Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints

ZHEN TU, RUNTONG LI, and YONG LI*, Tsinghua University, China
GANG WANG, Virginia Tech, American
DI WU, Hunan University, China
PAN HUI, HKUST, China and University of Helsinki, Finland
LI SU, DEPENG JIN, Tsinghua University, Tsinghua National Laboratory for Information Science and Technology (TNLIST), China

Understanding mobile app usage has become instrumental to service providers to optimize their online services. Meanwhile, there is a growing privacy concern that users' app usage may *uniquely* reveal who they are. In this paper, we seek to understand how likely a user can be uniquely re-identified in the crowd by the apps she uses. We systematically quantify the *uniqueness* of app usage via large-scale empirical measurements. By collaborating with a major cellular network provider, we obtained a city-scale anonymized dataset on mobile app traffic (1.37 million users, 2000 apps, 9.4 billion network connection records). Through extensive analysis, we show that the *set of apps* that a user has installed is already highly unique. For users with more than 10 apps, 88% of them can be uniquely re-identified by 4 random apps. The uniqueness level is even higher if we consider when and where the apps are used. We also observe that user attributes (*e.g.*, gender, social activity, and mobility patterns) all have an impact on the uniqueness of app usage. Our work takes the first step towards understanding the unique app usage patterns for a large user population, paving the way for further research to develop privacy-protection techniques and building personalized online services.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; *Mobile and wireless security*; Pseudonymity, anonymity and untraceability;

Additional Key Words and Phrases: Mobile Apps; Usage Patterns; User Privacy

ACM Reference Format:
Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, and Li Su, Depeng Jin. 2018. Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 138 (September 2018), 23 pages. https://doi.org/10.1145/3264948

1 INTRODUCTION

Mobile applications (apps) have enabled highly convenient and ubiquitous access to Internet services. Today's Internet users are spending more time on their mobile apps, rather than using traditional websites [21, 53]. For app developers and network service providers, it has become instrumentally important to understand how the apps are used under various temporal and location contexts to provide higher-quality and personalized

[*] This is the corresponding author

Authors' addresses: Zhen Tu, tuz16@mails.tsinghua.edu.cn; Runtong Li; Yong Li, liyong07@tsinghua.edu.cn, Tsinghua University, China; Gang Wang, gangwang@vt.edu, Virginia Tech, American; Di Wu, Hunan University, China; Pan Hui, HKUST, China, University of Helsinki, Finland; Li Su, Depeng Jin, Tsinghua University, Tsinghua National Laboratory for Information Science and Technology (TNLIST), China.

2474-9567/2018/9-ART138 $15.00
https://doi.org/10.1145/3264948

services [9, 15, 53, 58]. While various agencies (*e.g.*, advertisers, network providers, online services) are collecting app usage data [20], there is a growing concern about the privacy implications of mining or sharing such datasets. More specifically, different users may use different sets of apps due to their personal interest and preference. Even for users that use the same or similar apps, their usage patterns across location and time may vary [6, 53]. The concern is such app-level and behavior-level characteristics can uniquely reflect who the user is, which can be used to re-identify users from the "anonymized" app usage datasets.

The critical question is *how likely a user can be uniquely re-identified by her apps.* To answer this question, however, there are a number of practical challenges. First, researchers often have very limited access to the app usage data of a large user population. Most existing works on app usage analysis rely on small-scale datasets collected from volunteers who are willing to install a special monitoring app to record their app usage [5, 8, 23]. However, for the "uniqueness analysis" here, the result wouldn't be meaningful unless the dataset covers a large user population. Second, to understand the unique patterns of app usage, the information of temporal and spatial context is important. This requires collecting simultaneous traces on app usage, location, and time, which makes the data collection even more challenging [6, 53]. Related works have looked into the uniqueness of human behavior in different platforms of online services, including mobility patterns [32, 32, 34, 57], web browsing histories [13, 46, 56] and mobile device sensor fingerprints [7, 50]. However, in terms of app usage behavior, rarely do researchers have the data to measure the uniqueness of individual users' app usage behavior. Despite a large body of studies on app usage modeling [6, 9, 15, 16, 22, 24, 25, 27], the privacy perspective has not been well-understood, especially the upper bound of privacy leakage over large user populations.

In this paper, we perform a large-scale empirical measurement to quantify the *uniqueness* of app usage patterns for users in a metropolitan city. By collaborating with a major cellular network provider, we obtained an anonymized dataset which contains app network traffic data from 1.37 million users over 2000 apps during 1 week in one of the largest city in China. This dataset contains in total 9.4 billion network access records generated by users who use the app to send network requests to 9,800 cellular towers in the city, providing a unique opportunity to *quantify* the uniqueness of users' app usage behavior at a large scale.

Based on this dataset, we seek to answer the following research questions: *First*, how likely can adversaries re-identify a user based on the set of apps she uses? *Second*, how unique a user is in terms of *when* and *where* the user uses certain apps? *Third*, how does the uniqueness of app usage differ for users of different demographics (gender) and correlate with their online and offline behaviors? Through extensive analysis, our work has produced a number of surprising findings. Below, we summarize the key results and our contributions.

- First, by analyzing this city-scale dataset, we find that *the set of apps* that users have installed is highly unique. Among users with more than 10 apps, 88% of the users can be uniquely re-identified by a random selection of just 4 apps; 98% of the users can be uniquely re-identified by 6 random apps; and 97% of the users can be uniquely re-identified by their top 6 most frequently used apps.
- Second, we find that the uniqueness of app usage increases when we consider the spatial and temporal characteristics. For example, 80% of the users can be uniquely re-identified by 3 random data records (characterized by the used app coupled with the location and time information). The uniqueness is even higher for a sparser dataset which only contains 10% of the app usage records. In addition, we observe that users' uniqueness of app usage increases significantly during the weekdays compared to that of the weekend.
- Finally, we further examine the correlation between the uniqueness of app usage and the user attributes. We observe that gender, social activity and mobility patterns all have an impact. For example, male users, users who are less socially active, and suburb residents are more "unique" in their app usage. Surprisingly, users with a smaller movement range are also more unique in their used apps. Further analysis shows that

these less "mobilized" users prefer using various chatting and online shopping apps which makes their app usage more diverse than other groups (*i.e.*, more distinguishable).

Our work makes a concrete first step towards understanding the unique app usage patterns for a large user population. The implications are two folds. First, the result indicates that users' app usage is highly unique. Future research is needed to develop privacy-protection techniques in order to facilitate the mining (or even sharing) of the app usage datasets. Second, the high uniqueness of app usage patterns indicates that mobile users should not be treated as homogeneous user population [6]. The distinct user-level characteristics should be carefully considered and modeled in order to build personalized online services.

The rest of the paper is organized as the following. We first describe the dataset and our methodology in Section 2. Then in Section 3 we analyze the uniqueness of individual app usage fingerprints. Section 4 compares the difference of uniqueness between different user groups. Section 5 reviews the related work and discusses privacy-preserving solutions. Finally, Section 6 concludes the paper.

2 DATASET AND METHODOLOGY

2.1 Smartphone App Usage Dataset

We use a city-scale app usage dataset from a large-scale smartphone user population to analyze the unique individual app usage behavior and investigate the potential risks of privacy leakage.

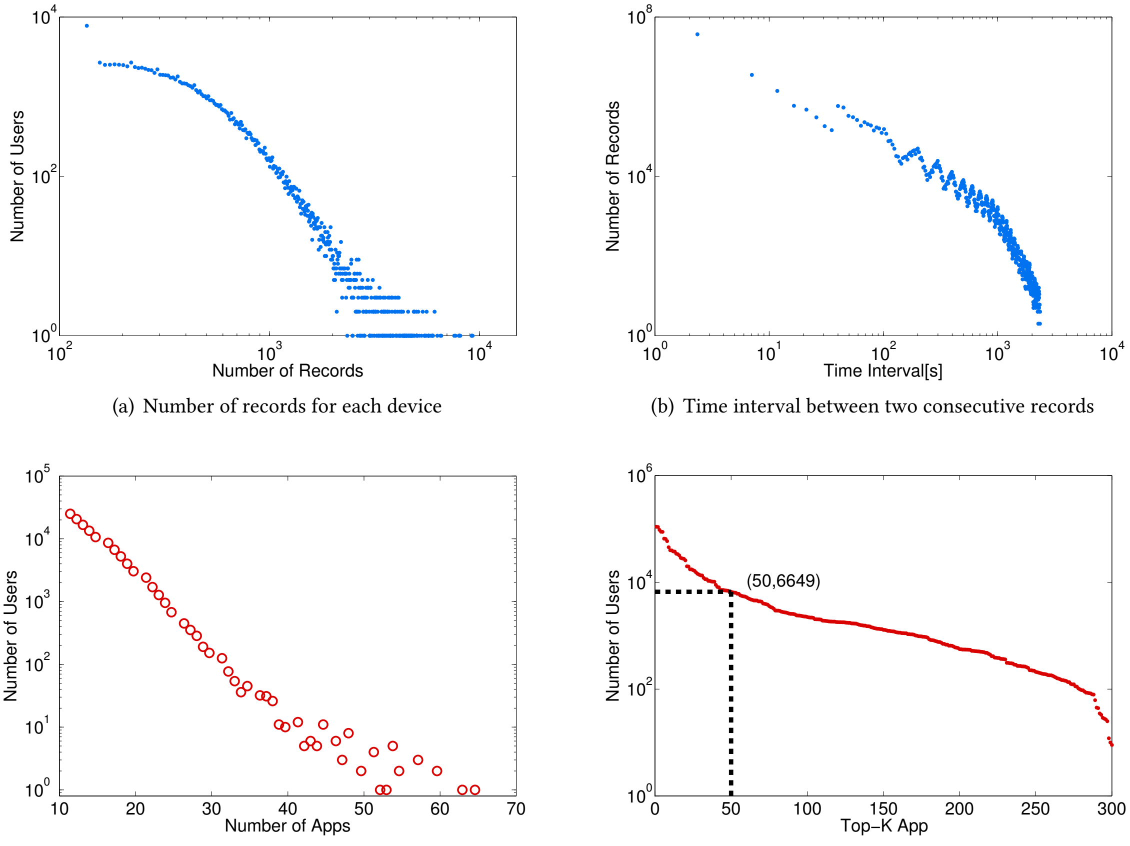
Data Collection and Pre-processing. We obtain a mobile app usage dataset by collaborating with a major ISP in China. During April 20-26 in 2016, our collaborator from the ISP collected cellular network traffic from the city of Shanghai, one of the major metropolitan cities in China. The dataset contains users' access records, *i.e.*, records generated when users issue a network connection request to the cellular towers. Each record contains the anonymized userID, the starting and ending time of the connection, the cellular tower ID, the GPS location of the cellular tower, and the header information of the HTTP and HTTPS requests.

