Methods for Location Privacy: A comparative overview

Kostantinos Chatzikokolakis
CNRS, École Polytechnique, University of Paris Saclay, France

Ehab ElSalamouny
Faculty of Computers and Informatics, Suez Canal University, Egypt

Catuscia Palamidessi
INRIA, University of Paris Saclay, France

Anna Pazii
INRIA, University of Paris Saclay, France

Full text available at: http://dx.doi.org/10.1561/3300000017
Editorial Scope

Topics

Foundations and Trends® in Privacy and Security publishes survey and tutorial articles in the following topics:

- Access control
- Accountability
- Anonymity
- Application security
- Artificial intelligence methods in security and privacy
- Authentication
- Big data analytics and privacy
- Cloud security
- Cyber-physical systems security and privacy
- Distributed systems security and privacy
- Embedded systems security and privacy
- Forensics
- Hardware security
- Human factors in security and privacy
- Information flow
- Intrusion detection
- Malware
- Metrics
- Mobile security and privacy
- Language-based security and privacy
- Network security
- Privacy-preserving systems
- Protocol security
- Security and privacy policies
- Security architectures
- System security
- Web security and privacy

Information for Librarians

Foundations and Trends® in Privacy and Security, 2017, Volume 1, 4 issues. ISSN paper version pending, ISSN online version pending. Also available as a combined paper and online subscription.
Methods for Location Privacy: A comparative overview

Kostantinos Chatzikokolakis  
CNRS, École Polytechnique,  
University of Paris Saclay, France

Ehab ElSalamouny  
Faculty of Computers and Informatics,  
Suez Canal University, Egypt

Catuscia Palamidessi  
INRIA,  
University of Paris Saclay, France

Anna Pazii  
INRIA,  
University of Paris Saclay, France
## Contents

1 The problems of privacy in location-based services 2
   1.1 Classification of threats 4
   1.2 Identification of the user from his traces 5
   1.3 The users’ point of view 12

2 Deterministic methods 14
   2.1 Deterministic Spatial Obfuscation 15
   2.2 Deterministic Spatial Cloaking 15
   2.3 Criticism of the spatial cloaking approach 26

3 Randomized methods 29
   3.1 Differential Privacy 29
   3.2 Protection of identity 31
   3.3 Protection of location 33

4 Conclusion 47

Acknowledgements 48

References 49
Abstract

The growing popularity of Location-Based Services, allowing for the collection of huge amounts of information regarding users’ locations, has started raising serious privacy concerns. In this survey we analyze the various kinds of privacy breaches that may arise in connection with the use of location-based services, and we consider and compare some of the mechanisms and the metrics that have been proposed to protect the user’s privacy, focusing in particular on the comparison between probabilistic spatial obfuscation techniques.
In recent years, the growing popularity of mobile devices equipped with GPS chips, in combination with the increasing availability of wireless data connections, has led to a growing use of Location-Based Services (LBSs), namely applications in which a user obtains, typically in real-time, a service related to his current location. Recent studies of the Pew Research Center show that in 2017, 77% of the adult population of the US owns a smartphone (in comparison with 35% in 2011) [63], and according to the same institution’s last survey about LBSs, in 2013, a high percentage (74%) of the smartphone owners used services based on their location [99]. Examples of LBSs include mapping applications (e.g. Google Maps), Points of Interest (POI) retrieval (e.g. AroundMe), coupon/discount providers (e.g. GroupOn) and location-aware social networks (e.g. Foursquare).

LBS providers often collect and store users’ locations and mobility traces (sequences of spatio-temporal points representing the users’ itineraries), for the purpose of further utilization, possibly by a third-party. For instance, they can be used for statistical analyses, such as finding typical mobility patterns and popular places [74, 97], or they can be made public to provide additional services to users, such as traffic information [44].
While LBSs have demonstrated to provide enormous benefits to individuals and society, the growing exposure of users’ location information raises important privacy issues. Not only the experts, but also the population at large are becoming increasingly aware of the risks, due to the repeated cases of violations and leaks that keep appearing on the news. For instance, on April 20th, 2011 it was discovered that the iPhone was storing and collecting location data about the user, syncing them with iTunes and transmitting them to Apple, all without the user’s knowledge. More recently, the Guardian has revealed, on the basis of the documents provided by Edward Snowden, that the NSA and the GCHQ have been using certain smartphone apps, such as the wildly popular Angry Birds game, to collect users’ private information such as age, gender and location [6], again without the users’ knowledge. Another case regards the Tinder application, which was found sharing the exact latitude and longitude co-ordinates of users as well as their birth dates and Facebook IDs [73]; even after the initial problem was fixed, it was still sharing more accurate location data than intended, as users could be located to within 100 feet of their present location [26].

