileges and the number of ad units in the application.

A number of machine learning techniques were tested using the collected data. Linear methods, such as perceptron, adaptive neuron, and logistic regression were tested. We also carried out an experiment with nonlinear models: support vector machines and decision trees. In our expert opinion, the most plausible were the results of classification using logistic regression. The graph with the results of the classification of APK files using this method is shown below.

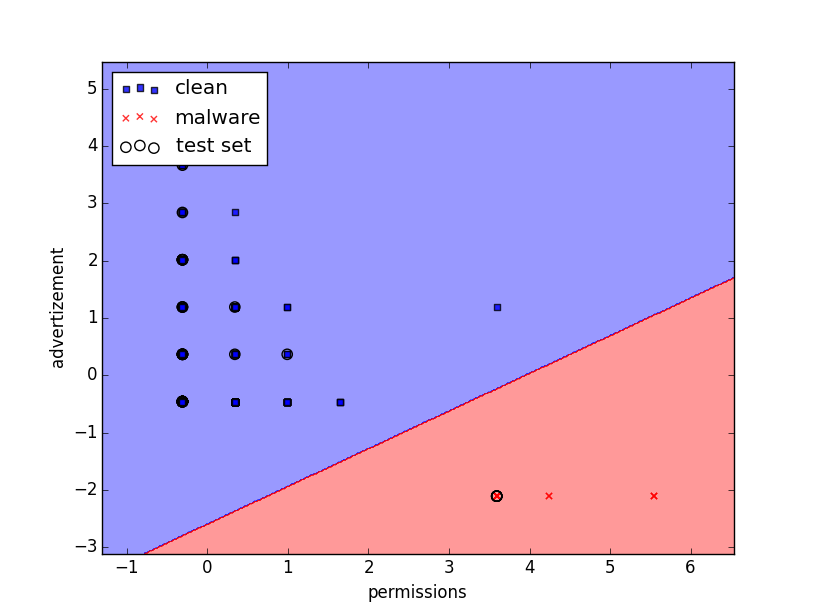


Figure 1. Results of the logistic regression classification algorithm

The vertical axis of the graph represents the number of advertising units in an Android application, the horizontal axis shows the number of privileges characteristic of malicious programs from Table 1.

Based on the results obtained, it was concluded that Android applications containing a large number of advertising modules are more likely to be secure. Dangerous applications can be recognized by the presence of a large number of security-related privileges in the manifest.

The problem is this technique does not allow you to separate secure applications that do not have ad modules from malicious applications that do not require a large number of privileges. Further research will focus on finding new signs of Android applications, which could be used to solve this issue.

## References

Tokareva N. N. *On the History of Cryptography in Russia*. 2012.

Speransky M. N. *Cryptography in South-Slavic and Russian Written Sources.* 1929.

Soboleva T. A. *Cryptography Throughout Russian History*. Moscow, OLMA-PRESS, 2002.

Larin D. A. *Soviet Cryptographic Service During the Great Patriotic War*. 2011.

*Criminal Code of the Russian Federation.* Moscow, Prospekt, KnoRus, 2011.

<https://eugene.kaspersky.com/2011/11/02/the-man-who-found-stuxnet-sergey-ulasen-in-the-spotlight/>

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data. A Revolution That Will Transform How We Live, Work, and Think*. Mann, Ivanov & Ferber, Moscow, 2014.

M. Meier, D. Reinhardt, S. Wendzel (Hrsg.): *Lecture Notes in Informatics (LNI).* Sicherheit, 2016. *Gesellschaft fur Informatik.* Bonn, 2016.