The Long Road to Computational Location Privacy: A Survey

Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Lionel Brunie

The Long Road to Computational Location Privacy: A Survey*

Vincent Primault
University College London
United Kingdom
v.primault@ucl.ac.uk

Antoine Boutet
INSA Lyon, Inria, CITI
Villeurbanne, France
antoine.boutet@insa-lyon.fr

Sonia Ben Mokhtar
INSA Lyon, CNRS, LIRIS
Villeurbanne, France
sonia.benmokhtar@insa-lyon.fr

Lionel Brinoe
INSA Lyon, CNRS, LIRIS
Villeurbanne, France
lionel.brunie@insa-lyon.fr

***Abstract*—The widespread adoption of continuously connected smartphones and tablets developed the usage of mobile applications, among which many use location to provide geolocated services. These services provide new prospects for users: getting directions to work in the morning, leaving a check-in at a restaurant at noon and checking next day's weather in the evening are possible right from any mobile device embedding a GPS chip. In these location-based applications, the user's location is sent to a server, which uses them to provide contextual and personalised answers. However, nothing prevents the latter from gathering, analysing and possibly sharing the collected information, which opens the door to many privacy threats. Indeed, mobility data can reveal sensitive information about users, among which one's home, work place or even religious and political preferences. For this reason, many privacy-preserving mechanisms have been proposed these last years to enhance location privacy while using geolocated services. This article surveys and organises contributions in this area from classical building blocks to the most recent developments of privacy threats and location privacy-preserving mechanisms. We divide the protection mechanisms between online and offline use cases, and organise them into six categories depending on the nature of their algorithm. Moreover, this article surveys the evaluation metrics used to assess protection mechanisms in terms of privacy, utility and performance. Finally, open challenges and new directions to address the problem of computational location privacy are pointed out and discussed.**

I. Introduction

More and more users are carrying a handheld device such as a smartphone or a tablet and using it to access a wide variety of services on the go. Using these services, users can consult their bank accounts' balance from anywhere, book a table in a restaurant at any moment, track the status of their flight in real time or get a notification when their friends are nearby. This is possible thanks to the wide range of sensors available on these devices, giving them access to some knowledge about their environment. They are able to determine their location in real time and then use it to interact with geolocated services, often called *Location-Based Services* (LBSs for short). These services provide a contextual and personalised information depending on the current user's location. A multitude of LBSs have emerged these last years. We give here a non-exhaustive list of common usages that have been enabled by the rise of LBSs.

- *Directions & navigation applications:* These services allow users to get directions to (almost) any destination, and then to navigate towards it by simply following spoken instructions. Location data is used to provide real-time directions, recalculated as the user is moving. Well-known players here include Google Maps [1] and Waze [2].
- *Weather applications:* These services provide current weather conditions as well as forecasts. Location data is used to give the user relevant information for the city he is currently located in. Yahoo! Weather [3] is an application providing such a service on Android and iOS.
- *Venue finders:* These services give users information about interesting places in their vicinity. Most of the time, they include recommendations based on the experience of other users. Location data is used to show only places in immediate user's neighbourhood. Foursquare [4] and Yelp [5] are two applications helping to find such interesting places, with an added social dimension.
- *Social games:* These services turn any urban walk into an ever-changing game, where each new place becomes a new playground. Location data is used to make the game evolve depending on the user's city and his immediate surroundings, sometimes allowing to compete with nearby other users. Examples of such games are Pokemon GO [6] and City Domination [7].
- *Crowd-sensing applications:* These services enable participatory sensing, where a crowd of users use their smartphones to monitor their environment and share their results through an LBS server. Crowd-sensing benefits to a large variety of domains such as smart cities (e.g., the traffic monitoring application Nericell [8]) or health monitoring (e.g., PEIR [9]). APISENSE [10] and Funf [11] are two applications allowing to run crowd-sensing campaigns.

*This paper is to appear in IEEE Communications Surveys & Tutorials.

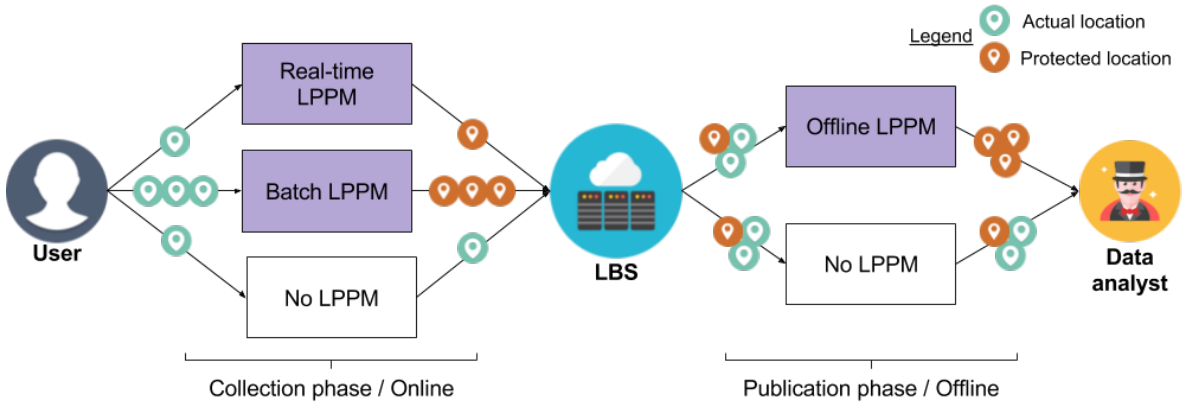


Fig. 1: Depending on the use case (i.e., real-time, batch, or offline), an LPPM can operate on location data during the collection or the publication phase.

Whatever their exact nature these services require users to disclose their location in order to make the application working as expected. This location disclosure nevertheless causes loss of control from users on their privacy and consequently allows LBSs to store the mobility and all places users are visiting over time. The sequence of all locations known to belong to a single user, along with the time at which the user was seen at each location, is called a *mobility trace*. Users are often not aware of the quantity of sensitive knowledge that can be inferred from their mobility traces. Analysing mobility traces of users can reveal their *points of interest* [12], which are meaningful places such as home or work. It can also reveal the other users they frequently meet [13], or lead to predicting their future mobility [14]. By combining mobility traces with semantic information [15], it is also possible to infer the actual user's activity (e.g., working, shopping, watching a film) or its transportation mode if the user is on the move. Besides the continuous tracking of a user's activities, points of interest can lead to leak even more sensitive information such as religious, political or sexual beliefs if one regularly goes to the headquarters of a political party, a worship place or a lesbian bar, respectively. As an example, it is possible to find out which taxi drivers are Muslim by correlating the time at which they are in pause with mandatory prayer times [16].

Needless to say, this large amount of mobility data is a gold mine for many companies willing to learn more information about users. The market related to LBSs is indeed enormous: the total revenue of the US-only LBS industry was already estimated to $75 billion in 2012 [17]. This high value of the mobility data leads many applications and companies to commercially exploit the collected data for analysis, profiling, marketing, behavioural targeting, or simply to sell this information to external parties. Moreover, the location of users is tracked and collected by many mobile applications with and without their consent [18], [19], [20], [21], [22], which aggravates the privacy threats related to sharing mobility data, voluntarily or not.

To mitigate these privacy problems, many *location privacy protection mechanisms* (LPPMs for short, or just *protection mechanisms* in this article) have been proposed in the literature. Their goal is to protect location privacy of users while still allowing them to enjoy geolocated services. There is a rich literature about existing LPPMs. Some of them are rather generic and can adapt to a lot of situations while others are very specific to a single use case. LPPMs rely on a wide array of techniques, ranging from data perturbation (e.g., [23], [24], [25]) to data encryption (e.g., [26], [27], [28]), and including fake data generation (e.g., [29], [30], [31]). In this survey, we distinguish between three classes of use cases for LPPMs as illustrated in Figure 1.

In *real-time use cases*, users query an LBS and expect an immediate answer. We include in this category the usage of navigation applications, weather applications, venue finders and social games. The main challenge for real-time LPPMs (e.g., [23], [24], [28]) is that they only have at their disposal actual and historical locations; they obviously do not know the future state of the system.

Offline use cases come into play once an LBS has collected mobility data and wants to publish it, whether it is for commercial or non-profit purposes (e.g., sharing mobility data with a marketing company or to release a dataset as open data). Instead of protecting locations on-the-fly, offline LPPMs (e.g., [32], [33], [34]) protect whole mobility datasets at once, possibly leveraging the knowledge of the behaviour of all users in the system to apply more efficient and subtle schemes.

In *batch use cases*, users regularly send their data to an LBS (e.g., every hour) and expect it to publish back aggregated results. This use case is typically adopted by crowd-sensing applications [35] and is a middle-ground situation between real-time and offline use cases. Batch LPPMs differ from real-time LPPMs in that they are less sensitive to latency, and they send more data at once. They also differ from offline LPPMs because they do not have

the global knowledge.

Figure 1 depicts an overview of the three aforementioned uses cases as well as the involved entities. As shown, we distinguish between two phases when using an LPPM: what happens *online* during the *collection*, i.e., between a user and an LBS, and what happens *offline*, during the *publication*, i.e., between an LBS and an analyst. Depending on their nature, LBSs fall either in the real-time or batch use cases. Furthermore, offline use cases appear as soon as one of these LBSs is willing to publish and to protect the gathered mobility data.

Protecting mobility data with an LPPM obviously improves the privacy but also impacts the quality of the resulting data. To evaluate LPPMs and to compare them together, researchers have proposed a large variety of metrics. These metrics can be divided in three categories. *Privacy metrics* quantify the level of privacy a user can expect while using a given LPPM. One popular way to evaluate privacy is to evaluate the effect of a privacy attack while being protected by an LPPM (e.g., [36]). *Utility metrics* measure the usefulness (also called quality of service) that can still be obtained while using an LPPM, which largely depends on the targeted LBS and the considered use case. There is an inherent trade-off between privacy and utility. Indeed, if no mobility data is sent, privacy is perfectly preserved, while the utility is null. Conversely, sending unprotected data results in a perfect utility at the cost of no privacy. Finally, *performance metrics* measure the algorithm efficiency or the cost of a given LPPM. Typical performance metrics are the execution time, the ability to scale or the tolerance to faults. Performance metrics are orthogonal and do not participate to the privacy and utility trade-off, but still are important because they impact the usability of LPPMs.

With the advent of pervasive computing offering new possibilities for tracking and collecting the location and the mobility of users, computational location privacy has become essential. We are only interested in computational location privacy, i.e., in threats and countermeasures performed by algorithms. This survey is not related to non-computational threats, which would for example come from a manual inspection and reasoning on mobility data, performed by a human brain. Computational location privacy has already been reviewed in several surveys. While both Krumm [37] and Shin et al. [38] published general surveys, Terrovitis [39] and Wernke et al. [40] followed an approach focused on location privacy attacks, and Chow et al. [41] focus on online LPPMs. However, only few of these articles address the evaluation of protection mechanisms, and when it is actually discussed, only privacy is considered without the complementary utility and performance metrics. Moreover, previous surveys often focus either on the online or the offline scenario but not both. In the context of Cognitive Radio Networks (CRNs), Grissa and al. [42] provide a comprehensive survey that investigates the various location privacy risks and threats as well, as countermeasures that have been proposed in the literature to cope with these location privacy issues. Lastly, Cottrill [43] uses a multidisciplinary approach to discuss location privacy and to ensure that the protection of private information is directly addressed from each of the relevant standpoints of law, policy, and technology.

In this survey, we provide an up-to-date vision over computational location privacy, including recent works like differentially private approaches [23] or privacy-by-design LBS architectures [44]. Figure 2 depicts the different areas covered by our survey. We review practical privacy attacks on mobility data, and survey state-of-the-art protection mechanisms. As presented in Figure 2, we organise LPPMs into three use cases (i.e., real-time, batch, or offline), and into six categories depending on the kind of algorithm (i.e., mix-zones, generalization-based, dummies-based, perturbation-based, protocol-based and rule-based). We furthermore discuss the evaluation of LPPMs with privacy, utility, and performance metrics, and report how the presented protection mechanisms have been evaluated by their authors. We also consider their architecture and associated impacts (e.g., added latency or integrability). Lastly, from the lessons we learned from our experience, we discuss open challenges and emerging visions in computation location privacy.

In this survey, we do not consider protection mechanisms only using anonymity, also called pseudonymity. These solutions consist in removing the link between an individual and its data by using a pseudonym instead of his real identity. Indeed, only relying on pseudonymization to protect data is not enough and this practice has resulted in several well-known privacy breaches these last years. (e.g., the identification of the governor of Massachusetts in an "anonymised" health dataset [45] or the re-identification of users from AOL web search logs [46] or a Netflix dataset [47]).

The remainder of this survey is structured as follows. We present in Section II the practical threats associated with location disclosure, before reviewing the evaluation metrics used to assess LPPMs in Section III. We present the different architectures adopted by LPPMs in Section IV. We then survey state-of-the-art protection mechanisms in Section V. We finally present some open challenges in Section VI before concluding in Section VII.

II. Privacy threats

Although LBSs potentially provide useful services, users are not always aware of the risks associated with the disclosure of their location during their daily life. For instance, the goal of the website *Please Rob Me* [48] is to "raise awareness about over-sharing". They use geolocated tweets to infer whether a user is at home or not, and hence if the way is free for potential thieves. In this section, we present the most important considered adversary models and the main practical threats for a user related to the exploitation of her mobility traces.

