COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

A review on smartphone usage data for user identification and user profiling

Syafira Fitri Auliya*, Lukito Edi Nugroho, Noor Akhmad Setiawan

Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Article history:
Received: 27 April 2021 / Received in revised form: 24 May 2021 / Accepted: 29 May 2021

Abstract

The amount of retrievable smartphone data is escalating; while some apps on the smartphone are evidently exploiting and leaking users' data. These phenomena potentially violate privacy and personal data protection laws as various studies have showed that technologies such as artificial intelligence could transform smartphone data into personal data by generating user identification and user profiling. User identification identifies specific users among the data based upon the users' characteristics and users profiling generates users' traits (e.g. age and personality) by exploring how data is correlated with personal information. Nevertheless, the comprehensive review papers discussing both of the topics are limited. This paper thus aims to provide a comprehensive review of user identification and user profiling using smartphone data. Compared to the existing review papers, this paper has a broader lens by reviewing the general applications of smartphone data before focusing on smartphone usage data. This paper also discusses some possible data sources that can be used in this research topic.

Keywords: Smartphone Usage Data; Privacy; User Identification; User Profiling

1. Introduction

The increasing Internet penetration and the escalating number of smartphone embedding sensors enhance the amount of retrievable near-real-time human behaviour dataset [1]. This abundant data about people has become more economically essential as many Internet platforms such as Google and Facebook get their monetary value from capturing their visitors' behaviour [2].

The European Unions' General Data Protection Regulation (GDPR) as the most prominent notable personal data protection law has attempted to put the boundaries to regulate the transfer and usage of people's personal data. However, the development of technologies makes big loopholes in the attempts to protect privacy and personal data. For instance, emerging technologies such as artificial intelligence and big data analytic are currently able to reveal sensitive information from the combinations of 'non-personal data' (not restricted in most protection laws) and other public data [2]. In addition, most people are not aware of the importance of their data [3,4]. Meanwhile, data leaks potentially come from various sources such as web access, cloud storage, or smartphone data.

Smartphone data is one of the most prominent resources of data leakage. Many studies attempted to identify potential smartphone data leakage. For instance, the static analysis of smartphone data proposed in [5,6] revealed that few apps on Android platform leaked phone ID, sent geographic location, and used ad/analytic library, phone number, and even the SIM card serial number to other parties. Meanwhile, various studies showed that these smartphone data could violate privacy by generating user identification and user profiling.

User identification identifies or differentiates specific users among the dataset using the users' characteristics, e.g. by using usage data [7-12], sensor data [13-16], and user input data [17-21]. User profiling, meanwhile generates users' profiles or traits (such as age, gender, income, and personality) by exploring how data are correlated with user personal information to extract key features and describe their characteristics, for instance in [22-28]. As any information capable of identifying or generating information about a person is categorized as 'personal data' [29], smartphone data that can identify the user in user identification and generate user information in user profiling should be started to legally entitled to the protection from exploitation and leakage.

In fact, the current review papers on user identification and user profiling using smartphone data are limited. To the best of the authors' knowledge, the relevant review paper on this topic is only studied in [30], focused on user profiling using smartphone applications data. However, applications data are only a small part of smartphone data which has a broader scope, such as usage data, sensor data, and user input data [31]. In addition, the existing review paper focuses only on user profiling without discussing user identification. Therefore, this paper aims to provide a general overview of smartphone data, especially smartphone usage data that can potentially be used for user identification and user profiling.

The rest of this paper is organized as follows section 2 is presented to explore the discussion about privacy issues and some instances of privacy violation cases. In section 3, the smartphone data taxonomy are explained with a list of the related studies and the state-of-the-art in user identification and user profiling using each category of smartphone data.

* Corresponding author.
Email: firauliya@gmail.com
https://doi.org/10.21924/cst.6.1.2021.363

Furthermore, the more detailed features and methods utilized in smartphone user identification and user profiling are explained in section 4. Section 5 is to summarise some possible data sources used in this research topic before providing the conclusion in section 6.

2. The Discussion about Privacy

2.1. A brief history

The discussion related to privacy has been so long developed and one of the significant milestones is Samuel Warren and Louis Brandeis's study 'the Right to Privacy' in the Harvard Law Review in 1890. During the technology development at that moment (especially photographs to support gossip column in newspapers), they believed that everyone has 'the right to be let alone'. Thus, their idea were widely accepted by both the law and the public and in turn made the right to privacy as the fundamental concept in society [32]. The importance of privacy subsequently was written in the Universal Declaration of Human Right as an essential part representing a civilization [33]. Decades later, Gellert and Gutwirth mentioned that the likelihood of privacy being violated is increased in the digital era. They thus highlighted the need for protecting personal data (they described personal data as any information able to identify a person) to ensure privacy [29].

The necessity to protect personal data is also widely realized around the world. Since 1995, the European Union (EU) adopted the European Data Protection Directive [34]. The union also made a breakthrough contribution to the data protection effort by ratified the General Data Protection Regulation (GDPR) in 2016. Not merely regulating the provisions and requirements to process the personal data of individuals located in the EU and the European Economic Area (EEA), GDPR also addressed the transfer of their citizens' personal data outside their jurisdiction area. The GDPR is then inspired many countries to also formally establish their data protection regulations.

2.2. Privacy violation cases

Regardless of the continuous attempts to protect privacy and personal data, the development of technology makes a big loophole in these attempts. For instance, emerging technologies such as artificial intelligence and big data analytic are currently able to reveal sensitive information from combinations of non-personal data (not restricted in any protection laws) and other public data [2].

The most notable instance of privacy violation cases using emerging technologies is the Cambridge Analytica involving Facebook. In 2015, Facebook was revealed to give an unauthorized access to personally identifiable information (PII) of more than 87 million users to Cambridge Analytica [35]. Combining OSEAN (openness, conscientiousness, extraversion, agreeableness, and neuroticism) psychological test in which users volunteered taken along with secret access to their Facebook friends' data through the Facebook Open API, Cambridge Analytica got an access to millions of Facebook data. Without the data owners' further concern, the data paired with other private and public data revealed many valuable individuals behaviour patterns. The behaviour patterns were then utilized and believed to influence the outcome of the US 2015 elections and Brexit 2016 votes [36].

