One Experience Collecting Sensitive Mobile Data

Yuan Niu
UC Davis
yuan.niu@gmail.com

Elaine Shi
PARC
eshi@parc.com

Richard Chow
PARC
rchow@parc.com

Philippe Golle
PARC
pgolle@parc.com

Markus Jakobsson
PARC
markus.jakobsson@gmail.com

ABSTRACT

We report on our efforts to collect behavioral data based on activities recorded by phones. We recruited Android device owners and offered entry into a raffle for participants. Our application was distributed from the Android Market, and its placement there unexpectedly helped us find participants from casual users browsing for free applications. We collected data from 267 total participants who gave us varying amounts of data.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factors; K.4.1 [**Public Policy Issues**]: Privacy

General Terms

Human Factors

Keywords

user study, privacy, Android

1. INTRODUCTION

We previously introduced Implicit Authentication [6], the notion of authenticating a mobile device user based on her past behavior. The idea is that a user's recent behavior might be compared with past training data from the same user in order to make better security decisions. To test this idea, we attempted to collect behavioral data from mobile device users. The behavioral data we wanted to collect was quite sensitive, including location information and call logs.

We report on our efforts to collect this data. Not surprisingly, we had a difficult time recruiting users. One reason was that we opted to develop our data collection agent on the Android platform [1], and the number of Android users was relatively small in mid-2009. The Android Market, however, was already quite active, and in a sense our data collection efforts were saved by virtue of our application being part of the Android Market. It turned out that casual browsers of the Android Market would install our data collection agent, which had at least the virtue of being free. Many would soon uninstall, but some would continue to run our data collection agent. For instance, of the 57 users who gave us at least 2 weeks of data, at least 44 were not recruited by us and most likely came across the application in the Market.

We also describe some notable aspects of the design for the study. In particular, we chose a data obfuscation strategy that balanced user privacy and the need to collect useful behavioral data features. We also allowed users to delete any data collected, and below we summarize users' behavior with respect to deletion.

Usable Security Experiment Reports (USER) Workshop 2010, July 14–16, 2010, Redmond, WA USA

2. DATA COLLECTION METHODOLOGY

We collected the four types of data: location through GPS and rough estimation, phone numbers and times, SMS numbers and times, and browser activity (just the part of the URL before the first slash and the time of visit). We created an Android application to collect this data. Android was ideal for our purposes because it allowed background processes. For the most part, our application ran in the background and at regular intervals, pulled and uploaded information so as not to disrupt a user's regular habits. Our intention was to create an application that, once installed, would run silently in the background.

2.1 Obfuscation

To preserve our participants' privacy, we obfuscated as much data as we could. We decided to use keyed one-way hashes to increase user privacy. Using unkeyed hashes would have allowed us to match phone numbers or URLs across users, but is susceptible to a brute-force dictionary attack.

We generated a random unique key for each user at the time of install, stored only on the user device. This key was never uploaded to our servers. When events were recorded, they were hashed using the key. We obfuscated all data except for location data, as previous work has demonstrated the difficulty in obfuscating location data while preserving utility (for example, see [5, 7]).

2.2 Data Deletion

To give users a chance to remove data that they were not comfortable sharing, we provided options to delete data. Deletions were done in the form of uploaded requests to remove intervals of data. We promised users that before any human eyes examined the data, a script would process the deletion requests and remove from permanent storage

any data intended for deletion. We intended for every user to participate for 4 weeks, and we promised users that we would not look at any data until 5 weeks after they joined the study. In practice, many users participated longer than 4 weeks. For data beyond 4 weeks, we always ensured that users had at least 1 week grace period for data deletion, that is, we would not process any data until at least one week after its upload.

2.3 Recruitment

We recruited Android owners primarily through emails and postings on forums and social network sites. We always linked to our website [3], which contained information about installing and uninstalling the application, as well as the consent form and instructions on how to delete data and disable GPS.