In order to identify the corresponding app for each record, we inspect and analyze the corresponding HTTP headers, using the destination domain and the user-agent as the app identifier. We rely on a systematic tool: SAMPLE [55] to generate the conjunctive rules to match specific apps. SAMPLE applies supervised learning over a set of labelled data streams to generate the rules. It has been shown that SAMPLE can identify over 90% of the apps with a 99% accuracy on average [55]. We learn the conjunctive rules by manually operating a small set of applications to generate data streams. Then we crawled the most popular 2,000 apps across App Store (iOS apps) and Google Play (Android apps) and matched most of our traffic records with these apps. Note that some apps use HTTPS protocols for critical functions (*e.g.*, log-in), but parts of their traffic still use HTTP. In our dataset, more than 95% of the traffic uses HTTP at the time of data collection. We are able to map up to 90% of our traffic records to the specific apps. We believe the labelled dataset, although does not cover all the traffic, is sufficient for our analysis. In total, our final dataset contains 9.4 billion access records from 1.37 million unique mobile users. The access records cover 2,000 different apps and 9,800 cellular towers. The key data statistics are summarized in Table 1.

To examine the data quality, we plot the distribution of the number of data records per user in Fig. 1 (a). We observe that it follows a power-law distribution with the number of records ranging from dozens to several thousand. The most active user owns 9276 records and the average number of records is over 400. In addition, we also plot the distribution of the average time interval between two consecutive records of each user in Fig. 1 (b). This distribution also follows a power-law distribution with a cut-off. 98% of time intervals are below 30 minutes. Fig. 1 (c) further shows the distribution of the number of unique apps per user in the dataset. We show that most users have used dozens of apps during the one week period. In addition, we show the number of users for the most frequently-used 300 apps in Fig. 1 (d). We observe that the top-one app has more than 100,000 users, and even the 50th most popular app has over 6,000 users. This result shows that our dataset covers very popular apps.

Table 1. Major information and key features of our utilized dataset.

Source	Location	Time Duration	Records	Users	Apps	Cellular Towers
Cellular network	Shanghai, China	20th-26th,April,2016	9.4 billions	1.37 millions	2,000	9,800



(a) Number of records for each device

(b) Time interval between two consecutive records

(c) Number of apps for each user

(d) Number of users for the Top-K app

Fig. 1. Illustration of the key statistical characteristics of our dataset.

Our analysis so far shows that this is a large-scale and yet fine-grained dataset to study app usage. Although the dataset comes from one single ISP, it has a sufficient coverage of the mobile users in a city (1.37 million users). More importantly, it contains abundant app usage records to reflect these users' behaviors (9.4 billion records among 2,000 apps). This dataset provides a unique opportunity to quantify the uniqueness of users' app usage behavior at a large scale.

User Demographic Information. One of our goals is to understand the uniqueness of app usage patterns with respect to different user demographics. For this purpose, the ISP collaborator also helps us to look into the network request data, and match with user data from the most popular social network app in China (Sina Weibo).

Based on the Weibo ID in the network request data, the ISP operator is able to link the cellular access data to the users' Weibo profile to obtain demographic information for a subset of users (over 32,000). The demographic information includes gender, city, number of Weibo followers, number of following users, and number of posted status. This enables a deeper analysis on users' app usage.

Ethics. We are very much aware of the privacy implications of using ISP data and social network data for our research. We have taken active steps to ensure ethical procedures for dealing with such sensitive data. First, our dataset is collected through the collaboration with the ISP. All the personal identification information have been stripped (or replaced by a random string) by the ISP operator before handling to us. We never had the direct access to the actual Weibo ID or the mobile user ID. Second, we store all the data in a secure local server behind a firewall protected by strict authentication mechanisms. Only the authors of this paper who are regulated by the non-disclosure agreements have access to the data. Finally, the ISP operator oversees the data processing on the server. Our work has received the approval from the ISP and our local institution.

2.2 Distinguishing Smartphone Users by App Usage

For service providers, a comprehensive analysis of individual users' app usage helps to provide a better understanding of user preference. The result can further help to implement personalized services to benefit users. Given the significant commercial and research values of app usage datasets, service providers may be inclined to publish or share the datasets. Considering the high sensitivity of app usage data [33], we argue that there is a strong need to quantify the uniqueness of app usage for individual users to measure the risk of potential privacy leakage. This is a crucial step before publishing the datasets for research or sharing them to the third parties.

In this paper, we consider the most likely privacy leakage: re-identifying an individual user from an anonymized dataset utilizing external information, *i.e.*, re-identification attack. This attack has been studied in other contexts such as analyzing user mobility traces [32, 34, 57]. In our case, the external information refers to a small number of observations that adversaries have about the victim (*e.g.*, through the victim's social media profile or observing the victim in real life). The external information can be a few records about *when and where the victim once used an app* or *some apps that the victims have on the phone.* By matching the external information with the records in the anonymized dataset, the adversary may successfully re-identify the victim and obtain all his/her records, *i.e.*, direct privacy leakage. Moreover, existing studies have demonstrated the feasibility to further infer users' demographic information such as gender, age and income from their app usage data [30, 42, 43], *i.e.*, indirect privacy leakage.

We focus on direct privacy leakage by assessing the *uniqueness* of the app usage records. For example, after matching the external information with the anonymized records, the adversary will generate a set of candidate users. If the candidate set only contains 1 user, then the adversary re-identifies the targeted user successfully. To quantify how likely a user will be re-identified, we need to measure how unique a user's trace is given a small sample of external observations. Then across a large user population, we can measure the percentage of uniquely re-identifiable users. The "uniqueness" metric is a common way to assess the intrinsic re-identification risk of the dataset [34].

More formally, given an anonymized dataset, we denote the user set as $U = \{U_i\}$ and the corresponding record set as $R = \{R_i\}$ with R_i representing the records of U_i. A fixed (small) number of records sampled from user's real-world records are regarded as the external information, denoted as $E = \{E_i\}$ with E_i representing an external observation that the attacker has about U_i. The percentage of unique users, denoted as P, can be calculated as follows:

$$C_i = \begin{cases} 1 & \left|\left\{R_j : R_j \bigcap E_i = E_i\right\}\right| = 1 \\ 0 & otherwise \end{cases}, \ P = \sum_i C_i/|U|, \tag{1}$$

where C_i represents whether U_i can be re-identified and $|*|$ denotes the size of the set $*$.

In the following, we perform a comprehensive analysis of the individual user's app usage behavior using our city-scale dataset. More specifically, we measure the uniqueness of app fingerprints at both the app-level and behavior-level. At the app-level, we quantify the uniqueness of users by only considering the *set of apps* that the users have. At the behavior-level, we enrich the data records by considering *when* and *where* the apps are used, and assess the impact of spatial and temporal contexts to the app usage uniqueness. In addition, we cross-examine the uniqueness of app usage fingerprints for different user groups (divided by gender, social activity level, and mobility patterns).

3 UNIQUE IN THE APP USAGE FINGERPRINTS

3.1 Simply App Usage Fingerprints

We investigate a simple scenario of app usage fingerprints, where the victim user's apps are known or inferred by a third party. In practice, it is not uncommon for advertisers or app marketplaces to collect information about users' apps on the phone [3, 36]. Such information can be utilized as the external information to re-identify the users from the anonymized dataset.

To quantify the uniqueness of app usage, we start by examining how *similar* users are in terms of their apps. We use two common metrics to measure the difference between installed apps of two users: *Hamming Distance* [51] and *Jaccard Distance* [52]. Hamming Distance measures the minimum number of substitutions required to change the used app set of one user into another 's [51]. Hamming Distance between user U_i and U_j can be defined as follows:

$$HD_{ij} = \left|A_i \cup A_j\right| - \left|A_i \cap A_j\right|, \tag{2}$$

with A_i and A_j representing the set of used apps of U_i and U_j, respectively.

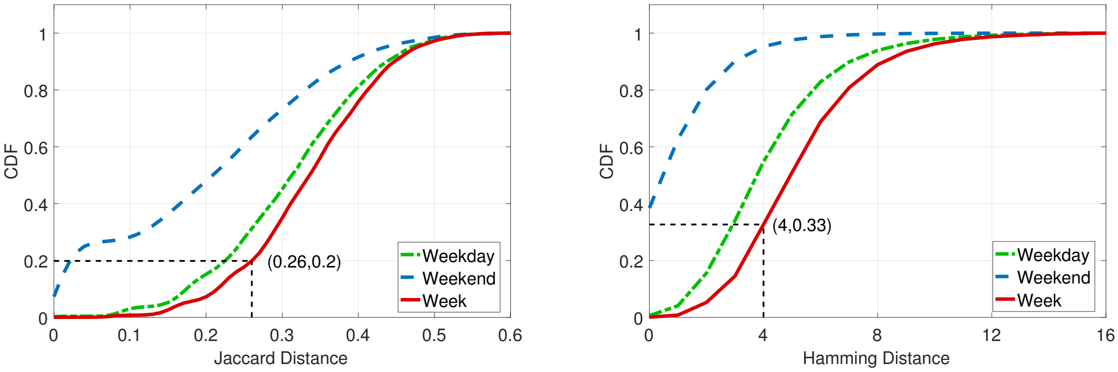
Jaccard Distance measures dissimilarity between two sets [52] and the calculation of Jaccard Distance between user U_i and U_j is as follows:

$$JD_{ij} = \frac{\left|A_i \cup A_j\right| - \left|A_i \cap A_j\right|}{\left|A_i \cup A_j\right|}. \tag{3}$$

Both metrics are useful to measure the differences of the app usage of two users. For example, if two users both used 10 apps and shared 6 common apps, their Hamming Distance is 8 (14-6) and Jaccard Distance is 0.57 (8/14).