With the increasing smartphone penetration and the development of various embedding sensors, the amount of data people have generated from a daily activity is tremendous. As if to acknowledge its significance in the modern world, data are currently considered 'the new gold'. However, as mentioned, most people are not aware of their data importance. In contrasts, evidence proved that daily smartphone activity is continuously threatening privacy. A study on over 21 million lines of source code from more than a thousand free Android apps using automated tests and manual inspections unveiled several trends in [5]. Of them, 33 apps leaked Phone IDs. Besides, 13 apps sent the geographic location to the network and/or advertisers, and 51% was included in an ad/analytic library. The infamous TaintDroid project also revealed that out of the total 30 monitored apps, 15 apps reported users' location, seven apps collected the device ID, and some cases disclosed the phone number and the SIM card serial number [6]. Thus we believe that what people do on their apps will unconsciously reveal too much undesired information about them.

3. Smartphone Data Taxonomy

A smartphone consists of multiple components that can be classified into four distinct categories: device (hardware), connectivity (technology to provide connectivity), applications, and data (information stored and used on the smartphone) [37]. Regarding the smartphone data, according to the study proposed in [31], smartphone data taxonomy is categorized into seven categories according to the data source, as seen in figure 1 and elaborated in table 1.

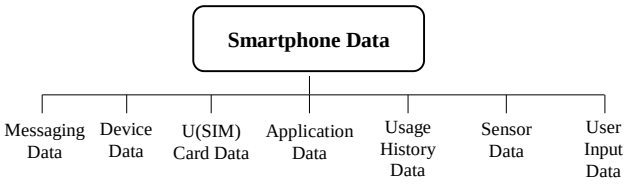

Fig. 1. Smartphone data taxonomy

Messaging Data is obtained from the mobile phone carriers' messaging services (SMS, EMS, and MMS) and electronic message (chat and email). It consists of messaging logs that contain the information about its receiver, sender, time or date of delivery, and attachment included. Study about identification and generating of smartphone users' profile using this messaging data (especially messaging log) is uncommon but technically possible. In the later paragraphs, we introduce user identification and profiling using call log details. Both messaging and call logs are available in Call Detail Record (CDR), a data record generated when a phone is connected to a network. CDR consists of the encrypted phone number, base transceiver station (BTS)s' identity, date and time of the call, call duration, and SMS metadata. Although user identification or user profiling using CDR mostly uses the call log details because the call log contains more details, a similar approach should also possible to use SMS metadata.

Device Data is the data about the device and operating system that are not related to any third party. It consists of IMEI, Wi-Fi MAC address, and device serial number that are

Table 1. Smartphone data taxonomy according to the data sources

Category	Definition	Examples
Messaging Data	Data obtained from messaging service (incl. SMS) and chat/email	Messaging logs
Device Data	Data of the device and the operating system not related to the third party	IMEI, MAC Address, serial number
(U)SIM Card Data	Specific information to uniquely identify the user by the telecommunication carrier	Mobile subscriber identification number, the integrated circuit card identifier
Application Data	Data accessible by apps and necessary for their execution	Configuration files, temporal data, logs data
Usage History Data	Log data related to the usage of the smartphone	Call logs, browsing history log, network connection history logs, event logs
Sensor Data	Data related to sensors	Data about location, temperature, direction
User Input Data	Data produced from the interaction of the user with its smartphone	User gesture, button presses, keystrokes

critical identifiers in which the information about them already reveals the phone's identity. Because this data is valuable to describe one's identity effortlessly, it must be protected from any leak. Therefore, information such as IMEI is usually only accessible after receiving explicit permissions from the user during app installation. The study that attempts to identify and generate user profile using device data is unheard.

(U)SIM Card Data contains specific user information to be uniquely identified by the telecommunication carriers. Some examples are international mobile subscriber identity, integrated circuit card identity, and mobile subscriber identification number. Similar to Device Data, the study that attempts to identify and generate user profile using (U) SIM Card Data is unheard because of its data confidentiality.

Application Data is the data accessible by applications. The apps need to access the data for their execution, such as configuration file, logs, and temporal data. User identification or user profiling using these files is also unheard. We believe that two of the reasons are that the inhomogeneous data across devices and thick application-level encryption (especially in apps like Whatsapp and Telegram) prevent the attacker from putting effort to get this data and choose other data sources instead.

Usage History Data is the log data related to the phone utilization. Some examples are call logs, browsing history logs, network connection history logs, and the operating system's event logs. With further details explained in section 4, studies used usage history data for identifying user or device such as by using: web browsing behavior [7], call-log [8,9], application behavior [10], and the set of apps installed [11,12]. Meanwhile, user profiling using this data is such as by using call-log [22,25] and the set of apps installed [26,27].

Sensor data is the data generated by sensors on a smartphone, such as a camera, GPS, compass, accelerometer, and microphone. An accelerometer was used in [13] for user identification and user authentication by observing people daily activities, like walking, jogging, and climbing stairs and in [28] for user profiling to recognize daily life activity. SenGuard project proposed in [14] used voice, location, multi-touch, and locomotion to enable user identification service on smartphone continuously. GPS mobility data was used in [15] to observe the distance function between a trajectory and sampled points and in [16] to observe the similarity of users' trajectory from various data sources for user identification.

User Input Data is created from an interaction between users and their smartphones, such as keystrokes and user gestures. Keystrokes were analyzed in [17] (key hold time, error rate, and diagraphs) and in [18] (incl. duration, time since the last key, frequent key, and infrequent key) to identify smartphone user. Touchscreen gestures used in [19,21] also obtained high accuracy to continuously identify the users.

