Do You Get What You Pay For? Comparing The Privacy Behaviors of Free vs. Paid Apps

Catherine Han
UC Berkeley
catherinehan@berkeley.edu

Álvaro Feal
IMDEA Networks Institute
alvaro.feal@imdea.org

Irwin Reyes
ICSI
ioreyes@icsi.@berkeley.edu

Kenneth A. Bamberger
UC Berkeley
kbamberger@berkeley.edu

Amit Elazari Bar On
UC Berkeley
amit.elazari@berkeley.edu

Serge Egelman
ICSI / UC Berkeley
egelman@cs.berkeley.edu

Joel Reardon
University of Calgary
joel.reardon@ucalgary.ca

Narseo Vallina-Rodriguez
IMDEA Networks Institute / ICSI
narseo@icsi.berkeley.edu

Abstract

It is commonly assumed that the availability of *"free"* mobile apps comes at the cost of consumer privacy, and that paying for apps could offer consumers protection from behavioral advertising and long-term tracking. This work empirically evaluates the validity of this assumption by investigating the degree to which *"free"* apps and their paid premium versions differ in their data collection behaviors and privacy practices. We compare pairs of free and paid apps using a combination of static and dynamic analysis, and examine the differences in the privacy policies within pairs. We analyzed 1,505 pairs of free Android apps and their paid counterparts. Our results show that there is no clear evidence that paying for an app will guarantee protection from extensive data collection. Specifically, 48% of the paid versions reused all of the same third-party libraries as their free versions, while 56% of the paid versions inherited all of the free versions' Android permissions to access sensitive device resources. Additionally, our dynamic analysis reveals that 38% of the paid apps exhibit all of the same data collection and transmission behaviors as their free counterparts, and less than 1% of the pairs have policies that differ between free and paid versions.

Introduction

In the advent of mobile applications, it has become apparent that users often trade their privacy for *"free"* apps [2].

The question, however, remains unanswered for paid apps—are consumers of paid apps truly safe from extensive user profiling and tracking? Users paying for apps expect them to be of higher quality compared to free versions [5], and a common selling point to that end is the removal of ads in paid versions. However, even if an app does not display ads, it may still perform invasive tracking for the purpose of serving highly-targeted ads in *other* apps.

Exploring if app behaviors comport with user expectations and if *"ad-free"* representations may be misleading consumers can inform regulators, policymakers, and consumers alike. Potentially misleading representations may run afoul of the FTC's prohibitions against deceptive practices and state laws prohibiting unfair business practices, as well as general privacy regulations, such as the GDPR and CCPA. Finally, such inquiry can also inform economic models exploring the viability of "pay for privacy" consumer protection models [1].

To that end, we explore the differences and similarities in the implementation and data collection practices of free Android apps and their paid counterparts offered on the Google Play Store, across 1,505 pairs of apps. On average, at least 10,000 users have installed each pair of apps. We measured their prospective differences along three key aspects: different third-party libraries—which may be used for advertising and tracking—bundled with the apps, the nature of the permissions they access, and the types of sensitive data shared with third-party services.

Methodology

In this analysis, we generalize different app monetization models into two overarching categories: we define "free apps" as those that are available for download on the app store at no up-front cost; and we define "paid apps" as apps that require a one-time payment to download.

App Corpus

The Google Play Store does not reliably link free apps to their paid versions, or even indicate if a corresponding paid version exists at all. Therefore, we created a labeling task on Amazon Mechanical Turk to construct our corpus. We presented workers with a free app and a list of all paid apps from the same developer. If the free app did not have a corresponding paid version, workers were instructed to select the "None" option. We presented each free app to three different workers, then manually adjudicated the responses for agreement and correctness. Workers were paid $0.10 for each match in consensus with the others. This process yielded 1,505 pairs of free apps and their paid counterparts.[1]

