Methods for Location Privacy: A comparative overview

Konstantinos Chatzikokolakis, Ehab Elsalamouny, Catuscia Palamidessi, Anna Pazii

Foundations and Trends® in Privacy and Security
Vol. 1, No. 4 (2017) 199–257

DOI: 10.1561/3300000017

now
the essence of knowledge

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

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.

K. Chatzikokolakis, E. ElSalamouny, C. Palamidessi and A. Pazii. *Methods for Location Privacy: A comparative overview*. Foundations and Trends® in Privacy and Security, vol. 1, no. 4, pp. 199–257, 2017.
DOI: 10.1561/3300000017.

1

The problems of privacy in location-based services

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

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 infering 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

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-

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][1], 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

[1] Althoug [68] refers to CabSpotting, the citation is relative to a mobility traces dataset called CRAWDAD.

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.

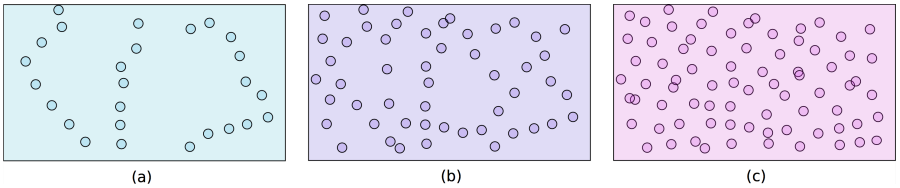


Figure 1.1: Traces in a low (a), medium (b), and high density area (c)

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

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

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-

ditional knowledge. 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 deanonymization 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

to be used for research purposes, and £20 if the data were to be used commercially. In [49] the author says that he could we 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.

2

Deterministic methods

In general, all computational methods for privacy protection are based on degrading the precision of information, in order to confuse the adversary. Now, the main problem with privacy is that in general we cannot distinguish between the adversary and a legitimate party (service provider, other users, ...), because every legitimate party constitutes in principle a potential threat for the sensitive information. Of course, a less precise information is in general less useful also for the legitimate partners, hence there is typically a trade-off between the privacy degree that one wishes to obtain, and the utility of the information. Most of the research on methods for privacy protection has to take utility into account, and aim at achieving a good trade-off.

In the particular case of location data, the protection is obtained essentially in two ways: *spatial obfuscation* and *spatial cloaking*. In both cases, we can have deterministic or probabilistic approaches. In this section we review the main deterministic ones that have been proposed in the literature.

2.1 Deterministic Spatial Obfuscation

Spatial obfuscation approaches preserve privacy by reducing the precision of the position sent from the user to the LBS, and this can be done at the user's site without the intervention of a trusted third party, which is an important advantage of this class of approaches. A naive way of doing it would be, of course, simply to reduce the granularity of the location information: the user could report, instead of the exact coordinates, the "zone" in which he is situated. Or, following a similar idea, he could reduce the precision of the coordinates. This method however is not very robust, because it is subject to triangulation attacks: A user sending two consecutive signals from different zones would reveal that he is close to the border between them, and three consecutive signals from different zones would reveal his position quite accurately. People have therefore investigated more effective solutions.

Duckham and Kulik [27] proposed a method based on the idea of sending to the LBS a set of locations containing also the user's true location. Then, the LBS would provide the services relative to each location, and the user would choose the right one. Much of the effort in [27] was focused on how to generate such set, and how the LBS should reply, in order to not degrade too much the quality of service.

Cheng et al. [17] and Ardagna et al. [4] proposed a method based on the idea of sending an area, dynamically calculated around the user's position, instead of the precise position.

For the sake of accuracy we should mention that also these last three methods contain a probabilistic aspect, in that they require that the user real location has uniform distribution on the set of locations (respectively, on the area) sent to the LBS.

Most of the deterministic methods, anyway, are based on spatial cloaking, and the remaining of this section will be dedicate to them.

2.2 Deterministic Spatial Cloaking

Spatial cloaking, first proposed in [41], is based on one of the most popular methods in the anonymity literature: *group-anonymity*. The idea is to make an individual indistinguishable from a group of other indi-

viduals. This is achieved by reporting a cloaked area, large enough to contain the group size necessary to meet the intended anonymity constraint. In order to limit the size of the cloaked area, some proposals have considered to combine it with temporal cloaking as well. Furthermore, in order to reduce the linkability between requests belonging to the same user's trajectory, [8] proposed the so-called *mix-zones*, where the users' pseudonyms get renewed. The reason why they are called mix-zones is that the temporal order in which the users enter and exit these zone must be obfuscated (*mixed*), otherwise the adversary could be able to link the new pseudonym to the old one.

All the above measures related to spatial cloaking need, of course, the intervention of a trusted party that acts as an *anonymity server*.

In order to explain the technical aspects of spatial cloaking, we start by reviewing two of the most successful approaches to group anonymity: k-anonymity [71, 70, 84] and a refinement of it called l-diversity [54]. These have inspired most of the techniques for spatial cloaking proposed in the literature of location privacy.

2.2.1 k-anonymity

Following the seminal paper [41], the k-anonymity approach to location privacy has attracted a lot of interest, mainly due to its simplicity. There has been a lot of research dedicated to increase the efficiency and reducing the cost of k-anonymity schemes [37, 58, 38, 86, 93], adapting the original architecture to different scenarios [72], moving from a centralized server to distributed ones [98], considering mobile P2P environments [21], and extending the method to trajectories [9]. However, all these proposals share the same key principles (of the k-anonymity approach), which will be illustrated in this section using the basic k-anonymity model of [41].

The k-anonymity method was originally proposed by Samarati and Sweeney in the contest of database sanitization [71, 70, 85]. They started from the observation that, in order to anonymize a database, it is not sufficient to remove the individuals' names, because other attributes present in the database (such as address, birth dates, gender, etc.) may be linked with publicly available information (for instance, a

voters' registers), and make it possible to re-identify the individuals in a large number of cases. These attributes are called *quasi-identifiers*.

As a remedy to the problem of re-identification by linking, Samarati and Sweeney proposed to obfuscate the quasi-identifiers (in addition to removing the names), so that each particular combination would be shared by a group of at least k people. More precisely, k-anonymity means that each individual in the database is indistinguishable, with respect to the quasi-identifiers, from $k-1$ other individuals. Clearly, privacy protection increases with the privacy parameter k. There are various methods to achieve k-anonymity: generalization of an attribute (for example postal addresses can be generalized to the street or to the city, depending on how much we need to generalize), suppression of an attribute, or addition of dummy records. Typically, the k-anonymity-sanitization involves a combination of them.

We illustrate the k-anonymity method with an example suitable for location privacy. Here, the goal of the adversary is to find out about the request (query) that a specific user has issued to a LBS. More precisely, we assume that the adversary aims at identifying the user who has issued the query, and to learn information (as accurate as possible) about it: location, time, type of query, etc. The purpose of k-anonymity is to make these tasks impossible or at least difficult.

As mentioned above, the approach of k-anonymity originated from the field of database privacy. In the case of location privacy, we can consider that the entries in the database are the requests send from the users to the LBS. Table 2.1 illustrates an example of such a database for a set of requests: each row contains the user's identity, the exact position, the query time and the type of query.

Let us assume that the quasi-identifiers are the location and the time of the query. In order to ensure the k-anonymity property, we need to ensure that the user's query be indistinguishable from those of at least $k-1$ other users. To achieve this goal, the identities of these k users are removed from the queries, and their location-time pair is obfuscated to be the same location-area and time-windows, large enough to contain the users' actual locations. Table 2.2 shows the result of k-anonymizing Table 2.1, for $k=2$.

Table 2.1: Queries

User's id	Location	Query	Time
person 1	49.413521, 21.316322	Grocery store	2016-09-21 15:05:37
person 2	49.417653, 21.316890	Hotel	2016-09-21 15:07:00
person 3	49.413123, 21.316876	Night club	2016-09-21 15:08:11
person 4	49.413098, 21.316485	Gas station	2016-09-21 15:14:52

Table 2.2: The table resulting from 2-anonymizing Table 2.1

Cloaked Location	Query	Cloaked Time
49.413-49.418, 21.316-21.317	Grocery store	2016-09-21 15:05-15:10
49.413-49.418, 21.316-21.317	Hotel	2016-09-21 15:05-15:10
49.413-49.414, 21.316-21.317	Night club	2016-09-21 15:08-15:15
49.413-49.414, 21.316-21.317	Gas station	2016-09-21 15:08-15:15

The k-anonymity model of [41] consists of the mobile users, an LBS, and an anonymity server, as depicted in Figure 2.1. The anonymity server is an entity trusted by the users that mediates the queries between the users and the untrusted LBS. The users send their queries $\langle i, q, l, t\rangle$ to the anonymity server, where i is the id of the user, q is the query, l is the location, and t is the time at which the query is generated. To cloak a query's location the anonymity server removes the identity of the user. Furthermore, it obfuscates the location l and the time t at which the queries were generated. To achieve this aim, the server constructs a cloaking region $R = ([x_1, x_2], [y_1, y_2], [t_1, t_2])$ such that there are at least k users in R whose location $l = (x, y)$ at time t satisfies $x_1 \leq x \leq x_2$, $y_1 \leq y \leq y_2$, and $t_1 \leq t \leq t_2$. $[x_1, x_2]$ and $[y_1, y_2]$ represent a two dimensional area where the subject is located, while $[t_1, t_2]$ represents the time period during which the subject was at this area. The server then sends the anonymized queries to the LBS, and the latter sends back the query responses to the server, which will forward them to the corresponding users.

