Evaluating Malware Mitigation by Android Market Operators

Yosuke KIKUCHI[†] Hiroshi MORI[†] Hiroki NAKANO[†] Katsunari YOSHIOKA[†]
Tsutomu MATSUMOTO[†] Michel VAN EETEN[†‡]
[†]*Yokohama National University, Japan* [‡]*Delft University of Technology, the Netherlands*

Abstract

All Android markets are confronted with malicious apps, but they differ in how effective they deal with them. In this study, we evaluate the mitigation efforts of Google Play and four third-party markets. We define three metrics and measure how sensitive they are to different detection results from anti-virus vendors. Malware presence in three third-party markets – Liqucn, eoeMarket and Mumayi – is around ten times higher than in Google Play and Freeware Lovers. Searching for certain keywords in Google Play leads leads to a fifty times higher malware rate than those for popular apps. Finally, we measure malware survival times and find that Google Play seems to be the only market that effectively removes malware, though it contains a cluster of apps flagged as adware and malware over long time periods. This points to different incentives for app markets, anti-virus vendors and users.

1 Introduction

In recent years, Android has become the dominant mobile operating system. Unsurprisingly, this also means that it has become more attractive for criminals to target Android devices. A key attack vector to get malware on these devices is via Android app markets, as they are the main mechanism through which users obtain their apps. Criminals put malware-laden apps in these markets and then use a variety of ways to get users to install them.

Next to "Google Play", the official market operated by Google, there are many third-party markets. All markets contain some degree of Android malware mixed among the regular applications [13][23][24]. The security of end users depends to a significant degree on whether these app markets mitigate malicious apps and, if they do, how effective they are. Mitigation can be done by preventing malicious apps from entering the market, by reducing the exposure of users to these apps, and by removing them from the market after they are discovered. We develop metrics to measure each of these activities.

We evaluate the malware mitigation of five of these markets: Google Play and four third-party app markets. We extend earlier work on infection rates of markets with several new contributions:

- We detail how sensitive malware presence and mitigation metrics for app stores are to detection rates of anti-virus (AV) solutions in VirusTotal;
- We find that malware rates in three third party app stores – Liqucn, eoeMarket and Mumayi – are more than ten times higher than in the Google Play store and Freeware Lovers. The malware rate in Google Play is relatively low, but fifty times higher in the search results for certain keywords;
- We present the first study of malware survival times in app stores. We find that only Google Play seems to act against malware that is detected after entering, but even this store contains long-lived malware. A portion of these apps are adware and they might remain in the store because of divergent security incentives among app markets, AV vendors and users.

After discussing related work (Section 2), we describe the method of app collection and malware identification used for this study (Section 3). Next, we define and calculate three metrics to evaluate malware mitigation by app market operators and discuss the results for different markets (Section 4). In Section 5, we discuss the limitations of the study. Section 6 summmarizes the main conclusions.

2 Related Work

A significant amount of work has studied methods to detect malware. Zhou and Jiang [23] has provided an early quantification of the problematic success rate of detecting Android malware by AV vendors. We extend this work with a more recent and more comprehensive look at detection rates by vendors.

Other work [19] [11] has focused on infection rates of the devices themselves, which has also been studied by Google itself [14].

Our study focuses on the app markets. Zhou et al. [22] has studied six third-party markets and concluded that

5% to 13% of apps on third-party markets are repackaged from Google Play, sometimes for malicious purposes. In a more recent in-depth study [24], they found that infection rates can vary by more than a factor of 10, from 0.02% for Google Play and alternative marketplaces ranging from 0.20% to 0.47%. Chakradeo et al. [12] have studied three third-party markets, different from the ones included in our study, to evaluate their malware detection technique. They include VirusTotal and Androguard[1] detections for comparative purposes. On average, they find the alternative markets to have an infection rate of 0.52%. Vidas and Cristin [20] have studied the relation between repackaging of apps and malware. They found wildly varying malware rates for markets, from zero to 100%.

Similar work on malware rates has been done by AV and hardware vendors (e.g., [13], [5]). Google itself published malware rates for the Play store as a proportion of overall app downloads by users [14].

We complement these studies by combining different metrics, and adding a new metric on survival times, to evaluate the effectiveness of malware mitigation by market operators. We also measure how sensitive these metrics, and thus the prevalence rates reported in earlier work, are to differences in detection methods.

3 Method

We sampled apps from the selected Android markets and then downloaded them, including their meta-data such as the number of downloads and the date when they were last updated. After downloading, we submitted the apps to VirusTotal and logged which AV vendors had flagged the app as benign or malicious and, in case of the latter, with what labels.

From these properties, we calculate three security metrics for each market:

1. the malware presence ratio (percentage of all collected apps that are detected as malware);

2. the malware download ratio (percentage of all downloads of the collected apps belonging to apps that are detected as malware);

3. the survival period of malware (how long apps detected as malicious remain in the app store).