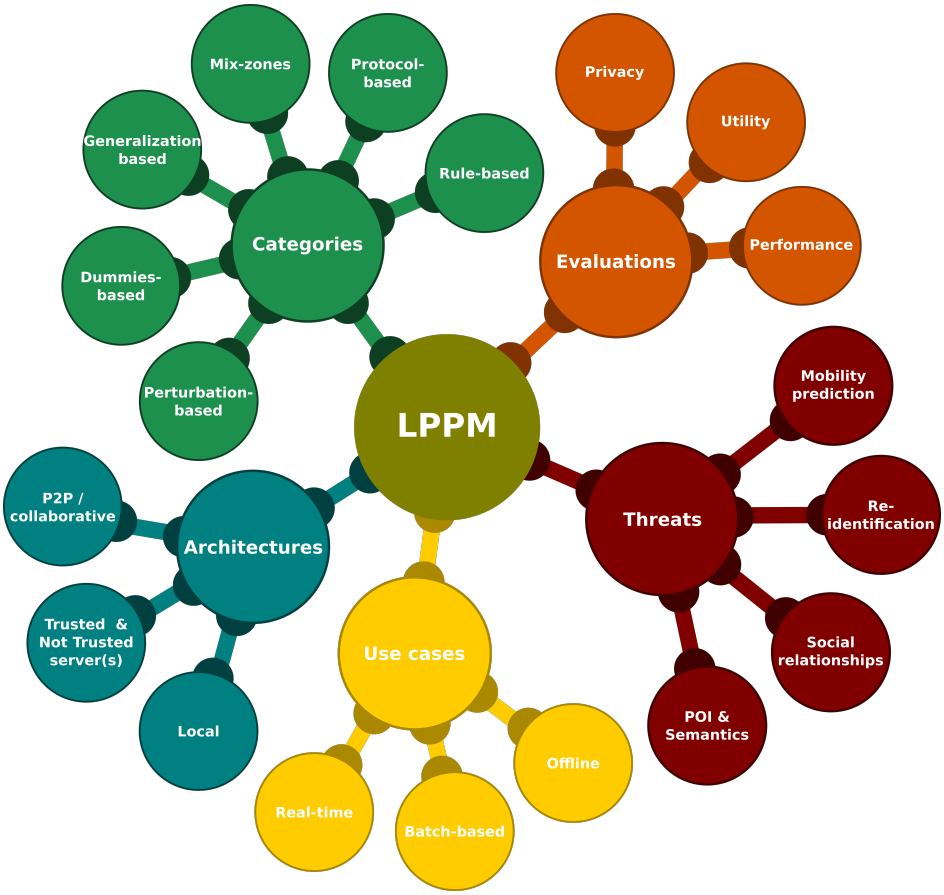


Fig. 2: Our survey reviews practical threats associated with location disclosure and surveys state-of-the-art LPPMs over different use cases (i.e., protecting data in real-time, operating by batches, or offline). LPPMs are structured in six categories and can adopt different architectures. We also detail their evaluation in terms of privacy and utility, as well as performance.

A. *Adversary models*

An adversary is anyone who can have an access to mobility data of one or several users. According to Figure 1, the adversary can be either the LBS itself or a data scientist who gets a dataset after it was published. In most of the cases, the adversary is considered to be *honest-but-curious* [49]. This means that the data scientist or the LBS behaves correctly (i.e., it provides the expected service) but it may exploit in all possible ways the information it receives. In particular, we assume that the adversary has access to any database containing additional information about the semantic of places or the associated activities, the topology of the road network, details about the public transportation lines (e.g., map, timetables), and so on. In addition, we also assume that the adversary may be able to collect external knowledge about each user in the system, modelled as a set of past mobility data. For instance, it can be used to run state-of-the-art re-identification attacks in order to re-associate the received data to a known user, or to predict future mobility. Moreover, it is also assumed that the LBS may identify that the client is relying on a LPPM while communicating with it. In some cases, it is further assumed that the adversary may collude with the proxy (i.e., typically when the proxy is not trusted) or different nodes in the system (e.g., using a Sybil attack [50]) in order to learn more information about the users.

B. *Points of interest & semantics*

Points of interest (POIs for short) are spatially delimited places where users spend some time. POIs can be home or work places, but also a swimming pool, a school or a cinema for instance. But they can also be even more sensitive places like a religious monument where a user regularly goes, the headquarters of a political party he is involved in or a hospital he is cured in. They can be extracted from mobility traces quite easily by using simple clustering algorithms like the ones presented in [51], [52]. Figure 3 shows an example of the behaviour of such an algorithm designed to extract POIs. Using APIs such as the Google Places [53] for instance can also provide details about each POI such as its precise address, the associated activity or the shops located at this location.

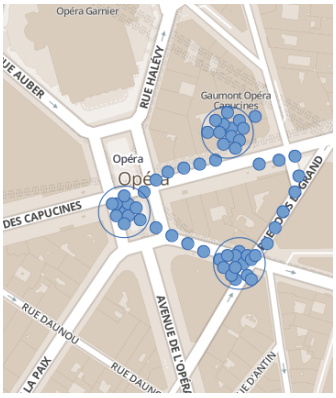


Fig. 3: Three POIs have been extracted from this mobility trace by using a clustering algorithm.

Gambs et al. [12] made an attack on a dataset containing mobility traces of taxi drivers in the San Francisco Bay Area. By finding points where the taxi's GPS sensor was off for a long period of time (e.g. 2 hours), they were able to infer POIs of the drivers. In some cases, they were able to locate a home and even to confirm it by using a satellite view of the area showing the presence of a yellow cab parked in front of the supposed driver's home. It was possible to infer a plausible home in a small neighbourhood for 20 out of 90 mobility traces analysed. Deneau [16] created a visualisation tool to analyse the active and inactive periods of taxi drivers over the day. By correlating their time of inactivity with the five times of prayer per day observed by practising Muslims, it was possible to identify drivers that could be Muslims.

The new development of machine learning brings huge security risks on location privacy. For instance, machine learning has showed his effectiveness to learn the semantics of some place. Krumm introduced *Placer* [15], a system using machine learning to automatically label places into 14 categories (home, work, shopping, transportation, place of worship, etc.). Author reported an overall accuracy of 73 %, mostly thanks to home and work places which are the easiest ones to label because this is where people spend most of their time. Riederer et al. developed *FindYou* [54], which aims to raise user awareness on the privacy issues surrounding the collection and use of location data. FindYou allow users to import their own location data from popular social networks and audit them. By leveraging the knowledge provided by the US Census Bureau and unsupervised machine learning methods, this personal location privacy auditing tool provides prediction on the home place of users as well as their age, income and ethnicity. Huguenin et al. [55] leveraged machine learning to infer the motivation (e.g., "Inform about activity", "Share mood" or "Wish people to join me") behind check-ins of users on Foursquare. They achieved up to 63 % of accuracy when predicting a coarse-grained motivation.

C. Social relationships

With several mobility traces, it is possible to compare them and infer relationships between users. The idea is very simple: if two (or more) persons spend some time within the same area at the same moment, they are likely to be related by some social link. Bilogrevic et al. [56] studied this threat by using malicious Wi-Fi access points deployed on the EPFL campus (in Switzerland) that were able to locate devices communicating with them. By setting appropriate thresholds, they could detect meetings between people. Then, they used training data to get a characterisation of the social link between students, thus finding out if they were classmates, friends or other depending on the place where they met and the time they spent together. With the optimal experimental parameters, the authors experimentally obtained a true positives rate of 60 % and a false positives rate of 20 %. Similarly, Wang et al. [57] studied how access points could infer social relationships between users. They designed a decision tree classifying relationships by using the interaction time, inferred activity type (i.e., home, work, leisure) and physical closeness of people. Their classifier exhibits a success rate above 85 %.

D. Re-identification

Mobility traces can ultimately lead to re-identifying physical users, i.e., associating an identity to each trace. Krumm [58] used two months of mobility traces and tried to infer users' home address with various heuristics. By using white pages, it was possible in some cases to associate a person to a mobility trace. The most accurate heuristic gave 9 correct re-identifications out of 172 drivers. Gambs et al. [59] proposed a re-identification approach based on mobility Markov chains. They are used to model mobility patterns of users, more specifically the transitions between POIs. They designed different distance measures to quantify the similarity between two Markov chains and used them to re-identify users. They achieved up to 45 % of good matching, which was significantly better than other state-of-the-art attacks. Tockar [60] showed it was possible to stalk at celebrities by using a taxi trips dataset and some specifically crafted queries. By extracting drop-off addresses of people frequently spending their night in a club, and by using Google and Facebook to get more information about these addresses, he was able to pinpoint certain individuals with a high probability. Pyrgelis [61] et al. studied the feasibility of membership inference attacks on aggregate mobility datasets (i.e., datasets featuring how many users where within specific regions during specific periods of time). They modeled the problem as a game, and trained a machine learning classifier on prior knowledge and used it to infer whether particular individuals were part of a mobility dataset. Powerful attackers reached an Area Under Curve up to 0.83 or even 1.0, depending on how much background knowledge they have at their disposal.

These results are mainly possible due to the high degree of uniqueness of human mobility. Indeed, De Montjoye et al. [62] showed that only four randomly chosen spatio-temporal points are sufficient to almost uniquely identify a user hidden among the crowd. It means that the mobility of every individual acts like a fingerprint, even among a large number of users (they used a dataset containing 1.5 million users). In the same manner, Golle et al. [63] studied the uniqueness of the home/work pair. When revealing the census district where people live and work, it is possible to uniquely identify 5% of them, and for 40% of them, they are only hidden among 9 other people. If the census block where they live and work is disclosed, nearly all people can be uniquely identified. Zang et al. [64] improved the previous study by considering top-N locations of a large set of call data records of a US nation-wide cell operator. With the most precise location that can be grabbed from these records, which is the sector where the person stands, it is possible to uniquely identify 35% of the users by using their top-two locations and 85% of them by using their top-three locations. We also demonstrated the highly unique nature of mobility traces from different sensors [65]. Among other observations, authors noticed that the temporal dimension is as discriminative as the spatial dimension and that the uniqueness degree of mobility traces is user-dependent. They also pointed out that it is still possible to re-identify users with a high success rate using an appropriate attack even if the data was protected by classical LPPMs. Manousakas et al. [66] showed it might be even worse, as even if removing all spatial and temporal information from the graph of visited places, the topology of the latter graph is itself often uniquely identifying. They evaluated their approach with a dataset of 1500 users and found out that about 60 % of the users are uniquely identifiable from their top-10 locations, and this percentage increases to 93 % in the case of a directed graph. With a directed graph, 19 locations are needed to uniquely identify each of the 1500 users.

Recently, Wang et al. [67] explored the discrepancies between the theory and practice of re-identification attacks. They leveraged a large ground-truth dataset containing 2 million users, and two smaller external datasets collected over the same population to match against the former. They first evaluated seven state-of-the-art algorithms, showing that with those datasets, the best performing ones only achieved a re-identification rate of 20 %, which is far from the theoretical bound announced by De Montjoye [62] and others. By further analysing the results, it appeared that there was large spatio-temporal mismatches, whose effect was underestimated. Finally, they proposed four new re-identification attacks addressing the previously mentioned concerns, improving the re-identification of 17 % as compared to previous state-of-the-art algorithms.

E. Future mobility prediction

Knowing past mobility of a user can help to model his habits and hence allow to predict where he will be in a future time. Noulas et al. [68] focused on Foursquare check-ins. They extracted features from users mobility and used machine learning algorithms to predict places where users were likely to leave the next check-in. Precision was maximal during morning and at noon, when they achieved an accuracy of 65 %. It was harder to predict the next check-in in the night, during which accuracy dropped to 50 %. Sadilek and Krumm [14] proposed *Far Out*, a system to predict the location of a user in the long term, i.e., in a far away future date and within a time window of one hour. They leveraged Fourier analysis and principal component analysis to extract repetitive patterns from mobility data. These patterns were associated to a week day and a hour in the day. Their system featured an accuracy in their predictions ranging from 77 % to 93 %. Another threat coming from the analysis of mobility traces is the inference of users' mobility patterns. Gambs et al. [69] modelled movement habits of people by using Markov chains. Each frequent POI becomes a state in the chain and a probability is assigned to each possible transition. They split the day into several temporal slices and differentiated between week days and week-ends. With a dataset spanning over a sufficient period of time, it became possible to predict future users' movements. Agir et al.[70] studied the prediction of the next check-in of users by using Bayesian networks. They evaluated their approach with data of 1065 users, collected from tweets that have been generated from Foursquare. Across 6 major cities, the median accuracy when predicting the next check-in is between 100 and 150 metres. The authors also studied the impact of semantic information on this kind of attack. They found out that by incorporating semantics, the median accuracy of their predictions decreased by 10 to 115 metres.

III. Evaluating LPPMs

Unfortunately, there is no standard to evaluate and compare LPPMs. To the best of our knowledge, only the work of Shokri et al. [36] focuses on the evaluation of LPPMs. Although it is only interested in quantifying privacy, it defines solid foundations towards building a complete evaluation methodology. In this section, we review the different evaluation metrics used in the literature to assess LPPMs in a quantitative manner. We start by introducing two classical privacy notions in Section III-A. We then group and present evaluation metrics through three complementary categories, namely privacy metrics in Section III-B, utility metrics in Section III-C and performance metrics in Section III-D. We discuss the inherent trade-off between privacy and utility in Section III-E. We conclude this section by surveying mobility datasets commonly used to conduct practical evaluations in Section III-F.

A. Privacy models

Two general privacy models have emerged and have been widely adopted by the community, and are still the foundation for most of subsequent works [71]. Those models propose generic privacy guarantees that were originally not specific to location privacy, but have been later successfully applied to location privacy. In this subsection (and only this one), the concept of dataset is not limited to mobility datasets but to generic datasets, i.e., a list of records with attributes.

1) k-anonymity: The model of k-anonymity has been introduced by Sweeney in 2002 [45]. The idea is to prevent one to uniquely identify individuals from a small subset of their attributes, called a *quasi-identifier*. The subset of attributes to protect, which is not part of the quasi-identifier, form the *sensitive attributes*. For instance, within medical records, the birth date, sex and zip code triplet is a quasi-identifier that is enough to uniquely identify some individuals, while the disease is a sensitive attribute. k-anonymity states that to be protected, a user must be indistinguishable among at least $k-1$ other users. To achieve that, all k indistinguishable users must have the same values for all attributes forming their quasi-identifier. This makes them look similar and forms what is called an *anonymity group*. Therefore, the probability of an attacker without external knowledge to re-identify someone among k similar users is at most $1/k$.