For our analysis, we filter out the inactive users who have fewer than 10 apps. For each user, we measure her distance to the *nearest other user* within the population. If the distance to the nearest user is higher, then this user is more unique. We plot the Cumulative Distribution Function (CDF) of Hamming Distance and Jaccard Distance in Fig. 2 (a) and (b), respectively. Our dataset covers a one-week period, which allows us to compare user behavior during the weekdays (5 days), weekends (2 days), and the whole week (7 days). More specifically, we compare the difference of app usage between users during the whole week, on weekdays and on weekend, respectively. Regarding the app usage during the whole week, as shown in Fig. 2, almost all the users have at least 13% different apps compare to their most similar user in the population. About 80% of the users own more than 26% unique apps compared to their "nearest" users. Fig. 2 (b) shows a similar trend. We observe that more than 67% of the users have a Hamming Distance bigger than 4 to their "nearest" users.

Fig. 2 (a) and (b) also show the differences between weekdays and weekends. There is a big gap between the green line and the blue line, indicating that users' app usage is more "unique" during weekdays compared to that of the weekends. Generally speaking, on weekdays, the used apps are likely to be related to the nature of the users' professions. Intuitively, the commonly used apps among different professions vary. During the weekends, we suspect user activities become more "synchronized" to enjoy the break by listening to music,



(a) Jaccard Distance between the nearest two users

(b) Hamming Distance between the nearest two users

Fig. 2. Results of the differences between users' apps by utilizing the metrics of Hamming Distance and Jaccard Distance.

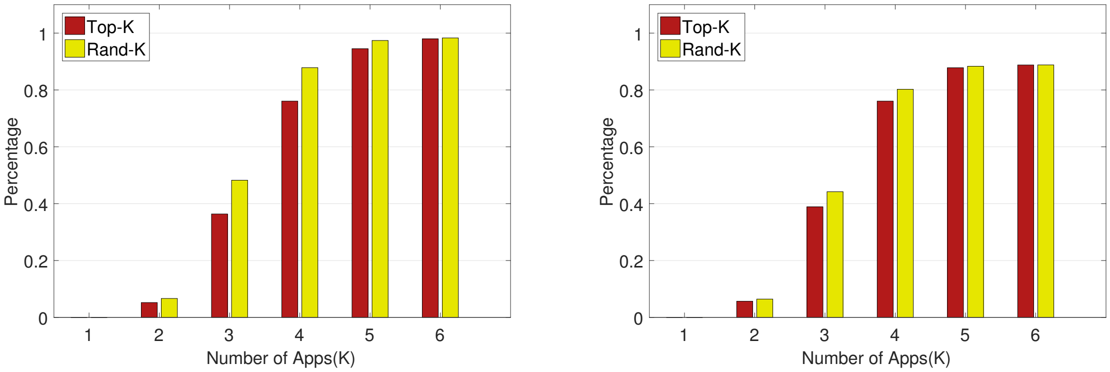
Table 2. Top apps used by users (and % of users) on the weekdays and on the weekends (sorted based on weekend statistics).

App	Category	Weekday	Weekend
QQMusic	music	11.57%	22.25%
XianYu	shopping	2.25%	4.64%
JingDo	shopping	1.33%	1.96%
Wechat	chatting	19.87%	1.41%
WangzheRongYao	game	0.49%	0.78%
LuShiChuanShuo	game	0.63%	0.77%
DiDaPinChe	travel	2.23%	0.32%
GaoDeMap	travel	5.72%	0.56%

watching videos or going shopping. To validate our intuition, we examine the most frequently used 50 apps on weekdays and weekends. As shown in Table 2, we find that music apps (QQMusic), online shopping apps (XianYu, JingDo), gaming apps (WangZheRongYao, LuShiChuanShuo) are used more frequently and more widely during the weekends compared to that on weekdays. The app usage is more disperse during weekdays with a focus on online chatting apps (Wechat) and travel apps (DiDaPinChe, GaoDeMap). The results suggest users are more distinguishable during weekdays.

Next, we further assess the uniqueness of app usage fingerprints with different levels of external information. More specifically, we investigate the possibility of distinguishing individual users by choosing only a small subset of apps in their app lists. We use two different strategies to obtain the subset of apps: selecting the top K frequently-used apps (Top-K), and randomly selecting K apps (Rand-K). Note that apps in the Top-K set are already ordered by their usage frequency. For comparison, we also rank the apps in the random-K set based on usage-frequency.

Fig. 3 (a) shows the app usage uniqueness based on the Top-K apps. We observe that 76% of users can be re-identified by the Top-4 apps. By considering the Top-6 apps, we show that about 97% of the users are distinguishable. Interesting, random-K returns an even higher level of data uniqueness. For example, over 88% of



(a) 100% of app records

(b) 10% of app records

Fig. 3. Results of the uniqueness of app usage fingerprints by considering 100% or only 10% of app records in our dataset.

the users can be distinguished by the Rand-4 apps. About 98% of the users are uniquely distinguishable using Rand-6 apps. The result makes sense since Rand-*K* apps are more likely to show user interests and preferences. Instead, the Top-*K* apps are more likely to contain the few vastly popular apps that are used by almost everyone (*e.g.*, WeiChat). Therefore, even with the same number of apps, Rand-*K* apps can achieve a higher probability of re-identifying a user.

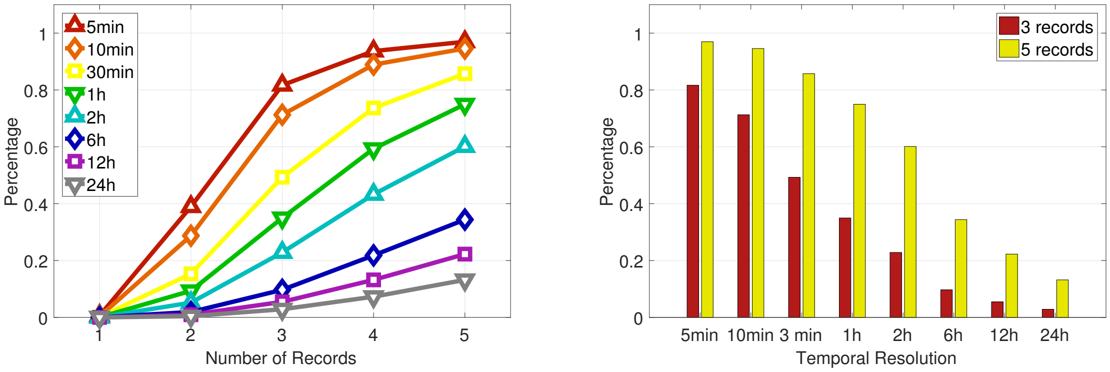
The result in Fig. 3 (a) is based on 100% of our dataset. However, in practice, data owners may only have a partial record of users' app usage. Therefore, in addition to Fig. 3 (a), we also measure the uniqueness of app fingerprints when only partial app usage records are known. As shown in Fig. 3 (b), we randomly select 10% of the app records to form a smaller dataset. Then we investigate the uniqueness of app fingerprints measured by the Top-*K* and Random-*K* methods. We observe that Rand-4 apps are still able to uniquely distinguish 80% of the users. Top-6/Rand-6 can uniquely distinguish 88% of the users. The level of uniqueness only drops slightly compared to Fig. 3 (a), indicating that the privacy leakage of a partial dataset is still very severe.

In summary, we have three important observations. First, the users' app usage fingerprints are quite unique. 88% users can be distinguished by just 4 random apps. In addition, even if a third-party only has the access to 10% of app usage records, 4 random apps can still distinguish 80% of the users. Second, apps used on weekdays are more likely to uniquely distinguish users compare to apps used on weekends. Third, users are more likely to be uniquely distinguished by *a random subset of their apps* compared to *their most-frequently used apps*. Overall, our result suggests that the app usage is quite unique, and privacy-preserving schemes are needed before sharing such datasets.

3.2 Spatial or Temporal App Usage Fingerprints

Next, we investigate how the *time* and *location* of the app usage would affect the uniqueness measurement. For example, if the time or location information is obtained, *i.e.*, several records about when the target user uses a specific app, or where the targeted user uses a specific app, the adversaries may or may not have a better chance to uniquely identify a user.

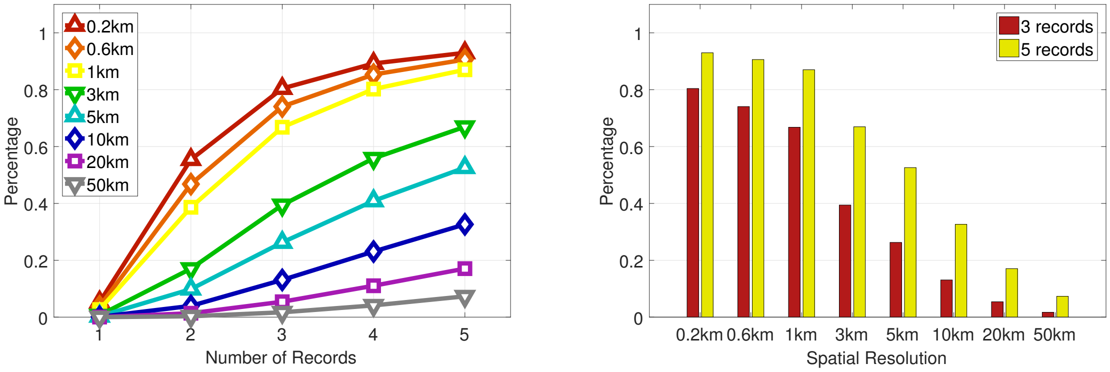
In order to perform the above analysis, we need to first define a way to match temporal and spatial records. This requires defining the temporal and spatial thresholds (or resolution) for the matching. More specifically, we define a temporal resolution threshold T_t. If the time difference between two records is smaller than T_t, then



(a) App uniqueness with time information

(b) The impact of temporal resolution

Fig. 4. Results of the uniqueness of app usage fingerprints with time information by considering 100% of app records in our dataset.



