From Fingerprint to Footprint: Revealing Physical World Privacy Leakage by Cyberspace Cookie Logs

Huandong Wang[1],Chen Gao[1], Yong Li[1], Zhi-Li Zhang[2], Depeng Jin[1]
[1]Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
[2]University of Minnesota
liyong07@tsinghua.edu.cn

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.

KEYWORDS

Privacy; Cookie; Trajectories; Cyber-Physical Systems

1 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 [18, 21], and correlated and inferred – especially coupled with network traffic – to build a mosaic of various personal traits and activities [43].

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

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132998

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 several previous studies have shown [32, 41]. 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 will frequent a shopping mall in the dead of a night; whereas a location that are 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 exaction location information is available. 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.

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 2 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 three *popular online services in China, namely, QQ (online instant messenger), Weibo (online social network), Tmall and Taobao (online shopping sites).* A total of 470 million records containing 28.0 million distinct user IDs and spanning nearly the whole city are used in our study.

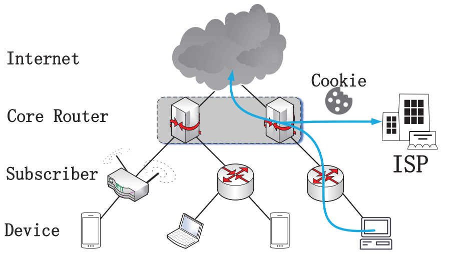


Figure 1: Framework for extracting cookies from packets.

The contributions of our study are summarized below:

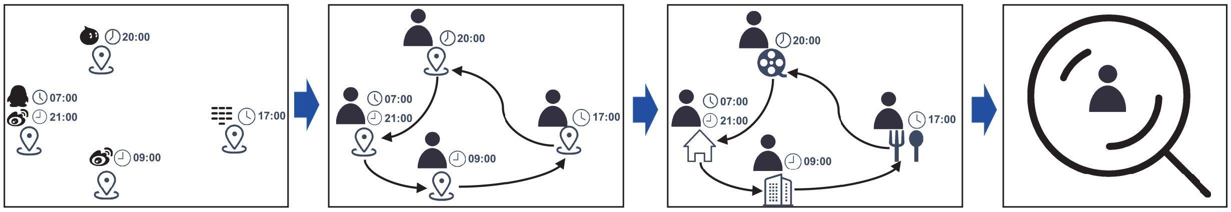
- We develop a user detection system to discover users' identifies in multiple cyberspace by utilizing the spatio-temporal locality. By checking against the ground-truth data, we validate that our algorithm achieves high accuracy with F1-score over 0.75.
- We develop a location-classification system 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.

The remainder of the paper is structured as follows. In Section 2, we describe collection and processing of the datasets used in our study. In Section 3, we motivate and provide a high-level overview of our proposed system. In Section 4, we develop a probabilistic approach using Gauss-Markov human mobility model to detect the online IDs for each user. In Section 5, we devise a location-classification system based on the location entropy. By combining the user mobility trajectories and physical context, in Section 6 we demonstrate that the user physical privacy can be revealed via cyber privacy leakage. After discussing related work in Section 7, we summarize our main findings in Section 8.

2 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.

The diagram of the first process is shown in Figure 1. Each subscriber can access the Internet via a broadband Internet connection at home or through various WiFi access points deployed by the ISP across the city. 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 where each HTTP session is generated. 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



(a) Original Dataset (b) User Detection (c) Location Classification (d) Privacy Analysis

Figure 2: Framework of the user tracking system.

be obtained. However, only 14% of packets in China use HTTPS [1], which indicates the overwhelming majority of users' cookies can be extracted from HTTP packets. On the other hand, though the datasets used in our study are collected by the ISP, the ISP is not the only possible attacker, since the cyberspace *fingerprints* of users is a ubiquitous content and can be easily obtained by different attackers. A company-level attacker, e.g., a service provider, can infer users' locations through packets uploaded by the applications installed in the their mobile phones. An individual attacker can also infer users' locations through crawling their publicly available online check-ins. However, the datasets collected by ISP give us the most comprehensive view of this kind of privacy leakage. Thus, we mainly focus on this dataset in our study.

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[22]. Inspired by this idea, we turn on the Chrome Developer Tools [13] and displays HTTP request/response headers containing cookies we need. Take the request header shown in Figure 3 as an example. It is generated when a user wants to login the Weibo account through www.weibo.com. Fields such as Accept and User-Agent have been ignored for simplicity. As we can observe, the ID is involved in the Path field, i.e., "<User ID>" in this request. 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 1. As mentioned in the introduction, our study focuses on three representative online services in China. 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 28.0 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. The large-scale datasets guarantee the credibility of our analyses.

