A Lightweight User Tracking Method for App Providers

Remo Manuel Frey
ETH Zurich
Weinbergstrasse 56/58
8092 Zurich
+41 44 632 48 18
rfrey@ethz.ch

Runhua Xu
ETH Zurich
Weinbergstrasse 56/58
8092 Zurich
+41 44 632 82 56
rxu@ethz.ch

Alexander Ilic
University of St. Gallen
Dufourstrasse 40a
9000 St. Gallen
+41 71 224 73 00
alexander.ilic@unisg.ch

ABSTRACT
Since 2013, Google and Apple no longer allow app providers to use the persistent device identifiers (Android ID and UDID) for user tracking on mobile devices. Other tracking options provoke either severe privacy concerns, need additional hardware or are only practicable by a limited number of companies. In this paper, we present a lightweight method that overcomes these weaknesses by using the set of installed apps on a device to create a unique fingerprint. The method was evaluated in a field study with 2410 users and 175,658 installed apps in total. The sets of these installed apps are unique in 99.75% of all inspected users. Furthermore, by reducing the granularity from apps to app categories to lessen users' privacy concerns, the results remain highly unique with an identification rate of 96.22%. Since the information of installed apps and app categories on each device is freely available for any app developer, the method is a valuable instrument for app providers.

CCS Concepts
• **Information systems→Data mining**
• **Information systems→Personalization**

Keywords
User Tracking; Mobile Analytics; Reality-Mining; Mobile Apps; Mobile Privacy

1. INTRODUCTION
Reality-Mining of a user infers user habits, behavior and needs by applying data-mining algorithms to information collected by mobile devices. According to MIT Technical Review [10], the technology is one of the "10 technologies most likely to change the way we live". Tracking users over a certain period of time and gather personal data enables new marketing and business opportunities. Custom-tailored content like personalized prices, product recommendations, and search results may be presented to consumers, based on observed activities on their devices [13].

As a pre-condition of providing such advanced services, a unique identifier is required to represent each smartphone user thereby distinguishing her from others. However, the leading players in the worldwide app business, Google and Apple, replace persistent device identifiers (Android ID and UDID) with resettable identifiers (Advertising ID and IDFA) due to the increasing privacy concerns. Also, non-anonymous data like name, (email) address, social media content [2], and phone call activities [7] cannot be used without user consent in both research and practice. Recent research reveals possibilities to use data that is considered less sensitive like antenna signals [6]. Nevertheless, such data is only available to limited entities like phone telecommunication companies and manufacturers.

Achara et al. [1] collected the open-accessible lists of running apps on mobile devices over more than seven month. They assumed that recording over a long period of time is likely to sum up to the set of all installed apps of a user. They admitted that the set might not be a complete set of all installed apps. Nevertheless, based on this set, they derived a fingerprint that is unique in 99% of all cases. Due to the fact that an app provider cannot wait several months until it is able to track its users, we propose to use an available public API from Google to get the complete list of installed app in one snapshot. The proposed approach was evaluated from the perspective of an app provider and a field study was conducted with a self-developed app. Since an app provider is interested to reduce privacy concerns of its users, we additionally analyze the usage of app categories instead of apps. For instance, the usage of dating and pregnancy apps is often considered as confidential. In Google's app store, these apps are just classified as 'Lifestyle' apps which is far less sensitive.

The present work is structured as follows: We provide an overview of related research first. Then, we explain the method and our study design, followed by the result section where we compare the results between apps and app categories. Finally, we raise the discussion about privacy and describe what we intend to do in future work.

