2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC), and IEEE Intl Conference on Intelligent Data and Security (IDS)

An Analysis of Differential Privacy Research in Location Data

Fatima Zahra Errounda
Electrical and Computer Engineering Department
Concordia University
Montreal, Canada
Email: f_errou@encs.concordia.ca

Yan Liu
Electrical and Computer Engineering Department
Concordia University
Montreal, Canada
Email: yan.liu@concordia.ca

***Abstract*—Location data is becoming ubiquitous with the spread of smart devices, and social media geo-tagged feeds. However, sharing location data may lead to serious privacy risks that must not be overlooked. Differential privacy is the standard technique that provides strong privacy guarantees regardless of the adversary's side information. Usually, this is achieved by adding noise to the true result of the statistical query extracted from the data. However, a straight forward application of differential privacy to location data is not always possible. The growing interest in designing solutions to achieve differential privacy that take into account the characteristics of location data is evident from the substantial number of works done in this field. This paper briefly reviews research works done in differential privacy targeted toward location data from the data flow perspective, including the collection, aggregation, and mining. Our goal is to help newcomers to the field to better understand the state-of-the art by providing a research map that highlights the different challenges in designing frameworks, as well as novel approaches, that tackle the characteristics of location data. We identify multiple challenges to the application of differential privacy to location data, such as the calibration of the added noise to assure utility, finding the optimal spatial division to release the location aggregate per region while balancing privacy and utility. We also discuss the future directions concluded from the analysis.**

I. Introduction

With the help of sensing technology advancement and the popularity of location-based applications, location data is becoming ubiquitous. Global positioning systems (GPS), wireless sensor networks (WSN), and social media geo-tagged locations generate daily data that covers millions of peoples and large distances. Typically, service providers collect users location data then store, analyze, and share it with other third parties. The collected data is characterized by the 5Vs (Volume, Velocity, Variety, Value and Veracity) of big data.

Given the tremendous business opportunities, it is not a surprise that big companies such as Google, Apple, Microsoft, and Facebook are interested in users' locations. However, location sharing comes with privacy risks that might lead to stalking, fraud, or kidnapping. Companies in the US are legally required to disclose their data collection and usage practices, and can be fined in case of compliance failure [1]. In 2011, a class action was filed against Microsoft before the Seattle Court in the US for tracking the location of its mobile customers using Windows Phone 7 despite the customers' requests not to be tracked.

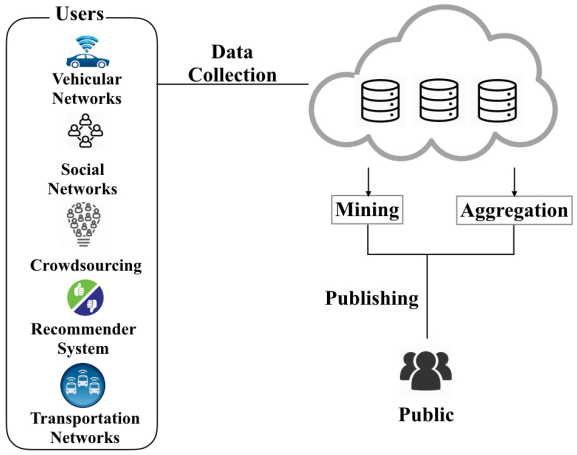


Fig. 1: Location Data Flow

Interest in location privacy is very high, and there are many techniques to achieve it. Traditional location privacy include anonymity [2], data obfuscation [3], computational Private Information Retrieval (PIR) [4]. However, many attacks point out these privacy techniques weaknesses. For example, anonymity is vulnerable to background and inference attacks where the adversary analyzes data in order to gain knowledge about an individual [5] (the user's location/identity/other sensitive information). Obfuscation techniques are susceptible to localization attacks where the adversary estimates the true user's location based on the obfuscated location [6]. PIR techniques may fall short against access pattern attacks where the attacker can intercept a user's query and response pairs and combine these access patterns with background knowledge to infer the user's identity [7].