4. Smartphone Usage Data for User Identification and User Profiling

Smartphone usage history data (for simplicity, called

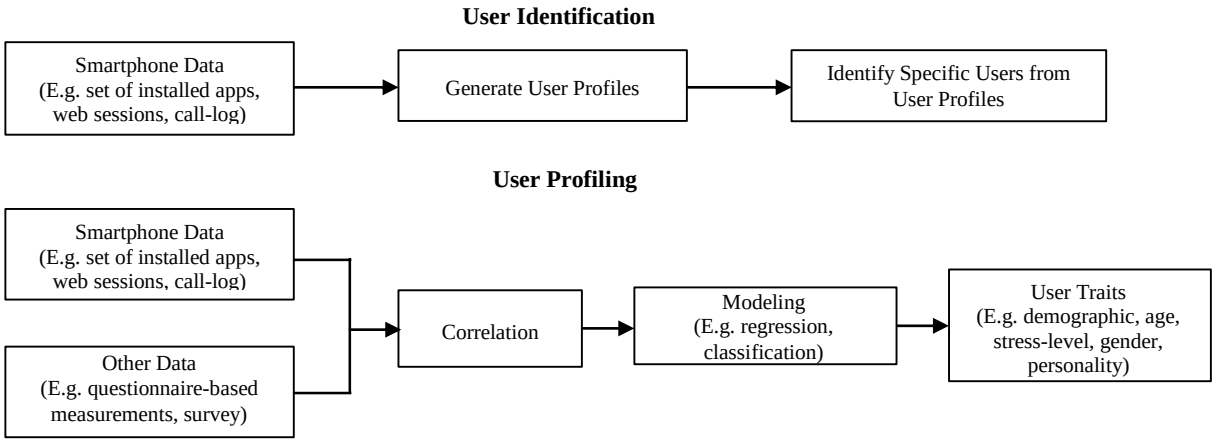

Fig. 2. User identification and user profiling general frameworks

"smartphone usage data" from now on) is the log data related to the phone utilization, e.g. call logs, browsing history logs, and network connection history logs. This section explores user identification and user profiling using especially smartphone usage data, including their features and methods utilized.

User identification is identifying or differentiating a specific user among the data based on the user' characteristics. Meanwhile, user profiling using smartphone data generates a user's profiles or traits (e.g. age, gender, income, and personality) by exploring how data is correlated with user personal information to extract key features and describe users' characteristics. The general idea of user identification and user profiling is available in figure 2 and described further in section 4.2 and section 4.4. In brief, in user identification, the "User Profile" of all users in the dataset is generated. This User Profile acts as a fingerprint that differentiates the user from others in identification processes. Meanwhile, in user profiling, a model is constructed by comparing the data from the respondents' smartphone with their answers in the questionnaire-based personality measurement. By exploring the model, the correlations are made to predict the traits that can be inferred from these respondents.

4.1. User identification

Table 2 presents the example of studies about user identification using smartphone data. In the study proposed in [7], the web user behavioural profiling, user profiles were generated based on the session of the users accessing specific sites. The profiles were then used to identify a user that anonymously accessed the website in the future. This study set 300 sessions as the minimum cut-off to choose a user for the training dataset. To avoid accuracy bias due to an imbalance in input length, only the first 300 sessions for each user were included in the training. This research had 2,798 qualified users and accurate identification rate of 87.36%.

A large dataset was used in [8] to identify users by matching the statistics of users' behavioural patterns. They used call-data records (CDR), web browsing history, and GPS dataset in the report with the latter dataset outside the scope of this section. The study used the CDR dataset of almost 50.000 random customers in Ivory Coast for two weeks. Using the location of the antenna where the user was connected when making the call, their method gave an accuracy of 21.1% (one-fifth correct identification). The study also observed 121 active website users for two weeks to analyse the users' web behaviour. From a total of 83,219 different websites visited by these users, they could correctly identify 50% of users by considering the top most popular websites. The accuracy dropped if fewer websites are considering.

Fifteen months of smartphone mobility data for 1.5 million individual were analysed in [9] to prove that human mobility can trace people mobility with high accuracy. The study used the location of the antenna (the maximal half-distance between antennas) recorded every time a user received or initiated a phone call or text messenger. This project concluded that four types of information about individuals' spatiotemporal points were sufficient to identify 95% of the individuals.

Application behaviour was analysed in [10]. The study recorded the network traffic from 20 users devices using tcpdump for eight hours. The study monitored the usage of the top 14 free Android applications and used 'burstiness' to distinguish each app. Burst is the idle periods before short peaks of incoming and outgoing data transfers. A single burst consists of a sequence of packets mainly from the (e.g.) TCP connection. Using the number of bursts to identify users based on their app usage behaviour, the study obtained 90% identification accuracy. Moreover, they also concluded that the eavesdropper required only 15 minutes to capture traffic and get the same accuracy using their method.

Identifying users based on their set of Android applications installed was studied in [11] and [12]. Collaborating with a major ISP In China, the study proposed in [11] obtained the data of 1.37 million. The dataset contained an anonymised user ID, connection timestamp and duration, the cellular tower ID, the location of a cellular tower, and the header information of the HTTP and HTTPS requests. The SAMPLE tool was then used to identify the app match with each HTTP request. The ISP collaborator also provides information about the user Weibo (the most popular social network app in China) includes gender, city, and users' activity in Weibo. The study concluded that 88% of users could be identified by four random apps and even higher when considering when and where the apps were used.

