Digital Communications and Networks 8 (2022) 324–332

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

CPFinder: Finding an unknown caller's profession from anonymized mobile phone data

Jiaquan Zhang [a], Hui Chen [b,*], Xiaoming Yao [c], Xiaoming Fu [a]

[a] Institution of Computer Science, University of Göttingen, Göttingen, 37077, Germany
[b] Beijing Foreign Studies University, Beijing, 100089, China
[c] Cloud Branch, China Telecom, Beijing, 100033, China

ARTICLE INFO

Keywords:
Mobile big data
Profession prediction
Machine learning
Classification
Privacy protection

ABSTRACT

Identifying an unfamiliar caller's profession is important to protect citizens' personal safety and property. Owing to the limited data protection of various popular online services in some countries, such as taxi hailing and ordering takeouts, many users presently encounter an increasing number of phone calls from strangers. The situation may be aggravated when criminals pretend to be such service delivery staff, threatening the user individuals as well as the society. In addition, numerous people experience excessive digital marketing and fraudulent phone calls because of personal information leakage. However, previous works on malicious call detection only focused on binary classification, which does not work for the identification of multiple professions. We observed that web service requests issued from users' mobile phones might exhibit their application preferences, spatial and temporal patterns, and other profession-related information. This offers researchers and engineers a hint to identify unfamiliar callers. In fact, some previous works already leveraged raw data from mobile phones (which includes sensitive information) for personality studies. However, accessing users' mobile phone raw data may violate the more and more strict private data protection policies and regulations (e.g., General Data Protection Regulation). We observe that appropriate statistical methods can offer an effective means to eliminate private information and preserve personal characteristics, thus enabling the identification of the types of mobile phone callers without privacy concerns. In this paper, we develop CPFinder —- a system that exploits privacy-preserving mobile data to automatically identify callers who are divided into four categories of users: taxi drivers, delivery and takeouts staffs, telemarketers and fraudsters, and normal users (other professions). Our evaluation of an anonymized dataset of 1,282 users over a period of 3 months in Shanghai City shows that the CPFinder can achieve accuracies of more than 75.0% and 92.4% for multiclass and binary classifications, respectively.

1. Introduction

With the wide use of online services, such as online shopping, food delivery orders, and taxi hailing, there is a significant annual growth in platform-to-consumer delivery (8.2%) [1], restaurant-to-consumer delivery (6.8%) [2], ride and taxi hailing (17.5%) [3]. As a result, numerous people frequently receive unfamiliar phone calls from delivery people and taxi drivers. Moreover, many people are suffering from digital marketing and fraudulent phone calls due to the leakage of personal information from phone number related services. Criminals may pretend to be delivery men and contact victims by making phone calls, which could cause personal safety crises and property loss. Similarly, fraudulent phone and telemarketing calls cause millions of financial losses yearly [4]. Therefore, it is critical for mobile phone users to recognize the professions of callers to protect themselves from potential dangers and losses. The phone identification systems, such as Baidu phone number labeling[1] and 360 phone number query [2], offer an effective way of solving the aforementioned problems by retrieving information about each phone call accumulated before. The telephone identification system originally developed was only used to identify fraudulent calls. With the

* Corresponding author.
E-mail addresses: jiaquan.zhang@cs.uni-goettingen.de (J. Zhang), chenhui@bfsu.edu.cn (H. Chen), yaoxm@chinatelecom.cn (X. Yao), fu@cs.uni-goettingen.de (X. Fu).

[1] https://haoma.baidu.com/mark.
[2] https://chaxun.360.cn/chaxun/tel.

https://doi.org/10.1016/j.dcan.2021.08.003
Received 29 January 2021; Received in revised form 22 July 2021; Accepted 10 August 2021
Available online 17 August 2021

emergence of smart phones, the telephone identification system was extended to cover new types of telephone users, such as delivery men and taxi drivers. However, the currently off-the-shelf phone identification systems face the following challenges, especially concerning efficiency and accuracy. First, these phone identification systems are based on manual and subjective reports and annotations from end-users. As shown in Fig. 1, a phone number is labeled for advertisement purposes and has been reported by nine users, as shown in the Baidu search. However, owing to the laziness and unwillingness of many end users, the annotation rate is very low (0.154), and a high percentage of incorrect labels are generated [5]. Similar observations are also made in Google searches. The incorrect labels could bring much inconvenience for normal users, e.g., their phone calls may be refused by others if they are mislabeled as fraud phones. To make the service more accurate and avoid wrong information due to incorrect labels, only the phone numbers with sufficient reported times could be released to the public by phone labeling services providers. Therefore, there will be a long gap in releasing the label information since a phone number was first labeled. To overcome these problems and improve the performance of the phone labeling systems, we developed *CPFinder*, a new machine learning-based system to automatically identify phone labels by leveraging only anonymized mobile phone cellular data without the need for any manual interventions.

Smart phones exchange large amounts of data traffic and application (APP) request data through the cellular base station. These requests could provide lots of useful information to enable researchers and engineers to study people's daily behaviors and living habits, which is highly related to people's occupation [6–10]. However, privacy protection remains a critical issue when handling personal data [11,12]. Most personality studies use raw mobile phone data, including sensitive personal information, such as locations, application logs, and charging records. Often, more detailed personal data in the survey are also included. The major privacy concern in these studies is believed to be addressed by encrypting personal IDs, such that the owners of the data are claimed to be protected. However, a user could be re-identified even though the labels are invisible [10,12] as long as the input data remains unchanged. Personal locations, application usages, and activities could be used to trace and (re-)identify a person.

As personal data protection policies become more stringent, using sensitive information for personality identification may violate various data protection regulations. For example, General Data Protection Regulation (GDPR Art. 71) mentioned that private data, such as personal preferences or interests, reliability or behaviour and locations, should be processed by appropriate mathematical or statistical procedures for privacy concerns. Compared with where people visit (exact coordinates) and what apps people use (app logs), how people move and use smart phones are less sensitive. Typically, these statistical parameters could not be used to reversely trace any personal information and will not violate privacy protection regulations. As GDPR recital 71 recommends [13], appropriate statistical procedures should be used when analyzing personal data to prevent potential risks of retrieving personal interest and privacy.

In this work, based on our analysis of an anonymous dataset from one of the largest telecommunication operators in China, we develop CPFinder, a machine learning-based intelligent phone user profession identification framework, which can identify phone labels among four categories, including normal users, taxi drivers, takeouts and delivery, fraud, and telemarketing. According to the information provided by various apps' network traffic and the professional attributes of each category, we develop comprehensive features via statistical analysis, and some nonsensitive parameters based on privacy eliminated data are created to model users' patterns. We show that only with the mobile phone user data of a single day, the professional category of a user could be identified with high accuracy. The key highlights of the proposed framework are as follows:

- It uses privacy-preserved personal data as the input of user identification, which enables end-device data processing and complies with data protection regulations.
- It contains a multiclass classification module, which complements previous binary classification methods. Binary classification is also embedded in the system, and outperforms some previous works [4,6].
- It significantly reduces the number of features and the requirement for data size, which helps improve the efficiency by 10 times compared with feature-rich methods.