Differential privacy leverages the concept of indistinguishability to overcome these shortcomings. Unlike other privacy techniques, it is immune to many of the privacy attacks, such as linking and correlation attacks. This is because differential privacy abstracts from the adversary's side information. Furthermore, differential privacy is immune to post-processing:

978-1-7281-0006-7/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataSecurity-HPSC-IDS.2019.00021

if private data is released with differential privacy, then an adversary cannot increase privacy loss simply by observing the released private data [8]. However, a straight forward application of differential privacy to location data is not always possible. For example, the concept of neighbourhood, as described in the traditional differential privacy setting where two datasets are neighbouring if they differ in one user, is difficult to map to the location data collection setting where there is only one user. Therefore, it is necessary to conduct an investigation to understand the challenges in applying differential privacy methods on location data and provide directions for researchers to select appropriate methods for specific problems.

Our objective in this paper is to provide readers with a broad perspective of the adoption of differential privacy to location data. We map the found work to the data flow shown in Fig.1. More specifically, our **contributions** are:

1) Provide a holistic status report on the adoption of differential privacy to location data.

2) Identify the challenges faced during the different stages of the data flow.

3) Compare the advantages and disadvantage of each solution.

4) Identify research gaps and offer future research directions.

The rest of the paper is organized as follow: Section II introduces the review method adopted in the paper and the background on differential privacy. The review of the differential privacy solutions during data collection are given in section III, followed by the solutions during data analysis in Section IV. Section V presents the future work deduced from the reviewed solutions. Finally Section VI concludes the paper.

II. Background

In this section, we present the general guidelines we followed during to collect the works to be analyzed, then we introduce the basics of differential privacy.

A. The Review Method

This review is conducted based on the guidelines defined by Kitchenham and Charters [9]. These guidelines include: research questions elaboration, the search strategy, and the selection of primary studies. We use a taxonomy to guide the reader as illustrated in 2. Based on the PICOC (Population, Intervention, Comparison, Outcomes, Context) method suggested by Petticrew and Roberts [10], we define the general concepts in order to formulate the research questions. The targeted population is the privacy researchers and practitioners in the field of location data. The identified intervention is the application of differential privacy to location data. The desired outcome is a research map that identifies and describes the differential privacy solutions in the different stages of data flow.

The following research questions emerged:

RQ1 How are the concept of differential privacy adapted during location data collection?

RQ2 What are the challenges faced in the spatial division methods that allow the release of location aggregates?

RQ3 What are the challenges in mining location data to identify points-of-interests?

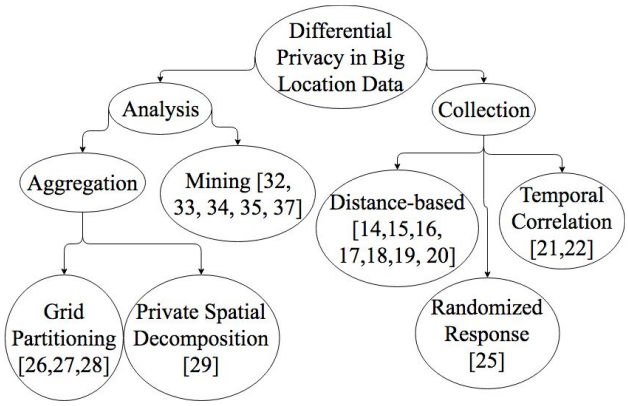


Fig. 2: Taxonomy

To detect as many of the relevant literature as possible, we base our search strategy on the keywords related to the population and the intervention. The search terms included concepts related to:

1) Differential privacy: Indistinguishability

2) Location data: Moving object, mining points-of-interest, mobile users, mobile network, location aggregate.

We use the search terms in five online databases for scientific journals, conference proceedings, and technical papers: Springer (https://link.springer.com/), Elsevier (https://www.elsevier.com), Science Direct (http://www.sciencedirect.com), ACM (http://dl.acm.org), and IEEE Xplore (http://ieeexplore.ieee.org). We perform a review of the search result and retain the most relevant according to their number of citation extracted from Googles Scholar to limit the scope of our review.

B. Differential Privacy

Differential privacy [8] is originally formulated in the context of statistical databases. It is based on the concept of neighbourhood defined as follows:

Definition 1 (Neighbourhood)**.** Two datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighbours if they differ in at most one element. Meaning that one dataset is a subset of the other and the larger dataset contains exactly one additional row.

