# Balancing Data Collection and User Privacy in the Mobile Environment

David Choffnes
*Khoury College of Computer Sciences, Northeastern University*
choffnes@ccs.neu.edu

2019

## INTRODUCTION

There has been a dramatic shift toward using mobile devices such as smartphones and tablets as the primary interface to access Internet services. Unlike their fixed-line counterparts, these devices also offer ubiquitous mobile connectivity via WiFi and cellular data access, and are equipped with a wide array of sensors (e.g., GPS, camera, and microphone).

The combination of rich sensors and ubiquitous connectivity make these devices ideal for continuous scientific data collection. Unlike their desktop and laptop counterparts, mobile devices contain more sensors and can more easily collect/transmit data about the world around them even while their people are not actively using the devices (e.g., they are in pockets or purses). For instance, there are myriad apps that can detect how often a user moves, which locations they visit, and what is their current heart rate.

While such information is valuable for scientific and medical reasons, there are significant concerns associated with gathering personally identifiable information (PII) and/or personal health information (PHI). In fact, we have already seen how commercial apps extensively track users and share their PII with other parties over the Internet (Ren et al. 2016; Ren et al. 2018), and even though users have in many cases consented to some form of data collection, they are generally unaware of this behavior (Consolvo et al. 2010). Worse, there have been substantial breaches of databases containing extremely sensitive PII, including those by Facebook, Equifax, Google (Plus), and others. In short, even the companies we trust most with our data can be hacked, and thus one can expect the same will happen to any sufficiently interesting store of sensitive information given a long enough time.

In this white paper, I briefly explore these potential benefits and risks of data collection in the mobile environment. As a researcher whose research projects (Choffnes and Bustamante 2008; Nikravesh et al. 2015; Kakhki et al. 2015) have gathered data from more than one million users (including more than 100,000 mobile users) and who has studied privacy leaks from commercial services (Ren et al. 2016; Ren et al. 2018; Pan et al. 2018; Leung et al. 2016), I bring to the table my experience, reasoning, and guidelines for such data collection and risk mitigation.

## DATA COLLECTION FROM MOBILE DEVICES

There are several ways to access information from mobile devices, and each platform imposes different constraints on whether such data is available, and if so, what whether user permission is required to access the data. Note that this list is intended to cover common ways of gathering data from mobile devices, but is not an endorsement of using any such techniques. In the next section I will discuss the relevant privacy concerns and mitigation strategies.

### *Identifiers and Time*

As with any data-collection process, it is important to be able to link records to the same individual over time. For mobile platforms, this can be done using device- or app-unique identifiers and timestamps, neither of which require any permissions.

### *Internet access*

To collect data, it is generally necessary to transmit it over the Internet. Doing so generally does not require user permission on iOS, but it does require permission on Android.

### *Sensor data*

Android and iOS place different restriction on what data from sensors (e.g., GPS, network status) can be accessed. On Android, for example. reading the network state (e.g., what network the device is connected to) is considered a "normal" permission, while GPS is considered a "dangerous" permission[1] because it can be used to violate users' privacy and potentially cause harm (e.g., stalking). Other similarly "dangerous" sensor permissions include body sensors, microphone and camera. Note that on iOS, access to sensors such as GPS, camera, microphone, and body sensors are similarly restricted.

### *User activity monitoring*

When it comes to identifying which apps have been installed by a user, or what content they are visiting, most OSes impose strict constraints. There are several levels at which this can be done:

**List of installed apps.** On Android it is possible to get a list of installed apps on a device without any permission from the user, while on iOS it is not allowed at all.

**Sniffing app activity.** On Android, any app can receive any broadcasted `Intent`, unless the broadcasting app imposes restrictions. This can be used to learn whether certain apps are used, and whether certain activities are occurring, but it applies only to those cases where activities use broadcast. As such, this approach can miss an unbounded amount of activity. This approach is not available on iOS.

**Monitoring network traffic.** All mobile OSes support the ability to redirect network traffic via a virtual private network (VPN). While generally intended for end users to access resources in protected (and private) networks, a VPN connection can also be used as a proxy for redirecting network traffic to the public Internet. As such, many approaches (including my ReCon project (Ren et al. 2016)) use VPN

[1] https://developer.android.com/guide/topics/permissions/overview#permission-groups

proxies to capture network traffic from a device without needing any special permissions. Instead, the user must install a VPN configuration that directs their network traffic to a proxy, e.g., a server running in a lab or in the cloud.

An alternative solution is to use Android's `VPNService` or iOS's `Network Extension` feature to capture network traffic on the device. These features are designed for accessing private networks, but again can be used for network monitoring. Both of these options require the equivalent of "dangerous" permissions, and thus require users to allow the behavior via a dialog that indicates doing so is risky.

