Evaluating the Privacy Risk of Location-Based Services

Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux

LCA1, EPFL, Switzerland
firstname.lastname@epfl.ch

Abstract. In modern mobile networks, users increasingly share their location with third-parties in return for location-based services. In this way, users obtain services customized to their location. Yet, such communications leak location information about users. Even if users make use of pseudonyms, the operators of location-based services may be able to identify them and thus affect their privacy. In this paper, we provide an analysis of the erosion of privacy caused by the use of location-based services. To do so, we experiment with real mobility traces and measure the dynamics of user privacy. This paper thus details and quantifies the privacy risks induced by the use of location-based services.

1 Introduction

In traditional cellular networks, users share their location with their network operator in order to obtain voice and data services pervasively. With the emergence of data services, users increasingly share their location with other parties such as location-based services (LBSs). Specifically, users first obtain their location by relying on the localization capability of their mobile device (e.g., GPS or wireless triangulation), share it with LBSs and then obtain customized services based on their location. Yet, unlike cellular operators, LBSs are mainly provided for free and generate revenue with location-based advertisement. Hence, there is a key difference between the business models of LBS providers and cellular operators: LBS providers aim at profiling their users in order to serve tailored advertisement.

Subscribers of cellular networks know that their personal data is contractually protected. On the contrary, users of LBSs often lack an understanding of the privacy implications caused by the use of LBSs [18]. Some users protect their privacy by hiding behind pseudonyms that are mostly invariant over time (e.g., twitter usernames). Previous works identified privacy threats induced by the use of LBSs and proposed mechanisms to protect user privacy. Essentially, these mechanisms rely on trusted third-party servers that anonymize requests to LBSs. However, such privacy-preserving mechanisms are not widely available and users continue sharing their location information unprotected with third-parties. Similarly, previous work usually considers the worst-case scenario in which users continuously upload their location to third-parties (e.g., traffic monitoring systems). Yet, with most LBSs, users do not share their location continuously but instead, connect *episodically* to LBSs depending on their needs and thus reveal a few location samples of their entire trajectory. For example, a localized search on Google Maps [7] only reveals a location sample upon *manually* connecting to the service.

In this work, we consider a model that matches the common use of LBSs: we do not assume the presence of privacy-preserving mechanisms and consider that users access LBSs on a regular basis (but not continuously). In this setting, we aim at understanding the privacy risk caused by LBSs. To do so, we experiment with real mobility traces and investigate the *dynamics* of user privacy in such systems by measuring the *erosion* of user privacy. In particular, we evaluate the success of LBSs in predicting the true identity of pseudonymous users and their points of interest based on small samples of mobility traces. Our results explore the relation between the *type* and *quantity* of data collected by LBSs and their ability to de-anonymize and profile users. We quantify the potential of these threats by carrying out an experimentation based on real mobility traces from two cities, one in Sweden and one in Switzerland. We show that LBS providers are able to uniquely identify users and accurately profile them based on a small number of location samples observed from the users. Users with a strong routine face higher privacy risk, especially if their routine does not coincide with that of others. To the best of our knowledge, this work is among the first to investigate the erosion of privacy caused by the sharing of location samples with LBSs using real mobility traces and to quantify the privacy risk.

2 State of the Art

Location-based services [1,7,13,29] offer users to connect with friends, to discover their environment or to optimize their mobility. In most services, users share their location episodically when connecting to the service. Some services such as road-traffic monitoring systems require users to continuously share their location. In this work, we consider users that manually share their location and thus only reveal samples of their mobility.

The IETF Geopriv working group [3] aims at delivering specifications that will help implementing location-aware protocols in a privacy-conscious fashion. It proposes to use independent location servers that deliver data to LBSs according to privacy policies defined by users. In other words, it provides user control over the sharing of their location data. In this paper, we complement the IETF proposal by enabling to quantify the privacy threat induced by the sharing of specific location data with LBSs.

Privacy-preserving mechanisms impede LBSs from tracking and identifying their users [5,20,21,22,24,31,42]. In general, the proposed mechanisms either run on third-party anonymizing servers, or directly on mobile devices. Most mechanisms alter the user identifier or the content of location samples. For example, *anonymization* techniques remove the identifier from the user requests, and *obfuscation* techniques blur the location information. The effectiveness of privacy-preserving mechanisms is usually evaluated by measuring the level of privacy [11,37,39,40]. Most existing works consider worst-case scenarios in which users continuously share their location with LBSs. However, privacy-preserving mechanisms are rarely used in practice. One reason may be that users do not perceive the privacy threat because they are not intensively sharing their location. As this is rarely the case in practice, in this work, we aim at clarifying the privacy threat when users reveal samples of their mobility manually and do not make use of privacy-preserving mechanisms.

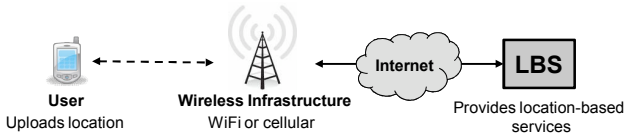


Fig. 1: System model. Users episodically upload their location wirelessly to LBSs.