3.1 Data Collection

F-Secure's 2014 Theat Report covers the trends in Android malware over the second half of 2013. We selected the largest markets as identified in the report, though which markets are dominant changes dramatically over time. Liqucn, eoeMarket and Mumayi were once among the "household names" and the "top players" in China [10] [9]. In just a year or two, they were overtaken by app markets created by tech giants and smartphone developers. Because we wanted to track survival times of malware in app markets, we need markets with a presence and user base over a longer time period. For this reason, we studied the five dominant markets as identified by F-Secure in early 2014, rather than the recent top players.

In order to collect free apps from the Android Markets automatically, we developed web crawlers for each market to identify and download apps and then submit them to VirusTotal. We tried to download the alternative markets in full. Google Play is too large to fully download, so we collected two samples.

Not all apps could be downloaded and a tiny fraction of the downloaded apps could not be submitted to Virustotal because of a file size limitation (64 MB). For each market, we specify the number of apps that were successfully downloaded and submitted for evaluation.

Since there is no way to enumerate all apps in the Google Play (GP) store, a true random sample is impossible. We sampled GP in two ways. First, we downloaded all 14,580 apps which were listed as the top 540 popular free apps in all 27 categories, such as tools, entertainment, etcetera. In the end, we could download 12,280 apps (84.2%), because of limitations imposed by Google Play related to Android OS versions or locale of the client device. Note that all downloads were attempted from Japan. We call this dataset GP_PO.

For comparative purposes, we also sampled apps via PlayDrone dataset[21][15] to draw a more randomized sample. The set is from 2014 and contains 1.5M apps. We randomly selected 15,000 free apps. Since PlayDrone was collected in 2014, we then checked whether these apps still exist in the Play store at the time of the study, since we cannot tell when the other have been removed. Our set is the 8,966 apps that remain in the store. We call it GP_PD. It is no longer random, as the removed apps might be biased in unknown ways, such as having a higher probability of being malware.

Note that both GP samples have their own strengths and weaknesses. GP_PO captures better what apps users are likely to encounter, while GP_PD is probably more representative of the overall market.

For Mumayi (MY), we collected apps via their web pages in the market. The URLs are constructed in a simple format, with a decimal number as ID for each app. Based on this format, we enumerated all possible IDs and found 93,286 unique apps. We could download only 19,794 apps (21.2%) for a variety of reasons, such as links no longer being available or requiring user registration. The language barrier and black-box nature of the market operation prevented us from identifying all the causes for the failures. While the IDs of successfully downloaded apps are distributed all across the ID space without any discernible bias, we have to treat this sample as potentially biased. While the data can still provide useful insights, we have treat our findings for Mumayi with a bit more scepticism.

From eoeMarket (EM), Liqucn (LQ) and Freeware Lovers (FL), we respectively found 10,422, 7,261 and 3,683 apps by exhaustively crawling all web pages of the markets. We successfully downloaded and evaluated most of them, namely, 10,203 free apps (97.9%), 6,884 free apps (94.8%), 3,652 free apps (99.2%), respectively.

Table 1 summarizes the evaluated app markets, the number of apps downloaded and evaluated by Virustotal, and their collection period.

Table 1: Evaluated App Markets, Number of Collected Apps, and Collection Periods

Market	Collected Apps	Collection Period
GP_PO	12,280	2015/09/21 - 2015/09/28
GP_PD	8,966	2014/08/07 - 2014/11/03
MY[8]	19,794	2015/07/14 - 2015/08/09
LQ[6]	6,884	2015/09/02 - 2015/09/04
EM[3]	10,203	2015/09/15 - 2015/09/19
FL[4]	3,652	2015/06/20 - 2015/06/23

3.2 Positive Detection Ratio

We use VirusTotal to track which AV-vendors have identified an app as malware. Needless to say, detection efforts, and thus VirusTotal, suffer from false negatives as well as false positives. It should also not come as a surprise that certain apps are flagged as malware by some vendors, but not by others. In other words, apps have different positive detection rates as being malicious.

We define the *Positive Detection Ratio (PDR(S, a))* of a certain application a by the inspection of the anti-virus product group S by the following equation.

$$PDR(S,a) = |\{av \in S | av(a) = malicious\}|/|S| \quad (1)$$

In (1), *av (a) = malicious* means that an AV product *av* has detected application a as malicious. S={*ALYac, AVG, AVware, Ad-Aware, Aegislab, Agnitum, Ahnlab-V3, Alibaba, Antiy-AVL, Arcabit, Avast, Avira, Baidu-International, BitDefender, Bkav, ByteHero, CAT- QuickHeal, CMC, ClamAV, Comodo, Cyren, DrWeb, ESET- NOD32, Emisoft, F-Prot, F-Secure, Fortinet, GData, Ikarus, Jiangmin, K7Antivirus, K7GW, Kaspersky, Kingsoft, Malwarebytes, McAfee, McAfee-GW-Edition, Microworld-eScan, Microsoft, NANO-Antivirus, Panda, Qihoo-360, Rising, SUPER-AntiSpyware, Sophos, Symantec, Tencent, TheHacker, TotalDefence, TrendMicro, TrendMicro-Housecall, VBA32, VIPRE, ViRobot, Zillya, Zoner, nProtect*} in case of $|S| = 57 (|S| \leq 57)$.