Definition 1: Let d be a sequence of records with n attributes $a_1, ..., a_n$ and $Q_d = \{a_i, ..., a_j\} \subseteq \{a_1, ..., a_n\}$ be the quasi-identifier associated with d. Let d_k be the k-th record of d and $r[Q_d]$ the projection of record $r \in d$ on Q_d, i.e., the $|Q_d|$-tuple formed of values for only the attributes of Q_d in r. d is said to satisfy k-anonymity if and only if each unique sequence of values in the quasi-identifier appears with at least k occurrences in d, or formally:

$$\forall s \in \{r[Q_d] \,|\, r \in d\}, |\{i \in \mathbb{N} \,|\, d_i[Q_d] = s\}| \geq k$$

TABLE I: A dataset with k-anonymity where $k = 2$.

Birth	Sex	Zip	Disease
1970	M	0247	Migraine
1970	M	0247	Chest pain
1970	F	0247	Asthma
1970	F	0247	Migraine
1970	F	0247	Asthma
1969	M	0232	Appendicitis
1969	M	0232	Appendicitis

For example, Table I shows a sample of a medical dataset exposing a k-anonymity guarantee, where the quasi-identifier is $\{Birth, Sex, Zip\}$ and the sensitive attributes are $\{Disease\}$, for $k = 2$. Here, there are three unique $\{Birth, Sex, Zip\}$ triplets, i.e., $\langle 1970, M, 0247\rangle$, $\langle 1970, F, 0247\rangle$ and $\langle 1969, M, 0232\rangle$. For each of those triplets, there are respectively two, three and two different records. Consequently, there is a minimum of two different records for each triplet of values taken by the quasi-identifier: this table guarantees 2-anonymity. This way, knowing the birth year, sex and zip code of some individual should not leak his disease, as there is at least one other person with the same quasi-identifier.

However, despite providing 2-anonymity, there is a problem in Table I for male patients born in 1969 and living in the area with 0232 zip code (i.e., the last two records). Indeed, they share the same value for their sensitive attribute (i.e., they have the same disease), which leaves them unprotected. This concern has been addressed by the introduction of ℓ-diversity [72]. It extends k-anonymity by additionally enforcing that within anonymity groups, there should be at least ℓ well-represented values. More precisely, it enforces a particular distribution of values for sensitive attributes across each anonymity group. This well-represented notion is formally defined in three different ways in [72]. The simplest one is called distinct ℓ-diversity and states that there must be at least ℓ distinct values for each sensitive field for each anonymity group.

t-closeness [73] is a further extension of ℓ-diversity Instead of just guaranteeing a good representation of sensitive values, this approach enforces that the distribution of every sensitive attribute inside anonymity groups must be the same than the distribution of this attribute in the whole dataset, modulo a threshold t.

2) Differential privacy: Differential privacy is a more recent model introduced by Dwork [74] defining a formal and provable privacy guarantee. The idea is that an aggregate result computed over a dataset should be almost the same whether or not a single element is present inside the dataset. In other words, the addition or removal of one single element shall not change significantly the probability of any outcome of an aggregate function. Unlike k-anonymity, the differential privacy definition is not affected by the external knowledge an attacker may have.

Definition 2: Let $\epsilon \in \mathbb{R}^{+*}$ and $\mathcal{K}$ be a randomized function that takes a dataset as input. Let $image(\mathcal{K})$ be the image of $\mathcal{K}$. $\mathcal{K}$ gives ϵ-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and for all $S \subseteq image(\mathcal{K})$,

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{K}(D_2) \in S]$$

For example, Table II shows two versions of a sample dataset listing whether individuals are subject to chronic migraines. Let us suppose that an analyst has access to these two datasets, and to a query Q that takes a dataset as input and returns the number of persons having chronic migraines. By computing $Q(D_2) - Q(D_1) = 3 - 2 = 1$, our curious analyst can infer that Joe is indeed subject to chronic migraines.

Several methods have been proposed to practically achieve differential privacy. We present one of them, called the Laplace mechanism, that can is used for numerical values, and hence in the location privacy context. It relies

TABLE II: Two datasets differing on one single element.

Name	Has chronic migraines
Agatha	True
Anna	False
John	True
Mark	False
Mary	False

(a) Dataset D_1, without Joe.

Name	Has chronic migraines
Agatha	True
Anna	False
Joe	True
John	True
Mark	False
Mary	False

(b) Dataset D_2, with Joe.

on adding random noise, whose magnitude depends on the *sensitivity* of the query function issued on the dataset. Intuitively, the sensitivity of a query function quantifies the impact that the addition or removal of a single element of a dataset could have on the output of this function.

Definition 3: Let f be a function that takes a dataset as input and produces a vector of reals, i.e., $f : \mathcal{D} \longrightarrow \mathbb{R}^n, n \in \mathbb{N}$. Let D_1 and D_2 be two datasets differing on at most one element. The sensitivity of f is noted Δf and defined, for all such datasets D_1 and D_2, as:

$$\Delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1.$$

The sensitivity is defined independently of the underlying data, and only depends on the function under consideration. In particular, for queries that are counting records (such as Q in our previous example), $\Delta Q = 1$ because the addition or removal of a single record affects the count result by increasing or decreasing its value by 1. Then, the Laplace mechanism adds Laplacian noise with mean 0 and scale parameter $\Delta f/\epsilon$ to the query's result[1]. Consequently, the ϵ-differentially privacy version of Q is defined as $\hat{Q}(D) = Q(D) + Y$, where $Y \sim \text{Lap}(1/\epsilon)$. That way, computing $Q(D_2) - Q(D_1)$ does not automatically result in 1, because of the added Laplacian noise. The Laplace mechanism is of course only suitable for queries producing numerical results; another method exists for categorical values [75], but it is outside of the scope of this paper.

Differential privacy supports the composition of functions, and the potential information leakage resulting of this composition can be quantified. In the general case, when applying n randomized independent algorithms $\mathcal{K}_1, ..., \mathcal{K}_n$ that provide $\epsilon_1, ..., \epsilon_n$-differential privacy, any composition of those algorithms provides $(\Sigma_i \epsilon_i)$-differential privacy. This is known as *sequential composition*. This protection model assumes that each analyst has a global privacy budget. In an interactive mode (i.e., online LPPMs), each time she issues an ϵ-differentially private query, his privacy budget is reduced by ϵ. Once the budget is totally consumed, all subsequent queries from this analyst should be rejected. It models the fact that once an information is learnt, it cannot be forgotten. In practice, determining this privacy budget and its instantiation (global, per user, etc.) remains largely an open question. Recent works (e.g., [76]) address this question.

Differential privacy has generated an important literature these last few years with new models and inter-model connections [71], as well as new techniques such as randomised response [77] and its combination with sampling [78] which achieves zero-knowledge privacy [79] (a privacy bound tighter than differential privacy).

[1]Proof of this is provided in [74].

B. Privacy metrics

To quantify the level of protection offered by an LPPM, we identify three categories of privacy metrics.

- *Formal guarantee* metrics adopt a theoretical approach to quantify the effect of an LPPM on mobility data. They use a well-defined and unambiguous framework to guarantee that a protected dataset has a certain level of privacy. As of now, there are two such guarantees commonly offered by LPPMs: k-anonymity and differential privacy (cf. Section III-A). *k-anonymity*, applied to location privacy, states that during a given time window and inside a given area, there should be at least k users. LPPMs then take different approaches to enforce this guarantee, for example by allowing users to specify the size of these areas or time windows as parameters, or by automatically adjusting them, such as they contain k users. *ϵ-differential privacy* has been instantiated differently by different LPPMs. Usually, instead of protecting the presence or absence of individual users, as it is the case with classical differential privacy, LPPMs attempt to protect the presence or absence of individual locations. Hence, the goal is not anymore to hide that a user is part of a dataset, but to hide where she went.
- *Data distortion* metrics compare privacy-related properties of mobility data before and after applying an LPPM on it. Indeed, using an LPPM is expected to hide sensitive information that was otherwise possible to obtain from actual mobility data. Examples of such metrics include computing the entropy of protected data or evaluating whether POIs can still be retrieved.
- *Attack correctness* metrics evaluate the impact of a location privacy attack that could be ran by an adversary in order to gain knowledge about users (see Section II for the list of potential attacks). Shokri et al. [36] did an extensive work on the usage of

attacks to quantify location privacy. They distinguish between three axes when evaluating the effectiveness of an attack: certainty, accuracy and correctness. Certainty is about the ambiguity of the attack's result; for example there is some uncertainty if a re-identification attack outputs three possible users, while the uncertainty is null if the same attack outputs a single user (independently of whether it is the correct answer). Accuracy is about taking into account that the attacker does not have unlimited computational resources; consequently, the output of his attack may be only an approximate response, e.g., by only taking into account a sample of all data at his disposal. Correctness quantifies the distance between the attack's result and the truth; it is what actually quantifies location privacy. An LPPM is expected to mitigate privacy attacks and lower (or even suppress) their harmful effects. As opposed to data distortion metrics, attack correctness metrics do not compare the effect of an attack before and after applying an LPPM, but rather evaluate directly the attack on a protected dataset, and use the actual dataset as ground truth to evaluate whether the attack was successful.

Very recently, a survey has specifically focused on reviewing and discussing privacy metrics [80].

C. Utility metrics

To evaluate the quality of protected mobility data, we identify two categories of utility metrics.

- *Data distortion* metrics compare utility-related properties of mobility before and after applying an LPPM on it. Indeed, we expect that the LPPM will not distort all properties of a dataset and make it unusable. Examples of such metrics include evaluating the spatial and temporal imprecision and comparing the covered area. It is of purpose that we name this category the same way as for privacy metrics, because they do represent the same thing, but applied on different properties (privacy- or utility-related). If we go even further, it happens that some data distortion metrics are used one time as a privacy metric and the other time as a utility metric[2].
- *Task distortion* metrics compare the result of some practical task on the data before and after applying an LPPM. For instance, these metrics can be interested in data mining tasks or analytics queries.

D. Performance metrics

Protecting mobility data can be resources greedy. To evaluate the performance an LPPM, four categories of metrics are commonly used.

[2]A common example is a metric whose goal is to compare the distance between actual locations and protected locations. It can be viewed either as a privacy metric, because by distorting locations we hide where users were, or as a utility metric, if the LBS that we use or the task that the analyst wants to run requires spatial precision.

- *Execution time* is a simple quantification of the time it takes for an LPPM to protect data. Of course, it does not have the same impact for real-time use cases, where a response is expected in a very short time frame (a few milliseconds, a few seconds at most), than for batch or offline use cases that do not expect an immediate answer. However, even for the latter, it is of importance as computational resources have a cost ("time is money"). This execution time can be measured in various ways, for example in seconds or in CPU cycles.
- *Communication overhead* metrics quantify the negative impact of applying an LPPM on the quantity of information that will be produced and exchanged through the network in online use cases. For online use cases, some LPPMs need to exchange more messages, or more answers are received from the LBS. Obviously, it has an impact on the execution time, but it can be measured separately. For offline use cases it is related to the size of the protected dataset; if bigger or more complex that the actual one, it can slow down the job of analysts and affect their experience when working with the dataset.
- *Energy overhead* metrics measure the negative impact on the battery lifetime implied by using a given LPPM, when running it as an application on a mobile device. It is important to be quantified because it impacts the usability and adoption by end users. It is only applicable to online LPPMs.
- *Scalability* measures how well an LPPM can face a high workload. For online LPPMs, scalability metrics are mostly related to the capability of handling a high volume of concurrent requests, while for offline LPPMs it concerns the ability to deal with datasets of large sizes.

E. Trade-off between utility and privacy

Protecting mobility data by an LPPM improves the privacy but impacts the quality of the resulting data: this is the trade-off between privacy and utility. The more the information is altered (e.g., modified or deleted), the less the protected data may be exploited. The configuration of this trade-off (i.e., defining the levels of privacy and utility required for the protected data) closely depends on the considered use case. For instance, a weather application requires a less precise location (it can accommodate with only a city name) than a navigation application (it needs to know in which street a users is located). Consequently, both the privacy and the utility evaluations should be tailored to fit the actual use case.

LPPMs are usually configured through various configuration parameters, which greatly impact the resulting privacy and utility, and it is an uneasy task to correctly define LPPM configuration parameters. As an example, Wait for Me [32] takes at least five parameters, with some labeled as the "initial maximum radius used in clustering"

TABLE III: Common datasets of mobility traces.

Dataset	Location	Time span	#users	#events
Cabspotting	San Francisco, USA	1 month	536	11 million
MDC	Geneva, Switzerland	3 years	185	11 million
Privamov	Lyon, France	15 months	100	156 million
Geolife	Beijing, China	5,5 years	178	25 million
T-Drive	Beijing, China	1 week	10,357	15 million
Brightkite	Worldwide	1.5 years	58,228	4 million
Gowalla	Worldwide	1.5 years	196,591	6 million

or the "global maximum trash size". While there are useful to fine-tune the behaviour of the algorithm, we do not expect final users to read the paper to understand what the trash is or how the clustering works. Even the single ϵ parameter of geo-indistinguishability [23] is tricky to configure, because it is expressed in meters^{-1} and its impact is exponential. Similarly, it is difficult for a final user who knows (usually) nothing about differential privacy to set it appropriately.

Recent works have been conducted to select and configure the best LPPM to use according to a set of objectives set by the user in term of utility and privacy. For instance, ALP [81] relies on a greedy approach that iteratively evaluates the privacy and utility, thus refining the values of configuration parameters at each step. While ALP can be used for online and offline LPPMs, PULP [82] proposes a framework allowing to automatically choose and configure offline LPPMs. To do that, PULP explores and models the dependency that exists, for different LPPMs, between their configuration parameters and the privacy/utility metrics that one wants to maximise.