Table 2. User identification using smartphone usage data

Features	Identification Method	Reference
Consecutive web sessions with known user IDs	Euclidean Distance and *k*-Nearest Neighbours (*k*NN)	[7]
Call-log: antenna location *Web behaviour*: number of visitors to certain sites	Min-weight max matching with the weight metric	[8]
Antenna location	Deductive Disclosure	[9]
Side-channel features (e.g. packet size, byte ratio) from network traffic of popular apps	*k*NN and Support Vector Machine (SVM)	[10]
Set of Installed Apps	Hamming Distance and Jaccard Distance	[11]
Set of Installed Apps	Logistic Regression	[12]

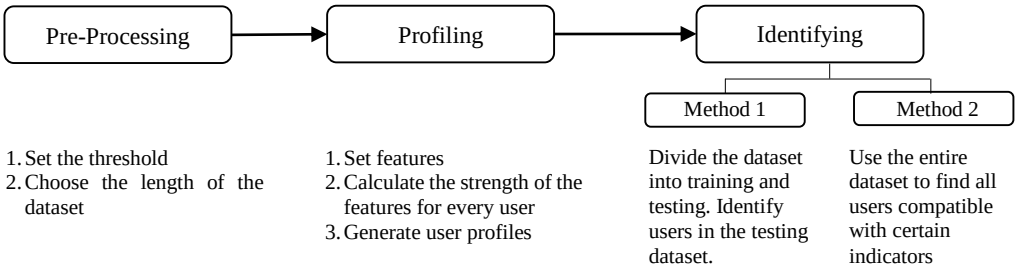

Fig. 3. User identification steps using smartphone usage data

Meanwhile, the study proposed in [12] used the dataset from the Menthal project. There were 46.736 users giving information about their phone usage (daily time spent, number of unlocking, frequency and duration of usage per app), SMS, phone call, GPS location, and answered a questionnaire about their basic demographic information. Using only the set of top 60 most frequently used apps, this study could identify 99.4% of users in the dataset. Of them, 95.27% even had the closest different user with Hamming distance >10, meaning that they must change the behaviour on at least ten different apps to achieve anonymity.

4.2. User identification methods

As seen in figure 3, from the user identification studies discussed in section 4.1, three main steps are required to identify users from their smartphone usage data.

Pre-Processing. The criteria or thresholds to distinguish the relevant dataset from the others should be set. For example, the study proposed in [7] set the minimum cut-off by 300 sessions, meaning that only the first 300 sessions of each user were included in the training process. Users with less than 300 sessions of recorded data were excluded from the following process. Another example in [8] restricted the observation only to active users in the two observation weeks. In the web behaviour analysis, they also deleted all URLs that did not have a favicon from their dataset. A favicon is a small icon associated with a certain website within the same domain. As an example, "news.yahoo.com" and "mail.yahoo.com" have different encrypted names but the same favicon identifier (e.g. "1") in the database. In this pre-processing stage, the length of the dataset is also determined. While the larger datasets generate the more extensive training and testing dataset, it also increases the computational resources consumed.

Profiling. The features used in user identification are determined and calculated. It begins by defining essential features and selecting the top few features from them. It is then followed by calculating the strength of each feature in the selected features. For instance, the study in [7] used (1), the relative pattern strength of pattern p_j and user u_i after treating the outliers by dividing the 0% to the 90% quartile into bins and move all outlier to the last quartile. The example in [10] used (2), as the relative mutual information between a feature F_i and the user U to calculate their features strength.

$$rps(p_j|u_i) = \frac{|D_{u_i}^{p_j}|/|D_{u_i}|}{|D^{p_j}|/|D|} \quad (1)$$

$$rMI(F_i;U) = 1 - \frac{H(U|F_i)}{H(U)} \quad (2)$$

$rps(p_j|u_i)$ is the relative pattern strength of pattern p_j and user u_i. $|D_{u_i}^{p_j}|$ is the number of sessions from user u_i that have the behavioural pattern p_j. $|D_{u_i}|$ is the total number of sessions from user u_i. $|D^{p_j}|$ is the total number of sessions from all users that contain behavioural pattern p_j. $|D|$ is the total length of the dataset. $rMI(F_i;U)$ is the relative mutual information between a feature F_i and the user U. The entropy $H(U)$ quantifies the uncertainty about the user U. The conditional entropy $H(U|F_i)$ quantifies the remaining uncertainty if the value of the feature F_i is known. The difference between $H(U)$ and $H(U|F_i)$ is maximal if the features fully determine the user U.

Subsequently, the User Profiles using the value of (1) or (2) for every user u_i (i = 1… N, N is the total users in the dataset) are generated. The User Profiles can be formulated in (3) and (4).

$$r_1 = \{a_1, a_2, \dots, a_K\} \quad (3)$$

$$a_i = rfs(f_j|u_i) \quad (4)$$

r_1 is a user profile consisting of a_1 to a_K. K is the total number of users in the dataset and a_i is the relative feature strength of feature f_j and user u_i.

Identify. The identification stage can be classified into two general approaches. The first approach is by separating the dataset in half for training and testing (validation). The features for each user are then calculated for every user in the dataset. The studies using this approach were proposed in [7,8,10]. The second approach is by using the entire dataset to find all compatible users with a certain indicator, for instance, using a brute force method like in [9] or statistical method such as in [11] and [12].

For the first approach, after separating the dataset and calculating the users' features, the study in [7] used Euclidean Distance (5) to create the list of distances between every user in the dataset. It compared the distance of the users in the training and validation dataset. Meanwhile, the study in [8] divided the dataset into an unlabelled histogram (testing) and labelled histogram (training). Each row in the histogram contained the user identity and the value of several features. A min-weight max matching with the weight metric was then used to match each row in the labelled histogram and unlabelled histogram. Another example in [10] used *k*NN and SVM to solve the multiclass classification problem with n users in

Table 3. Studies about user identification traits using smartphone usage data

Dataset	Classification Method	Traits Generated
Application (e.g. number of uses of office apps), Bluetooth, SMS, Calls	SVM	Big Five Personality Traits
Basic phone use (e.g., number of calls), active user behaviours, location, regularity, diversity	SVM	Big Five Personality Traits
Telecommunication logs (e.g. call duration), online social networks (e.g. number of Facebook friends), physical proximity.	SVM	Big Five Personality Traits
Call info, images, SMS info, contact, SDK version	Various Algorithms (incl. Random Forest and SVM)	Credit Score
Set of Installed Apps	SVM	Gender, age, race, relationship status, children, income
Set of Installed Apps	SVM	Religion, relationship status, spoken language, country, children

which each problem consisted of 50% features from the training dataset and 50% features from the testing dataset.

$$d_{r_1,r_2} = \sqrt{\sum_{j=1}^{K}(a_j - b_j)} \quad (5)$$

d_{r_1,r_2} is the vector distance between two profiles r_1 and r_2 as mentioned in (3). K is the total number of users in the dataset, a_j and b_j is two users where their distances are being calculated.