The total number of anti-virus solutions making up the PDR percentage varies a little, as some AV products do not provide any test result, positive or negative, for some of the apps. Overall, the average number of the anti-virus products that rendered a verdict for an app was 56.

4 Evaluating App Markets

4.1 Malware Presence Ratio

To evaluate the security of app markets, prior work has predominantly focused on the fraction of all apps in a market that are malware, a metric that has also been referred to as the "infection rate" [24]. Along these lines,

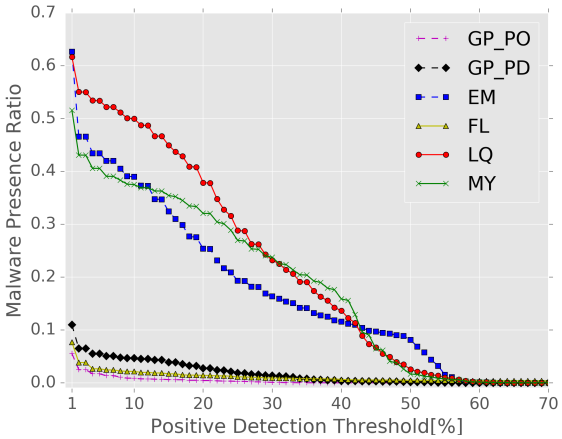

Figure 1: Malware Presence Ratio for Each Market

we define *Malware Presence Ratio (m, S, M)* as:

$$MPR(m,S,M) = |\{a \in M | PDR(S,a) > m\}|/|M| \quad (2)$$

where M is the set of collected apps for a market and m is the chosen threshold value for the positive detection ratio. (Note that a and S are the same as before: apps and the set of AV-products that have rated the app).

Figure 1 visualizes the Malware Presence Ratio (MPR) of all five evaluated markets for different threshold values m of the Positive Detection Ratio. As we can clearly see, different PDR thresholds lead to different malware presence ratios.

The variance of MPR is smallest for GP_PO, GP_PD and FL. They have low ratios, even at low positive detection thresholds. The three other markets have very high MPRs at low thresholds. Even at higher thresholds, the portion of apps flagged as malicious is disconcerting, though it eventually drops to zero.

This raises an important question: when can an app be reliably considered malicious? At very low thresholds, e.g., when less than 10% of AV products label an app as malware, we are likely to incur a significant false positive rate. On the other hand, if we want to avoid false positives by waiting for more consensus among AV-products, say a PDR of 70%, we end up with a malware ratio of zero. Prior research and also Google's own analysis [14] have established beyond doubt that the true malware rate is non-zero, so high threshold values clearly suffer from an unacceptable rate of false negatives.

If we assume that 10% is too low and we know that 70% is too high, how can we then evaluate app markets? Rather than setting a specific threshold value, we will show how the different security metrics we developed vary across different PDR values. Readers can then assess the sensitivity of the metric to those values as well as focus on the range that they find sensible.

For the evaluation of app stores, this still leaves us with the question of when the combined signal of AV vendors is the most reliable – or rather, the least unreliable. To balance the trade-off of false positives versus false negatives, we draw some heuristics from prior work.

The F-Secure report referenced earlier [13] found that Google Play had an MPR of 0.1%. While this was calculated against a different sample of apps than ours, it provides a useful point of reference. An MPR of 0.1% corresponds to a PDR threshold value of 30%. Zhou et al. [24] found a much lower MPR of 0.02%, based on a specific detection approach developed by the authors. This corresponds to a PDR of 34%.

Within the range of 10 to 70%, this prior work suggests the most reliable PDR is likely closer to 10% than to 70%. Another reason to gravitate towards that side of the range is that app detection is unevenly distributed across the AV products. The vendors that invest more in the Android platform will be able to identify more malware than those that don't. When we rank the vendors from those with the most detections to those with the least, the first one third of the vendors (19 of 56) do 72% of all detections. In contrast, the last third of vendors only add only 1% more detections.

These heuristics do not point to a precise value, of course, but they do suggest that a PDR of 30% is a reasonable approximation of when the detection by AV vendors combine to a reliable signal. As mentioned above, this PDR corresponds to a malware presence ratio of 0.1% for GP_PO and 1.38%, so one order of magnitude higher, for GP_PD. One explanation could be that popular apps are more often checked by the app store, AV vendors and users. FL is in the same range as GP_PD: 0.93%. The MPRs of LQ, MY, and EM are more worrying: 23.2%, 23.8%, and 16.0%, respectively.

We should note that balancing false positives and false negatives – i.e., setting a reliable PDR – is not just a problem for external evaluations of app markets, but also for the mitigation practices of the market operators themselves. These findings suggest that malware detection cannot rely on a single source, either the market operator's own methods or those of an AV vendor. While that might sounds uncontroversial, it implies that mitigation policies have to contend with an uncertain set of signals with which to drive decisions to block or remove apps from the store.