F. Mobility datasets

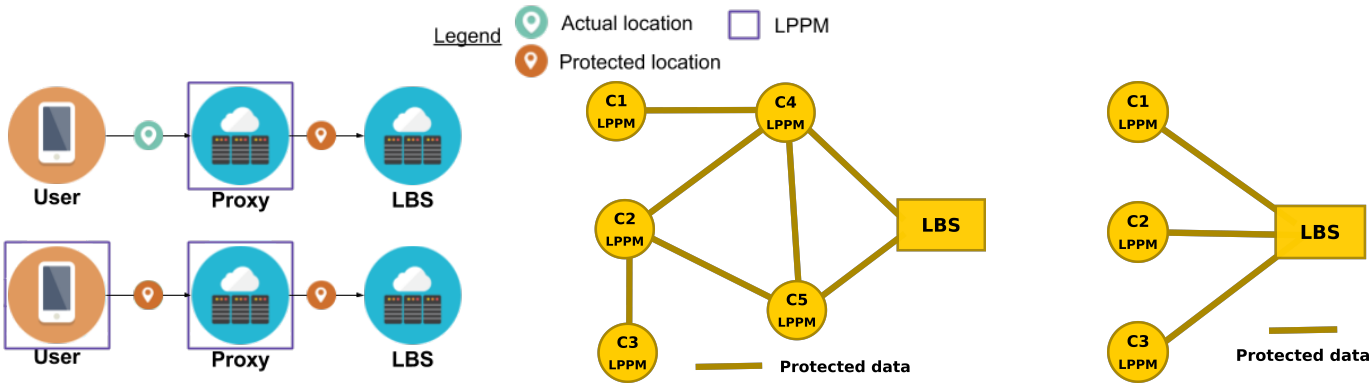
To evaluate and compare protection mechanisms, we need to assess their effectiveness on mobility datasets, preferably including real mobility data of real users. To this end, several initiatives have been conducted to publicly provide datasets coming from real-life data collections. Table III lists some of the most common datasets used in the literature.

The Cabspotting dataset [83] contains GPS traces of taxi cabs in San Francisco (USA), collected in May 2008. The Geolife dataset [84] gathers GPS trajectories collected from April 2007 to August 2012 in Beijing (China). The MDC dataset [85], [86] involves 182 volunteers equipped with smartphones running a data collection software in the Lake Geneva region (Switzerland), collected between 2099 and 2011. A privacy protection scheme based on k-anonymity has been performed on the raw data before releasing the MDC dataset. As described in [85], this privacy preserving operation includes many manual operations which have obviously an impact on the outcome of LPPMs, but these impacts are difficult to fully understand. It includes not only locations coming from the GPS sensor, but also data from various other sensors (e.g., accelerometer, battery). The Privamov dataset also gathers mobility data from multiple sensors (i.e., accelerometer, WiFi, cellular network). This data collection took place from October 2014 to January 2016 and involves 100 students and staff from various campuses in the city of Lyon equipped with smartphones [87]. T-Drive [88], [89] is another dataset collected in Beijing and featuring taxi drivers. It features a high number of users (more than 10,000) over a very short period of time (one week).

Other datasets come from geolocated social networks, rather than from a custom data collection campaign ran by academics. These social networks allowed users to leave "check-ins" to places where they went, thus allowing to build sparse mobility traces for these users. Two datasets are available in this category, coming from the (now closed) Brightkite and Gowalla [90], [91] social networks. They contain 4 million (respectively 6 million) check-ins collected between February 2009 and October 2010. These datasets present a different kind of mobility data along with relationships between users in the network.

Lastly, another approach is to use *synthetic datasets*, i.e., randomly generated mobility datasets. Such generators include *BerlinMOD* [92], Brinkhoff's generator [93] and Hermoupolis [94]. Because they are generated, these datasets may not always model very realistically the human mobility and all the hazards attached to it. However, this approach allows to use datasets of any size, for example to assess the scalability of algorithms to large-scale datasets. Indeed, very few researchers have access to extremely large datasets (e.g., the dataset with 1.5 million individuals used by De Montjoye et al. [62] or the dataset of 2.1 million individuals used by Wang et al. [67]).

Past experimental results have demonstrated that datasets can have an important impact on the evaluation of an LPPM. LPPMs providing k-anonymity are among the most sensitive to this aspect. For example, it is far easier to likely provide k-anonymity with an important value of k with a large dataset than a small one. Similarly, the sparsity of datasets will also be of high importance in k-anonymous LPPMs, because this protection scheme is particularly sensitive to the co-location of users (i.e., users being at the same place at the same time). From one evaluation [32] to another [95] of the same LPPM, results can indeed largely vary, due to the considered



(a) Trusted and Not Trusted Third Party (TTP up, NTTP bottom) – Online

(b) Peer-to-Peer (P2P) – Online

(c) Local – Online or Offline

Fig. 4: Four different architectures can be adopted by LPPMs: Trusted Third Party (TTP), Non-Trusted Third Party (NTTP), Peer-to-Peer (P2P), and Local.

datasets and their associated features. In the former case, the dataset is generated using Brinkhoff's generator, and represents one day of mobility with 4.7 million events. In the latter case, the real-life datasets Geolife, Cabspotting and MDC are used, which contain between 11 and 25 million events, but scattered across several weeks or even years. Another factor, besides the size of a datasets, is the kind of users that it contains. Indeed, the taxi drivers inside the Cabspotting dataset will likely have a different behaviour that the ordinary people who are part of the MDC dataset.

IV. Architectures of LPPMs

In addition to their different use cases, LPPMs can leverage four different architectures. These architectures are depicted in Figure 4. The local architecture is used by both online and offline LPPMs, while the Trusted Third Party (TTP), the Non-Trusted Third Party (NTTP), and Peer-to-Peer (P2P) architectures are only used by online LPPMs.

- The *TTP* architecture requires a trusted third party proxy server. It means there is an external entity that has access to the actual data coming from all users.
- The *NTTP* architecture still involves one or several third party servers but they do not need to be trusted. However, the LPPM is designed in such a manner that this third party cannot represent a privacy threat, even if malicious or colluding with the LBS. In this scheme, the behaviour of the LPPM is usually split and implemented on both the client and the proxy server.
- The *P2P* architecture requires no external server, but it requires users devices taking part in the system to exchange information in a peer-to-peer fashion in order to protect their data. Such LPPMs engage users devices in a collaborative privacy protocol before they send their data to an LBS.
- The *local* architecture does not require any communication with another party to protect data. LPPMs entirely autonomous and process everything locally, on the device on which they are executed (i.e., a users devices or a server operated by the LBS). They may need access to external databases, in which case the latter are expected to be entirely available locally.

V. Location privacy preservation

In order to mitigate location privacy threats, LPPMs have been introduced. Their goal is to transform mobility data in order to protect it and prevent threats such as the ones presented in Section II. As presented in Section I, we distinguish between two main use cases of LPPMs: *online* and *offline*. In the online case, the LPPM protects either on-the-fly or by batch the mobility data before it even reaches the LBS. In the offline case, the LPPM is applied on an entire dataset before its publication.

In this section, we survey existing works about LPPMs, and organise them into six categories. We summarise this categorisation in Table IV and Table V for online and offline LPPMs, respectively. Moreover, we indicate for each LPPM its architecture and the categories of metrics that were used to evaluate it by its authors. For the sake of completeness, we distinguish in these tables between differential privacy and k-anonymity for privacy formal guarantees, and mention when an ad-hoc metric was used to evaluate LPPMs. Ad-hoc metrics encompass metrics that do not fit in our classification, usually because they measure something that is unique to the considered LPPM (e.g., something related to its algorithm and that cannot be made generic to all LPPMs).

A. Mix-zones

Mix-zones are a concept introduced by Beresford and Stajano. [144], taking its roots in the seminal work of Chaum [145] about mix networks. The mix-zones model applies to mobile users communicating with LBSs, by

TABLE IV: List of *online* LPPMs studied in this survey, with metrics used by their authors to evaluate them.

		Privacy					Utility			Performance			
Protection mechanism	Architecture	Differential privacy	k-anonymity	Attack correctness	Data distortion	Ad-hoc metric	Data distortion	Query distortion	Ad-hoc metric	Execution time	Communication cost	Energy overhead	Scalability
Mix-zones													
Beresford et al. [96]	TTP			✓						✓			
Freudiger et al. [97]	TTP			✓					✓				
Traffic-aware mix-zones [98]	TTP			✓					✓	✓			
MobMix [99]	TTP			✓						✓			✓
Gong et al. [100]	TTP/P2P								✓				✓
Generalization-based mechanisms													
CliqueCloak [101]	TTP		✓					✓		✓	✓		
Casper [102]	TTP		✓							✓			
P2P cloaking [103]	P2P		✓					✓		✓	✓		
PRIVÉ [24]	P2P		✓			✓				✓	✓		✓
PrivacyGrid [104]	TTP		✓					✓		✓			✓
Xu and Cai [105]	TTP					✓			✓	✓		✓	✓
Agir et al. [106]	Local				✓	✓	✓						
Ngo et al. [107]	Local	✓			✓				✓	✓			
ReverseCloak [108]	TTP		✓	✓		✓				✓			
Huguenin et al. [55]	Local							✓	✓				
Dummies-based mechanisms													
Realistic fake trips [109]	Local			✓				✓	✓		✓		
Synthetic fake trips [110]	Local			✓									
Kido et al. [31]	Local		✓	✓									
You et al. [111]	Local		✓	✓									
MobiPriv [112]	TTP		✓		✓			✓		✓			
SpotME [29]	Local		✓					✓		✓			
Kato et al. [113]	Local		✓		✓	✓							
SybilQuery [114]	Local		✓			✓				✓	✓		
Perturbation-based mechanisms													
Geo-indistinguishability [23]	Local	✓						✓			✓		
Path cloaking [115]	TTP				✓		✓						
CAP [116]	Local		✓					✓		✓			
Temporal clustering [117]	TTP		✓				✓	✓		✓			
Location truncation [118]	Local							✓					
Bordenabe et al. [119]	Local	✓		✓			✓		✓	✓			
Oya et al. [120]	Local	✓		✓	✓		✓						
Predictive geo-indistinguishability [121]	Local	✓					✓		✓				
Elastic geo-indistinguishability [122]	Local	✓			✓		✓						
LocLok [123]	Local	✓											
PIVE [124]	Local	✓		✓		✓	✓						
Protocol-based mechanisms													
Louis, Lester and Pierre [26]	P2P									✓			
PrivStats [28]	P2P							✓		✓			✓
MobiCrowd [125]	P2P					✓							
C-Hide&Seek, C-Hide&Hash [27]	NTTP							✓		✓	✓		
SRide [126]	NTTP									✓	✓		✓
PIR [127]	NTTP									✓			
Trust No One [128]	NTTP												
Narayanan et al. [129]	P2P/TTP									✓			
Koi [44]	NTTP												✓
Zerosquare [130]	NTTP									✓			
Outsourced garbled circuit [131]	NTTP									✓	✓		✓
Rule-based mechanisms													
ipShield [132]	Local									✓		✓	
LP-Guardian [133]	Local						✓			✓		✓	

TABLE V: List of *offline* LPPMs studied in this survey, with metrics used by their authors to evaluate them.

	Privacy					Utility			Perf.	
Protection mechanism	Differential privacy	k-anonymity	Attack correctness	Data distortion	Ad-hoc metric	Data distortion	Query distortion	Ad-hoc metric	Execution time	Scalability
Generalization-based mechanisms										
Nergiz et al. [134]	✓		✓					✓	✓	
Never Walk Alone [135]		✓		✓			✓		✓	
Wait for Me [32]		✓		✓		✓	✓		✓	✓
Yarovoy et al. [136]		✓				✓	✓			
Differentially private grids [137]	✓						✓			
GLOVE [34]		✓				✓				
Gramaglia et al. [138]		✓				✓		✓		
Dummies-based mechanisms										
Realistic fake trips [109]										
Synthetic fake trips [110]			✓				✓	✓		
Hermes++ [30]		✓					✓	✓	✓	
Perturbation-based mechanisms										
Geo-indistinguishability [23]	✓						✓			
Path confusion [139]			✓		✓	✓				
PINQ [140]	✓									
Chen et al. [141]	✓						✓		✓	✓
Jiang et al. [25]	✓					✓				
DP-WHERE [33]	✓					✓				
Acs et al. [142]	✓					✓				
Riboni et al. [143]	✓	✓				✓	✓			
Promesse [95]				✓		✓	✓		✓	✓

using a pseudonym instead of their real identity (e.g., real name, IP or MAC address). In this context, a mix-zone is an area where movements of users are not tracked, and consequently where users cannot communicate with an LBS. When a user leaves a mix-zone, she receives a new pseudonym chosen among the pseudonyms of users inside the mix-zone. It means that when k users are inside a mix-zone at the same time, their identities will be shuffled, providing some of k-anonymity and resulting in an attacker's confusion.

1) Online mechanisms: The initial model was further refined by Beresford and Stajano [96], providing a more formal mathematical model as well as a location privacy metric, from the point of view of the attacker. A question that arises with mix-zones is where to place them. This has been tackled by Freudiger et al. [97], with the goal to maximise location privacy while taking into account the negative impact of mix-zones on utility. Practically, they use a new metric called mobility profiles to theoretically compute the effectiveness of a mix-zone, and then solve the mix-zones placement problem as an optimisation problem. Another solution to the mix-zones placement problem was proposed by Liu et al. [98]. They model the city as a graph, where nodes are venues (i.e., places of interest inside a city such as monuments, restaurants, cinemas, etc.) and the road network is used to create edges connecting those venues. On one hand, an LBS can have side information on this graph and use it to re-identify users. On the other hand, information about traffic is used to compute the optimal placement of mix-zones as an optimisation problem. *MobiMix* proposed by Palanisamy and Liu [99] is another solution leveraging the road network for optimising the mix-zones placement. They consider the speed of users as a side channel that could be used to re-identify them. They also propose different manners to construct mix-zones, designed to defeat timing attacks. Gong et al. [100] proposed a socially-aware way to exchange pseudonyms. They model the decision of changing a pseudonym as a game, which takes into account social ties between users.

2) Discussion: Overall, mix-zones require an important number of users to be effective. Indeed, if too few users participate to the system, it is not very likely that they will meet at any time during the day. We believe that this critical mass of users is too important to make mix-zones usable for individual users willing to protect their privacy in online use cases. Moreover, mix-zones need to rely on a third party to provide the pseudonyms, and a trusted third party to handle the swapping of pseudonyms. Introducing

another trusted party, in place of the LBS, does not appear to be a desirable property.

B. Generalization-based mechanisms

Generalization methods have been successfully applied to provide k-anonymity in the location privacy context, through the concept of spatial cloaking introduced by Gruteser et al. [146]. Broadly speaking, the idea is to report an coarser information instead of the exact location of users. Besides the effect of reducing the precision of the information, this also allows to create cloaking areas in which at least k users are at any given moment.