Capturing network traffic allows you to see any activities from apps that entail accessing the Internet. As such it does not help with apps that do not use the Internet. Further, these approaches do not allow you to easily tell which apps are installed nor what content was visited. For the former, mapping network traffic to the app that generated it is an open problem, though such "app fingerprinting" is an active area of research. For the latter, it is important to note that increasing fractions of network traffic are encrypted, meaning the contents of that traffic are hidden from the observer. The unencrypted clues that remain (e.g., DNS lookups, Server Name Indicator in the TLS handshake) cannot in general uniquely identify the corresponding app or website, though this is certainly an area of research.

In certain cases, one can ask the user to install a self-signed root certificate that will enable decryption of such traffic (using a man-in-the-middle, or MITM, approach), but this potentially exposes all of a users' sensitive data to the observer, may permit malicious communication by weakening TLS security guarantees, and might cause apps to stop working (e.g., Facebook will not accept connections that have been MITMed). For all these reasons, MITM should be avoided on research subjects (but is reasonable to use for lab testing on study personnel).

**Building a custom web browser.** To monitor web browsing activity, one can simply build on top of existing open-source browsers such as Chromium and Firefox. When properly instrumented, such browsers can gather data about all content visited by a user. Of course, these browsing histories are generally considered very sensitive information, so it would be essential to provide proper consent and safeguards to protect this information.

**Screen monitoring.** In general, it is not possible for researchers to record the entire mobile device screen while a user interacts with different apps. While such screen capture permissions exist, they are tightly safeguarded by Android and iOS and generally never given to third parties.

Recently, my team discovered that app developers *can record the screen of the app(s) they develop*, without requiring any permission or consent from the user (Pan et al. 2018). This is true both on iOS and Android. While recent changes to the Android terms of service seem to bar such screen recording in some cases, there is at the moment no way for Android or iOS to prevent it using OS permissions.

**Custom OS.** Many of the privacy/security constraints that prevent user monitoring can be evaded by using a custom version of the Android, built off their open source code. However, doing so poses numerous challenges to deployment, including device compatibility and difficulty for users to install (potentially leading to bricking their devices).

## PRIVACY RISKS AND MITIGATION

There is a clear tension between the desire to collect as much data as possible pertinent to an investigation, and protecting user privacy. On the one hand, one may wish to collect detailed information to answer research questions, casting a wide net of data collection to avoid realizing at analysis time that critical information had been omitted. On the other hand, users want to be sure that only the data that is necessary is collected, that it cannot be used to identify them, and that it does not fall into the wrong hands.

The techniques described in the previous section open the door to a number of threats to user privacy that warrant careful attention. Below I discuss several such considerations.

**Anonymity.** As is standard practice with human subjects research, individual identifiers that are not based on any of a subject's PII should be used to link multiple data records to the same subject. As such, it is best not to use identifiers like a mobile device's IMEI or advertising ID because these can be used by other parties to link to the individual. Instead, the identifier should be randomly generated by the study personnel in a way that does not take any PII as input.

While such identifiers are necessary to protect subjects' identities, they are not sufficient. Other data can uniquely identify subjects, such as a series of timestamps indicating when a device was active, and geolocations of the device over time. Taken together, locations with timestamps can be used to identify an individual's place of residence, work, and other points of interest that are unique to the individual. To mitigate this issue, it is common to coarsen the granularity of data or omit either a location or timestamp, so that multiple individuals share the same properties (e.g., to provide $k$-anonymity).

Other metadata can impact anonymity. For example, the user-agent string generated by mobile browsers may uniquely identify individuals by their devices, as might the list of apps installed. As such, the study personnel should monitor the uniqueness of such data as it is collected, and transform the data as needed to provide anonymity.

It should be noted that there are more sophisticated approaches to providing anonymity. For example, one can add noise to samples in a way that prevents an adversary from identifying the true value for an individual, but that allows the researcher to accurately calculate (true) aggregate statistics over those noisy samples.

**Network traffic collection.** Gathering full network traces from subjects' devices poses enormous privacy and security risks. As part of my team's prior research, we found that network traffic contains extremely sensitive information such as locations, relationship status, and even usernames and passwords—*even in some cases unencrypted*. Thus, for our study we opted to collect only a limit number of bytes at the beginning of a flow, enough to identify whether sensitive data was in network traffic, and we modified those flows to delete detected personal information. In general, it is best to identify which parts of network traffic are necessary for a study *a priori*, then take steps to minimize network traffic collection to contain only that information.