MPR for Google Play via Keyword Search. While the MPR for GP_PO is low, we have explored whether this risk was perhaps concentrated, and thus higher, in certain areas of the market. For this purpose, we collected another sample of apps, this time via keyword searches.

We used 155 popular keywords listed on the keyword ranking sites [7] [17] [2], and 11 keywords in the security report by Symantec [18]. We then collected the 40 apps that are shown on the first page of search results. Figure 2 shows the keywords with the highest MPR, calculated for the threshold $m = 30\%$.

Keywords like "sms" and "adult" lead users to apps with an MPR that is more than 50 times higher than the rate for GP_PO. Figure 3 shows how the MPR of several top keywords varies over different positive detection ratios. To illustrate, the search results for the keyword "sms" contain at least one app that is considered malicious by nearly 50% of all AV products. In short, these

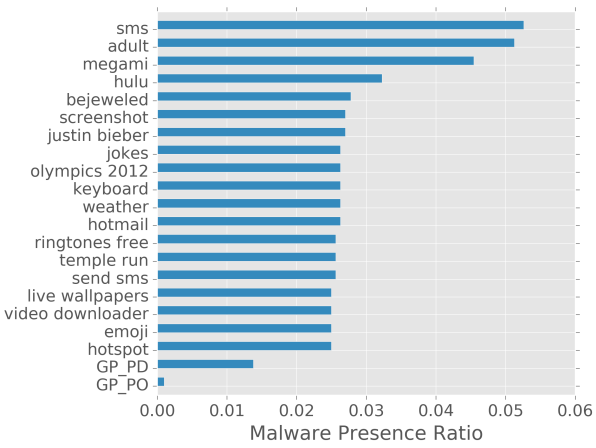

Figure 2: Malware Presence Ratio for Google Play for different keywords, using a Positive Detection Threshold $m = 30\%$. "All Categories" is the MPR for the total set.

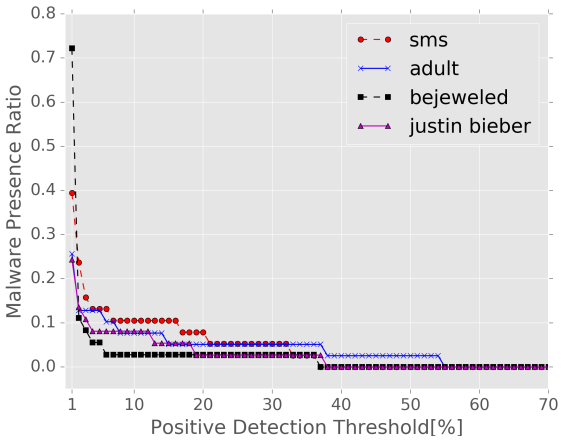

Figure 3: Malware Presence Ratio for Google Play for Top Keywords

results imply that the risk of encountering malware is very unevenly distributed across the market.

4.2 Malware Download Ratio

In the MPR metric, all apps are treated equal–also the ones lurking in the margins of the store, where users might barely engage with them. This is why a second metric is valuable: malware downloads as a fraction of all downloads. We define the *Malware Download Ratio MDR (m, s, M)* as follows:

$$MDR(m,S,M) = \frac{\sum_{a \in M'} DL(a)}{\sum_{a \in M} DL(a)} \quad (3)$$

Here, *DL (a)* is the total number of downloads from the last update of the app a. Also, *M'* is a set of all apps in the app set M which are determined as malware by the threshold m. Namely, $M' = \{a \in M | PDR(S,a) > m\} \subset M$. Figure 4 shows how MDRs vary across PDRs in dif-

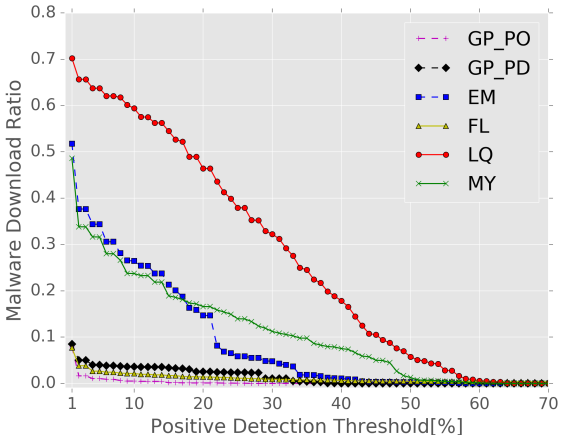

Figure 4: Malware Download Ratio for Each Market

ferent markets. In FL, the number of downloads is unpublished, which means it is excluded from this comparison. For reference, we calculated the index on the assumption that each app was downloaded same times.

Even a cursory glance at Figure 4 tells us that the mitigation practices of LQ, MY, and EM look insufficient. With the positive detection threshold of 30%, respectively 32.2%, 11.2%, and 4.86% of the total downloads in the three markets (LQ, MY, and EM) are considered malicious. LQ is especially poor in terms of mitigation performance as measured by the number of downloads. The fact that the MDR of these markets is lower than their MPR means that popular apps are less likely to contain malware. It is one thing to get malware into the market, it is another to get users to download it.