Figure 3: A cookie of online services for example.

```
GET /u/<User ID>/home?wvr=5 HTTP/1.1
Host: www.weibo.com
Cookie: SUS=SID-<User ID>-1462809518-GZ-kegic-9be6
28ae4bc14c92b0ee9200543cc7f8
```

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.

3 SYSTEM OVERVIEW AND ROADMAP

Culling user online IDs from the cookies, we obtain the login records of massive online IDs with the corresponding time and locations as shown in Figure 2(a). Our goal is to reveal privacy leakage for 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, we 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 we infer the privacy leakage from the physical world footprint is the third challenge. To meet these challenges, we design our system with three modules as shown in Figure 2, 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. Trajectory of a single online ID is only a subset of mobility records left by its user, and thus we cannot only completely reveal the potential privacy leakage of users only based on their single online ID. Thus, in order to obtain the bound of users' privacy leakage from their login records, we must link all the online IDs for each user together to obtain the universal set of its mobility records. Thus, as shown in Figure 2(b), the first module of our system is to link all the online IDs belonging to each user, which is discussed in detail in Section 4.

On the other hand, physical context can provide rich information about users' behavior. For example, given the physical context, we can infer what people are doing through where users are located in physical world. Further, we can infer what people are going to do through where users are moving to. Through these behaviors, more privacy of users is exposed. Meanwhile, locations of places

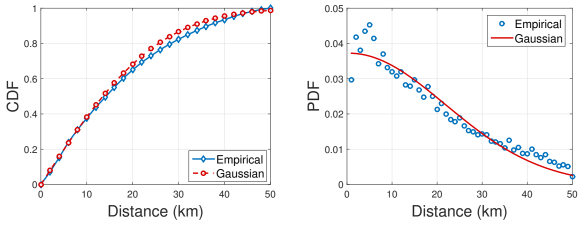
Table 1: Service types and regular expressions to match IDs.

Services	Type	Website	RegExp	Example

QQ	Instant messengers (IM)	qq.com	pt2gguin=o(\d+)
			o_cookie=(\d+)
				pt2gguin=
				o<User ID>

Taobao	E-commerce (EC)	tmall.com
		taobao.com
			lgc=(\w+)
			cna=(\w+)
				lgc=
				<User ID>

Weibo	Online social networks (OSN)	weibo.com	SID-(\w+) fid%(\w+)	SUS=SID- <User ID>



(a) (b)

Figure 4: Distribution of distance between users' login records of adjacent hours.

such as home and office are also critically important privacies for users. Thus, as shown in Figure 2(c), we investigate the physical context of each location in the second module of our system, and this module will be discussed in detail in Section 5.

Finally, by combining the all-round login records and physical context, we can thoroughly analyze the trajectories of users, and reveal the main privacy leakages. These are discussed in Section 6.

4 PHYSICAL WORLD USER DETECTION

In our system, we *do not* know the true user identity – e.g. the device ID – behind these online user IDs. In fact, in the routers of the backbone network, we have no user identification information. 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 [32, 41]. Inspired by this idea, we propose an algorithm which maximizes the likelihood of observed login records of online IDs based on Gauss-Markov human mobility model to solve this problem.

4.1 Model and Problem Formulation

We first propose a mathematical model and formulation for the problem. Let $\mathbf{A}$ represent the set of all online IDs in our dataset, and let $2^{\mathbf{A}}$ denote the power set of $\mathbf{A}$, i.e., the set of all subsets of $\mathbf{A}$. Given any online ID $a \in \mathbf{A}$, we define its mobility records as $R^a = \{(l_1, t_1), (l_2, t_2), ...\}$, where (l_i, t_i) represents a login record in location l_i at time slot t_i that was traced by cookie. Moreover, for a cluster of online IDs U, we define their mobility records $R^U = \{R^a | a \in U\}$. Then $R^{\mathbf{A}} = \{R^a | a \in \mathbf{A}\}$ represents the set of mobility records of all online IDs. Let $t(a) \in \mathbf{T}$ denote the type of online ID $a \in \mathbf{A}$, where $\mathbf{T}$={IM, OSN, EC} is set of all online ID types.

Definition 1 (Partition of A) Let $p = \{U_1, U_2, ..., U_n\}$, where $\forall k = 1, ..., n,\ U_k \in 2^{\mathbf{A}}$. We further define p as a partition of $\mathbf{A}$ if following four conditions hold: **(1)** $\emptyset \notin p$, **(2)** $\cup_{U \in p} U = \mathbf{A}$, **(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 a_1, $a_2 \in U$, then $t(a_1) \neq t(a_2)$.