1) Online mechanisms: Gedik and Liu introduced *CliqueCloak* [101], which is a system generating cloaking areas on-the-fly, as messages arrive, before sending them to an LBS. Users specify a value for k, the maximum size of the cloaking area and the maximum time between the engine transmits the message to the LBS. Because the engine needs enough messages to enforce k-anonymity, messages can be delayed or cancelled if there are not enough queries coming from the same area. *Casper* [102] is a spatial cloaking architecture proposed by Mokbel et al. It uses a location anonymizer (i.e., a trusted third party) which knows locations of all users and their privacy parameters (a k parameter and a minimal cloaking area). When a user sends his query to the anonymizer, the latter transforms it into a cloaked query and forwards it to the query processor. The query processor needs to be able to understand such cloaked queries. The response is then sent to the anonymizer, which refines it by using the actual user location and sends the final response to the user. Chow et al. proposed *P2P cloaking* [103], which is essentially an improvement of Casper. Like in Casper, users specify a privacy profile with a k parameter and a minimal cloaking area size. However, instead of using a central trusted anonymizer, a peer-to-peer protocol enables nearby peers to generate cloaking areas. Clients can then send themselves the query including the cloaking area, instead of their exact location, to an LBS. *PRIVÉ* [24] is a decentralised architecture proposed by Ghinita et al. Instead of having a central server building cloaking areas, a hierarchical and distributed index is used. This index defines groups of at least k users in the same vicinity (i.e., cloaking areas). *PrivacyGrid* [104] is a solution introduced by Bamba et al. focusing on speed and effectiveness. It is an anonymisation proxy providing k-anonymity and ℓ-diversity thanks to a dynamic grid cloaking algorithm. It allows users to specify individually their requirements in terms of both privacy and utility. However, this LPPM can fail to deliver a query to an LBS; the probability of a query to be effectively protected can be increased by allowing to delay its sending. Xu and Cai [105] proposed a "feeling-based" approach to location privacy. The main idea is that users specify their privacy requirements by defining a public region that they would feel comfortable to be reported as their location. The goal is to replace traditional models in which users have to specify parameters, such as the k of k-anonymous LPPMs, which may not be very expressive to them. The authors then propose a solution, based on a trusted server, to build cloaking areas matching such a privacy requirement of each user. Li and Palanisamy introduced *ReverseCloak* [108], an LPPM providing k-anonymity for users moving inside a road network. This approach differs from previous work in that it provides a multi-level reversible privacy model. It means that different LBSs may have different access levels, and thus access more or less granular information.

Agir et. al [106] introduced an adaptive mechanism to dynamically change the size of obfuscated areas hiding the exact location of users. More precisely, the proposed solution locally evaluates the privacy level and enlarges the area until a target privacy level is achieved, or the information is too distorted (in which case the location cannot be released). Ngo and Kim [107] proposed a protection mechanism trying to optimise the average size of cloaked areas generated by Hilbert curve methods. They define a new privacy metric for cloaking areas, relying on ϵ-differential privacy. They explore a notion of identifiability, quantifying the probability of an attacker to identify the user's location from the cloaking area, that help them to choose a value for the ϵ parameter. Huguenin et al. [55] studied the effect of using generalization-based LPPMs while using LBSs such as Foursquare [4] to leave check-ins. Towards this end, they used a predictive model to quantify the effect of generalization on perceived utility. They were able to predict with a small error the loss of utility caused by the generalization protection, allowing to implement efficiently LPPMs featuring both good privacy and utility properties.

2) Offline mechanisms: Nergiz et al. [134] proposed an algorithm to segment trajectories into groups of points providing k-anonymity. Then, they introduced a randomised reconstruction algorithm that uses the cloaked areas obtained from the previous step to recreate trajectories, giving a protected version of the original trajectories. Abul et al. proposed *Never Walk Alone* [135], whose idea is to guarantee that at every instant there is at least k users walking at a given distance of the others, thus creating cylinders within which users move. This radius exploits the inherent incertitude that comes with GPS measurements to avoid distorting the data too much. This mechanism has been later improved by *Wait4Me* [32], which is more generic with respect to the input dataset it can protect, and scales better to large datasets.

When people move, they essentially move from one place to another, which are often POIs. The list of these places can been considered as a quasi-identifier, which can be protected with a k-anonymity guarantee. Yarovoy et al. [136] tackled the problem of creating optimal anonymisation groups for moving objects, which unlike traditional databases may not be disjoint. They consider an attacker model where the latter is building an attack graph giving

relationships between objects in the protected database and their identities in the raw database. Qardaji et al. [137] studied the usage of grids to partition the continuous space into a discrete domain. They started with static grids and proposed a method to choose the grid's size, and then proposed a new approach using an adaptive grid. Gramaglia and Fiore [34] proposed a measure of the anonymisability of mobility datasets, based on their spatio-temporal similarity. Then, they introduced *GLOVE*, an adaptive protection mechanism providing k-anonymity while reducing the utility loss. It iteratively merges mobility traces that are the most similar, with respect to the previously introduced metric, until all groups of traces contain at least k users. Gramaglia et al. [138] introduced $k^{\tau,\epsilon}$-anonymity, which is a model extending k-anonymity to include temporal information. In this setup, an attacker may have some background knowledge covering a continuous period of time of at most τ, and is allowed to discover the mobility records of a targeted user for a period of at most ϵ (disjoint from τ). The authors propose a way to reach $k^{\tau,\epsilon}$-anonymity by relying on a spatio-temporal generalization.

3) Discussion: Overall, generalization-based LPPMs have the advantage to propose an easy to understand privacy model (e.g., k-anonymity). This category of LPPMs is most adapted to offline scenarios with deterministic approaches. Indeed, providing different protected versions of the same dataset with an non-deterministic approach can reveal additional information at each release. Moreover, as outlined with *Wait for Me*, these approaches faces to a scalability issue when the size of the dataset increases.

In online use cases, this protection scheme suffers from the same weakness than mix-zones, specifically the requirement to have enough users to be effective. In addition, generalization-based LPPMs usually do not work with GPS coordinates but with areas or trajectory which is not immediately usable by existing LSBs.

C. Dummies-based mechanisms

Instead of relying on other users to be hidden among them and obtain k-anonymity, as with generalization-based approaches, it is possible to generate fake users, called dummies. In this scheme, the attacker may be aware that there are dummies inside the data it got, but the challenge here is to generate realistic fake data, ideally indistinguishable from the real data.

1) Online mechanisms: Dummies-based protection mechanisms are mostly used for online usage in the literature. The basic idea is for each user to send multiple queries to an LBS, instead of a single one. One of those queries contain his actual location, while the others contain fake locations. Consequently, the LBS is not able to determine exactly where the user really is located. Kido et al. [31] were the first to introduce a protection mechanism using dummies. They simply split the space into regions of a fixed size and generate dummies in neighbouring regions. You et al. [111] proposed another method to create fake trajectories. They generate endpoints randomly and then generate trajectories between these new endpoints with two methods. A first method generates a trajectory randomly by using vertical, horizontal and diagonal random speeds. A second method intends to force intersections between trajectories of dummies and the real user's trajectory. In that case, a dummy trajectory is obtained by rotating the real trajectory around a given point. Stenneth et al.[112] presented *MobiPriv*, which uses an anonymisation proxy through which all queries transit before being sent to an LBS. Similarly to centralized protection mechanisms presented in Section V-B, the proxy of MobiPriv ensures k-anonymity by generating realistic looking dummies. It also leverages a history of previous queries to prevent attacks using the intersection of multiple queries' results to infer new knowledge. *SpotME*, proposed by Quercia et al. [29], also generates dummies within discrete regions. However, this solution is specific to one use case: counting users inside regions. Users report to be or not within a region according to a randomized mechanism. The LBS can then "reverse" this mechanism and compute how many users are really within a region with a high probability. If users are honest, there is at most an error of 11 % in the final result. Lato et al. [113] presented a method to generate dummies that acknowledge the fact that users make stops during their mobility, while previous work often consider generating fake trajectories between two endpoints. They assume user's movements are known in advance and use this knowledge to reduce user's traceability (i.e., increasing the confusion of an LBS about which of the dummies are the real users). More specifically, each time the user stops, there is a possibility that a dummy can stop at the same place, thus creating a crossing between multiple paths and increasing the confusion.

Shankar et al. introduced *SybilQuery* [114], a solution to generate real-looking fake trips, especially suited for navigation applications. Knowing the real trajectory that a user will do, SybilQuery will generate fake trips starting from and ending to different locations, but preserving properties such as the length of the trip and the semantics of the areas where endpoints are located (e.g., residential vs business areas). When a user moves and sends queries to an LBS to get directions, fake users will also move along fake trips. Krumm [109] proposed a probabilistic model to generate fake driving trips. Endpoints are chosen according to some probability model. A route planner is used to generate a trajectory between trips, with some randomness injected into trajectory to prevent the optimal path to be always selected (indeed, users do not always follow the best path when driving). Speeds are also drawn from a probability model, and some noise is finally added to each point to simulate GPS noise. Bindschaedler et Shokri [110] presented a way to generate synthetic mobility traces that share statistical features with real traces in a privacy-preserving way. These synthetic traces are designed to be used instead of the real traces, thus presumably leaking

no sensitive information. They build a mobility model for each trace and an aggregate probabilistic mobility model about the entire dataset, and use it to synthetise fake traces from these models; which must satisfy a privacy test before being released.

2) Offline mechanisms: In an offline context, dummies can also be used to provide artificial k-anonymity when data would otherwise be discarded because there is no other user with a similar behaviour, although it is less used because it introduces an obvious error for analysts. Pelekis et al. proposed *Hermes++* [30], which is a privacy-preserving query engine. It relies on the injection of dummies in query results, these dummies being designed to follow the behaviour of actual users. This engine also has an auditing module that is able to detect if a sequence of queries can be harmful for the privacy of individuals (i.e., if trying to track users over time).

3) Discussion: The main issue with dummies-based LPPMs is their ability to produce real-looking dummies. Indeed, a study of Peddinti et al. [147] showed that *SybilQuery* is very vulnerable to attacks based on machine learning. They developed an algorithm able to correlate traces, and tested it against a dataset containing data of 85 taxi drivers around San Francisco. SybilQuery was configured with $k = 5$, which means that each mobility event generated by a driver was hidden among four other dummy events. In the case of an attacker having access to a previous mobility dataset (i.e., forming the training dataset), their algorithm re-identified 93 % of the users. Furthermore, some of these algorithms (e.g., [114], [109]) use an extensive amount of external knowledge, such as a graph modeling the road network, a route planner or census statistics about the population. However, using and processing important external knowledges in online use cases represents a limiting factor for a usage on mobile device with limited computational and storage capacities.

D. Perturbation-based mechanisms

While protection mechanisms providing k-anonymity try to hide a user inside a crowd, other mechanisms rely on the alteration of the data to be sent to an LBS to protect it. In this case the challenge is to have a trade-off between privacy (i.e., the data needs to be distorted enough to be protected) and the utility (i.e., if the data is too distorted, results from the LBS will be unusable). Most mechanisms work by adding some (often random) noise to the underlying raw data.

1) Online mechanisms: Hoh et al. [115] proposed a mechanism playing with the confusion of an attacker. They designed a privacy metric called *time-to-confusion*, quantifying the duration for which a given user can be tracked by an attacker. They developed an LPPM using an third party server which aims to maximise this metric, in the context of traffic monitoring. Pingley et al. presented *CAP* [116]. This solution protects two channels by which a curious LBS could obtain information: the location contained inside the query and the user's IP address. For the latter an improved routing algorithm in the Tor anonymising network [148] is used, while for the former Hilbert curves are leveraged to generate fake locations close to the real one. Assam and Seidl [117] proposed a mechanism enforcing k-anonymity through temporal clustering of streams of mobility data. Micinski et al. [118] studied the effect of location truncation, i.e., reducing the precision of latitude/longitude coordinates by dropping decimals, on a nearby venues finder application on Android. This is a simple protection mechanism that can still be effective to protect someone's exact location.

Differential privacy has been generalised for location privacy by Andres et al. under the notion of *geo-indistinguishability* [23]. Geo-indistinguishability is a formal notion of location privacy that bounds the probability of two points to be reported locations of the same real location within a given radius. Thus, a user can quantify the level of privacy she wants within a specific area. Practically, it is done through the ϵ parameter (the lower ϵ, the higher the noise), resulting in ϵ-geo-indistinguishability. Authors proposed a way to provide geo-indistinguishability by adding noise drawn from a planar Laplace distribution to a real location. Bordenabe et al. [119] proposed a method to construct an LPPM enforcing geo-indistinguishability that maximises the utility. To achieve that, they relied on linear programming techniques, and proposed a way to reduce the number of constraints, this improving the time required to build the LPPM. Oya et al. [120] also studied the design of optimal LPPMs, including geo-indistinguishability among others. They argue that, besides metrics based on attack correctness, auxiliary metrics should be taken into account when evaluating an LPPM. Consequently, they introduced two such metrics, showed that one single metric is not sufficient to assess the efficiency of an LPPM, and used these additional metrics to design a new LPPM. Due to temporal correlations between a user's locations, differential privacy proposed in geo-indistinguishability can be problematic because each time a location is protected, some more privacy is lost. In other words, protecting a trace of n locations with ϵ-geo-indistinguishability results at the end in $n\epsilon$-geo-indistinguishability. To overcome this limitation, Chatzikokolakis et al. proposed a predictive mechanism [121] providing geo-indistinguishably, using prediction to avoid spending too much budget for each location protection. With two different ways of spending the privacy budget this gives a substantial improvement over the original geo-indistinguishable mechanism. The same authors also proposed another extension of geo-indistinguishability [122], which leverages contextual information to calibrate the amount of noise applied to disturb the mobility traces. They consider two levels of sensitivity if the user is located in an urban environment, where there is a high density of venues around, or in a sparse countryside area.

