Journal of The Korea Institute of Information Security & Cryptology
VOL.27, NO.6, Dec. 2017

ISSN 1598-3986(Print)
ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2017.27.6.1491

소셜 기반 안드로이드 마켓에서 악성 앱 경향성 분석*

오하영,†‡ 구은희
아주대학교

Trend Analysis of Malwares in Social Information Based Android Market*

Hayoung Oh,†‡ EunHee Goo
Ajou University

요 약

스마트폰의 사용 및 다양한 앱들의 출시 등이 급격하게 증가됨에 따라 악성 앱들도 많이 증가하였고 이로 인한 피해가 속출하고 있다. 안드로이드 앱들이 등록되는 구글 마켓은 앱 등록 규정이 있음에도 불구하고 정상적인 앱들과 악성 앱들이 불가피하게 동시에 존재한다. 특히, 소셜 네트워크가 활성화됨에 따라 안드로이드 구글 마켓에서도 다양한 형태로 보이지 않게 사용자들이 소셜 정보망을 맺고 평점, 다운로드 수 및 인지도 정보 등이 참고 되어 앱 다운로드 수에 반영되고 있다. 결과, 일반 사용자들이 단순히 평점, 인기도, 인기 있는 댓글 및 인지도 높은 카테고리 앱 등만 반영하여 앱을 선택하게 되면 악성 앱 다운로드로 인해 때로는 큰 피해를 볼 수 있다. 따라서 본 연구는 실제 운용되고 있는 안드로이드 마켓에서 장기간 소셜 정보를 직접 크롤링하고 분석하여 악성 앱의 경향성을 처음으로 분석했다.

ABSTRACT

As the use of smartphones and the launch of various apps have increased rapidly, the number of malicious apps has also increased, and the damage is continuing. The Google Market where Android apps are registered is inevitably present at the same time as normal apps and malicious apps even though there are regulations for app registration. Especially, as social networks are activated, users are connected with social networks, and the ratings, downloads and awareness information are reflected in the number of downloaded apps. As a result, when users choose their apps by simply reflecting ratings, popularity, popular comments, and highly-categorized apps, malicious app downloads can sometimes cause significant harm. Therefore, this study first analyzed the tendency of malicious apps by directly crawling and analyzing long-term social information in the currently active Android market.

Keywords: Social Information based Android Market, Advanced Malware, Benign, Tendency Analysis

I. Introduction

Numerous wireless network technologies and the increasing expansion of smartphone usage has† enhanced the development of social network that puts the basis on Recommender Systems (RSs). Yet there has been an increase of

Received(07. 13. 2017), Modified(10. 23. 2017) , Accepted(11. 06. 2017)

* This work was supported by the Ajou University research fund and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1A1B03035557).

† 주저자, hyoh79@gmail.com
‡ 교신저자, hyoh79@gmail.com(Corresponding author)

tendency which abuses this phenomenon by using malicious apps. This paper has been conducted in order to analyze actual gathered data of apps, show that attackers show the tendency to spread malicious apps via items that show high rating value or high recognition of social information, and present the importance of researching strong social network service based on these tendencies.

A social network is an open network in which anyone can join a social network to reflect the opinions of others. Therefore, anyone using a social network can refer to ratings and comments of other users of the social network, and reflect the information in the decision of the user. In recent years, there has been a phenomenon of Sybil attack that maliciously utilizes the characteristics of the open social information network to physically create a plurality of fake accounts, and manipulate the tendency of the normal users of the social network.

Thus, in this paper, we analyze that the percentage of malicious Android apps rated by owners of these Sybil accounts is much higher than the ratio of normal users′ normal apps. This is because, unlike existing malicious app spreaders simply creating malicious apps and expecting to download them at random, they are now becoming more intelligent to finally download malicious apps after configuring them to look like the normal users with reputable apps using social network information.

Especially, when using social network, it is shown that the rating manipulation such as target item, filler item, and average items is set to a tendency similar to that of normal users.

In section 2, we introduce the system environment of gathering and analyzing data. In section 3 and 4, we explain the proposed scheme and evaluate the performance. Finally, section 5 concludes the paper.

II. Gathering and analyzing data

2.1 Background

We gather and analyze real app data of two types from android market with scrapy and beatifulsoup4 of Python. First samples of market apps for this paper have been gathered by subjecting GooglePlay, a standard market, for 12 months, from Jan. to Dec. in 2016. And for the second dataset, we newly crawled the samples of market apps from Jan. 2017 to July. 2017. Crawling methods based on app samples have been chosen by downloading apps which have been frequently downloaded. Each app sample gathered from standard market were 28,193 in total, and in order to decrease unequal distribution, we have prevented the gathering of overlapping apps through figuring out names of packages.