In the final stage, the testing (validation) dataset users were matched with the value of Euclidean Distance in the training dataset. In other words, to identify user u_{vi}, they compared $d_{r_{vi},r_{tj}}$ and $d_{r_{ti},r_{tj}}$. $d_{r_{vi},r_{tj}}$ is the distance between the vector user u_{vi} (the user we wanted to identify) and user u_{tj} (all users in the training dataset, with j=1…N). Meanwhile, $d_{r_{ti},r_{tj}}$ is the distance between user u_{ti} and u_{ti} (where i, j =1…N in training dataset). The users were ranked according to the distance between their profile and user u_{vi}, with the rank was dynamic according to the compared profiles (similar to k-nearest neighbour (*k*NN) method with k=1). The user with the highest rank (smallest Euclidean Distance) was then identified as the proposed user u_{vi}.

Meanwhile, for the second approach, the study in [9] used the entire dataset to generate, with a brute force method, $S(I_p)$ as the set of users whose mobility traces were compatible with I_p. I_p is the information available to attackers such as "7 am – 8 am at Spatio-temporal point A" and "10 am – 11 am at Spatio-temporal point B". The conclusion was made (meaning that a specific individual was found out/identified) when $S(I_p) = 1$. Similarly, the study in [12] used the entire dataset to find the Hamming distance between app signatures. Hamming distance (7) depicts a similarity between two users; thus, a user is anonymous if the Hamming distance = 0 while a user can be identified if the Hamming distance = 1. Alongside Hamming distance, the study in [11] also used Jaccard Distance (8) to measure dissimilarity between two sets and plots the Cumulative Distribution Function.

Hamming Distance and Jaccard Distance between user u_i and u_j can be respectively defined as follows:

$$HD_{ij} = |A_i \cup A_j| - |A_i \cap A_j| \quad (7)$$

$$JD_{ij} = \frac{|A_i \cup A_j| - |A_i \cap A_j|}{|A_i \cup A_j|} \quad (8)$$

HD_{ij} is the Hamming Distance between user u_i and u_j. JD_{ij} is the Jaccard Distance between user u_i and u_j. A_i and A_j are respectively the set of apps used by user u_i and u_j

4.3. User profiling

Smartphone usage data also can be used for user profiling (often referred to as "user identification traits" or "user fingerprinting"). Some studies, such as proposed in [38-41] revealed the communal fingerprint while others, such as the examples in [22-27] revealed the individual fingerprint from the dataset. The communal fingerprint is the general trend of the people in the dataset, such as daily activity pattern within people in the same work area, land use pattern in some different areas, or people movement pattern. Meanwhile, individual fingerprint reveals information about specific individuals in the dataset. The communal fingerprint is outside this paper's scope. Thus this paper focuses on the individual fingerprint. The summary of these related works is present in table 3.

Smartphone data to determine individuals' Big Five Personality Traits was used in [22,23,24]. In the pioneer study in [22], they obtained eight months of data from 83 participants given Nokia N95 phones. Most of them had not owned a mobile phone before. They thus were asked to fulfill an online TIPI questionnaire as the self-perceived personality. The correlations obtained from the study were between 59.8% - 75.9%. Meanwhile, the study in [23] used data from 69 participants equipped with an open sensing framework running in Android named *Funf*. It monitored several indicators: basic phone use such as number of calls, active user behaviors (e.g., number of calls initiated, time to answer a text), location (radius of gyration and number of places calls made), regularity (routine), and diversity (call entropy and number of interactions by number of contacts ratio). This study used SVM to classify and validate using 10-fold cross-validation. They thus predicted whether smartphone users were low, average, or high in each of the Big Five Personality. The results were that they could predict each of the personality better than random for 29%-56%. The study in [24] analyzed questionnaire-based data from 636 freshman student in TU Denmark. The study could predict the extraversion trait well (35.6% higher than by a null model) but gave lower results for other Big Five Personality traits. Similar to the other two first mentioned studies, this study also used SVM as the model to predict the classification label Y from the feature vectors X.

Using a dataset from an anonymous European consumer lending company offering a digital loan application submitted by a mobile application, the study in [25] examined the dataset of 2.503 customers having a loan. From the dataset, they separated the trustworthy customer (1.516 customers without any debts that exceeded the 90-days limit) and the untrustworthy customer (987 customers who delayed their payment). Features used in the study came from users' device, such as monthly average number of calls, the average number of images per month, number of contacts, and SDK version. Using various algorithms such as logistic regression, decision tree, random forest, SVM, and Neural Network (NN) as the classification methods, this study obtained AUC (Area under the Curve) of 0.51-0.59.

Traits identifications based on users' set of installed Android applications were studied in [26,27]. The study in [26] collected data from 218 volunteers and converted the differences between app installation patterns among male and female users into features. The features were then used to build a linear SVM classifier to predict users' gender with 70% accuracy. Meanwhile, the study in [27] used a similar methodology. They collected data from 231 volunteers that pre-installed an app giving information about the installed apps on users' phone. The volunteers answered a questionnaire about religion, relationship status, spoken language, origin and residence country, including whether they had a child aged under 10. The research compared the dataset obtained from these volunteers with two popular sites where users publicly shared their installed app lists ("Appbrain" and "Appaware") to ensure the dataset's representativeness. Using SVM classifier, the best results obtained was over 90% of precision and 75% of recall.

