# Revealing Physical World Privacy Leakage by Cyberspace Cookie Logs

Huandong Wang, *Member, IEEE*, Chen Gao, Yong Li, *Senior Member, IEEE*, Zhi-Li Zhang, *Fellow, IEEE*, and Depeng Jin, *Member, IEEE*

***Abstract*—It is well-known that online services resort to various cookies to track users through users' online service identifiers (IDs) – in other words, when users access online services, various "fingerprints" are left behind in the cyberspace. As they roam around in the physical world while accessing online services via mobile devices, users also leave a series of "footprints" – i.e., hints about their physical locations – in the physical world. This poses a potent new threat to user privacy: one can potentially correlate the "fingerprints" left by the users in the cyberspace with "footprints" left in the physical world to infer and reveal leakage of user physical world privacy, such as frequent user locations or mobility trajectories in the physical world – we refer to this problem as *user physical world privacy leakage via user cyberspace privacy leakage*. In this paper we address the following fundamental question: what kind – and how much – of user *physical world privacy* might be leaked if we could get hold of such diverse network datasets *even without any physical location information*. In order to conduct an in-depth investigation of these questions, we utilize the network data collected via a DPI system at the routers within one of the largest Internet operator in Shanghai, China over a duration of one month. We decompose the fundamental question into the three problems: i) linkage of various online user IDs belonging to the same person via mobility pattern mining; ii) physical location classification via aggregate user mobility patterns over time; and iii) tracking user physical mobility. By developing novel and effective methods for solving each of these problems, we demonstrate that the question of user physical world privacy leakage via user cyberspace privacy leakage is not hypothetical, but indeed poses a real potent threat to user privacy.**

***Index Terms*—Privacy, spatio-temporal trajectories, identity linkage, location classification.**

Manuscript received March 29, 2019; revised September 4, 2019 and February 26, 2020; accepted April 25, 2020. Date of publication July 31, 2020; date of current version December 9, 2020. This work was supported in part by The National Key Research and Development Program of China under grant 2018YFB1800804, the National Nature Science Foundation of China under U1936217, 61971267, 61972223, 61941117, 61861136003, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology. The associate editor coordinating the review of this article and approving it for publication was M. Conti. *(Corresponding author: Yong Li.)*

Huandong Wang, Chen Gao, Yong Li, and Depeng Jin are with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn).

Zhi-Li Zhang is with the Department of Computer Science, University of Minnesota, Minneapolis, MN 55416 USA.

Digital Object Identifier 10.1109/TNSM.2020.3013335

## I. Introduction

SMART phones and other mobile devices have made it easy for users to access various online services nearly everywhere and at any time – literally with a few touches of fingertip – whether on the go, at home, school or work. Online services such as social networks, messaging apps or e-commerce sites typically require users to create online user identifiers (IDs) to login and access their services. Due to the stateless nature of the HTTP protocol, it is well known that HTTP requests and responses often contain cookies as part of the HTTP headers that embed user online ID information. This is despite the fact that the HTTP payload itself may be encrypted. Hence in this sense, users leave a variety of "fingerprints" in the cyber world. Previous studies have shown a wide range of highly sensitive personal attributes and information such as age, gender, photos, friends, sexual orientation, ethnicity, religious and political views, hobbies, activities, even emotions, can be culled from online social network profiles and activities [1], [2], and correlated and inferred – especially coupled with network traffic – to build a mosaic of various personal traits and activities [3].

As they roam around in the physical world while accessing online services, users also leave a series of "footprints" – i.e., hints about their physical locations – in the physical world. This poses a potent new threat to user privacy – *leakage of user physical world privacy*: *one can potentially correlate the "fingerprints" left by the users in the cyberspace with "footprints" left in the physical world to infer and reveal information about users in physical world, such as frequent user locations or mobility trajectories in the physical world!* To demonstrate that this problem of *user physical world privacy leakage via user cyberspace privacy leakage* is plausible, we make the *weakest* assumption about the (physical) location information: we simply assume that we have access to a (diverse) collection of deep packet inspection (DPI) data of a number of broadband subscribers, each of which is associated with a physical location (of certain geographical resolution), e.g., a WiFi access point, or a broadband interface; *but* we do *not* have information regarding the nature of the physical location (e.g., whether they are residential, business or downtown commercial districts), not to mention the GPS location coordinates. We further assume that each network data record (e.g., an HTTP session) are time-stamped, and the collection of network datasets has a large *geographical* span as well as *temporal* span that cover the mobility and other physical activities of a significant portion of users. Users can employ multiple and different online user IDs to access online services – these

are the so-called cyberspace *fingerprints*; here we assume that we do *not* know the true user identity – e.g., the mobile phone number or the device ID – behind these online user IDs. The fundamental question we are interested in answering is: what kind – and how much – of user *physical world* information might be leaked if we could get hold of such diverse network datasets *even without any physical location information*; or is this concern merely hypothetical?

Given the above assumptions, we decompose this question of *user physical world privacy leakage via user cyberspace privacy leakage* into three sub-problems: i) Is it possible to link various online service IDs belonging to the same user together, using only mobility patterns of users across multiple locations over time, but without the precise location information? Here a key insight is that users' daily mobility patterns are fairly predictable, e.g., two frequent locations are home and work/school, as previous studies have shown [4]. Hence the question becomes whether such predictability can be exploited to link together various user online IDs. ii) Assuming that we could link together various online user IDs of a significant portion of users, could we then use such information to classify the physical locations that users are associated with, such as residential, business, entertainment, etc.? Here the intuition is that the time, frequency and duration that various users visit a location can reveal the nature of a location or provide other contextual information about a location. For example, very few people tend to visit shopping malls in late night, because they are usually closed before 9 pm; whereas a location that is associated with many people throughout the evening and night would likely be a residential place. Lastly, iii) with answers to i) and ii), we would like to develop an effective method to piece together and track users' physical world trajectories and activities.

We remark that many of today's Internet service providers (ISPs) collect and store various sources of network traffic data for legitimate business reasons (e.g., for service billing, network management, traffic engineering and performance monitoring). It therefore is possible that stored network traffic data might be hacked and stolen, despite the fact that location information might have been encrypted, anonymized or removed. This is not notwithstanding that a powerful third party, e.g., a crime syndicate, a rogue employee of an Internet operator or a state agent of an authoritarian government, or any other "big brother" entities, could possibly directly tap into the wire or force an ISP to surrender (e.g., via subpoena) to get access to such data. In this case, the physical locations might even be available to the third party, yielding a simpler version of the problem that what we try to address in this paper; in other words, the sub-problem ii) becomes trivial, when exact location information is available. However, this is a very limited situation. In more cases, we should consider the sub-problem ii). From the perspective of network measurement, the problem we attempt in this paper is also highly relevant: an affirmative answer to the fundamental question posed above suggests that merely encrypted or anonymized user id (e.g., phone numbers) and location information (e.g., GPS coordinates) associated with network datasets is insufficient – not only user cyberspace privacy but also the user physical world privacy could be mined and inferred, thus leaked.

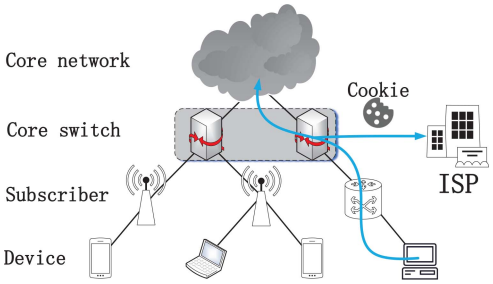



Fig. 1. Framework for extracting cookies from packets.

In order to conduct an in-depth investigation of these questions, we utilize the network traffic data collected via a DPI system at the routers within one of the largest Internet operators in Shanghai, China over a duration of one month. Only cookies in the HTTP header traffic which contain users' online service IDs during the online login process are collected and used in our study (see Section II for more details)– no payload or other personally identifiable information is collected. For scalability, we also limit ourselves with only the user online IDs of four *popular online services in China, namely, QQ (online instant messenger), Weibo (online social network), Tmall and Taobao (online shopping sites), and Dianping (online review site)*. A total of 470 million records containing 32.0 million distinct user IDs and spanning nearly the whole city are used in our study. The contributions of our study are summarized below:

- We develop a user detection system to discover users' identifiers in multiple cyberspace by utilizing the spatio-temporal locality. By checking against the ground-truth data, we validate that our algorithm achieves high accuracy with performance gain over 0.1 compared with baselines.
- We develop a location-classification system that is able to divide millions of locations into three types: residential, business and entertainment. Our results achieve F1-score of 0.78 and highly coincide with the POI distribution, indicating the effectivity of our system.
- We systematically analyze the obtained all-round mobility trajectories with physical context of over 10 million users, and reveal their main privacy leakage in terms of time, locations and services.

A conference version of this paper was published in [5]. Compared with the conference version, we propose a new probabilistic mobility model for identity linkage without requiring location information, which better satisfies our system model. In addition, additional important baselines, experimental analysis, and discussion about our proposed algorithms are supplemented in this version.

## II. Data Collection and Processing

The datasets used in our study were collected from the core routers of a major Internet service provider (ISP) in Shanghai, China. They are obtained through the two processes: extracting the cookie data from the DPI system, and culling user IDs from the cookies.

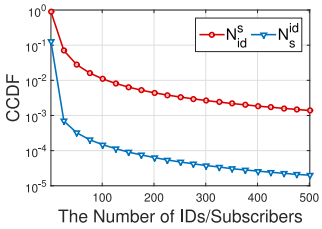
```
GET /<User ID> HTTP/1.1
Host: user.qzone.qq.com
Cookie: pt2gguin=o<User ID>||_qz_referrer=
www.baidu.com||qz_screen=1920x1080
```

```
GET / HTTP/1.1
Host: www.taobao.com
Scheme: https
Cookie: thw=cn||cna=<User ID>
```

Fig. 2. A Cookie of different online services for example.

TABLE I
DATASET SUMMARY

| # Records | # Online IDs | # Subscribers | Duration |
|---|---|---|---|
| 470 million | 32.7 million | 3.4 million | One month |




| Items | Mean | Standard Deviation |
|---|---|---|
| $N^s_{\mathrm{id}}$ | 12.70 | 122.28 |
| $N^{\mathrm{id}}_{\mathrm{s}}$ | 1.31 | 41.68 |

Fig. 3. The relations between online IDs and subscribers, where $N^s_{\mathrm{id}}$ denotes #IDs appeared in one subscriber, $N^{\mathrm{id}}_{\mathrm{s}}$ denotes #subscribers that one online ID accessed.

The diagram of the first process is shown in Fig. 1. This system records users' Internet accessing activities via broadband subscribers across the city. Each subscriber is associated to a physical locations, e.g., a broadband Internet connection at a user' home or a WiFi access point. By deploying network monitoring and packet capture tools on the core routers of the ISP, we extract cookies generated by users. In addition, the ISP maintains a separate (billing) database, namely, the identification of each broadband subscriber, which can tell us in which broadband subscriber each HTTP session is generated. However, we still do not know the GPS coordinate of each broadband subscriber. Thus, we believe this dataset provides a meaningful representation of users' cyberspace Internet accessing behavior. Combining these two data sources, we can correlate cookies extracted from the data packets with the corresponding broadband subscribers. One issue is that cookies in HTTPS session cannot be obtained. It is found that only 14% of top websites in China use HTTPS in 2015 [6], which indicates the overwhelming majority of users' cookies can be extracted from HTTP packets. This number has increased to 60% in 2019 [6]. However, we observe that these 60% websites still use HTTP for parts of the Internet requests, while they use HTTPS only to transmit sensitive information. Overall, the privacy leakage through HTTP protocol is still a considerable problem. In addition, for services based on HTTP protocol, all the users' cookies can be extracted.