To incentivize users, we offered them entry into a raffle for a $150 gift certificate to `amazon.com` if they provided us with a minimum of two weeks worth of data. We also promised participants access to all obfuscated data, which was all data except for location data. For every 10 users, there would be one raffled certificate with a limit of 10 certificates total. Participation in the raffle required that users sign up themselves because we required an email address and did not automatically extracting an email address from the user's device.

Users were free to uninstall at any time and instructions on how to do so were clearly stated on both the consent form (hosted on our study webpage) and on the webpage itself.

As most of our users were remote and obtaining signed consent forms was impractical, we considered the act of installing the application to be consent for participation (see Appendix A). Recruited users would have first visited our study website before installing. However, we did not anticipate the flood of installs from the Android Market. Our IRB application did not explicitly cover this situation, but we felt that these users were legitimate participants - the description for the application in the marketplace and popup during installation stated that users of the application would be involved in a study and what data would be collected, even if users did not read the consent form in its entirety.

The study lasted 8 weeks, beginning on August 10, 2009, and ending on October 6, 2009, although we continued to receive data after the study ended from users who did not uninstall the application. On October 6, we sent out an end-of-study email to users who signed up for the raffle, informing them that our study has ended, and whether they won the raffle. We did not send out end-of-study emails to non-rafflers, as we did not know their email addresses.

2.4 Android Market

To make it as easy as possible for potential subjects to install our application, we placed it in the Android Market [2] under the *Demos* section. This had an unexpected side effect in that unrecruited users downloaded the application as well.

Along with the application, we posted a contact email, our study webpage, and the following blurb:

> "This application is intended to gather data for a PARC (parc.com) project. Data includes your location (GPS, Wi-Fi, cell tower) as well as obfuscated versions of your phone calls in or out, SMS numbers in or out, and browser URLs. Contact us with problems or question. http://rachel-epn.parc.xerox.com has more information."

To provide information for users who did not read the blurb, we popped up a message at the time of first install (See Appendix B for the message) to inform them that they were eligible to participate in a raffle and remind them that this was a data collecting application. This message could be displayed again by pressing a button on the application's primary screen. We emphasized the raffle by mentioning it first to encourage participation.

3. USER BEHAVIOR

We report findings on user behavior and incentives.

A total of 363 people downloaded the application over a period of approximately 8 months. At the time of this writing, 36 users still have the application installed, although we are receiving data from only a few of them. Among these 363 downloads, we received data only from 267 users, probably owing to immediate uninstalls.

A total of 20 out of the 267 users signed up for the raffle. Among these 20, we were able to map 18 of them to users in our database by matching up phone numbers or device IMEIs. We found that 3 out of the 18 registered for the raffle after the end-of-study, i.e., October 6th, 2009. Therefore, only 15 users participated in the raffle.

3.1 Installation

To study users' incentives for installing our application, we track how many users installed the application during each time period. Figure 1 shows how many users installed the application in each week for the duration of the study. The figure also shows the breakdown of these installs into rafflers and non-rafflers. The study was planned to last 8 weeks, beginning on August 10, 2009, and ending on October 6, 2009, when we sent out an end-of-study email to users who signed up for the raffle, informing them that our study has ended, and whether they won the raffle. However, we continued to receive data after the study ended from users who did not uninstall the application. Roughly 200 users installed the application in Week 1 through Week 8. After the study ended, 77 users installed the application. Among the 200 people who installed the application during Week 1 through 8, more than half of them installed in the first two weeks. The spike in the number of new installs in the first two weeks may be explained by the following three factors: 1) We sent email to our connections immediately after the study started, encouraging them to install our application. 2) Whenever a new Android application is released, it appears towards the top of the "Newly Released" list in the Android Market. This makes it easier for random users to discover the application and install it. 3) We also posted an advertisement on the Android development forum around that time and encouraged users to install to enter the raffle.

As the application remained in the Android Market after the study ended, 77 more users installed the application after the end of study over a period of approximately 6 months (not shown in Figure 1), and 3 out of the 77 signed up for the raffle, although the raffle already took place by then.

