Inference of User Demographics and Habits from Seemingly Benign Smartphone Sensors

Manar Safi[1], Irwin Reyes[2], Serge Egelman[1,2]

[1]UC Berkeley, [2]ICSI

ABSTRACT

Modern smartphones come equipped with a suite of high-fidelity motion and environmental sensors: accelerometers, gyroscopes, and light sensors, among others. In order to protect user privacy, the dominant mobile platforms, Android and iOS, have placed restrictions on certain device resources deemed "sensitive." Apps are required to ask the user for permission before accessing those capabilities; for instance, geolocation and phone calling functionality. These access controls don't extend to environmental and motion sensors though. These sensors are considered "benign," so apps are able to collect measurements from these data sources without informing the user of this usage.

In this ongoing work, we consider that smartphones are highly personal devices, typically used by just one person, and remains in very close proximity to – if not in direct contact with the body of – the sole user for much of the phone's operation. We explore the feasibility of using unrestricted "benign" sensor data to infer private details about the user, such as their daily schedules, income levels, and relationship status. We propose our method of sensor data collection, analysis, and validation with user-reported ground truth to build a model to infer user information. Preliminary internal testing supports the possibility of relationships between motion sensors and certain everyday activities. Further analysis is warranted to explore the general implications of these intuitions.

1. INTRODUCTION

Over the past decade, smartphones have evolved from high-end business tools and luxury goods, to commodity items accessible to most consumers. In the United States, smartphones reached nearly 80% penetration among all mobile phone subscribers by Q1 2016 [9]. Apple's iOS and Google's Android platforms dominate this market, combining to power over 95% of American smartphones [9]. As smartphone capabilities improve and ownership costs drop, these mobile devices serve as the primary portal for a number of consumer needs: communications, entertainment, news and weather, documenting life, and the like.

Part of what enables smartphones to be such versatile and intuitive tools for so many users is their suite of onboard high-fidelity environmental and motion sensors. These sensors include accelerometers to measure motion, light sensors to detect the level of ambient lighting, and magnetometers to determine a phone's orientation relative to the Earth's poles [8] [6]. Smartphones' rich sensing capabilities have allowed software developers to create innovations like augmented reality applications [10] [4] and power-conserving adaptive brightness [2].

The mobile software industry that develops these products relies on advertising revenue for their business models. As users heavily prefer free apps over paid ones, developers have implemented monetization strategies like in-app premium features and displaying targeted advertisements to specific audiences [17]. App developers that opt for in-app advertisements frequently do so through third-party advertising networks. Ad networks provide prebuilt libraries that handle ad retrieval, presentation, and user analytics. Mobile advertising revenue depends on tracking individual users' interests and backgrounds [12]. Advertising companies have been observed maintaining persistent identfiers to piece together personal profiles of consumers across services over time [3].

In order to stem potential concerns over the collection of personal information on smartphones by third parties, mobile platforms have permissions systems that regulate access to sensitive data [11] [7]. These protect highly personal data like contact lists, geolocation data, and SMS and phone call activity. Permissions systems require app developers to declare their use of personal information. Developers must request the user's approval prior to accessing that data. On older versions of Android before 6.0 "Marshmallow," users review and grant app permissions at the time of install. On iOS and current versions of Android, users are prompted for permission the first time an app attempts to access the protected resource.

Despite these robust permissions frameworks, mobile operating systems don't require user consent – or even user awareness – for an app to collect measurements

from the host device's environmental and motion sensors. Smartphones have attained majority market share among all mobile phone users in the US, and advertisers seek to target such a large audience through superior user fingerprinting and profiling. There is strong economic incentive to tap into such a rich, unrestricted, and highly personal data source as device sensors to silently infer valuable details about the user.

Moving forward, sensor access will no longer just be limited to native apps for a given mobile platform. HTML5 proposes methods for dynamic websites to access host device sensor data [1]. This opens up the possibility of increasingly personal and persistent user fingerprinting, as web applications can run as invisible elements on sites the user visits, able to collect sensor measurements without alerting the user. As websites are generally designed to be platform-agnostic, this would allow advertisers and malicious developers to collect this sensor data with greater ease, regardless of the underlying device or operating system.