Without privacy-preserving mechanisms, location information enables the identification of mobile users [6]. Partridge and Golle [19] could identify most of the US working population using approximate home and work locations. In other words, a small amount of information (i.e., two location samples) may be sufficient to uniquely identify most users. Similarly, Beresford and Stajano [4] identified all users in continuous location traces by examining where users spent most of their time. Using GPS traces from vehicles, two studies by Hoh *et al.* [23] and Krumm [25] found the home addresses of most drivers. De Mulder *et al.* [32] could identify mobile users in a GSM cellular network from pre-existing location profiles by using statistical identification processes. In [30], Ma *et al.* study the erosion of privacy caused by published anonymous mobility traces and show that an adversary can rapidly relate location samples to published anonymous traces. In this work, we push further the analysis of the de-anonymization threat by considering an adversary that learns only few location samples of users real trajectories and that does not have access to anonymized traces.

3 System Model

We present the assumptions regarding LBSs and the associated privacy threats.

3.1 Network Model

We study a network (Fig. 1) that involves mobile users equipped with wireless devices, third-parties running LBSs and a wireless infrastructure. Wireless devices feature localization technology such as GPS or wireless triangulation that lets users locate themselves. The geographic *location* of a user is denoted by $l = (lon, lat)$, where lon is the longitude, and lat is the latitude. The wireless infrastructure relies on technology such as WiFi, GSM or 3G to let users connect to the Internet. LBSs are operated by independent third-parties that provide services based on the location of mobile users.

Cellphone users send their location together with a service request to LBSs through the wireless infrastructure. For each request sent, users may identify themselves to the LBS using proper credentials. In general, we assume that users are identified with *pseudonyms* (i.e., fictitious identifiers), such as their username, their HTTP cookie or their IP address: some services may require users to register and provide the corresponding username and password, whereas others may use HTTP cookies to recognize users.

LBSs provide users with services using the location information from requests. LBSs store the information collected about their users in a database. As defined in [38], each location sample is called an *event* denoted by $< i, t, l >$, where i is the pseudonym of a user, t is the time instance at which the event occurred, and l is the location of the user. The collection of events from a user forms a *mobility trace*.

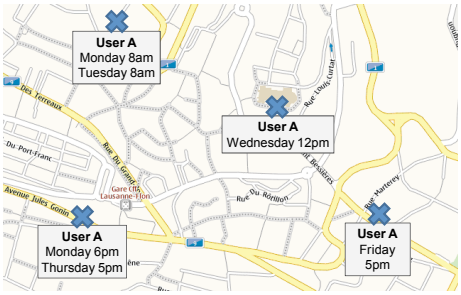


Fig. 2: Threat model. The LBS learns multiple locations of user A over several days. It can then infer the activities of user A and possibly obtain his true identity.

3.2 Threat Model

LBS operators passively collect information about the locations of pseudonymous users over time. For example, an LBS can observe location samples of user A over the course of weeks (Fig. 2). A priori, the LBS also knows the map over which users move and has access to geographic information systems that provide details such as points of interest in maps [9]. In addition, the LBS may use the increasing amount of location information available from other sources such as public transportation systems [36].

We consider that the LBS aims at obtaining the true identify of its users and their points of interest. To do so, the LBS studies the collected information. Even if communications are pseudonymous, the spatio-temporal correlation of location traces may serve as a *quasi-identifier* [5,6,10,19]. This is a significant threat to user privacy as such information can help the LBS to profile users.In this work, we investigate the ability of LBSs to identify users and their points of interest based on the collected information.

4 Privacy Erosion

In this section, we describe the process of *privacy erosion* caused by the use of LBSs. To do so, we first discuss how users tend to share their location with LBSs. Then, we explain how LBS operators can obtain the true identity of their users and their points of interest from the collected information.

4.1 Collection of Traces by LBSs

Depending on the provided service, LBSs collect location samples about their users. For example, typical services (such as Foursquare, Google Maps, Gowalla, or Twitter) offer users to connect with friends, to discover their environment or to optimize their mobility. Depending on their needs, users access such LBSs at different times and from different locations, thus revealing multiple location samples. In general, users of these services do not continuously share their locations. Instead, users have to manually decide to share their location with their mobile devices. We classify most popular LBSs into three broad categories and describe the information shared by users in each case.

Localized Search Many LBSs enable users to search for local services around a specific location (e.g., localized Google Search [7]). Localized searches offer mobile subscribers spontaneous access to nearby services. Hence, a user location acts as a spatial query to the LBS. For example, users can obtain the location of nearby businesses, products, events, restaurants, movie theaters or other local information depending on the type of information provided by the LBS.

Such localized searches help users navigate unfamiliar regions and discover unknown places. Thus, users episodically connect to LBSs, revealing samples of their mobility. LBSs obtain statistical information about the visited locations of mobile users and learn popular locations and user habits. Yet, LBSs do not learn the actual activity of mobile users (e.g., the name of the visited restaurant) as they do not know the decision of the user about the provided information.

Street Directions Another popular use of LBSs consists in finding a route between two locations. Typically, a user shares its location with an LBS and requests the shortest route to another location (e.g., Google Maps).

Users of such services usually reveal their current location and a potential destination. Hence, users may leak their home/work locations to LBSs in addition to samples of their mobility. This enables LBSs to obtain statistical information about the preferred origins and destinations of mobile users.

Location Check-ins A novel type of location-based service offers users to check-in to specific *places* in return for information related to the visited location [13]. For example, it can be used to check into shops, restaurants or museums. It allows users to meet other users that share similar interests and to discover new aspects of their city through recommendations [29].

With such services, users not only precisely reveal their location (GPS coordinates), but also their intention. Indeed, the LBS can learn the current activity of its users. Users can check-in to public places, but also private homes.