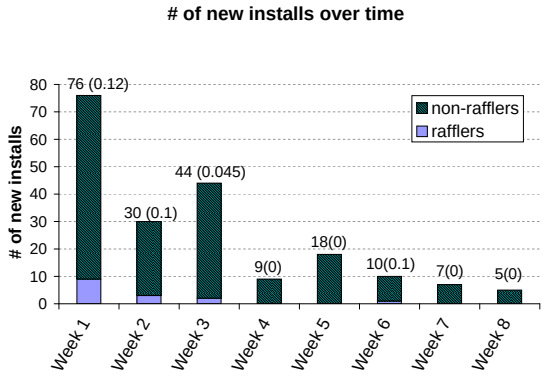


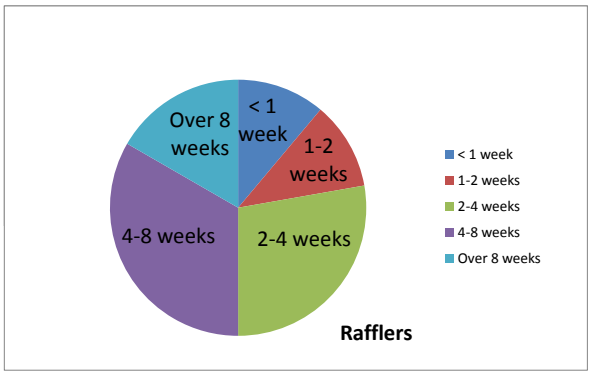
Figure 1: Number of new installs over time. The first number on top of each bar is the total number of users who installed the application during the specified week. The following number in the parenthesis is the fraction that joined the raffle.

3.2 Duration of Participation

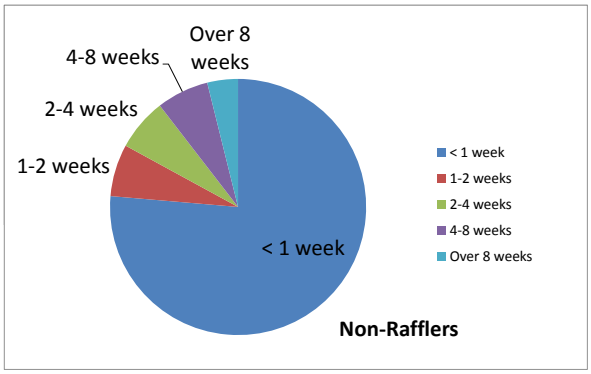
Users participated in our study for a varying amount of days. About half of the users uninstalled the application on the first day, while the longest participant stayed for more than 200 days.

Below are some hypotheses on what motivated users to uninstall the application: 1) On deciding that the application is not doing anything useful, users uninstalled it. This may explain why half of the users uninstalled it on the first day. 2) About 57 out of all users uninstalled the application within the first week. These users may have given our application more time, but on deciding that it is not interesting or useful after a few days, they uninstalled it. Also, some users may have been annoyed by our application causing the battery to drain faster – users needed to recharge their device on a daily basis. That may be another reason why users chose to uninstall the application. In fact, both the above hypotheses were confirmed by users' feedback in the Android Market. Users left comments complaining that the application is not useful, or is draining batteries too quickly. 3) Our application works only with phones running Android version 1.5. Users who successfully installed our application initially and later upgraded their OS version may have uninstalled our application due to the incompatibility of the application. However, we did not gather sufficient statistics in the raffle, only 2 of them uninstalled the app on October 6 – immediately after we sent out the end-of-study email. We did not send the end-of-study email to non-rafflers, as we did not to verify or infer how many users were affected by this problem. 4) Only a few users out of the 15 for whom we had email addresses uninstalled the application upon receiving the end-of-study email.