Traditional definition of partition only needs three conditions (1)~(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 P as the set of all partitions.

Assume there is no shared online ID among different users. Then, there is an inherent partition of $\mathbf{A}$ composed of the true set of online IDs for each user denoted as p_{true}. Our problem, i.e., detecting all online IDs of each user, can be transformed to finding a partition p for $\mathbf{A}$ that are closed to p_{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 $\{a_i\}_{i=1}^{k}$, detecting all other online IDs belonging to the same user with them. That is, find a partition p, where elements involving $\{a_i\}_{i=1}^{k}$ approaches to elements involving $\{a_i\}_{i=1}^{k}$ in p_{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. To obtain the mobility model, we plot the cumulative distribution function (CDF) and probability distribution function (PDF) of the distance between login records of adjacent hours for all IDs in Figure 4(a) and (b), respectively. By fitting analysis, we find the empirical distribution can be approximated well by a Gauss distribution with $\sigma = 21.43$, with the average R-squared statistics of 99.85%. Thus, the mobility distance of users can be well approximated by a Gaussian stochastic variable.

Inspired by this observation, we assume the movement of users follows the Gauss-Markov model, i.e., the location of one user in the next time slot only depends on its current location, and the moving distance follows a Gaussian distribution. Then, the conditional probability of a login record given its time-adjacent login record can be calculated as follows,

$$p((l_2, t+1)|(l_1, t)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{d^2(l_1, l_2)}{2\sigma^2}),$$

Then, using Markov property, we can obtain the distribution of location between Δt hours as follows,

$$p((l_2, t+\Delta t)|(l_1, t)) = \frac{1}{\sqrt{2\pi\Delta t\sigma^2}} \exp(-\frac{d^2(l_1, l_2)}{2\Delta t\sigma^2}).$$

Without loss of generality, for a user's mobility records $R = \{(l_1, t_1), (l_2, t_2), ..., (l_n, t_n)\}$ with $t_1 < t_2 < ... < t_n$, its probability can be computed as:

$$p(R) = \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi\Delta t_i \sigma^2}} exp(\frac{\Delta d_i^2}{2\pi\Delta t_i \sigma^2}), \tag{1}$$

where $\Delta d_i = d(l_{i+1}, l_i)$ and $\Delta t_i = t_{i+1} - t_i$.

Intuitively, if IDs belonging to different users are linked by mistake, their merged trajectory is unreasonable, e.g., there may exist very large distance gap Δd between records of very small time gap Δt, leading to a small probability. Then, based on this model, we can make a Bayesian inference about the relationship of IDs.

4.2 Detection Method

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

$$\widehat{p} = \text{argmax}_{p \in P} P(p|R^{\mathbf{A}}).$$

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

$$P(p|R^{\mathbf{A}}) = \frac{P(R^{\mathbf{A}}|p)P(p)}{P(R^{\mathbf{A}})}.$$

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

$$P(R^{\mathbf{A}}|p) = \prod\nolimits_{U \in p} P(R^U|U),$$

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 (1) 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:

$$P(p|R^{\mathbf{A}}) \propto \prod_{U \in p} P(R^U|U)P(U).$$

We further assume users own each type of online ID independently with Bernoulli distribution with probability θ_t for $t \in \mathbf{T}$. 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, even finding an optimal partition for a subset of $\mathbf{A}$ 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 $\mathbf{A}$, 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:

$$q(C) = \log P(R^C|C)P(C).$$

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

$$\Delta q(C, a) = q(C \cup a) - q(C) - q(a).$$

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.

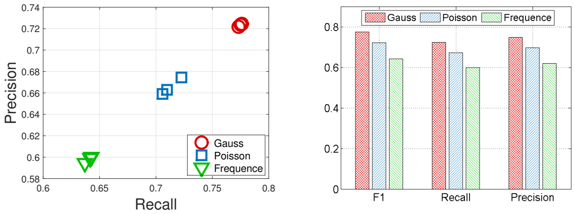
4.3 Performance Evaluation

To evaluate the accuracy of our proposed user detection algorithm, we need some ground-truth data for the validation. By a questionnaire survey, ISP obtained all online IDs of about 3000 users. The results are encrypted with the same encryption function as the DPI data by the ISP, and thus they can be matched with each other. We use this data as the ground truth in our study. For each user, we pick up one ID for a selected service type as the initial node, and use our algorithm to detect all IDs belonging to this user. After obtaining a set of IDs, we compare it with the ground-truth data by looking at other types of IDs.