The risks are even greater when using MITM to inspect the contents of encrypted traffic. This can reveal substantial amounts of subject PII, and depending on the implementation of MITM, can enable certain attacks on TLS (e.g., spoofing using invalid certificates) to succeed. It is strongly recommend not to use MITM on participants' network traffic when the proxy resides outside of the participant's device—the research team's proxy server then becomes a prime target for attack to steal subjects' personal data or launch any other number of attacks. While doing MITM on the device itself (e.g., using a `VPNService`) mitigates the threat to a proxy server, it still can weaken TLS security guarantees for the subject if not done carefully. Finally, note that such decryption of data, when used on messaging applications, can violate federal laws against wiretapping and two-party consent.

**Web browsing histories.** The set of web pages, or even the set of domains, that a subject visits is also incredibly sensitive. Such online activity can reveal a great deal about a subject, including interests, socioeconomic status, medical conditions, and more. In addition, it may consist of embarrassing or even illegal online activity. To help mitigate this issue, it is helpful to whitelist a set of categories of websites that are of interest to the research team, and only track categories instead of specific sites.

**Screen monitoring.** Recording how a user interacts with a researcher's app can pose few privacy risks if the app does not entail interaction with, or inputting, PII. If that is not the case, such monitoring is a form of surveillance that can reveal a great deal about an individual. It is generally not recommended to use this approach unless the user's information is protected.

**Practical concerns.** Outside of privacy and security, there are number of other practical concerns when collecting data from mobile devices. It is important that the data collection does not consume so much power that the subject's battery drains fast enough to empty before the typical nightly recharge. Even draining battery faster than usual can cause anxiety. Many users have a limited number of bytes that they can transmit per month over a cellular connection. Thus, transmitting data to collection servers is best done over Wifi where such constraints are either absent or less of an issue. Any data collection should be done off the critical path for any of a device's/app's functionality; in other words, any failures or errors from data collection should not impact the user experience.

## CONCLUSION

In this white paper, I discussed a number of ways that information can be collected about subjects for research studies, and how these may pose substantial privacy and security risks to those subjects. I proposed a number of mitigation strategies and other considerations for such data collection. Note that this document is not intended to be a complete treatment of the topic, but rather focus on a summary of key concerns for discussion.

# References

Choffnes, David R., and Fabián E. Bustamante. 2008. "Taming the Torrent: A Practical Approach to Reducing Cross-ISP Traffic in Peer-to-Peer Systems." In *Proc. of ACM SIGCOMM*.

Consolvo, Sunny, Jaeyeon Jung, Ben Greenstein, Pauline Powledge, Gabriel Maganis, and Daniel Avrahami. 2010. "The Wi-Fi Privacy Ticker: Improving Awareness & Control of Personal Information Exposure on Wi-Fi." In *Proc. of UbiComp*.

Kakhki, Arash Molavi, Abbas Razaghpanah, Anke Li, Hyungjoon Koo, Rajesh Golani, David R. Choffnes, Phillipa Gill, and Alan Mislove. 2015. "Identifying Traffic Differentiation in Mobile Networks." In *Proc. of IMC*. doi:10.1145/2815675.2815691. http://doi.acm.org/10.1145/2815675.2815691.

Leung, Christophe, Jingjing Ren, David Choffnes, and Christo Wilson. 2016. "Should You Use the App for That? Comparing the Privacy Implications of Web- and App-Based Online Services." In *Proc. of IMC*. http://dx.doi.org/10.1145/2815675.2815691.

Nikravesh, Ashkan, Hongyi Yao, Shichang Xu, David R. Choffnes, and Zhuoqing Morley Mao. 2015. "Mobilyzer: An Open Platform for Controllable Mobile Network Measurements." In *Proc. of MobiSys*. doi:10.1145/2742647.2742670. http://doi.acm.org/10.1145/2742647.2742670.

Pan, Elleen, Jingjing Ren, Martina Lindorfer, Christo Wilson, and David Choffnes. 2018. "Panoptispy: Characterizing Audio and Video Exfiltration from Android Applications." In *Proc. of PETS*.

Ren, Jingjing, Martina Lindorfer, Daniel J. Dubois, Ashwin Rao, David R. Choffnes, and Narseo Vallina-Rodriguez. 2018. "Bug Fixes, Improvements, ... and Privacy Leaks – A Longitudinal Study of PII Leaks Across Android App Versions." In *Proc. of NDSS*.

Ren, Jingjing, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David R. Choffnes. 2016. "ReCon: Revealing and Controlling Privacy Leaks in Mobile Network Traffic." In *Proc. of MobiSys*.