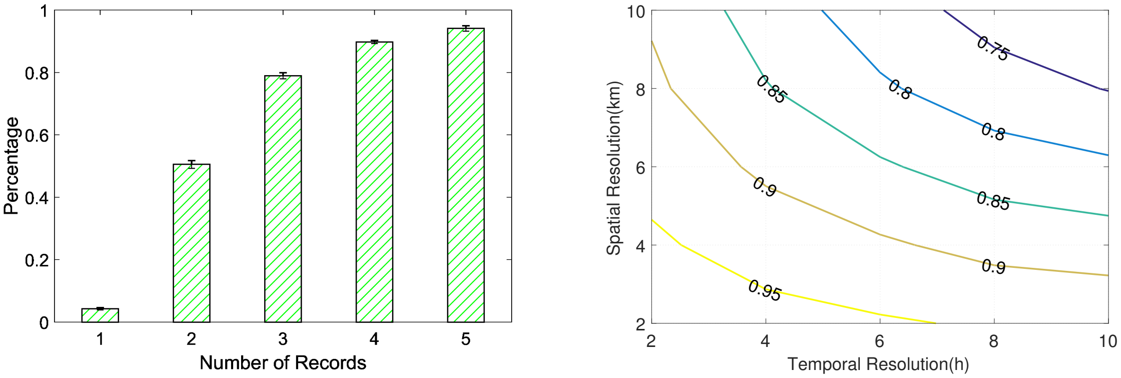
(a) App uniqueness with location information

(b) The impact of spatial resolution

Fig. 5. Results of the uniqueness of app usage fingerprints with location information by considering 100% of app records in our dataset.

we regard the two records temporally matches. Similarly, we define a spatial resolution threshold T_s. If two geolocation records have a distance smaller than T_s, we would consider the two location records match with each other.

Fig. 4 shows the percentage of unique users regarding app usage with respect to the time information. In Fig. 4 (a), the lines with different colors represent different time resolutions (T_t), ranging from 5 minutes to 24 hours. Using a fine-grained temporal resolution (5 mins), we show that over 82% of the users can be distinguished by only 3 randomly selected records. When we consider 5 records, 98% of the users are unique. If we lower the temporal resolution to 1 hour, then fewer users can be uniquely re-identified. About 35% of the users are still unique given 3 random records; 75% of the users are considered unique based on 5 random records.



(a) App uniqueness with time and locations

(b) Uniqueness under different data resolutions

Fig. 6. Results of the uniqueness of app fingerprints with both time and location information.

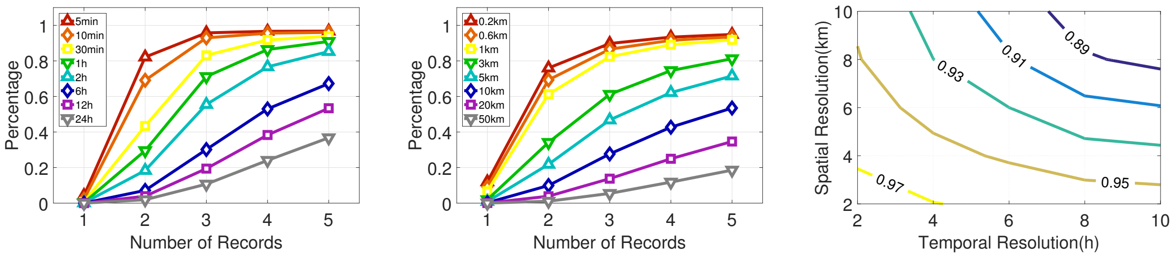
Note that when we gradually change the time resolution from 2h to 24h, the uniqueness decreases rapidly from 60% to 13%, even when we consider 5 random records as external information. Fig. 4 (b) clearly shows the impact of temporal resolution towards the app fingerprint uniqueness. There is a near linear decreasing tendency of uniqueness level when the temporal resolution changes from 5 minutes to one day.

To measure the uniqueness of app usage with respect to the spatial information, we plot Fig. 5. In Fig. 5 (a), we show that using a spatial resolution of 0.2km, 80% of the users can be uniquely re-identified using 3 random records; 95% of the users are unique based on 5 random records. As expected, when the spatial resolution becomes coarser, the level of uniqueness also decreases. For example, with a resolution of 3km, the percentage of the unique users falls below 40% based on 3 random records. Even with 5 records, the percentage of unique users is now below 70%. If we take 50km as the threshold, only 10% of the users are distinguishable by 5 records. Fig. 5 (b) shows the same trend with a different angle. Cleary, the uniqueness of app fingerprint decreases rapidly when we lower the spatial resolution.

In summary, our results reveal that mobile users have distinguishing temporal and spatial patterns in their app usage. Given a reasonable set of time and spatial resolutions, we show that over 80% of the users can be uniquely distinguished by only 3 randomly selected records (regardless spatial or temporal records). Compared with the app-level information (section 3.1), the spatial and temporal information shows a much stronger distinguishing power. In addition, we show that lowering the time and spatial resolutions help to decrease data uniqueness. From the privacy protection perspective, it is a desired step (although not sufficient) to lower the data resolutions before sharing the data.

3.3 Spatiotemporal App Usage Fingerprints

Now we analyze the uniqueness of app usage fingerprints by considering the time and location information together. We assume the following external information obtained by the adversaries include several records about when and where the target user uses a specific app. In practice, it is very likely that adversaries can only obtain a rough estimation rather than the exact fine-grained data records. To this end, we lower the temporal and spatial resolutions to 5h and 5km respectively. We run our experiments under such coarse-grained data resolution to avoid overestimating the risk of privacy leakage.



(a) App uniqueness with time

(b) App uniqueness with locations

(c) App uniqueness with time and locations

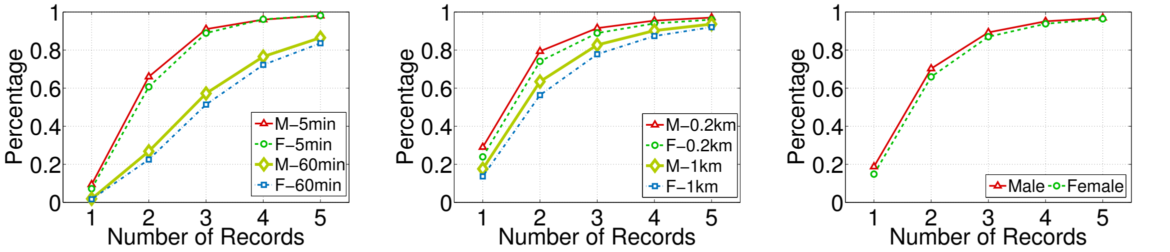
Fig. 7. Results of the uniqueness of app fingerprints with time and location information by considering only **10%** of app records in our dataset.

Fig. 6 (a) shows the percentage of the uniquely distinguishable users using different numbers of records. Recall that each record is characterized by a location, a timestamp and the app name. We show that using 2 random records, we are able to re-identify half of the users. This is a much higher uniqueness compared to that using a single source of information. We also show the standard deviation of the results through multiple rounds of repeated experiments. The small variance also confirms the stability of the uniqueness measure.

We further examine the impact of different spatial and temporal resolutions in Fig. 6 (b). Here we randomly select 4 records as the external information. A series of contour lines show the descending data uniqueness with respect to the more coarse-grained data resolutions. The yellow line represents different combinations of spatial and temporal resolutions with the uniqueness level of 95%. For example, using 7hour temporal resolution and 2km spatial resolution, the yellow line shows that 95% of users can be uniquely re-identified. From Fig. 6 (b), we observe that when the temporal resolution is lowered to 10h and the spatial resolution is set to 8km, the uniqueness is less than 20%. A figure like this can help us to choose suitable combinations of spatial and temporal resolutions to meet the specific requirement of privacy protection.

After analyzing the upper bound of privacy leakage, now we further measure the uniqueness in a constrained setting. More specifically, we take only 10% of the data records from each individual user to measure the uniqueness in a smaller and sparser dataset. The results are shown in Fig. 7. Similar to before, lowering temporal or spatial resolution helps to decrease the uniqueness of app fingerprints in this setting. For example, using 5 random records, the percentage of unique users decreases from 98% down to 38% when the temporal resolution changes from 5 mins to 24 hours. Similarly, the uniqueness level also decreases by 75% when we lower the spatial resolution from 0.2 km to 50 km. If we compare with the results from the full dataset, we find that that app fingerprints in this smaller datasets are more unique. For example, with 24h resolution and 5 records, we can distinguish 38% of the users, while the corresponding number is 15% using the full dataset. Fig. 7 (c) further confirms that, under the same spatial and temporal resolutions, the uniqueness is higher for the smaller dataset than that of the full dataset (*e.g.*, 0.91 vs 0.80). The intuition is that in a smaller dataset, it is less likely for a user to be exactly the same as another user due to the overly sparse data records. To some extend, the results also indicate that analyzing a smaller dataset may have biased conclusions when measuring the uniqueness of user behavior.

In summary, we show that the uniqueness of app usage can be significantly boosted by considering the time and location information. With only 2 random records, almost half of the users can be uniquely re-identified (even under a relative coarse grained spatial-temporal resolution). We also show that a sparser dataset with only 10% of users' app records produces an even higher uniqueness level. To really reduce the potential privacy leakage, the resolution of the spatial-temporal data records should be severely reduced.