We compare our algorithm with other two state-of-the-art approaches, which are described as follows:

Algorithm 1: Algorithm 1

```
Input: The set of IDs $\mathbf{A}$, the type $t(v) \in \mathbf{T}$ for all IDs $v \in \mathbf{A}$,
  and an initial online ID $u_0$.
Output: $C$, the cluster of IDs belonging to same user.
Initialize: $C \leftarrow \{u_0\}; T_C \leftarrow \mathbf{T}$;
while $\max_{u \in A \setminus C} \Delta q(C, u) > 0$ and $T_C \neq \emptyset$ do
    $u_{\max} = \mathrm{argmax}_{u \in \mathbf{A} \setminus C} \Delta q(C, u)$;
    $C = C \cup \{u_{\max}\}$;
    $T_C = T_C \setminus \{t(u_{\max})\}$.
```



(a) Precision and Recall.

(b) Best Performance.

Figure 5: Performance of different solutions.

Poisson-based approach: Riederer et al. [34] 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. Then, they find the maximum weighted matching of online IDs as the results.

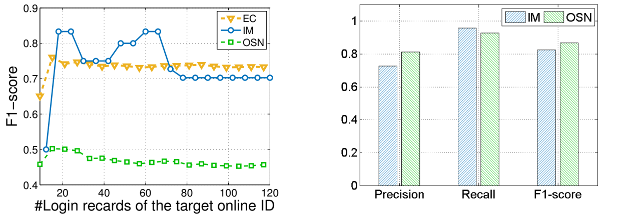
Frequency-based approach: Rossi et al. [35] use the frequency of login to approximate the probability of visit, which is represented as: $P(l|R^U) = \frac{N_l^U + \alpha}{\sum_{l \in L} N_l^U + \alpha|L|}$. N_l^U is the number of login records of user (online ID) U at location l. In addition, α>0 is 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 $\prod_{(l,t) \in R^a} P(l|R^U)$.

There are some tunable parameters, σ and θ_t, in our model that both have physical meaning and can be estimated directly from the ground truth. Thus, we evaluate the performance of our algorithm by 3-fold cross-validation. That is, we split our ground truth into three equal sets, and use each two sets to estimate the parameters, and evaluate the performance of the algorithm to the left set under the obtained parameters.

We use three key metrics in binary classification to quantify the detection accuracy, i.e., precision, recall, and F1-score[31]. 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 $F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. The results are shown in Figure 5 and Figure 6.

In Figure 5, we compare the performance of our proposed algorithm with the two state-of-the-art algorithms. Figure 5(a) shows the precision-recall plots with time granularity of 1, 2 and 3 hours respectively. From the result, we can observe that the time granularity does not influence the performance of our proposed algorithm, while it has larger influence on the performance of the other two approaches. In addition, we present the best precision, recall and F1-score in Figure 5(b). It shows that our algorithm outperforms others in terms of all metrics. Specifically, the best F1-score of our algorithm is over 0.75, improving 0.07 and 0.20 compared with two state-of-art algorithms, respectively.

In Figure 6, we evaluate our algorithm with different types of online IDs. Figure 6(a) shows the performance with different types of intial IDs, in which the F1-score is evaluated as the function of the number of login records of the target users. As we can observe, users with more login records can be detected with higher accuracy, especially for users with records less than 20. When there are



(a) Performance vs.#login records for different types of intial IDs.

(b) Best performance with intial EC accounts.

Figure 6: Performance of different types of IDs.

80 or more records, the F1-score turns to be stable. On the other hand, as for the overall performance, user detection with initial IDs of EC shows better performance with stable F1-score above 0.73, while that of OSN shows the worse performance, with stable F1-score of 0.46. 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 Figure 6(b), detection to all types of online IDs shows F1-score larger than 0.8. Another interesting observation is that performance of OSN accounts, in turn, shows the 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, more information about the user can be obtained by investigating their OSN accounts, 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 Gauss-Markov human mobility. Results show that our solution achieves good performance. That is, the overall F1-score is about 0.75, and for IM and EC account, the F1-score can reach 0.85.