A major source of concern about location privacy lies in the realization that with sufficiently accurate data, it is possible to precisely locate a user and track his movements throughout the day [18], giving rise to a variety of malicious activities such as robbing or stalking. For instance, in Wisconsin there were episodes of men tracking women with GPS or other location devices [60]. In California, records from automatic toll booths on bridges were used in divorce proceedings to prove claims about suspicious movements of spouses [82]. The application “Girls Around Me”, combined social media and location information to find nearby women who did not necessarily agree to be found, allowing to access their Facebook profiles with a single click [11]. Particularly worrisome is the perspective of potential combination with the users’ most sensitive information, such as sexual orientation.

To some extent, the research and the experimentation on privacy contribute to raise the awareness about the practical risks. For instance, the website “Please Rob Me” [65] aggregates location check-ins and
The problems of privacy in location-based services

presents them as “robbery opportunities”, pointing out the fact that publicly announcing one’s location effectively reveals to the world that they are not home.

1.1 Classification of threats

Following [35], we classify the concerns about the leakage of location information into three major kinds of threats:

Tracking Threat: An adversary collecting continuously the location updates of the user might be able to identify the user’s mobility patterns (frequently traveled routes) and predict his present and future location with high accuracy by leveraging typical mobility habits [47, 94].

Identification Threat: The adversary can use the user’s traces as quasi-identifiers to reveal his identity in an anonymized dataset. This may happen even if the adversary accesses the user’s location only sporadically, since he might be able to infer his frequently visited locations, such as home and work. This is the most studied kind of threat in the literature, we expand on it in the next section.

Profiling Threat: Mobility traces, and in particular the points of interest that can be extracted from them, typically contain semantic information that the adversary can use for profiling, that is for inferring a variety of (often sensitive) information about the user. Examples include health clinics, religious places, areas which may revel his sexual inclinations, etc. [5]. The practice of location profiling is likely to increase in the future, as marketers are becoming more and more aware of its potential to gain visibility of consumer behavior in the real world, and to help targeting their marketing efforts. Indeed, location profiling seems to provide insights into offline activity at a level comparable to that of web or mobile app analytics for online activity. There are already various companies that provide this kind of services: for instance, Urban Airship [89] offers tools that produce audience profiles by
1.2 Identification of the user from his traces

combining in-app behaviors, user preferences, and location. Mobility data are particularly useful, since brands can segment users based on their current or past location.

1.2 Identification of the user from his traces

In this section we focus on the threat constituted by using location data for fingerprinting the user, namely for finding out the identity of the person who has originated the data. In short, the problem raises by the fact that mobility traces may be unique to an individual, and they can therefore allow identifying that individual like the ridges on his finger. Apart from uniqueness, temporal correlation is also crucial for fingerprinting, allowing an anonymized trace to be identified based on mobility data about the same individual that have been previously recorded.

1.2.1 Uniqueness of human mobility traces.

There have been various statistical studies aimed at showing the uniqueness of human mobility traces. One of the most remarkable ones is that of de Montjoye et al. [23], measuring uniqueness in the following way. Given a set of points $P$, and a set of traces $T$, we say that $P$ identifies a unique trace in $T$ if there is exactly one trace in $T$ that contains $P$. Then, the uniqueness of $T$ is defined as the percentage of traces in $T$ that are uniquely identified by a set of $n$ points drawn randomly from a random trace in $T$ (where $n$ is a parameter). They examined fifteen months of human mobility traces generated by 1.5 million of individuals, who were users of a certain mobile phone operator. Each time a user interacted with the network by initiating or receiving a call or a text message, the location of the connecting antenna was recorded in the dataset together with the time of the event, and linked to previous location-time points of the same user already in the dataset, via the user id, so to form a trace (one trace for each user). The experiments showed that human mobility traces are highly unique: In fact, with the temporal granularity fixed to an hour, and the spatial granularity equal to that given by the carrier’s antennas, 4 spatio-temporal points, ran-
The problems of privacy in location-based services

domly drawn from a trace, were enough to uniquely identify the trace in 95% of the cases. They also observed that the uniqueness of mobility traces decays approximately as the $1/10$ power of the spatial and temporal resolution. Hence, they concluded that even coarse datasets provide little anonymity.