(a) App uniqueness with time information (b) App uniqueness with location information (c) App uniqueness with time and locations

Fig. 8. Results of uniqueness differences between men and women.

4 UNIQUENESS OF DIFFERENT USERS

So far, through a quantitative assessment, we have demonstrated the highly distinguishable app usage patterns of individuals that can lead to privacy leakages. In this section, we want to further break down the uniqueness measurement by considering different user groups. Our goal is to investigate whether there are different patterns in terms of the app usage uniqueness for users of different gender, social activity levels, and mobility habits. Studying the impact of user attributes on app usage is important. In many existing privacy-preserving schemes [4, 19, 29, 48], the system designers often assume all the individuals have equal risks of privacy leakage and provide the same protection strategy. However, if the behavioral uniqueness of different user groups varies significantly, more customized strategies may be necessary to preserve user privacy. For example, uber drivers' movement traces may be different from normal users, making them more distinguishable. By understanding the behavioral statistics of different user groups, we seek to provide results to inform more personalized privacy protection mechanisms.

Based on the user demographics obtained from online social networks (Sina Weibo), we aim to measure the app fingerprint uniqueness for the following user groups:

- Male and female users;
- User groups of a low/medium/high activity level in the social network;
- User groups of different mobility patterns.

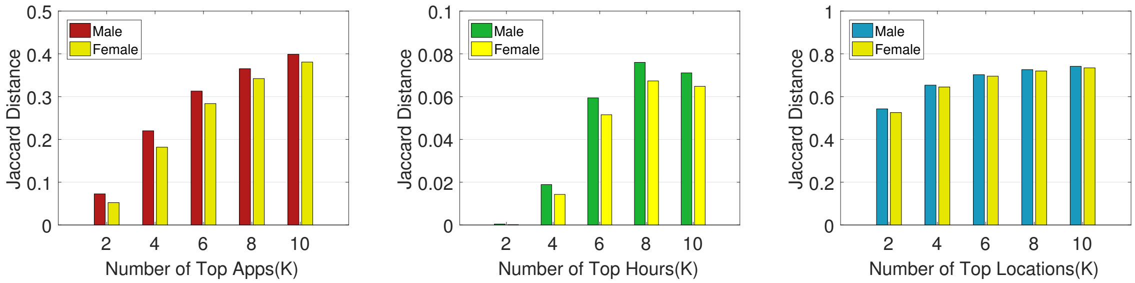
In the following, we measure the uniqueness of app usage fingerprints and compare the results of different user groups. Through the analysis, we also seek to infer the underlying reasons for the observed differences.

4.1 Gender

We start by analyzing the potential influence of *gender*. Here we consider app usage along with time and location information to measure the uniqueness. We compare the results of female and male user groups in Fig. 8. Under two different time resolutions, Fig. 8 (a) compares the app uniqueness between male and female users. With a 5-min temporal resolution, the results for male and female users are quite close. When considering a coarser-grained 60-min resolution, male users tend to have a higher uniqueness level. For example, 2 random records are able to distinguish 26.8% of male users, which is 5.3% higher than that of female users. If we use 3 random records, 58% of male users can be distinguished, which is 3% higher than that of female users. Similar results can also be observed in Fig. 8 (b), where male users show a higher level of uniqueness in app usage and location. In Fig. 8 (c), we look at the app usage fingerprints with both time and location information. We show that the gap between female and male becomes smaller (male users' uniqueness level is still higher).

Table 3. The significance of uniqueness differences between men and women, in the case that 2 random records are given to distinguish app users.

Data Resolution	Uniqueness Results		p-value	H
	Male	Female		
5min	[0.6677, 0.6585, 0.6723, 0.6618, 0.6640]	[0.6008, 0.6046, 0.6058, 0.6035, 0.6068]	1.24e-8	1
60min	[0.2666, 0.2642, 0.2674, 0.2701, 0.2703]	[0.2265, 0.2225, 0.2230, 0.2217, 0.2296]	1.33e-8	1
200m	[0.7877, 0.7923, 0.7900, 0.7919, 0.7982]	[0.7381, 0.7468, 0.7406, 0.7384, 0.7423]	2.24e-8	1
1km	[0.6410, 0.6345, 0.6434, 0.6379, 0.6363]	[0.5630, 0.5679, 0.5623, 0.5643, 0.5595]	4.09e-10	1
5h+5km	[0.7052, 0.7137, 0.7047, 0.7022, 0.6971]	[0.6416, 0.6391, 0.6411, 0.6418, 0.6363]	1.65e-8	1

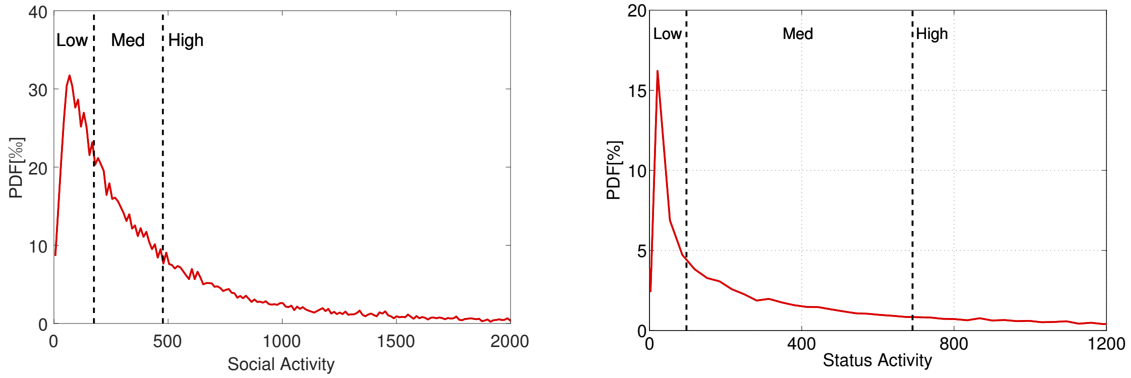


(a) Jaccard Distance of simply app usage (b) Jaccard Distance of temporal app usage (c) Jaccard Distance of spatial app usage

Fig. 9. Results of differences between men and women's app usage behaviors in multiple domains.

To examine the statistical significance of the observed differences, we perform a standard *t-test* to compare the mean values of the app usage uniqueness for the two user groups. For this *t-test*, the null hypothesis is that the means are not different. Setting a certain significant level, the returned value $H = 1$ means *t-test* rejects the null hypothesis at that significance level. We set the significant level to 5%, and compare the differences of the uniqueness metric for male and female groups. We repeated our experiments five times and take the mean to avoid biases. For simplicity, we present the results where we use 2 random records to compute the uniqueness, as shown in Table 3. We observe $H = 1$ for all cases and the p-value is much smaller than 0.05, which confirm the statistical significance of the observed differences. In other words, the differences between male and female are not observed by chance. In addition, we performed additional experiments where we take different settings into consideration (*e.g.*, number of records, temporal and spatial resolutions). The results are consistent with the above and omit them for brevity. In this way, we confirm that there is a significant difference between female and male users' uniqueness level of their app usage.

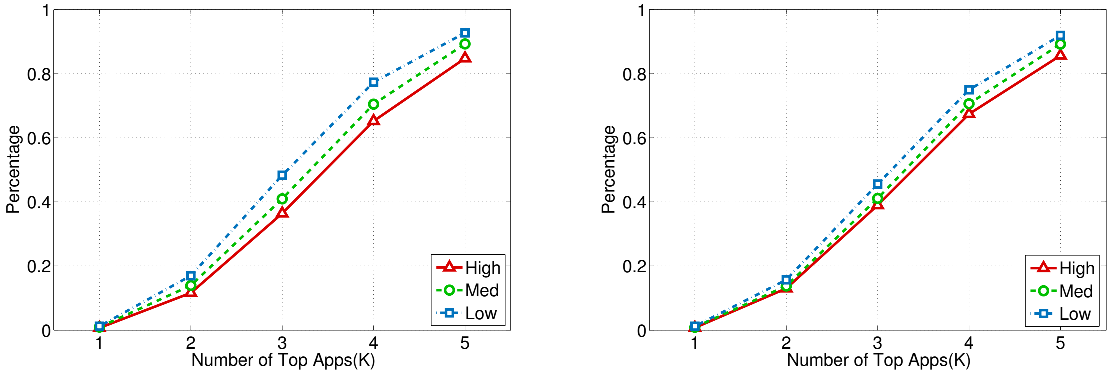
To explore the possible reasons behind the observed differences, we break down the analysis results based on their used app, and time and location information. More specifically, we compare male and female users in their used apps, and time and locations of app usage to examine where the difference comes from. To this end, we extract the Top-K apps, Top-K hours and Top-K locations from each user's data records. Then we calculate the Jaccard Distance between the given user and the "nearest" user in our dataset. The perform the analysis for male and female users separately and the results are presented in Fig. 9. As shown in Fig. 9 (a), we observe that "Top-K apps" exhibit the larger dissimilarity between male and female users. For example, when measuring the top-4 apps, male users have a 4% higher Jaccard Distance than that of female users. This indicates that male users are more likely to use a diverse set of apps.



(a) PDF of social activity

(b) PDF of status activity

Fig. 10. Results of Probability Distribution Function(PDF) of user activity in social network.



(a) Uniqueness for groups of different social activity

(b) Uniqueness for groups of different status activity

Fig. 11. Results of uniqueness differences between user groups of different activity in social network.

In Fig. 9 (b), we further investigate users' preferred hours in a day to use apps. Again, male users still have a higher Jaccard Distance, indicating that their temporal patterns are more diverse. However, compared with the results of "top-K app" analysis (Fig. 9 (a)), the differences are much smaller. Similarly, Fig. 9 (c) shows the Jaccard Distance in terms of top-K locations. We show that male and female users are very close and don't present apparent differences.