To account the temporal correlation of mobility traces, Xiao and Xiong [149] also proposed a mechanism based on differential privacy and a new measure to evaluate the sensitivity of each protected location. The same authors later proposed *LocLok* [123], an LPPM taking into account temporal correlations via a hidden Markov model. In a nutshell, their mechanism maintains a hidden Markov model of possible actual locations each time a protected location is released, and then generates protected locations via a differentially private method taking into account this Markov model. Yu et al. introduced *PIVE* [124], whose goal is to enforce at the same time two privacy guarantees: geo-indistinguishability and protection against adversary attacks. Their solution works in two steps: first a protection location set is generated from the actual location and a minimum attacker error specified by the user. Then a protected location, having differential privacy guarantees, is generated from the protection location set.

2) Offline mechanisms: Hoh and Gruteser [139] introduced the idea of path confusion as an LPPM. The idea is to make closer users' paths to cross when they are close enough, to augment the confusion of an adversary about which path belong to which user. They formulate and solve this problem as a constrained non-linear optimisation problem. We proposed *Promesse* [95], which is specifically designed to hide the POIs of users. It works on mobility traces by smoothing speed, i.e., making the speed appear as being constant. Therefore, users seem to be always moving, thus making it more difficult to guess were they stopped and what their POIs are.

PINQ (for Privacy INtegrated Queries) [140] is an analytics platform allowing to execute queries against a data source while preserving privacy through differential privacy. The data analyst writes his queries, specifies privacy budget ϵ that can be consumed, and the platform automatically takes care of returning results satisfying differentially private guarantees. One of the proposed examples illustrates geo-located queries and shows that PINQ can be successfully applied in this context. Chen et al. [141] protected public transportation usage data, which can be seen for each user as a sequence of places (metro/bus stations) she went to. They built a method to protect such data in a differentially private way and evaluated their mechanism by studying the impact of their protection on range queries and sequential pattern mining. Differential privacy has been used by Jiang et al. [25] to protect ships' trajectories. Endpoints of trajectories are preserved while intermediate locations are altered by adding some noise satisfying differential privacy guarantees. *DP-WHERE* [33] is a method introduced by Mir et al. to generate synthetic Call Detail Records (CDRs) in a differentially private way. They start by building a model of real CDRs, formed of several histograms, and then add noise to each of them to achieve differential privacy. A synthetic CDR can be generated by using the private versions of the histograms. Authors of geo-indistinguishability [23] also presented an offline usage of their protection mechanism. Acs et al. [142] proposed a mechanism to protect spatio-temporal densities datasets, which reports counts of active users within small areas for given time windows. Such data can be obtained for instance from call data records that are gathered by mobile phone operators. The authors proposed an approach that adapts to the original data in order to guarantee differential privacy with the highest possible utility. Counting users is one interesting thing to do with mobility data, but we want to publish entire trajectories and allow more mining tasks to be performed. Riboni and Bettini [143] introduced a way to publish check-in data (e.g., from Swarm [150]) in a differentially private manner, with the goal to allow venue recommandation from this data. They start by filtering check-ins that fall within regions of fixed size where a single user did too many check-ins, thus indicating that it may be an important or sensitive venue for him. Then, some noise is added to the number of check-ins at each venue to enforce differential privacy, before to release these statistical values.

3) Discussion: Online perturbation-based LPPMs have the advantage to be working on a local architecture, i.e., they do not need an external trusted party, as opposed to generalization-based LPPMs which very often require one. A large number of them also rely on differential privacy which can be performed locally. However, an open question with this privacy model in an interactive mode (i.e., online use cases) represents the management of the privacy budget ϵ (e.g., [151]). In addition, as the meaning of ϵ is less intuitive than the meaning of k (in k-anonymity), choosing the proper value of this parameter may also be less clear from a user point of view. Lastly, some perturbation-based LPPMs do not rely on any formal privacy guarantee, which makes their guarantees harder to justify in practice.

E. Protocol-based mechanisms

Protection mechanisms falling in previous categories rely on the alteration of mobility data in order to protect information. The solutions adopted in this category is to propose protocols which preserve privacy by design. These protocol-based mechanisms are generally more specific (e.g., getting nearby friends, counting people) but can achieve the best privacy guarantees. They largely rely on encryption schemes offering strong privacy guaranties for specific use cases.

1) Online mechanisms: Louis, Lester and Pierre are three protocols proposed by Zhong et al. [26] that can be used to locate nearby friends in a privacy-preserving way. They are multi-parties protocols: a user, say Alice, initiates a communication with another user, say Bob, and tries to learn if she is within a given radius r from herself. Depending on the protocol, Alice learns nothing, Bob's exact location, Bob's exact distance or Bob's grid cell distance. On the other side Bob learns nothing, Alice's exact location and/or the radius r. Popa

et al. [28] introduced *PrivStats*, a system that can be used to collect location-based aggregate statistics within defined geographic areas. Users collaborate to send pre-aggregated and encrypted data to the LBS, which allows to hide the number of tuples and the time at which they were collected. The LBS receives a constant number of (encrypted) values at fixed time intervals, combines them by using homomorphic encryption and asks a user to decrypt the final aggregate value. Authors also propose a privacy-preserving accountability protocol without any trusted party to prevent clients from cheating. *MobiCrowd* [125] is a protection mechanism designed by Shokri et al. relying on collaboration between users to avoid querying an LBS if the information is already available on another nearby user's device. Each requested piece of information is stored inside a local buffer on users' devices and are given an expiration time. Users issuing queries first broadcast it to neighbours in an attempt to get the answer without contacting the LBS. Mascetti et al. [27] proposed a protocol for proximity notifications of nearby friends in a space divided into cells. Two protocols are introduced in the article. With *C-Hide&Seek*, each user sends to an LBS her current encrypted cell identifier, the encryption key being shared between each pair of friends. This way, users can learn in which cell their friends are, even if they are not nearby. With *C-Hide&Hash*, each user sends a hash of her salted location to the LBS, the salt being the key shared between each pair of friends. By computing the hashes of cells around and running a private set intersection protocol with the LBS, users can learn which friends are nearby, but not their distance. Narayanan et al. [129] introduced another protocol for testing the proximity of friends. They transformed the problem from proximity testing to equality testing, and then presented two protocols for private equality testing, one using a peer-to-peer architecture and the other relying on a trusted server. They finally introduced another solution relying on nearby location tags (e.g., WiFi broadcast packets) detected by users, who may then run a private set intersection protocol to infer their proximity. *SRide* [126] is a privacy-preserving ridesharing system proposed by Aïvodji et al. The goal of such a system is to prevent the ridesharing platform to learn sensitive information about the origin and destination of the users' trips. They leverage tools such as homomorphic encryption and secure multiparty computation; the ridesharing platform is still needed but do not have to be trusted anymore. Overall, the SRide protocol can be executed in 5 to 9 seconds, with a communication overhead comprised between 3 and 6 MB.

Private information retrieval (PIR), first theorised by Chor et al. [152], is a schema allowing someone to retrieve a row from a database without letting it know what she wants to retrieve. Ghinita et al. [127] proposed to apply PIR to N-nearest neighbours spatial queries, that can be used for example to look for nearby venues (i.e., restaurants, monuments, etc.). They introduced a way to index spatial information in a PIR-compliant way by using Hilbert space-filling curves. *Trust No One* [128], proposed by Jaiswal and Nandi, uses *pseudonymised locations* to represent locations by an identifier without revealing actual coordinates and *pseudonymised identifiers* to represent entities by an identifier. These two pseudonyms are generated by two different entities, respectively a mobile operator and an LBS. Finally, a decentralized matching service, that does not know anything about location or identity of entities, has the responsibility to answer queries. *Koi* [44] is a platform proposed by Guha et al. It relies on two non-colluding servers, namely the *matcher* and the *combiner*. The matcher knows about entities and locations but nothing about links between them (i.e., which location belongs to which entity). The combiner knows the mapping between entities and locations but nothing about actual content of these entities and locations. A communication protocol between the matcher and the combiner allows to answer queries by performing a privacy-preserving matching. Instead of directly querying Koi, mobile devices set up triggers reacting to some events (e.g., getting notified when there is a restaurant at less than 500 meters around me). Application developers must hence create event-centric applications instead of location-centric applications. Pidcock and Hengartner proposed *Zerosquare* [130], which relies on two non-colluding servers, one of them storing a user-indexed database and the other a location-indexed database. Moreover, some cloud components owned by service providers can be allowed individually by each user to access data contained in the location-indexed database. This is the mobile device itself that queries the two databases and join information coming from each of them. Garbled circuits where theorised by Yao [153] and allow two parts to privately evaluate the result of a generic function. Carter et al. [131] proposed a way to outsource the evaluation of such garbled circuits. Since they require a high computational power, outsourcing their evaluation in the cloud allows to speed up the processing and let a mobile device use garbled circuit despite a low computational capacity. The challenge is to preserve privacy guarantees even with an untrusted cloud. As an example they implemented a privacy-preserving navigation application that mainly consists in a Dijkstra shortest-path algorithm used to privately get directions between two (private) points while taking into account (private) hazards that can occur along the path.

2) Discussion: Overall, protocol-based LPPMs exhibit the best privacy and utility trade-off. This is because they are tailored to answer more specific use cases, instead of trying to solve a large range of use cases. But it also implies that such approaches may be too specific, e.g., [129], [27] are designed only to detect nearby friends. They also do not interact with existing LBSs, because they essentially replace them. While this gives to the LPPM designer much more freedom, it may seriously slow down their adoption, as new infrastructures need to be deployed to support

them. Moreover, because of the cost of the cryptographic primitives, such approaches may still be practically unusable because of their algorithmic complexity which impacts the execution time. For example, finding navigation directions using outsourced garbled circuits [131] takes about 15 minutes, and for a road network graph composed of only 100 vertices.

F. Rule-based mechanisms

Some believe that one-size-fits-all protection mechanisms are unrealistic. This is why some protection mechanisms implement various state-of-the-art solutions and follow a set of rules to decide of the most appropriate countermeasure to take in the current situation.

1) Online mechanisms: Chakraborty et al. proposed *ipShield* [132], which is a framework, implemented on Android, leveraging a rules engine to protect location privacy. Users define which threats they want to be protected against, with a priority level. The system then leverages a database of inference attacks to recommend protection rules to apply on each sensor (i.e., not only the GPS but also the accelerometer, the gyroscope, etc.). Users can also define their own rules, using contextual information and specifying actions to take on sensor data. *LP-Guardian* [133] is a software running on Android proposed by Fawaz and Shin to protect location privacy of Android smartphones users. They designed a framework to protect privacy against different threats: tracking threat, identification threat and profiling threat. It uses a decision tree to decide which action to perform in a given situation, leveraging the context (e.g., application being used, location) and a combination of statically-defined and user-defined rules.

2) Discussion: As rule-based mechanisms essentially rely on a composition of LPPMs from other categories, they inherit the associated pros and cons. Rule-based mechanisms can cover a wider range of use cases, however the side effects of these compositions can also jeopardise the privacy protections.

VI. Open challenges

Although the literature in the location privacy field is quite large, there are still several open challenges. We present in this section some of them based on the lessons we have learned from our experience. The first two challenges are more research-oriented, while the three last challenges are more technical

A. Quantifying location privacy

Evaluating the efficiency of a protection mechanism is not an easy task. However, as it appears by looking at Table IV and Table V, that there is a high heterogeneity when it comes to the metrics used to evaluate LPPMs, although it is possible to categorise them in a small number of categories. This makes it very difficult to fairly evaluate and compare LPPMs. We believe that an important research direction is to be able to have a common framework to evaluate LPPMs, by using a set of well-defined and accepted metrics, across all three dimensions of privacy, utility and performance.

Few works have been done in this direction. Shokri et al. [36] were the first to propose a formal framework to evaluate the efficiency of a protection mechanism by using the impact of a location privacy attack. They formally define three dimensions when evaluating privacy: the accuracy, the certainty and the correctness. They introduced different attacks and used their framework to evaluate a few protection mechanisms. It is worth noting that their framework is available as an open-source C++ tool [154]. However, this work is only applicable to a subset of LPPMs, which fit into a probabilistic framework, and is only interested in evaluating privacy. Later the ALP framework [81] (available as an open source tool [155]) proposed a more generic support to configure LPPMs from a set of privacy and utility objectives. However, the number of available metrics and the definition of the objectives are still somewhat limited, and the convergence to appropriate configuration parameters is not ensured.

Those works both introduce a model and a number of metrics to evaluate privacy and utility of LPPMs. Indeed, despite formal guarantees are needed in most contexts, they do not always translate well how an LPPM behaves in practice, as it has been shown in several works (e.g., [156], [157], [158]). On the privacy side, we advocate (similarly to, e.g., [124]) that metrics relying on adversary attacks should be considered as complimentary to the formal guarantees. The emergence of data anonymization and de-anonymization challenges [159] is promising to propose new protection schemes and assess existing ones. On the utility side, besides the classical information theoretic metrics such as the entropy, there is a need to consider more application-driven use cases.

B. Towards new protection mechanisms

Differential privacy has met a large interest and generated an important literature since its introduction in 2006 [71], [77]. Several recent works on location privacy use and apply this approach for the protection of geolocated information (e.g., [143], [121], [141]). We believe that this privacy model is still promising and will continue to generate an important literature in the next few years. As noticed in Section V-D, how to manage the ϵ privacy budget in an interactive mode still remains an important question. But besides its wide adoption and interest in the research community, other guarantees are still worth being used such as l-diversity and t-closeness. Chatzikokolakis et al. [160] explore more in depth the deterministic (e.g., k-anonymity) and non-deterministic (e.g., differential privacy) methods that can be used to design modern LPPMs.

Another promising track of research concerns the composition of LPPMs (e.g., combining k-anonymity with l-diversity where $l < k$). We previously classified such approaches as rule-based LPPMs (Section V-F). These ap-

proaches rely on different LPPMs and use one or another depending on the actual situation. Because there is no one-size-fits-all LPPM, combining existing LPPMs allows to cover a larger variety of use cases. However, quantifying the guarantee offered by the composition of heterogeneous LPPMs is challenging.