5 LOCATION CLASSIFICATION

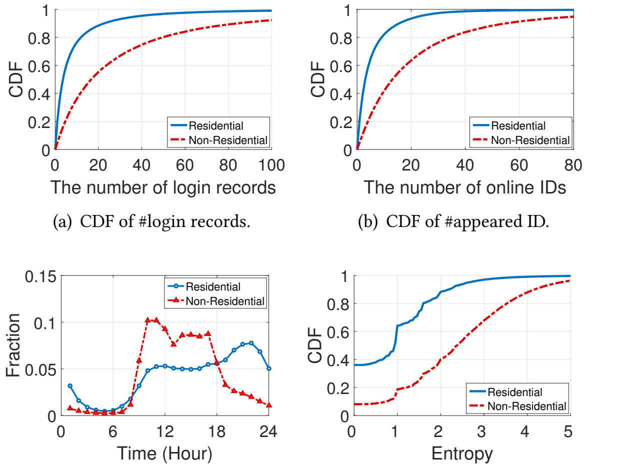
In order to characterize the physical-world behaviors of users, we need to infer the types of the locations that online IDs login. On the other hand, since in our system the subscriber identity is encrypted, and the associated location information is insufficient, we cannot obtain their types directly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

5.1 Distinguishing Residential and Non-Residential Locations

From the subscribers' registration information, we obtain 10,000 locations with type of residential or non-residential. We use them as training set, which enables us to utilize supervised learning algorithm to distinguish residential or non-residential locations.

Table 2: Precision and recall 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



(a) CDF of #login records.

(b) CDF of #appeared ID.

(c) 24 Hours' Distribution.

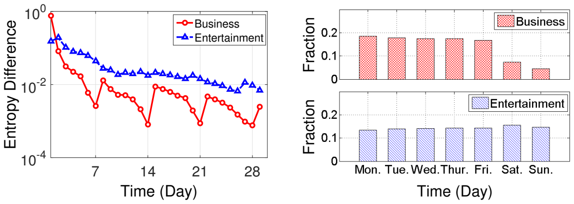
(d) CDF of location entropy.

Figure 7: Distribution of features for residential and non-residential locations.

Through detailedly investigating the character of residential and non-residential locations, we select four features, 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 Figure 7.

As we can observe from Figure 7(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 Figure 7(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 have been found that places such as the university campus, shopping and dining districts, have high entropy, while residential areas have low entropy [4]. In fact, we plot the distribution of entropy of different types of locations in Figure 7(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.

Using these features above, we apply three mainstream supervised learning algorithms, i.e., logistic regression (LR), support vector machine (SVM), random forest (RF) [30]. 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 [40]. Random forest is an ensemble learning method that operates by constructing and combining a



(a) Entropy difference.

(b) Frequency of one week

Figure 8: Distribution of features for business and entertainment locations.

multitude of weak learners with random subset of the features and training set. Due to the 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 2. 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. The results validate the feasibility and correctness of the selected features.

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.

5.2 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 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:

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

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 and applying clustering algorithm, 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 Figure 8(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 Figure 8(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.

Table 3: Normalized POI distribution around subscribers.

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

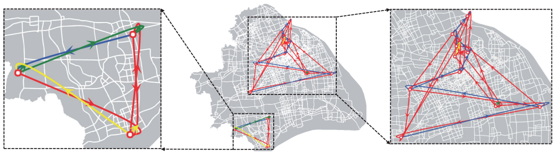


Figure 9: Trajectory for example, in which red line indicates the complete trajectory of all online IDs, and line with other colors indicates trajectory of some single online ID.

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 specific point location of a certain function such as restaurant or shopping mall. An area's POI distribution can reflect its function. Specifically, we study the four types of POI within 200m of locations in the two clusters, and show their average normalized value in Table 3. To be better compared with, the number of POI is normalized by the mean value of the corresponding type. 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 learning, we are able to successfully divide locations into three types, i.e., residential, business and entertainment location. These three types of locations have covered most places where people access the Internet all around the city.

6 USER PRIVACY ANALYSIS

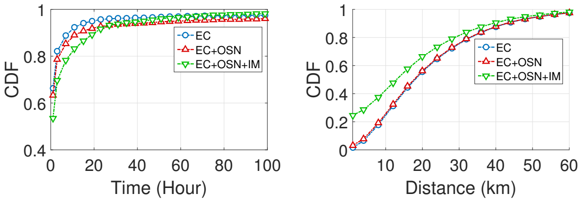
Having linked the online IDs belonging to the same user, we are able to derive the complete trajectories of physical users. Moreover, we can infer the physical context of locations from the behavior of online IDs around them, which provides rich information about users' behavior. By combining the mobility trajectories and physical context, in this section we provide a thorough analysis of user physical world privacy leakage via cyberspace. We first provide basic analysis about the obtained user trajectories in terms of quality. Then, we focus on the privacy bound, i.e., the uniqueness of our obtained trajectories.