The second step is to cull user IDs from the cookies. It has been discovered that although cookies are often opaque strings with hidden semantics known only to the party setting the cookie, they may include visible identity information [7]. Inspired by this idea, we turn on the Chrome Developer Tools [8] and display HTTP request/response headers containing cookies we need. Take the first request header shown in Fig. 2 as an example. This request header is generated when we want to login their accounts through the website user.qzone.qq.com. Fields such as Accept and User-Agent have been ignored for simplicity. As we can observe, the ID of the user is involved in the Path field, i.e., "/<User ID>" in this request. On the other hand, in the cookie, there are many visible information we can obtain. For example, "qz_screen=1920x1080" indicates the resolution of the user's browser, and "pt2gguin=o<User ID>" contains the ID of the user. Thus, we can extract the ID for QQ of the user by matching the field "pt2gguin=" in cookies of QQ. The second request header contains cookie for the Taobao service. Similarly, we can find the ID of the user in the cookie field. Overall, by performing regular expressions matching to these cookies, we obtain account IDs of users for different online services. In more detail, we list the regular expressions used to extract online IDs and related examples in Table II. As mentioned in the introduction, our study focuses on four representative online services in China, i.e., QQ (online instant messenger), Weibo (online social network), Taobao (online shopping site), and Dianping (online review site). All of them are the leading and most popular ones among the corresponding categories in China.

By sniffing the traffic of millions of broadband subscribers, we capture the login actions when users access these services. The data collection was from Nov. 1 to Nov. 30, 2015, involving over 3.4 million broadband subscribers and 32.0 million online IDs. There are 470 million entries in our dataset. Each entry contains following fields: name of the online service, online ID, identity of the broadband subscriber, and login time. Take <Weibo, 123456, 789, 2015112113> as an example. It records a user with ID 123456 logs in Weibo at 1PM Nov. 21, 2015, and the identity of the subscriber is 789. Table I presents a summary of the dataset. The large-scale datasets guarantee the credibility of our analyses of *user physical world privacy leakage via user cyberspace privacy leakage*.

We now provide an informative overview of the data set. We are interested in the following four metrics, i.e., the number of online IDs that appear in one subscriber, the number of subscribers accessed by one online ID. According to Fig. 3, we can observe that there are in average 12.70 online IDs that appear in a single subscriber, and a single online ID accesses 1.31 subscribers in average.

To preserve user privacy, the online ID and subscriber identity in our datasets and aforementioned cookies are all anonymized. The real online ID and subscriber identity are never made available to, or utilized by us. In addition, there is also no payload collected in the dataset. The usage of the datasets is authorized by the ISP.

## III. THREAT MODEL

### A. Attack Scenarios

Though the datasets used in our study are collected by the ISP, the ISP is not the only possible attacker, since the

TABLE II
SERVICE TYPES AND REGULAR EXPRESSIONS TO MATCH IDS

| Services | Type | Website | RegExp | Example | Quantity |
|---|---|---|---|---|---|
| QQ | Instant messengers (IM) | qq.com | pt2gguin=o(\d+)<br>o_cookie=(\d+) | pt2gguin=o<User ID> | 11M |
| Taobao | E-commerce (EC) | tmall.com<br>taobao.com | lgc=(\w+)<br>cna=(\w+) | lgc=<User ID> | 15M |
| Weibo | Online social networks (OSN) | weibo.com | SID-(\w+)<br>fid%(\w+) | SUS=SID-<User ID>-1446074239-JA<br>-j90ae-cf43c6fa7a7ecf6fa3460f3f3611b724 | 2M |
| Dianping | Online review (OR) | dianping.com | ([^0-9]{1}[1][^1)[3-8]<br>{1}\d{9}($\|[^0-9]1) | pvsid=<br><User ID> | 4M |

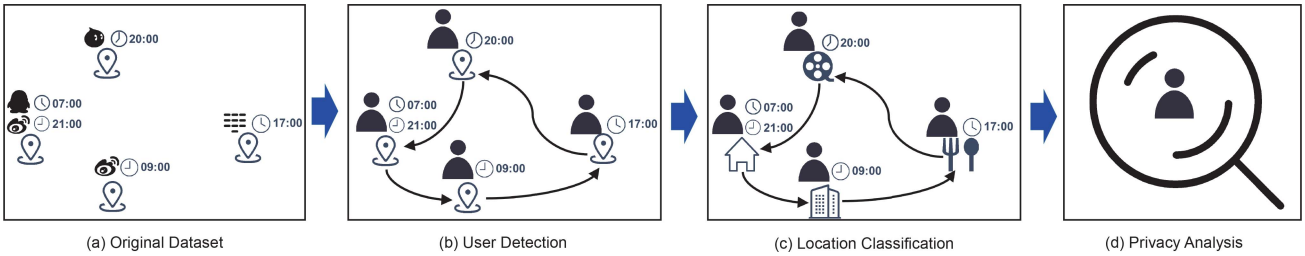



Fig. 4. Framework of the attacker model.

cyberspace *fingerprints* of users is a ubiquitous content and can be easily obtained by different attackers. Another strong attacker is the company-level attacker, e.g., a service provider, who can infer users' locations through packets uploaded by the applications installed in their mobile phones. For example, a user may have two IDs of two services provided by one company. The company can link the two IDs based on their mobility patterns and obtain users' behavior in more detail, which leads to more privacy leakage of users. Compared with the ISP, the company-level attackers are only limited in the number of services of which the user IDs they can obtain, which are still a big potential threat to user privacy. An individual attacker can also infer users' locations through crawling their publicly available online check-ins from multiple location-based services. However, this kind of user trajectories are sparse and noisy [9]. Thus, individual attackers are the weakest attackers. Overall, the datasets collected by ISPs give us the most comprehensive view of this kind of privacy leakage. For example, the threat of company-level attackers can be analyzed by only considering a part of services in the ISP dataset. Thus, we mainly focus on the ISP dataset in our study.

### B. Attacker Model

As described in Section II, by culling user online IDs from the cookies, attackers obtain the login records of massive online IDs with the corresponding time and locations as shown in Fig. 4(a). The goal of the attacker is to obtain privacy information of physical users based on them. However, it is not a trivial task in terms of three major challenges. First, users' multiple identifiers are extracted without cross linking, while it is quite normal for a physical user to have multiple IDs of different services. Second, attackers have known few background or context about physical locations, which are critically important for inferring the privacy leakage of the physical world mobility. Last but not the least, how can they infer the privacy leakage from the physical world footprint is the third challenge. To meet these challenges, we design the attacker model with three modules as shown in Fig. 4, which are discussed as follows.

In order to uncover physical world privacy leakage as much as possible from cyberspace cookie records of users, a basic question must be answered, namely which online IDs belong to the same users? In physical world, it is quite normal for an individual user to have multiple IDs for different online services. The trajectory of a single online ID is only a subset of mobility records left by the corresponding physical user. Thus, in order to obtain the bound of users' privacy leakage from their login records, attackers must first link all the online IDs for each user together to obtain the universal set of its mobility records. Thus, as shown in Fig. 4(b), the first module of the system is to link all the online IDs belonging to each user, which is discussed in detail in Section IV.

On the other hand, physical context can provide rich information about users' behavior. For example, given the physical context, attackers can infer what people are doing through where users are located in physical world. Further, attackers can infer what people are going to do through where users are moving to. Through these behaviors, more privacy information of users is exposed. Meanwhile, locations of places such as home and office are also sensitive information for users. Thus, as shown in Fig. 4(c), attackers investigate the physical context of each location in the second module, and this module will be discussed in detail in Section V.

Finally, by combining the all-round login records and physical context, attackers can thoroughly analyze users' privacy information as shown in Fig. 4(d). Then, we will formally define different types of attack and users' privacy leakage in the following section.

### C. Privacy Metrics

We consider users' physical world privacy leakage in terms of three kinds of attack in this work.

The first kind of attack is to find out an activity or an event, which is described by a spatio-temporal point, in users' trajectories. We refer to this kind of attack as localization attack. The privacy leakage in terms of localization attack can be quantified by the quality of the obtained trajectories, including their density, spatial and temporal resolutions, etc.

The second kind of attack is the semantic attack. In semantic attack, adversaries aim to acquire individual's behavior or motivation at some spatio-temporal points by using location semantic information. For example, whether users visit locations of hospital is very sensitive for users. We introduce the coverage rate of locations with different semantics for all trajectories as the privacy metric in terms of semantic attack. Specifically, the coverage rate is defined as the percentage of users of which the trajectories have at least one residential, business or entertainment location, respectively. Higher coverage rate indicates larger semantic information of users' trajectories are leaked.

The third kind of attack is the re-identification attack. In the re-identification attack, adversaries seek to identify individuals in this dataset based on external information (e.g., users' check-ins on online social networks), or use this dataset as the external information to identify users in other datasets. We use uniqueness [10] as the metric of privacy leakage in terms of re-identification attack. Specifically, it is to estimate the number of points necessary to uniquely identify the mobility trace of an individual. If the uniqueness of trajectories is high, the dataset has a high risk of being re-identifiable using external datasets, and it also has a strong ability to identify users in other external datasets [10].

These three kinds of attack cover main physical world privacy leakage by cyberspace cookie logs. We will evaluate users' privacy leakage in terms of them in detail in Section VI.

## IV. Physical World User Detection

In our system, we *do not* know the identifier of physical user – e.g., the device ID – behind these online user IDs. We only know the identifier of subscribers, which correspond to different physical locations. Thus, the extracted users' multiple online IDs are not linked together, while it is quite normal for a physical user to have multiple IDs. In order to characterize users' privacy leakage in a comprehensive way, users' multiple IDs should be linked together to be analyzed. On the other hand, users' daily mobility patterns have been discovered to be fairly predictable [4]. Inspired by this idea, we propose an algorithm which maximizes the likelihood of observed records of online IDs based on Markov human mobility model to solve this problem.

### A. Model and Problem Formulation

We first propose a mathematical model and formulation for the problem. Let $\boldsymbol{A}$ represent the set of all online IDs in our dataset, and let $2^{\boldsymbol{A}}$ denote the power set of $\boldsymbol{A}$, i.e., the set of all subsets of $\boldsymbol{A}$. Given any online ID $u \in \boldsymbol{A}$, we define its mobility records as $R^u = \{(l_1, t_1), (l_2, t_2), \ldots, (l_{N_u}, t_{N_u})\}$, where $(l_i, t_i)$ represents a login record in location $l_i$ at time slot $t_i$ that was traced by cookie, and $N_u$ is the number of records for ID $u$. Without loss of generality, we further assume $t_1 < t_2 < \cdots < t_{N_u}$. In addition, we denote $\boldsymbol{L}$ as the set of locations. Moreover, for a cluster of online IDs $U$, we define their mobility records $R^U = \{R^u | u \in U\}$. Then $R^{\boldsymbol{A}} = \{R^u | u \in \boldsymbol{A}\}$ represents the set of mobility records of all online IDs. Let $s(u) \in \boldsymbol{S}$ denote the type of online ID $u \in \boldsymbol{A}$, where $\boldsymbol{S} = \{\mathrm{IM}, \mathrm{OSN}, \mathrm{EC}, \mathrm{OR}\}$ is set of ID types.

*Definition 1 (Partition of $\boldsymbol{A}$):* Let $p = \{U_1, U_2, \ldots, U_n\}$, where $\forall k = 1, \ldots, n,\ U_k \in 2^{\boldsymbol{A}}$. We further define $p$ as a partition of $\boldsymbol{A}$ if following four conditions hold: **(1)** $\emptyset \notin p$, **(2)** $\cup_{U \in p} U = \boldsymbol{A}$, **(3)** if $U_1,\ U_2 \in p$, and $U_1 \neq U_2$, then $U_1 \cap U_2 = \emptyset$, and **(4)** $\forall U \in p$, if $u_1,\ u_2 \in U$, then $s(u_1) \neq s(u_2)$.