A randomization mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if the addition or removal of a single element in the dataset $\mathcal{D}$ does not change the probability of the mechanism outcome by more than some small amount. More formally:

Definition 2 (Differential Privacy)**.** [8] A randomization mechanism $\mathcal{M}$ gives ϵ-differential privacy for all neighbour-

ing datasets $\mathcal{D}$ and $\mathcal{D}'$ and for every set of outputs $\mathcal{O} \in Range(\mathcal{M})$, if $\mathcal{M}$ satisfies:

$$Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \times Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}]$$

1) Popular Mechanisms: To achieve differential privacy, two popular mechanisms are proposed in Dwork [8], they perturb the original query q result using random noise that is calibrated with the privacy budget ϵ and the global sensitivity Δq defined as below [11]:

Definition 3 (Global Sensitivity)**.** The global sensitivity of a query $q : \mathcal{D} \rightarrow R$ is:

$$\Delta q = \max_{\mathcal{D},\mathcal{D}'} \parallel q(\mathcal{D}) - q(\mathcal{D}') \parallel_1$$

for all neighbouring $\mathcal{D}$ and $\mathcal{D}'$.

Laplace mechanism is based on the Laplace distribution: $Lap(x \mid b) = \frac{1}{2b} e^{-\frac{|x|}{b}}$ where b is the scale factor. The mechanism uses a scale factor of $\frac{\Delta q}{\epsilon}$.

Laplace mechanism adds noises to the real output using Laplace distribution as follows:

Definition 4 (Laplace Mechanism)**.** For a query q, a mechanism $\mathcal{M}(x) = q(x) + Lap(\frac{\Delta q}{\epsilon})$ satisfies ϵ-differential privacy

In many applications, the query q may map datasets to strings, strategies, or trees, which makes the addition of noise no longer sensible. McSherry and Talwar [12] propose the exponential mechanism to tackle this issue. First, it defines a score function $u : \mathcal{D} \times R \rightarrow \mathbb{R}$ that gives a real valued score to each possible output of the query function. The higher the score, the better the utility. Then, the exponential mechanism selects an output with a probability proportional to the sensitivity of the score function using the exponential distribution as follows: $e^{\left(\frac{\epsilon u(d,r)}{2\Delta u}\right)}$.

2) Properties: It is important to note that an adversary cannot increase her knowledge about the private data of an individual using the output of the privacy mechanism. That is, an adversary cannot combine a data-independent mapping f with a ϵ-differentially private mechanism $\mathcal{M}$ to gain more knowledge. In other words $f \circ \mathcal{M}$ is also a ϵ-differentially private mechanism.

McSherry et al. [13] introduce two types of mechanism compositions: sequential and parallel. When a sequence of computations provides differential privacy in isolation, the final privacy guarantee is said to be the sum of each ϵ-differential privacy. On the other hand, when the input data is partitioned into disjoint sets, independent of the original data, the composition is said to be parallel and the final privacy depends on the worst computation guarantee of the sequence. The two compositions are formalized as follows:

Theorem 1 (Sequential Composition)**.** *Consider mechanisms $\mathcal{M}_i$ that provide ϵ_i-differential privacy. A sequence of $\mathcal{M}_i$ over a dataset $\mathcal{D}$ provides $\Sigma\epsilon_i$-differential privacy.*

Theorem 2 (Parallel Composition)**.** *Consider mechanisms $\mathcal{M}_i$ that provide ϵ-differential privacy. A sequence of $\mathcal{M}_i$ over a set of disjoint datasets $\mathcal{D}_i$ provides ϵ-differential privacy.*

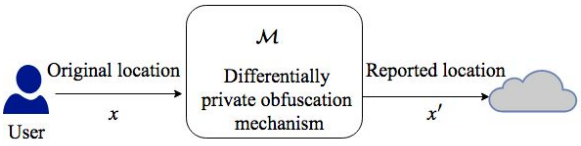


Fig. 3: Differentially private mechanism for sharing locations

III. Location Collection

Differential privacy is applied during the collection of location data. In this setting, the data collector is not considered trustworthy, therefore differential privacy is applied on every user's location before it reaches the data collector. Since there is a single user involved, the concept of neighbourhood as introduced in Definition 1 cannot be implemented in this situation. Therefore, the presented solutions guarantee indistinguishability between different locations of a single user. Meaning, they aim at preventing an adversary from accurately guessing the true location of the users x, based on the perturbed location x'. This is achieved using a differentially private mechanism $\mathcal{M}$ as depicted in Fig.3. In this section, we summarize the types of mechanism to achieve this based on the nature of the location characteristic used to guarantee indistinguishability between a user's location into three categories: distance-based methods, temporal correlation methods, and randomization-based methods.