2. RELATED WORK

The literature on inferring sensitive details about the user have tended to focus on discrete activities and individual sensors. Some of the existing research rely on data from peripheral devices tied to the smartphone. For instance, Liu 2015 [13] uses accelerometer data from wearable devices to build classifiers that distinguish between such activities as driving, eating, and using the phone. In the case that the user is driving, accelerometer measurements are then used to calculate which way the steering wheel is pointed, and by proxy determine if the user is driving unsafely. Parate 2014 [16] similarly leverages sensors on wearables to detect arm motions associated with smoking a cigarette. For our research, we limit our scope to sensors (and other unprivileged data sources) available on smartphones themselves, as consumers with wearable devices currently represent just a fraction of smartphone users in the US [**?**].

Other research efforts have also focused on strictly smartphone-based data sources. Owosu 2012 [15] demonstrates a proof-of-concept side-channel attack to infer short on-screen keyboard input sequences (e.g., passwords, n ames, search queries) from phone accelerometer readings; a classifier is trained on known accelerometer measurements and screen tap locations. Michalevsky 2015 [14] shows how battery voltage and current information – data provided by the phone without special restrictions – can be used to infer user location and direction; this research observes that battery activity is (in part) a function of cell network signal strength, and signal strength is sufficiently stable and predictable for a given location. Our work expands on this by exploring relationships between multiple onboard sensors and other unrestricted data (e.g., battery state and system state broadcasts), instead of just individual data sources.

Data Sources
Accelerometer
Ambient temperature
Barometric pressure
Battery state
Gyroscope
Humidity
Light sensor
Proximity sensor
Rotation vector
Screen state
Step counter
System-wide broadcasts

Table 1: "Benign" data sources and sensors recorded, where available

3. METHODOLOGY

We intend to recruit approximately 200 participants in the United States through Amazon Mechanical Turk (MTURK). In order to mitigate attrition, we target experienced MTURK workers with at least a 95% job completion rate over a history of at least 500 accepted tasks. We require participants to have fairly recent Android devices. Current models (i.e., from the past 3 years) support low-power continuous monitoring of sensors. Targeting phones with this capability ensures a rich dataset without incurring significant battery overhead. All available sensors listed in Table 1 are monitored. Per MTURK policy, all workers are at least 18 years of age.

We devise a payment schedule to increase the likelihood that subjects will actively participate through the study's 7-day data collection period. A $1 payment is issued upon completion of an online entry survey, which triggers the data collection. Additional $1 payments are available for each activity survey, offered each evening for a maximum of 6 times. At the end of the collection period, the user completes an exit survey and receives a final $2 bonus for their full participation.

3.1 Survey and location ground truth

We establish ground truth through surveys (entry, daily activity, and exit) and geolocation data. Our entry survey asks participants for their demographic information (e.g., income, age, weight, occupation, marital status, etc.). The daily surveys gather information about the participant's activities that day (e.g., working hours, physical activity, transportation, etc.). The exit survey asks the same questions as the entry form in order to validate the initial responses. We selected survey questions based on what marketers find valuable [5]

and what individuals might consider sensitive.

Geolocation data provides additional ground truth to validate inferences about work and home schedules.

3.2 Analysis

After we collect sufficient sensor and ground truth data from a diverse set of participants, we then test possible inferences one by one; some user details may be more easily inferred from sensor measurements than others. We define a successful inference as a predictive correlation between some collection of sensor measurements and a sensitive detail about the user, as seen in the ground truth.

Our proposed inferences broadly fall into two categories: what the user does (i.e., habits and activities), and who the user is (i.e., demographic details). One intuition we have is that there is a relationship between physical activity and income levels; a white-collar professional is more likely to have a sedentary lifestyle than a retail sector worker, for instance. Physical activity could be measured through smartphone motion sensors like accelerometers, gyroscopes, and step counters. Another potential inference to test is detecting whether the user is at home, at work, or elsewhere; we suspect that individuals most often charge their phones at work and at home, and do so in roughly the same place out of habit. Battery charging state, device orientation, and magnetometer data could be validated against geolocation ground truth to test this inference. Headphone connection state might also allow inferences about sleeping habits and physical activity.