Traditional definition of partition only needs three conditions (1)$\sim$(3) hold. In our case, we use the definition of partition to represent a user detection result, where each set $U$ in partition $p$ represents all online IDs belonging to one physical user. Thus, we add the condition **(4)** to limit that each set in a partition has at most one ID of each type. In addition, we define $\mathcal{P}$ as the set of all partitions.

Assume there is no shared online ID among different users. Then, there is an inherent partition of $\boldsymbol{A}$ composed of the true set of online IDs for each user denoted as $p_{\mathrm{true}}$. Our problem, i.e., detecting all online IDs of each user, can be transformed to finding a partition $p$ for $\boldsymbol{A}$ that is close to $p_{\mathrm{true}}$ as much as possible. However, in most cases, we only need to detect online IDs for a part of users, or even one user. Thus, by using the target online IDs as the identifications of our target "users", the problem is transformed to: for a list of online IDs $\{u_i\}_{i=1}^k$, detecting all other online IDs belonging to the same user with them. That is, find a partition $p$, where elements involving $\{u_i\}_{i=1}^k$ approaches to elements involving $\{u_i\}_{i=1}^k$ in $p_{\mathrm{true}}$ as close as possible.

In order to formally analyze our problem, it is necessary to build a mobility model which describes how users move and produce login records. Human mobility modelling has been studied in a number of works [11], [12]. Brockmann *et al.* [11] model human mobility as a Lévy flight, where the length of spatial displacement of individuals follows power-law distribution. Cho *et al.* [12] model human movements as periodic movement between only two latent states, which represent users' home and work place, respectively. Further, it assumes space coordinates together with timestamps follow a three-dimensional Gaussian mixture model with two components corresponding to the two latent states. However, these mobility models require accurate GPS coordinates of locations, which is not available in our scenario.

In the conference version of this paper [5], we model users' spatial displacement to follow Gaussian distribution, which still need GPS coordinates of locations. Thus, we relax the constraint of unknown physical location information, and utilize the GPS coordinates of locations provided by the ISP. Instead, in this journal version, we assume the movement of users follows Markov model, which is widely used to model human behavior, and do not need GPS coordinates of locations. Specifically, it models the movements of human as transitions among definite and countable states, and each state corresponds to a location. For arbitrary user (online ID) $u \in \boldsymbol{A}$, there exists a unique transition matrix $T_u$ of

size $|\boldsymbol{L}| \times |\boldsymbol{L}|$, where $|\boldsymbol{L}|$ is the total number of locations. Then, location of each mobility record is assumed to be only dependent on the last mobility record. Denote $T_u(l,k)$ as the probability that user *u* move from location *l* to location *k* in adjacent records. Then, the conditional probability of a mobility record given its time-adjacent record can be calculated as follows,

$$p((l_{i+1}, t_{i+1})|(l_i, t_i), T_u) = T_u(l_i, l_{i+1}).$$

Then, the probability of all the mobility records of user *u*, i.e., $R^u = \{(l_1^u, t_1^u), (l_2^u, t_2^u), \ldots, (l_{N_u}^u, t_{N_u}^u)\}$ can be computed as:

$$p(R^u|T_u) = \prod_{i=1}^{N_u-1} T_u\big(l_i^u, l_{i+1}^u\big) = \prod_{l,k\in\boldsymbol{L}} T_u^{c_{lk}^u}(l,k), \quad (1)$$

where $c_{lk}^u$ is the observed transition counts between location *l* and *k* of user *u*. It can be calculated by

$$c_{lk}^u = \sum_{i=1}^{N_u-1} I(l_i^u = l) I\big(l_{i+1}^u = k\big), \quad (2)$$

where $I(\cdot)$ is defined to be an indicator function of the logical expression with *I*(*true*) = 1 and *I*(*false*) = 0.

On the other hand, considering that transition matrix $T_u$ is usually unknown, we adopt Bayesian probability methods, where $T_u(l,k)$ for each location *l* and *k* is regarded as a random variable. Further, by denoting $T_u(l,:)$ as the $|\boldsymbol{L}|$-sized vector $[T_u(l,1), T_u(l,2), \ldots, T_u(l,|\boldsymbol{L}|)]^T$, we use the common conjugate prior distribution to model $T_u$ [13], [14], which can be expressed as follow:

$$T_u(l,:) \sim Dirichlet(\cdot|\beta_0). \quad (3)$$

where $Dirichlet(\cdot|\beta_0)$ is Dirichlet distribution with hyperparameter $\beta_0$ [15]. Specifically, $\beta_0$ is used to describe how much we believe this prior. By following the recommended setting in [15], we set $\beta_0 = 0.1$. Based on this prior distribution, the probability density function (PDF) of $T_u$ can be calculated as follows:

$$p(T_u) = \prod_{l\in\boldsymbol{L}} Dirichlet(T_u(l,:)|\beta_0). \quad (4)$$

Finally, the unconditional probability of observing mobility records $R^u$ is computed by integrating $T_u$ as follow:

$$P(R^u) = \int p(R^u|T_u)p(T_u)dT_u = \lambda(\beta_0)\prod_{l\in L} B(b_l), \quad (5)$$

where $B(\cdot)$ is the multivariate Beta function, and $\lambda(\beta_0)$ is a function of $\beta_0$ and independent with $R^u$. $b_l$ is an $|\boldsymbol{L}|$–sized vector with $b_l(k) = c_{lk} + \beta_0$ for each $k \in \boldsymbol{L}$.

Different with Gaussian distribution used in [5], the mobility model based on (4), (5) considers personalized mobility patterns of each user. For example, in the viewpoint of aggregated user mobility, there are actually less users that move a large distance in a short time. However, for a certain user, if we observe that he travel between two distant locations in a short time every day, this is a personalized mobility pattern of him. Thus, his corresponding transition probability should be large. A better solution to consider the dependence of distance and the personalized mobility pattern of users simultaneously is to use a better prior probability to the transition matrix which considers the dependence of distance. However, it requires the background or context of physical location, which is in conflict with our model. Thus, we do not consider the dependence to the distance in (4).

| ID | Trajectories | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $u_1$ | $l_A$ | $l_A$ | $l_B$ | $l_A$ | $l_A$ | $l_B$ | $l_B$ | $l_A$ | $l_B$ | $l_A$ | |
| $u_2$ | $l_A$ | $l_A$ | $l_A$ | $l_A$ | $l_A$ | $l_A$ | $l_A$ | $l_A$ | $l_B$ | $l_B$ | $l_B$ |
| $u_3$ | $l_A$ | $l_B$ | $l_B$ | $l_A$ | $l_A$ | $l_A$ | $l_B$ | $l_A$ | $l_A$ | $l_B$ | $l_B$ |

(a) Trajectories of IDs

| $u_1$ | |
|---|---|
| 0.40 | 0.60 |
| 0.60 | 0.40 |

| $u_2$ | |
|---|---|
| 0.87 | 0.13 |
| 0.05 | 0.95 |

| $u_3$ | |
|---|---|
| 0.50 | 0.50 |
| 0.50 | 0.50 |

(b) Posterior expectation of transition matrixes, where first column and row correspond $l_A$ and second column and row correspond $l_B$.

| Candidate IDs | $\Delta q$ |
|---|---|
| $u_1, u_2$ | -0.73 |
| $u_1, u_3$ | 3.71 |

(c) Change of log-likelihood, where prior $P(p)$ is ignored.

Fig. 5. Three example IDs to be matched.

For a cluster of ID *U*, conditioned on that all IDs in *U* belong to the same user, we also use (5) to compute the probability of observing their mobility records $R^U$.

Intuitively, according to the property of multivariate Beta function, merged mobility records of IDs with similar distribution of transition counts have larger $P(R^U)$. Because different users have different mobility patterns, the distribution of transition counts of their IDs are far from each other, leading to a small probability $P(R^U)$ when putting their IDs in the same set *U*. Then, based on this model, we can make a Bayesian inference about the relationship of IDs.

### *B. Detection Method*

The true partition $p_{\text{true}}$ can be approximated by the partition $p$ that maximizes the posterior probability of

$$\widehat{p} = \underset{p\in\mathcal{P}}{\operatorname{argmax}}\, P(p|R^{\boldsymbol{A}}). \quad (6)$$

By applying Bayes' theorem to it, we can obtain:

$$P\Big(p|R^{\boldsymbol{A}}\Big) = \frac{P\big(R^{\boldsymbol{A}}|p\big)P(p)}{P\big(R^{\boldsymbol{A}}\big)}. \quad (7)$$

In terms of $P(R^{\boldsymbol{A}}|p)$, we assume the login records are produced independently by different users. Thus we have

$$P\Big(R^{\boldsymbol{A}}|p\Big) = \prod_{U\in p} P\Big(R^U|U\Big), \quad (8)$$

where $P(R^U|U)$ is the probability that the mobility records in $R^U$ occur under the condition that they belong to the same user. Under the proposed mobility model, this probability can be computed by applying (5) to the merged mobility records of the user, $R = \bigcup_{a\in U} R^a$. In addition, we further assume that prior $P(p)$ is only dependent on the online IDs of each user, i.e., $P(p) = \prod_{U\in p} P(U)$. Then, we have:

$$P\Big(p|R^{\boldsymbol{A}}\Big) \propto \prod_{U\in p} P\Big(R^U|U\Big)P(U). \quad (9)$$

We further assume users own each type of online ID independently with Bernoulli distribution with probability $\theta_t$ for $t \in \boldsymbol{T}$. That is, $P(U) = \prod_{t\in T} \theta_t^{I_t(U)}(1-\theta_t)^{(1-I_t(U))}$, where $I_t(U)$ is the indicator function of whether $U$ contains online ID of type $t$.

However, the computation time of finding a partition of the over 30 million online IDs over a modern metropolis as Shanghai that maximizes the posterior probability is intolerable for us. Since it is NP-hard [16], even finding an optimal partition for a subset of $\boldsymbol{A}$ is intractable. Thus, we alternatively compute the set of online IDs $U$ which maximizes the local likelihood of the target online ID in a greedy way, which is described in detail in Algorithm 1. It starts from the target online ID as the initial node of the target cluster $C$. Then, it works in an iterative way to discover prospective nodes belonging to the same physical user. In each iteration round, among all nodes in $\boldsymbol{A}$, it picks the node with the maximum increase to the local log likelihood by involving it to the current cluster $C$, in which the local log likelihood is defined as:

$$q(C) = \log P\Big(R^C|C\Big)P(C). \tag{10}$$

Then, the change of the local log likelihood by involving ID $a$, denoted by $\Delta q(C, a)$, can be calculated as follows,

$$\Delta q(C, a) = q(C \cup a) - q(C) - q(a). \tag{11}$$

Since $\Delta q$ is compared with 0 in Algorithm 1, by changing parameters in the prior distribution $P(p)$, we can adjust the threshold of the stop condition of the algorithm.

After that, the algorithm updates the target cluster by adding the picked node, and continues to select another node. This process is repeated until no increase of the local log likelihood can occur, and $C$ is outputted as the detected result.

Let $n$ and $k$ denote the total number of IDs and service types, respectively. Then, for each user, we need to find its $k$ IDs at most among $n$ IDs. It takes $O(n\log n)$ time to find the ID with largest $\Delta q$ each time. Thus, the total computational time is $O(kn\log n)$. Further, if we restrict the potential IDs of the target user within those that have been appeared in the same locations with the target ID. Denote $\tau$ as the maximum number of IDs appeared in the same locations. The computational complexity of the greed algorithm is only $O(k\tau\log\tau)$. Thus, it significantly reduces the computational time.