6.1 Quality of Users' Trajectories

We now present some basic analysis about users' trajectories in terms of their quality.

1) Example case study: We first present two examples about the obtained trajectories in Figure 9. As we can observe, 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 IDs, we study the spatial



(a)

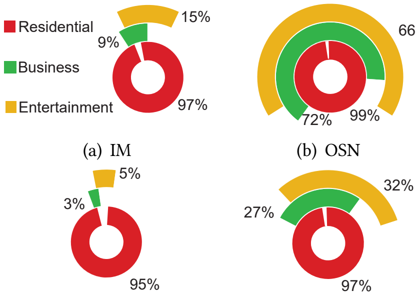
(b)

Figure 10: Distance and time gap between adjacent records.

and temporal distance between adjacent records in the merged trajectories compared with unmerged trajectories. The results are shown in Figure 10. 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 spatial distance, it is reduced by 25%. These results demonstrate that by combining different types of online IDs together, the quality of trajectories is significantly improved.

3) Coverage rate for different location context: Another important metric to quantify the benefit is the coverage rate of three types of locations for trajectories, i.e., the percentage of users of which the trajectories have at least one residential, business or entertainment location, respectively. The results are shown in Figure 11. 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 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 1. 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.

4) Basic mobility metrics: In Figure 12, we present the complementary cumulative distribution function (CCDF) of the obtained trajectories in terms of two mobility metrics, including radius of gyration[12] and login distance from home. The radius of gyration is shown in Figure 12(a), which is the mean square root of the distance of each point in the trajectory to its center of mass, and can be computed as $r_g = \sqrt{\Sigma_{i=1}^{n}(r_i - r_{cm})^2/n}$, where r_i represents the ith login recorded, 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 [12]. 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 Figure 12(b). According to [3], the check-in distance from home for Brightkite, Gowalla



(a) IM

(b) OSN

(c) EC

(d) Comprehensive

Figure 11: Covered locations of different types of IDs.

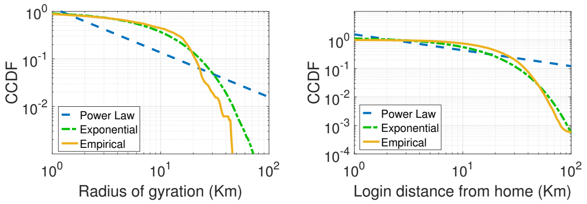
and the cell phones follows power-law distribution within around 100km. However, as we can observe, empirical distribution of login distance from home in our dataset is well approximated by exponential distribution rather than power-law distribution. Specifically, the average R-squared statistics between the empirical distribution and exponential distribution is as high as 0.9827, while for power-law distribution, it is only 0.6788. Overall, the users' trajectories inferred from the cookies information are very different from trajectories obtained from base station or GPS positions. Thus, it is a new observation method of human behavior that is worth studying.

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 spacial resolution, and has a higher coverage rate of three kinds of locations, which provides a more comprehensive view of the user behavior. In addition, it is very different from trajectories directly obtained from the physical world in terms of mobility metrics including radius of gyration and login distance from home, indicating it also provides a different view of the user behavior of research value.

6.2 Privacy Bounds of Users' Trajectories

Uniqueness of trajectory is a well-recognized metric to measure its privacy bounds, which is introduced by Montjoye et al. [5]. 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 [5]. Thus, in order to analyze the privacy bounds of cyberspace fingerprints, we mainly focus on the uniqueness of trajectories in this section.

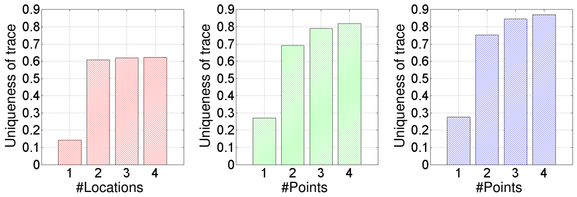
1) Overall privacy bound: We first analyze uniqueness of our obtained trajectories from three aspects: top N locations, random N spatio-temporal points, continuous N spatio-temporal points with temporal resolution of 3 hours for N from 1 to 4. The obtained results are shown in Figure 13. As we can observe, top 4 locations



(a)

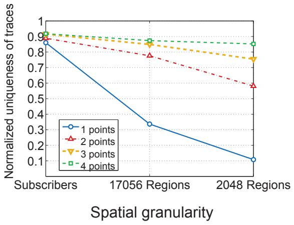
(b)

Figure 12: Basic mobility metrics of obtained trajectories.