Similar to what we can found for MPR, CP and FL seem to be doing much better than the other three with low MDR at any positive detection threshold.

4.3 Malware Survival Period

The *survival period* of an app is the time that an app flagged as malware has been present in the app store. To calculate a survival period, we count the number of days between the date when we downloaded the app and the date when it was last updated.

For a market with active mitigation, one would expect malicious apps to be removed from the store and removed faster if more AV vendors are identifying an app as malicious. To put it differently, the higher the positive detection ratio of an app, the shorter its survival time is expected to be.

Figures 5 to 10 show the survival times for apps in the evaluated markets. Each dot is an app. For context, we also visualize the total number of apps submitted to the store in the same period as the malware, as well as the average PDR for that period. These trend lines help to understand the vertical bands of apps with similar survival times: at those moments, a large number of apps were added to the store. Sometimes these were predominantly malicious apps, suggesting a campaign by the attackers. In other cases, the bands are caused by an overall high submission rate of new apps, which also brings a certain portion of malware into the store.

The figures show that all app stores contain long-lived malware, also Google Play. From manually inspecting the labels assigned by the AV vendors, we saw that some of these apps might be adware. Ideally, we would like to separate adware from other malware. In practice, the labels are too inconsistent to allow this. To illustrate: eoeMarket has 1670 apps with a PDR¿30%. Of these, 1267 apps have at least one adware label, but just 42 of them have more than half of the labels indicating adware. The same patterns holds for other the markets. In other words, the bulk contains mixed labels for adware and malware. This might reflect inconsistencies in detection and labeling, but also the fact that some apps technically combine, for example, Trojan and adware functionality.

To cope with the inconsistent labels, while still exploring the ratio of adware to other malware, we apply a color gradient from dark blue (only malware, no adware labels) to white (more than 25% labels are adware).

Google Play. For GP_PO and GP_PD market, we see the expected pattern (Figure 5 and 6). Apps with higher positive detection ratios are concentrated on the left side – which means that they have relatively short survival times. On the other hand, we also see a rather surprising cluster of apps with detection ratios of 20 to 50 percent that survive for hundreds of days in the app store. How is this possible if Google does active malware mitigation?

A significant portion of the apps with high PDR and long survival times are flagged as some form of adware. This label covers a variety of practices. Some apps aggressively push advertisements to the user. A notorious example is Airpush. It embed advertisements in the notification bar of the device, among other places. Many users are annoyed by this behavior, but cannot identify which app is causing it. In response, a developer has started offering an app called Airpush Detector to remediate the issue. It was installed between 5-10 million times.

More nefarious in-app ads sometimes pose as system prompts, trying to trick the user to install other apps, for example by claiming their Whatsapp has 'expired' and leading them to whatsapap.com to install an 'update'.

Some of the adware were labeled as Potentially Unwanted Applications (PUA), signaling that it is dependent on user preferences whether this app displays unacceptable behavior or not. For example, apps might overreach in terms of permissions. An app called 'Battery Improve' claims to help maximizes battery usage, but it also requires permission to collect a variety of personal and device identifiers that are not required for its purported functionality and might expose the user to spam or other unwanted consequences.

We suspect that many of these apps operate within the Terms of Service of the store, which would mean Google has no clear grounds for removing them, even if they would want to. To the extent that the apps display nefarious behavior, they often rely on social engineering, rather than malicious code. In short: these apps operate

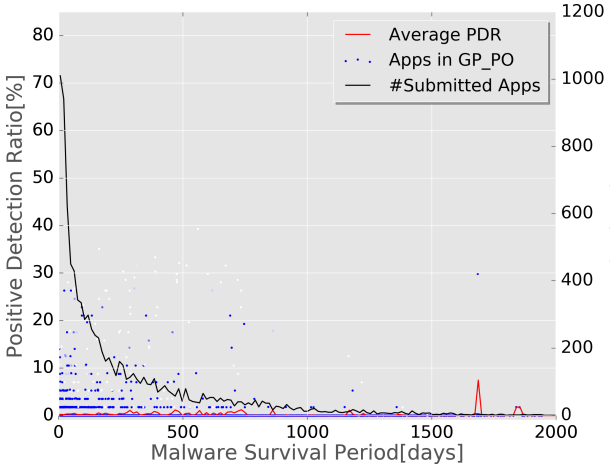

Figure 5: Positive Detection Ratio and Survival Period for Popular Apps of Google Play(*GP_PO*)