To illustrate the model, let us consider three example IDs with mobility records shown in Fig. 5. For simplicity, there are only two locations $l_A$ and $l_B$ in this example. Their trajectories are shown in Fig. 5(a), where each column represents a time bin. Note that elements of different trajectories in the same column do not represent the same time bin. Posterior expectation of their transition matrix estimated based on $P(T^u|R^u)$ is shown in Fig. 5(b). Finally, the change of log likelihood by matching them is shown in Fig. 5(c), where prior $P(p)$ is ignored. We can observe that $u_1$ and $u_3$ tend to switch more frequently between $l_A$ and $l_B$, and their transition matrixes are also more similar compared with $u_2$, and we have $\Delta q(u_1, u_3) > \Delta q(u_1, u_2)$. Thus, we tend to match $u_1$ and $u_3$ as IDs belonging to the same user.

**Algorithm 1:** MLink Algorithm

**Input:** Network $G = (V, E)$, the set of types of online IDs $\boldsymbol{S}$, the type $s(v) \in \boldsymbol{S}$ for all online ID $v \in \boldsymbol{A}$, and an initial online ID $u_0$.
**Output:** $C$, the cluster of online IDs belonging to the owner of $u_0$.
**Initialize:**
  $C \leftarrow \{u_0\}$;
  $S_C \leftarrow \boldsymbol{S}$;
**while** $\max_{u\in\boldsymbol{A}\setminus C} \Delta q(C, u) > 0$ *and* $S_C \neq \emptyset$ **do**
  $u_{\max} = \text{argmax}_{u\in\boldsymbol{A}\setminus C}\Delta q(C, u)$;
  $C = C \cup \{u_{\max}\}$;
  $S_C = S_C\setminus\{s(u_{\max})\}$.

### C. Performance Evaluation

*1) Compared Algorithms:* Since our proposed identity linkage algorithm is based on Markov model, we denote it as MLink. We compare our proposed MLink algorithm with other three state-of-the-art approaches, which are described as follows:

*POIS:* Riederer *et al.* [17] assume visit of each user to a place follows the Poisson distribution, and an action (e.g., login) of each service occurs independently with the Bernoulli distribution. Based on this mobility model, they compute a score for every candidate pair of online IDs. They find the maximum weighted matching of online IDs as the results. In addition, it filters out IDs by the "eccentricity" factor $\epsilon$ , which is defined as the threshold for the weight gap between the best and second-best IDs.

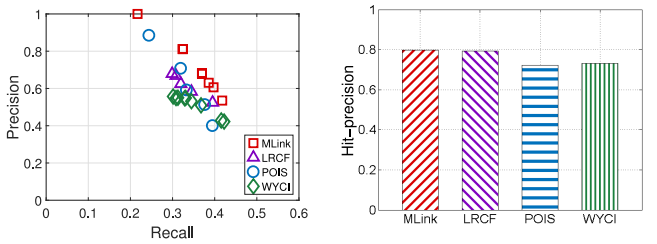
*WYCI:* Rossi and Musolesi [18] use the frequency of login in different locations to approximate the probability of visits to these locations, which is represented as: $P(l|R^U) = \frac{N_l^U+\alpha}{\sum_{l\in L} N_l^U+\alpha|L|}$. $N_l^U$ is the number of login records of user (online ID) $U$ at location $l$, i.e., the number of elements in $\{t|(l,t) \in R^U\}$. In addition, $\alpha > 0$ is the smoothing parameter and $|L|$ is the number of locations in the dataset, which are used to eliminate zero probabilities. Their target is to find online ID $a$ maximizing the probability $\prod_{(l,t)\in R^a} P(l|R^U)$.

*LRCF:* Goga *et al.* [19] further consider the popularity of different regions. Specifically, they apply the *term frequency - inverse document frequency (TF-IDF)* [20] weighting scheme to the histograms, i.e., $\Lambda_u(l) = N_l^u/\log(\text{IDF}(l))$, where $\text{IDF}(r) = \sum_{u\in\boldsymbol{A}} N_l^u$ is the number of records in location $l$ of the whole dataset. Then, they measure the cosine similarity between $\Lambda_u$ and $\Lambda_v$ as follow:

$$\text{similarity}(u, v) = \Lambda_u^T\Lambda_v/\|\Lambda_u\|\|\Lambda_v\|.$$

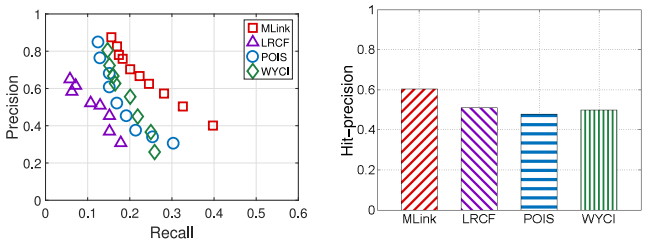
Users (online IDs) $u$ and $v$ with largest similarity are linked.

*2) Experiment Setting:* To evaluate the accuracy of our proposed user detection algorithm, we need some ground-truth data for the validation. By a questionnaire survey, ISP obtains all online IDs of 3000 users, where 65%, 29%, 58%, 68% of them have IM, OSN, EC, OR accounts, respectively. The results are encrypted with the same encryption function as the DPI data by the ISP, and thus they can be matched




(a) Precision and Recall (b) Hit-precision

Fig. 6. Performance of different solutions (IM vs. OR).




(a) Precision and Recall (b) Hit-precision

Fig. 7. Performance of different solutions (Twitter vs. Foursquare).

with each other. We use this data as the ground truth in our study. Note that the ground data is only used for evaluating the performance, implementing the identity linage algorithm does not require such background knowledge.

On the other hand, we also evaluate the performance of our proposed algorithm on public dataset collected in previous work by Zhang *et al.* [21]. On Foursquare, users may display their Twitter account information, which makes it possible to obtain the ground-truth mapping between Twitter IDs and Foursquare IDs. In total, this dataset contains 385 users with location check-ins on both sides (770 online IDs), and totally 24,556 location check-ins collected from both Twitter and Foursquare.

We use four key metrics in binary classification to quantify the detection accuracy, i.e., precision, recall, F1-score, and hit-precision [22]. Specifically in our problem, precision is defined as the fraction of online IDs detected by our algorithm that are included in the ground-truth data, and recall is defined as the fraction of online IDs in the ground-truth data that are successfully retrieved. F1-score is the harmonic mean of precision and recall, which is defined as $F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. Hit-precision is another widely-used metric to evaluate the performance of identity linkage in terms of the ranking. Specifically, the hit-precision of top-$k$ candidates is defined as follows,

$$h(x) = \begin{cases} \frac{k-(x-1)}{k}, & \text{if } k \geq x \geq 1, \\ 0, & \text{if } x > k. \end{cases} \tag{12}$$

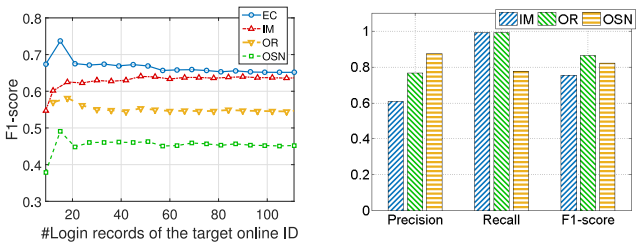
where $x$ is the rank of truly matched ID in the top-$k$ candidates. In addition, we set $k$ as 10 by default.

*3) Experiment Results:* We first evaluate the performance of our proposed algorithm on the ISP dataset. Specifically, we focus on the performance of linking the IM account and OR account belonging to the same user, and show the performance in Fig. 6. Fig. 6(a) shows the precision-recall trade-off by adjusting parameters (e.g., $\theta$ in MLink, $\epsilon$ in POIS). From the result, we can observe that our proposed algorithm outperforms other algorithms, since it lies in the top right corner, meaning it has larger precision when recall is equal and larger recall when precision is equal. In addition, the largest performance gap of our algorithm achieves 0.13 in terms of precision when recall is about 0.37. The main reason is that our proposed method better models users' transition patterns between locations based on Bayesian probabilistic model, while users' transition patterns between locations are ignored in the competing algorithms. Then, Fig. 6(b) shows the hit-precision of different algorithms on ISP dataset, we can observe that our proposed algorithm also has the best performance.

In Fig. 7, we conduct the same experiments using the Twitter-Foursquare dataset, whose quality is found to be worse than the ISP data – check-in data is much sparser. The precision-recall curve and hit-precision become lower than that of Fig. 6, but the trend is still consistent: our algorithm outperforms the baselines. In addition, the performance gain of our algorithm becomes larger. For example, precision is increased by over 0.20 when recall is 0.30, and hit-precision is increased by 0.10, indicating that our proposed algorithm is more robust to data quality. The underlying reason is that when trajectory data is sparse, the prior distribution used in our method will help to model users' mobility, leading to a better performance.

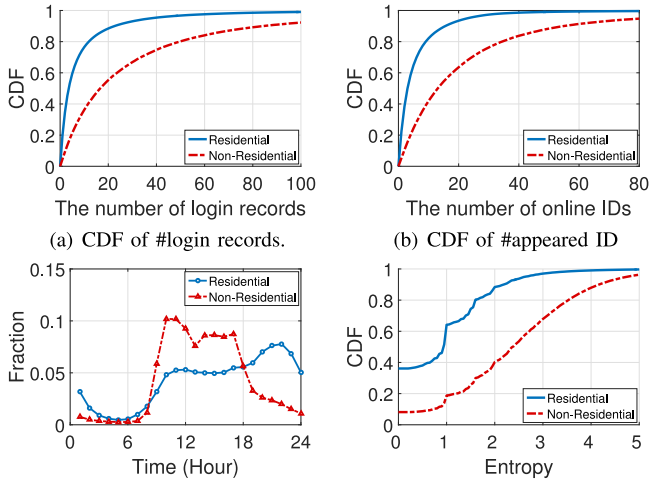
In Fig. 8, we evaluate our algorithm with different types of online IDs. Fig. 8(a) shows the performance with different types of initial IDs, in which the F1-score is evaluated as the function of the number of login records of the target IDs. As we can observe, user detection with initial IDs of EC shows better performance with stable F1-score above 0.65, while that of OSN shows the worse performance, with stable F1-score of 0.45. One possible reason of this result is the imbalanced number of different types of IDs as shown in Table II. For example, when using OSN accounts as the initial IDs, there are more candidate IDs of other services, leading to the worst performance. From another point of view, we fix the initial IDs to be EC accounts, and study the performance of the detection to other types of online IDs. As we can observe from Fig. 8(b), detection to all types of online IDs shows F1-score larger than 0.7. Another interesting observation is that performance of OSN accounts, in turn, shows relative better performance. The main reason is users' different behavior for using different types of online IDs. For example, people tend to use their EC accounts in more private places such as home, while they tend to use their OSN accounts more in public places. Thus, transition matrixes estimated based on EC accounts are more likely to be biased with insufficient information of public places, while estimation based on trajectories of OSN accounts is more comprehensive, leading to the better performance.

Overall, in order to detect all online IDs belonging to the same user, we propose a user detection algorithm of optimizing the local likelihood under Markov human mobility. The proposed MLink algorithm does not require the accurate GPS coordinates or other background knowledge of physical




(a) Performance vs.#login records for different types of intial IDs. (b) Best performance with intial EC accounts.

Fig. 8. Performance of different types of IDs.




(a) CDF of #login records. (b) CDF of #appeared ID

(c) 24 Hours' Distribution. (d) CDF of location entropy.

Fig. 9. Distribution of features for residential and non-residential locations.

locations, which exactly solves the identity linkage problem in our scenario. However, such requirements also limit the ability of MLink algorithm in modeling the intention or utilizing other physical features such as velocity or diameter of users' movements [5], and it is also hard to model the spatial or temporal mistakes of different datasets without physical background [23], which are compromises have to be made in our problem scenario. Experimental results show that our solution achieves good performance. Specifically, its precision beats baselines by 0.13 on the ISP dataset, and its hit-precision beats baselines by 0.1 on the Twitter-Foursquare dataset.