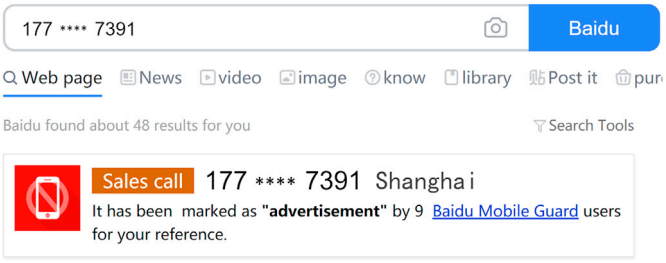


Fig. 1. Baidu phone label search result.

2. Related work

Mobile phone data have received increasing attention from data science researchers, especially in social computing areas [7–9,14–16]. Some studies have analyzed the mobility patterns of populations on the city level [17]. They found people's moving distances follow a scale-free distribution. Public health [18] and regional air pollution [19] could also be inferred by mobile phone data. These studies analyzed group behaviors and the corresponding statistical parameters of certain districts. With the development and wide deployment of high-speed cellular networking technologies, users' data volume significantly increases, which makes it possible for researchers to study personal characteristics based on big data analysis [20,21]. As a result, studies on profession, human activity, age, and gender prediction have become widespread. For example, mobile phone data are used to identify personalities and could implement multiclass classification [7,8]. In particular, the usage of apps has become the main source feature for identifying personal characteristics [10,21]. Besides phone-based data, these works also included other information, such as age and gender from surveys or questionnaires. From these works' perspectives, simply relying on artificially collected data is inefficient and unsuitable for real-time identification.

There are also many studies on detecting harassing or fraudulent phone callers using mobile phone data for run-time applications, which are mostly considered as binary classification problems [22–24]. Furthermore, most of these studies used Call Detail Records (CDR) data to derive their predictive models [22,25,26]. CDR data are very useful for fraudulent phone detection because fraudsters significantly differ from normal users in terms of dialing properties. Meanwhile, content analysis based on natural language processing or artificial intelligence methods has recently attracted attention [27,28]. However, these types of data are not enough to identify strange callers' professions, which is a multiclass classification problem. People of different professions, such as delivery people, taxi drivers, and telemarketers, share similar dialing properties as they have to frequently communicate with customers. CDR data could also provide location information estimates based on the base stations' locations [9,15,25]. However, there is a large deviation in these locations because each tower is in a fixed position and with a coverage of up to 1-km radius in most urban areas. Liao et al. [29] show GPS data could provide more precise location information, enabling people to build personal maps for mobility pattern analysis.

3. Problem statement

In this paper, we develop a method to allow mobile phone users to

identify a caller's profession immediately when they receive a call from an unknown number. The data used for identification are primarily the web service requests issued from the devices and recorded by telecommunication operators, who could push the identification results to end-users when necessary. Suppose there is a group of mobile phone users U, and each user has a phone number p_i, a label of profession l_i, and mobile phone records R_i, where $U = \{[p_1, l_1, R_1], [p_2, l_2, R_2], \ldots, [p_n, l_n, R_n]\}$. The mobile phone records are web services requests, that is, $R_i = \{[time, request], [time, request], \ldots\}$. *Time* is in the form of *YYYYMMDDHHmmSS* with a precision of seconds, and the request is always formed as *www.xxx.com/ … longitude = xx& latitude = xx/ ….* The problem is to use the data of R_i to identify the label of profession l_i without any external data resource. Currently, four categories of professions are included, where $l_i \in$ {*Normal*, *Driver*, *Delivery &*; *Takeouts*, *Telemarketing &*; *Fraud*}.

Based on the information of web services requests, three kinds of highly profession-related and sensitive information-eliminated features are constructed: mobility pattern, request volume, and app preferences.

3.1. Mobility pattern

A user's mobility pattern characterizes the user's overall movements throughout the day. A statistical parameter, Standard Deviation (SD) is applied to describe mobility. Because the location records are not always complete, using statistical parameters could maximally preserve mobility characteristics. The four categories of professions have very different moving ranges on every single working day.

3.2. Request volume

Request volume tells the number of requests at different times, which could indicate people's activeness. The more requests, the more active a user is. Different professions have significantly different active times. Request volume is different from data volume, as one request of video may generate much more data transmission than several requests of pure text. Therefore, the number of requests is much more suitable to characterize activeness.

3.3. App preferences

App preferences indicate how frequently each App is used. The apps here are web services extracted from the main domain names of the requests because the requests could be generated by browsers or many other apps. The main domain name is extracted by eliminating the prefix and suffix representing the same institution (i.e., didi taxi out of www.poiservice.diditaxt.com). Therefore, two different domains, www.common.diditaxi.com.cn, and www.poiservice.diditaxi.com are referred to as identical main domain names, which indicates a root related web service. Each profession has its own several frequently used services, which results in its unique app preferences pattern.

3.4. Time span of data

The time span of the data is also considered to be a factor for modeling. Each of the above-mentioned features could be computed using the data of an arbitrary number of days, but at least one day is needed. Generally, if more days are included in the data, a better identification performance will be achieved. The problem is how much the accuracy improves with the increasing of the time length of data. It addresses the question of identifying efficiency, which cares about if the data of one single day is sufficient to identify a caller's profession. The overall privacy preservation processing for CPFinder is shown in Fig. 2.

4. Mobile phone users' identification framework

4.1. Mobility pattern

Researchers found that mobility trajectories could be useful in identifying personalities [10,15]. In these works, users' exact coordinates are used to locate their work and residential places. Exact coordinates could provide the possibility of tracing where users are and thus may face personal data protection issues. However, using some statistical parameters to characterize people's mobility patterns, instead of using exact locations, could avoid violating any data protection policies. To ensure data privacy protection, any information about how people move cannot be used to trace personal locations.

Our analysis shows mobility pattern is strongly related to users' professions, especially on workdays. This coincides with the following intuitive scenario: during holidays, people's activities are self-related with high randomness, which means it is hard to find a common mobility pattern for people with an identical label if considering their mobility in leisure time. For this reason, we first study users' mobility patterns of active time (6:00–24:00) on workdays. As shown in Fig. 3, three typical users with different labels are presented with their locations in one week. Not surprisingly, each category of users has its unique mobility pattern: taxi drivers cover a large area and have large ranges in all directions, takeouts and delivery staffs mostly serve in certain regions with limited areas, whereas telemarketing and fraud phone callers mostly show a strip-like commuting pattern (which indicates their commutes between their residences and offices). Normal users are excluded in the figure because normal users' mobility patterns are not similar and are of large bias among different entities.