2. RELATED WORK ON MOBILE USER TRACKING
Several methods are proposed in current research to accurately distinguish a user from others without getting a unique identifier and any other non-anonymous data. For instance, a recent study proved that sampling four spatial-temporal points from each mobile user's antennas is enough to uniquely identify 95% of all the users [6]. Several studies investigate web-based fingerprinting [3,8], i.e. they analyze browser activities, like user's page visits. With the increasing importance of apps over the last decade, browser-related approaches may lose relevance. Quattrone et al. [11] revealed that smartphone diagnostic information that is not considered as sensitive (like hardware statistics and system settings) could be further processed to identify each user at an accuracy of 94%. Similarly, Olejnik et al. [9] read out the battery status of a smartphone. They showed that battery status could

CF'16, May 16-19, 2016, Como, Italy
ACM 978-1-4503-4128-8/16/05.
DOI: http://dx.doi.org/10.1145/2903150.2903484

serve as a finger-printable surface to accurately represent each smartphone user in short time intervals. In addition, data from power amplifiers, oscillators and signal mixers on a mobile phone also provides possibilities to uniquely identify each mobile device and its user [5]. As mentioned in the introduction, Achara et al. [1] describe the uniqueness apps on smartphones. In contrast to them, we catch only a single snapshot from each device.

Recent studies proofed that the set of installed apps is a valuable information source for user profiling [12,14] and other delicate actions like prediction of current life events [4]. Thus, the method based on apps raises again concerns about privacy protection and we suggest to use the less sensitive app categories instead of apps. These considerations lead us to the following research question:

How unique is a fingerprint of mobile devices based on the sets of installed apps and app categories?

3. METHOD

3.1. Recording Installed Apps and App Categories

Google's operating system Android provides a public API called 'android.content.pm' to retrieve information about installed apps on mobile devices. (Apple closed the access in its operating system version iOS 9.) Each app that is installed on an Android device is able to access the data through the API. The retrieved list of apps can either be used to find out more about the user's properties [12,14] or to produce a fingerprint from the device [1].

In the present work, we go further and reduce the granularity from app to app category level in order to investigate if even the installed categories are still highly unique.

To the best of our knowledge, there is no scientific categorization of mobile app. As 'Google Play Store' is the leading market for Android apps, we thus choose its categorization as reference in this work. This results to 42 categories in total. Our prototype automatically visits Google's online description of each observed app and downloads the corresponding app category from there.

3.2. Test App

To demonstrate the potential for app providers, we developed a test app and put it in Google Play Store. The app was categorized as a 'Lifestyle' app and was described as a personality test game. Figure 1 shows two screenshots of the app. Each user gives answers to personality measurements (as shown in Figure 1 left) to compare her personality traits with the average of other people who have already participated in the game (as shown in Figure 1 right). Additionally, the app also collects demographic data from the users by displaying questions about gender and age.

Before downloading and using the app, the user had to accept the condition that her data can be used for our research analysis. When a user opens the app for the first time, a background process reads the list of installed apps on the device and does a lookup to get the corresponding app categories on Google's online description.

As a data cleaning step, double entries coming from uncommon user actions (like installing the app twice) are removed.

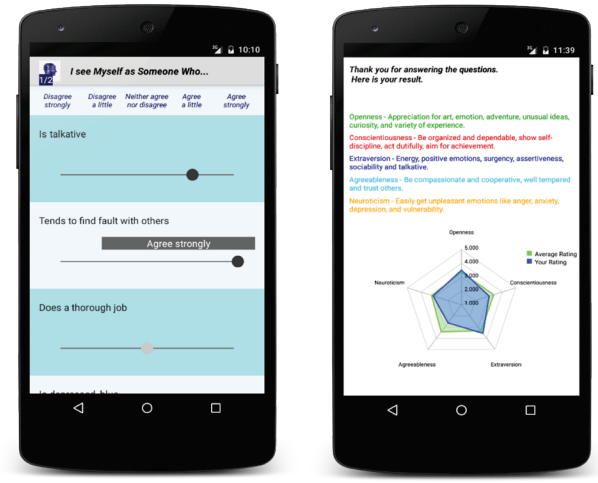

Figure 1. Two screenshots of the test app.

3.3. Definition of Uniqueness

To answer the research question, we give a short definition of uniqueness.