1) Distance-based Methods: Distance-based methods, [14], [15], and [16], are based on the idea that farther apart locations should have less privacy influence on each other. It means that through perturbation any two locations within a given distance produce observations with similar distributions, and thus attackers have no way to learn the user' real location. The distance between these locations is used to relax the guarantee of indistinguishability (Definition 2). For a privacy a privacy mechanism $\mathcal{M}$, for all user locations x and y, distance is used as follows:

$$P(\mathcal{M}(x'|x) \leq e^{\epsilon d_{\mathcal{X}}(x,y)} P(\mathcal{M}(x'|y)) \tag{1}$$

where x' is the set of reported locations, and $d_{\mathcal{X}}$ is a distance metric.

This definition is coined geo-indistinguishability [15]. In other words, for a fixed radius r, the location owner enjoys a privacy guarantee of ϵr, and an observer cannot distinguish between two originating locations within r based on the observed location. To achieve geo-indistinguishability, a privacy mechanism must output locations with a probability that decreases exponentially as the points get further from the owner's original location. If we consider x_0 the original location, then the planar Laplacian centered at x_0 satisfies geo-indistinguishability. In that case, the planar Laplacian probability distribution D_ϵ is defined as follows:

$$D_\epsilon(r, \theta) = \frac{\epsilon^2}{2\pi} r e^{-\epsilon r} \tag{2}$$

where r and θ are two random variable that represent the distance and the angle between the reported location x' and the original location x_0 respectively and ϵ is the privacy budget.

Geo-indistinguishability attracts the interest of researchers generating a rich body of literature. Many improvements are made to the original definition. For example, in terms of utility, a new randomization algorithm is proposed in [17] based on linear programming techniques to release locations with better utility than the planar Laplacian mechanism. Utility is defined in terms of the expected distance between the real location and the reported location.

In terms of correlation between a single user's locations in repeated location collection, [18] introduces an R-tree that caches the previously released location obfuscations to reuse in future releases. Another way to incorporate correlation is presented in [19], [20] by using it to predict a location that is close to the user's actual location based on the previously released ones. The predicted location is released to decrease the privacy budget consumption and therefore increasing the degree of privacy guaranteed and the privacy budget is used only when predicted locations are far from the user's actual location.

Discussion: Distance-based mechanisms present an intuitive way to incorporate location characteristics into the guarantee of differential privacy by calibrating the degree of indistinguishability between locations based on their proximity. However, since the temporal aspect of the locations is not taken into account, these mechanisms are best suited for sporadic uses.

2) Temporal Correlation Methods: The privacy mechanism presented in this subsection ([21] and [22]) report a location $x_t{}'$ at time t calculated based on the temporal correlation between the original locations x_t and x_{t-1} at time t and $t-1$ respectively. The set of all possible locations that a user may occupy is built based on the temporal correlation model. The set of most probable locations S_t is then constructed with a given threshold α as follows:

$$S_t = min\{x_i | \sum_{x_i} p_t{}^-[i] \geq 1 - \alpha\} \tag{3}$$

where $p_t{}^-[i]$ is the prior probability of a user's location at time t.

The true location x_t is assumed to be part of S_t set, and the privacy mechanism then reports one of the locations in S_t with ϵ-differential privacy guarantee. This is formalized as follows:

$$Pr[\mathcal{M}(x_1) = x_t{}'] \leq e^{\epsilon} \times Pr[\mathcal{M}(x_2 = x_t{}')] \tag{4}$$

where $\mathcal{M}$ is the privacy mechanism, x_1 and x_2 are locations in S_t, ϵ is the privacy budget, and $x_t{}'$) is the reported location.

Using Laplace distribution to achieve differential privacy in this case might lead to over-perturbation of the location because of the high sensitivity of the distance between locations. Instead, the authors in [21] introduce a mechanism that transforms the locations in the S_t set to an isomorphic shape to reduce the sensitivity before uniformly sampling a location that is used to report the noisy location $x_t{}'$.

Discussion: Temporal correlation methods are based on the prior probability of a user's movement to a certain location. Let us consider a user at location A at time $t-1$, when the user moves to another location at time t, the privacy mechanism must know the probabilities assigned to all his potential movement from location A, however, this is not always possible in big location data with millions of possible locations. A potential solution to this problem is presented in [23] where the distance-based and temporal correlation approaches are combined by defining the (α,r)-dataset at a time t as a set of probable (i.e., cumulative prior probability is more than 1- α) landmarks or anchor point a user might visit in a personalized privacy circle with radius r.