Based on the above analysis, we propose an SD-based parameter to denote the mobility patterns of different professions. The parameter contains the ranges of 12 evenly distributed directions on the flat. For a given set of N points and a certain direction d (angled with x-axis of θ), the range of the points in direction d is as follows:

$$R_\theta^2 = \frac{1}{N} \cdot \sum_1^N \left(x_i \cos\theta + y_i \sin\theta - x\cos\theta + y\sin\theta \right)^2 \quad (1)$$

where N is the number of points, x_i and y_i are the longitude and latitude of each point.

For each user, his/her daily mobility pattern is characterized by a vector, containing ranges of several different but evenly distributed directions. In addition, the range vector is sorted from maximum to minimum to eliminate the directional bias. For a certain pattern, its sorted range vector remains the same no matter how it rotates. As shown in Fig. 4(a), the sorted ranges of 12 directions significantly differ among the three professions. Drivers and telemarketers have a decreasing trend of the sorted ranges, but driver's range is much larger than telemarketing and fraud users. Delivery staff has a more balanced patterns, where there are rare differences in all directions. For the normal users, their patterns are somewhere between drivers and delivery people, and no unified pattern could be found because their trajectories vary among different people. However, they are different from the other professions in terms of the following features. The number of directions used for characterizing a person's mobility pattern needs investigation in terms of both accuracy and efficiency. As shown in Fig. 3, when increasing the number of directions, the accuracy remains almost unchanged after more than 12 directions, which means using more than 12 directions will only increase the computational complexity without bringing any improvement to the performance.

4.2. User data volume in different time periods

Data volume can indicate the frequency of using mobile phones. The higher the data volume, the more network requests are generated, which also means the mobile phone is more frequently used. The data volumes

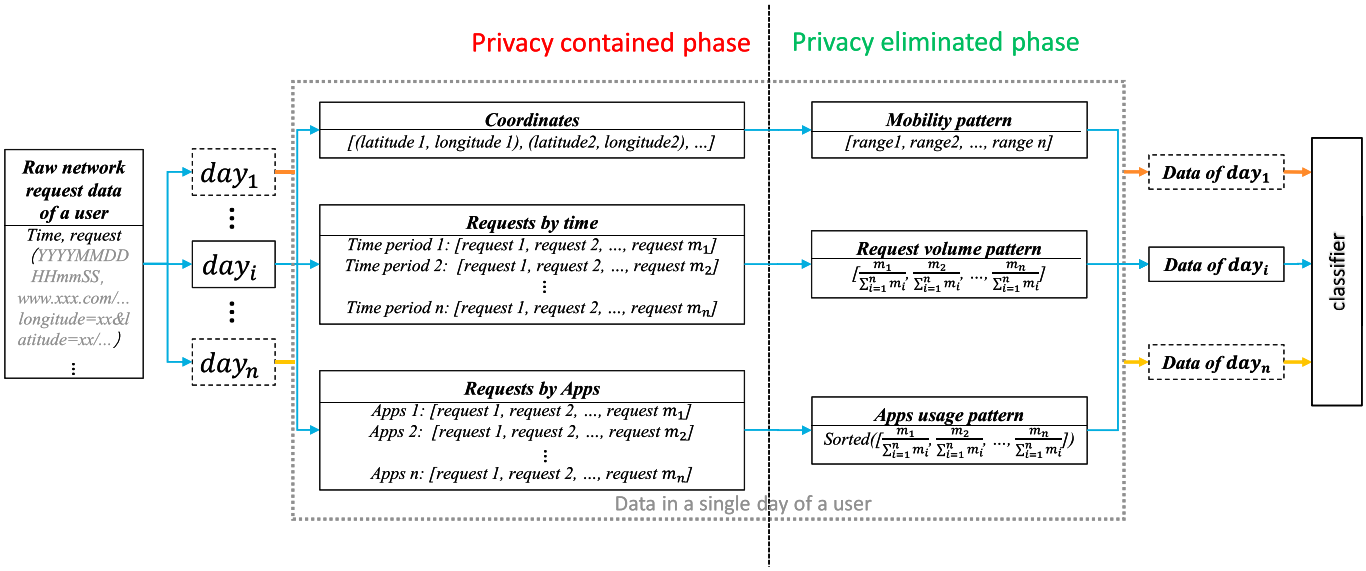


Fig. 2. CPFinder privacy preservation process.

in different time slices are also related to professions. Each profession has its unique pattern of active duration. As shown in Fig. 5, a day is divided into 6 equal time slices: 6:00–9:00, 9:00–12:00, 12:00–15:00, 15:00–18:00, 18:00–21:00, 21:00–24:00. Data between 0:00 a.m. and 6:00 a.m. is excluded because there are too few records during this period. Since some apps may automatically generate network requests when executing in the background, this kind of execution usually generates many repeated requests within a second. These large amounts of repeated requests, which are generated within a short period, could significantly interfere with the analysis of data usage in different periods. Therefore, if a request has many duplicated records for each timestamp, the duplicated records will be deleted. Fig. 5 shows the distribution of a number of network requests for all professions. Drivers and telemarketers share a very similar data usage pattern; they are more active than the other two professions after 18:00. For delivery staff, their data volume is somehow evenly distributed in different time slices and is more stable than the others. For normal users, the data volume distribution is slightly different from the delivery people's. For each user, his/her data volume

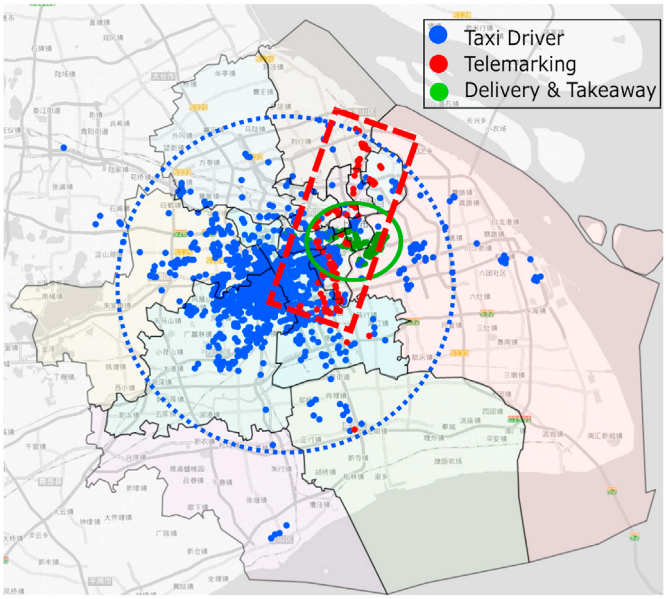


Fig. 3. Locations of three different professions.