## V. Location Classification

In order to characterize the physical-world behaviors of users, we need to infer the types of the locations that online IDs login. On the other hand, since in our system the subscriber identity is encrypted, and the associated location information is insufficient, we cannot obtain their types directly through API provided by map services. Thus, in this section we develop a location-classification system that labels locations with different types based on the aggregate user behavior.

### A. Distinguishing Residential and Non-Residential Locations

From the subscribers' registration information, we obtain 10,000 locations with type of residential or non-residential, which was obtained by the ISP when users subscribed the broadband service. We use the 10,000 locations as training set, which enables us to utilize supervised learning algorithm to distinguish residential or non-residential locations.

Using this information as training set does not have conflict with our assumption that adversaries do not have any physical location information, since we still do not know any information of the majority of locations (over 99.6%). In addition, it is easy for adversaries to manually find some residual and non-residual locations as training set to develop this modular. In the worst case, adversaries can still use unsupervised learning methods, which will be introduced in Section V-B.

Through detailedly investigating the character of residential and non-residential locations, we select four features of each location $L$, i.e., the number of online IDs $N_u(L)$, the number of login records $N_l(L)$, the location entropy $E(L)$, and the day-night login frequency ratio $R_d(L)$ to distinguish residential and non-residential locations, of which the distributions are shown in Fig. 9.

As we can observe from Fig. 9(a) and (b), non-residential locations have much more login records and appeared online IDs compared with residential locations, which coincides with our empirical knowledge. On the other hand, from Fig. 9(c), we can observe that the temporal distributions of login records in these two types of locations show much difference. Specifically, non-residential locations have more login records from 8AM to 6PM (working time), while residential locations have more login records from 6PM to 0AM (leisure time). Thus, we use the ratio of the login number in these two time periods as the feature. Another important feature is the entropy. For a location $L$, its entropy $E(L)$ can be calculated by $E(L) = -\sum_{u \in U_L} P_L(u) \log P_L(u)$, in which $U_L$ is the set of online IDs appeared in $L$, and $P_L(u)$ is the probability of any online ID appeared in $L$ is $u$. It has been found that places such as the university campus, shopping and dining districts, have high entropy, while residential areas have low entropy [24]. In fact, we plot the distribution of entropy of different types of locations in Fig. 9(d). As we can observe, non-residential locations have much larger entropy compared with residential locations, indicating its effectiveness in distinguishing two types of locations. Overall, all these features we select show much difference between two types of locations. Thus, we use them in our classifier to distinguish residential and non-residential locations.

For each type of online IDs, we can compute its value for $N_l(L)$, $R_d(L)$, $E(L)$ and $N_u(L)$. There are 4 types of online IDs in our dataset. Thus, we have $4 \times 4 = 16$ features in total. Using these features, we apply three mainstream supervised learning algorithms, i.e., logistic regression (LR), support vector machine (SVM), random forest (RF) [25]. Specifically, logistic regression is a generalized linear model. Compared with it, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [26]. Random forest is an ensemble learning method that operates by constructing and combining a multitude of weak learners with random subset of the features and training set. Due to the

TABLE III
PERFORMANCE FOR DISTINGUISHING RESIDENTIAL AND NON-RESIDENTIAL LOCATIONS

| Algorithm | Residential | | | Non-residential | | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Prec. | Recall | F1 |
| LR | 0.92 | 0.97 | 0.94 | 0.79 | 0.58 | 0.67 |
| SVM | 0.97 | 0.91 | 0.94 | 0.66 | 0.88 | 0.75 |
| RF | 0.95 | 0.95 | 0.95 | 0.78 | 0.78 | 0.78 |

randomness and ensemble, it can better deal with redundant features.

We make a 10-fold cross-validation on the ground truth data, and show the results in Table III. As we can observe, random forest algorithm has the best performance with F1-score of 0.78 for non-residential locations and 0.95 for residential locations. Since about 90% locations in our ground truth are residential locations, it is not surprising that performance for distinguishing residential locations are much better. Overall, the results validate the feasibility and correctness of the selected features.

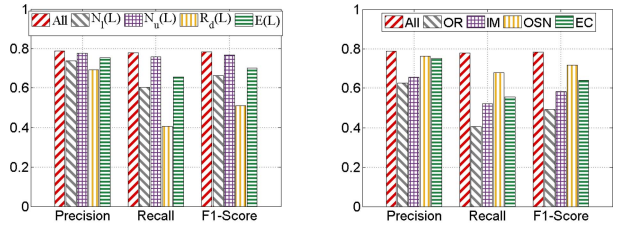
To compare the effectiveness of different features in the classifiers, we evaluate the performance of only using one feature or records of one services in Fig. 10(a) and (b), respectively. Since the majority of locations are residential locations, we focus on the performance of distinguishing non-residential locations in this experiment. From Fig. 10(a), we can observe that data-night ratio $R_d(L)$ has the worst performance to distinguish non-residential locations, since non-residential locations are also possible to have similar $R_d(L)$ with residential locations, e.g., hotel or bar. On the other hand, location entropy $E(L)$ and number of online IDs $N_u(L)$ are the most effective two features to distinguish non-residential locations, since it contains more information about the number of people in each location. In terms of different types of IDs shown in Fig. 10(b), features of OR accounts show the worst performance. The most possible reason is that the corresponding online review service only serves for a subset of non-residential locations. For the non-residential locations not served by the OR service, users do not use their OR accounts at these locations. Thus, using records of OR accounts cannot distinguish these locations. Similarly, records of OSN and EC accounts contain more information about the number of people in each location. Therefore, they are most effective to distinguish non-residential locations.

Overall, by elaborately selecting 4 features and the best classifier among three mainstream supervised learning algorithms, we divide millions of locations into residential and non-residential types.

### B. Clustering Non-Residential Locations

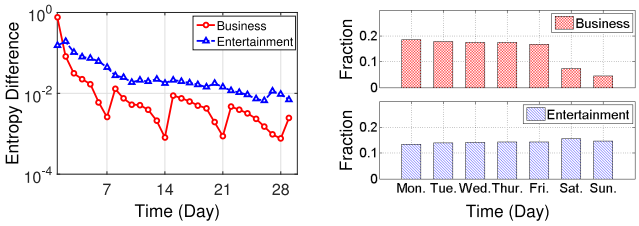
We further investigate location types among non-residential locations. Since we do not have ground truth information about non-residential locations, instead of supervised learning method, we choose to use an unsupervised learning method.

As discussed before, entropy plays an important role in distinguishing different types of locations. Thus, we further investigate entropy in temporal dimension. Specifically, we use the entropy of one location of the duration from one day to




(a) Different features (b) Different services

Fig. 10. Effectiveness of different features and services for distinguishing non-residential locations.




(a) Entropy difference. (b) Frequency of one week

Fig. 11. Distribution of features for business and entertainment locations.

the whole 30 day as a 30 dimensional vector, and then get the difference between adjacent elements, which is referred to as the entropy difference, and defined as follow:

$$D_i(L) = E_i(L) - E_{i-1}(L), \tag{13}$$

where $E_i(L)$ is the entropy of location $L$ during the first $i$ days, and $E_0(L)$ is set to be 0. Intuitively, the entropy difference can describe the difference of appeared IDs between adjacent days. Using it as a feature, we apply a clustering algorithm, i.e., hierarchical clustering [27]. The basic idea of hierarchical clustering is iteratively merging the nearest two clusters. It first considers each input point as a cluster and then bottom-up iteratively merges the nearest two clusters until the stop condition is met. In the clustering, we use the Euclidean distance as the distance metric and define the distance between clusters as average-linkage distance. Based on this method, we divide non-residential locations with stable and unstable appeared IDs into two clusters.

The distribution of entropy difference for the two clusters is shown in Fig. 11(a). As we can observe, the entropy difference of the first cluster reduces fast, which indicates the appeared IDs are stable. In addition, it shows a periodic variation with a cycle of one week, indicating periodicity of users. As for the other cluster of location, its entropy difference remains high throughout the month, indicating the appeared online IDs are unstable. Then in Fig. 11(b), we plot login number of different days in one week. As we can observe, there are more login records at weekdays for locations in the first cluster. As for the second cluster, there are more login records at weekends than those at weekdays. By combining results from these two figures, we label the first cluster as business locations, and the second cluster as entertainment locations.

To validate our conjecture that locations in the two clusters are corresponding to business and entertainment locations, we study the POI distribution around these locations. POI is a

TABLE IV
NORMALIZED POI DISTRIBUTION AROUND SUBSCRIBERS

| POI \ Cluster | Business | Entertainment |
|---|---|---|
| Office Building | 1.5634 | 0.9574 |
| Factory | 1.0435 | 0.4534 |
| Restaurant | 0.5635 | 1.3765 |
| Hotel | 0.4657 | 1.0343 |

TABLE V
SUMMARY OF LOCATION CLASSIFICATION

| Location Type | Residential | Business | Entertainment |
|---|---|---|---|
| Number | 3,271,529 | 36,071 | 55,200 |

specific point location of a certain function such as restaurant or shopping mall, and can be open accessed through the APIs of map service providers. An area's POI distribution can reflect its function. Note that in order to obtain the POI distribution of locations, we need GPS coordinates of these locations. However, this information is only used in evaluating the performance, while implementing location classification algorithms does not require such background knowledge.

Specifically, we crawl four types of POI within 200m of locations in the two clusters through APIs of the most popular map service in China, i.e., BaiduMap, and show their average normalized value in Table IV. As we can observe, the number of business POI, office Building and factory, around locations of the first cluster are much higher than that of the second one, while the number of entertainment POI, restaurant and hotel, around locations in the second cluster have larger values, indicating the correctness of our conjecture. Thus, we conclude the two clusters of locations to be business locations and entertainment locations.

Overall, by using the location entropy as the main feature and applying both supervised and unsupervised method, we are able to successfully divide locations into three types, i.e., residential, business and entertainment location. The inner formation mechanism of the selected features and the types of locations are ignored in our method. However, investigating this problem requires more fine-grained user trajectories with physical context, which is out of the scope of this paper. Thus, we leave it for future work. We summarize the number of different locations obtained based on the location classification module in Table V. Note that location classification helps to the semantic attack, which cannot be protected by the privacy model of $k$-anonymity. These three types of locations have covered most places where people access the Internet all around the city.

## VI. PRIVACY ANALYSIS

Having linked the online IDs belonging to the same user, attackers are able to derive the complete trajectories of physical users from the mobility records of the whole 32 million IDs. Moreover, attackers can infer the physical context of locations from the behavior of online IDs around them, which provides rich information about users' behavior. By combining the mobility trajectories and physical context, in this section we provide a thorough analysis of user physical world privacy leakage via cyberspace.

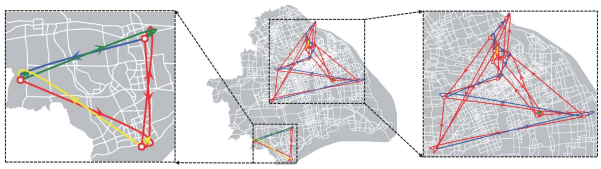


Fig. 12. Trajectory for example, in which red line indicates the complete trajectory of all online IDs, and line with other colors indicates trajectory of some single online ID.

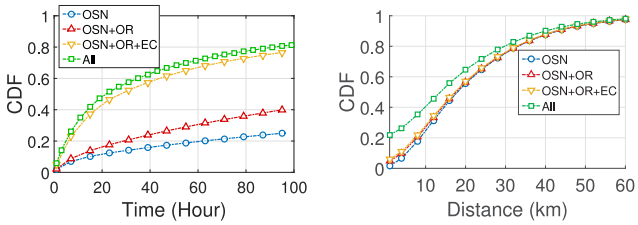



Fig. 13. Distance and time gap between adjacent records.