(a) Top (b) Random (c) Continuous

Figure 13: Uniqueness of traces for top N locations, random N spatio-temporal points, and continuous N spatio-temporal points.



(a) (b)

Figure 14: Spatial difference of the privacy bound.

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 our obtained trajectories is high and most users are likely to be re-identifiable by using only a few outside locations.

2) Spatial difference: In Figure 14, we analyze the uniqueness of trajectories with respect to the spatial granularity and location types. The privacy bound of trajectories with different spatial granularity, which includes dividing the whole city into 2048 regions, 17056 regions, and over 3 million subscribers, respectively. The results are shown in Figure 14(a). 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 Figure 14(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.

3) Different types of online IDs: Next, we analyze the influence of different types of IDs on the privacy leakage. As shown in Figure 15(a), the uniqueness of trajectories for IM accounts is the strongest, indicating it contains the most information about users, while EC accounts is the weakest. Further, the process of merging records of multiple online IDs is shown in Figure 15(b). 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.

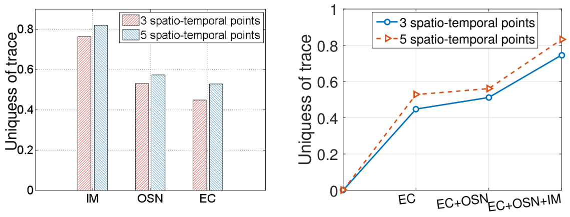
In summary, the obtained user trajectories are highly unique. Even when spatial granularity is very low, 4 points are sufficient to uniquely identify 88% users, indicating that it is easy for the attacker to re-identify the trajectory of a targeted individual and make a big threat to users' privacy.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

7 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 [19, 20, 22, 25, 29] 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 left when browsing the web service. Furthermore, it has been shown [18, 23] that additional personal or private information about users, e.g., sexual orientation, etc. can also be inferred from the digital records left online. In order to preserve privacy, a number of technical solutions have been proposed [2, 8, 33, 39]. All these studies highlight various aspects of cyberspace user privacy leakage. In contrast, our work calls attention to another aspect of user privacy leakage – physical world privacy leakage when accessing the cyberspace web services.

In terms of human mobility, recent extensive studies focus on discovering individual mobility patterns[12], revealing mobility prediction limits[37] and building accurate mobility model[9, 36]. On the other hand, individual mobility is revealed with high uniqueness to distinguish each other even in a large population[5, 6, 44] These investigations call attention to the privacy risks inferred from human mobility [5, 14, 16, 27, 44], along with a number of privacy-preserving techniques [7, 14, 26, 27, 38]. 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 are inferred from the cookies information when we accessing the web.

In our system, user detection and location classification are other two key workflows. In terms of user detection, linking accounts of the same user across datasets are recognized as an important open problem [15, 17, 28, 34]. Most existing solutions rely on either utilizing different portions of the same dataset[11, 15, 17] or observing the same behavior across thematically similar domains[10, 28]. The only approach proposed to date that is able to provide generic and self-tunable solution is POIS[34] by utilizing the temporal-spatial behaviors of humans. However, this approach can only match user identifiers of two domains. Moreover, the mobility assumption in the built model and proposed theory are far from reality, and we pick it as a benchmark for our comparative analyses. In terms of location classification, it is also a hot topic recently[42], especially



(a) Uniqueness of users (b) EC-based merging

Figure 15: Privacy bounds of different types of IDs.

in location-aware social networks[3, 4, 24]. Different from previous solutions, we combine the supervised and unsupervised learning method to suit the dataset, which achieves better performance.

8 CONCLUSIONS

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 *user physical-world privacy leakages via user cyberspace privacy leakage.*

ACKNOWLEDGMENTS

This work is supported by research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology, in part by US DTRA grant HDTRA1-14-1-0040, DoD ARO MURI Award W911NF-12-1-0385, and NSF grant CNS-1411636.

REFERENCES

[1] Amazon. [n. d.]. *Alexa's digital marketing tools.* http://www.alexa.com/topsites/countries/CN.

[2] Randy Baden, Adam Bender, Neil Spring, Bobby Bhattacharjee, and Daniel Starin. 2009. Persona: an online social network with user-defined privacy. In *Proc. ACM SIGCOMM.*

[3] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proc. ACM SIGKDD.*

[4] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the gap between physical location and online social networks. In *Proc. ACM Ubicomp.*

[5] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013).

[6] Yves-Alexandre de Montjoye, Laura Radaelli, and Vivek Kumar Singh. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.

