Paper Review CS846, Mar. 06
Mining apps for anomalies **and** *Mining apps for abnormal usage of sensitive data*

Wenhan Zhu (Cosmos)
w65zhu@uwaterloo.ca
University of Waterloo
Waterloo, Canada

ABSTRACT

This week's chapter from the book being reviewed here is **Mining apps for anomalies** by *Andreas Zeller* and **Mining apps for abnormal usage of sensitive data** by *Vitalii Avdiienko, Konstantin Kuznetsov, Alessandra Gorla, Andreas Zeller, Steven Arzt, Siegfried Rasthofer and Eric Bodden.*

Zeller's chapter is a brief overview of the research on mining application behavior to determine malicious apps.

Avdiienko et al. is the paper that *Zeller*'s chapter covered. It went through in detail how tracking data flow in *Android* apps can identify malware with high success rate.

KEYWORDS

paper review

1 SUMMARY

Since both the chapter and paper is on the same topic, I'll summarise both of them together.

It is a very important thing to make sure that the apps on our phones does what it claims to do. However, it is not always the case in real life. Previous studies on the topic focus on finding the similarities of apps with existing malware to determine whether the app is harmful for example comparing byte code to find malicious byte code. Due to the nature of the topic, databases of apps is not easily shareable, and the best practice is for each research group to mine their own copy of apps from the app store. On the other hand, database of malware are quite common since malware authors are unlikely to fire a legal claim to the applications. The main idea here is to look at the dissimilarities between normal software and malware. The authors looked at the flow of sensitive data in popular android applications and compared it to the flow of sensitive data in malware and found a significant difference between them. Normal apps that uses sensitive data often have the data flow sinked to some fixed number of places such as log. In android, *intent* is used to communicate between components, the authors didn't follow the intents but labeled it as a sink. Malware tend to have sensitive data sink to different places most likely places that is not supposed to be access by the app description. A example app that requires messaging but it's not an message app. A particular discovery in the paper is that about 25% of the malware the authors used have data sinked to SMS while only 1% of popular normal apps used it as a sink. However, due to the source of the database it's unknown if it's a good representation of malware. Using the data flow of sensitive information, the authors was able to create a classifier that can successfully identify malware with high percentage (90%) and low recall (20%). This paper showed a new way to identify malware by looking at data flow of sensitive information.

2 THOUGHTS

We have been fighting against viruses and malware since it's introduction to computing devices. With the rise of the internet era and the mass adoption of smart phones. The main source of malware have been moved from PC to mobile and the internet. Leaking of information happens on websites and mobile devices. I still remember that a few years back, there was a breach in Apple store due to developers using a compromised version of XCode that injected malicious code segments in their apps affection millions of users in China. And the reason for using the compromised version of XCode is actually quite funny, back then Apple did not have servers in China, so the download speed for XCode is extremely slow, so people would download it from third party sources, and unfortunately, one common sharing website was compromised and it distributed a malicious version of XCode. I think the results of this study can be used by application distributors to monitor the behavior of applications so that more malware can be detected before distribution. The method discussed in the paper seems deployable in production and computing power also seems reasonable.

Collecting user data seems to be the main focus recently by large companies. Reports show that *Amazon* is tracking every event when browsing their website. And the competition for getting more user data let to some bad competition. For example, *Facebook* having shadow profiles that's impossible to delete and *Google Chrome*'s delete all cookies means delete all cookies that's not *Google*'s. This paper's approach can't not capture the idea of these information leaks and some of them considered malicious by advocates on privacy. There is still a long way to go to combat internet privacy.

3 RATINGS

I would rate *Zeller*'s chapter 4/5. It's a nice summary over the content and it give needed information to understand the process.

I would rate *Avdiienko et al.*'s work a 4/5. It introduces novel idea and the execution is pretty nice. However, some of the data used does not seem perfect and the results is not 100% convincing due to it. However, personally I could not find a way to improve it given the constraints.

4 REFERENCES

(1) Apple Store Malware https://www.washingtonpost.com/news/the-switch/wp/2015/09/21/apples-app-store-was-infected-w

ith-malware-from-china/?noredirect=on&utm_term=.c6cf4c87b9b4
(2) Facebook Shadow Profile http://theconversation.com/shadow-profiles-facebook-knows-about-you-even-if-youre-not-on-facebook-94804
(3) Chrome keep Google's Cookies https://www.bleepingcomputer.com/news/google/chrome-69-keeps-googles-cookies-after-you-clear-browser-data/