Evaluation will use methods established in the literature: feature selection based on our intuitions, selection and training of viable supervised learning models (e.g, SVMs, artificial neural nets, nearest neighbor), and cross-validation to reduce spurrious correlations.

4. PRELIMINARY RESULTS

We conducted internal tests of our data collection app to verify that it works, as well as generate some preliminary data to examine for possible inferences. The app was installed on a researcher's personal Nexus 5 for about 24 hours on a typical weekday. The Nexus 5 supports low-power continuous monitoring for the accelerometer and gyroscope. This model is equipped with all the sensors in Table 1, except for ambient temperature and relative humidity.

Figures 1 and 2 show two points (denoted by the green line in the graph) in this subjects's morning commute: a walking commute in progress, and shortly after arriving at the office, respectively. The map is ground truth geolocation data at that time. The graphs show accelerometer (blue) and step counter (orange) activity from early in the morning through the afternoon.

Minimal motion sensor activity is observed during the early morning hours, when the subject is asleep and not actively using the phone. By contrast, the accelerometer and step counter both record high levels of activity while the subject is commuting by foot that morning (as shown in the ground truth data). A period of inactivity follows this commute, corresponding to when the subject arrives at the destination. Similar patterns of increased accelerometer and step counter activity occur later in the day, corresponding to the walk back home.

Although this is just one individual in a well-controlled observation, this limited data suggests some correlation between sensor measurements and certain everyday activities. Broader testing is needed to validate more general inferences across a variety of individuals and devices.

5. CONCLUSION

We present our ongoing work on inferring sensitive user details from "benign" smartphone sensor data. This project explores sensitive user information that might be leaked through a combination of smartphone data sources outside the protection of the OS's permissions system. We've developed a data collection app to gather sensor measurements and ground truth data, and will use this information to test our intuitions on the relationships between unrestricted sensor data and private user details. More general results will be presented in the future as this ongoing project continues.

6. REFERENCES

[1] DeviceOrientation Event Specification. `http://w3c.github.io/deviceorientation/spec-source-orientation.html`.

[2] Get the most life from your battery - Nexus help. `https://support.google.com/nexus/answer/6090612?hl=en`.

[3] Haystack: Preliminary Results — The ICSI Haystack Blog. `https://haystack.mobi/wordpress/index.php/2015/12/11/haystack-preliminary-results/`.

[4] Homepage — Pokemon Go. `http://www.pokemongo.com/`.

[5] How much is your personal data worth? - FT.com. `http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html#axzz4BDh1fipu`.

[6] iPhone 7 - Technical Specifications - Apple. `http://www.apple.com/iphone-7/specs/`.

[7] Requesting Permission - Interaction - iOS Human Interface Guidelines. `https://developer.apple.com/ios/human-interface-guidelines/interaction/requesting-permission/`.

[8] Sensor types — Android Open Source Project. `https://source.android.com/devices/sensors/sensor-types.html`.

[9] `comScoreReportsJanuary2016U.S.SmartphoneSubscriberMarketShar` comScore,Inc. `https://www.comscore.com/Insights/Rankings/comScore-Reports-January-2016-US-Smartphone-Subscriber-Market-Share`.

[10] What is Yelp's Monocle feature — Support Center — Help. `http://www.yelp-support.com/article/What-`