4.4. The methods for user profiling

All of the above studies compared smartphone data with questionnaire-based personality measurement. They thus made a correlation between them, for example, by making a classification model. The model was used to predict the traits that could be obtained from smartphone data.

As seen in table 3, all selected studies used Support Vector Machine (SVM) to obtain users' traits from their smartphone usage data. SVM is an algorithm that creates a line or a hyperplane that separates the data input classes. It is favourable and superior in classification problems because of its extraordinary generalization capability, optimal solution, and discriminative power [42].

The input in SVM is the selected smartphone features. To select the most suitable features relevant to the dataset, the above studies used various approaches. For example, the study in [23] selected the most relevant features using a greedy method. This method ranked all features based on their squared weight and removed the worst feature at each of the iterations. The removal stopped when the worst feature subset was less than 3 degrades the performance and reported of the three highest-ranked features. Another example in [24] chose the features with the strongest correlations with their targeted traits.

Some of the studies also combined several features into new ones, such as the entropy of contact (the ratio between "total number of contacts" and "the relative frequency he/she interacts with them"). For example, the study in [23] used (7) and another study in [24] used (8).

$$H(a-c) = -\sum_c f_c \log f_c \tag{7}$$

$$S_u = \sum_c \frac{f_c}{f_t} \log_2 \frac{f_c}{f_t} \tag{8}$$

$H(a-c)$ is the entropy of the contacts between user a and a contact c. f_c is the frequency at which the user a communicates with contact c. S_u is the entropy of user u. f_t is the total number of interaction.

5. Some Possible Methods to Obtain Smartphone Usage Data

Various methods can obtain the smartphone usage data, which we here have classified into three categories based on the data sources: from telecommunication carrier, using additional apps, and retrieving data from operation system.

5.1. From telecommunication carrier

For usage tracking and billing purposes, telecommunication operators generate Call Detail Record (CDR). CDR is triggered every time a subscriber uses service such as calling and messaging [43]. It contains basic information about mobile phone usage, likely the caller and recipient's cell towers, the identities of sources (point of origin), the identities of destination (endpoints), the duration of each call, the amount bill of each call, and the subscriber's billing information (total amount and time period). The accuracy of CDR to pinpoint people varies according to the expected traffic and terrain. This variety is caused by the difference in the cell towers distance as in rural areas cell towers are spaced 2-3 kilometres apart, while the distance is only 400-800 metres in densely populated areas [44].

For privacy proposes, the exemplary International Telecommunication Union (ITU)'s report in Liberia stated in [44] revealed few data anonymization process. Before the dataset was given to Data Analysis Partner (DAP), the telecommunication operators transferred the data to Local Collaborator Centres (LCOs). LCOs ran anonymization software and manually processed each data file to ensure that all privacy-related identifiable information were removed. The anonymized CDR thus was transferred to DAP to be used in more profound analysis.

CDR is not publicly available. Researchers who can analyse it always collaborate with telecommunication operators to obtain the dataset.

5.2. Using additional-apps

To obtain smartphone usage data, numerous tracking applications are available in markets. For example, in Android platform, *StayFree*, *App Usage*, *Time Tracker*, and *My Phone Time* are some of the popular apps on the Google Play Store proficient in gathering smartphone usage data. These apps can visualize daily and weekly usage data from the installed apps starting from the moment the user installed them. However, as for April 2021, only *App Usage* and *My Phone Time* allowed their data to be extracted into CSV file in their free account.

While *Time Tracker* is also capable of extracting CSV file, this feature is only accessible in their premium paid account.

The usage history information provided by these apps is various. The CSV file retrieved from *My Phone Time* and *App Usage* already contains information about app name, starting time, and usage duration every time a new app is opened on the user's phone. Both tracking applications are recording conscious and unconscious activities carried out on the user's phone. For example, the dataset presents activities opening *Whatsapp*, *Chrome*, and a camera that users deliberately open and *Permission controller*, *System UI*, and *Android System* that run in the background.

Overall, the information provided by *App Usage* is more complete than *My Phone Time*. Although the dataset from *My Phone Time* contains app package name that abstains in *App Usage*, *App Usage* can also extract usage trends (daily, weekly, monthly). The frequency users check their phone and location history are inaccessible in *My Phone Time*.

5.3. Retrieving from operation system

Because of the GDPR privacy regulation, since 2018, Apple and Google have allowed users to also download all the data about them kept by the Operation System (OS). Apple provides data associated with the user's Apple ID including sign-in record, calendars, photos, documents, and record of a retail transaction, as seen in table 4. Meanwhile, Google, with a service named Google Takeout, issues data linked with Google Account with more enormous details including detailed data about browser search history, location history, and user activity data/app usage history, as seen in table 5.

Table 4. Few data retrievable on apple device

Category	Description
Apple Media Services Information	Activities in App Store, iTunes Store, Apple Books, Apple Music, and Podcasts
Apple ID Account and Device Information	Includes AppleID, name, emails (masked), Phones number (masked), address, device name, device IMEI, device serial number, and device last heartbeat IP
Apple Online and Retail Stores Activity	Includes subscription status and device used
AppleCare	Includes serial number, purchase date, and shipped date
Game Centre activity	Activities about gaming sessions
Maps	
Marketing	Information that Apple has used to contact the user for marketing reasons
iCloud	It includes photos, contacts, calendars, notes, bookmarks, files, and email stored in iCloud

Table 5. Few data retrievable on android device (google takeout)

Category	Description
Android Device Configuration Service	Includes AndrioidID, MEID, IMEI, Serial number, MAC address, and device attributes
Chrome	Includes title, URL, and timestamp of each web activity using Google Chrome
Drive	Files stored in Google Drive
Google Photos	Files stored in Google Photos
Google Play	Data about app installed, ratings, order. Also includes movie, games, and books accessed.
Google Translator	Documents translated using Google Translator Toolkit
Location History	Users location data includes latitude, longitude, place ID, address, name, and a location confidence number.
Mail	Messages and attachment in Gmail account
My Activity	Including detailed activities on the phone (app used and timestamp) as well as searching history on Chrome, Gmail, YouTube, Google Map,
YouTube	List of YouTube activities