In summary, depending on the provided service, users share different *samples* of their mobility. In order to take this into account, in the following, we consider that LBSs obtain various *type* of location samples out of users' whereabouts.

4.2 Attacks by LBSs

The spatial and temporal information contained in mobility traces may serve as location-based quasi-identifiers [6,10]: an LBS may obtain the true identity and points of interests of its pseudonymous users from the collected mobility traces.

Location-Based Quasi-Identifiers Quasi-identifiers were introduced by Delenius [10] in the context of databases. They characterize a set of attributes that in combination can be linked to identify to whom the database refers (see [33] for more). In [6], Bettini *et al.* extend the concept to mobile networking. They consider the possibility to identify users based on spatio-temporal data and propose the concept of location-based quasi-identifiers. They describe how a sequence of spatio-temporal constraints may specify a mobility pattern that serve as a unique identifier. For example, as already mentioned,

Golle and Partridge [19] identify location-based quasi-identifiers by showing that home and work locations uniquely identify most of the US population. Hence, if an LBS learns users' home and work locations, it can obtain their identity with high probability.

In this work, we assume that the LBS succumbs to the temptation of finding the identify of its users. To do so, we consider that the LBS uses the home and work locations of users as location-based quasi-identifiers.

Inferring Home and Work Locations Previous works investigated the problem of characterizing and extracting important places from pseudonymous location data. These works propose various algorithms to infer important locations based on the spatial and temporal evidence of the location data. We group the existing works in two categories. In the first category [2,23,25], the authors use *clustering* algorithms to infer the *homes* of mobile users. For example in [25], Krumm proposes four different clustering techniques to identify the homes of mobile users in vehicular traces: traces are pseudonymous and contain time-stamped latitudes and longitudes. Similarly, in [23], Hoh *et al.* propose a k-mean clustering algorithm to identify the homes of mobile users in anonymous vehicular traces: traces do not have pseudonyms, but contain the speed of vehicles, in addition to their location. In the second category [27,28], Liao *et al.* propose *machine learning* algorithms to infer the different type of *activities* from mobile users data (e.g., home, work, shops, restaurants). Based on pedestrian GPS traces, the authors are able to identify (among other things) the *home and work locations* of mobile users.

We rely on previous work to derive an algorithm that exclusively infers the *home* and *work* locations of mobile users based on spatial and temporal constraints of pseudonymous location traces. The algorithm operates in two steps: first, it clusters spatially the events to identify frequently visited regions; second, it temporally clusters the events to identify home and work locations.

The spatial clustering of the events uses a variant of the k-means algorithm as defined in [2]: it starts from one random location and a radius. All events within the radius of the location are marked as potential members of the cluster. The mean of these points is computed and is taken as the new centre point. The process is repeated until the mean stops changing. Then, all the points within the radius are placed in the cluster and removed from consideration. The procedure repeats until no events remain. The number of points falling into a cluster corresponds to its weight and is stored along with the cluster location. Clusters with a large weight represent frequently visited locations.

Based on the output of the spatial filtering, the algorithm then uses temporal evidence as a criterion to further refine the possible home/work locations. In practice, users have different temporal patterns depending on their activities (e.g., students). The algorithm considers simple heuristics that apply to the vast majority of users. For example, most users spend the night at home and commute in the beginning/end of the day. In order to apply the temporal evidence, the algorithm considers all events in each cluster and labels them as home or work. Some events may remain unlabeled if they do not match any temporal criterion. The algorithm considers two temporal criteria. First, the algorithm checks the duration of stay at each location. To do so, it computes the time difference between the arrival of a trip at a certain location, and the departure time of the following trip. A user that stays more than 1 hour in a certain location over night is likely to have spent the night at home. Hence, the algorithm labels events occurring at

such location as home events. Second, the algorithm labels events occurring after 9am and before 5pm as possible work events. Finally, for each cluster, the algorithm checks the number of events labelled home or work and deduces the most probable home and work locations.

Inferring User Points of Interest Usually, LBSs use the content of queries to infer the points of interest of their users. Yet, LBSs may further profile users by analyzing the location of multiple queries and inferring users' points of interest.

We use the spatial clustering algorithm defined above to obtain the possible points of interest of users that we call *uPOIs*: a uPOI is a location regularly visited by a user. For each identified uPOI, we store the number of visits of the user and derive the probability P_v^i that a user i visits a specific uPOI, i.e., the number of visits to a uPOI normalized by the total number of visits to uPOIs.

Metrics The real home/work addresses are unavailable in our data sets. Hence, we apply our algorithm to the original mobility traces and derive a baseline of home/work locations. We then evaluate the probability of success of the LBS by comparing the baseline to the outcome of our algorithm on the *samples* of location data collected by the LBS. In other words, we compare the home/work location pairs predicted from the sampled traces with the baseline. In practice, it is complicated to obtain the real home and work locations of users (i.e., the baseline ground truth) in mobility traces without threatening their privacy. Because no real ground truth is available, this approach does not guarantee that we have identified the real home/work locations. Yet, it allows us to compare the effectiveness of the attack in various conditions.

The probability P_s of a successful identification by the LBS is then:

$$P_s = \frac{\text{Number of home/work pairs correctly guessed}}{\text{Total number of home/work pairs}} \tag{1}$$

This probability measures the ability of LBSs to find the home/work locations from sampled traces and thus uniquely identify users. This metric relies on the assumption that home/work location pairs uniquely identify users [19]: it provides an upper-bound on the identification threat as home/work location pairs may in practice be insufficient to identify users especially in the presence of uncertainty about the home/work locations.