S_i is the set of all installed apps or categories, respectively on the mobile device of a user i. The notation $S_1, S_2, \ldots, S_n$ describes the sets of all n users. Thus, the common mathematic definition for the symmetric distance d between the sets S_i and S_j from two users is

$$d(S_i, S_j) = \left|(S_i \backslash S_j) \cup (S_j \backslash S_i)\right| \quad (1)$$
Distance

To use a set S_i as unique fingerprint, we are interested in the minimum distance to all other $n-1$ sets, which is

$$d_{min}(S_i) = \min_{S_k \in \{S_1, S_2, \ldots, S_n\}/S_i} (d(S_i, S_k)) \quad (2)$$
Minimal Distance

If d_{min} of a set S_i is equal to zero, there is another set with exactly the same apps or categories, respectively, which makes the fingerprint not unique. Greater than zero means that the set is unique and the device and its owner are uniquely identifiable. Thus, the definition of a unique identifier is

$$d_{min}(S_i) > 0 \quad (3)$$
Uniqueness

It is clear that an individual's set changes over time. If our method aims to be able to track devices and users over a certain period of time, it must be robust against ongoing installations and de-installations of apps or categories, respectively. Therefore, we strengthen our definition of uniqueness: A fingerprint is only unique if the distance to all other fingerprints is greater than a given size b. The formal definition is

$$d_{min}(S_i) > b \quad (4)$$
Extended Uniqueness

In other words, devices who are unique under this definition stay unique after b app installations or de-installations in any case.

4. RESULTS

4.1. Participants

Our test app was published on Google Play Store on March 27, 2015. Until June 6, there were 2428 participants in total. The data on demographics was not used in the model and are just provided for illustration purposes of the sample population only as seen in Table 1. Based on the data cleaning step described in the previous section, 18 participants are removed.

Table 1. Characteristics of participants in the study (N=2410).

Gender	Female	74.9%
	Male	22.8%
	No Answer	2.2%
Age	10 - 19	25.1%
	20 - 29	47.3%
	30 - 39	16.5%
	40 - 49	7.1%
	50 - 59	1.4%
	60 - 69	0.2%
	No Answer	2.3%

4.2. Installations of Apps and App Categories

Figure 2 shows the distribution of apps and app categories installed by participants on their mobile devices. Due to the fact that the number of app categories is limited to 42, the distribution of the app categories is narrower than that of the apps. The standard deviation is 27 for the apps and 5 for the app categories. In total, we observed 175,658 installed apps, among which 12,681 apps are distinct. Each user installed 73 apps on average, which belongs to 19 app categories on average.

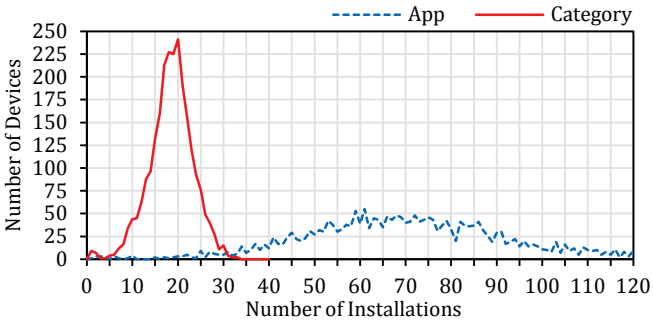

Figure 2. Distribution of installed apps and app categories per device (N=2410).

4.3. Uniqueness

Figure 3 shows the distribution of minimal distance d_{min} between all sets of apps as described in Equation 2. The median of d_{min} is 31. Only six out of the 2410 participants have a zero distance on installed apps. Thus, we have an identification rate of 99.75%. We obtain similarly good results for app categories, as shown in Figure 4. There are 91 participants with zero distance on installed categories. The identification rate is 96.22%, which is still higher than most results reported by previous research.