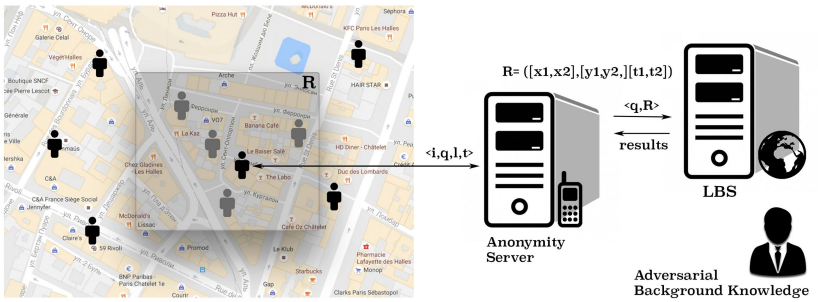


Figure 2.1: k-anonymity model

Sometimes it may happen that in the area of the user to be protected there are no enough users to form a set of k. In this case, we can use fake users and dummy queries. This technique involves generating the necessary number of suitably selected dummy queries, and sending these queries to the service provider. "Suitably selected" means that the dummy requests must look likely to be real queries from the point of view of the attacker. Any side information that allows to rule out any of those queries as having low probability of being real, would fail the purpose [3]. An alternative approach to this problem, illustrated in the next section, is to adapt the size of the cloak so to satisfy the anonymity constraint.

2.2.2 Adaptive-Interval Cloaking

Adaptive-interval cloaking was proposed by Gruteser [41]. The purpose of this method is to achieve anonymity independently from the population density. The idea is to decrease the accuracy of the spatial and/or temporal data so that the resulting cloak contains enough individual requests to satisfy the anonymity constraint.

The algorithm for adaptive spatial cloaking proposed in [41] is illustrated in Table 2.3. The algorithm takes in input the current position of the requester, the coordinates of the area covered by the anonymity

Table 2.3: Adaptive-interval cloaking algorithm. The algorithm computes an area (quadrant) that includes the actual requester and enough potential requesters to satisfy the anonymity constraint k_{min}[41].

1.	Initialize the quadrants q and q_{prev} as the total area covered by the anonymizer
2.	Initialize a traffic vector with the current positions of all known vehicles
3.	Initialize p as the position of the requester vehicle
4.	If the number of vehicles in traffic vector $< k_{min}$, then return the previous quadrant q_{prev}
5.	Divide q into quadrants of equal size
6.	Set q_{prev} to q
7.	Set q to the quadrant that includes p
8.	Remove all vehicles outside q from the traffic vector
9.	Repeat from Step 2

server, and the current positions of all other subjects in the area. The desired degree of anonymity is expressed by a parameter k_{min}. In summary, the algorithm subdivides the area (quadrant) around the subject's position according to some fixed equal-size partition strategy until the number of subjects in the area falls below the constraint k_{min}. The previous quadrant, which still meets the constraint, is then returned.

As a refinement of adaptive spatial cloaking, [41] proposed the combination with temporal cloaking. The purpose was to reveal spatial coordinates with more accuracy, while reducing the accuracy in time. The spatio-temporal cloaking algorithm is provided with an additional parameter: the desired spatial resolution. It determines the monitoring area by dividing the space until the specified resolution is reached. The algorithm then monitors vehicle movements across this area, and delays the request until k_{min} vehicles have visited the area chosen for the requester. Time interval $[t_1,t_2]$ is then computed by setting t_2 to the current time, and t_1 to the time of request minus a random cloaking factor. The area and the time interval are then returned.

2.2.3 *l*-diversity

Sometimes it may happen the k users of a k-anonymity group all have the same value for a sensitive attribute. In this case, being indistinguishable from the other members of the group is of no use, because the entire group leaks the sensitive information. To cope with this problem, [54] proposed the principle of l-diversity. It is a rather subtle concept, and, in order to explain it properly, we need to introduce some technical notions.

Usually all the knowledge at the disposal of the adversary can be helpful to discover sensitive information about the user. In general, we can distinguish the knowledge of the adversary into *prior* and *posterior*. The prior knowledge, also called *background knowledge*, or *side knowledge*, is what the adversary knows before exploiting any observation on the system or on the database.

In general we cannot prevent the adversary from having prior information, but we should be able to control the additional information he gets by observing the published table. The ideal situation is when the knowledge of the adversary does not increase, i.e. when prior and posterior coincide. In general this is impossible to achieve, so the goal is to make the increase as small as possible. [54] showed that the increase of the adversary's knowledge, and more precisely, the increase of the probability to correctly identify the individual's sensitive value, is directly linked to the lack of diversity in the observed values of the sensitive attribute in the group to which the individual belong. More precisely, let q be the value (after sanitization) of the quasi-identifier of the tuple representing the individual under consideration. We call q-block the set of tuples of all the individuals that share the same q value. The l-diversity principle states that the q-block should contain at least l "well-represented" values (i.e., occurring with high frequency) for the sensitive attribute.

2.2.4 Location diversity

In this section we consider l-diversity in the context of location privacy. The goal is to hide sensitive information about the request, which may

not be limited to the coordinates. Indeed, location often contains direct semantic information in terms of location objects such as shops, schools, hospitals, restaurants, churches, etc. These location objects may be sensitive because they may allow to infer indirect information about the individual, such as his job, hobby, religion, etc. In order to protect this sensitive information, It means, in addition to k-anonymity, the server needs to ensure that there are at least l location objects that can be associated to his query [7, 95].

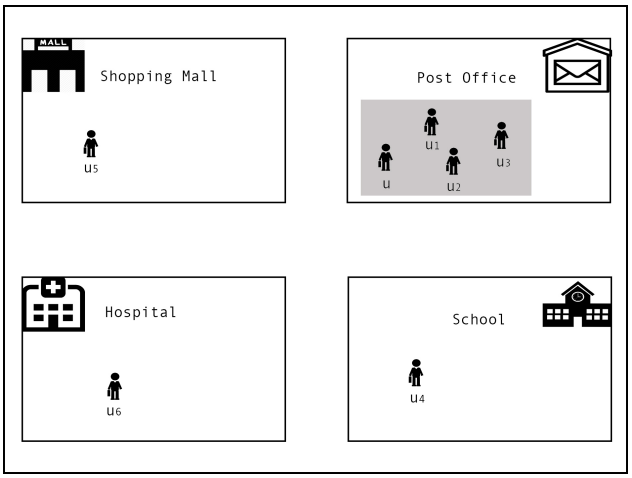
To better illustrate the idea, consider the following example: A courier u working at a post office uses a LBS to find out the best route for some delivery. To avoid identification queries are issued through an anonymity server, which keeps a map of the current users (Figure 2.2a). The server ensures k-anonymity by sending to the LBS a cloaking region contain 4 users (the grey area in Figure 2.2a). However, the cloaking region contains only the post office (where alls these users work), so the adversary may deduce that with high probability u works at the post office.

To address the above problem, [7, 95] considered the principle of *location l-diversity*, which guarantees that the query can be associated with at least l semantically different location objects, so that each of these has probability $1/l$ to be real one. In the example, if we want to achieve l-diversity with $l = 4$, the server needs to consider three other regions containing different semantic objects, and send to the LBS the a group of queries contain the one from u, plus three from fake users u_a, u_b, u_c located in the other regions. Figure 2.2b illustrates the idea.

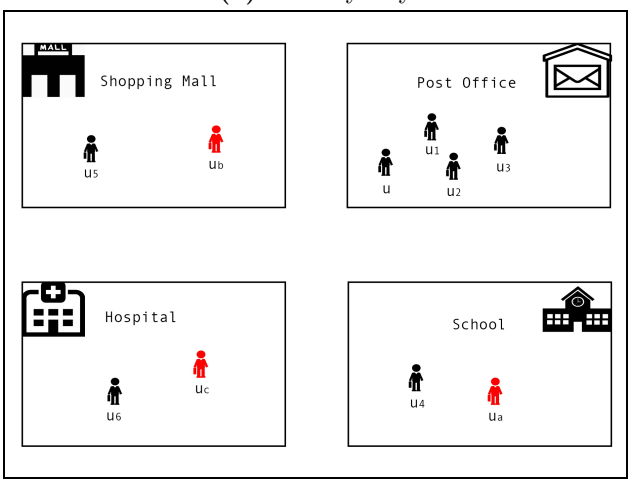
When the LBS returns all results to the anonymity server, the latter filters out the results for the fake users and returns to u the relevant result.

2.2.5 Mix zones

When a moving user sends consecutive requests to an LBS, the corresponding sequence of locations (trajectory) may leak more information than each individual location, and cause a quick degradation of the level of protection, even if each individual location is sanitized. For example, the approximate home location may not identify a single user, but the



(a) k-anonymity



(b) Location diversity

Figure 2.2: Location diversity versus k-anonymity

pair home-work locations, even if approximate, could be unique. Locations belonging to the same trajectories may be linked even if requests are anonymized, because they still must contain a pseudonym in order for the LBS to indicate to the server the request it is answering to. In principle, then, two requests issues along some user's trajectory could be linked via their pseudonym.

To prevent the above problem, [8] proposed the so-called *mix-zones*. The idea is to change frequently each users' pseudonym to a new, unused one, and to do it in such way that the new pseudonym cannot be linked to the old one. To achieve this goal, the server distinguish between *application zones* and *mix zones*. The former are areas in which users typically issue requests to the LBS. For instance, airports, banks, coffee shops, etc. The latter are used to assign new, unused pseudonyms to the users inside them. While they are inside a mix zone, the users cannot send requests to the LBS. In this way, the users going into a mix-zone cannot be linked to those who will come out. Figure 2.3 illustrates the idea: users u_1, u_2 and u_3 enter the mix-zone with pseudonyms x, y and z, and will exit with new pseudonyms s, q and h, respectively.

Obviously, we need to obfuscate also the temporal order in which the user enter and exit the mix zone, otherwise the adversary could try to infer the link. For instance, if they all travel at similar speeds, then it would be reasonable to assume that the first user in will be the first out.