3) Randomized Response Method: This type of mechanism uses randomized response algorithm to generate a location x' based on original location x. The randomized response algorithm is introduced in [24] where the user responds either truthfully or the opposite answer to a sensitive question depending on a coin flip. It is commonly used in surveys of people's "yes or no" opinions about a sensitive question. Because the user gives a randomized response, the surveyors cannot determine the individual's true answer, but can still extract useful statistics [24].

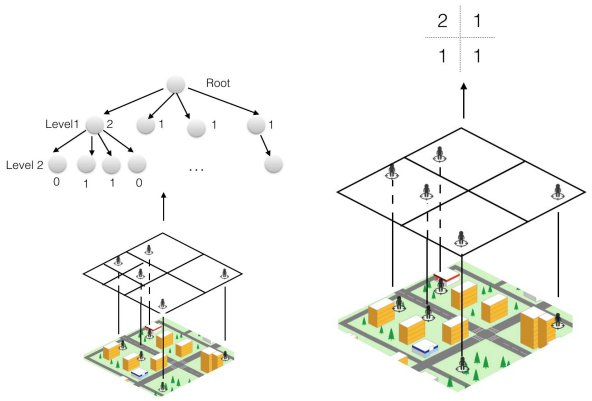
Let us consider a location that is represented as a vector of binary values where each index corresponds to a known location and the bit value shows the presence or absence (i.e., 1 or 0 respectively) of the user in the given location. An example of this type of location is indoor positioning data, where installed devices in fixed locations send signals to users, and depending on the signal strength the user identifies her current indoor positioning. Then, each user locally applies the randomized response mechanism on the bits in the location vector before sharing it with a data collector as introduced in [25]. In this case, the relationship is between all user location pairs x_1 and x_2 and the reported location x' is translated as below:

$$\frac{P(\mathcal{M}(x_1) = O)}{P(\mathcal{M}(x_2) = O)} = e^{\epsilon} \tag{5}$$

where $\mathcal{M}$ is the privacy mechanism, ϵ is the privacy budget, and $x' \in O$, and O is any output of $\mathcal{M}$.

The privacy mechanism $\mathcal{M}$ guarantees that an adversary cannot guess if the user is at location x_1 or x_2 based on the observed location x'.

Discussion: The randomized response methods work well when the set of possible locations is fixed and the user's movements are constrained such as in indoor positioning systems. However, it does not take into account the temporal correlation between the user's locations, which can lead to privacy breaches. Especially that the privacy guarantee by a factor of up to $e^{T\epsilon}$ for T consecutive location collection due to the sequential composition property of differential privacy (Theorem 1). This means that the indistinguishability degree



(a) Private Spatial Decomposition (b) Grid partitioning

Fig. 4: An example of the two spatial division approaches: (a) the space is divided using a quad-tree, the number of users present in each cell is recorded in the corresponding node (b) the space as divided using a grid of 2 by 2 and the corresponding number of users

decreases as the frequency of collection T increases reducing the effectiveness of the privacy method.

IV. Location Data Analysis

For location data analysis, users share their true locations with the data collector, who is considered trustworthy. Then the data collector uses the location data (i) to calculate aggregates such as the number of locations in all the cells of the location space, and (ii) to mine for user's behavioural movement patterns. We summarize these uses in this section.

A. Aggregate Sharing

In this subsection, we focus on the count of users per cell of the location space. The concept of neighbourhood is applied as introduced in Definition 1 and the global sensitivity (Definition 3) is calculated accordingly to calibrate the noise to the count. However, since this type of counts are the basics for answering range queries, means, medians, the challenge in designing the space division method is in finding the optimal division that achieves utility. For example, when answering a range query based on the counts calculated from the spatial division, the accumulated noise from the subregions covered by the range query reduces the query utility. In this subsection we review two methods for spatial division as illustrated in Fig.4.

1) Grid Partitioning: One way to release the number of users located on a map is by laying a grid over the whole location space, then counting the number of users in each cell of the grid. Laplace noise is then added to each count with sensitivity 1, since adding or removing a user changes the counts in each cell of the grid by a maximum of 1. The challenge then is to select the grid granularity that induces the lowest amount of noise and therefore assures utility. With the uniform grid partitioning presented in [26], the selected granularity is proportional to the total number of locations in the dataset, the privacy budget, and a constant that reflects how uniformly the locations are distributed on the grid cells.