Song et al. [83] conducted similar experiments on a dataset of location-time data generated by about a million users over a period of a week. They considered the same notion of unique identification as de Montjoye et al., except that they calculated the percentage of identification on all the traces instead than some randomly drawn subset. The location of each individual was recorded every fifteen minutes. The spatial resolution of the data (i.e., the minimum distance between two locations) was about 0.11 km, while the diameter of the whole area (i.e., the largest distance between two locations) was about 49 km. Their results confirm that, even with a low resolution, location traces can be identified with only a few spatio-temporal points. In particular, they show that 2 points are enough to uniquely identify a trace in 60% of the cases.

It is important to note that the implicit notion of attack considered in the above works presupposes that the adversary is provided with points that he had “previously seen” in a trace, and the only challenge (for the adversary) is to be able to distinguish which trace. In contrast, Rossi et al. [68] considered the threat posed by a “previously unseen” set of points. Namely, they assume that the attacker has already collected a set of traces $T$ from some community of users, one trace per user, and then, given a set of additional points $P$ produced by one of the users during his trajectory, they try to re-identify the user by looking for the closest trace, namely the trace in $T$ with the smallest Hausdorff distance from $P$. They experimented with three real-world datasets GPS mobility traces: CabSpotting [64], CenceMe [59] and GeoLife [55]. The location data in these datasets have high spatial resolution (GPS coordinates up to 5 or 6 digits precision). As for the temporal resolution, in GeoLife and CabSpotting locations are recorded

\[\text{Although [68] refers to CabSpotting, the citation is relative to a mobility traces dataset called CRAWDAD.}\]
1.2. Identification of the user from his traces

at a time interval of 1 – 5 seconds, while for CenceMe it is 1 hour. Concerning the experiments methodology, they randomly partitioned each dataset into a training set and a test set, where each trace contained 50% of the original GPS points. Then, they used the training set as the traces $T$ to identify, using sets of points $P$ extracted randomly from traces in the test set. They showed that, thanks to the high precision of the GPS coordinates, on GeoLife and CenceMe just 1 spatio-temporal point is enough to identify 90% and 96% of the traces, respectively. With 2 points, these percentages reach 94% and 99%. The results for CabSpotting are significantly lower: 60% for 2 points. The difference is probably due to the nature of the data: GeoLife and CenceMe contain traces left by users during their daily routines, while CabSpotting are traces of taxi drivers in the San Francisco Bay area. The first two contain many personal and thus unique locations, such as home and workplace locations, while the latter is characterized by the presence of common taxi routes and locations associated to taxi ranks.

1.2.2 Reconstructing traces from location samples

Typically, there can be various users repeatedly updating and sending their positions on the map to some LBS. Hence, collecting these locations may result in a mix-up of traces left by different individuals. Un-mixing the locations, i.e., reconstructing the individual traces, can be done easily when the data are associated to some invariant attribute, like, for instance, a pseudonym. Even when the data are completely anonymous, however, the traces can often still be reconstructed by linking the location samples. Clearly, the higher is the sample frequency compared to the users’ density in the area, the easier it is to recognize a trace (cfr. Figure[1.1]). In fact, the next point in a trajectory will be at a distance determined by the speed of the user and the time in between the two updates. The reconstruction of a trace can also be facilitated by correlating location samples with likely routes on a map. Finally, the task can be enhanced by using a model of typical trajectories constructed on the basis of prior observations on the population movements.
The first attempt to reconstruct the traces from completely anonymized mobility data (i.e., without any pseudonyms) was by Gruteser et al. [42]. They used a multi-tracking algorithm to identify individual mobility traces from a collection of anonymized location samples generated by multiple users. They tested their algorithm on a collection of GPS traces generated by the students of a university campus, and their experiments showed that often individuals used to travel along the same unique route and could therefore be re-identified. Their system however was prone to misclassification of crossing paths, as it was unable to determine whether the paths of different individuals actually crossed or just touched.

More recently, Tsoukaneri et al. [87] developed a mechanism called Comber which is able to disentangle the traces by using a generic, empirically derived histogram of user speeds. The authors evaluated Comber with two different datasets, MDC [45] and GeoLife [55], which consist both of GPS-based mobility traces (collected in Lausanne and Beijing, respectively). Each of these datasets span more than a year and include location information of about 180 users. Their results show that Comber is able to infer the original traces of the users with more than 90% accuracy.