Evaluating Apps

We looked for similarities across pairs of free and paid apps along three dimensions: (1) the portion of third-party packages found in the free app that are also included in the paid version; (2) the portion of Android permissions declared by the free app also declared by the paid app; and (3) the portion of sensitive network transmissions performed by the free app also seen in the paid app. We believe these four aspects are a good representation of apps' data collection and sharing behaviors. We employed the following methods to evaluate these:

Dynamic Analysis: We used dynamic analysis methods derived from earlier work [3] to automatically evaluate apps by executing them in an instrumented environment (deployed on identical Nexus 5X smartphones) that captures apps' network traffic. We attempted to control for differ-

[1] https://github.com/io-reyes/play-store-purchase/blob/master/data/pairs-conpro.csv

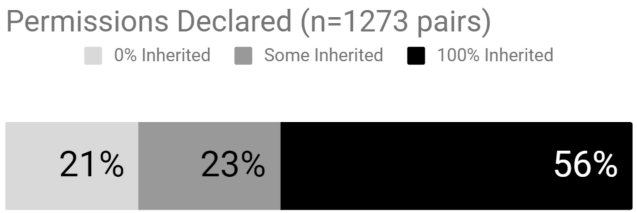

Figure 1: Frequency of Android permissions inherited between free/paid pairs, where the free app requested at least one Android permission.

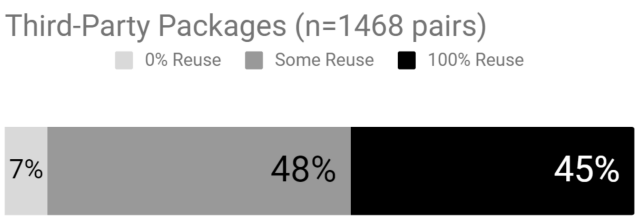

Figure 2: Frequency of third-party package reuse among free/paid pairs, where the free app had at least one third-party package.

ences in app execution by providing both apps with the same random input stream at the same time.

At the end of each paired execution, we analyzed the captured network data to identify which sensitive data types were sent to which remote services—services that could be for ads, profiling, crash reporting, etc. We focused on detecting the transmission of sensitive data that can be used to uniquely track a user over time and across different services: persistent identifiers, such as the Android Advertising ID (AAID), IMEI, and Wi-Fi MAC address; as well as personally identifiable information (PII).

Analysis

This work focuses on measurable differences in privacy between free and paid versions, so all presented comparisons are conditioned on the free app having at least one observation for any of the corresponding metrics.

Declared Android Permissions

The Android permission system serves to protect user privacy. Apps must hold appropriate permissions to use various device resources (Internet access and information about the device) and access sensitive user data (phone number).

Of the 1,505 pairs in our corpus, 1,273 had free versions that declare at least one Android permission (either regular permissions or “dangerous”). In 56% of these pairs, the paid version (Figure 1) declared all of the same permissions held by the free version. That is, paid apps held all the same privileges as free versions in a majority of the time that any permissions are declared. The most common permissions that both the paid and free versions requested were the ones that gave access to network state, Internet, and writing to external storage.

Bundled Third-Party Packages

The use of third-party code is common practice in software engineering to expedite development. In mobile apps, third-party libraries allow for pre-built functionality like graphics rendering, advertising, and analytics, among others.

Of the 1,505 pairs in our corpus, 1,468 had at least one third-party package in the free version. Of these (Figure 2), we observed that 45% of paid apps contained the same third-party libraries as the free versions, while 7% of paid apps showed no third-party libraries carried over from the free version. Although we acknowledge that our analysis does not account for third-party libraries included but not

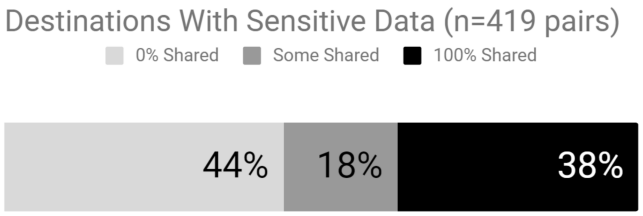

Figure 3: Frequency of unique domain destinations shared between free/paid pairs, where the free app transmitted sensitive data to at least one domain.