The anonymity provided by a mix zone is measured in terms of the unlinkability between the old and new pseudonyms. For instance, [62] proposed the following definition: A set of users U in a mix-zone Z is k-anonymized iff:

1. U contains at least k users.

2. All users in U are in Z at a point in time, i.e., all users in U must enter Z before any user in U exits.

3. Each user in U spends a random duration of time inside Z.[1]

[1]The random duration here should guarantee that all members of the anonymity set U are equally likely to exit from any of the exit points at a given time. This also corresponds to the maximum entropy (uncertainty) of guessing the original

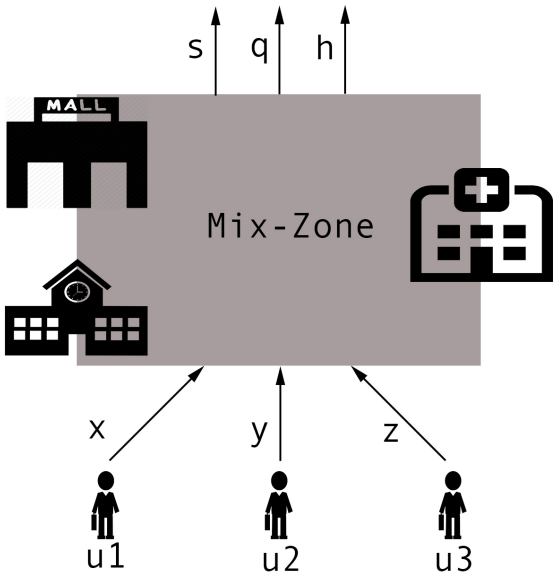


Figure 2.3: A mix-zone with three users

4. The probability of every user in U entering through an entry point of Z is equally likely to exit in any of the exit points of Z.

Point 4 above is the most difficult to ensure. In particular, vehicle movements depend on many factors, such as travel speed, traffic density, road conditions, traffic lights, etc. Thus a vehicle usually is not inside a mix-zone for a random amount of time. For example, Figure 2.4a represents a mix zone that is placed at the intersection of four roads segments $Seg1$, $Seg2$, $Seg3$ and $Seg4$ (in street networks, mix zones are typically placed at crossroads). An adversary knows that a vehicle exiting the mix zone at $Seg3_{out}$ will probably have entered from either $Seg1_{in}$ or $Seg2_{in}$ or $Seg4_{in}$. Now, $Seg3_{out}$ is closer to $Seg2_{in}$ than $Seg1_{in}$ or $Seg4_{in}$, so the adversary could use this information to link events at road segment $Seg3_{out}$ with either $Seg1_{in}$ or $Seg2_{in}/Seg4_{in}$.

pseudonym of the exiting user.

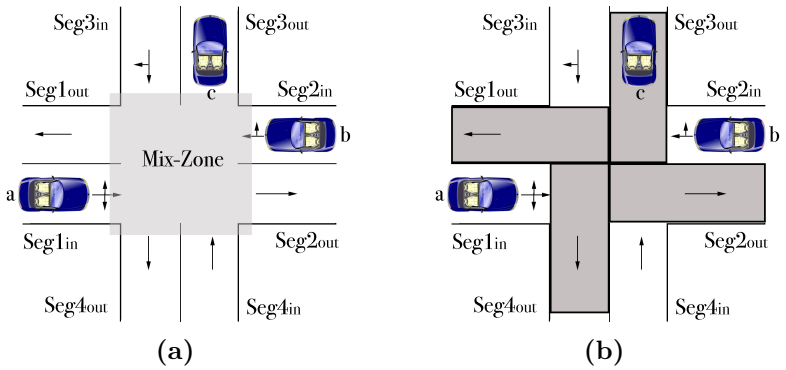


Figure 2.4: A vehicular mix-zone(a) and non-rectangular, adaptive vehicular mix-zones (b)

A more effective way to construct mix-zones is shown in Figure 2.4b. The idea is to construct non-rectangular, adaptive mix-zones that start from the centre of the crossroad on the outgoing road segments [62]. The length of the mix-zone along each outgoing segment depends on the average speed of the road segment, the minimum pairwise entropy threshold and the size of the chosen time window. The pair-wise entropy is computed for every pair of users a and b in an anonymity set U by considering a and b to be the only members in U and determining the linkability between their old and new pseudonyms [20].

2.3 Criticism of the spatial cloaking approach

The spatial cloaking method based on k-anonymity has been criticized in various papers, and most thoroughly in [76, 81]. In [76], the authors argue that the k-anonymity is not a suitable metric for capturing most kinds of privacy that are relevant for location-based applications. In [81] the authors focus on two privacy properties: *query anonymity* and *location privacy*. Achieving query anonymity means concealing the link between the user and his query, while location privacy refers to the link between the user and his spatio-temporal coordinates. In principle, these are the two main goals of the spatial cloaking approaches. However, [76] shows that, while k-anonymity-based cloaking can help

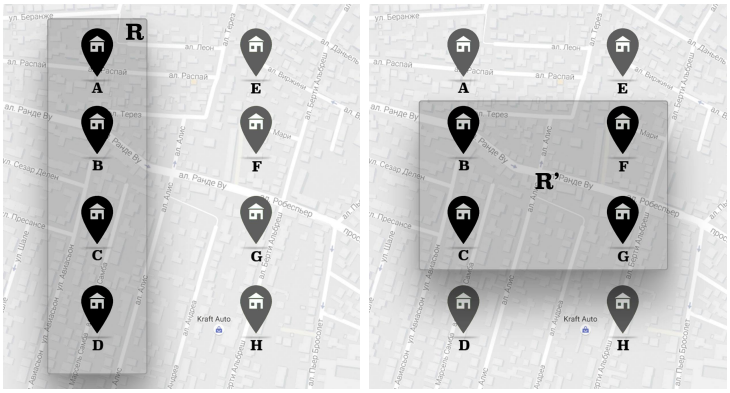
with query anonymity, it may be ineffective, or even counterproductive, with location privacy.

First, the authors of [81] argue that the parameter k in the k-anonymity property is not in itself relevant for location privacy. In fact, we could have a small cloaking region with a large k (for instance, when all the k users are inside the same building), or, vice versa, a large cloaking region with a small k. Clearly the degree of location privacy is low in the first case, and high in the second one. Thus it does not increase with k alone.

Second, the authors of [81] show that, in presence of certain adversarial background information, the cloaking method does not help to increase location privacy, and may even decrease it. We describe in the following their example.

Assume the adversary has statistical information about the users' home location and that he knows that with a high probability all users are at home at a certain time (for instance, late in the evening). Let us consider the neighborhood shown in Figure 2.5a, and assume that we want to achieve 4-anonymity. Suppose that B sends a request $\langle q, R\rangle$ to the LBS, where q is the query and R the cloaking region that covers the user's position. If the adversary intercepts the request, since he is uncertain of the current location of users, he can only use the available background information to infer who/where is the sender of q. Thus, user B is (or is not) 4-anonymous independently of whether or not A, C, and D are currently using the system, or even present at their home locations. In conclusion, it is not necessary for the anonymity server to compute the region R on-the-fly on the basis of the presence of A, C, and D in R. Using the available background information region R could be pre-computed and used uniquely based on the user's location, regardless of whether or not $k - 1$ other users are currently in the vicinity.

On the other hand, consider now the situation illustrated in Figure 2.5b, where only users B, C, F and G are active (i.e., using the system), and assume that the adversary has the same information as in the previous case. When user B sends a request, the anonymity server, following the algorithm for spatial cloaking with $k = 4$, will



(a) **(b)**

Figure 2.5: Examples where the use of cloaking is pointless (a) and harmful (b)

forward $\langle q, R' \rangle$ to the LBS. Assuming that R' is larger than R, since the server tries to minimize the cloaking area, the adversary will learn that B, C, F and G are currently in their home locations, and that A and D are either inactive or absent. In this case, although the query q is still 4-anonymous, the information revealed to the adversary by the cloaking method itself causes a decrease of the location privacy of A, B, C, D, F and G.

3

Randomized methods

In this section we give a comparative overview of the randomized methods for location privacy. Like in the case of the deterministic methods, they can be classified into two groups: those that protect the identity of the users, and those which aim at obfuscating the position of the user. Among the most investigated random mechanisms are those based on the popular notion of differential privacy [30]. The reasons for the success of this notion are essentially the fact that it is much more robust to composition attacks than other notions (especially the deterministic ones), and that it does not depend on the side knowledge of the adversary. These advantages are preserved also in the case of its application in the field of location privacy.

3.1 Differential Privacy

Differential privacy was originally proposed in the area of statistical databases [30, 28, 29, 31]. The goal is to protect an individual's data while publishing aggregate information about the database. This is obtained by adding controlled noise to the query outcome, in such a way that modifying a single user's data will have a negligible effect on the (noisy) reported answer.

More precisely, let $\mathcal{K}$ be a (noisy) mechanism for answering a certain query on a generic database D, and let $P[\mathcal{K}(D) \in S]$ denote the probability that the answer given by $\mathcal{K}$ to the query, on D, is in a set of values S. Then we say that $\mathcal{K}$ satisfies ϵ-differential privacy (where ϵ is a parameter quantifying the level of privacy) if for every pairs of *adjacent* databases D and D' (i.e., differing only for the value of a single individual), we have:

$$\frac{P[\mathcal{K}(D) \in S]}{P[\mathcal{K}(D') \in S]} \leq e^{\epsilon}. \tag{3.1}$$

Note that the smaller is ϵ, the more indistinguishable D and D' are, with respect to the query.