We also evaluate the normalized *anonymity set* of the home and work pairs of mobile users. To do so, we compute the number of home/work locations that are in a certain radius from the home/work location of a certain user i. For every user i, we define its home location as h_i and its work location as w_i. For each user i, we have:

$$A^i_{home} = \frac{1}{|h|} \sum_{j \neq i} 1_{|h_j - h_i| < R_A} \tag{2}$$

$$A^i_{work} = \frac{1}{|w|} \sum_{j \neq i} 1_{|w_j - w_i| < R_A} \tag{3}$$

where R_A specifies the radius considered for the anonymity set.

We measure the ability of LBSs to infer uPOIs by considering for each user i, the number of uPOIs correctly inferred. For every user i, we have:

$$P^i_{uPOI} = \frac{\text{Number of uPOIs correctly guessed}}{\text{Number of uPOIs}} \tag{4}$$

We also use the notion of Kullback-Leibler divergence [26] to measure the ability of the adversary to guess the probability of each user visiting specific uPOIs. For every user i, we have:

$$D_{KL}(P^i_v||Q^i_v) = \sum_j P^i_v(j) \log \frac{P^i_v(j)}{Q^i_v(j)} \tag{5}$$

where P^i_v is the actual probability that user i visits specific uPOIs and Q^i_v is the probability guessed by the adversary.

5 Evaluation

We present our methodology to evaluate the erosion of privacy caused by LBSs.

5.1 Setup

We start from data sets of real mobility traces. The data sets contain the location of users at a high granularity. Because users usually reveal only a few location samples to LBS operators, we artificially reduce the information available to the LBSs by selecting a few events from the traces. Then, we consider various de-anonymization attacks on the location traces. In practice, we load mobility traces in Matlab and apply the algorithm described in Section 4.2. We repeat every analysis 100 times and consider the average.

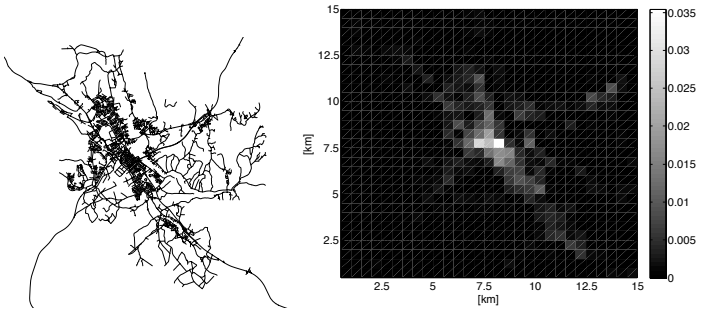
5.2 Mobility Traces

There exist several publicly available data sets of human mobility. For example, there are mobility traces of taxis [35], of student mobility in campus [12], or of sport activities [34]. Yet, most of these data sets have a limited applicability to our problem because the mobility of users is tied to specific scenarios (e.g, taxis, campus).

In this work, we consider two data sets representing normal activities of users in cities. These mobility traces contain several *trips* for each user. A trip defines a trajectory of a user going from one source location to a destination (e.g., a user commuting from home to work). Users move on a map following road constraints.

Borlange Data Set The city of Borlange is a middle-sized (15×15km^2) Swedish city of approximately 46000 inhabitants. Borlange has 3077 road intersections interconnected by 7459 roads (Fig. 3 (a)). The data set was collected over two years (1999-2001) as part of an experiment on traffic congestion that took place there.[1] About 200 private *cars* (with one driver per car) within a 25 km radius around the city center were equipped with a GPS device. At regular intervals (approximately every 5 seconds), the position, time and speed of each vehicle was recorded and stored. Mostly because of GPS accuracy issues, many observed trips did not match the Borlange map. The data was thus manually verified and corrected using road fitting algorithms for a subset of 24 vehicles resulting in a total of 420814 "clean" trips (see [14] for more details). This data set was obtained by civil engineers and used to analyze the route choices of mobile users.

[1] The data set is available at http://icapeople.epfl.ch/freudiger/borlange.zip .



(a) (b)

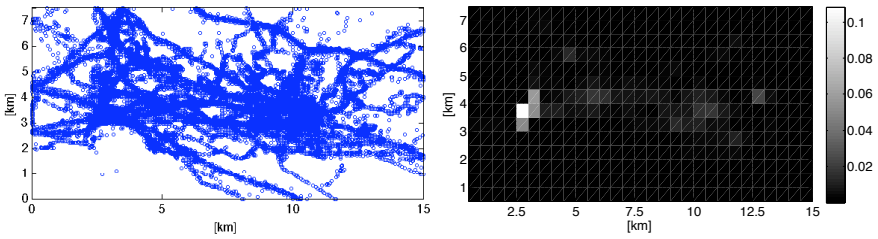
Fig. 3: Borlange data set. (a) Map of Borlange, Sweden. The city has 46000 inhabitants and spreads over $15 \times 15\text{km}^2$. (b) Spatial histogram showing the density of users per cell $c(z)$.