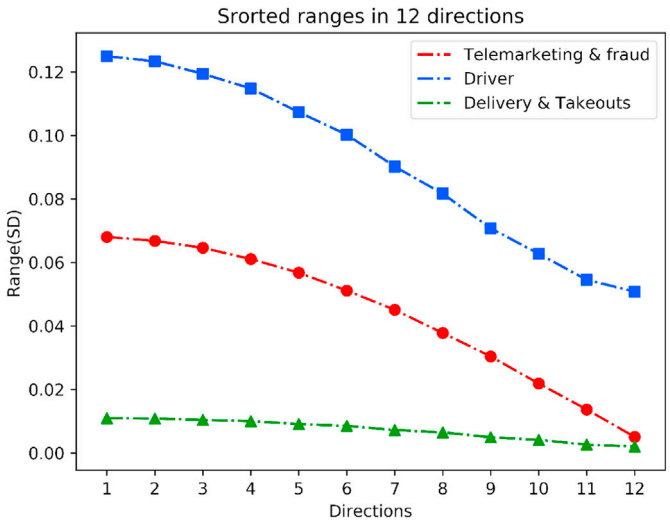
pattern in a single day is created by calculating the daily network request distributions, which is a vector of six elements. The ith element is given by $v_i = n_i/n_{total}$, where n_i is the number of network requests in the time slice and n_{total} is the number of requests in an entire day.

4.3. Apps preference distribution

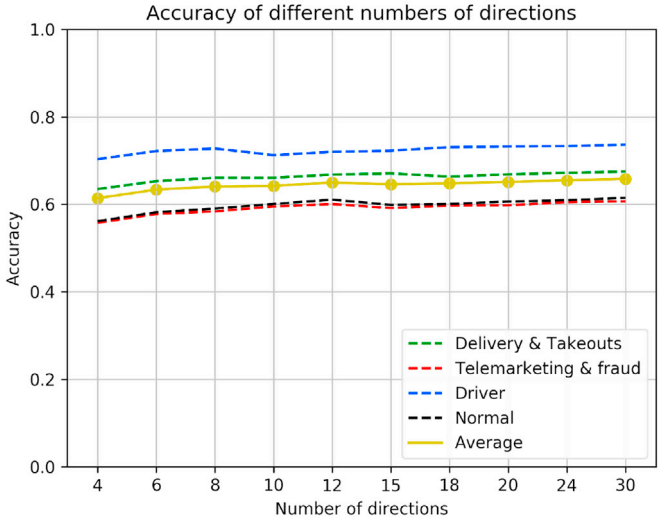
Mobile phone app usage plays an important role in personality studies [7,8,10,30]. These studies attempted to figure out what apps people use, what categories of apps each person prefers to use, as well as how someone uses different apps. Apps log containing what apps are used, as well as when and how long an App is used, could easily be leveraged by others to forward advertisements to certain devices. For privacy concerns, the plaintext of app names should not be contained in any parameters. In addition, the exact apps' names could not be directly used as inputs unless they are transferred as numeric values, e.g., mapping each App to an integer. Moreover, as there are thousands of apps used by people, it is very complex to create users' app preference patterns when considering the exact app. Further, as there are new apps released frequently, it is impossible to maintain the mapping list when considering the new apps. A more efficient and more accurate way to characterize users' App preferences is finding the distribution of 10 mostly used apps, regardless of the exact hostnames. Each network request is related to a web service, which can be found from the host. Fig. 6 shows the normalized daily app usage distribution (sorted from the most frequently used to the least frequently used) of four categories. Except for normal users, other professions have a steep descent in their apps' usage distributions, which means they have several frequently used apps or services. In particular, delivery people have the least frequently used apps or services. Meanwhile, the normal users show a more balanced distribution. The best number of most frequently used apps is also investigated in Fig. 6(b). With more than ten apps, the percentages of the four categories are all extremely low and close to each other. As a result, it is unnecessary to include more than 10 apps. Moreover, the accuracy tends to decrease slightly when including more than ten apps.

4.4. CPFinder system implementation

Based on the above analysis, the users' data will not contain sensitive personal information after privacy-preserving processing. The privacy preserved data could then be stored out of the centralized secure server and even be transmitted to end devices. Previous methods use original



(a) Sorted standard deviations of 12 directions



(b) Accuracy of different numbers of directions

Fig. 4. Users' mobility patterns in active time of workdays and its statistical results.

data that contain sensitive personal information to build features, resulting in the limitations that the entire procedure (including both training and identifying phases) should be operated within a secured server. Then, only the identification results will be directly transmitted to the end-users. Considering that there are numerous cell phone calls every second, generating users' data and identifying the labels for all phone calls are a big challenge for the server.

However, if the identification task could be pushed to be done in each end device, the workload of the centralized server could be substantially reduced in terms of both computational amount and storage. Transmitting the input data from a secure server to end devices is feasible only when the input data do not contain any sensitive personal information, in case of not violating some data protection policies and regulations. Distributing the privacy-preserved data is acceptable as the data cannot be used for any malicious purpose. As shown in Fig. 7, the private user-generated data, which is stored in the centralized secure server, is first processed to eliminate sensitive information (by Fig. 2) and exported as privacy-preserved data. Then, the data could be distributed to servers with lower security levels or even unsecured servers. The servers could train a model with the privacy-preserved data periodically, and the identification models in the end devices could be kept updated when a newly trained model is available. When a device receives a call from an unfamiliar number, the server could transmit the corresponding data of the number to the device simultaneously for identification.

The CPFinder system is also flexible and efficient, as it does not require excessive computational and network resources. For the secure database maintaining original user data, the main computation is processing the user data into a privacy-preserved form and transmitting the processed data to distributed insecure clouds. The processing of the user data only has to traverse each user just once, and the time complexity is linear. For each user, the additional data output by the module is only 28 float numbers and a phone number, which is much smaller than the size of original data and is portable for transmission. For the clouds that hold non-private user data, the main computation is training the models and transmitting the identification models to the end devices. As the above tasks only have to be done periodically, they do not require powerful CPUs or large bandwidths. The only real-time application is when an end device receives an unfamiliar phone call, and the end device has to request the data of the specific users from the clouds and identify the label. However, the data query only needs one hash and map, and the main body of transmission only contains 28 float numbers.