Differential privacy has a Bayesian interpretation. Formally, let R be a record in the database D, and D^- the rest of the database when we remove R, i.e., $D^- = D \setminus R$. Let $P[R = r \mid D^- = d]$ be the probability that the value of R is r given that the value of D^- is d (*prior* probability), and $P[R = r \mid D^- = d, \mathcal{K}(D) \in S]$ be the same probability conditioned additionally on the value of the reported answer $\mathcal{K}(D)$ (*posterior* probability). Then, a mechanism $\mathcal{K}$ is ϵ-differentially private if and only if, for any D, d, R, r, and S:

$$e^{-\epsilon} \leq \frac{P[R = r \mid D^- = d, \mathcal{K}(D) \in S]}{P[R = r \mid D^- = d]} \leq e^{\epsilon} \tag{3.2}$$

Note that Inequalities (3.2) essentially mean that knowledge of the adversary about an individual record R, when he knows the values of all other individuals, does not increase significantly by knowing $\mathcal{K}(D)$. In other words, differential privacy establishes a bound on what the adversary can lean about the value of an individual by knowing $\mathcal{K}(D)$.

As already mentioned, one of the advantages of differential privacy is that it does not depend on the prior (attacker's side information), hence a differentially private mechanism can be designed without making any assumption about the knowledge of the adversary.

Even more important, differential privacy is robust with respect to composition attacks: if we query n ϵ-differentially private mechanisms on two adjacent databases D and D', the bound on the ratio of the

probabilities becomes $e^{n\epsilon}$. Namely, the composition of n ϵ-differentially private mechanisms gives a $(n\,\epsilon)$-differentially private mechanism. This means that the level of privacy (as expressed by the parameter ϵ) decreases linearly with n. In reality, one could argue that the privacy level is represented by the bound on the probability ratio, and that therefore it decreases exponentially with n. Still, it will never be the case that D and D' could be completely distinguished by composing more queries (to distinguish them completely the probabilities for D and D' should become 1 and 0 respectively, i.e., the the probability ratio should become infinite). This is a crucial advantage with respect to all other known methods. Another advantage is that having the exact formula for the privacy degradation allows to plan in advance the amount of noise to add at each query: if our goal is to achieve ϵ-differential privacy, and we want to ask n queries, then we have to apply to the single queries a noise with parameter $\epsilon' = \epsilon/n$. In case we do not know n, then we can think of ϵ as a privacy *budget* which gets gradually consumed by each query: when the budget is depleted, no more queries are allowed.

Finally, differential privacy is very easy to implement: A typical way to achieve it is to add controlled random noise to the query output, for example noise drawn from a Laplace distribution. Furthermore, it does not need a trusted third party between the database's curator and the person querying the database: the noise can be computed and added by the curator.

3.2 Protection of identity

Most of the works that have used differential privacy to protect the user's identity in the context of location-based applications have considered a scenario where *aggregate* information about several users is published. In such situation, differential privacy can be applied just like in the case of databases. For instance, Machanavajjhala et al. [53] used a synthetic data generation technique to publish statistical information about commuting patterns in a differentially private way. Ho et al. [43] used a spatial decomposition technique to ensure differential privacy in a database with location pattern mining capabilities. Chen et al.

[16] used variable-length n-grams to disclose sequential data, such as mobility traces, in a differentially private way.

One of the most impressive work along these lines is represented by DP-WHERE [57], which builds on a previous proposal called WHERE (Work and Home Extracted REgions) [46]. WHERE is an approach to model mobility traces and predict human densities over time at the geographic scale of metropolitan areas. Some potential uses are, for instance, to explore what-if scenarios regarding changes in residential density, telecommuting facilities, etc. WHERE uses real CDRs (Call Detail Records) to infer probability distributions on location attributes, and then, on the basis of these distributions, it produces synthetic CDRs for a synthetic population. The data sets generated in this way present a better privacy properties than the original (simply anonymized) real data, and still retain a good utility, as demonstrated by experiments conducted on billions of location samples for hundreds of thousands of cell phones in the New York and Los Angeles metropolitan areas.

DP-WHERE adds a further level of protection to WHERE by adding controlled noise to the probability distributions generated in the first phase. This is done by counting the instances for each possible attribute and value (for instance, the number of phone owners living in a certain district, as deduced by the billing address) and then adding Laplacian noise to the result. Afterwards, these noisy distributions are used like in WHERE to generate synthetic CDRs. The utility-preservation of DP-WHERE has been validated against WHERE and against the real CDRs, using about 1 billion CDRs involving over 250,000 phones, and generating traces within 30 consecutive days in an area of about $14,000$ squared miles around New York City. As measure of utility, they used the Earth Mover Distance (EMD) [69] which can be described intuitively as the minimum amount of "energy" required to transform one probability distribution into another, and normalized so to be measured in miles. For instance, concerning the distribution on home-work commuting distances, with cell size of $0.01°$ latitude $\times 0.01°$ longitude (about 0.6×0.6 squared miles) and $\epsilon = 0.33$, the EMDs between the real distribution and those of DP-WHERE and WHERE

turned out to be about 3.5 and 3.2 miles, respectively. Considering the large covered area and the relatively coarse granularity of the cells, these results seem quite encouraging.

3.3 Protection of location

We consider now the case in which we want to protect the location privacy of a single user by using probabilistic noise. We are not concerned here with protecting the user's identity, we want just to obfuscate his location.

Some researchers have tried to apply differential privacy to this problem, but differential privacy as such require data from a set of individuals. Considering a set consisting of one single user only does not give good results, because the definition would require that any change in the user's location should have negligible effect on the published output, making it impossible to communicate any useful information to the service provider. In other words, privacy would have a price too high in terms of utility loss. To overcome this issue, Dewri [25] proposed a mix of differential privacy and k-anonymity, by fixing an anonymity set of k locations and requiring that the probability to report the same obfuscated location z from any of these k locations should be similar (up to e^{ϵ}). To achieve this property, [25] showed that it is sufficient to add Laplace noise to each Cartesian coordinate independently. There are however two problems with the proposed property: first, the choice of the anonymity set crucially affects the resulting privacy; outside this set no privacy is guaranteed at all. Second, the property itself is rather weak; reporting the geometric median (or any deterministic function) of the k locations would satisfy the same definition, although the privacy guarantee would be substantially lower than with Laplace noise. Nevertheless, Dewri's intuition of using Laplace noise for location privacy is valid, and [25] provides extensive experimental analysis supporting this claim.

Most of the methods for location privacy aiming at obfuscating a single-user location, however, use the simple and general framework illustrated in the next section.

3.3.1 The general framework

The seminal work of Reza et al. [77] provided a fundamental contribution to the theory of location privacy: it established a rigorous framework for defining obfuscation mechanisms, and proposed various natural and general notions of privacy and utility metrics. These metrics have been widely used as a common denominator for evaluating mechanisms for privacy protection in many of the works in this area. We illustrate here some of the main elements of this framework, focussing mainly on the notions that will be useful in the rest of this survey. We also simplify the definitions by considering only single locations, along the lines of the companion work [78], while the original paper [77] considered traces.

A common way of achieving location privacy is to apply a *location obfuscation* mechanism, that is a probabilistic function $K : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ where $\mathcal{X}$ is the set of possible locations, and $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$. K takes a location x as input, and produces a *reported location* z which is communicated to the service provider. In general $\mathcal{X}$ is considered to be finite, in which case K can be represented by a stochastic matrix, where k_{xy} is the probability to report y from location x.

A prior distribution $\pi \in \mathcal{P}(\mathcal{X})$ on the set of locations can be viewed either as modelling the behaviour of the user (the *user profile*), or as capturing the adversary's *side information* about the user. Given a prior π and a metric d on $\mathcal{X}$, the expected distance between the real and the reported location is:

$$\textsc{ExpDist}(K, \pi, d) = \textstyle\sum_{x,y} \pi_x k_{xy} d(x, y) \tag{3.3}$$

From the user's point of view, we want to quantify the service's *quality loss* (QL) induced by the mechanism K. Given a *quality metric* d_Q on locations, such that $d_Q(x, z)$ measures how much the quality decreases by reporting z when the real location is x (the Euclidean metric d_2 being a typical choice), the quality loss can be naturally defined as the expected distance between the real and reported locations, that is

$$\mathrm{QL}(K, \pi, d_Q) = \textsc{ExpDist}(K, \pi, d_Q) \tag{3.4}$$

The QL can also be viewed as the (inverse of the) utility of the mechanism, hence we will also call it *utility loss*.

Similarly, we want to quantify the *privacy* provided by K. A natural approach is to consider a Bayesian adversary with some prior information π, trying to remap y back to a guessed location $\hat{x}$. A remapping strategy can be modeled by a stochastic matrix H, where $h_{y\hat{x}}$ is the probability to map y to $\hat{x}$. Then the privacy of the mechanism can be defined as the expected error of an adversary under the best possible remapping:

$$\textsc{AdvError}(K, \pi, d_A) = \min_H \textsc{ExpDist}(KH, \pi, d_A) \tag{3.5}$$

Note that the composition KH of K and H is itself a mechanism. Similarly to d_Q, the metric $d_A(x, \hat{x})$ captures the adversary's loss when he guesses $\hat{x}$ while the real location is x. Note that d_Q and d_A can be different, but not necessarily. The canonical choice is to use the Euclidean distance for both.

Apart from the use of d_Q vs d_A, the main difference between $\mathrm{QL}(K, \pi, d_Q)$ and $\textsc{AdvError}(K, \pi, d_A)$ is that the first is defined simply as the expected loss, without remapping. It seems natural, indeed, not to expect the service provider to have knowledge about the user's prior.

In the rest of this survey we present the two main lines of research in the area of the definition of mechanisms for location obfuscation: the *game-theoretic* approach, leading to optimal mechanisms, and the so-called *geo-indistinguishability* framework. Although historically the game-theoretic approach started to be explored first, in the next section we start with geo-indistinguishability, because we need this notion to explain some of the game-theoretic approaches which have been developed more recently.