Lausanne Data Set The Lausanne area in Switzerland is a region of $15 \times 7\text{km}^2$ of approximately 120000 inhabitants (Fig. 4 (a)). In September 2009, Nokia began running a data collection campaign in Lausanne area. Around 150 *users* are equipped with GPS-enabled Nokia phones that record their daily activities and upload them on a central database. Among other things, the phones measure the GPS locations of users at regular intervals (approximately every 10 seconds). In July 2010, we took a snapshot of the database containing traces of 143 users tracked over 12 months.[2] Note that the database contains traces of pedestrians, but also of users in cars, buses and trains. It has thus a larger diversity in terms of mobility patterns than the Borlange data set. We focus on the traces that start and finish in the Lausanne area and obtain around 106600 trips.

In order to evaluate the statistical relevance of the mobility traces, we compute statistics of mobility in the data sets. We divide the whole region of Borlange/Lausanne into square cells of equal size (500×500m). and evaluate the distribution of users' visits in each cell. We define a variable C_z that counts the number of events among all users that happen in each cell z. For each cell, we compute the empirical probability that an event falls into the cell z, $c(z) = \frac{C_z}{\sum_x C_x}$. In Figure 3 (b), we show the density map (i.e., the set of cells with their corresponding $c(z)$) for the Borlange data set. We observe that the activity of users is concentrated in a few regions. We observe a similar distribution in the Lausanne data set (Fig. 4 (b)). Yet, in the latter, there is a small bias towards one location (the EPFL campus), indicating that many users from the experiment share the same work place. The Lausanne data set reflects scenarios in which many users share the same work place, for example, downtown of a large city.

In Figure 5, we show the empirical Cumulative Distribution Function (CDF) of $c(z)$ for both data sets in semi-log scale. We observe that the CDF increases linearly, indicating a heavy-tailed distribution of user density. This confirms that some cells have a density much above the average. Our observations about the heavy-tailed distribution match existing results in the literature on mobility traces [8,34,41] and confirm the statistical relevance of the data sets. Intuitively, the heavy-tailed distribution may indicate that users are easily identifiable as they share few locations.

[2] The data set is not publicly available.



(a) (b)

Fig. 4: Lausanne data set. (a) Map of Lausanne area, Switzerland. The city has 120000 inhabitants and spreads over $15 \times 7\text{km}^2$. (b) Spatial histogram showing the density of users per cell $c(z)$.

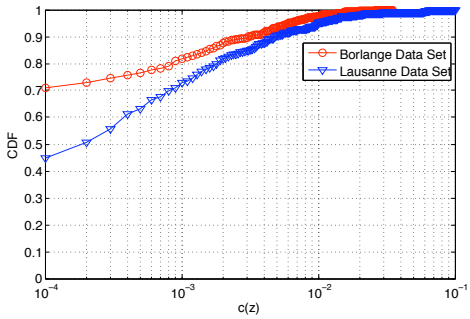


Fig. 5: Empirical CDF of $c(z)$ in semi-log scale. We observe a linear behavior for both data sets indicating a heavy-tailed distribution of user density in the network.

5.3 Modeling the Collection of Traces by LBSs

We start from mobility traces containing location samples at high granularity. As described in Section 4.1, the *type* and *quantity* of location information collected by LBSs depends on the services and their usage. To take this into account, we select a few events from the entire traces in various ways. Each selected event effectively represents a *query* to LBSs.

Uniform Selection (UF) We select events uniformly at random from the set of all possible events of each user. This captures scenarios in which users are likely to use an LBS anytime and anywhere.

Home/Work Selection (HW) We distinguish between three types of events: *home*, *work* and *miscellaneous*. Home and work events refer to queries made from home and work, respectively, whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples uniformly in each set corresponding to *home/work* events with probability ρ and miscellaneous events with probability $1-\rho$. A large ρ captures scenarios in which users access LBSs mostly from home and work (e.g., street directions), whereas a small ρ captures scenarios in which users access LBSs mostly on the go (e.g., localized search or location check-ins).

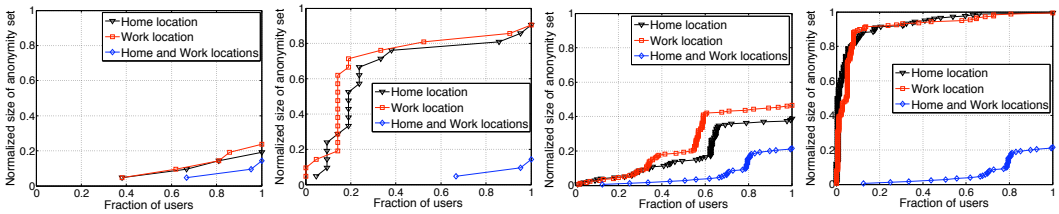


Fig. 6: Normalized size of anonymity set A^i_{home}, A^i_{work} and $A^i_{homeWork}$. Borlange with (a) $R_A = 1$km and (b) $R_A = 5$km. Lausanne with (c) $R_A = 1$km and (d) $R_A = 5$km.

Points of Interest Selection (PO) We distinguish between two types of events: *cPOIs* and *miscellaneous*. cPOI events refer to queries made from regions of a city with many points of interest (e.g., POIs of the city), whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples uniformly in each set corresponding to cPOI events with probability ρ. A large ρ captures scenarios in which users access LBSs mostly from popular locations (i.e., localized search), whereas a small ρ captures scenarios in which users access LBSs mostly in unpopular areas such as residential areas.

Preferred Selection (PF) We distinguish between two types of events: *preferred* and *miscellaneous*. Preferred events refer to queries made from locations frequently visited by each user (i.e., uPOIs), whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples corresponding to preferred events with probability ρ. A large ρ captures scenarios in which users access LBSs mostly during their routine, whereas a small ρ captures scenarios in which users access LBSs mostly in unfamiliar areas.