According to the existing research [1-10], malicious apps are being distributed via standard market. That is, apps developers can register Apps in android market through the secure standard procedure. But outliers or attackers can still upload abnormal Apps or malware by avoiding the sophisticated proof of the Android market with a variety of approaches.

Normally the detection methods of the malicous apps can be classified into static and dynamic approaches explicitly. In case of the static method, it only depends on a static program (source code or binary). Therefore it cannot consider the dynamic part of the program against a malware′s semantic signature and that of a new binary. On the other hand, even though the dynamic method improves the limits of the static method, it cannot still protect the damage by misusing the hidden effect of recent social network implicitly.

Therefore in this paper by subjecting the crawled data, VirusTotal [11] has been applied, while diagnosis of malicious apps used in the experiment has been used by the list of those created from F-Secure [12] considering the statistics information of the social network.

F-Secure is anti virus and malware protection program for PC developed in a Finnish cyber security and privacy company based in Helsinki, Finland. Using F-Secure we can verify the malware different from normal Apps.

VirusTotal is a website aggregating many antivirus products and online scan engines[4][5] to check for viruses that the user's own antivirus may have missed, or to verify against any false positives. It is by the Spanish security company Hispasec Sistemas.

In this paper, we use VirusTotal as well as F-Secure for verifying the result of detecting malware Apps of the proposed scheme.

[6] and [7] shows the examples of VirusTotal's own capability. [6] explains about applying behavioral detection on android-based devices in the following scenario: Files up to 128 MB can be uploaded to the website or sent via email. [7] shows the tendency of virus towards taming privilege-escalation attacks on Android. Anti-virus software vendors can receive copies of files that were flagged by other scans but passed by their own engine, to help improve their software and, by extension. [8] is about that users can also scan suspect URLs and search through the VirusTotal. VirusTotal for dynamic analysis of malware uses Cuckoo sandbox. In summary, VirusTotal was selected by PC World as one of the best 100 products of 2007 [9].

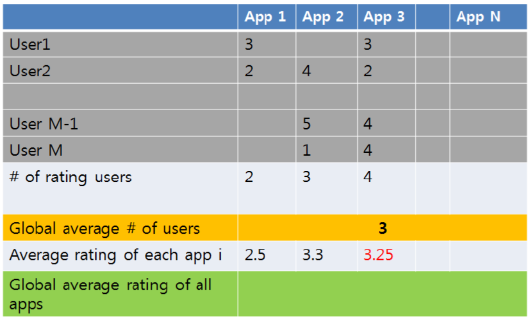
2.2 System model

In this paper, to account only for accounts rated higher than the average number of ratings for all apps, we need to include only the preprocessed information, which leaves only accounts that satisfy Eq. (1) in a crawling matrix such as Fig. 1 similar to the previous work [13].

$$\text{Globalratingof all apps}(G_r) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\text{rating}_{ij} \qquad (1)$$

In Equation (1), *i* denotes an *i*-th user and *j* denotes a *j*-th app. As a result, only the user accounts ranked above the average number of the data as shown in Fig. 1 are kept in the matrix.

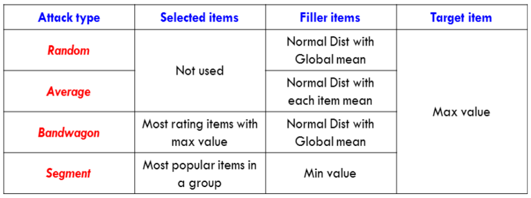
Generally, in the recommendation system of social network, items are classified into highly recognized popular items, target items, and other filler items. A normal user and Sybil freely rate selected item, target item and filler item. In Apps recommendation system, the items are Apps.



	App 1	App 2	App 3		App N
User1	3		3		
User2	2	4	2		
User M-1		5	4		
User M		1	4		
# of rating users	2	3	4		
Global average # of users	**3**				
Average rating of each app i	2.5	3.3	3.25		
Global average rating of all apps					

Fig. 1. Matrix with the rating values (R_{MxN})

Table. 1. Various types of Sybil attack



Attack type	Selected items	Filler items	Target item
Random	Not used	Normal Dist with Global mean	Max value
Average		Normal Dist with each item mean	
Bandwagon	Most rating items with max value	Normal Dist with Global mean	
Segment	Most popular items in a group	Min value	

If Sybil is highly rated only on the target item, it will easily be detected if it raises awareness and sales. Thus, Sybils try to appear in intelligent form as shown in Table 1, taking into account item-specific averages for selected items, target items, and filler items to look as similar as possible to normal users.