In summary, our results show that male users have more distinguishable app usage behaviors than female users. Our breakdown analysis suggests that the differences are more likely to be caused by the choices of apps. Male users tend to use a more diverse set of apps (that are different from other people).

4.2 Activity in Social Network

With the increasing popularity of social networking applications, social network activity level is also an important attribute of individuals. We consider two metrics to represent user activity level in online social networks: social activity and status activity. More specifically, we define the number of friends, including following users and followers, as the social activity. We also regard the number of posted status as status activity. Using these two metrics, we divide the users into three groups respectively, *i.e.*, user groups of high/medium/low level of social activity and status activity. For the fairness of the measurement, we set the thresholds so that each group contains a relatively equal number of users.

Fig. 10 shows the Probability Distribution Function (PDF) of user activity in social networks. We observe the PDF first reaches a peak and then follows an exponential distribution. The ranges of social activity for low/medium/high groups are [0,175), [175,477) and [477,2000]. The distribution of status activity also shows a similar distribution and the ranges for each group are [0,97), [97,691) and [691,1200]. Considering the app usage uniqueness (Top-K apps), we plot Fig. 11 (a) and (b).

Surprisingly, we observe the user groups with a high social activity level and user groups with a high status activity level both have a lower level of uniqueness. For example, in Fig. 11 (a), 38% of the high-activity users can be distinguished by Top-3 apps, while the percentage of unique users among medium-activity and low-activity groups is 40% and 46% respectively. The trend is similar in Fig. 11 (b) (the differences between the three groups are smaller). A possible explanation is that highly active users are likely to follow hot trends and use popular apps more often. This helps to reduce their overall uniqueness of app usage fingerprints, and mix themselves into the "crowd". We have examined the detailed app lists of different user groups, which helps to verify our intuition.

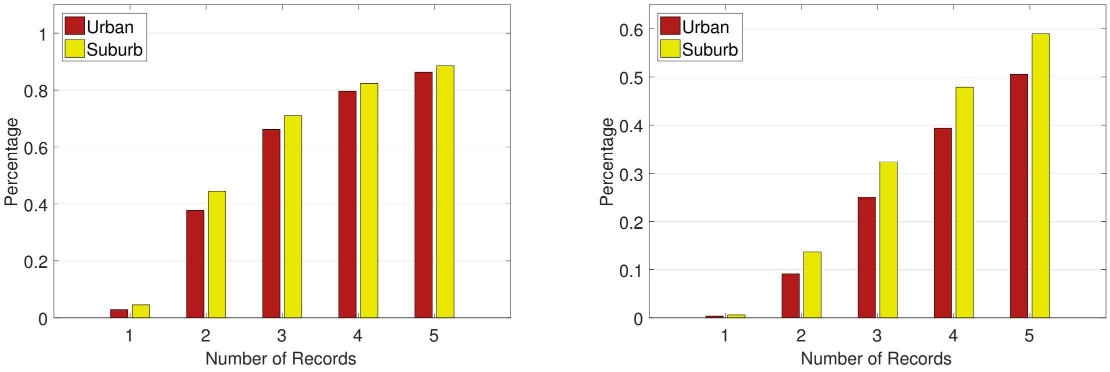
In summary, our results show that users with a high activity level in social network are likely to be *less unique* in their app usage. A possible explanation is that active users are more likely to use popular mobile apps, which helps to reduce their uniqueness.

4.3 Mobility Patterns

The third factor we investigate is the mobility patterns. Many studies have shown that mobility patterns of human movement trajectories can reflect personal habits of individual users [35, 41, 49]. Especially in metropolitan cities, different mobility patterns may reflect the individual's profession or lifestyles. To this end, we investigate whether there are key differences in the app usage uniqueness between user groups of diverse mobility patterns.

For this analysis, we divide users into two groups based on their residential locations: urban residents and suburb residents. Generally speaking, trajectories of suburb residents are more unique, since the "crowd size" in suburb areas is smaller for people to hide. To this end, we explore users' app usage uniqueness along with spatial information. More specifically, with a spatial resolution of 1 km and 5 km, we compare app usage uniqueness (based on app and location) of urban and suburb users in Fig. 12. As expected, suburb users have a higher uniqueness than urban users under both 1 km and 5 km resolutions. Under the 5 km resolution, the differences are larger. Since many suburb users have to travel a long distance to work, they are more likely to have unique app records given specific locations. In addition, from 1km to 5km, urban users' app uniqueness decreases more rapidly. This is because trajectories of urban users are more "concentrated" and their app usage uniqueness (with location) is more sensitive to spatial resolution change. On the contrary, suburb users' trajectories are more "decentralized", and thus app uniqueness is still high even the spatial resolution is decreased to 5 km.

Besides the simple division of urban/suburb users, We further adopt two common metrics to represent individual mobility in trajectory: *Radius of Gyration* [39] and *Mobility Entropy* [37], to divide users into different groups and investigate whether there exists some differences regarding app uniqueness. For an individual, radius of gyration



(a) Uniqueness with locations (resolution: 1km)

(b) Uniqueness with locations (resolution: 5km)

Fig. 12. Results of uniqueness (app+location) differences between urban and suburb residents.

Table 4. Top-10 apps and their penetrations for three user groups with different radius of gyration.

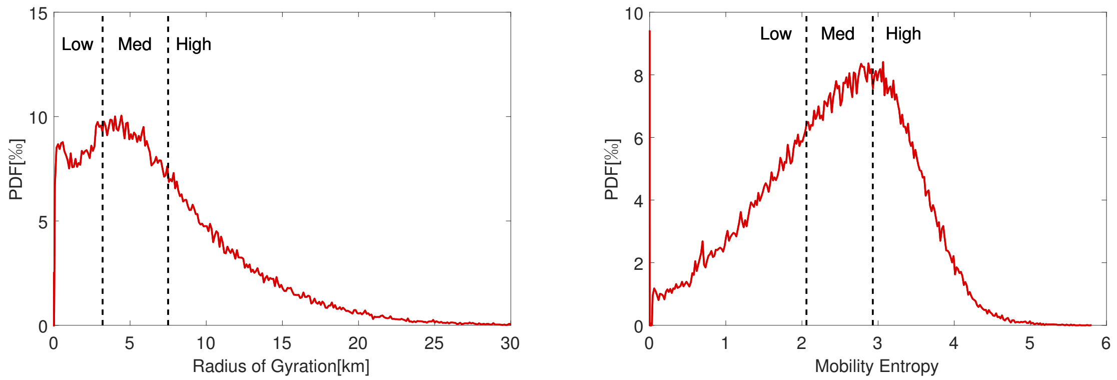
App	**User Group**		
Top-*K*	High	Medium	Low
1	Weibo (7.13%)	**Wechat (10.95%)**	**Wechat (14.51%)**
2	**Wechat (6.55%)**	Weibo (6.56%)	QQ (5.76%)
3	QQ (6.45%)	QQ (5.98%)	**Taobao (5.31%)**
4	JinRiTouTiao (6.13%)	QQMusic (5.26%)	Weibo (5.04%)
5	DiDaPinChe (5.34%)	DiDaPinChe (5.22%)	GaoDeMap (4.52%)
6	GaoDeMap (5.17%)	QQNews (4.90%)	QQNews (3.90%)
7	QQMusic (4.64%)	GaoDeMap (4.84%)	QQMusic (3.89%)
8	QQNews (4.56%)	JinRiTouTiao (4.32%)	DiDaPinChe (3.87%)
9	iQiYi (3.69%)	**Taobao** (4.27%)	MobileMap (3.64%)
10	WangYiNews (3.22%)	iQiYi (3.05%)	**DaZhongDianPing (3.58%)**

for his trajectory is calculated as follows:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r_i} - \mathbf{r_{mean}})}, \tag{4}$$

where $\mathbf{r_i}$ is the position of the i-th locatipn and $\mathbf{r_{mean}}$ is the mean position of all the visited locations. As for entropy, we first count the visited frequency for each location and then calculate the entropy of those visited frequencies.

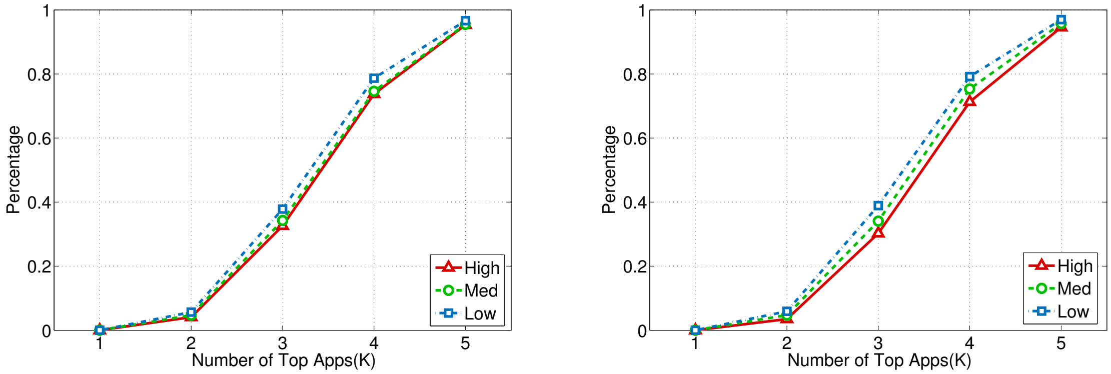
Fig. 13 shows the distribution of user's radius of gyration, and mobility entropy. As shown in Fig. 13 (a), the range of radius of gyration for low/medium/high groups are [0,3.2km), [3.2km,5.5km) and [7.5km,30km]. In Fig. 13 (b), the distribution of mobility entropy seems to be a normal distribution and the ranges for each group are [0,2.06), [2.06,2.93) and [2.93,6]. Here we also show the uniqueness of the app usage considering Top-*K* used apps. Fig. 14 (a) and (b) show the uniqueness measurements for different groups. Interestingly, we find that users



(a) PDF of radius of gyration

(b) PDF of mobility entropy

Fig. 13. Results of Probability Distribution Function (PDF) of user mobility in trajectory.