3.3.2 Geo-indistinguishability

The notion of geo-indistinguishability, proposed in [3], is based on an extension of differential privacy to arbitrary metrics [12]. Using the notation introduced in § 3.3.1, let k_{xy} represent the conditional probability that the sanitization mechanism sends to the LSB the location y

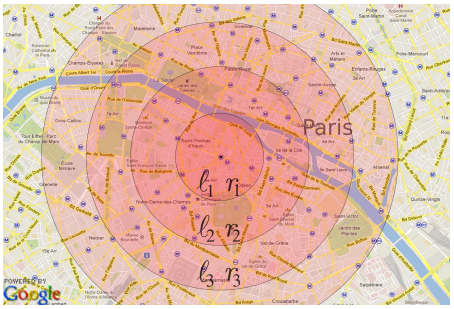


Figure 3.1: Geo-indistinguishability: the level ℓ of privacy varies with r.

when the real location is x. Also, for any pair of locations x, z, let $d(x, z)$ represent the distance between them. We say that the mechanism satisfies ϵ-geo-indistinguishability (where ϵ is a parameter representing the level of privacy) if for any locations x, y, z we have:

$$\frac{k_{xy}}{k_{zy}} \leq e^{\epsilon\, d(x,z)} \tag{3.6}$$

Intuitively, this means that the user's location x is not "distinguishable" (up to a certain level) from any other location z within a radius r from x, where the level of distinguishability $\ell = \epsilon\, r$ grows with the distance r. In other words, an attacker can determine that the user is in – say – Paris rather than London, and be reasonably confident that he is in the Quartier Latin, but cannot tell where exactly in the Quartier Latin he is: the protection increases exponentially as the distance from the real location decreases. Figure 3.1 illustrates the situation for a user located in "Café Les Deux Magots" (at the center of the map), which is a popular destination in the Quartier Latin: darker areas means higher level of indistinguishability.

Geo-indistinguishability is closely related to differential privacy. In fact, the inequalities (3.1) and (3.6) can be derived from each other, if we interpret x, z as databases, and $d(x, z)$ as the Hamming distance between x and z (i.e., the number of entries in which they differ), and we assume that the set of possible values is discreet. As such, geo-indistinguishability inherits the appealing properties of differential privacy: it is independent from the prior of the adversary, and robust

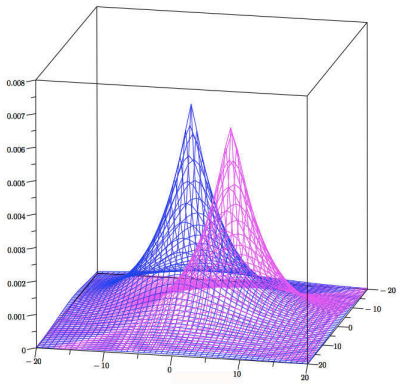


Figure 3.2: The probability density function of two planar Laplace distributions, centered at $(-2, -4)$ and at $(5, 3)$ respectively, with $\epsilon = 1/5$.

with respect to composition. It also has a natural Bayesian interpretation, in that it establishes a bound on the amount of information (relative to the prior information) on the real location that the adversary can acquire by learning the reported location. Finally, and most important, it does not need a third trusted party to be achieved: it can be implemented at the user's end simply by adding noise to the real location, using for instance a planar Laplace as the noise function. Such function, illustrated in Figure 3.2, is centered in the real location of the user, and decreases exponentially with the distance from it. This means that a user located in x report a location y with a probability that decreases exponentially with the distance between x and y. Using the triangular property of metric spaces, it's easy to see that this implies the geo-indistinguishability property.

The generation of Laplace noise can be done efficiently and at low-cost using an analytic expression, so the mechanism can be implemented easily even in a computationally limited device such as a smart phone.

Thanks to the above properties, geo-indistinguishability via the Laplace mechanism has been adopted as the basis or as a component of several tools and frameworks for location privacy, including: Location Guard [50], LP-Guardian [35], LP-Doctor [34], the system for secure

nearby-friends discovery in [51], SpatialVision QGIS plugin [67], and it is one of the possible input methods in STAC [66]. Furthermore, the PIM mechanism [92] can be considered an extension of the planar Laplace to the case of traces (aka trajectories).

When the possible locations of the users are modeled by a continuous region, the Laplacian was formally proved in [33] to provide the minimal noise required to satisfy geo-indistinguishability on this region. However, it is not the only way to implement geo-indistinguishability efficiently. On a discrete map one can also use a planar variant of the geometric mechanism [39], and, if the map is bounded, the exponential mechanism (cfr. [14] for its use in location privacy) and the tight-constraints mechanism [32] are applicable as well. In § 3.3.5 we will see that these mechanisms actually outperform the Laplacian when they are applied on a discrete grid.

Although geo-indistinguishability presents various appealing aspects, it has the problem of treating space in a uniform way, imposing the addition of the same amount of noise everywhere on the map. This is sometimes undesirable: for instance, a user located in a small island in the middle of a lake should generate much more noise to conceal his location, so to include (as reported points) also other locations on the ground, because the adversary knows that it is unlikely that the user is in the water. In order to cope with this problem, [14] proposed to use "elastic distinguishability metric" that warps the geometrical distance, capturing the different degrees of density (of likely locations) of each area. As a consequence, the obtained mechanism adapts the level of noise so to achieve the same degree of privacy everywhere.

3.3.3 The problem of traces

The method of geo-indistinguishability was designed for protecting the user's location when making a request to a LBS. In practice, however, a user rarely performs a *single* request. Typically, he performs different activities throughout the day: for instance he might have lunch, do some shopping, visit friends, etc. During these activities, the user may produce several requests: searching for restaurants, getting driving directions, finding friends nearby, and so on. For each request, a new location

needs to be reported to the service provider. If we sanitize the requests one by one by independently adding noise at the moment when each of them is executed, then the privacy degrades as the number of requests increases, due to the *correlation* between the locations. Technically, the mechanism independently applying ϵ-geo-indistinguishable noise to a sequence of n location (trace) satisfies $(n\,\epsilon)$-geo-indistinguishability. This is exactly the same phenomenon that happens with differential privacy when we have repeated queries (cfr. § 3.1). Note anyway that any obfuscation mechanism is bound to cause privacy loss when used repeatedly; geo-indistinguishability, like differential privacy, has the advantage of allowing to directly quantify this loss terms of the number of repetitions.

In order to mitigate this degradation and protect the privacy of traces, [13] had the intuition that the correlations between successive locations in a trace could actually be exploited, so to avoid reporting all the time a new sanitized location. The idea is to use a *prediction function* that tries to guess the new location based on the previously reported locations. The proposed mechanism then tests the quality of the predicted location: in case of success it reports the prediction, otherwise it reports the new location sanitized with new noise. Note that reporting the predicted location does not increase the knowledge of the adversary, since he knows the previous locations, so he could calculate the predicted location as well.

The above method helps to mitigate the degradation of privacy when the requests have to be sanitized "on the fly", like in the example illustrated above, where the used has to issue a new request whenever he needs a new service. One problem in this scenario is that, in general, the user does do not know in advance how many requests he is going to issue, so he cannot plan what should be the level ϵ' of privacy to be used in each application of the mechanism so to ensure a certain level ϵ of privacy at the end. One solution (not too satisfactory) could be to use a privacy budget, like in differential privacy, and disallow further requests when all the budget is consumed. Another, less dramatic, solution could consist in increasing the noise progressively at each application of the mechanism. For example, if we aim at achieving the privacy level ϵ, we

could use as parameters the geometric series $\{\epsilon/2^n\}_n$ since $\sum_1^\infty \epsilon/2^n = \epsilon$.

On the other hand, if all the locations of the trace to be sanitized are known in advance, then there may be better methods to sanitize it, possibly providing a better trade-off between privacy and utility. This scenario is considered for instance in [92], which proposes to add Laplace noise directly to the convex hull of the trace.

3.3.4 The game-theoretic approach and optimal mechanisms

As argued in § 2, it is important to achieve a good trade off between privacy and utility. A major line of research in the area of randomized mechanisms for privacy has been devoted to designing the optimal mechanism, i.e., computing the noise function that gives the optimal trade-off between privacy and utility.

The first paper to undertake the challenge of producing an optimal mechanism was [80]. The authors of this seminal paper proposed an intriguing interpretation of the problem in terms of 0-sum Stackelberg games. In such games, a leader (the user) and a follower (the attacker) interact strategically, trying to maximize their own payoff. The leader decides on his strategy, i.e., the distributions $k_{x\cdot}$ (cfr. § 3.3.1), knowing that it will be observed by the follower, who will optimize his choice based on this observation. We assume that that the adversary knows the choice of the distributions $k_{x\cdot}$ and will use this knowledge to improve his attack effectiveness, by a judicious choice of the remappings $h_{x\cdot}$ (cfr. § 3.3.1). On the other hand, the user has to take into account the adversary's choice of the $h_{x\cdot}$'s when choosing the distributions $k_{x\cdot}$'s for his mechanism. The fact that these are 0-sum games means that the user's gain coincides with the adversary's loss. The payoff function, i.e., the adversary loss considered in [80], is the privacy measure defined in (3.5) as the expected error of a Bayesian adversary. In addition, [80] extended the classic formulation of a Stackelberg game with an extra constraint to ensure that the quality loss QL, as defined in (3.4), is not greater than a threshold established by the user. The Stackelberg equilibrium (aka the saddle point strategy) of the game, under the QL constraint, defines exactly the optimal mechanism in the tradeoff curve between the privacy and the quality of service requirements.

Note that the optimal mechanism defined in this way depends by definition on the prior knowledge of the adversary, because the expected error of a Bayesian adversary (which is the payoff function) depends on the prior, and consequently the saddle point depends on it as well. This contrasts with the geo-indistinguishable mechanisms that, as explained in § 3.3.2, are independent from the adversary.

In order to compute the optimal mechanism, [80] shows that the game can be reduced to a linear optimization program, where the variables are the strategies of the user and the adversary, k_{xy} and h_{yz} respectively. The solution of the program gives the equilibrium strategy, and the values of $k_{x\cdot}$ in this strategy are the distributions defining the mechanism.