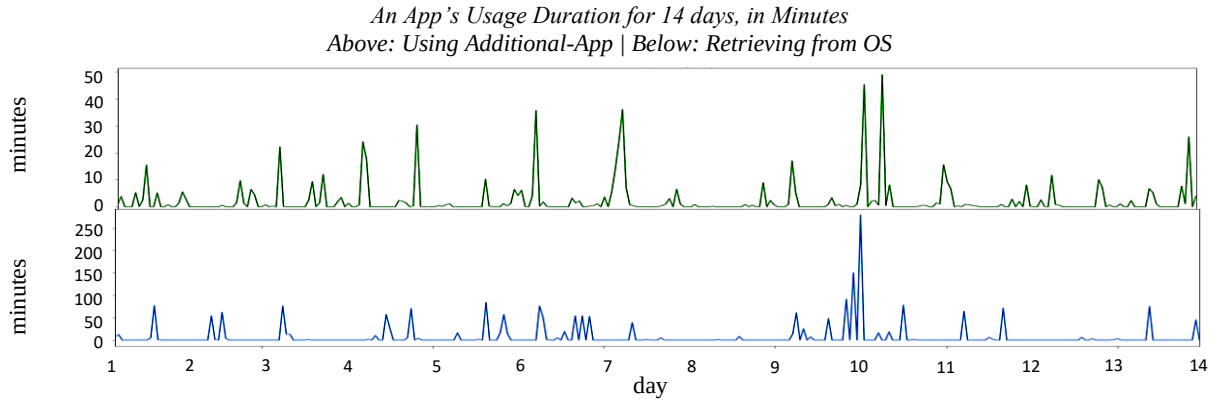

Fig. 4 The illustration of apps' usage duration using an additional app (above graph) and retrieving from the OS (below graph)

By retrieving data from the OS, it is possible to generate abundant data as the OS automatically records since the first time a user activates their phone – unlike the previous method requiring an additional installation action from users. However, we suggested to scrutinizing the accuracy of this data. As seen in figure 4, to illustrate, we compared the usage duration of a specific app using *App Usage* vs. retrieving it from the *Google Takeout*. Compared with the real-time activities (above graph, green line), the duration recorded by the OS (below graph, blue line) differs in, such as the duration and the recording interval. We assumed that, to not burden the phones' hardware resource usage, the users' data is only sent and stored to the OS storage with minimum interval, thus not entirely capturing full activities of the users.

6. Conclusion

The increasing penetration of the Internet and the advances of smartphone embedding sensors wider the variety of data generated by phones. However, these data are in some cases exploited or leaked to other parties without explicit consent from users. Although the data is anomyzed thus not containing obvious identifies data, emerging technologies such as artificial intelligence and big data analytic are potentially able to reveal users' personal data and violate privacy, for example, by conducting user identification and user profiling. As the related studies are limited, this study presents a comprehensive review of user identification and user profiling using various smartphone data (e.g. application data, usage data, sensor data, and user input data). Focusing on smartphone usage data, this study revealed that the current studies on this field generated a high accuracy and precision for user identification and user profiling. This study also discussed three possible methods to obtain smartphone usage data: collaborating with telecommunication operator, asking users to install additional apps, and retrieving from operation system (e.g. Apple and Google). The first method required tactical skills and resources, while data from the second method were limited only for the duration of users installing the apps. Although the third method promises to obtain abundant data, the accuracy of data stored by the OS should be scrutinized because, unlike the second method that record real-time activities with the users' consciousness to install the apps, the OS-stored data recording run only in the background. We assumed that the data were only sent and stored to the OS storage with minimum interval to not burden the users' phone hardware resource usage. Further studies about user identification and user profiling using the OS-stored data are required for real implementation in more elaborated settings.