(a) Uniqueness for groups of different Radius of Gyration

(b) Uniqueness for groups of different Entropy

Fig. 14. Results of uniqueness differences between user groups of different mobility in trajectory.

with a small radius of gyration and a low mobility entropy are likely to be more unique. In Fig. 14 (a), considering the Top-4 apps, we observe that users with a small radius of gyration have a 4% higher uniqueness level than the other two groups. In Fig. 14 (b), the differences among these groups are even bigger when we consider Top-3 or Top-4 apps. This is again counter-intuitive: users with a high mobility entropy turns out to be less unique compared to those with lower mobility entropy.

In order to examine the underlying reasons, we look into the Top-30 apps and the percentage of users that use them in each group. We observe some interesting differences. For brevity, in Table 4, we demonstrate the detailed information of the Top-10 apps for the three groups with different radius of gyration. First, we observe that Wechat, a popular online chatting app, increases its percentage of coverage when the user group's mobility level gradually decreases. This indicates that users with a low mobility level are more dependent on online chatting apps such as Wechat to stay in touch with their friends. Secondly, we observe that Taobao, one of the biggest online shopping apps in China, is at the 3rd place for low-mobility user group, and then goes to the 9th place for the medium-mobility user group. Taobao is even dropped out of top-1o for the high-mobility group. One popular app DaZhongDianPing only appears in the top 10 of the low-activity user group. This app provides a service of delivering take-out food to user's home. Apps like Taobao and DaZhongDianPing are more widely used by people who move within a smaller range indicating that they prefer shopping through online services. Such habits make their app usage more easily distinguishable.

In summary, our measurements show that suburb residents have a more unique spatial app usage fingerprint, compared to that of urban residents. We also find that users who are less "mobilized" show a higher level of uniqueness, which is counter-intuitive. A further analysis on their preferred apps shows that that people in the "low mobility" group are more dependent on online chatting apps to keep in touch with friends, and are more likely to buy products and have food delivered through online services. These habits make them more distinguishable.

Overall, our results in this section show that the uniqueness of app usage patterns clearly varies among different user groups. User demographics, social activity level and mobility patterns all contribute to these differences. To the best of our knowledge, this is the first work to study the influence of user attributes on the uniqueness of app usage fingerprints. Our results provide valuable insights to better understand app usage preference, and inform more customized privacy-preserving schemes for app usage datasets.

5 RELATED WORK AND DISCUSSION

5.1 Human Behavior Uniqueness

Due to the diversity of individual lifestyle and preferences, users leave unique behavior fingerprints in different platforms of online services. Current research has studied the uniqueness of human behavior from various aspects, including mobility patterns, web browsing histories, and mobile device sensor fingerprints.

In terms of mobility patterns, existing works studied the uniqueness of various mobility datasets to reveal the potential risk of privacy leakage in the re-identification attacks [32, 34, 57]. Results have demonstrated that individual mobility trajectories are quite very unique. Regarding user demographics, Montjoye et al. further studied the impact of gender and income on the likelihood of re-identification for individual trajectories [34]. As for web browsing behavior, empirical studies showed that the rich information that the browsers send to a website can help to identify users even without cookies [13, 56]. In addition, Su et al. showed that it is possible to de-anonymize web browsing users with social networks [46]. For mobile sensor fingerprints, Bojinov et al. demonstrated that it is possible to identify a user by only utilizing the accelerometer imperfections and distortions in the speaker-microphone system [7]. Moreover, Weiss et al. showed that accelerometer information of phones is able to predict user demographics, such as gender, height and weight [50].

However, in terms of mobile app usage behavior, very limited work has studied the uniqueness of app usage fingerprints. Blaszkiewicz et al. measured the difference of app usage patterns by simply comparing the used apps between mobile users [6]. They do not investigate the upper bound of privacy leakage in the re-identification scenario and does not take the time and location information into consideration due to limitations on their data. In contrast, our work studies the uniqueness of app usage behavior along with spatial and temporal information on a large-scale dataset. Our results provide a comprehensive assessment of its potential risk of re-identification. In addition, we also evaluate the impact of user attributes, *i.e.*, gender, social activity level, and mobility patterns, on the uniqueness of the app usage behavior.

5.2 App Usage Behavior Modeling

In recent years, a large body of studies have sought to understand smartphone users' app usage behaviors. For example, Li et al. studied how individuals download, install and use different applications [25], Falaki et al.

investigated how many daily interactions individuals have with different apps [15]; [9, 16] looked into how app usage varies with contexts; and other studies investigated how to predict which apps users are likely to install or use [16, 22, 24, 27, 44].

To study individual app usage, Xu et al. identified traffic from distinct marketplace apps based on HTTP signatures and presented aggregate results on their spatial and temporal prevalence, locality, and correlation [53]. In addition, other studies focused on discovering app usage patterns: Jones et al. studied how often individuals revisit a specific app [23]. Rachuri et al. found that pairwise apps that were frequently used together [38]. Furthermore, existing works also focused on understanding usage characteristics of different user groups. Zhao et al. discovered different kinds of smartphone users through their application usage behaviors [58], Blaszkiewicz et al. differentiated smartphone users by their installed app and Top-60 apps [6], and Malmi et al. demonstrated user groups with different demographics differ from each other in app usage behaviors [30].

However, most of these works are based on small-scale datasets which inevitably suffer from major biases. Due to the limited scale, these studies cannot analyze the privacy and anonymity perspective in terms of the app usage. In our work, we focus on the unique individual app usage behaviors along with spatial and temporal information. Our primary focus is the potential risks of privacy leakage on a much larger mobile user population.

5.3 Mobile Data Privacy Protection

To understand how to anonymized mobility datasets, existing works focus on defending against re-identification attack to protect user privacy. Many privacy-preserving schemes are proposed to reduce the uniqueness of traces, such as *k-anonymity* [18, 48], *l-diversity* [28, 47], *t-closeness* [26, 40]. A number of specific techniques have been proposed: identifier replacement [45], generalization [18, 31], suppression [17], or perturbations and permutations [1, 10], to make sure the released datasets comply with the guideline of *k-anonymity* and *l-diversity*. In addition, many works studied how to protect data privacy by satisfying differential privacy, which protects the membership information of individuals in the dataset. Dwork et al. [11] formally demonstrated that two useful methods to guarantee differential privacy by adding noise and adopting exponential mechanism. [54, 59] summarized the diverse release mechanisms of differential private data publishing. In addition, Acs et al. [2] implemented a differential privacy scheme on aggregated population density data. The system provides a provable privacy guarantee that the adversaries cannot determine whether a given mobility record is in the dataset or not. Furthermore, Dwork et al. [12] proposed a privacy framework that prevented the adversaries to infer the membership information of an individual given the statistic information of a DNA dataset.

In this paper, our main purpose is to investigate the upper bound of privacy leakage in anonymous app usage data. Our results show that the basic generalization technique, *i.e.*, lowering the spatial and temporal resolutions, can help to reduce the data uniqueness and the potential privacy leakage. Now we give a more detailed discussion about possible protection solutions to prevent privacy leakage from anonymous app usage data.

- *Perturbation*: Perturbation is a popular privacy preserving solution to protect data privacy, which requires adding noises to the original data [1, 10]. On the one hand, for our app usage data, we can add random noises to achieve differential privacy to protect user privacy [11, 54, 59]. For example, when publishing anonymous app usage data, we may adopt RAPPOR [14], a randomized response mechanism utilized by Chrome of Google, to guarantee differential privacy. More specifically, to answer the question whether user i has used app j, we can first flip a coin and answer truthfully if the coin comes up heads. If the coin comes up tails, we can always answer "No". In this way, we can conduct experiments for all the users and apps, and finally gather all the results to publish an anonymous app usage dataset. This published dataset achieves differential privacy and can be utilized to analyze aggregated results such as the trend of app usage, but it can prevent the leakage of user privacy. On the other, we may disturb important app usage information and eliminate sensitive records. For example, noises can be added to financial and health-care app records, or very unique app records with sensitive time and location information. In this way, perturbation guarantees the most important and sensitive information is not leaked to protect user privacy as much as possible.

- *Generalization*: Generalization is also a widely adopted protection technique to preserve mobile users' privacy before releasing mobility datasets [18, 31]. For our app usage data along with spatial and temporal information, we can also adopt this method to reduce the uniqueness of individual app fingerprints. Our experiments in Section 3 have proved that lowering spatial or temporal resolutions can decrease the percentage of unique app users effectively, which makes it easier to achieve *k-anonymity*. In addition, as for app domain, we can also degrade the detailed app name into a rather vague category that the app belongs to, *e.g.*, from "Wechat" app to "Social Network" category. This coarser app usage data will lower the risk of users being re-identified from the anonymous datasets. The published datasets can still be useful for many applications such as app recommendation systems. In addition, coarse-grained app fingerprints are also less sensitive because they do not give away the information about the exact apps or when/where the apps are used.

In the future, we plan to design advanced privacy protection schemes to better protect app usage datasets. We will explore the trade-offs between the privacy protection (*i.e.*, *K-anonymity* or differential privacy) and the data utility, for different application domains.

5.4 Limitations of This Work

Our dataset provides a rare opportunity to study the uniqueness of app usage across a large user population. However, there are still limitations. First, our dataset does not cover all the app usage behaviors of an individual user. Second, certain apps may use HTTPS protocols for every network request which will not be recorded by our dataset. Third, the dataset does not cover apps that make absolutely no network request, or apps that make networks requests solely through the WiFi network. Finally, our dataset comes from a single network service provider in China. In general, most of the data-plan services from different providers are pretty similar and there should be some similarities regarding how people use mobile apps. We leave further explorations across different network providers to future work.