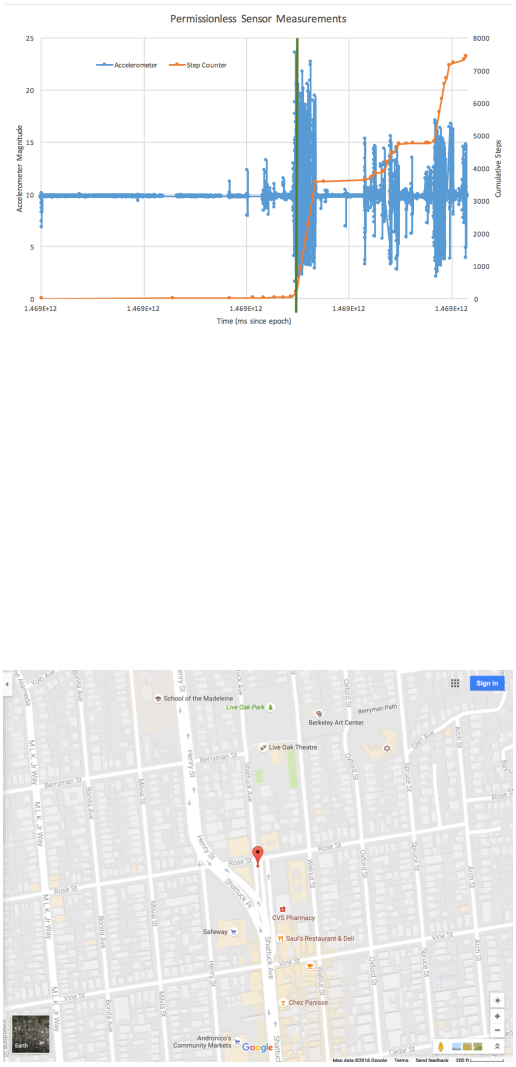

Figure 1: Morning commute by walking, sensors (top) and truth (bottom). Mapped location occurred at the time noted by the green line.

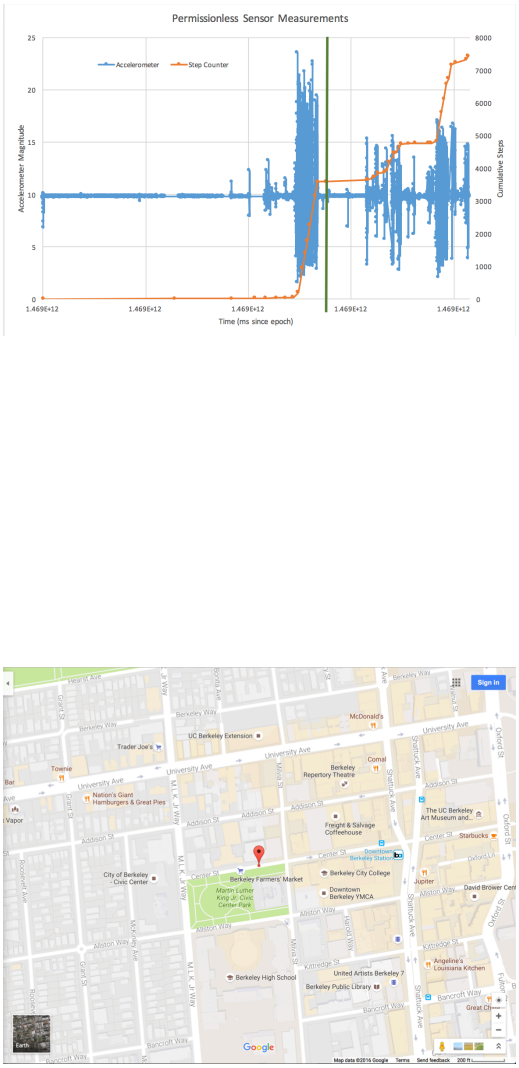

Figure 2: Morning arrival at work, sensors (top) and truth (bottom). Mapped location occurred at the time noted by the green line.

is-Yelp-s-Monocle-feature?l=en_US.

[11] Working with System Permissions — Android Developers. https://developer.android.com/training/permissions/index.html.

[12] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Best paper – follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 141–148, New York, NY, USA, 2013. ACM.

[13] L. Liu, C. Karatas, H. Li, S. Tan, M. Gruteser, J. Yang, Y. Chen, and R. P. Martin. Toward detection of unsafe driving with wearables. In *Proceedings of the 2015 Workshop on Wearable Systems and Applications*, WearSys '15, pages 27–32, New York, NY, USA, 2015. ACM.

[14] Y. Michalevsky, A. Schulman, G. A. Veerapandian, D. Boneh, and G. Nakibly. Powerspy: Location tracking using mobile device power analysis. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 785–800, Washington, D.C., Aug. 2015. USENIX Association.

[15] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang. Accessory: Password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, HotMobile '12, pages 9:1–9:6, New York, NY, USA, 2012. ACM.

[16] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 149–161. ACM, 2014.

[17] T. Pierce. Generating revenue from various business models. In *Appreneur*, pages 27–30. Springer, 2013.