How long users participated in our study may be directly related to their motivation. Our study shows that users who signed up for the raffle were well-motivated to participate longer in our study. Our analysis shows that the rafflers participated for an average of 32 days, with a median of 27 days, and a minimum of 5 days. By contrast, the non-rafflers participated for an average of 10 days, with a median of 1 day. Given that half of the users uninstalled the application on the first day, it is likely that many users who found the application at random on the Android Market did not know the existence of the raffle or were uninterested. Our application did display a welcome message immediately after the users installed the application to encourage entry into the raffle. However, some users may have simply dismissed the message without reading.



(a) Rafflers



(b) Non-Rafflers

Figure 2: How long rafflers/non-rafflers participated in the study.

3.3 Data Deletion

As part of the study, we offered users the ability to delete any data collected, up until one week after the end of the study. The deletion was done through a GUI that was part of the data collection application, see Figure 3. We kept track of deletion requests, in particular the user requesting

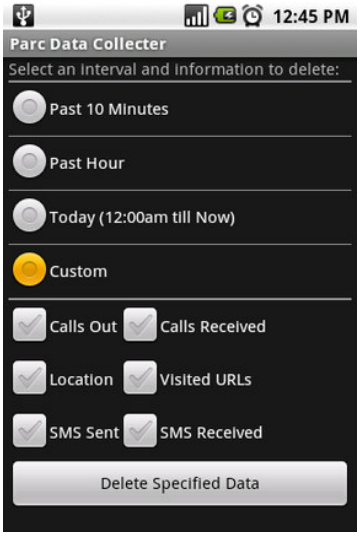


Figure 3: Users could delete any data collected through a UI that was part of the data collection application.

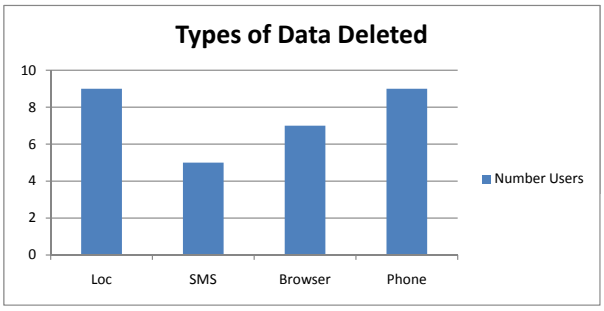


Figure 4: Types of data deleted

the deletion, the type of data deleted, and the time period of the deleted data.

A total of 12 users deleted data. One user in particular was very vigilant about deleting data, especially browser data, with 86 separate deletion requests. Another user decided to essentially opt out of the study by deleting all his data, specifying a period of greater than one month and all types of data. The remaining users who deleted data made just a few (and often just one) specific deletion request. Only one user who deleted data was a raffler. Figure 4 shows the types of data deleted. Phone numbers were deleted as often as locations, leading one to question whether the users who deleted data understood or trusted our obfuscation process - with our obfuscation process, phone numbers cannot be reconstructed but GPS coordinates are not obfuscated at all.

Figure 5 shows the time periods chosen for deletion. One can select the last 10 minutes, the last hour, the day (up to the moment), or a custom time period. Nobody selected

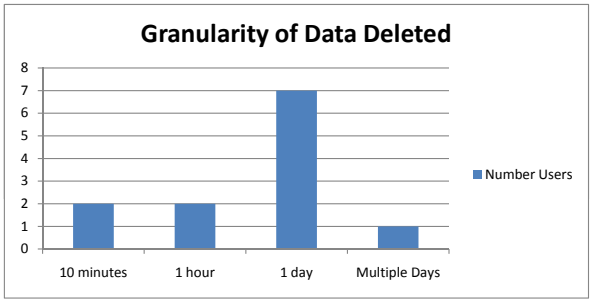


Figure 5: Time periods chosen to be deleted.

a custom time period, except the user who deleted all his data. There seemed to be a preferred mode of deleting a period of one day.

12 users make up only a small fraction of all the users. It seems the majority of users either deemed our obfuscation measures sufficient or simply did not care about releasing the data.