1.2.3 Linking traces to users’ identity

There has been a lot of research showing that it is possible to infer user identities from anonymous traces, especially when the traces are pseudonymized (i.e. the real identity has been replaced by a pseudonym) rather than completely anonymized. Beresford and Stajano [8] already pointed out that the re-identification risks of LBS’ users
1.2. Identification of the user from his traces

employing pseudonyms: they showed that almost all location traces of AT&T Labs Cambridge employees collected from the Active Bat system could be correctly identified by knowing the office positions of the workers and by keeping track of the frequency of visits of a given pseudonym to each office.

Many of the attacks on pseudonymized traces are, like the above, based on observing the frequent presence of the pseudonyms in specific locations that can be easily linked to a certain individual, like home or office. For instance, Krumm [48] proposed various algorithms to infer the user’s home address, and used a web search engine in order to reveal the real identities of the subjects. Notably, Golle and Partridge [40], using US census data, showed that knowing both locations of an individual’s home and workplace with the precision of a census block allowed to uniquely identify most of the U.S. working population. Furthermore, even with the lower granularity of a census track, although the average size of the anonymity set (i.e., the number of people sharing the same pair) went up to 21, the location data of people who lived and worked in different regions could still be easily re-identified.

A further study [96] investigated call records rather than census data, using a data set of more than 30 billion call records made by 25 million cell phone users in the US. They considered the “top N” locations for each user, inferred from the call records, and different levels of granularity, ranging from a cell sector to whole cell (where cell and cell sector are location units used by the phone company) to the zip code, city, county and state. They analysed a variety of different factors potentially impacting the size of the anonymity set, such as the distance between the top N locations, the geographic environment (rural vs urban), and social information (whether the size of the user’s social network is large or small). Their result showed that, while the top 1 location does not typically yield small anonymity sets, the top 2 and top 3 locations do, at least at the sector or cell-level granularity. For example, with top 3 locations, 85% of the users are identifiable at the sector level, 50% at the cell level, and 35% at the zip code level.

Even when the location data are completely anonymized (i.e., no pseudonym is used), though, it is still possible to retrieve the user’s
The problems of privacy in location-based services

identity by means of modern machine learning technologies if the attacker disposes of side information about the user. Several works in the literature have investigated this problem, particularly in the case in which a database of users’ profiles in the form of previously collected traces, called the training set, is available to the adversary. The work by Rossi et al. [68] mentioned in § 1.2.1 went in this direction; however it did not use the full power of machine learning techniques, and it was more focused in the uniqueness of traces rather than re-identification of the user. In general, the idea is that the adversary will use the training set to build a representation of the users’ typical movements. Thus each user will be associated to a mathematical model of his past traces, playing the role of a signature. This model can be, for instance, a Markov chain, but other models have been investigated as well. Then the attacker will collect one or more of the victim’s (sanitized) traces, the testing set, from which he will build a model as well. The latter is then compared to the models of the training set, according to some similarity criterion, and the user profile most likely to correspond to the target user is finally selected.

De Mulder et al. [24] investigated this kind of attack on mobility traces generated by a GSM cellular network. They developed two methods based on different models and on the cosine similarity measure, and evaluated them on the Reality Mining dataset made available by the MIT Media Lab, which consists of the location traces of one hundred human subjects at MIT during the 2004–2005 academic year, collected using one hundred instrumented Nokia 6600 smart phones. With the best of those methods, they were able to re-identify about 80% of the users. It is to be noted that a trace generated by a GSM network is formed by the sequence of all cells that the user has visited along his path, i.e., it is not possible to skip cells by “jumping” to a non-adjacent cell. This may affect the success rate when compared with the case in which the traces consist of locations generated dynamically with, say, a GPS.

Ma et al. [52] considered also two kinds of adversaries: passive ones, retrieving the testing set from a public source, and active ones that can deliberately participate or perturb the data collection phase to gain ad-
1.2. Identification of the user from his traces

The authors used four different estimators to measure the similarity between mobility traces: the Maximum Likelihood Estimator, relying on the Euclidean distance, the Minimum Square Approach, computing the sum of the square of the difference between the traces, the Basic Approach, which assumes that the traces might been perturbed by uniform noise, and the Weighted Exponential Approach, which is similar to the previous one except that no assumption is made on the type of noise generated. The authors tested their methods on two datasets: the CRAWDAD repository [64], recording the movements of San Francisco YellowCabs, and a collection of traces generated by the public buses in Shanghai city. They obtained a success rate of de-anonymization of 80% to 90%, even in the presence of noise.