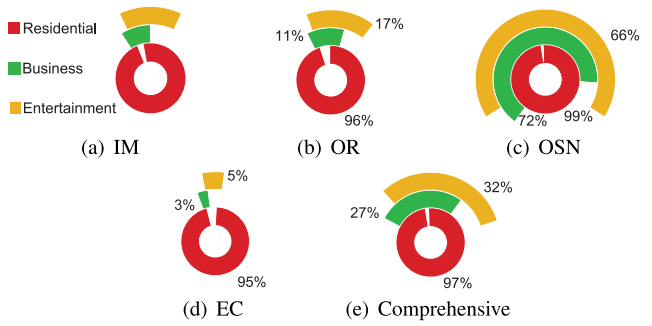
### A. Privacy Leakage Through Localization Attack

As introduced in Section III-C, attackers in localization attack aim to find out activities or events, which is described by spatio-temporal points, as many as possible in users' trajectories. The corresponding privacy leakage can be quantified by the quality of the obtained trajectories. Thus, we now present analysis about the quality of users' trajectories.

*1) Examples Case Study:* We first present two examples about the obtained trajectories in Fig. 12. As we can observe from Fig. 12, before merging the trajectories, we only know a part of places users have visited, and their retrieved trajectories are not complete. After merging login records of multiple types of online IDs belonging to the same user, almost complete trajectories of the users can be retrieved. Thus, benefiting from it, more information about users is obtained.

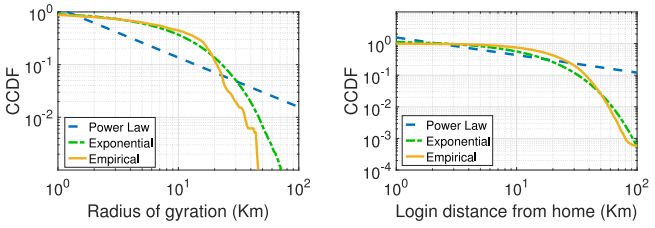
*2) Spatial and Temporal Resolution:* In order to measure the benefit obtained from merging different online IDs, we then study the spatial and temporal distance between adjacent records in the merged trajectories compared with unmerged trajectories. The results are shown in Fig. 13. As we can observe, by merging login records of multiple online IDs, the average time gap and spatial distance between adjacent records are obviously reduced. Specifically, the average time gap is reduced by 50%, i.e., from about 18 hours to 8 hours by merging login records from EC accounts to all types of IDs. In terms of the average spatial distance, it is also reduced by about 25%. These results demonstrate that by combining different types of online IDs together, the quality of the obtained trajectories is significantly improved.

*3) Basic Mobility Metrics:* In Fig. 15, we present the complementary cumulative distribution function (CCDF) of the obtained trajectories in terms of two mobility metrics, including radius of gyration [28] and login distance from home. The radius of gyration is shown in Fig. 15(a), which is the mean square root of the distance of each point in the trajectory to its center of mass, and can be formally defined as $r_g = \sqrt{\Sigma_{i=1}^{n}(r_i - r_{cm})^2/n}$, where $r_i$ represents the $i$th login



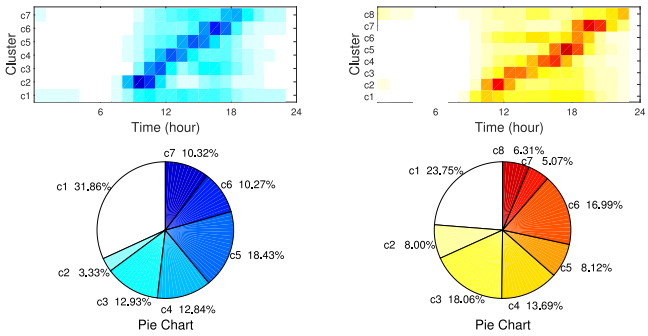

(a) IM (b) OR (c) OSN (d) EC (e) Comprehensive

Fig. 14. Ratio for covered locations of users for each type.




(a) (b)

Fig. 15. Basic mobility metrics of obtained trajectories.




(a) Time to work (b) Time to go home

Fig. 16. User clusters for their commute behavior.

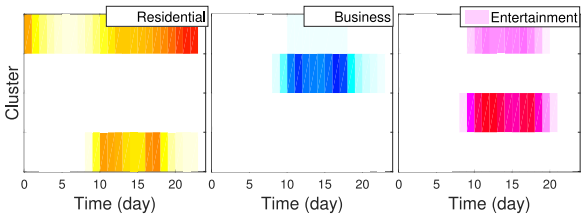



Fig. 17. User clusters of their standing time in different places.

recorded for $i \in \{1, \ldots, n\}$, $n$ is the total number of points in the trajectory, and $r_{cm} = 1/n \sum_{i=1}^{n} r_i$ is the center of mass of the trajectory. It characterizes the range of movement of each user. In the trajectories obtained from cell phone when users receive a call or a text message, the radius of gyration follows the truncated power-law distribution [28]. However, in our dataset, the distribution of radius of gyration can be approximated with the exponential distribution better. Another important indicator is login distance from home, which is shown in Fig. 15(b). According to [12], the check-in distance from home for Brightkite, Gowalla and the cell phones follows power-law distribution within around 100km. However, as we can observe, empirical distribution of login distance from home in our dataset is well approximated by exponential distribution rather than power-law distribution. Specifically, the average R-squared statistics between the empirical distribution and exponential distribution is as high as 0.9827, while for power-law distribution, it is only 0.6788. The fast decay rate compared with power-law distribution indicates the disadvantage of the obtained user trajectories in localization attack at places far away from users' center of activity.

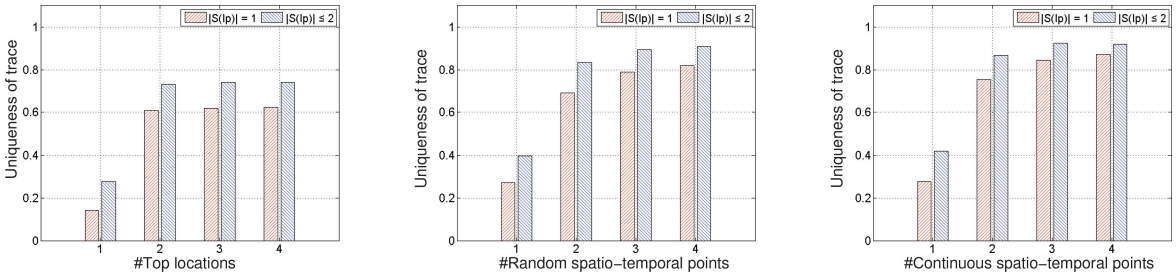
In summary, after linking the IDs of the same users and classifying different kinds of locations, we obtain more complete and meaningful user trajectories. The new user trajectory has a higher temporal and spatial resolution, which leads to more privacy leakage of users.

### *B. Privacy Leakage Through Semantic Attack*

As for semantic attack, adversaries aim to acquire individual's behavior or motivation at spatio-temporal points by using location semantic information. Thus, we first characterize users' privacy leakage by using the coverage rate of three types of locations for trajectories. Then, we focus on switching between different types of locations for users, and analyze their concrete privacy leakages, including when they go to work and go home, where and when they live, work, and relax. For example, if a user goes home from work, he might leave a pair of login records in business and residential locations, respectively. We have 3 types of locations. Thus, we have $3 \times 3 = 9$ switching types in total.

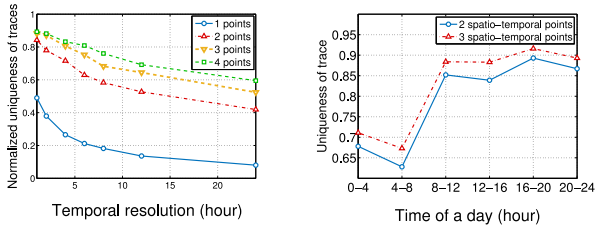
*1) Coverage Rate for Different Location Context:* We introduce coverage rate of three types of locations for trajectories, which is defined as the percentage of users of which the trajectories have at least one residential, business or entertainment location, respectively. The results are shown in Fig. 14, where "comprehensive" represents trajectories combining mobility records of all IDs belonging to the same users. As we can observe, coverage rates of residential locations for all types of online IDs are more than 95%. However, the coverage rates of business and entertainment locations for different types of online IDs are very diverse. EC accounts have the smallest coverage, indicating people tend to use their EC accounts in more private places; while OSN accounts have the highest coverage rate, indicating people tend to use their OSN accounts in more public places. Though OSN accounts have the highest coverage rate for business and entertainment locations, the number of OSN accounts is the smallest, as shown in Table II. Thus, by merging online IDs belonging to the same user, the coverage rate is balanced, however, smaller than OSN accounts, but increases 2-3 times compared with IM and EC accounts. In addition, the total number of covered users is significantly improved. By combining them together, we can obtain the trajectories covering more locations of all types, and thus characterize users in a more comprehensive way.

*2) Commute Analysis:* We study the switch between home and work places, which indicates users' commuting behavior. Using the hourly time series of the normalized frequency of




(a) (b) (c)

Fig. 18. The uniqueness of traces with respect to top $p$ locations, random $p$ spatio-temporal points, continuous $p$ spatiotemporal points. Red bars represent the fraction of unique traces, i.e., $|S(I_p)| = 1$, and blue bars represent the fraction of $|S(I_p)| \leq 2$, where $S(I_p)$ denote the anonymity set of the sub-trajectory $I_p$.




(a) Different temporal resolution (b) Different time periods

Fig. 19. Temporal difference of the privacy bound.

switches between home and work places as features and applying the k-means clustering algorithm [25], we successfully divide users into 7 clusters, where users commute at similar time are divided into one cluster. The clusters and their temporal distribution of the time to work as well as their percentage are shown in Fig. 16(a). Users in the clusters c2 to c7 tend to go to work from about 8AM to 6PM, respectively, while in the cluster c1, the time to go to work is dispersedly distributed all over the day, indicating users in this cluster may not go to work. The number of users in it is also the largest, which is about 31.86% of all users. Then, Fig. 16(b), we implement the same method to the time to go home of users. The results are similar with Fig. 16(a). Differently, the earliest time to go home is at around 12PM. By this method, the daily activity and schedule of users are exposed to us, which is critically sensitive information for users.

*3) Standing Time Analysis:* Based on location context obtained based on our algorithm proposed in Section V, we analyze the standing time of users in different type of places, including residence, business, and entertainment. Using it as feature and unsupervised learning method, we divide the users into four clusters, of which the results are shown in Fig. 17. User in the first three clusters tend to spend most of their time in residential, business and entertainment locations, respectively. Thus, we refer these users as residential-dominated, business-dominated and entertainment-dominated users. Take the business-dominated users for example. Their most login records are generated in business locations, indicating that they continuously stay in work places with the longest time. As for users in the last cluster, we can observe their standing time distributes in multiple types of locations. Thus, we refer them as comprehensive users. For any users, if we know which cluster they belong to, we know what and where they spend their time on doing every day.

Overall, the obtain users' physical world trajectories have a higher coverage rate of three kinds of locations. In addition, much detailed privacy can be extracted from users' trajectories. For example, when they commute, where and when they live, work, and relax, demonstrating that *user physical world privacy leakage in terms of semantic attack via user cyberspace privacy leakage* is a real potent threat.