3.4 User Feedback

We assumed that anyone who downloaded the application would also sign up for the raffle, thus providing us with an e-mail address. Because of the unanticipated users who found the application through Android Market, we found ourselves with users providing us with data beyond the end of the study and no way to contact them. Although our application did extract the phone number associated with each device, so as not to disrupt our subjects, in our consent form we had promised participants that those numbers were only used to verify their presence in the study, and we would not contact them through any means but e-mail. Despite removing the application from the Android Market, there are still 36 users to date who have it installed, though we are not receiving data. It is possible that these users have upgraded their Android devices, or have disabled uploading entirely.

We contacted participants for whom we had email addresses and asked them to fill out a survey. Only six users responded. They were most annoyed by drain on battery life because of GPS. Users were aware that they could delete data, and one mentioned in particular that he/she deleted location data because it could be used as an identifier. We also examined the few comments regarding our application from the Android Market. Some users noted its incompatibility with other versions of Android OS, and others were confused as to the purpose of the application.

4. LESSONS LEARNED

We now summarize our findings on user behavior in the study, and discuss lessons learned.

The Android Market seems to be a good place to recruit for user studies, as random users who found the application in the Android Market contributed the majority of the data

in the study. Other application download sites may work as well, but some have the disadvantage of requiring a lengthy vetting process, such as Apple's App Store [4].

We found that the majority of users who successfully installed our application immediately uninstalled on the same day. While this was not surprising, given the intrinsic interest of such applications for the average user, it was unexpected that even the users who kept the application installed did not participate in our raffle. In fact, only 20 out of 267 users signed up for the raffle. It could be that many of the users ignored the welcome message and did not notice the existence of the raffle.

Had we made the existence of the raffle more prominent, more users may have been motivated to stay longer in the study – as our study shows that rafflers clearly participated much longer than non-rafflers.

Location was the only data type which we did not obfuscate. Surprisingly, deletions seem to have happened equally often for each type of data. We are not sure if users understood the description of our data obfuscation techniques in the consent form. Only 12 out of 267 users chose to delete data – it remains unclear whether they were unconcerned about privacy, or were happy with our obfuscation technique, or did not know that the deletion functionality existed.

For others wishing to take advantage of the many users on an application download site, we recommend that the researcher make sure the users understand the risks and incentives for participating in the study. In our case, at the time of install, instead of popping up a message explaining the purpose of the application, we would have displayed the consent form and required that all potential users enter at least an email address to signal their understanding of the study and its incentives.

5. ACKNOWLEDGMENTS

We gratefully thank the anonymous users who participated in our study, without whom this work would of course not be possible.

6. REFERENCES

[1] Android. `http://www.android.com`.

[2] Android Market. `http://www.android.com/market`.

[3] Implicit authentication study website. `http://rachel-epn.parc.xerox.com`.

[4] iPhone App Store. `http://www.apple.com/iphone/iphone-3gs/app-store.html`.

[5] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, pages 161–171, New York, NY, USA, 2007. ACM.

[6] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit Authentication for Mobile Devices. In *HotSec '09: Proceedings of the 4th USENIX Workshop on Hot Topics in Security*, 2009.

[7] J. Krumm. Inference attacks on location tracks. In *In Proceedings of the Fifth International Conference on Pervasive Computing (Pervasive), volume 4480 of LNCS*, pages 127–143. Springer-Verlag, 2007.

APPENDIX

A. CONSENT FORM

Consent Form

Thank you for participating in this study, which aims to explore new authentication techniques using various types of data collected from the user's device. The data will be used to develop techniques to augment traditional password authentication of a user.