[7] Josep Domingo-Ferrer and Rolando Trujillo-Rasua. 2012. Microaggregation-and permutation-based anonymization of movement data. *Information Sciences* 208 (2012), 55–80.

[8] Lujun Fang and Kristen LeFevre. 2010. Privacy wizards for social networking sites. In *Proc. WWW.*

[9] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2013. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proc. ACM CIKM.*

[10] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proc. WWW.*

[11] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. 2015. On the reliability of profile matching across large online social networks. In *Proc. ACM SIGKDD.*

[12] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

[13] Google. [n. d.]. *Chrome developer tools.* https://developer.chrome.com/devtools/.

[14] Marco Gramaglia and Marco Fiore. 2015. Hiding Mobile Traffic Fingerprints with GLOVE. In *Proc. ACM CoNEXT.*

[15] Ehsan Kazemi, S Hamed Hassani, and Matthias Grossglauser. 2015. Growing a graph matching from a handful of seeds. *Proceedings of the VLDB Endowment* 8, 10 (2015), 1010–1021.

[16] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. 2005. Protection of location privacy using dummies for location-based services. In *Proc. ICDE Workshops.*

[17] Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment* 7, 5 (2014), 377–388.

[18] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *PNAS* 110, 15 (2013), 5802–5805.

[19] Balachander Krishnamurthy, Konstantin Naryshkin, and Craig E Wills. 2011. Privacy leakage vs. Protection measures: the growing disconnect. In *Proc. W2SP.*

[20] Balachander Krishnamurthy and Craig Wills. 2009. Privacy diffusion on the web: a longitudinal perspective. In *Proc. WWW.*

[21] Balachander Krishnamurthy and Craig E Wills. 2006. Generating a privacy footprint on the internet. In *Proc. ACM IMC.*

[22] Balachander Krishnamurthy and Craig E Wills. 2009. On the leakage of personally identifiable information via online social networks. In *Proc. ACM WOSN.*

[23] Stevens Le Blond, Chao Zhang, Arnaud Legout, Keith Ross, and Walid Dabbous. 2011. I know where you are and what you are sharing: exploiting P2P communications to invade users' privacy. In *Proc. ACM IMC.*

[24] Xutao Li, Tuan-Anh Nguyen Pham, Gao Cong, Quan Yuan, Xiao-Li Li, and Shonali Krishnaswamy. 2015. Where you instagram? Associating your instagram photos with points of interest. In *Proc. ACM CIKM.*

[25] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing facebook privacy settings: user expectations vs. reality. In *Proc. ACM IMC.*

[26] Noman Mohammed, Benjamin Fung, and Mourad Debbabi. 2009. Walking in the crowd: anonymizing trajectory data for pattern analysis. In *Proc. ACM CIKM.*

[27] Anna Monreale, Gennady L Andrienko, Natalia V Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. 2010. Movement Data Anonymity through Generalization. *Transactions on Data Privacy* 3, 2 (2010), 91–121.

[28] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proc. IEEE SP.*

[29] Jahna Otterbacher. 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proc. ACM CIKM.*

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

[31] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[32] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. 2012. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review* 16, 3 (2012), 33–44.

[33] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. 2015. Recon: Revealing and controlling privacy leaks in mobile network traffic. In *Proc. MobiSys.*

[34] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking Users Across Domains with Location Data: Theory and Validation. In *Proc. WWW.*

[35] Luca Rossi and Mirco Musolesi. 2014. It's the way you check-in: identifying users in location-based social networks. In *Proc. ACM COSN.*

[36] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (2010), 818–823.

[37] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[38] Yi Song, Daniel Dahlmeier, and Stéphane Bressan. 2014. Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data. In *ACM PIR*. 19–24.

[39] Yi Song, Panagiotis Karras, Sadegh Nobari, Giorgos Cheliotis, Mingqiang Xue, and Stéphane Bressan. 2012. Discretionary social network data revelation with a user-centric utility guarantee. In *Proc. ACM CIKM.*

[40] Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9, 3 (1999), 293–300.

[41] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic, and Antonio Nucci. 2009. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In *Proc. ACM IMC.*

[42] Chuang Wang, Xing Xie, Lee Wang, Yansheng Lu, and Wei-Ying Ma. 2005. Web resource geographic location classification and detection. In *Proc. WWW.*

[43] Ning Xia, Han Hee Song, Yong Liao, Marios Iliofotou, Antonio Nucci, Zhi-Li Zhang, and Aleksandar Kuzmanovic. 2013. Mosaic: Quantifying privacy leakage in mobile networks. In *Proc. ACM SIGCOMM.*

[44] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proc. ACM MobiCom.*