5. Evaluation results and discussions

5.1. Dataset

The dataset of user mobile cellular data comes from a major telecommunication operator in China. Each record contains three fields: phone number, time stamp, and network request. The data contains 1,282 users in Shanghai City and covers the period from November 1, 2016, to February 16, 2017. In the total of 108 days, call records during the 35 Chinese public holidays are excluded due to their different nature from ordinary days. For the phone label data, if a phone number is labeled, the annotation information could be easily found by searching the phone number via the Baidu search engine (Fig. 2). An annotation is the detailed description of the phone number, which directly indicates the profession of the user. In addition, each labeled number would be assigned with a root annotation, which is formatted textual data containing several different types of professions. Originally, the online datasets typically contain the following annotations: fraud, harassment, illegality, insurance, finance management, intermediary, recruiting hunter, takeouts, delivery, driver, and customer service. These root annotations could be divided into three categories: taxi drivers, delivery, and harassment. Delivery includes express delivery and takeout delivery, whereas harassment is a general designation of harassment, fraud, and telemarketing. For numbers that are not annotated, they are regarded as normal users. The label crawling is done by the telecommunication operator within their internal server for privacy concerns, and the phone numbers are replaced by the categories of professionals for later processing.

Although there are only 1,282 users, each user has 73 days of data and could generate up to 73 different instances. However, the instances of a single user are only used in the training or testing phase because instances of the same user may show significant similarities and will seriously impact the performance of the models. If a testing instance belongs to a user whose other instances were used in the training phase, the testing instance will be correctly identified with a high probability. We seek to find a general pattern for individuals of the same profession rather than particular individuals. As a result, applying the instances of an identical person both in the training and testing phases will not make much sense, and a better testing data selection scheme should be based on persons rather than instances. Several regression and classification algorithms are used to evaluate the performance of the model. Each algorithm is fine-pretested to present its best performance. Table 1 shows

the distribution of user quantities in different categories. Considering the unbalanced distribution of the user quantities, 50 users are randomly selected from each category; a total of 200 users are formed for testing.

5.2. Evaluation results: multiclass classification

Table 2 shows the identification accuracy of all algorithms, the accuracy of each category of people is also presented. Each result is an average of 100 experiments with cross-validations. Except for the logistic regression, other classification algorithms perform very similarly. Random forest could achieve the highest overall accuracy of 75.64%. The taxi drivers, delivery and takeouts are more accurately identified, whereas the identification accuracy for the other two groups is slightly lower. The result shows our approach outperforms the performance of the approach presented in Ref. [6], whose accuracy is 70.4% for unemployment identification (and lower for exact profession identification) (see Table 3).

However, because a user's data on every single day is not always complete, merging the data of several days together may influence the accuracy. Therefore, we also investigate how time duration impacts the performance of the model by combining data from different numbers of days. The combination is not the average of the statistical parameters from different days, but is the one-time calculation of all data within these days. For example, the mobility pattern of a two-day combined data is the range of all locations during the two days. We study the impact of combining data in the training phase while the test data remains for the duration of one day. Reversely, the combination of testing data without combination of training data is also investigated. Fig. 8 shows the identification accuracy of combining training data over different time lengths.

The identification performance does not change too much when the number of days of data combination increases, which means that combining data for the training phase makes less sense. However, the accuracy slightly improves when the number of days for data combination increases. The improvement is not significant even when comparing the accuracy of combining 7 days of data together with non-combination data. In general, we can see from our data that using the data of a single day is sufficient to identify a strange caller's profession with a good performance, even though adding data of more days could somehow improve the accuracy.

The latitudes and longitudes extracted from mobile cellular network requests could indicate the exact positions of the users, but the dataset could not fully contain all locations or trajectories of a user because not each network request contains location related information. Generally, as users could control the access of location services for each APP, most people prefer to turn off the location services for most apps except map-related applications such as ride-hailing applications (e.g., Uber and DiDi). The incomplete location information brings the bias between the actual and recorded patterns of a user, which turns out to influence the performance of the proposed method. This is why the accuracy slightly improves with increasing the number of days for data combination. In a long period of nearly 3 months, some users may change their professions. These changes negatively impact the accuracy of identifying callers' professions, and it is hard to verify how serious is the impact because how and when people change their professions are not recorded.

Moreover, the deviation of personalities between individuals of the same label also negatively influences the performance of the model. This challenge holds true for most personality-related studies, especially for multiclass classification problems [6–8]. Overall, CPFinder achieves relatively good performances based on such limited information and outperforms some previous methods.

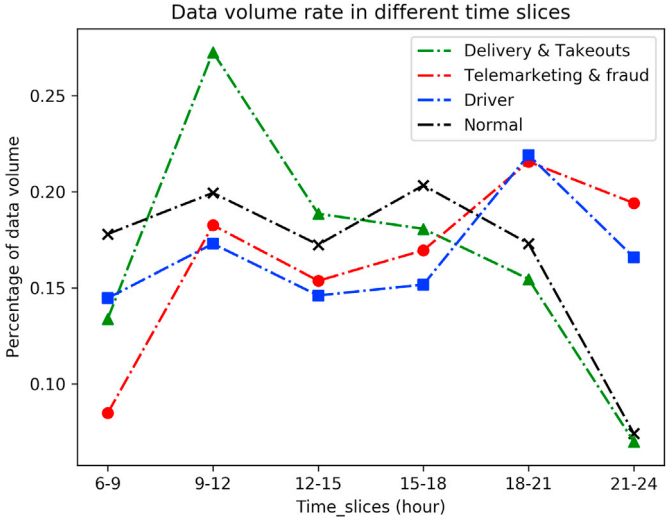
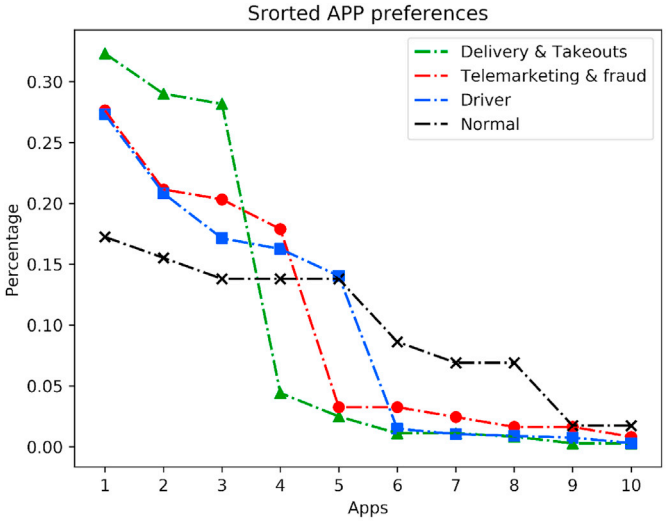
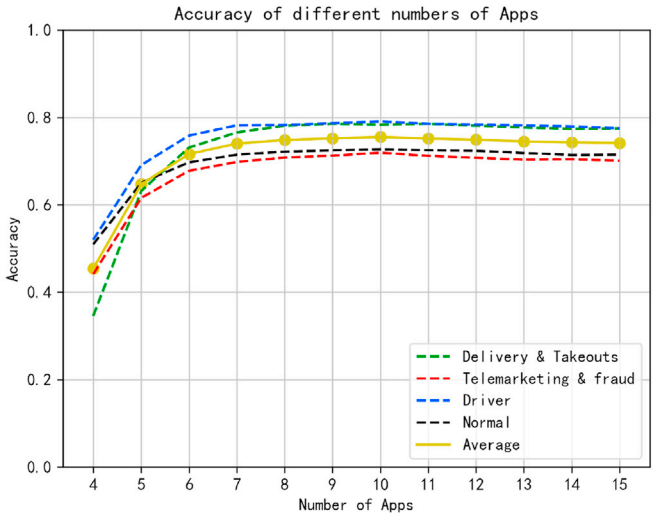


Fig. 5. Data volume distribution in different time slices.