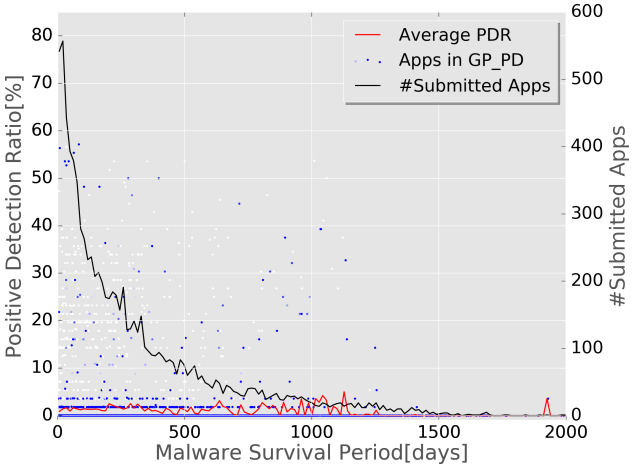

Figure 6: Positive Detection Ratio and Survival Period for Randomly Selected Apps of Google Play(*GP_PD*)

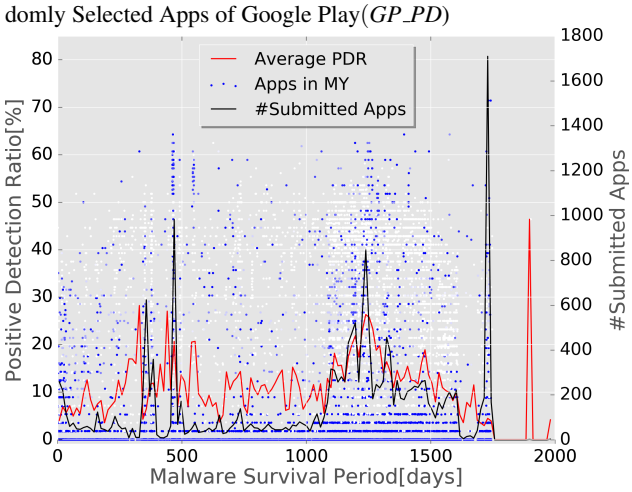

Figure 7: Positive Detection Ratio and Survival Period for Mumayi

in a grey zone in terms of security.

The fact that these apps are flagged by AV vendors point to a divergence in incentives between the vendors, the users and the market operators. Ads, including third-party advertisement libraries and networks, are a core part of the app market and the different revenue models it supports. The market operator has an incentive to allow and even enable advertisement-related revenue for developers who operate within the terms of service. The AV vendors, on the other hand, have an incentive to be more strict in evaluating apps, since their revenue depends on users paying for the added value of using a dedicated AV solution beyond the default security measures of the market. The users operate under information asymmetry *viz a viz* the app developers and markets: theoretically they could evaluate carefully the permissions and risks of every app they install, but in practice this is very difficult. Their incentive is to outsource this problem to the AV vendors, which reinforces the incentives of the latter to also flag apps as malicious that behave differently than users expect or prefer, rather than pose actual security risks.

A more puzzling, and potentially alarming, finding is that there is also a cluster of long-lived malware that is not adware. We can only speculate as to the cause of this. Perhaps Google relies exclusively on its own detection techniques which have not picked up on these apps. Perhaps the AV vendors detected these apps as malicious well after they have been submitted to the store and checked by Google. In these cases, it might make sense to leverage the signals from VirusTotal, which is actually a subsidiary of Google. On the other hand, these apps could also be false positives by the AV vendors, perhaps relying on similar detection heuristics. Even in combination, AV products have been found to produce false positives [12].

Alternative Markets. Figures 7 to 10 show the survival times for apps in the alternative markets. The alternative markets show a dramatically different picture than the Play store: apps with a high PDR stay in the store over long periods, often years.

In MY (Figure 7), one can clearly see the constant submission of apps between 1000 to 1600 days before our study with a relatively high average PDR. During this period, many malicious apps seem to have been added and then left untouched by the market for nearly three years. Also after that period, apps with a high PDR have been constantly added and we cannot see any sign of the market acting against them.

In LQ (Figure 8), there is higher density of apps with a survival period of less than 782 days compared to more than 782 days, as app submissions increased after that moment. The average PDR for period since then is around 20%, implying that a considerable portion of the apps submitted to the market during the period is malicious. Similar to MY, we do not see any sign of the market removing malicious apps.

Figure (Figure 9) suggest that EM has suffered two bulk malware submissions at about 250 days and 300 days before the study. The submitted apps have an average PDR of nearly 50%, showing the bulk of the submissions was malicious. As in the other markets, it is