In [79], the full version of [80], the authors extended the above model to cover inference attacks that are applied to individual locations in the user's trace at various times, e.g. past, current, or future positions, instead of focusing only on his current location. The adversary exploits his knowledge about the correlation between the positions of the user along her trace to construct a prior upon which he makes his strategy of localization attack.

We note that in [80, 79] the authors fix the utility and and optimize the privacy. Two subsequent works, [10] and [75], have considered instead the opposite approach of fixing the privacy and optimizing the utility. Furthermore, [10] considered as privacy constraints the inequalities expressing geo-indistinguishability, while [75] considered three different types of constraints: geo-indistinguishability, AdvError, and the combination of the two. Interestingly, the resulting optimal mechanisms showed similar utilities in all these three cases. This is probably due to the fact that geo-indistinguishability is strictly related to Bayesian inference.

As an example of how the corresponding linear programs look like, we illustrate in Table 3.1 the linear program of [10]. The notation is the same as in § 3.3.1, and k_{xy} represent the variables of the program. It is interesting to note that, when the constraints are those for geo-indistinguishability, it is indifferent to include a remapping function in the target QL. In fact [10] showed that in their setting there al-

Table 3.1: The linear program of [10].

Minimize:	$\mathrm{QL}(K, \pi, d_Q)$	
Subject to:	$k_{xy} \leq e^{\epsilon d_{\mathcal{X}}(x,z)} k_{zy}$	$x, y, z \in \mathcal{X}$
	$\sum_{y \in \mathcal{X}} k_{xy} = 1$	$x \in \mathcal{X}$
	$k_{xy} \geq 0$	$x, y \in \mathcal{X}$

ways exists a so-called *direct* optimal mechanism, i.e., a mechanism for which QL coincides with the expected quality loss of an LBS allowed to apply an optimal remapping. The main disadvantage of the optimal approach is that it is based on linear programming, and therefore it is computationally very expensive, and not feasible when the number of locations is high. To address this problem, Bordenabe et al. proposed a method to reduce the complexity at the price of renouncing to perfect optimality [10]. The idea is to reduce the number of the constraints expressing geo-indistinguishability by considering only a subset of them, corresponding to a spanning tree in the underlying graph. The missing constraints are implied by the triangular inequality, in combination with a dilation factor. In this way, the number of constraints becomes $\mathcal{O}|\mathcal{X}|^2$, i.e., quadratic on the number of locations, in contrast with the $\mathcal{O}|\mathcal{X}|^3$ of the original program. In the same work, Bordenabe et al. showed with various examples that the decrease of utility with respect to the original program is not too significant, while the gain in computation time is considerable. Still, even with the reduction to $\mathcal{O}|\mathcal{X}|^2$ constraints, solving the linear program is unfeasible when the cardinality of $\mathcal{X}$ is of the order of hundreds.

3.3.5 Discussion

We now discuss and compare the various randomized approaches. First, we contrast the optimal mechanism with the others to see how significant is the difference in terms of the privacy/utility trade-off. Second, we reflect on the meaning of the measures defined in § 3.3.1 and discuss to what extent they provide a useful notion of privacy protection and utility.

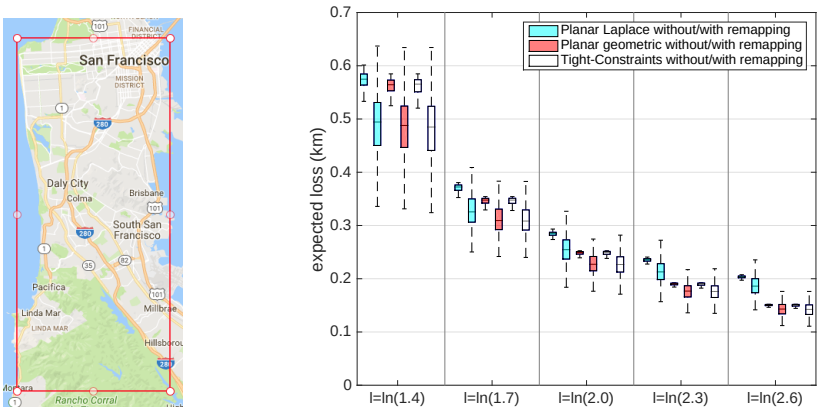
Comparison We compare the optimal and the geo-indistinguishability -based approaches with respect to the trade-off between privacy and utility loss as defined in § 3.3.1.

The advantages and disadvantages of the above approaches could be, at a first thought, resumed as follows: Generating the noise via an "immediate" method (Laplace, geometric, exponential, tight-constraints)[1] is efficient, but there is no guarantee of optimality. Generating the noise via linear programming techniques, on the other hand, is computationally expensive, and not feasible for more than about hundred locations, but it gives the optimal trade-off between privacy and utility.

A recent paper [15], however, has revisited this judgement and argued that the greater utility of the optimal method depends mainly on the use of the background knowledge and consequent remapping. The immediate methods in themselves are independent from the background knowledge, but, if the background knowledge is known (as it is assumed in the optimal method), then it can be used to enhance the immediate methods via an optimal remapping, thus increasing their utility. A key observation is that remapping (as any post-processing) does not affect the property of geo-indistinguishability, hence the privacy constraints are preserved by this transformation. Note also that the remapping can be applied at the user's end, without any modification of the service provider.

Some of the resulting methods (the remapping-enhanced Laplace, geometric, and tight-constraints) were evaluated by [15] on real-world datasets from the Gowalla and Brightkite social networks. In particular, Figure 3.3b illustrates the boxplot of the utility loss of the various methods applied to the Gowalla dataset on a rectangular region of size 12 km $\times$ 28 km covering most of the San Francisco peninsula (Figure 3.3a, region delimited by the red line). Here, l represents the degree of privacy. More precisely, $l = \epsilon\, d$ where $d = 0.1$ km is the size of a location. As we can see from the figure, the experiments show consistently

[1] These methods are called "direct" in [15]. In this survey we use the term "immediate" instead, to avoid confusion, since the adjective "direct" was already used earlier to denote a mechanism that incorporates the optimal remapping.



(a) **(b)**

Figure 3.3: Evaluation of the immediate methods with and without remapping

a significant increase of utility when the remapping is applied.[2] Also, as anticipated in § 3.3.2, we can see that both the geometric and the tight-constraints mechanisms perform better than the Laplace on a discrete grid, and that the tight-constraints one performs slightly better than the geometric one.

The next set of experiments in [15] focused on the comparison between the remapping-enhanced immediate methods and the optimal mechanism. Since the latter can be computed only for a small number of locations, they had to consider a relatively small area. Figure 3.4 displays the boxplot for the utilities on a region in San Francisco of 2 km $\times$ 2 km. The area is divided in a grid of 10×10 square cells (locations), each of size 0.2 km $\times$ 0.2 km. As we can see, the remapped methods achieve an utility close to the optimal one.

Concerning the (un-)feasibility of the optimal mechanism: some researchers claim that it be scaled to large areas by making the grid coarser. We argue, however, that the optimality provided by this approach is questionable: in a grid the real locations are approximated

[2]Interestingly, [61] has shown that the application of an optimal remapping can turn any obfuscation mechanism into an optimal one in terms of average adversarial error. Namely, the resulting mechanism has maximal AdvError among all the mechanisms that have the same quality loss QL.

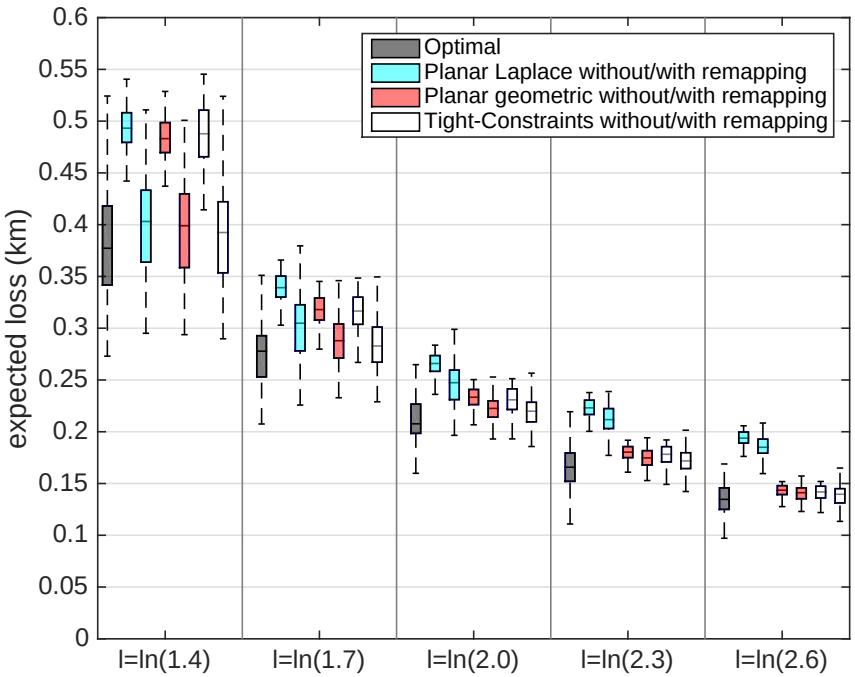


Figure 3.4: A Comparison between the optimal mechanism and the immediate ones constructed on a grid of 100 cells covering a region in San Francisco

by the center of the cell in which they are situated, hence, the coarser is the grid, the looser is this approximation. The optimal mechanism constructed on a coarse grid is optimal for that grid, but it is no longer optimal when considering the quality of service on the "real" continuous surface. In fact, [15] shows that on a grid of square cells of size 2 km $\times$ 2 km the "optimal" method has a much worse utility than the immediate methods. This is due to the fact that in such a grid a real location can be up to $\sqrt{2}$ km away from its best approximation. Another important observation made in [15] is that, as the grid becomes coarser, the optimal mechanism tends to approximate the deterministic 'cloaking' mechanism.