Yet another way to tackle the lack of a one-size-fits-all LPPM, instead of composing existing LPPMs, is to dynamically alter the level of protection offered by single LPPM. Because not every data is equally sensitive and needs to be similarly protected, such approaches bring adaptivity. Dynamically adapting the offered privacy level avoids over protecting data, and consequently provides a better utility. This has been explored by some previous works (e.g, [106], [122], [55], [81]). However, few of these works take into account the semantics of visited places. For example, it may be more important to protect that a user went to a hospital than to protect that he is shopping inside a mall. Providing adaptivity with respect to the semantics of some place (Section II-B for more on the semantics aspects) could be a smart way to provide users with a tailored LPPM.

Recently, we have seen the development of privacy-by-design approaches [44], [130]. Privacy-by-design has been theorised by the information and privacy commissioner of Ontario, Canada [161]. In a nutshell, it relies on seven core principles: proactivity, privacy as the default setting, privacy embedded in the design, full functionality, end-to-end security, visibility/transparency and user-centricity. In other words, it advocates for systems where privacy is integrated since the beginning as a requirement and by default, where the interests of the user come first, and without sacrificing the quality of service. Despite seeming utopian, this goal is actually reachable as soon as we throw away the LBS stack as we know it today. With privacy-by-design architectures, there is no need anymore to alter mobility data, as the LBS itself integrates privacy as a first class citizen.

C. Datasets

As shown in Section III-F, the research community has at its disposal a few real-life mobility datasets to evaluate its work. However, despite, several initiatives that have been conducted to publicly provide datasets coming from real-life data collections, all these datasets remain small and involve a limited number of users. This lack of large datasets strongly limit the ability of researchers to test their solutions under real condition. Providing golden standards in terms of large mobility data collections is definitely appealing and would be very useful to better compare LPPMs. There is a real need to share methodologies and tools around those collections, and make them available to the research community. Some efforts are already going into that direction, such as the Funf [11] and APISENSE [10] platforms, or the Crawdad [162] community. Lastly, open data initiatives followed by many organisations and cities (e.g., the city of Montreal provides trajectory data [163]) are also promising to provide open and large datasets.

D. Users awareness

Most of the users are not aware about the risk related to the exploitation of their mobility data. There is a lack of tools to improve users' awareness on this. To give an example, *Please Rob Me* [48] is a website whose goal is to "raise awareness about over-sharing", by showing it is possible to infer from geo-located tweets whether users are at home. Moreover, people are not aware of the value of their mobility data, certainly because they do not know the amount of knowledge that can be derived from it. A study showed that people would share their mobility trace in exchange of a little amount of money (the median was £10 or £20 for a commercial usage in [164]) or a gift (1 % of chances to win a US$200 MP3 player in [37]). We advocate it is one of the mission of researchers to raise awareness on societal problems such as privacy. Besides talks targeted towards the general public, tools could be developed to highlight privacy issues and the benefits of using an LPPM. *FindYou* [54] is an example of such a tool that allowed users to visualise what could be inferred from the data collected by online LBSs such as Foursquare, Instagram or Twitter. To go a step further, it would be very interesting to additionally show the impact of an LPPM on the collected data, and the benefits it brings to users' privacy.

E. Implementation effort

To be used, protection mechanisms obviously need to be implemented and made available. Very few solutions are freely downloadable and usable without reimplementing them from scratch. A notable work is *ipShield* [132], which is actually implemented on the Android platform (though not necessarily installable trivially by end-users, because it is tightly integrated in the Android kernel). Geo-indistinguishability [23] has also been implemented by its authors as a browser extension [165] working with several popular browsers. This extension easily allows users to benefit from some privacy when using geolocated services through their Web browser. Another example is Aircloak [166], a project that aims to propose a trusted sensitive data collection architecture with privacy-preserving querying capabilities. By using several layers of noise, as well as maintaining a history of previous queries, the application is able to detect combinations of queries that could result in a privacy leak and prevent them. With ACCIO [167], we proposed an experimental platform to experiment with location privacy. This platform implements several state-of-the-art mobility data manipulation routines, privacy and utility metrics, and LPPMs such as geo-indistinguishability [23] or Wait4Me [32].

Lastly, we have also seen large companies taking steps to actually implement privacy-preserving measures in their

products. Apple for example, uses differential privacy for some machine learning applications [168], such as the keyboard suggestions. Google also successfully integrated RAPPOR [169] into its Chrome browser, allowing to get usage statistics in a privacy preserving way. The latter also relies on differential privacy, although it requires a large number of users to behave properly (the experiments were conducted with 1 million users).

VII. Conclusion

In this article, we surveyed the latest works about computational location privacy. At the best of our knowledge, it is the first survey to propose a unified view on both online and offline protection mechanisms, and putting the evaluation metrics as first-class citizens. This shows that online and offline protection mechanisms can be based on the same underlying primitives (e.g., differential privacy), while providing appropriate algorithms suited for the considered use case. While literature is already rich in various protection mechanisms, we outlined the lack of standard methods to compare these mechanisms.

According to the title of this survey, there is still a long road until location privacy can be democratised, both politically (i.e., being accepted in the users' mind) and technically (i.e., having production-quality software). But some recent theoretical and practical works are encouraging and show the way to what future research in location privacy could be.

References

[1] "Google maps website," Online at: https://maps.google.com.

[2] Waze Mobile Ltd., "Waze website," Online at: https://www.waze.com.

[3] Yahoo!, Inc., "Yahoo! Weather website," Online at: https://mobile.yahoo.com/weather/.

[4] Foursquare Labs, Inc., "Foursquare website," Online at: https://www.foursquare.com.

[5] Yelp, Inc., "Yelp website," Online at: https://www.yelp.com.

[6] Niantic, Inc., "Pokemon GO website," http://pokemongo.nianticlabs.com.

[7] City Domination GmbH & Co. KG, "City Domination website," Online at: http://www.citydomination.games.

[8] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *SenSys*, 2008, pp. 323–336.

[9] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "Peir, the personal environmental impact report, as a platform for participatory sensing systems research," in *MobiSys*, 2009, pp. 55–68.

[10] N. Haderer, R. Rouvoy, and L. Seinturier, "Dynamic Deployment of Sensing Experiments in the Wild Using Smartphones," in *DAIS*, vol. LNCS-7891, 2013, pp. 43–56.

[11] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive Mobile Computing*, vol. 7, no. 6, pp. 643–659, Dec. 2011.

[12] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Show Me How You Move and I Will Tell You Who You Are," *Transactions on Data Privacy*, vol. 4, no. 2, pp. 103–126, Aug. 2011.

[13] K. Sharad and G. Danezis, "An automated social graph deanonymization technique," in *WPES*, 2014, pp. 47–58.

[14] A. Sadilek and J. Krumm, "Far out: Predicting long-term human mobility," in *AAAI*, 2012, pp. 814–820.

[15] J. Krumm and D. Rouhana, "Placer: Semantic Place Labels from Diary Data," in *UbiComp*, 2013, pp. 163–172.

[16] L. Franceschi-Bicchierai, "Redditor cracks anonymous data trove to pinpoint muslim cab drivers," Online at: http://mashable.com/2015/01/28/redditor-muslim-cab-drivers/, Jan. 2015.

[17] H. Henttu, J.-M. Izaret, and D. Potere, "Geospatial Services: A $1.6 Trillion Growth Engine for the U.S. Economy," Online at: http://www.bcg.com/documents/file109372.pdf, 2012.

[18] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: An Information-flow Tracking System for Realtime Privacy Monitoring on Smartphones," in *OSDI*, 2010, pp. 1–6.

[19] J. P. Achara, F. Baudot, C. Castelluccia, G. Delcroix, and V. Roca, "Mobilitics: Analyzing Privacy Leaks in Smartphones," *ERCIM News*, vol. 2013, no. 93, 2013.

[20] M. Eskandari, B. Kessler, M. Ahmad, A. S. de Oliveira, and B. Crispo, "Analyzing remote server locations for personal data transfers in mobile apps," *PETS*, vol. 2017, no. 1, pp. 118–131, 2017.

[21] J. Zang, K. Dummit, J. Graves, P. Lisker, and L. Sweeney, "Who knows what about me? a survey of behind the scenes personal data sharing to third parties by mobile apps," in *Technology Science*, 2015.

[22] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal, "Your location has been shared 5,398 times!: A field study on mobile app privacy nudging," in *CHI*, 2015, pp. 787–796.

[23] M. E. Andrès, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential Privacy for Location-based Systems," in *CCS*, 2013, pp. 901–914.

[24] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems," in *WWW*, 2007, pp. 371–380.

[25] K. Jiang, D. Shao, S. Bressan, T. Kister, and K.-L. Tan, "Publishing Trajectories with Differential Privacy Guarantees," in *SSDBM*, 2013, pp. 12:1–12:12.

[26] G. Zhong, I. Goldberg, and U. Hengartner, "Louis, Lester and Pierre: Three Protocols for Location Privacy," in *PETS*, 2007, pp. 62–76.

[27] S. Mascetti, D. Freni, C. Bettini, X. Wang, and S. Jajodia, "Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies," *The VLDB Journal*, vol. 20, no. 4, pp. 541–566, 2011.

[28] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li, "Privacy and accountability for location-based aggregate statistics," in *CCS*, 2011, pp. 653–666.

[29] D. Quercia, I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft, "SpotME If You Can: Randomized Responses for Location Obfuscation on Mobile Phones," in *ICDCS*, 2011, pp. 363–372.

[30] N. Pelekis, A. Gkoulalas-Divanis, M. Vodas, D. Kopanaki, and Y. Theodoridis, "Privacy-aware Querying over Sensitive Trajectory Data," in *CIKM*, 2011, pp. 895–904.

[31] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of Location Privacy Using Dummies for Location-based Services," in *ICDE Workshops*, 2005, p. 1248.

[32] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Information Systems*, vol. 35, no. 8, pp. 884–910, 2010.

[33] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, "DP-WHERE: Differentially private modeling of human mobility," in *BigData*, Oct 2013, pp. 580–588.

[34] M. Gramaglia and M. Fiore, "Hiding Mobile Traffic Fingerprints with GLOVE," in *CoNEXT*, 2015, pp. 26:1–26:13.

[35] H. Mousa, S. B. Mokhtar, O. Hasan, O. Younes, M. Hadhoud, and L. Brunie, "Trust management and reputation systems in mobile participatory sensing applications," *Computer Networks*, vol. 90, no. C, pp. 49–73, Oct. 2015.

[36] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *S&P*, 2011, pp. 247–262.

[37] J. Krumm, "A Survey of Computational Location Privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, Aug. 2009.

[38] K. G. Shin, X. Ju, Z. Chen, and X. Hu, "Privacy protection for users of location-based services," *IEEE Wireless Communications*, vol. 19, no. 1, pp. 30–39, Feb. 2012.

[39] M. Terrovitis, "Privacy preservation in the dissemination of location data," *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 1, pp. 6–18, 2011.

[40] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A Classification of Location Privacy Attacks and Approaches," *Personal Ubiquitous Computing*, vol. 18, no. 1, pp. 163–175, Jan. 2014.

[41] C.-Y. Chow and M. F. Mokbel, "Trajectory privacy in location-based services and data publication," *SIGKDD Explorations Newsletter*, vol. 13, no. 1, pp. 19–29, Aug. 2011.

[42] M. Grissa, B. Hamdaoui, and A. A. Yavuza, "Location privacy in cognitive radio networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1726–1760, 2017.

[43] C. D. Cottrill, "Location privacy: Who protects?" *URISA Journal-Urban and Regional Information Systems Association*, vol. 23, no. 2, pp. 49–59, 2011.

[44] S. Guha, M. Jain, and V. N. Padmanabhan, "Koi: A Location-Privacy Platform for Smartphone Apps," in *NSDI*, 2012, pp. 183–196.

[45] L. Sweeney, "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[46] M. Barbaro and T. Zeller Jr, "A face is exposed for aol searcher no. 4417749," Online at: http://www.nytimes.com/2006/08/09/technology/09aol.html, Aug. 2006.

[47] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *S&P*, 2008, pp. 111–125.

[48] "Please Rob Me website," Online at: http://pleaserobme.com.

[49] O. Goldreich, "Cryptography and cryptographic protocols," *Distributed Computing*, vol. 16, no. 2-3, pp. 177–199, Sep. 2003.

[50] J. R. Douceur, *The Sybil Attack*, 2002, pp. 251–260.

[51] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering Personal Gazetteers: An Interactive Clustering Approach," in *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, 2004, pp. 266–273.

[52] R. Hariharan and K. Toyama, "Project Lachesis: parsing and modeling location histories," in *Geographic Information Science*, 2004, pp. 106–124.

[53] "Google place api documentation," Online at: https://developers.google.com/places.

[54] C. Riederer, D. Echickson, S. Huang, and A. Chaintreau, "Findyou: A personal location privacy auditing tool," in *WWW*, 2016, pp. 243–246.

[55] K. Huguenin, I. Bilogrevic, J. S. Machado, S. Mihaila, R. Shokri, I. Dacosta, and J.-P. Hubaux, "A predictive model for user motivation and utility implications of privacy protection mechanisms in location check-ins," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 760–774, April 2017.

[56] I. Bilogrevic, K. Huguenin, M. Jadliwala, F. Lopez, J.-P. Hubaux, P. Ginzboorg, and V. Niemi, "Inferring Social Ties in Academic Networks Using Short-Range Wireless Communications," in *WPES*, 2013, pp. 179–188.

[57] C. Wang, C. Wang, Y. Chen, L. Xie, and S. Lu, "Smartphone privacy leakage of social relationships and demographics from surrounding access points," in *ICDCS*, June 2017, pp. 678–688.

[58] J. Krumm, "Inference Attacks on Location Tracks," in *PerCom*, 2007, pp. 127–143.

[59] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "De-anonymization attack on geolocated data," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, 2014.

[60] A. Tockar, "Riding with the stars: Passenger privacy in the nyc taxicab dataset," Online at: https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/, September 2014.

[61] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Knock Knock, WhoâĂŹs There? Membership Inference on Aggregate Location Data," in *NDSS*, 2018.

[62] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, no. 1376, 2013.

[63] P. Golle and K. Partridge, "On the Anonymity of Home/Work Location Pairs," in *PerCom*, 2009, pp. 390–397.

[64] H. Zang and J. Bolot, "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study," in *MobiCom*, 2011, pp. 145–156.