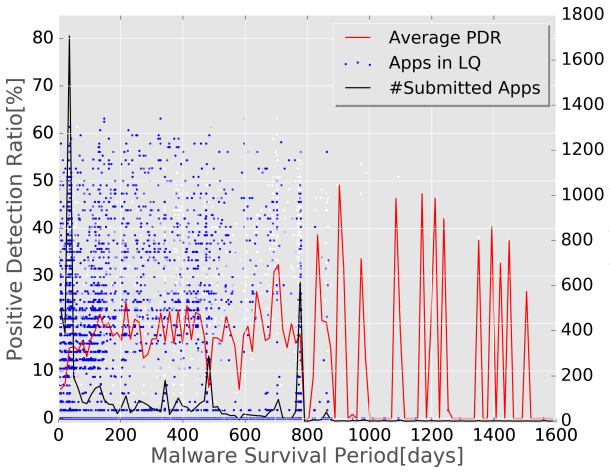

Figure 8: Positive Detection Ratio and Survival Period for Liqucn

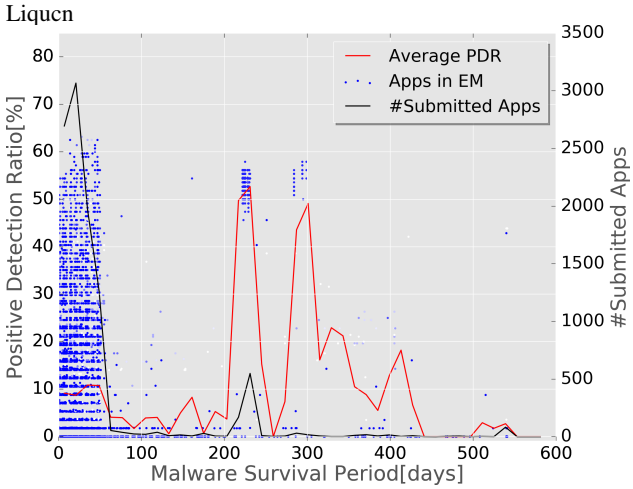

Figure 9: Positive Detection Ratio and Survival Period for Eoe-Market

clear that apps with very high PDR are left untouched in the market.

The blank period of FL (Figure 10) is caused by errors in the operation of the market, which blocked app developers from submitting their apps. The market contains apps with very high PDR on a long-term basis. In terms of MPR and MDR, FL comes close to the strong rates of GP_PO and GP_PD. Looking at the malware survival times, it seems that the low MPR and MDR stem from the fact that the market was not targeted by attackers, rather than by active mitigation by the market operator.

5 Limitations

To properly evaluate our results, we discuss the main limitations of our approach. First, and foremost, the internal validity of our results depend on VirusTotal – that is, on the detection result of AV products – to assess the maliciousness of apps. Weaknesses in detection methods of vendors undermines the degree in which the metrics approximate true malware levels.

Second, we simply counted the number of positive detections and did not delve deeply into the labels describing the purported malicious behavior, except for the anal-

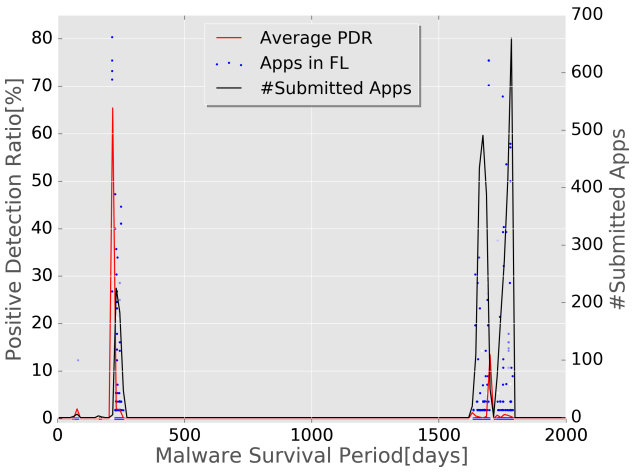

Figure 10: Positive Detection Ratio and Survival Period for Freeware Lovers

ysis of apps with a high PDR in Google Play (Section 4.3). Even this partial analysis of detection results revealed that malware is not a black-and-white category. Lurking underneath the AV detections is a set of policy decisions that merit further investigation. When we look at the Windows operating system, which has been dealing with malware for a much longer time, we can see how boundaries of what is considered malware shift over time. Recently, the Microsoft Malware Protection Center has updated its policies to decide when adware or, more neutrally, advertising programs, will be removed from the platform [16]. The idea is that to protect users, it is not enough to check whether the apps were installed with the user's consent. the platform will now be policed more aggressively. On other other end of the spectrum, it might be informative to study mitigation performance of markets against the most malicious apps, such as banking Trojans and ransomware.

Third, we do not know at what moment in time an AV vendor flagged an app as malware. This might well long be after submission to the market. We cannot expect that all malware can be detected by the AV at submission time. Furthermore, even if some AV solutions flag them as malicious, that might not be a reliable enough signal to bar entry to the market. That being said, the long survival times we have observed for some malware suggests that later detections can be leveraged more effectively for mitigation.

A fourth limitation concerns the external validity: to what extent are the app markets we studied representative for the Android ecosystem? For Google Play, the market leader, this is not an issue. The situation is more complicated for the alternative markets. As discussed in Sect. 3.1, we chose the four dominant third-party markets in 2013 as identified by F-Secure. Since that time, these markets have been overtaken by competitors. The fact that these evaluated markets became less competitive might have introduced a bias, as their lack of mitigation of malware could be caused by the fact their the owners no longer make serious investments in the market. In future research, we hope to compare these markets against ones that are currently successful.