### C. Privacy Leakage Through Re-Identification Attack

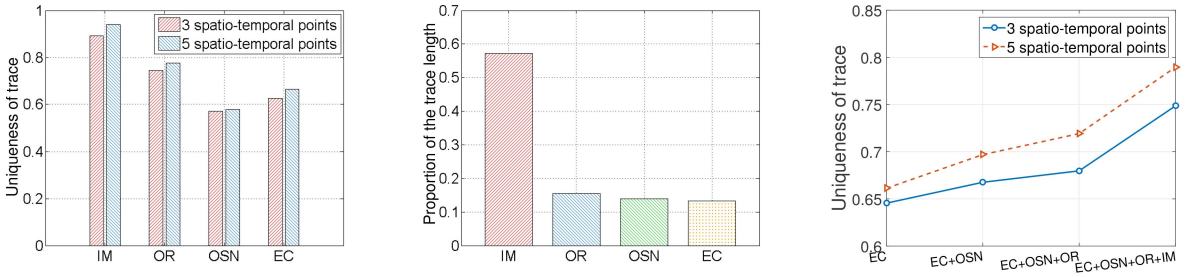
In this section, we adopt a well-recognized metric to measure the privacy leakage in terms of re-identification attack, which is introduced by Montjoye *et al.* [10] and has been used in a number of existing studies [29], [30]. Specifically, it is to estimate the number of points necessary to uniquely identify the mobility trace of an individual. If the uniqueness of trajectories is high, the mobility dataset is likely to be re-identifiable using information only on a few outside locations [10]. Thus, in order to analyze the privacy bounds of cyberspace fingerprints, we mainly focus on the uniqueness of trajectories in this section.

*1) Overall Privacy Bound:* We first analyze uniqueness of our obtained trajectories from three aspects: top $p$ locations, random $p$ spatio-temporal points, continuous $p$ spatio-temporal points with temporal resolution of 3 hours for $p$ from 1 to 4.

Specifically, let $I_p$ denote a sub-trajectory of a user with $p$ spatio-temporal points (for top $p$ locations, we do not consider temporal dimension). We define its anonymity set $S(I_p)$ as the subset of trajectories that match the $p$ spatio-temporal points composing $I_p$. Then, its uniqueness is characterized by $|S(I_p)|$, i.e., the size of its anonymity set. If $|S(I_p)| = 1$, its anonymity set only contains one trace, i.e., trajectory of its true owner. We define this sub-trajectory of $p$ points is unique, indicating that $p$ points are sufficient to re-identify this user. Then, we present the percentage of users satisfying $|S(I_p)| = 1$ and $|S(I_p)| \leq 2$ with different $p$ in Fig. 18.

As we can observe, top 4 locations can uniquely characterize 62% individuals, and 4 random spatio-temporal points are enough to identify 82% of the individuals, and 4 continuous spatio-temporal points are enough to identify 87% of the individuals, which indicates that uniqueness of trajectories is high in our dataset, and most users are likely to be re-identifiable by using only a few outside locations.

*2) Temporal Difference:* In Fig. 19, we analyze the uniqueness of users' trajectories at different time period and with



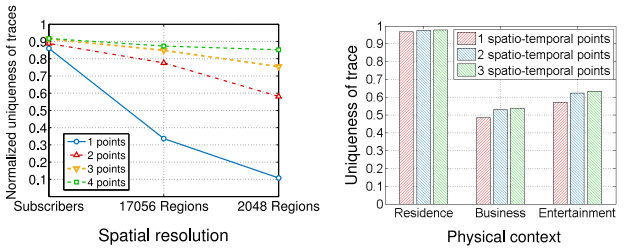

(a) Uniqueness of users (b) Component ratios (c) EC-based merging

Fig. 20. Privacy bounds of different types of online IDs.

different temporal resolution. The influence of temporal resolution to the privacy bound is shown in Fig. 19(a). It is not surprising that the uniqueness of trajectories is stronger with higher temporal resolution. For one spatio-temporal point, percentage of unique trajectories is reduced from about 50% to about 10% with temporal resolution from 1 hour to 1 day. However, with more spatio-temporal points, the influence of temporal resolution is reduced. For example, with 4 spatio-temporal points, the percentage of unique trajectories is only reduced from about 87% to about 60%. On the other hand, as we can observe from Fig. 19(b), privacy bound of trajectories at different time period are also diverse. Specifically, compared with nighttime, from 0AM to 8AM, trajectories at daytime expose more privacy of users. In addition, the trajectories during 4PM to 8PM expose most privacy of users.

*3) Spatial Difference:* We next analyze the uniqueness of trajectories with respect to the spatial resolution and location types. We first analyze the privacy bound of trajectories with different spatial resolution, which includes dividing the whole city into 2048 regions, 17056 regions, and over 3 million subscribers, respectively. The results are shown in Fig. 21(a). Similar with temporal resolution, a trivial observation is that with higher spatial granularity, more privacy of users is revealed. However, when there are 4 spatio-temporal points, the corresponding privacy bound is almost not influenced by the spatial granularity, remaining as high as 0.88 even when the city is only divided into 2048 regions, indicating that reducing spatial granularity does not work on preserving privacy under this condition. In addition, the uniqueness of points for locations of different types is shown in Fig. 21(b). Residential places have the highest uniqueness, followed by entertainment places, while business places expose the least information of users. It indicates that residential locations expose more privacy of users, which coincides with our empirical knowledge that home is more private places for users.

*4) Different Types of Online IDs:* Next, we analyze the influence of different types of IDs on the privacy leakage. As shown in Fig. 20(a), the uniqueness of trajectories for IM accounts is the strongest, indicating it contains the most information about users, while the uniqueness of trajectories for OSN accounts is the weakest. One possible reason is that users tend to use OSN accounts more in public places, which makes the corresponding trajectories less unique. Then, the component ratio of points for each type of online ID in the merged trajectory shown in Fig. 20(b). The component ratio for the IM account




(a) Different spatial resolution (b) Different physical context

Fig. 21. Spatial difference of the privacy bound.

is as high as 68%, while the component ratios for the other three types of accounts are all about 16%. Further, the process of merging records of multiple online IDs is shown in Fig. 20(c). By merging trajectories of different types of online IDs, their uniqueness is increased, indicating that by linking online IDs belonging to the same user together, more privacy of users is revealed.

In summary, the obtained user trajectories are highly unique. Even when spatial granularity is very low, 4 points are sufficient to uniquely identify 88% users, indicating that it is easy for the attacker to re-identify the trajectory of a targeted individual and make a big threat to users' privacy. In addition, the type of online IDs, the time period of day, and the physical context of locations show a big influence on the privacy bound, which can help to preserve user privacy in further work.

## VII. Related Work

The potential threat of user privacy leakage through online activities has attracted a lot of attention from the research community in the past decade. For example, it has been reported that a variety of personally identifiable information, i.e., age, gender, zipcode, address, or even real-name, can be leaked via HTTP headers, URIs, cookies that we leave when browsing the Web service [7], [31]–[33]. Furthermore, more underlaying personal privacy, i.e., sexual orientation, ethnicity, friendship, religious and political views, even personality traits, intelligence and happiness can also be inferred from the digital records we left [1], [34]–[37]. In order to preserve privacy, lots of technical solutions are proposed to control or reduce related sensitive information leakage [38]–[40]. For example, Sánchez *et al.* [41] proposed a system which selectively blocks or bypasses tracking on the browsed Web sites based on users'

privacy to achieve privacy-preserving advertising. However, on the other hand, recent studies find that personal information across different Web service of both the same kind [42], [43] and different kinds [2], [3], [44], [45] can be linked and aggregated, which aggravates the Web privacy leakage and related attacks. In addition, a number of studies investigated users' privacy behavior, e.g., users' privacy concern [46] and privacy fatigue [47], etc. All these studies highlight the possibility of personal information leakage from the cyberspace. In contrast, our work calls attention to another aspect of physical world privacy leakage when accessing the cyberspace Web services.

In terms of human mobility, recent extensive studies focus on discovering individual mobility patterns [28], revealing mobility prediction limits [48], and building accurate mobility model [49], [50] by the movement data trace collected by monitoring the devices' connected base station or GPS positions. On the other hand, individual mobility is revealed with high uniqueness to distinguish each other even in a large population, i.e., with only several locations visited most frequently [30] or several random spatial-temporal points [10], [51]. These investigations call attention to the privacy risks inferred from human mobility by re-identifying individuals from the spatial (only locations) [29], [30], [52] or temporal-spatial trajectories [10], along with a number of privacy-preserving techniques [29], [52]–[55]. All these works deal with the mobility understanding and privacy analysis based on the physical world directly observed human mobility. These are very different scenarios from the one we address here, because the mobility trajectories we investigate are inferred from the cookies information when we access the Web. In addition, as observed in this work, these kind of mobility trajectories is different from these directly obtained from the physical world.

In our system, user detection and location classification are other two key workflows. In terms of user identification, linking accounts of the same user across datasets are recognized as an important open problem studied in diversity contexts. A number of approaches are proposed to link user IDs based on datasets of graph structures, e.g., friendship graphs, contact graphs. Specifically, Korula and Lattanzi [56] and Zhang and Philip [57] focused on linking user IDs based on the friendship graph. Srivatsa and Hicks [58] focused on linking user IDs between friendship graphs and contacts graphs. Ji *et al.* [59], [60] further provided a theoretical analysis of linking users based on the graph structural data. Kazemi *et al.* [61] presented another graph matching algorithm that relies on smaller seeds than other approaches. Wang *et al.* [62], [63] clustered IDs belonging to the same users in one big contact graph of IDs. Zhou *et al.* [64] proposed an unsupervised method to link IDs of different social networks. In addition, other approaches focused on linking user IDs based on various user profile attributes and posted content by users. Goga *et al.* [19], [65] linked IDs based on user name, profile photos, writing styles, etc. Zafarani and Liu [66] linked IDs based on user names through behavioral modeling. Narayanan and Shmatikov [67] linked users of Netflix and IMDB based on the similarity of their movies ratings. Mu *et al.* [68] used "latent user space" for linking user profiles. Gao *et al.* [69] proposed an unsupervised method to link users based on their attributes and social features. However, these algorithms require information such as friendship graphs and user attributes. Thus, they are not applicable in our scenario. The only way to adapt these algorithms to the mobility trajectory data in our scenario is constructing the "contact graph" to model users' encountering with each other based on their mobility trajectories, but such adaption ignores users' daily mobility patterns, which limits the performance of linking user IDs. At the same time, a number of studies focused on linking user IDs based on trajectory data directly. Naini *et al.* [70] focused on linked users by matching the statistics of their trajectories. Riederer *et al.* [17] proposed an algorithm combining Poisson processes and maximum weight matching to link user IDs. Wang *et al.* [23] proposed an algorithm based on Gaussian mixture model which considers spatio-temporal mismatches between different datasets. Feng *et al.* [71] proposed a deep learning based algorithm to link user IDs, which used a co-attention mechanism to overcome the mismatches between different datasets. However, these approaches need accurate GPS coordinates of locations, which is not available in our scenario. In terms of location classification, it is also a hot topic recently [72], especially in location-aware social networks [12], [24], [73]. Different from previous solutions, we combine the supervised learning and un-supervised learning method to suit the dataset, which achieves better performance.

## VIII. Conclusion

In this work we have demonstrated that it is possible to infer and reveal user physical world privacy via cyberspace privacy leakage, namely, by correlating the cyber "fingerprints" (e.g., user IDs and other information contained in cookies) left by users in the cyberspace with the "footprints" (e.g., hints about physical locations) left by users in the physical world. We have developed a powerful privacy analysis system, which combines the login records of users and physical context information, and successfully reveals main privacy leakage of users. Our analysis unveils that cyberspace cookie logs contain high-quality user trajectories. In addition, most of user trajectories can be discovered and confirmed by leveraging only a few exogenous records of GPS coordinates. Furthermore, much detailed physical privacy of users can be inferred by applying some simple analytical methods to users' mobility trajectories. In summary, our study provides a systematic and comprehensive understanding of users' physical-world privacy leakages from their cookie logs.