In particular, the average attack is more difficult to be detected since it is characterized by the fact that the attacker can know the average score of each item, and the score of the filler items is filled with the average score of each item similar to the normal items.

III. Proposed Scheme

In this paper, for the first time to analyze the intelligent tendency of Android malware apps based on recent social network information, we consider the average rating trend of the social information of the malicious and normal apps. We assume that the probability of malicious apps are increased with Sybils. For example, in Figure. 1 if App 2 and App 3 are malwares, we can verify the probability of Sybil of the user 2 is higher than other users. For that we proposed Sybil metric as shown Eq. (2).

$$\mathrm{P}_{\mathrm{sybil}}(\mathrm{m}) = \left(\alpha \cdot C_m{}^{O} + \beta \cdot C_m{}^{F} + \gamma \cdot C_m{}^{S}\right) \div T_{rating}(m) \tag{2}$$

P$_{sybil\ (m)}$ is the probability that user m is an account of Sybil. Sybil selects a group of selected items and filler items at random to increase the rating of the target items.

On the other hand, a normal user will score each item on the remaining filler items according to the freely recognizable selected item and other personal tendencies without regard to the target items in particular. Therefore, P$_{sybil\ (m)}$ will have a large value because the user m is Sybil, and each group item, target, selected, and filler items will follow various Sibil Attack tendencies such as Table 1.

To do this, we multiply each group item by the weights α, β, ɣ and divide it by the total number of users m to obtain the probability that the user m is a Sybil account.

The optimal value of each weight is calculated with enough evaluations as α = 0.10, β = 0.35 and ɣ = 0.55 by considering the characteristics of the crawl data and that of the crawl by type, as a result of other studies that the average of Sybil user accounts is less than 8%. P$_{sybil\ (m)}$, which is the difference between normal user and Sybil, was verified as 0.3.

As social networks become active, regular users reflect ratings and comments on their apps from the social information of other users before purchasing the app. As we mentioned before, we assumed that there is a high probability that there will be malicious app developers among the Sibyl accounts that create a number of user accounts and rate the maximum value for the target items. To raise malicious apps on Google Play and raise awareness of those apps, the malicious app developers can make an intelligent attack that gives maximum value to target apps.

Therefore, finally to check how much of the Sibyl accounts are in each app, we define Eq. (3) as follows similar to [13].

$$\text{sybil ratio} = \frac{\sum_{i=1} \text{sybil}_{ID_i}}{\sum_{i=1} \text{sybil}_{ID_i} + \sum_{j=1} \text{normal}_{ID_j}} \qquad (3)$$

Sybil$_{IDi}$ refers to a Sybil account with a Sybil probability metric of at least 0.3 by applying Eq. (2), and normalIDj denotes a normal account with a value less than 0.3. As a result, the sybil ratio as shown in Eq. (3) for malicious app type is found, and if there is a proportional relation, it is known that there is a high correlation, and if there is no regularity, it is known that there is a low correlation.

IV. Experiments

4.1 Dataset from Jan. 2016 to Dec. 2016

Figure 2 shows the rate numbers of downloads of normal app and malicious app based on the crawled apps that have been sampled based on its frequency. Axis X shows average of download numbers, while axis Y shows percentage of apps that contain specific download numbers. Distribution of normal apps and malicious apps is being formed around download numbers from 30,000 to 300,000, and it has been figured out that the number of malicious apps, when compared to the number of normal apps, is much bigger. This shows that apps that have a higher number of download are the ones that are more known to the public, meaning that attackers using malicious apps could distribute malicious apps in an easier way.

Figure 3 shows the rate based on categories of normal apps and malicious

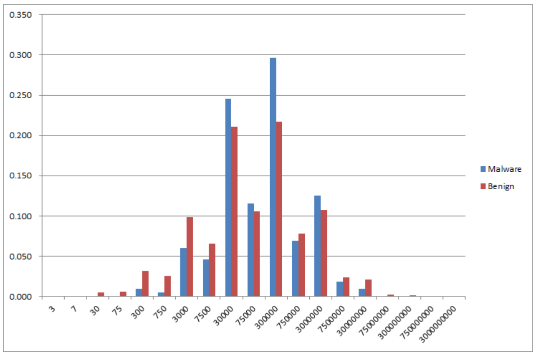


Fig. 2. Rate of normal/malicious apps based on download frequency of normal market in 2016