6 CONCLUSION

In this paper, we focus on understanding the uniqueness of individual app usage behaviors across a large user population. Based on a city-scale mobile app usage dataset, we perform an empirical measurement to quantify the *uniqueness* of app usage patterns. Extensive results demonstrate that the fingerprints of mobile app usage are highly unique, and user demographics, and users' online and offline behaviors all influence the uniqueness level. Our study is a first step towards understanding the unique app usage behaviors of users, which paves the way for developing the next generation of privacy protection schemes and supporting personalized online services for mobile users.

ACKNOWLEDGMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2017YFE0112300, the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

REFERENCES

[1] O. Abul, F. Bonchi, and M. Nanni. 2010. Anonymization of moving objects databases by clustering and perturbation. *Information Systems* (2010).

[2] G. Acs and C. Castelluccia. 2014. A case study: privacy preserving release of spatio-temporal density in paris. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*.

[3] APKsHub. 2017. *com.pp.assistant*. https://www.apkshub.com/app/com.pp.assistant.

[4] Bhuvan Bamba, Ling Liu, Peter Pesti, and Ting Wang. 2008. Supporting anonymous location queries in mobile environments with privacygrid. In *Proceedings of the 17th international conference on World Wide Web (WWW)*. 237–246.

[5] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. 2014. ProactiveTasks: the short of mobile device use sessions. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 243–252.

[6] Konrad Blaszkiewicz, Konrad Blaszkiewicz, Konrad Blaszkiewicz, and Alexander Markowetz. 2016. Differentiating smartphone users by app usage. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 519–523.

[7] Hristo Bojinov, Michalevsky Yan, Gabi Nakibly, and Boneh Dan. 2014. Mobile Device Identification via Sensor Fingerprinting. *Computer Science* (2014).

[8] Karen Church, Denzil Ferreira, Nikola Banovic, Kent Lyons, Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2015. Understanding the Challenges of Mobile Phone Usage Data. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*.

[9] Trinh Minh Tri Do, Jan Blom, and Daniel Gatica-Perez. 2011. Smartphone usage in the wild: a large-scale analysis of applications and context. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)*. 353–360.

[10] J. Domingo-Ferrer and R. Trujillo-Rasua. 2012. Microaggregation- and permutation-based anonymization of movement data. *Information Sciences* (2012).

[11] Cynthia Dwork. 2008. Differential privacy: a survey of results. In *International Conference on Theory and Applications of MODELS of Computation*. 1–19.

[12] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. 2015. Robust traceability from trace amounts. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 650–669.

[13] Peter Eckersley. 2010. How Unique Is Your Web Browser? *Lecture Notes in Computer Science* 6205 (2010), 1–18.

[14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, 1054–1067.

[15] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in smartphone usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications and Services (MobiSys)*. 179–194.

[16] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K. Dey. 2014. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 91–100.

[17] S. Garfinkel. 2006. Privacy Protection and RFID. In *Ubiquitous and Pervasive Commerce*. Springer.

[18] M. Gramaglia and M. Fiore. 2015. Hiding Mobile Traffic Fingerprints with GLOVE. *ACM CoNEXT* (2015).

[19] Marco Gruteser and Dirk Grunwald. 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MobiSys)*. 31–42.

[20] Lisa Gutermuth. 2018. *How to Understand What Info Mobile Apps Are Collecting About You*. http://www.slate.com/articles/technology/future_tense/2017/02/how_to_understand_what_info_mobile_apps_collect_about_you.html.

[21] Hackernoon. 2017. How Much Time Do People Spend on Their Mobile Phones in 2017? (2017). https://hackernoon.com/how-much-time-do-people-spend-on-their-mobile-phones-in-2017-e5f90a0b10a6.

[22] Ke Huang, Chunhui Zhang, Xiaoxiao Ma, and Guanling Chen. 2012. Predicting mobile application usage using contextual information. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 1059–1065.

[23] Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. 2015. Revisitation analysis of smartphone app use. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 1197–1208.

[24] Philip Leroux, Klaas Roobroeck, Bart Dhoedt, Piet Demeester, and Filip De Turck. 2013. Mobile application usage prediction through context-based learning. *Journal of Ambient Intelligence and Smart Environments* 5, 2 (2013), 213–235.

[25] Huoran Li, Xuan Lu, Xuanzhe Liu, Tao Xie, Kaigui Bian, Felix Xiaozhu Lin, Feng Feng, and Feng Feng. 2015. Characterizing Smartphone Usage Patterns from Millions of Android Users. In *Proceedings of the Conference on Internet Measurement Conference (IMC)*. 459–472.

[26] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings of the of International Conference on Data Engineering(ICDE)*. 106–115.

[27] Liao, ZhungXun, YiChin, Peng, WenChih, Lei, and PoRuey. 2013. On mining mobile apps usage behavior for predicting apps usage in smartphones. (2013), 609–618.

[28] A. Machanavajjhala, D. Kifer, J. Gehrke, et al. 2007. l-diversity: Privacy beyond k-anonymity. *ACM TKDD* (2007).

[29] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3.

[30] Eric Malmi and Ingmar Weber. 2016. You Are What Apps You Use: Demographic Prediction Based on User's Apps. (2016).

[31] A. Monreale, G. L. Andrienko, N. V. Andrienko, et al. 2010. Movement Data Anonymity through Generalization. *Transactions on Data Privacy* (2010).

[32] Yves Alexandre De Montjoye, CÃĺsar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3, 6 (2013), 1376.

[33] Yves Alexandre De Montjoye, Jordi Quoidbach, Florent Robic, and Alex Pentland. 2013. Predicting Personality Using Novel Mobile Phone-Based Metrics. (2013), 48–55.

[34] Yves Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex Sandy Pentland Pentland. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. (2015), 536–539.

[35] Moon Hee Park, Jin Hyuk Hong, and Sung Bae Cho. 2007. Location-based recommendation system using Bayesian user's preference model in mobile devices. In *Proceedings of the 4th International Conference on Ubiquitous Intelligence and Computing (UIC)*. 1130–1139.

[36] Google Play. 2017. *App Usage - Manage/Track Usage*. https://play.google.com/store/apps/details?id=com.a0soft.gphone.uninstaller&hl=en.

[37] Shao Meng Qin, Verkasalo Hannu, Mohtaschemi Mikael, Hartonen Tuomo, and Alava Mikko. 2012. Patterns, Entropy, and Predictability of Human Mobility and Life. *Plos One* 7, 12 (2012), e51353.

[38] Kiran K. Rachuri, Kiran K. Rachuri, Kiran K. Rachuri, Kiran K. Rachuri, Emmanuel Munguia Tapia, and Emmanuel Munguia Tapia. 2014. MobileMiner: mining your frequent patterns on your phone. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 389–400.

[39] A. Ravve. 1953. *Principles of Polymer Chemistry*. Cornell University Press. 2854–2854 pages.

[40] David Rebollomonedero, Jordi Forne, and Josep Domingoferrer. 2010. From t-Closeness-Like Privacy to Postrandomization via Information Theory. *IEEE Transactions on Knowledge & Data Engineering* 22, 11 (2010), 1623–1636.

[41] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Proceedings of the 5th International Conference on Web Search and Data Mining (WSDM)*. 723–732.

[42] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. [n. d.]. Predicting user traits from a snapshot of apps installed on a smartphone. *Mobile Computing and Communications Review (SIGMOBILE)* 18, 2 ([n. d.]).

[43] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. 2015. Your Installed Apps Reveal Your Gender and More! *Mobile Computing and Communications Review (SIGMOBILE)* 18, 3 (2015), 55–61.

[44] Choonsung Shin, Jin Hyuk Hong, and Anind K. Dey. 2012. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 173–182.

[45] Y. Song, D. Dahlmeier, and S. Bressan. 2014. Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data.. In *PIR@ SIGIR*.

[46] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 17th international conference on World Wide Web (WWW)*. 1261–1269.

[47] K. Sui, Y. Zhao, D. Liu, et al. 2016. Your Trajectory Privacy Can Be Breached Even If You Walk in Groups. *IEEE/ACM IWQoS* (2016).

[48] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.

[49] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 1100–1108.

[50] Gary M. Weiss and Jeffrey W. Lockhart. 2011. Identifying user traits by mining smart phone accelerometer data. In *International Workshop on Knowledge Discovery from Sensor Data*. 61–69.

[51] Wikipedia. 2017. *Hamming distance*. https://en.wikipedia.org/wiki/Hamming_distance.

[52] Wikipedia. 2017. *Jaccard index*. https://en.wikipedia.org/wiki/Jaccard_index.

[53] Qiang Xu, Jeffrey Erman, Alexandre Gerber, Zhuoqing Mao, Jeffrey Pang, and Shobha Venkataraman. 2011. Identifying diverse usage behaviors of smartphone apps. In *Proceedings of the Conference on Internet Measurement Conference (IMC)*. 329–344.

[54] Yin Yang, Zhenjie Zhang, Gerome Miklau, Marianne Winslett, and Xiaokui Xiao. 2012. Differential privacy in data publication and analysis. *Chinese Journal of Computers* 14, 1 (2012), 601–606.

[55] Hongyi Yao, Gyan Ranjan, Alok Tongaonkar, Yong Liao, and Zhuoqing Morley Mao. 2015. SAMPLES: Self Adaptive Mining of Persistent LExical Snippets for Classifying Mobile Application Traffic. In *Proceedings of the 21st International Conference on Mobile Computing and Networking (MobiCom)*. 439–451.

[56] Ting Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martin Abadi. 2012. Host Fingerprinting and Tracking on the Web:Privacy and Security Implications. 11, 1 (2012), 111 – 124.

[57] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work:a large-scale measurement study. In *Proceedings of the 17th International Conference on Mobile Computing and Networking (MobiCom)*. 145–156.

[58] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K. Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 498–509.

[59] Tianqing Zhu, Gang Li, Wanlei Zhou, and Philip S. Yu. 2017. Differentially Private Data Publishing and Analysis: A Survey. *IEEE Transactions on Knowledge & Data Engineering* 29, 8 (2017), 1619–1638.

Received November 2017; revised February 2018; accepted September 2018