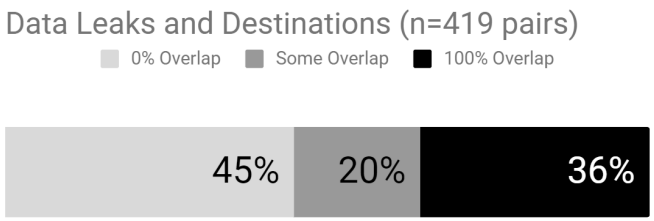

Figure 4: Frequency of unique sensitive data type-domain pairs shared between free/paid pairs of apps, where the free app transmitted sensitive data to at least one remote domain.

actually executed (i.e., dead code), these results show that developers may leave paying consumers exposed to the same potential for third-party data collection as found in free apps.

Based upon the library categorizations of LibRadar[4], we analyzed the types of third-party libraries present in free and paid versions of apps, focusing our attention on libraries associated with libraries labeled as “Advertising” and “Analytics.”

Focusing on advertising libraries specifically, LibRadar detected at least one ad library present in either the free or paid release (or both) in 831 pairs. Of those, there were 802 free apps where ad libraries were detected, while only 408 paid apps were found to contain ad libraries. This suggests that ad libraries are most likely present in either free versions of apps exclusively, or to a lesser extent, in both the free and paid versions of an app.

Network Transmissions

Third-party services bundled in apps routinely collect various data from users and their devices. For example, advertising networks collect persistent identifiers and personal information to better target users with ads relevant to them.

By observing all of the network traffic associated with an app, we can discern the types of sensitive data being transmitted and the recipient of that data.

Among the 1,505 pairs of apps that we examined, there were 419 pairs in which the free version transmitted sensitive data to online services over the Internet. Out of these 419 pairs, we observed that 44% of these pairs’ paid versions (Figure 3) did not communicate with any of the domains that the free version did, while 18% shared some destinations with the free version. Conversely, 38% of these pairs’ paid versions communicated with all of the same domains as the free version.

Conclusion

This paper presents a multi-dimensional analysis of the measurable benefits that consumers can expect to receive when paying for an app by employing both static and dynamic analysis, uniquely performing a large-scale, one-to-one comparison between a free version of an app and its paid counterpart. Our preliminary results show that the privacy benefits of paying for apps are tenuous at best, and are likely to mislead consumers, making it impossible for them to make informed decisions about their privacy.

Acknowledgements

This work was supported by the U.S. National Security Agency's Science of Security program (contract H98230-18-D-0006), the Department of Homeland Security (contract FA8750-18-2-0096), the National Science Foundation (grants CNS-1817248 and CNS-1564329), the European Union's Horizon 2020 Innovation Action program (grant Agreement No. 786741, SMOOTH Project), the Rose Foundation, the Data Transparency Lab, and the Center for Long-Term Cybersecurity at U.C. Berkeley.

REFERENCES

1. Amina Wagner, Nora Wessels, Peter Buxmann, Hanna Krasnova. 2018. Putting a Price Tag on Personal Information - A Literature Review. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.
2. Brian X. Chen. 2017. How to Protect Your Privacy as More Apps Harvest Your Data. (2017). https://web.archive.org/web/20190622220211/https://www.nytimes.com/2017/05/03/technology/personaltech/how-to-protect-your-privacy-as-more-apps-harvest-your-data.html
3. Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, Serge Egelman. 2018. "Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale. In *Proceedings on the 2018 Privacy Enhancing Technologies (PET2018)*. 63–83.
4. Ziang Ma, Haoyu Wang, Yao Guo, and Xiangqun Chen. 2016. LibRadar: fast and accurate detection of third-party libraries in Android apps. In *Proceedings of the 38th international conference on software engineering companion*. ACM, 653–656.
5. Matthew Panzarino. 2011. Why you should want to pay for apps. (2011). https://web.archive.org/web/20181129005820/https://thenextweb.com/apps/2011/04/24/why-you-should-want-to-pay-for-apps/