We tune the selection type using probability ρ. Note that home/work selection strategy with $\rho = 0.5$ is different from the uniform selection strategy: with $\rho = 0.5$ in home/work selection, home/work events and miscellaneous events have the same probability to be chosen, whereas with the uniform selection, all events have the same probability to be chosen. We consider various number of queries λ in order to model the quantity of data collected by LBSs. For example, a number of queries $\lambda = 60$ means that 60 samples of all location samples of each user are shared with the LBS.

5.4 Results

Unless otherwise stated, we consider that users share their location with the LBS with a 10 meters precision (i.e., GPS), that the clustering radius in the spatial clustering algorithm is 100 meters and that the adversary has a tolerable error margin of 50 meters to correctly guess a home/work/uPOI locations.

Size of Anonymity Set The graphs in Fig. 6 detail the size of the anonymity set for home locations, work locations, or both normalized with the number of users in the data

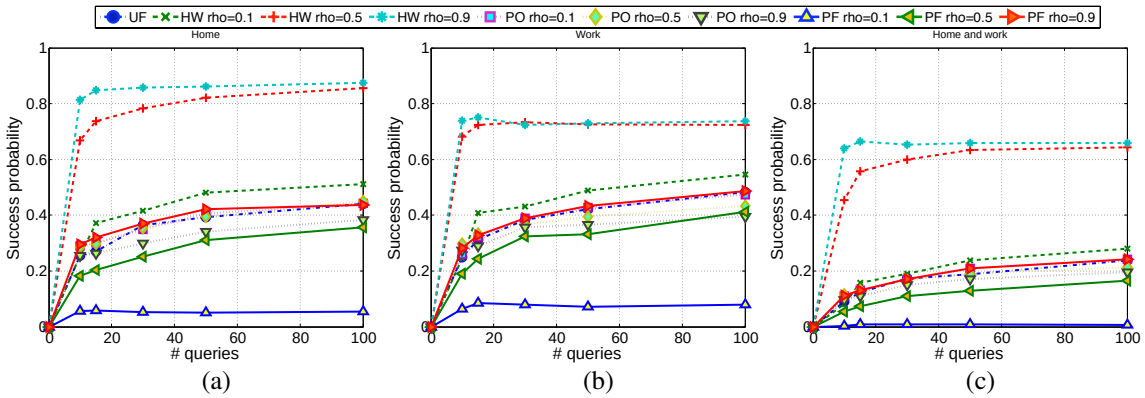


Fig. 7: Privacy erosion in Borlange with varying selection probability ρ and number of queries λ. (a) Home identification. (b) Work identification. (c) Home and work identification (P_s).

set. On the x-axis, the graphs show the fraction of users that has an anonymity set of less than a given normalized size on the y-axis. We consider two radius $R_A = 1$km and $R_A = 5$km. As predicted in [19], the anonymity set size is low especially when a small radius is used and revealing the home and work locations is much more identifying than only revealing one of them. In general, we observe that more users share a common work place than home. In the Lausanne data set, many users have a larger anonymity set than in the Borlange data set due to the larger number of users.

Privacy Erosion We evaluate the privacy erosion of users from the Borlange and Lausanne data sets in multiple scenarios. We measure the probability that an LBS successfully identifies the home location, the work location, or both. In the case of a successful home and work identification, the LBS successfully identifies its users. We consider different data collection scenarios as described earlier (UF, HW, PO and PF) with three selection probabilities: $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. We also vary λ, the amount of information shared with LBSs.

In Fig. 7 and Fig. 8, we show the erosion of privacy in Borlange and Lausanne for various ρ, λ and selection strategies. We observe that with HW selection, the probability of identification of a home, work, or home/work pair increases the fastest with respect to the number of sent queries indicating that LBSs uniquely identify users with few locations: in Borlange, if $\rho = 0.9$, 20 queries are sufficient to identify 65% of users. We observe that as ρ increases, so does the identification success. In Lausanne, the identification success is slightly higher but still leads to the same conclusions. We observe that PF selection with $\rho = 0.1$ makes de-anonymization particularly difficult. In this case, users share their location only in unfamiliar areas and it is thus difficult for LBSs to infer users' identity. For other selection strategies, the identification success saturates around 20 to 40% and increases slowly with the number of queries.

Inferred User Points of Interest Table 1 shows the average fraction of visits to uPOIs. Each uPOI identifies a region of 200 meters radius frequently visited by each user. In both data sets, the distribution is long tail showing that few uPOIs are frequently visited.