References

1. R. Mafrur, I. G. D. Nugraha, D. Choi, *Modeling and discovering human behavior from smartphone sensing life-log data for identification purpose*, Hum.-centric Comput. Inf. Sci. 5 (2015) 31.
2. T. Tjerk and Z. Mann, *Data Protection in the era of Artificial Intelligence - trends , existing solutions and recommendations for privacy-preserving technologies*, Springer, Cham. (2021) 153-175.
3. S. Spiekermann and J. Korunovska, *Towards a value theory for personal data*, J. Inf. Technol. 32 (2017) 62-84.
4. CISSReC, *Hasil survey lembaga riset CISSREC 'Tingkat Kesadaran Masyarakat Tentang Keamanan Informasi*, (2017).
5. W. Enck, D. Octeau, P. McDaniel, S. Chaudhuri, *A study of android application security*, Proc. 20th USENIX Secur. Symp., 2011.
6. W. Enck et al., *TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones*, ACM Trans. Computer Syst. 32 (2014) 1-15.
7. Y. Yang, *Web user behavioral profiling for user identificatio*n, Decis. Support Syst. 49 (2010) 261-271.
8. F. M. Naini, J. Unnikrishnan, P. Thiran, M. Vetterli, *Where you are is who you are: user identification by matching statistics*, IEEE Trans. Inf. Forensics Secur. 11 (2016) 358-372.
9. Y. A. De Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, *Unique in the Crowd: the privacy bounds of human mobility*, Sci. Rep. 3 (2013) 1376.
10. T. Stöber, M. Frank, J. Schmitt, I. Martinovic, *Who do you sync you are? Smartphone fingerprinting via application behaviour*, Proc. 6th ACM Conf. Secur. Priv. Wirel. Mob. Networks, 2013, pp. 7-12.
11. Z. Tu et al., *Your apps give you away*, Proc. ACM Interac. Mob. Wearable Ubiquitous Technol., 2, 2018.
12. P. Welke, I. Andone, K. Błaszkiewicz, A. Markowetz, *Differentiating smartphone users by app usage*, Proc. 2016 ACM Int. Joint Conf. Pervasive Ubiquitous Comput, 2016.
13. J. R. Kwapisz, G. M. Weiss, S. A. Moore, *Cell phone-based biometric identification*, IEEE 4th Int. Conf. on Biometrics: Theory App. Syst., 2010.
14. W. Shi, J. Yang, Y. Jiang, F. Yang, Y. Xiong, *SenGuard: passive user identification on smartphones using multiple sensors*, Int. Conf. Wirel. Mob. Comput., Networking Commun., 2011.
15. L. Rossi, J. Walker, M. Musolesi, *Spatio-temporal techniques for user identification by means of GPS mobility data*, EPJ Data Sci., 4 (2015) 11.
16. W. Cao, Z. Wu, D. Wang, J. Li, H. Wu, *Automatic user identification method across heterogeneous mobility data sources*, 2016 IEEE 32nd Int. Conf. Data Eng,, 2016.
17. S. Zahid, M. Shahzad, S. A. Khayam, M. Farooq, *Keystroke-based user identification on smart phones*, Lecture Notes Computer Sci., 5758 (2009) 224-243.
18. L. Sun, Y. Wang, B. Cao, P. S. Yu, W. Srisa-An, A. D. Leow, *Sequential keystroke behavioral biometrics for mobile user identification via multi-view deep learning*, Lecture Notes Computer Sci., 10536 (2017) 228-240.
19. T. Feng, J. Yang, Z. Yan, E. M. Tapia, and W. Shi, *TIPS: Context-aware implicit user identification using touch screen in uncontrolled environments*, Proc 15th Workshop Mob. Comput. Syst. and App., 2014.
20. T. Feng et al., *Continuous mobile authentication using touchscreen gestures*, 2012 IEEE Int. Conf. Technol. Homeland Secur., 2012.
21. C. Bo, L. Zhang, X. Y. Li, Q. Huang, Y. Wang, *SilentSense: silent user identification via touch and movement behavioral biometrics*, Proc. Ann. Int. Conf. Mob. Comput. Net 13 (2013) 187-190.
22. G. Chittaranjan, B. Jan, D. Gatica-Perez, *Who's who with big-five: analyzing and classifying personality traits with smartphones*, Proc. Int. Symp. Wearable Comput. 15 (2011) 29-36.
23. Y.A. de Montjoye, J. Quoidbach, F. Robic, A.S. Pentland, *Predicting personality using novel mobile phone-based metrics*, LNCS 7812 (2013) 48-55.
24. B. Mønsted, A. Mollgaard, J. Mathiesen, *Phone-based metric as a predictor for basic personality traits*, J. Res. Pers. 74 (2018) 16-22.
25. H. Ots, I. Liiv, and D. Tur, *Mobile phone usage data for credit scoring*, Comm. Computer Inf. Sci. 1243 (2020) 82-95.
26. S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, *Your installed apps reveal your gender and more!*, Proc. ACM MobiCom Workshop Secur. Privacy Mob. Environ., 2014.

27. S. Seneviratne, A. Seneviratne, P. Mohapatra, A. Mahanti, *Predicting user traits from a snapshot of apps installed on a smartphone*, ACM SIGMOBILE Mob. Comput. Commun. Rev. 18 (2014) 1-8.
28. I. M. Pires, N. M. Garcia, N. Pombo, F. Flórez-Revuelta, S. Spinsante, M. C. Teixeira, *Identification of activities of daily living through data fusion on motion and magnetic sensors embedded on mobile devices*, Pervasive Mob. Comput. 47 (2018) 78-93.
29. R. Gellert and S. Gutwirth, *Beyond accountability, the return to privacy?*, Manag. Privacy Through Account. (2012) 261-283.
30. S. Zhao et al., *User profiling from their use of smartphone applications: a survey*, Pervasive Mob. Comput., 59 (2019) 101052.
31. M. Alexios, *Smartphone spying tools*, MSc Thesis, University of London, United Kingdom, 2018.
32. A. Lukács, *What is privacy? the history and definition of privacy*, Budapest 2016 (2017) 256-265.
33. W. Djafar, *Hukum perlindungan data pribadi di indonesia: lanskap, urgensi dan kebutuhan pembaruan*, ELSAM (2019) 1-14.
34. European Parliament and the Council of the European Union, *Directive 95/EC of the european parliament and of the council*. 1995.
35. J. Isaak and Mina J. Hanna, *User data privacy: Facebook, Cambridge Analytica, and privacy protection*, Computer 51 (2018) 56-59.
36. E. Graham-Harrison and C. Cadwalladr, *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*, Guard., (2018) 1-5.
37. M. Theoharidou, A. Mylonas, D. Gritzalis, *A risk assessment method for smartphones*, IFIP Adv. Inf. Commun. Technol., 376 (2012) 443-456.
38. S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, C. Ratti, *Activity-aware map: identifying human daily activity pattern using mobile phone data*, Hum. Behav. Understanding 6219 (2010) 14-25.
39. V. Soto and E. Frías-Martínez, *Automated land use identification using cell-phone records*, Proc. 9th Int. Conf. Mob. Syst. Appl. Serv. Co-located Work., 2011.
40. B. C. Csáji et al., *Exploring the mobility of mobile phone users*, Phys. A Stat. Mech. its Appl. 392 (2013) 1459-1473.
41. V. Frias-Martinez and J. Virseda, *Cell phone analytics: scaling human behavior studies into the millions*, Inf. Technol. Int. Dev. 9 (2013) 35-50.
42. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, *A comprehensive survey on support vector machine classification: applications, challenges and trends*, Neurocomputing 408 (2020) 189-215.
43. UN Global Working Group on Big Data for Official Statistics, *Handbook on the use of mobile phone data for official statistics*, 2017.
44. M. D. Chinn and R. W. Fairlie, *ICT use in the developing world: an analysis of differences in computer and internet penetration*, Rev. Int. Econ., 18 (2010) 153-167.