Process.
In this study you will be asked to do the following:

1. Install an Android app that records times and numbers associated with phone calls and SMS, browser history, cell tower IDs, WiFi SSID (if available), GPS coordinates (if available), battery state, and USB state. Phone numbers, SSIDs (if available) and browser history entries will be obfuscated before being recorded on the device. The obfuscation method is described below.
2. Decide whether the app may upload data using the carrier network, WiFi, or both. You may want to choose WiFi if you would be charged for uploading data using the carrier network. This is the only configuration needed.
3. The app will automatically upload obfuscated and recorded data to a central server to which only the researchers in this study have access. The researchers will not access the information on the server until the end of the study.
4. The study lasts four weeks. At the end of the study, you will be asked to uninstall the app. You may uninstall the app before the end of the study if you choose to.
5. At any time during the study, and till one week after its end, you may erase data of your choice based on categories and time intervals. Data you erase will be removed from storage before any human examines the data and before any experiment or analysis are run on the data. At the end of the study, data that has not been erased will be copied from the server storage area to a machine managed by the experimenters.

Incentives.
All subjects will be enrolled in a raffle for $150 Amazon Gift Certificates. There will be one such certificate for each ten subjects who complete the study and submit at least two weeks of data. (This is subject to a limit of 10 certificates total. If there are more than 100 participants, the odds of getting a certificate start decreasing from 1 in 10.) If the total number of participants is not divisible by 10, we will make up the balance with dummy participants and hold the raffle as usual.

In addition, each subject who completes the participation will be given access to parts of the obfuscated data from all participants that complete their participation in the study. We will not give access to geo-location data, as that is difficult to obfuscate well.

Uninstallation.
To remove the application, users simply use the application manager provided as part of the operating system. The un-installation deletes all settings and files associated with the application as well as the application itself. The participant can quit the study and remove the application at any time.

Anonymization.
The obfuscation procedure depends on the type of data. We will obfuscate the following types of data, as indicated:

- Phone numbers: Keyed hash of the phone number at time of collection. This corresponds to a random number that the researchers cannot decode to infer the real number, but where two instances of the very same number – when recorded by one and the same participant – will correspond to the same random number. It will not be possible to compare numbers recorded by different participants.
- WiFi SSID: Keyed hash of the SSID at time of collection.
- Browser history: Keyed hash of the domain or subdomain at time of collection.

We will not obfuscate GPS data.

Here, each subject will use a unique key known only to his device. This key is not known by the researchers, and is randomly generated on the device at installation time. Given the one-way properties of hash functions, it will not be possible for the researchers (or anybody else) to infer the secret input data (such as a phone number) from the recorded data. However, it will be possible to identify reoccurrences of the same number for the same user (but not the same phone number in the context of two different users.)

We will make every effort to keep your data confidential. We will only publish aggregate data. Your participation in the study is completely voluntary. You have the right to withdraw from the study at any time without penalty or explanation, and you may refuse to answer any questions.

Questions?
If you have any questions about our research, please feel free to contact any of the following researchers:

Richard Chow, rchow@parc.com, 650-812-4739
Philippe Golle, pgolle@parc.com, 650-812-4459

Markus Jakobsson, mjakobss@parc.com, 650-812-4722
Yuan Niu, yniu@parc.com, 650-812-4461
Elaine Shi, Elaine.shi@parc.com, 650-812-4462

If you have any concerns at any time about your participation in this study or wish to report any objections or concerns about the conduct of the study, please report them either orally or in writing to:

Peter Pirolli
Chair: Human Subject Review Board
PARC Inc.
3333 Coyote Hill Road
Palo Alto CA 94304
+1 650 812 4483
pirolli@parc.com

Your approval.
I understand that by installing the Android app described in this consent form, I agree to participate in the study, and give PARC employees Richard Chow, Philippe Golle, Markus Jakobsson, Yuan Niu, Elaine Shi, and research associates permission to present data from this work including the options specified above in written and oral form, without further permission from me.

B. POPUP MESSAGE

Thank you for installing the PARC Implicit Authentication data collector! By participating in this study, you may be eligible to participate in a raffle for one of ten $150 Amazon.com gift certificates.

This application is for a PARC project that aims to explore new authentication techniques to augment traditional password authentication using data collected from a user's device.

To enter the prize drawing or to find out more information, including details on the types of data collected, and what it means to participate, visit `http://rachel-epn.parc.xerox.com`.

To close this message, use the back button on your phone or tap anywhere outside the message box.