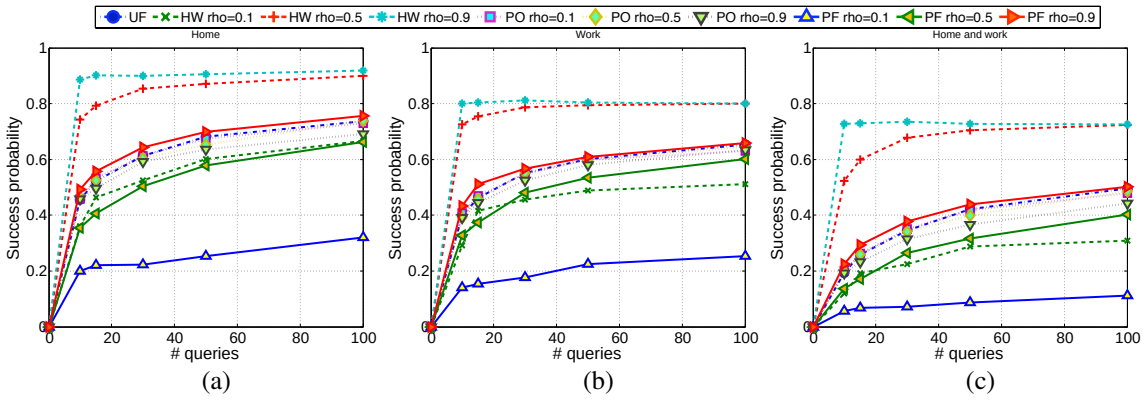


Fig. 8: Privacy erosion in Lausanne with varying selection probability ρ and number of queries λ. (a) Home identification. (b) Work identification. (c) Home and work identification (P_s).

Table 1: Average probability $E[P_v^i]$ of visiting a uPOI.

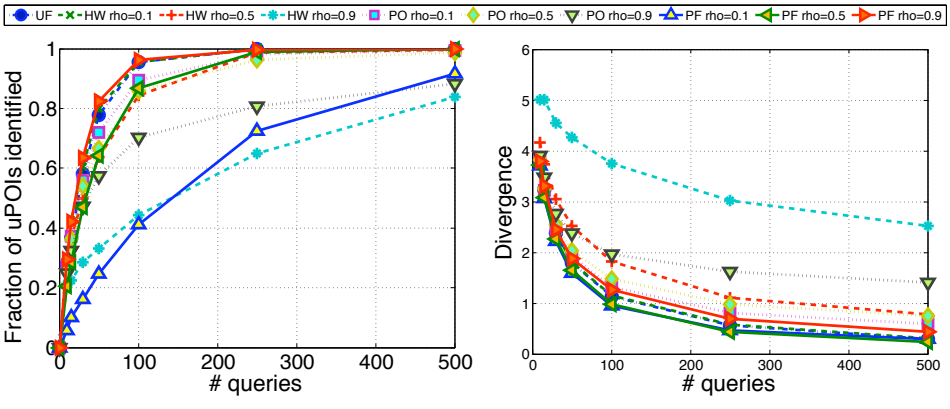
Data Set	**uPOIs**									
	1	2	3	4	5	6	7	8	9	10
Borlange	0.357	0.209	0.112	0.078	0.052	0.031	0.021	0.017	0.015	0.012
Lausanne	0.401	0.14	0.092	0.063	0.045	0.035	0.028	0.023	0.019	0.016

In Figure 10, we show the ability of LBSs to infer the top ten uPOIs of each user: we compute the average fraction of uPOIs identified $E[P_{uPOI}^i]$ within a 100 meters error margin and evaluate the average divergence $E[D_{KL}(P_v^i||Q_v^i]$. We observe that the adversary can infer a large number of uPOIs with a small number of samples: with 30 samples, it can learn up to 65% of uPOIs in the case of PF $\rho = 0.9$. The best selection strategies are PF $\rho = 0.9$, HW $\rho = 0.1$ and UF. Intuitively, revealing preferred visited locations reveals clusters, similarly, uniform across visited locations will have high probability to sample from frequently visited location. On the contrary, with PO $\rho = 0.9$, HW $\rho = 0.9$ or PF $\rho = 0.1$, the attack works less efficiently. Hence, even with a few location samples, the adversary is also able to infer most uPOIs.

In terms of divergence, a divergence of zero indicates a perfect match. We observe that the divergence decreases fast indicating that the adversary obtains a probability distribution similar to the true one and identifies the most probable uPOIs. We observe a similar behavior with the Lausanne data set. Note that the ability to infer uPOIs is at odds with the ability to infer users' identity: with HW $\rho = 0.9$, it is harder to identify uPOIs and easier to identify users.

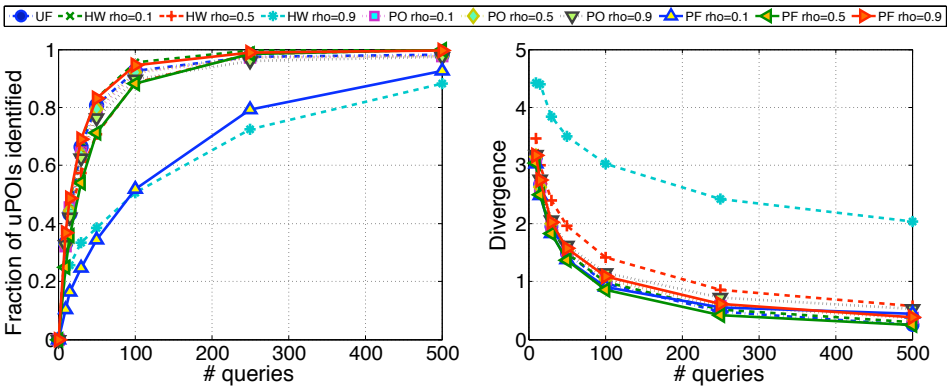
6 Conclusion

We have considered the problem of privacy erosion when using location-based services. We identify the quantity and type of location information that statistically helps LBSs find users' real identity and points of interest. In contrast with previous work (mostly showing that de-anonymization based on location information is possible), we push the understanding of the threat further by showing how de-anonymization depends on the collected data. We experiment with two real data sets of mobility traces, model the col-



(a) (b)

Fig. 9: Inferring the top ten uPOIs in Borlange data set. (a) Average fraction of uPOIs identified $E[P^i_{uPOI}]$. (b) Average divergence $E[D_{KL}(P^i_v||Q^i_v)]$.