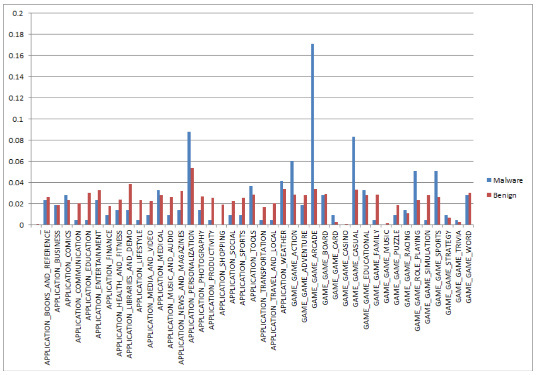


Fig. 3. Rate of normal/malicious apps based on category of normal market in 2016

apps which took basis on crawled apps by focusing frequency. In the case of malicious apps, distribution has been taken place through games including arcade, casual, and role-play, all of which people could enjoy during their pastime, and decoration apps including Dodol Launcher. In the case of decoration apps which distribute packaged apps in forms that include decorating fonts, the designated app does not get activated in the actual screen, making malicious apps become distributed without people noticing.

Figure 4 shows the rate between malicious apps and normal apps based on the category. It can be figured out that even though the number of malicious apps

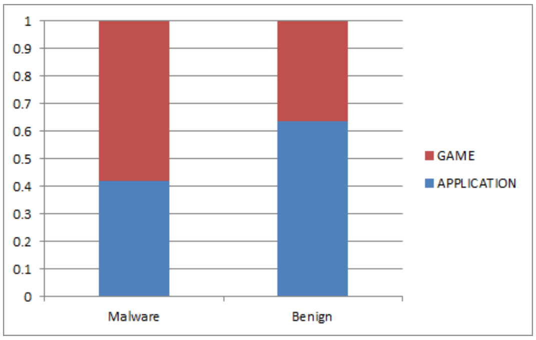


Fig. 4. Rate between normal/malicious apps based on category of normal market in 2016

makers is small, they tend to distribute malicious apps via game apps.

normal apps based on the rating. Due to crawling methods that focuses on frequency, rating of normal apps is relatively higher, yet rating of malicious apps show high rate within the range of 3.5 to 4.5. This indicates that apps with higher ratings could be recognized as apps favored by people, which could affect the number of download. Hence, makers of malicious apps could use it as a media that could distribute malicious apps in an easier way, which, in turn, could use them for distributing malicious apps by manually manipulating the rating of certain apps.

4.2 Dataset from Jan. 2017 to July. 2017

Figure 4 shows the rate of normal/malicious apps based on download frequency of normal market in 2017. Similar with the results of 2016, malware follows the normal distribution.

Figure 5 and 6 show the rate of normal/malicious apps based on category of normal market in 2017. It is a little different with the result of 2016 since most apps are mainly distributed in the application related apps. The important

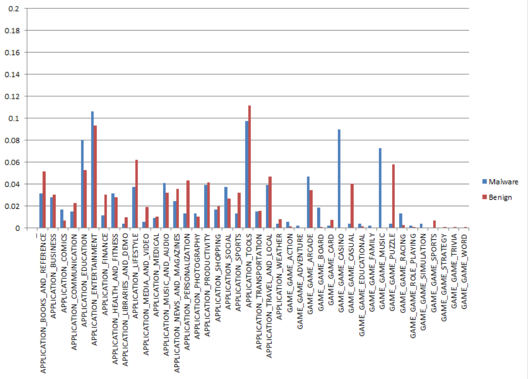


Fig. 5. Rate of normal/malicious apps based on category of normal market in 2017

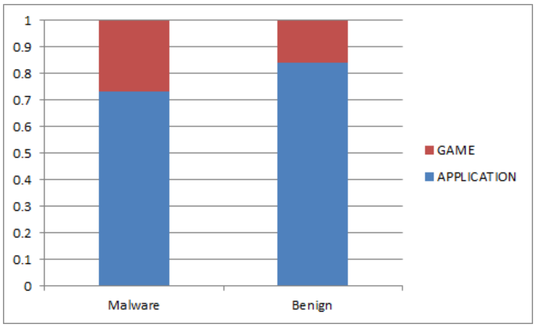


Fig. 6. Rate between normal/malicious apps based on category of normal market in 2017

founding from 2017 dataset is for many malware to move to application related normal apps based on this tendency based on the social information similar to the normal users.