On the meaning of the privacy and utility measures The authors of [61] have critically revisited the optimal method and the privacy

measure AdvError defined in § 3.3.1. They have observed that, for every level of utility, there is an optimal mechanism which does not provide adequate privacy protection, despite it (obviously) maximizes AdvError. They construct the mechanism as follows. Let z^* be a fixed location, and let Q^* be the utility of the mechanism that always reports z^*, independently from the real location. Then, choose the desired level of QL (quality loss) Q (with $Q \leq Q^*$) and set $\alpha = 1 - Q/Q^*$. Now, define the so-called *coin mechanism* f_{coin} as the mechanism that, when applied to a real location x, reports the same location x with probability α, and reports z^* with probability $1 - \alpha$. Surprisingly, it turns out that f_{coin} is an optimal mechanism in the sense of [80], namely it provides the maximal AdvError among all mechanisms with the same QL. Clearly, however, such a mechanism does not provide a good privacy protection according to common sense, since it completely reveals the location of the user with probability α.

The authors of [61] come to the conclusion that the unique criterion of maximizing AdvError is not sufficient to obtain a good mechanism for privacy protection, and propose to apply also other criteria, based on the amount of entropy, or on the worst-case utility loss. They also show that the geo-indistinguishability measures perform well with respect to these criteria.

Incidentally, the authors of [61] argue that also QL might be unsatisfactory as the only measure of utility, since it does not provide guarantees for the worst case: the distance between a real location and the reported one might be large even if the level of QL is considered sufficient. And clearly, a mechanism that performs well in average, but may fail to indicate the closest hospital when the user desperately needs it, would not be considered very reliable.

4

Conclusion

This monograph has revised some of the techniques for privacy protection, focusing in particular on the anonymity technologies (k-anonymity, l-diversity, spatio-temporal cloaking, mix zones) and randomized methods for location obfuscation (geo-indistinguishability and optimal methods). We have tried to present a comparative, critical view of the main approaches in the above areas, without pretending to be comprehensive and cover all aspects of location privacy: The problem is multi-faceted, and literature in this field is huge. For the interested reader, we recommend as further reading other surveys such as [49, 90, 91, 36, 88]. We also recommend [19] for a neat overview of the spatial cloaking approaches.

Acknowledgements

This work was partially supported by the ANR project REPAS. The work of Anna Pazii was supported by Renault R&D.

References

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Alessandro Acquisti, Leslie K. John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249 – 274, 2013.

[3] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: differential privacy for location-based systems. In *Proceedings of the 20th ACM Conference on Computer and Communications Security (CCS 2013)*, pages 901–914. ACM, 2013.

[4] Claudio Agostino Ardagna, Marco Cremonini, Ernesto Damiani, Sabrina De Capitani di Vimercati, and Pierangela Samarati. Location privacy protection through obfuscation-based techniques. In Steve Barker and Gail-Joon Ahn, editors, *Proc. of the 21st Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DAS)*, volume 4602 of *Lecture Notes in Computer Science*, pages 47–60. Springer, 2007.

[5] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[6] James Ball. Angry birds and 'leaky' phone apps targeted by NSA and GCHQ for user data. The Guardian, January 27, 2014. `http://www.theguardian.com/world/2014/jan/27/nsa-gchq-smartphone-app-angry-birds-personal-data`.

[7] Bhuvan Bamba, Ling Liu, Péter Pesti, and Ting Wang. Supporting anonymous location queries in mobile environments with privacygrid. In *Proc. of the 17th International Conference on World Wide Web (WWW)*, pages 237–246. ACM, 2008.

[8] Alastair R. Beresford and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[9] Claudio Bettini, Xiaoyang Sean Wang, and Sushil Jajodia. Protecting privacy against location-based personal identification. In *Proceeding of the 2nd Workshop on Secure Data Management (SDM 2005)*. Springer, 2005.

[10] Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 21th ACM Conference on Computer and Communications Security (CCS 2014)*, 2014.

[11] J. Brownlee. This Creepy App Isn't Just Stalking Women Without Their Knowledge, It's A Wake-Up Call About Facebook Privacy (Update), March 2012. `http://www.cultofmac.com/157641/`.

[12] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás E. Bordenabe, and Catuscia Palamidessi. Broadening the scope of Differential Privacy using metrics. In Emiliano De Cristofaro and Matthew Wright, editors, *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS 2013)*, volume 7981 of *Lecture Notes in Computer Science*, pages 82–102. Springer, 2013.

[13] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A predictive differentially-private mechanism for mobility traces. In E. De Cristofaro and S.J. Murdoch, editors, *Proceedings of the 14th International Symposium on Privacy Enhancing Technologies (PETS 2014)*, volume 8555 of *Lecture Notes in Computer Science*, pages 21–41. Springer, 2014.

[14] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. Constructing elastic distinguishability metrics for location privacy. *PoPETs*, 2015(2):156–170, 2015.

[15] Kostantinos Chatzikokolakis, Ehab ElSalamouny, and Catuscia Palamidessi. Efficient utility improvement for location privacy. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2017(4):308–328, 2017.

[16] Rui Chen, Gergely Ács, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In Ting Yu, George Danezis, and Virgil D. Gligor, editors, *Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS 2012)*, pages 638–649. ACM, 2012.

[17] Reynold Cheng, Yu Zhang, Elisa Bertino, and Sunil Prabhakar. Preserving user location privacy in mobile data management infrastructures. In George Danezis and Philippe Golle, editors, *Proceedings of the 6th International Workshop on Privacy Enhancing Technologies (PET)*, volume 4258 of *Lecture Notes in Computer Science*, pages 393–412. Springer, 2006.

[18] Anne Cheung. Location privacy: The challenges of mobile service devices. *Computer Law & Security Review*, 30(1):41–54, 2014.

[19] Chi-Yin Chow. Cloaking algorithms for location privacy. In Shashi Shekhar, Hui Xiong, and Xun Zhou, editors, *Encyclopedia of GIS.*, pages 229–235. Springer, 2017.

[20] Chi-Yin Chow and Mohamed F. Mokbel. Trajectory privacy in location-based services and data publication. *SIGKDD Explorations*, 13(1):19–29, 2011.

[21] Chi-Yin Chow, Mohamed F. Mokbel, and Xuan Liu. Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments. *Geoinformatica*, 15(2):351–380, April 2011.

[22] George Danezis, Stephen Lewis, and Ross J. Anderson. How much is location privacy worth? In *Proceedings of the 4th Annual Workshop on the Economics of Information Security, (WEIS 2005)*, 2005.

[23] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Nature Scientific Reports*, 3(1376), 03 2013.

[24] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. Identification via location-profiling in gsm networks. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (WPES 2008)*, pages 23–32. ACM, 2008.

[25] Rinku Dewri. Local differential perturbations: Location privacy under approximate knowledge attackers. *IEEE Transactions on Mobile Computing*, 99(PrePrints):1, 2012.

[26] Stuart Dredge. Tinder dating app was sharing more of users' location data than they realised. The Guardian, February 2014. `https://www.theguardian.com/technology/2014/feb/20/tinder-app-dating-data-location-sharing`.

[27] Matt Duckham and Lars Kulik. A formal model of obfuscation and negotiation for location privacy. In *Proc. of the Third International Conference on Pervasive Computing (PERVASIVE)*, volume 3468 of *Lecture Notes in Computer Science*, pages 152–170. Springer, 2005.

[28] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[29] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 371–380. ACM, May 31 - June 2 2009.

[30] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *In Proceedings of the Third Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[31] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[32] Ehab ElSalamouny, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Generalized differential privacy: Regions of priors that admit robust optimal mechanisms. In Franck van Breugel, Elham Kashefi, Catuscia Palamidessi, and Jan Rutten, editors, *Horizons of the Mind. A Tribute to Prakash Panangaden*, volume 8464 of *Lecture Notes in Computer Science*, pages 292–318. Springer International Publishing, 2014.

[33] Ehab ElSalamouny and Sébastien Gambs. Optimal noise functions for location privacy on continuous regions. *International Journal of Information Security*, 2017.

[34] Kassem Fawaz, Huan Feng, and Kang G. Shin. Anatomization and protection of mobile apps' location privacy threats. In Jaeyeon Jung and Thorsten Holz, editors, *Proceedings of the 24th USENIX Security Symposium, (USENIX Security 2015)*, pages 753–768. USENIX Association, 2015.

[35] Kassem Fawaz and Kang G. Shin. Location privacy protection for smartphone users. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS 2014)*, pages 239–250. ACM Press, 2014.

[36] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *J. Comput. Syst. Sci.*, 80(8):1597–1614, 2014.

[37] Bugra Gedik and Ling Liu. Location privacy in mobile systems: A personalized anonymization model. In *Proc. of the 25th International Conference on Distributed Computing Systems (ICDCS)*, pages 620–629. IEEE Computer Society, 2005.

[38] Bugra Gedik and Ling Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Trans. Mob. Comput.*, 2008.

[39] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC)*, pages 351–360. ACM, 2009.

[40] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe, editors, *Proceedings of the 7th International Conference on Pervasive Computing (Pervasive 2009)*, volume 5538 of *Lecture Notes in Computer Science*, pages 390–397. Springer-Verlag, Nara, Japan, May 2009.

[41] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proc. of the First International Conference on Mobile Systems, Applications, and Services (MobiSys)*. USENIX, 2003.

[42] Marco Gruteser and Baik Hoh. On the anonymity of periodic location samples. In Dieter Hutter and Markus Ullmann, editors, *Proceedings of the Second International Conference on Security in Pervasive Computing (SPC 2005)*, volume 3450 of *Lecture Notes in Computer Science*, pages 179–192. Springer, 2005.