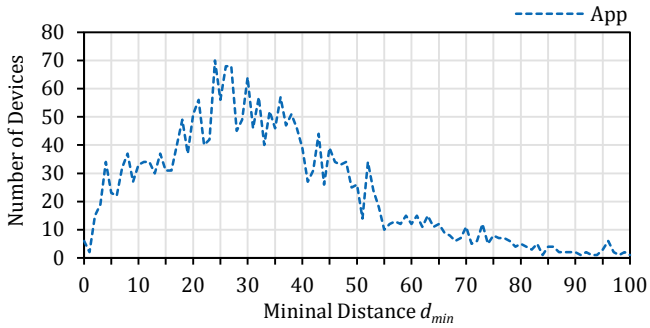

Figure 3. Distribution of the minimal distance d_{min} of installed apps (N=2410).

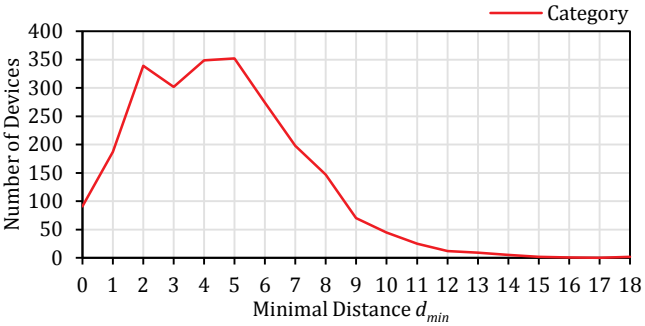

Figure 4. Distribution of the minimal distance d_{min} of installed app categories (N=2410).

4.4. Robustness

Finally, the robustness of the proposed method is tested. The previously defined 'Extended Uniqueness' is used in the test. Figure 5 shows how the identification rate decreases if the value of b in Equation 4 alters from 0 to 20.

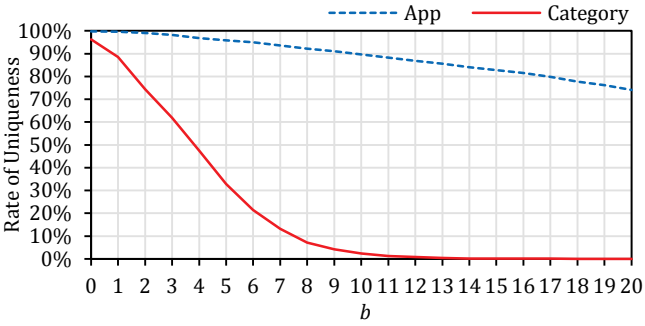

Figure 5. Rate of the extended uniqueness of installed apps and app categories (N=2410).

After 10 app installation or de-installation activities, the uniqueness for 90% of all devices is still guaranteed. However, a device will only become non-unique if the user installs and/or de-installs exactly the 10 apps that differentiate it from the other devices with a distance of 10. (Or, two users with distance of 10 install and/or de-install 5 apps each.) Taking millions of apps available on app market into account, the probability of having such a specific app installation pattern is very low. Therefore, the 90% level stands for the lowest bound our approach will deteriorate in theory. In reality, we believe that the actual uniqueness will stay high. Furthermore, people on average install apps every 19 days (based on analyzing our study data), which

means our estimated uniqueness will stay stable over a long period of time.

As expected from the previous results, app categories are weaker in producing unique fingerprints – one change activity will decrease the lowest bound of uniqueness to 90%. Based on our data, a change activity on category level happens only every 36 days. However, similar to the rationale on individual app level, the uniqueness on app category level is still acceptable in reality. Thus our research questions are answered.

5. DISCUSSION, LIMITATIONS, AND FUTURE WORK

Our analysis shows a reliable approach for mobile user tracking, which it can be used by any app publisher. With an identification rate of 99.75% for apps and 96.22% for app categories respectively, the method performs better than previous methods and has no hardware requirements like accelerometer, SIM card, GPS, or microphone. Since app categories are less sensitive but still highly unique, we recommend to use app categories instead of apps for fingerprinting.

From a technical point of view, the method does not need an explicit user permission. However, to use the present method, we strongly recommend to obtain the explicit consent from the user before using it regardless of whether apps or app categories are used. If the user accepts, the method supports a company to track her with the aim to offer her personalized content, products, and services – a great benefit for both.