6 Conclusion

Our analysis found that measurements of malware presence and download ratios are very sensitive to detection rates across AV products. This is not just a measurement problem, but also points to difficulty of developing mitigation policies by app market operators. What can we reasonably expect market operators to do? If they are overly zealous, they will remove benign apps and impact the livelihood of innocent developers. If they are overly cautious, they will expose their users to significant amounts of malware. There is currently only a rough understanding of where to strike that balance when it comes to using AV detections as signals for mitigation.

Based on heuristics, we argued that a detection of 30% is a reasonable approximation of when AV detections combine into a reliable signal. At a PDR of 30%, malware presence and download ratios are relatively low for GP_PO, GP_PD and FL, though the risk is unevenly distributed within the store. When users of Google Play search with specific keywords, they may have a 50 times higher probability of encountering malware than when selecting from the popular app list. For the three other app stores, EM, MY and LQ, the ratios are alarmingly high. Around one in five apps are flagged as malicious.

When it comes to evaluating market operators, the analysis of survival times of malicious apps has found very clear patterns that are robust to different positive detection rates. EM, MY and LQ do not seem to act at all against malware, even when PDRs are very high. Clearly malicious apps stay in the store over very long time frames, sometimes years.

Google Play is the only market that seems to conduct active malware removal, though we did find an intriguing cluster of adware and malware that has persisted over time. This cluster points to different incentives for market operators, AV vendors and users.

Notwithstanding the limitations discussed above, the metrics we defined have shown to be a practical means for evaluating malware mitigation by market operators. Repeated measurements, expanded across a wider range of app stores, will tell us if app stores are keeping up with the dynamic threat of mobile malware.

References

[1] Androguard. https://github.com/androguard/androguard.

[2] App annie app store stats google play top app charts united states overall. https://www.appannie.com/apps/google-play/top/.

[3] eoemarket. http://www.eoemarket.com/.

[4] Freeware lovers. http://www.freewarelovers.com/android/.

[5] Juniper networks third annual mobile threats report march 2012 through march 2013. http://www.juniper.net/us/en/local/pdf/additional-resources/jnpr-2012-mobile-threats-report.pdf/.

[6] Liqucn. http://www.liqucn.com/.

[7] The most popular app store keywords from chomp & google play. http://searchengineland.com/the-most-popular-app-store-keywords-from-chomp-google-play-135744.

[8] Mumayi. http://www.mumayi.com/.

[9] The top 10 android app stores in china 2015. http://technode.com/2015/09/22/ten-best-android-app-stores-china/.

[10] Understanding chinas mobile app system. http://theappentrepreneur.com/chinas-mobile-app-system.

[11] ASOKAN, N. On mobile malware infections. In *Proceedings of the 2014 ACM conference on Security and privacy in wireless & mobile networks* (2014), ACM, pp. 37–38.

[12] CHAKRADEO, S., REAVES, B., TRAYNOR, P., AND ENCK, W. Mast: triage for market-scale mobile malware analysis. In *Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks* (2013), ACM, pp. 13–24.

[13] F-SECURE. Threat report h2 2013. https://www.f-secure.com/documents/996508/1030743/Threat_Report_H2_2013.pdf.

[14] GOOGLE. Android security 2015 year in review. http://source.android.com/security/reports/Google_Android_Security_2015_Report_Final.pdf.

[15] JAKEJ. Playdrone apk's. https://archive.org/details/playdrone-apks/.

[16] MICROSOFT. Adware: A new approach. https://blogs.technet.microsoft.com/mmpc/2014/04/02/adware-a-new-approach/.

[17] SIMILARWEBPRO. Google play top 20 popular keywords in apr 2015. http://www.similar-web.jp/blog/archives/1004.

[18] SYMANTEC. Malicious websites by search term. http://securityresponse.symantec.com/threatreport/topic.jsp?id=threat_activity_trends&aid=malicious_websites_by_search_term.

[19] TRUONG, H. T. T., LAGERSPETZ, E., NURMI, P., OLINER, A. J., TARKOMA, S., ASOKAN, N., AND BHATTACHARYA, S. The company you keep: Mobile malware infection rates and inexpensive risk indicators. In *Proceedings of the 23rd international conference on World wide web* (2014), ACM, pp. 39–50.

[20] VIDAS, T., AND CHRISTIN, N. Sweetening android lemon markets: measuring and combating malware in application marketplaces. In *Proceedings of the third ACM conference on Data and application security and privacy* (2013), ACM, pp. 197–208.

[21] VIENNOT, N., GARCIA, E., AND NIEH, J. A measurement study of google play. In *ACM SIGMETRICS Performance Evaluation Review* (2014), vol. 42, ACM, pp. 221–233.

[22] ZHOU, W., ZHOU, Y., JIANG, X., AND NING, P. Detecting repackaged smartphone applications in third-party android marketplaces. In *Proceedings of the second ACM conference on Data and Application Security and Privacy* (2012), ACM, pp. 317–326.

[23] ZHOU, Y., AND JIANG, X. Dissecting android malware: Characterization and evolution. In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 95–109.

[24] ZHOU, Y., WANG, Z., ZHOU, W., AND JIANG, X. Hey, you, get off of my market: Detecting malicious apps in official and alternative android markets. In *NDSS* (2012).