(a) Sorted usage rate of 10 most frequently used apps



(b) Accuracy of different numbers of apps

Fig. 6. Apps preference.

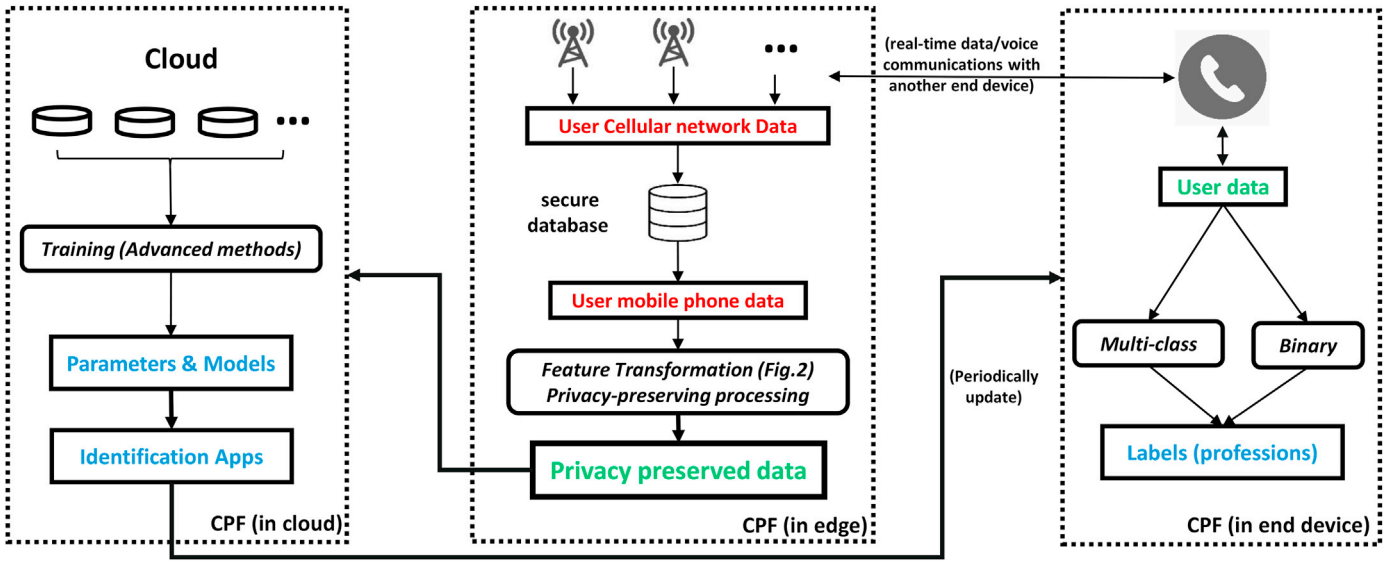


Fig. 7. CPFinder implementation: an overview.

5.3. Evaluation results: binary identification

Mobile phone users typically have some prior knowledge of incoming phone calls. If a person orders a taxi and receives a call later from an unfamiliar phone number, he/she just needs some information about whether the caller is a driver from the taxi company that he just ordered. In addition, there is a scenario where the user may need to quickly identify after receiving a call and knowing some basic information from the caller, e.g., the caller states as a driver or delivery staff. In such scenarios, a binary identification rather than multiclass classification can adequately meet the demands of providing information to identify the callers. Therefore, for a specific category of callers, a binary model regarding the other categories as negative labels can be created to serve as a special binary identification system, which typically requires quick and accurate identification. At the moment, each category should have its own model and be trained individually with the data. To demonstrate the advantage of the method, we compare the results with the work of using CDR data to identify harass calls (DeMalC) [5] and using real-time mobile phone data to predict unemployment and professions (TRP) [6]. Based on the results in Ref. [5], we adopt exactly the same algorithms, that is, GBDT [31] for training the models. All metrics are compared, including precision, recall, F1-score and overall accuracy, between our methods and others'. As there are no other metrics except the accuracy in TRP, only the overall accuracy of unemployment prediction is compared here.

First, CPFinder outperforms the DeMalC method when used for identifying harass calls. DeMalC is based on CDR data and some explicit information about mobile devices and calls, including IP addresses, base station tower locations, and call frequencies. As mentioned above, the data used for the inputs of this method contain much plaintext of sensitive personal information, which could be used to trace a person's trajectory or retrieve personal behavior. Moreover, using such data requires the approval of the data owners. Although our method outperforms DeMalC in identifying harassing phone calls, and the accuracy is improved by about 0.01, it contains fewer features and variables: CPFinder contains only 3 features and 28 variables in total, whereas DeMalC needs 7 features and up to 190 variables. Obviously, fewer variables could lead to higher efficiency in terms of shorter training time and lower identifying speed, as well as the faster transmission of the data. Our method also adopts privacy-preserved features. Hence the input data could be transmitted to end devices or third-party servers to further speed up the computational efficiency. The performances of two other categories are better than the category of harassment and could achieve an accuracy of more than 0.93 for driver identification; TRP also uses CDR data to generate feature-rich models for predicting personal employment status and professions. The best result TRP could achieve for predicting binary status (employed or unemployed) is 0.704. Moreover, the overall average accuracy of predicting personal professions is 0.675, which is worse than our original model used for multiclass identification.

Both DeMalC and TRP contain a large number of features (up to 160) and variables, which require high computational complexity. As is well known, the time complexity of the decision tree is $O(m)$, whereas for XGboost algorithm, it is $O(m^2)$, where m is the number of the attributes. As a result, our method will be more than six times faster than DeMalC and more than five times faster than TRP if applying algorithms with linear time complexity, and the time efficiency will be quadratically increased if algorithms with $O(m^2)$ time complexity are applied in the model.

In addition to the overall accuracy, the precision of our method also outperforms the counterpart methods. The overall precision of all three categories is around 0.85, and the false positive rate is about 0.15. The false-positive rate means that for each positive identification, it is not actually a positive sample with a probability of 15%. In whatever conditions, misidentifying negative to positive is a big problem, e.g., identifying normal phone numbers as harassing phone calls may make users miss important calls. However, reversely identifying harassing phone calls as normal calls is somehow less severe than missing important calls. Currently, the methods used in the paper are optimized by maximizing overall accuracy. If the precision is preferable to the accuracy, the optimization function could be changed to maximize the precision, where the false-positive rate will decrease, but the overall accuracy will also decrease. Therefore, there is a consideration of the balance between precision and accuracy.