## References

[1] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 110, no. 15, pp. 5802–5805, 2013.

[2] B. Krishnamurthy and C. E. Wills, "Generating a privacy footprint on the Internet," in *Proc. 6th ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2006, pp. 65–70.

[3] N. Xia *et al.*, "Mosaic: Quantifying privacy leakage in mobile networks," in *Proc. ACM Conf. SIGCOMM*, 2013, pp. 279–290.

[4] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3G network," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2009, pp. 267–279.

[5] H. Wang, C. Gao, Y. Li, Z.-L. Zhang, and D. Jin, "From fingerprint to footprint: Revealing physical world privacy leakage by cyberspace cookie logs," in *Proc. ACM Conf. Inf. Knowl. Manag. (CIKM)*, 2017, pp. 1209–1218.

[6] *Alexa's Digital Marketing Tools*, Amazon, Seattle, WA, USA. [Online]. Available: http://www.alexa.com/topsites/countries/CN

[7] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proc. 2nd ACM Workshop Online Soc. Netw. (WOSN)*, 2009. pp. 7–12.

[8] *Chrome DevTools*, Google, Mountain View, CA, USA. [Online]. Available: https://developer.chrome.com/devtools/

[9] G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, "'will check-in for badges': Understanding bias and misbehavior on location-based social networks," in *Proc. 10th Int. Conf. Web Soc. Media (ICWSM)*, 2016, pp. 417–426.

[10] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, p. 1376, Mar. 2013.

[11] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[12] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD)*, 2011, pp. 1082–1090.

[13] J.-H. Prinz *et al.*, "Markov models of molecular kinetics: Generation and validation," *J. Chem. Phys.*, vol. 134, no. 17, 2011, Art. no. 174105.

[14] B. Trendelkamp-Schroer and F. Noé, "Efficient bayesian estimation of Markov model transition matrices with given stationary distribution," *J. Chem. Phys.*, vol. 138, no. 16, 2013, Art. no. 04B612.

[15] K. P. Murphy, *Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.

[16] G. E. Andrews, *The Theory of Partitions*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[17] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, 2016, pp. 707–719.

[18] L. Rossi and M. Musolesi, "It's the way you check-in: Identifying users in location-based social networks," in *Proc. 2nd ACM Conf. Online Soc. Netw. (COSN)*, 2014, pp. 215–226.

[19] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 447–458.

[20] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[21] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *Proc. 7th ACM Int. Conf. Web Search Data Min. (WSDM)*, 2014, pp. 303–312.

[22] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explorations Newsletter*, vol. 18, no. 2, pp. 5–17, 2017.

[23] H. Wang, Y. Li, C. Gao, G. Wang, X. Tao, and D. Jin, "Anonymization and de-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," *IEEE Trans. Mobile Comput.*, early access, Nov. 11, 2019, doi: 10.1109/TMC.2019.2952774.

[24] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proc. 12th ACM Int. Conf. Ubiquitous Comput. (UbiComp)*, 2010, pp. 119–128.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[26] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[27] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Res.*, vol. 16, no. 22, pp. 10881–10890, 1988.

[28] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[29] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with GLOVE," in *Proc. 11th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, 2015, pp. 1–13.

[30] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2011, pp. 145–156.

[31] B. Krishnamurthy, K. Naryshkin, and C. E. Wills, "Privacy leakage vs. protection measures: The growing disconnect," in *Proc. Web 2.0 Security Privacy (W2SP)*, vol. 2, 2011, pp. 1–10.

[32] B. Krishnamurthy and C. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, 2009, pp. 541–550.

[33] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Analyzing facebook privacy settings: User expectations vs. reality," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf. (IMC)*, 2011, pp. 61–70.

[34] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul, "Who is .com?: Learning to parse WHOIS records," in *Proc. Internet Meas. Conf. (IMC)*, 2015, pp. 369–380.

[35] E. Pujol, O. Hohlfeld, and A. Feldmann, "Annoyed users: Ads and ad-block usage in the wild," in *Proc. Internet Meas. Conf. (IMC)*, 2015, pp. 93–106.

[36] S. Le Blond, C. Zhang, A. Legout, K. Ross, and W. Dabbous, "I know where you are and what you are sharing: Exploiting P2P communications to invade users' privacy," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf. (IMC)*, 2011, pp. 45–60.

[37] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez, "Follow the money: Understanding economics of online aggregation and advertising," in *Proc. Conf. Internet Meas. Conf. (IMC)*, 2013, pp. 141–148.

[38] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin, "Persona: An online social network with user-defined privacy," in *Proc. ACM SIGCOMM Conf. Data Commun.*, 2009, pp. 135–146.

[39] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 351–360.

[40] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes, "ReCon: Revealing and controlling PII leaks in mobile network traffic," in *Proc. 14th Annu. Int. Conf. Mobile Syst. Appl. Serv. (MobiSys)*, 2015, pp. 361–374.

[41] D. Sánchez and A. Viejo, "Privacy-preserving and advertising-friendly web surfing," *Comput. Commun.*, vol. 130, pp. 113–123, Oct. 2018.

[42] Y. Xie, F. Yu, and M. Abadi, "De-anonymizing the Internet using unreliable IDs," in *Proc. ACM SIGCOMM Conf. Data Commun.*, 2009, pp. 75–86.

[43] D. Irani, S. Webb, C. Pu, and K. Li, "Modeling unintended personal-information leakage from multiple online social networks," *Internet Comput.*, vol. 15, no. 3, pp. 13–19, 2011.

[44] S. Khemmarat, S. Saha, H. H. Song, M. Baldi, and L. Gao, "On understanding user interests through heterogeneous data sources," in *Proc. Int. Conf. Passive Active Netw. Meas. (PAM)*, 2014, pp. 272–274.

[45] H. Feng, K. Fawaz, and K. G. Shin, "LinkDroid: Reducing unregulated aggregation of app usage behaviors," in *Proc. 24th USENIX Conf. Security Symp. (USENIX Security)*, 2015, pp. 769–783.

[46] J. Kwon and M. E. Johnson, "The market effect of healthcare security: Do patients care about data breaches?" in *Proc. 14th Workshop Econ. Inf. Security (WEIS)*, 2015.

[47] H. Choi, J. Park, and Y. Jung, "The role of privacy fatigue in online privacy behavior," *Comput. Hum. Behav.*, vol. 81, pp. 42–51, Apr. 2018.

[48] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[49] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nat. Phys.*, vol. 6, no. 10, pp. 818–823, 2010.

[50] H. Gao, J. Tang, X. Hu, and H. Liu, "Modeling temporal effects of human mobile behavior on location-based social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, 2013, pp. 1673–1678.

[51] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[52] A. Monreale *et al.*, "Movement data anonymity through generalization," *Trans. Data Privacy*, vol. 3, no. 2, pp. 91–121, 2010.

[53] J. Domingo-Ferrer and R. Trujillo-Rasua, "Microaggregation- and permutation-based anonymization of movement data," *Inf. Sci.*, vol. 208, pp. 55–80, Nov. 2012.

[54] Y. Song, D. Dahlmeier, and S. Bressan, "Not so unique in the crowd: A simple and effective algorithm for anonymizing location data," in *Proc. ACM Property Insurance Rep. Nat. Conf. (PIR)*, 2014, pp. 19–24.

[55] N. Mohammed, B. Fung, and M. Debbabi, "Walking in the crowd: Anonymizing trajectory data for pattern analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manag. (CIKM)*, 2009, pp. 1441–1444.

[56] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proc. VLDB Endowment*, vol. 7, no. 5, pp. 377–388, 2014.

[57] J. Zhang and S. Y. Philip, "Multiple anonymized social networks alignment," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, Atlantic City, NJ, USA, 2015, pp. 599–608.

[58] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proc. ACM Conf. Comput. Commun. Security (CCS)*, 2012, pp. 628–637.
[59] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural data de-anonymization: Quantification, practice, and implications," in *Proc. ACM Conf. Comput. Commun. Security (CCS)*, 2014, pp. 1040–1053.
[60] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. A. Beyah, "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge," in *Proc. 22nd Annu. Netw. Distrib. Syst. Security Symp. (NDSS)*, 2015, pp. 1–15.
[61] E. Kazemi, S. H. Hassani, and M. Grossglauser, "Growing a graph matching from a handful of seeds," *Proc. VLDB Endownment*, vol. 8, no. 10, pp. 1010–1021, 2015.
[62] H. Wang, Y. Li, Y. Chen, and D. Jin, "Co-location social networks: Linking the physical world and cyberspace," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1028–1041, May 2019.
[63] H. Wang, Y. Li, G. Wang, and D. Jin, "You are how you move: Linking multiple user identities from massive mobility traces," in *Proc. SIAM Int. Conf. Data Min. (SDM)*, 2018, pp. 189–197.
[64] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.
[65] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD)*, 2015, pp. 1799–1808.
[66] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD)*, 2013, pp. 41–49.
[67] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy (SP)*, Oakland, CA, USA, 2008, pp. 111–125.
[68] X. Mu, F. Zhu, E.-P. Lim, J. Xiao, J. Wang, and Z.-H. Zhou, "User identity linkage by latent user space modelling," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD)*, 2016, pp. 1775–1784.
[69] M. Gao, E.-P. Lim, D. Lo, F. Zhu, P. K. Prasetyo, and A. Zhou, "CNL: Collective network linkage across heterogeneous social platforms," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, Atlantic City, NJ, USA, 2015, pp. 757–762.
[70] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016.
[71] J. Feng *et al.*, "DPLink: User identity linkage via deep neural network from heterogeneous mobility data," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 459–469.
[72] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma, "Web resource geographic location classification and detection," in *Proc. Spec. Interest Tracks Posters 14th Int. Conf. World Wide Web (WWW)*, 2005, pp. 1138–1139.
[73] X. Li, T.-A. N. Pham, G. Cong, Q. Yuan, X.-L. Li, and S. Krishnaswamy, "Where you instagram?: Associating your instagram photos with points of interest," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, 2015, pp. 1231–1240.

**Huandong Wang** (Member, IEEE) received the B.S. degrees in electronic engineering and mathematical sciences from Tsinghua University, Beijing, China, in 2014 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include software-defined networks, wireless ad hoc network, and mobile big data.

**Chen Gao** received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His work focuses specifically on the theory and applications of mobile big data and data mining.

**Yong Li** (Senior Member, IEEE) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012.

He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University. He has served as the General Chair, the TPC Chair, the SPC/TPC Member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 6900. Among them, ten are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia–Pacific Outstanding Young Researchers, Young Talent Program of China Association for Science and Technology, and the National Youth Talent Support Program.

**Zhi-Li Zhang** (Fellow, IEEE) received the B.S. degree in computer science from Nanjing University, Nanjing, China, in 1986, and the M.S. and Ph.D. degrees in computer science from the University of Massachusetts, Amherst, MA, USA, in 1992 and 1997, respectively.

In 1997, he joined the Computer Science and Engineering Faculty, University of Minnesota, Minneapolis, MN, USA, where he is currently a Qwest Chair Professor and Distinguished McKnight University Professor. His research interests lie broadly in computer communication and networks, Internet technology, multimedia, and emerging applications.

Dr. Zhang was a co-recipient of four Best Paper awards and has received a number of other awards. He has the co-chaired several conferences/workshops, including IEEE INFOCOM 2006 and served on the TPC of numerous conferences/workshops. He is a member of the Association for Computing Machinery.

**Depeng Jin** (Member, IEEE) received the B.S. and Ph.D. degrees in electronics engineering from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, where he is currently an Associate Professor and the Vice Chair of the Department of Electronic Engineering. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture. He was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002.