Both [52] and [24], however, took the samples to generate the testing set directly from the training set. Clearly such way of proceeding introduced a bias that may have resulted in an overly strong success rate in the re-identification results. In fact Gambs et al. [36] showed that there is a substantial difference in the success rate when the training set and the testing set are separated. They used a model based on Mobility Markov Chains, namely Markov chains where the states are locations. They considered various similarity measures between such chains, and tested their methods on several GPS datasets, including MDC and Geolife. For each individual, they split his mobility traces, chronologically ordered, into two disjoint parts of approximately the same size: the first half formed the training set, and the second half the testing set. Thus the training and the testing data were not only disjoint, but also separated in time. With such split, they were able to re-identify between 35% and 45% or the users. For comparison, they repeated the experiments also without splitting, i.e., using the same set of traces for training and for testing, and obtained, in this case, a success rate of almost 100%! Of course, this comparison is not completely fair because they used as testing set exactly the same as the training set, instead than a subset as in previous works. Nevertheless, such high success rate shows that (1) the training set and the testing set should be independent to avoid any bias, and (2) the Mobility Markov Chain obtained from the traces of a user is almost always unique to the user.
1.3 The users’ point of view

The users’ concerns about location privacy, and privacy in general, vary a lot from individual to individual, and depend on factors such as age, education, cultural background, etc. They also tend to evolve in time, and cases of privacy breaches that hit the news, like that of “Birds and ‘leaky’ phone apps” [6], can have a huge impact on the attitude of the population.

There have been several studies to assess people’s perceptions and attitude towards privacy. We mention in particular the empirical research conducted at CMU by Acquisti and his team, which provides a systematic analysis of several aspects of human behavior in relation to privacy. See [1] for a summary of their findings.

Concerning the specific case of location privacy, the concerns seem in general less strong than for other kinds of sensitive data (such as medical records, financial data, bank information etc.), and the studies give mixed results. For instance, in 2014 the authors of [35] interviewed 180 smartphone users, recruited through social network announcements and through Amazon Mechanical Turk. They chose Mechanical Turk workers who had achieved master qualification. They obtained the following statistics: 78% of the participants believed that apps accessing their location can pose privacy threats. Also, 85% of them reported that they care about who accesses their location information (in line with the 87% reported by the 2011 Microsoft survey [56]). Furthermore, 77% of the users were interested in installing a privacy protection mechanism. Finally, on the specific method based on the addition of random noise, 52% of the surveyed individuals stated no problem in supplying apps with imprecise location information to protect their privacy. Only 18% of the surveyed people objected to supplying apps with imprecise location information.

On the other hand, in contrast with the other kinds of sensitive data mentioned above (medical record etc.) there seem to be more willingness to renounce to location privacy in exchange of compensation. For instance, Danezis et al. [22] conducted a study on 74 undergraduates to find how much money they would require in order to share a month’s worth of their location data. The median price was £10 if the data were
1.3. The users’ point of view

to be used for research purposes, and £20 if the data were to be used commercially. In [49] the author says that he could easily convince over 250 people from his institution to give him two weeks of GPS data recorded in their car in return for a 1% chance of winning a US$ 200 MP3 player. He asked 97 of them if he could share their location data outside our institution, and only 20% said ‘no’. In contrast, in an experiment conducted by Acquisti et al. [2] on the privacy attitude towards payments, where people were offered the choice between a traceable gift card of 12 US$ or an anonymous gift card of 10 US$, about half of the people chose the second option. Incidentally, [2] main point is to show that people value their privacy differently, depending on how the choice privacy vs non-privacy is presented to them. In particular, people tend to assign a different value to their privacy depending on whether they would receive a compensation in order to disclose otherwise private information, or rather they would pay to protect otherwise public information.

In conclusion, location data seems to be less critical in the mind of many people than data like financial or medical ones, but this may be due to the lack of knowledge about the negative consequences of a location leak. In particular, about the fact that the location can help profiling the user with respect to more sensitive data. Furthermore, the attitude of people concerning the protection of location information may change during time, along with the general increase of privacy concerns. For example, a study in [1] showed that, in the last decade, the percentage of members in the Carnegie Mellon University Facebook network who chose to publicly reveal personal information had decreased steadily. For instance, over 80% of profiles publicly revealed their birthday in 2005, but less than 20% in 2011.
References


References


References


References


References


References


References


References


References


References


References