5.4. Capability of privacy-preserving

We compare our approach with other works (e.g., DeMalC and TRP) in terms of privacy preservation of user data. In addition to the advantage of reducing the dimensions of the input features and computational complexity, adopting privacy-preserved data in the training phase could also bring other benefits, such as enabling end device identification and looser requirements for data arrangements.

5.4.1. Prevent re-identification of personal data

According to the study in Ref. [10], mobile phone data could be used to identify a person. In other words, a person could be uniquely re-identified in the crowd by the apps he/she uses and the locations

Table 1
Phone labels distribution within the dataset.

Group name	Group Description	Number of users
Normal	Normal users	591
Drivers	Taxi drivers	176
Delivery	Delivery & Takeouts men	183
Harass	Telemarketing & Fraud people	332

he/she visits. However, the identification requires investigating a few frequently used apps by knowing their names and how they operate on the devices. In terms of locations, with additional information on coordinates, a person is entirely exposed to the public if such data is disclosed. Therefore, using original sensitive personal data brings a lot of risks to people's privacy, even in the training phase. So obviously, the original personal data should not be distributed outside the secure data centers to end devices. However, privacy-preserved data do not support identifying exactly who a person is, but only some non-private attributes, e.g., professions. DeMaLC requires city-level locations of the data, whereas TRP requires some information about a person's home district and the charge amount of each account. Consequently, the attributes of the two models do not comply with data protection when the data are distributed.

5.4.2. *Privacy considerations on data acquisition and processing*

According to various data protection regulations and policies (e.g., Art. 5 GDPR), collecting personal data for any purpose of processing should be conducted under the consent of the data owners. Consequently, the data used in the methods of DeMaIC and TRP, as well as the data in many studies that use personal mobile phone data for various purposes are mostly declared to be used under the agreement of the data owners. Although the data used in the case study in this paper also complies with the consent of the users, our model could be further generalized to the data without the users' additional agreement. Because users' requests with encryption through the network are public to all people, our approach just needs the information of some frequently used apps without knowing their exact names. For example, the name of apps could be hashed and captured by the network edge devices, and the same hash value indicates a unique app, and the data is somehow public. As a consequence, the users' preferences for apps could be collected and processed in a privacy-preserving manner without violating data privacy regulations. In contrast, as the names of the apps could not be reversely retrieved from the encrypted requests, the results obtained in previous methods relying on data containing plaintext of personal data are not easily reproducible since they are not applicable to privacy-preserved data.

6. Concluding remarks

We present CPFinder, a new methodological framework and system for identifying a mobile phone caller's profession using mobile cellular data, i.e., users' cellular network traffic recorded by telecommunication operators. Three kinds of privacy-preserving features are constructed based on the information provided by the cellular data: mobility pattern, data volume distribution, and app preferences. All features are formed by several statistical parameters, which exclude sensitive information such as coordinates and app logs. This private information could be used to reversely trace people's locations and living habits, thus eliminating the private information, even for input data, and it could avoid violating any data privacy regulation.

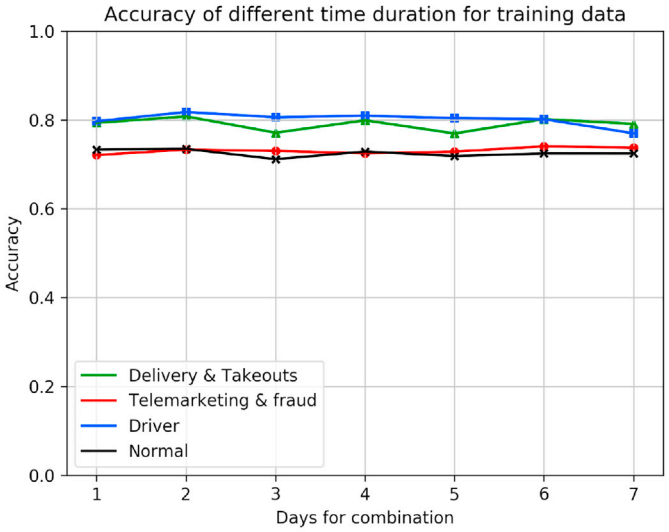
We apply several state-of-the-art classification and regression algorithms in our model; the experimental results of a dataset containing 1,282 users in Shanghai City prove that CPFinder could achieve an accuracy of 0.7564 (against the alternative methods that achieve accuracies up to 0.7358) when identifying four different categories of callers:

Table 2
Identification accuracy of different categories and different methods.

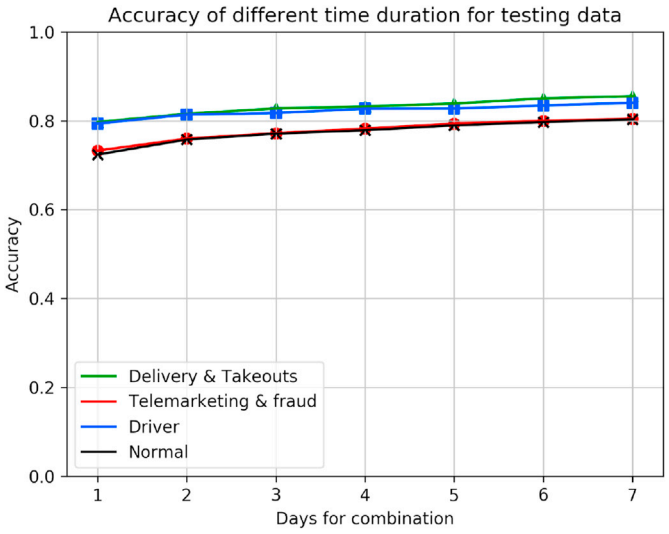
	Normal	Driver	Delivery	Harass	Overall
LR	0.5677	0.6284	0.6257	0.6113	0.6083
KNN	0.6970	0.7771	0.7705	0.6864	0.7328
RF	0.7300	0.7912	0.7884	0.7160	**0.7564**
SVM-RBF	0.7062	0.7625	0.7633	0.7112	0.7358
ADABOOST	0.6865	0.7404	0.7510	0.6888	0.7167
NN (MLP)	0.7216	0.7505	0.7366	0.7142	0.7307

Table 3
Experimental results of binary classification.

	Precision	Recall	F1-score	Accuracy
Driver (GBDT)	0.8672	0.8012	0.8329	**0.9351**
Harass (GBDT)	0.8338	0.7935	0.8132	**0.9287**
Delivery (GBDT)	0.8476	0.7790	0.8119	**0.9235**
DeMalC (GBDT) [5]	0.8395	0.7521	0.7934	0.9186
TRP (DNN) [6]	–	–	–	0.704



(a) Training data combination



(b) Testing data combination

Fig. 8. Accuracy of different time duration for training (a) and testing (b) data.

normal users (other professions), drivers, delivery and harass. If the binary classification is applied to the same dataset, CPFinder achieves an accuracy of 0.923+, outperforming two existing approaches with accuracies of 0.704 (TRP) and 0.9186 (DeMalC).