The method focuses on sets of apps and categories and not on device dependent data like installation time (which is also readable from the device). This leads to a great advantage compared to other methods: A user typically shares her apps to all his current devices (smartphones, tablets, and desktop). Even if there are device specific apps, we believe that there is a subset that is still unique compared to other users. Thus, we speculate that it is possible to identify and to track users over several devices. Moreover, if a user buys a new device and removes the old one, the apps are typically automatically migrated and the tracking remains possible. Existent solutions are not able to handle this issue. In further work, we expect to observe and track users using more than one device.

There are two limitations on this work. First, our samples are unbalanced in terms of age and gender. 75% of the samples are woman and 72% of the samples are younger than 30 years. Second, we analyzed an app with 2410 users, but many apps have much more users. More users leads to a smaller minimum distance d_{min} between the sets, i.e. the uniqueness decrease. In a future study, we plan to cooperate with our research partner thereby enabling us to measure the minimal distance of about two million sets and check if the method is still useful for app providers.

6. CONCLUSION

The contributions of our work are three-fold: First, we present a lightweight user tracking method based on installed apps and evaluate it in a large field study. Second, we raise concerns about privacy prevention and invoke further discussion about the access and use of mobile app logs by app providers. Third, we propose a less problematic solution based on app categories instead of apps. The results demonstrate that the uniqueness remains high but with less privacy concerns and thus, the method is a useful instrument for app providers.

7. REFERENCES

[1] Jagdish Prasad Achara, Gergely Acs, Claude Castelluccia, et al. 2015. On the unicity of smartphone applications. *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*: 27–36.

[2] David N. Chin and William R. Wright. 2014. Social Media Sources for Personality Profiling. *Proceedings of the 2nd Workshop Emotions and Personality in Personalized Services*, 613–620.

[3] Peter Eckersley. 2010. *How unique is your web browser?* Springer Berlin Heidelberg.

[4] Remo Manuel Frey, Runhua Xu, and Alexander Ilic. 2015. Reality-Mining with Smartphones: Detecting and Predicting Life Events based on App Installation Behavior. *Proceedings of the 36th International Conference on Information Systems (ICIS)*.

[5] Jakob Hasse, Thomas Gloe, and Martin Beck. 2013. Forensic Identification of GSM Mobile Phones. *ACM workshop on Information hiding and multimedia security*: 131–140.

[6] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific Reports* 3, 1376.

[7] Yves-alexandre De Montjoye, Jordi Quoidbach, and Florent Robic. 2013. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 48–55.

[8] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2014. On the workings and current practices of web-based device fingerprinting. *IEEE Security and Privacy* 12, 3: 28–36.

[9] Lukasz Olejnik, Gunes Acar, Claude Castelluccia, and Claudia Diaz. 2015. The leaking battery A privacy analysis of the HTML5 Battery Status API. *Cryptology ePrint Archive, Report 2015/616*.

[10] Alex Pentland. 2008. TR10: Reality Mining. *MIT Technology Review*.

[11] Anthony Quattrone, Tanusri Bhattacharya, Lars Kulik, Egemen Tanin, and James Bailey. 2014. Is This You? Identifying a Mobile User Using Only Diagnostic Features. *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, 240–243.

[12] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. 2014. Predicting User Traits from a Snapshot of Apps Installed on a Smartphone. *ACM SIGMOBILE Mobile Computing and Communications Review* 18, 4: 1–8.

[13] Phumisak Smutkupt, Donyaprueth Krairit, and Vatcharaporn Esichaikul. 2010. Mobile Marketing : Implications for MARKETING STRATEGIES. *International Journal of Mobile Marketing* 5, 2: 126–139.

[14] Runhua Xu, Remo Manuel Frey, Denis Vuckovac, and Alexander Ilic. 2015. Towards Understanding the Impact of Personality Traits on Mobile App Adoption - A Scalable Approach. *23rd European Conference on Information Systems*.