(a) (b)

Fig. 10: Inferring the top ten uPOIs in Lausanne data set. (a) Average fraction of uPOI identified $E[P^i_{uPOI}]$. (b) Average divergence $E[D_{KL}(P^i_v||Q^i_v)]$.

lection of traces by LBSs and implement various attacks. Our results show that in many scenarios a small amount of information shared with LBSs may enable to uniquely identify users. These results stem from the fact that the spatio-temporal correlation of location traces tends to be unique to individuals and persistent. We also show that in some scenarios, users have high privacy without using privacy-preserving mechanisms.

The results of this work can help prevent the false sense of anonymity that users of LBSs might have by increasing the awareness of location privacy threats. In particular, it may encourage users to stop revealing sensitive information to third-parties, such as their home and work locations, and adopt privacy-preserving mechanisms. These results notably question the ability of privacy-preserving mechanisms to obfuscate highly correlated information such as users' whereabouts. These results can thus help design more efficient privacy-preserving mechanisms [15,16,17] and may also encourage the use of distributed solutions in which users store maps and the related information directly on their mobile devices.

References

1. A. Aki. The discovery of a lifetime. http://www.aka-aki.com.
2. D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.
3. R. Barnes, A. Cooper, R. Sparks, and C. Jennings. IETF geographic location/privacy. http://www.ietf.org/dyn/wg/charter/geopriv-charter.html.
4. A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *Pervasive Computing, IEEE*, 2(1):46–55, 2003.
5. A. R. Beresford and F. Stajano. Mix zones: User privacy in location-aware services. In *PerSec*, March 2004.
6. C. Bettini, X. S. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. In *SDM*, 2005.
7. G. M. Blog. Finding places "near me now" is easier and faster than ever, 2010. http://googlemobile.blogspot.com/2010/01/finding-places-near-me-now-is-easier.html.
8. A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE TMC*, 6:606–620, 2007.
9. Cloudmade. Makes maps differently. http://cloudmade.com.
10. T. Dalenius. Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.
11. C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *PET*, 2002.
12. N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. In *National Academy of Sciences (PNAS)*, pages 15274–15278, 2009.
13. Foursquare. Check-in, find your friends, unlock your city. http://foursquare.com.
14. E. Frejinger. *Route choice analysis : data, models, algorithms and applications*. PhD thesis, EPFL,, 2008.
15. J. Freudiger, M. Manshaei, J.-Y. L. Boudec, and J.-P. Hubaux. On the age of pseudonyms in mobile ad hoc networks. In *Infocom*, 2010.
16. J. Freudiger, M. Manshaei, J.-P. Hubaux, and D. C. Parkes. On non-cooperative location privacy: A game-theoretic analysis. In *CCS*, 2009.
17. J. Freudiger, R. Shokri, and J.-P. Hubaux. On the optimal placement of mix zones. In *PETs*, 2009.
18. G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geotagging. In *HotSec*, 2010.
19. P. Golle and K. Partridge. On the anonymity of home/work location pairs. In *Pervasive*, 2009.
20. M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *MobiSys*, 2003.
21. B. Hoh and M. Gruteser. Protecting location privacy through path confusion. In *SECURECOMM*, 2005.
22. B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *MobiSys*, 2008.
23. B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *Pervasive Computing*, pages 38–46, 2006.
24. B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in GPS traces via uncertainty-aware path cloaking. In *CCS*, 2007.
25. J. Krumm. Inference attacks on location tracks. In *Pervasive*, 2007.
26. S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
27. L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational Markov networks. In *IJCAI*, 2005.
28. L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, (171):311–331, 2007.
29. Loopt. Discover the world around you. http://loopt.com.

30. C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao. Privacy vulnerability of published anonymous mobility traces. In *MobiCom*, 2010.
31. M. F. Mokbel, C.-Y. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, 2006.
32. Y. D. Mulder, G. Danezis, L. Batina, and B. Preneel. Identification via location-profiling in GSM networks. In *WPES*, 2008.
33. A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy*, 2009.
34. M. Piorkowski. Sampling urban mobility through on-line repositories of GPS tracks. In *HotPlanet*, 2009.
35. M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *ComsNets*, pages 1–10, 2009.
36. C. Roth, S. M. Kang, M. Batty, and M. Barthelemy. Commuting in a polycentric city. Technical report, CNRS, 2010.
37. A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In *PET*, 2002.
38. R. Shokri, J. Freudiger, and J.-P. Hubaux. A unified framework for location privacy. In *HotPETs*, 2010.
39. R. Shokri, J. Freudiger, M. Jadliwala, and J.-P. Hubaux. A Distortion-based Metric for Location Privacy. In *WPES*, 2009.
40. L. Sweeney. k-anonymity: A model for protecting privacy. *Uncertainty, Fuzziness and Knowledge-based systems*, 10:557–570, 2002.
41. M. Vojnovic and J.-Y. L. Boudec. Perfect simulation and stationarity of a class of mobility models. In *Infocom*, 2005.
42. S. Zhong, L. E. Li, Y. G. Liu, and Y. R. Yang. Privacy-preserving location-based services for mobile users in wireless networks. Technical report, SUNY at Buffalo, 2005.