V. Conclusion

Existing research regarding the search of malicious apps has not considered characteristics of social network including category, numbers of download, and rating change of gathered data, but only considered old classification system that focuses on similarity comparison and signature methods. However, when considering social network service and

RSs, malicious app attackers could perform organized attack by planting malicious code to the source code itself by recognizing information of RSs including numbers of download, rating values and information of apps category. This research has based actual data regarding RSs and showed the specific portion of malicious apps from numbers of download, rating values and distribution of apps category by having a basis of actual data regarding RSs, and set goal to show the importance of research and the importance of also considering figuring out the malicious apps for reliable social network service.

Further researches needed include those that focus on making credible social network service methods in which matrix of social network service has been combined with malicious apps detection method.

References

[1] Krebs B. Mobile malcoders pay to (Google) Play. Krebs on Security, 2013, http://bit.ly/1kranE5.

[2] “Gartner Says Annual Smartphone Sales Surpassed Sales of Feature Phones for the First Time in 2013,” Feb. 2014, https://www.gartner.com/newsroom/id/2335616

[3] “Security Apps under Permanent Stress,” Mar. 2014, http://www.avtest.org/en/ news/news-single-view/artikel/security-apps-underpermanent-stress.

[4] Enck, W., Ongtang, M., McDaniel, P., “On lightweight mobile phone application certification,” Proceedings of the 16th ACM conference on Computer and communications security, pp. 235 - 245, Nov. 2009.

[5] Pearce, P., Felt, A.P., Nunez, G., Wagner, D., “Addroid: Privilege separation for applications and advertisers in android,” Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, pp. 71 - 72, May. 2012.

[6] Shabtai, A., Elovici, Y., “Applying behavioral detection on android-based devices,” Mobile Wireless Middleware, Operating Systems, and Applications, pp. 235 - 249, vol. 48, Dec. 2010.

[7] Bugiel, S., Davi, L., Dmitrienko, A., Fischer, T., Sadeghi, A.-R., Shastry, B., “Towards taming privilege-escalation attacks on Android,” Proceedings of the 19th Annual Symposium on Network and Distributed System Security, pp. 1-18, Nov. 2012.

[8] Bose, A., Hu, X., Shin, K.G., Park, T., “Behavioral detection of malware on mobile handsets,” Proceedings of the 6th international conference on Mobile systems, applications, and services, pp. 225-238, June. 2008.

[9] Blasing, T., Batyuk, L., Schmidt, A.-D., Camtepe, S.A., Albayrak, S., “An android application sandbox system for suspicious software detection,” Proceedings of Malicious and Unwanted Software (MALWARE), pp. 55 - 62, Oct. 2010.

[10] Yang, C., Yegneswaran, V., Porras, P., Gu, G.,“Detecting money-stealing apps in alternative Android markets,” Proceedings of the 2012 ACM conference on Computer and communications security, pp. 1034 - 1036, Oct. 2012.

[11] VirusTotal . VirusTotal - Free Online Virus, Malware and URL Scanner. https://www.virustotal.com/en/; 2014.

[12] F-Secure . F-Secure, 25 years of the best protection in the world. http://www.f-secure.com/en/web/labs global/; 2014.

[13] Oh, “Relationship Analysis between Malware and Sybil for Android Apps Recommender System,” Journal of The Korea Institute of Information Security & Cryptology, pp. 1235-1241, vol. 26, no. 5, Oct. 2016.

〈저자소개〉

오 하 영 (Hayoung Oh) 정회원
2002년 2월: 덕성여자대학교 컴퓨터공학과 졸업
2006년 2월: 이화여자대학교 컴퓨터공학과 석사
2013년 2월: 서울대학교 컴퓨터공학과 박사
2010년 4월~2010년 10월: U.C. Berkeley 방문연구원
2013년 3월~2013년 8월: 서울시립대학교 연구교수
2013년 9월~2016년 8월: 숭실대학교 전자정보공학부 조교수
2016년 9월~현재: 아주대학교 다산학부대학 조교수
〈관심분야〉 소셜 정보망, 추천시스템, 무선 네트워크 및 비디오 스트리밍

구 은 희 (EunHee Goo) 정회원
2002년 8월: 단국대학교 전자컴퓨터공학부(공학사)
2004년 8월: 단국대학교 전자컴퓨터공학과(공학석사)
2009년 8월: 단국대학교 전자컴퓨터공학과(공학박사)
2011년 3월~2013년 2월: 서일대학교 정보통신과 강의전담 교수
2013년 3월~2014년 9월: ㈜도넛시스템 LSI 이미징사업부 책임 연구원
2014년 10월~2016년 8월: ㈜이너트론 이동통신연구소 수석 연구원
2016년 9월~현재: 아주대학교 다산학부대학 조교수
〈관심분야〉 정보보호, 암호 알고리즘, 서비스로서의 보안(ASCaaS), 소프트웨어 교육