[65] A. Boutet, S. Ben Mokhtar, and V. Primault, "Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets," LIRIS UMR CNRS 5205, Research Report, Oct. 2016.

[66] D. Manousakas, C. Mascolo, A. R. Beresford, D. Chan, and N. Sharma, "Quantifying Privacy Loss of Human Mobility Graph Topology," *PETS*, vol. 2018, no. 3, pp. 5 – 21, 2018.

[67] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *NDSS*, 2018.

[68] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining User Mobility Features for Next Place Prediction in Location-Based Services," in *ICDM*, 2012, pp. 1038–1043.

[69] S. Gambs, M.-O. Killijian, and M. N. n. del Prado Cortez, "Next Place Prediction Using Mobility Markov Chains," in *MPM*, 2012, pp. 3:1–3:6.

[70] B. AÄ§ÄšÅśr, K. Huguenin, U. Hengartner, and J.-P. Hubaux, "On the Privacy Implications of Location Semantics," *PETS*, vol. 2016, no. 4, pp. 165 – 183, 2016.

[71] J. Domingo-Ferrer, D. Sánchez, and J. Soria-Comas, *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, ser. Synthesis Lectures on Information Security, Privacy, & Trust. Morgan & Claypool Publishers, 2016.

[72] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[73] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *ICDE*, 2007, pp. 106–115.

[74] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, 2006, vol. 4052, pp. 1–12.

[75] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS*, 2007, pp. 94–103.

[76] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *CSF*, 2014, pp. 398–410.

[77] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," in *EDBT*, 2016.

[78] D. R. Krishnan, D. L. Quoc, P. Bhatotia, C. Fetzer, and R. Rodrigues, "Incapprox: A data analytics system for incremental approximate computing," in *WWW*, 2016, pp. 1133–1144.

[79] J. Gehrke, E. Lui, and R. Pass, "Towards privacy for social networks: A zero-knowledge based definition of privacy," in *TCC*, 2011, pp. 432–449.

[80] I. Wagner and D. Eckhoff, "Technical privacy metrics: a systematic survey," *ACM Computing Surveys*, vol. 51, no. 3, Jun. 2018.

[81] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie, "Adaptive Location Privacy with ALP," in *SRDS*, 2016, pp. 269–278.

[82] S. Cerf, V. Primault, A. Boutet, S. Ben Mokhtar, R. Birke, L. Y. Chen, S. Bouchenak, N. Marchand, and B. Robu, "Achieving privacy and utility trade-off in mobility database with PULP," in *SRDS*, 2017, pp. 164–173.

[83] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset epfl/mobility (v. 2009-02-24)," Online at: http://crawdad.org/epfl/mobility/20090224, Feb. 2009.

[84] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining Interesting Locations and Travel Sequences from GPS Trajectories," in *WWW*, 2009, pp. 791–800.

[85] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen, "From

big smartphone data to worldwide research: The mobile data challenge," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 752–771, Dec. 2013.

[86] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards rich mobile phone datasets: Lausanne data collection campaign," in *ICPS*, 2010.

[87] S. Ben Mokhtar, A. Boutet, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D 'alu, V. Primault, P. Raveneau, H. Rivano, and R. Stanica, "PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets," in *NetMob*, 2017.

[88] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving Directions Based on Taxi Trajectories," in *SIGSPATIAL*, 2010, pp. 99–108.

[89] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with Knowledge from the Physical World," in *KDD*, 2011, pp. 316–324.

[90] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-based Social Networks," in *KDD*, 2011, pp. 1082–1090.

[91] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

[92] C. Düntgen, T. Behr, and R. H. Güting, "Berlinmod: A benchmark for moving object databases," *The VLDB Journal*, vol. 18, no. 6, pp. 1335–1368, Dec. 2009.

[93] T. Brinkhoff, "A framework for generating network-based moving objects," *Geoinformatica*, vol. 6, no. 2, pp. 153–180, Jun. 2002.

[94] N. Pelekis, S. Sideridis, P. Tampakis, and Y. Theodoridis, "Hermoupolis: A semantic trajectory generator in the data science era," *SIGSPATIAL Special*, vol. 7, no. 1, pp. 19–26, May 2015.

[95] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie, "Time Distortion Anonymization for the Publication of Mobility Data with High Utility," in *TrustCom*, Aug. 2015.

[96] A. R. Beresford and F. Stajano, "Mix Zones: User Privacy in Location-aware Services," in *PerCom Workshops*, 2004, pp. 127–131.

[97] J. Freudiger, R. Shokri, and J.-P. Hubaux, "On the optimal placement of mix zones," in *PETS*, 2009, pp. 216–234.

[98] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," in *INFOCOM*, Mar. 2012, pp. 972–980.

[99] B. Palanisamy and L. Liu, "MobiMix: Protecting location privacy with mix-zones over road networks," in *ICDE*, Apr. 2011, pp. 494–505.

[100] X. Gong, X. Chen, K. Xing, D. H. Shin, M. Zhang, and J. Zhang, "From Social Group Utility Maximization to Personalized Location Privacy in Mobile Networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1703–1716, June 2017.

[101] B. Gedik and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model," in *ICDCS*, 2005, pp. 620–629.

[102] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The New Casper: Query Processing for Location Services Without Compromising Privacy," in *VLDB*, 2006, pp. 763–774.

[103] C.-Y. Chow, M. F. Mokbel, and X. Liu, "A Peer-to-peer Spatial Cloaking Algorithm for Anonymous Location-based Service," in *SIGSPATIAL*, 2006, pp. 171–178.

[104] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *WWW*, 2008, pp. 237–246.

[105] T. Xu and Y. Cai, "Feeling-based location privacy protection for location-based services," in *CCS*, 2009, pp. 348–357.

[106] B. Agir, T. Papaioannou, R. Narendula, K. Aberer, and J.-P. Hubaux, "User-side adaptive protection of location privacy in participatory sensing," *GeoInformatica*, vol. 18, no. 1, pp. 165–191, 2014.

[107] H. Ngo and J. Kim, "Location privacy via differential private perturbation of cloaking area," in *CSF*, July 2015, pp. 63–74.

[108] C. Li and B. Palanisamy, "Reversecloak: Protecting multi-level location privacy over road networks," in *CIKM*, 2015, pp. 673–682.

[109] J. Krumm, "Realistic driving trips for location privacy," in *Pervasive*, 2009, vol. 5538, pp. 25–41.

[110] V. Bindschaedler and R. Shokri, "Synthesizing Plausible Privacy-Preserving Location Traces," in *S&P*, 2016, pp. 546–563.

[111] T.-H. You, W.-C. Peng, and W.-C. Lee, "Protecting moving trajectories with dummies," in *MDM*, 2007, pp. 278–282.

[112] L. Stenneth, P. S. Yu, and O. Wolfson, "Mobile systems location privacy: "MobiPriv" a robust k anonymous system," in *WiMob*, 2010, pp. 54–63.

[113] R. Kato, M. Iwata, T. Hara, A. Suzuki, X. Xie, Y. Arase, and S. Nishio, "A dummy-based anonymization method based on user trajectory with pauses," in *SIGSPATIAL*, 2012, pp. 249–258.

[114] P. Shankar, V. Ganapathy, and L. Iftode, "Privately Querying Location-based Services with SybilQuery," in *UbiComp*, 2009, pp. 31–40.

[115] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *CCS*, 2007, pp. 161–171.

[116] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, "CAP: A Context-Aware Privacy Protection System for Location-Based Services," in *ICDCS*, June 2009, pp. 49–57.

[117] R. Assam and T. Seidl, "Preserving Privacy of Moving Objects via Temporal Clustering of Spatio-temporal Data Streams," in *SIGSPATIAL Workshops*, 2011, pp. 9–16.

[118] K. Micinski, P. Phelps, and J. S. Foster, "An Empirical Study of Location Truncation on Android," in *MOST*, May 2013, pp. 1–10.

[119] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *CCS*, 2014, pp. 251–262.

[120] S. Oya, C. Troncoso, and F. Pérez-González, "Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms," in *CCS*, 2017, pp. 1959–1972.

[121] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "A predictive differentially-private mechanism for mobility traces," in *PETS*, 2014, vol. 8555, pp. 21–41.

[122] ——, "Constructing elastic distinguishability metrics for location privacy," in *PETS*, vol. 2015, Jun. 2015, pp. 156–170.

[123] Y. Xiao, L. Xiong, S. Zhang, and Y. Cao, "Loclok: Location cloaking with differential privacy via hidden markov model," *VLDB*, vol. 10, no. 12, pp. 1901–1904, Aug. 2017.

[124] L. Yu, L. Liu, and C. Pu, "Dynamic differential location privacy with personalized error bounds," in *NDSS*, 2017.

[125] R. Shokri, P. Papadimitratos, G. Theodorakopoulos, and J.-P. Hubaux, "Collaborative location privacy," in *MASS*, Oct 2011, pp. 500–509.

[126] U. M. Aïvodji, K. Huguenin, M.-J. Huguet, and M.-O. Killijian, "SRide: A Privacy-Preserving Ridesharing System," in *WiSec*, 2018, pp. 40–50.

[127] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-l. Tan, "Private queries in location based services: anonymizers are not necessary," in *SIGMOD*, 2008, pp. 121–132.

[128] S. Jaiswal and A. Nandi, "Trust No One: A Decentralized Matching Service for Privacy in Location Based Services," in *MobiHeld*, 2010, pp. 51–56.

[129] A. Narayanan, N. Thiagarajan, M. Lakhani, M. Hamburg, and D. Boneh, "Location privacy via private proximity testing," in *NDSS*, 2011.

[130] S. Pidcock and U. Hengartner, "Zerosquare: A Privacy-Friendly Location Hub for Geosocial Applications," in *MOST*, May 2013.

[131] H. Carter, B. Mood, P. Traynor, and K. Butler, "Secure Outsourced Garbled Circuit Evaluation for Mobile Devices," in *USENIX Security*, 2013, pp. 289–304.

[132] S. Chakraborty, C. Shen, K. R. Raghavan, Y. Shoukry, M. Millar, and M. Srivastava, "ipShield: A Framework For Enforcing Context-Aware Privacy," in *NSDI*, Apr. 2014, pp. 143–156.

[133] K. Fawaz and K. G. Shin, "Location privacy protection for smartphone users," in *CCS*, 2014, pp. 239–250.

[134] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in *SIGSPATIAL Workshops*, 2008, pp. 52–61.

[135] O. Abul, F. Bonchi, and M. Nanni, "Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases," in *ICDE*, 2008, pp. 376–385.

[136] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Anonymizing Moving Objects: How to Hide a MOB in a Crowd?" in *EDBT*, 2009, pp. 72–83.

[137] N. Li, W. Yang, and W. Qardaji, "Differentially private grids for geospatial data," in *ICDE*, 2013, pp. 757–768.

[138] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," in *INFOCOM*, May 2017, pp. 1–9.

[139] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *SECURECOMM*, 2005, pp. 194–205.

[140] F. D. McSherry, "Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis," in *SIGMOD*, 2009, pp. 19–30.

[141] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou, "Differentially Private Transit Data Publication: A Case Study on the Montreal Transportation System," in *KDD*, 2012, pp. 213–221.

[142] G. Acs and C. Castelluccia, "A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris," in *KDD*, 2014, pp. 1679–1688.

[143] D. Riboni and C. Bettini, "Differentially-private release of check-in data for venue recommendation," in *PerCom*, March 2014, pp. 190–198.

[144] A. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, Jan. 2003.

[145] D. L. Chaum, "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, Feb. 1981.

[146] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking," in *MobiSys*, 2003, pp. 31–42.

[147] S. T. Peddinti and N. Saxena, "On the Limitations of Query Obfuscation Techniques for Location Privacy," in *UbiComp*, 2011, pp. 187–196.

[148] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in *USENIX Security*, vol. 13, 2004, pp. 21–21.

[149] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *CCS*, 2015, pp. 1298–1309.

[150] "Swarm website," Online at: https://www.swarmapp.com.

[151] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on macos 10.12," *CoRR*, vol. abs/1709.02753, 2017. [Online]. Available: http://arxiv.org/abs/1709.02753

[152] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private Information Retrieval," in *FOCS*, 1995, pp. 41–.

[153] A. Yao, "Protocols for secure computations," in *FOCS*, 1982, pp. 160—164.

[154] "Location-privacy meter tool," http://people.epfl.ch/reza.shokri.

[155] V. Primault, "Adaptive Location Privacy source code," Online at: https://github.com/privamov/alp.

[156] R. Shokri, C. Troncoso, and C. Diaz, "Unraveling an old cloak: k-anonymity for location privacy," in *WPES*, 2010, pp. 2–5.

[157] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie, "Differentially Private Location Privacy in Practice," in *MOST*, May 2014.

[158] S. Oya, C. Troncoso, and F. Pérez-González, "Is geo-indistinguishability what you are looking for?" in *WPES*, 2017, pp. 137–140.

[159] "Workshop on design issues for a data anonymization competition," https://petsymposium.org/2017/workshop.php.

[160] 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.

[161] Information and P. C. of Ontario (Canada), "Privacy by design," Online at: https://www.ipc.on.ca/privacy/protecting-personal-information/privacy-by-design/.

[162] C. Community, "Crawdad website," Online at: http://crawdad.cs.dartmouth.edu.

[163] "Déplacements mtl trajet 2016," Online at: http://donnees.ville.montreal.qc.ca/dataset/mtl-trajet (in French).

[164] G. Danezis, S. Lewis, and R. Anderson, "How much is location privacy worth," in *WEIS*, 2005.

[165] "Location guard source code," Online at: https://github.com/chatziko/location-guard.

[166] "Aircloak website," Online at: https://www.aircloak.com.

[167] V. Primault, M. Maouche, A. Boutet, S. B. Mokhtar, S. Bouchenak, and L. Brunie, "ACCIO: How to Make Location Privacy Experimentation Open and Easy," in *ICDCS*, 2018.

[168] A. D. P. Team, "Learning with privacy at scale," Online at: https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html, Dec. 2017.

[169] Ùlfar Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *CCS*, 2014, pp. 1054–1067.