[43] Shen-Shyang Ho and Shuhua Ruan. Differential privacy for location pattern mining. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS (SPRINGL)*, pages 17–24. ACM, 2011.

[44] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. Cartel: A distributed mobile sensor computing system. In *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems*, SenSys '06, pages 125–138. ACM, 2006.

[45] Mobile data challenge dataset. `https://www.idiap.ch/dataset/mdc`.

[46] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 239–252. ACM, 2012.

[47] Oliver Jan, Alan Horowitz, and Zhong-Ren Peng. Using global positioning system data to understand variations in path choice. *Transportation Research Record: Journal of the Transportation Research Board*, 1725:37–44, 2000.

[48] John Krumm. Inference attacks on location tracks. In Anthony LaMarca, Marc Langheinrich, and Khai N. Truong, editors, *Proceedings of the 5th International Conference on Pervasive Computing (Pervasive 2007)*, volume 4480 of *Lecture Notes in Computer Science*, pages 127–143. Springer, 2007.

[49] John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[50] Location guard. https://github.com/chatziko/location-guard.

[51] Changsha Ma and Chang Wen Chen. Nearby friend discovery with geo-indistinguishability to stalkers. *Procedia Computer Science*, 34:352 – 359, 2014.

[52] Chris Y. T. Ma, David K. Y. Yau, Nung Kwan Yip, and Nageswara S. V. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking, (MOBICOM 2010)*, pages 185–196, 2010.

[53] Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In Gustavo Alonso, José A. Blakeley, and Arbee L. P. Chen, editors, *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, México*, pages 277–286. IEEE, 2008.

[54] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[55] Microsoft Research (2012) GeoLife trajectories (v. 1.3) . `https://www.microsoft.com/en-us/download/details.aspx?id=52367`.

[56] Microsoft Trustworthy Computing. Location Based Services Usage and Perceptions Survey, January 2011. `https://www.microsoft.com/en-us/download/details.aspx?id=3250`.

[57] Darakhshan J. Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N. Wright. DP-WHERE: differentially private modeling of human mobility. In Xiaohua Hu, Tsau Young Lin, Vijay V. Raghavan, Benjamin W. Wah, Ricardo A. Baeza-Yates, Geoffrey C. Fox, Cyrus Shahabi, Matthew Smith, Qiang Yang, Rayid Ghani, Wei Fan, Ronny Lempel, and Raghunath Nambiar, editors, *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 580–588. IEEE, 2013.

[58] Mohamed F. Mokbel, Chi-Yin Chow, and Walid G. Aref. The new casper: Query processing for location services without compromising privacy. In Umeshwar Dayal, Kyu-Young Whang, David B. Lomet, Gustavo Alonso, Guy M. Lohman, Martin L. Kersten, Sang Kyun Cha, and Young-Kuk Kim, editors, *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pages 763–774. ACM, 2006.

[59] Mirco Musolesi, Mattia Piraccini, Kristof Fodor, Antonio Corradi, and Andrew T. Campbell. CRAWDAD data set dartmouth/cenceme (v. 2008-08-13). `http://crawdad.cs.dartmouth.edu/dartmouth/cenceme`, 2008.

[60] Kevin Orland. Stalker Victims Should Check For GPS. The Associated Press, February 2003. `http://www.cbsnews.com/news/stalker-victims-should-check-for-gps/`.

[61] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1959–1972. ACM, 2017.

[62] Balaji Palanisamy and Ling Liu. Mobimix: Protecting location privacy with mix-zones over road networks. In Serge Abiteboul, Klemens Böhm, Christoph Koch 0001, and Kian-Lee Tan, editors, *Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany*, pages 494–505. IEEE Computer Society, 2011.

[63] Pew research center: Internet & technology – mobile fact sheet, January 2017. `http://www.pewinternet.org/fact-sheet/mobile/`.

[64] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD data set epfl/mobility (v. 2009-02-24). `http://crawdad.cs.dartmouth.edu/epfl/mobility`.

[65] Please Rob Me. `http://pleaserobme.com/`.

[66] Layla Pournajaf, Li Xiong, Vaidy Sunderam, and Xiaofeng Xu. Stac: Spatial task assignment for crowd sensing with cloaked participant locations. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '15, pages 90:1–90:4. ACM, 2015.

[67] QGIS Processing provider plugin. `https://github.com/SpatialVision/differential_privacy`.

[68] Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science*, 4(11), 2015.

[69] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, Nov 2000.

[70] Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng*, 13(6):1010–1027, 2001.

[71] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In ACM, editor, *PODS '98. Proceedings of the ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems, June 1–3, 1998, Seattle, Washington*, pages 188–188. ACM Press, 1998.

[72] Krishna Sampigethaya, Mingyan Li, Leping Huang, and Radha Poovendran. AMOEBA: Robust location privacy scheme for VANET. *IEEE Journal on Selected Areas in Communications*, 25(8):1569–1589, 2007.

[73] Zachary M. Seward. Tinder's privacy breach lasted much longer than the company claimed. Quartz Media LLC, July 2013. `https://qz.com/107739/tinders-privacy-breach-lasted-much-longer-than-the-company-claimed/`.

[74] Shashi Shekhar, Viswanath Gunturi, Michael R. Evans, and KwangSoo Yang. Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '12, pages 1–6. ACM, 2012.

[75] Reza Shokri. Privacy games: Optimal user-centric data obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2015(2):299–315, 2015.

[76] Reza Shokri, Julien Freudiger, Murtuza Jadliwala, and Jean-Pierre Hubaux. A distortion-based metric for location privacy. In *Proceedings of the 8th ACM Workshop on Privacy in the Electronic Society (WPES)*, WPES '09, pages 21–30. ACM, 2009.

[77] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *IEEE Symposium on Security and Privacy*, pages 247–262. IEEE Computer Society, 2011.

[78] Reza Shokri, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Quantifying location privacy: The case of sporadic location exposure. In *Proceedings of the 11th International Privacy Enhancing Technologies Symposium (PETS 2011)*, volume 6794 of *Lecture Notes in Computer Science*, pages 57–76. Springer, 2011.

[79] Reza Shokri, George Theodorakopoulos, and Carmela Troncoso. Privacy games along location traces: A game-theoretic framework for optimizing location privacy. *ACM Transactions on Privacy and Security*, 19(4):11:1–11:31, 2017.

[80] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Protecting location privacy: optimal strategy against localization attacks. In Ting Yu, George Danezis, and Virgil D. Gligor, editors, *Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS 2012)*, pages 617–627. ACM, 2012.

[81] Reza Shokri, Carmela Troncoso, Claudia Díaz, Julien Freudiger, and Jean-Pierre Hubaux. Unraveling an old cloak: k-anonymity for location privacy. In Ehab Al-Shaer and Keith B. Frikken, editors, *Proceedings of the 2010 ACM Workshop on Privacy in the Electronic Society, WPES 2010, Chicago, Illinois, USA, October 4, 2010*, pages 115–118. ACM, 2010.

[82] John Simerman. FasTrak to courthouse. East Bay Times, 2007. http://www.eastbaytimes.com/2007/06/05/fastrak-to-courthouse/.

[83] Yi Song, Daniel Dahlmeier, and Stéphane Bressan. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In Luo Si and Hui Yang, editors, *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security*, volume 1225 of *CEUR Workshop Proceedings*, pages 19–24. CEUR-WS.org, 2014.

[84] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[85] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[86] Kar Way Tan, Yimin Lin, and Kyriakos Mouratidis. Spatial cloaking revisited: Distinguishing information leakage from anonymity. In *Proceedins of the 11th International Symposium on Advances in Spatial and Temporal Databases (SSTD 2009)*. Springer, 2009.

[87] Galini Tsoukaneri, George Theodorakopoulos, Hugh Leather, and Mahesh K. Marina. On the inference of user paths from anonymized mobility data. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P 2016)*, pages 199–213. IEEE, 2016.

[88] Amit Kumar Tyagi and N. Sreenath. A comparative study on privacy preserving techniques for location based services. *British Journal of Mathematics & Computer Science*, 10(4):1–25, 2015.

[89] Urban Airship. https://www.urbanairship.com/.

[90] Ting Wang and Ling Liu. *From Data Privacy to Location Privacy*, pages 217–246. Springer, Boston, MA, 2009.

[91] Marius Wernke, Pavel Skvortsov, Frank Dürr, and Kurt Rothermel. A classification of location privacy attacks and approaches. *Personal and Ubiquitous Computing*, 18(1):163–175, 2014.

[92] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In Indrajit Ray, Ninghui Li, and Christopher Kruegel, editors, *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS 2015)*, pages 1298–1309. ACM, 2015.

[93] Toby Xu and Ying Cai. Feeling-based location privacy protection for location-based services. In *Proceedings of the 2009 ACM Conference on Computer and Communications Security (CCS 2009)*. ACM, 2009.

[94] Andy Yuan Xue, Rui Zhang, Yu Zheng, Xing Xie, Jin Huang, and Zhenghua Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *29th IEEE International Conference on Data Engineering (ICDE)*, pages 254–265. IEEE, 2013.

[95] Mingqiang Xue, Panos Kalnis, and Hung Pung. Location diversity: Enhanced privacy protection in location based services. In *Proc. of the 4th International Symposium on Location and Context Awareness (LoCA)*, volume 5561 of *Lecture Notes in Computer Science*, pages 70–87. Springer, 2009.

[96] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom 2011)*, pages 145–156. ACM, 2011.

[97] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800. ACM, 2009.

[98] Ge Zhong and Urs Hengartner. A distributed k-anonymity protocol for location privacy. In *Proceedings of the Seventh Annual IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE Computer Society, March 2009.

[99] Kathryn Zickuhr. Pew research center: Internet & technology – location-based services, September 2013. `http://www.pewinternet.org/2013/09/12/location-based-services/`.