Combining data of different time lengths is also investigated for both training and testing phases to study how the amount of data affects the identification performance. The result demonstrates that the data of a single day is sufficient to identify a caller's profession, even though the accuracy slightly improves when using combined data of several days. Again, the overall performance outperforms some previous studies in terms of both accuracy and efficiency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 824019 and China Scholarship Council (CSC), as well as the Fundamental Research Funds for Central Universities (No. 2020JJ014, YY19SSK05).

References

[1] Platform-to-consumer delivery - worldwide — statista market forecast. https://www.statista.com/outlook/376/100/platform-to-consumer-delivery/worldwide. , 2022 (accessed 9 May, 2022).
[2] Restaurant-to-consumer delivery - worldwide — statista market forecast. https://www.statista.com/outlook/375/100/restaurant-to-consumer-delivery/worldwide, 2022(accessed 9 May, 2022).
[3] Ride-hailing & taxi - worldwide — statista market forecast. https://www.statista.com/outlook/368/100/ride-hailing-taxi/worldwide, 2022(accessed 9 May, 2022).
[4] A growing threat to your finances: cell-phone account fraud. https://www.consumerreports.org/scams-fraud/cell-phone-account-fraud, 2022 (accessed 9 May, 2022).
[5] Y. Li, D. Hou, A. Pan, Z. Gong, Demalc: a feature-rich machine learning framework for malicious call detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1559–1567.
[6] P. Sundsøy, J. Bjelland, B.-A. Reme, E. Jahani, E. Wetter, L. Bengtsson, Towards real-time prediction of unemployment and profession, in: International Conference on Social Informatics, Springer, 2017, pp. 14–23.
[7] G. Chittaranjan, J. Blom, D. Gatica-Perez, Mining large-scale smartphone data for personality studies, Personal Ubiquitous Comput. 17 (3) (2013) 433–450.
[8] Y.-A. de Montjoye, J. Quoidbach, F. Robic, A.S. Pentland, Predicting personality using novel mobile phone-based metrics, in: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2013, pp. 48–55.
[9] L. Gabrielli, B. Furletti, R. Trasarti, F. Giannotti, D. Pedreschi, City users' classification with mobile phone data, in: 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015, pp. 1007–1012.
[10] Z. Tu, R. Li, Y. Li, G. Wang, D. Wu, P. Hui, L. Su, D. Jin, Your apps give you away: distinguishing mobile users by their app usage fingerprints, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 2018, pp. 1–23, 3.
[11] K. Kenthapadi, I. Mironov, A.G. Thakurta, Privacy-preserving data mining in industry, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 840–841.
[12] D. Su, H.T. Huynh, Z. Chen, Y. Lu, W. Lu, Re-identification attack to privacy-preserving data analysis with noisy sample-mean, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1045–1053.
[13] Gdpr Recital 71, https://gdpr-info.eu/recitals/no-71/. 2022(accessed 9 May, 2022).
[14] W. Chen, Q. Gao, H. Xiong, Temporal predictability of online behavior in foursquare, Entropy 18 (8) (2016) 1–11, article number 296.
[15] Y. Yu, H. Chen, B. Ma, B.P.C. Yen, Utilizing geospatial information in cellular data usage for key location prediction, in: Proceedings of the 51st Hawaii International Conference on System Sciences, Scholar Space, 2018, pp. 981–988.
[16] E. Jahani, P. Sundsøy, J. Bjelland, L. Bengtsson, Y.-A. de Montjoye, et al., Improving official statistics in emerging markets using machine learning and mobile phone data, EPJ Data Science 6 (1) (2017) 3, 1–21.
[17] F. Wang, W. Chen, Y. Zhao, T. Gu, S. Gao, H. Bao, Adaptively exploring population mobility patterns in flow visualization, IEEE Transactions on Intelligent Transportation Systems. 18 (8) (2017) 2250–2259.
[18] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A.A. Salah, R. Benjamins, C. Cattuto, et al., Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, Science Advances 6 (23) (2020) 1–6.
[19] M. Nyhan, I. Kloog, R. Britter, C. Ratti, P. Koutrakis, Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data, J. Expo. Sci. Environ. Epidemiol. 29 (2) (2019) 238.
[20] K.B. Stecher, S. Counts, Spontaneous inference of personality traits and effects on memory for online profiles, in: Proceedings of International Conference on Weblogs & Social Media (ICWSM), Seattle, WA, USA, AAAI Press, 2008, pp. 118–126.
[21] G. Chittaranjan, J. Blom, D. Gatica-Perez, Who's who with big-five: analyzing and classifying personality traits with smartphones, in: 2011 15th Annual International Symposium on Wearable Computers, IEEE, 2011, pp. 29–36.
[22] M.U. Khan, S.A. Khan, Social networks identification and analysis using call detail records, in: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, 2009, pp. 192–196.
[23] L. Peng, R. Lin, Fraud phone calls analysis based on label propagation community detection algorithm, in: 2018 IEEE World Congress on Services (SERVICES), 2018, pp. 23–24.
[24] X. Li, Y. Liu, M. Zhang, S. Ma, Fraudulent support telephone number identification based on co-occurrence information on the web. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014, pp. 108–114.
[25] C. Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, X. Ma, Analyzing and geo-visualizing individual human mobility patterns using mobile call records, in: 2010 18th International Conference on Geoinformatics, IEEE, 2010, pp. 1–7.
[26] Y. Yuan, K. Ji, R. Sun, K. Ma, Z. Chen, L. Wang, An integration method of classifiers for abnormal phone detection, in: 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), IEEE, 2019, pp. 1–6.
[27] L. Peng, R. Lin, Fraud phone calls analysis based on label propagation community detection algorithm, in: 2018 IEEE World Congress on Services (SERVICES), IEEE, 2018, pp. 23–24.
[28] Q. Zhao, K. Chen, T. Li, Y. Yang, X. Wang, Detecting telecommunication fraud by understanding the contents of a call, Cybersecurity 1 (1) (2018) 1–12.
[29] L. Liao, D.J. Patterson, D. Fox, H. Kautz, Building personal maps from gps data, Annals of the New York Academy of Sciences 1093 (1) (2006) 249–265.
[30] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. Armenta Lopez de Vicuña, N. Oliver, Towards a psychographic user model from mobile phone usage, in: CHI'11 Extended Abstracts on Human Factors in Computing Systems, 2011, pp. 2191–2196.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, Journal of Machine Learning Research. 12 (85) (2011) 2825–2830.