A challenge arise with the uniform grid approach when answering range queries that cover multiple cells fully or partially. Since the cumulative noise is proportional to the number of cells involved in calculating the value of the range query, it is important to avoid over-partitioning sparse areas or under-partitioning dense areas. An adaptive grid partitioning is presented in [26], where a second level of grid partitioning is done on denser areas. The granularity of the second partitioning depends on the noisy count of the dense region. The total privacy budget ϵ is then distributed on the two levels of partitioning.

The authors in [27], [28] adopt the two-level grid partitioning approach to protect the location privacy of users who participate in a crowdsourcing system. An important difference with [26] is that in crowdsourcing systems the user identities can be unknown, and what matters is knowing if there is any user at a given region in order to assign him/her a task. This creates a constraint on the second level of the grid partitioning where it should guarantee that at least one user is in a cell. To cater to this constraint, the authors in [27], [28] propose a heuristic method to choose the granularity of the second level partitioning but only to the point where there is at least one user in a cell.

2) Private Spatial Decomposition: Spatial decomposition as introduced in [29] is a partitioning of the location space into smaller areas, indexed using a tree where each node represent a portion of the space and the tree leaves are the smallest cells in the space. The number of users is calculated at each node, including the leaves. When answering a query, the tree is traversed to identify the nodes that correspond to the regions covered by the query and the sum of corresponding nodes' counts is calculated. An advantage in using spatial decomposition is that the sensitivity of the counts (number of users in each node) when adding or removing a user is 1.

There are two main types of spatial decomposition: data independent and data dependent. Data independent decomposition does not consider the distribution of the locations in space. For example, quad-trees recursively split the location space into four equal regions regardless of the number of locations in each region. An example of data dependent decomposition is the kd-tree where the space is recursively split via lines passing through a median calculated based on the location distribution. However, in the case of data-dependent trees, an adversary can use information, such as the median in case of kd-trees, to gain more knowledge about users' locations. Therefore, the structure of the tree must also be private, meaning privately calculating private means.

Since the set of nodes of each level in the tree cover the whole space, the challenge is in distributing the privacy budget ϵ on the different levels of the tree. A uniform distribution of the privacy budget over the levels of the tree is presented in [29]. However, the utility decreases as the tree depth increases. A geometric budget allocation strategy increases the budget geometrically at each tree level so the leaf counts have the

highest accuracy.

Discussion: Grid partitioning is an intuitive solution to divide the location space into using an equi-width grid over the location data domain. However, this solution may not be adapted to sparse spatial domains. For example, suppose we use a $100m^2$ to represent the population distribution of Canada. Given that the population distribution has the highest density in urban areas, most of the values in rural areas are noisy counts of 0. If we want to calculate the population of an area that covers a rural region, the output is the sum of noisy counts which accumulates into a large error.

A common challenge in both data-dependent and data-independent private spatial decomposition [29] is the choice of the tree height. In fact, the query accuracy is influenced by the tree height, as more levels compete for the same budget amount. Another challenge is choosing the right structure for the dataset. The work in [30] compares the results of using different tree structures for private spatial decomposition on a recommendation system. They show that kd-tree performs best in skewed data (i.e., different levels of sparsity), while quadtree is better suited to uniform spatial distributions.

B. Mining

Differentially private mining is concerned with extending the current non-private mining algorithms to differentially private algorithms. Here, we focus on the most common mined information in location data, namely mining for points-of-interest.

Points-of-interests are characterized by two parameters: the encompassing geographic region of interest and the number of visits [31]. The challenge in adding noise mining for points-of-interests is in the magnitude of the global sensitivity. For example, mining points-of-interests of a population is based on the frequency visits counts. To calculate the global sensitivity of these counts, a user is removed the dataset and the maximum difference between the neighbouring datasets is calculated. However, in this case, the difference can range from 0 to the maximum visit frequency of the user. This hinders utility since a large global sensitivity results in utility loss due to the increased value of the added noise.

One way to reduce sensitivity is by using spatial decomposition [29] to calculate local (smaller) sensitivity per geographic region as presented in [32], [33], [34] and [35]. The local sensitivity is then used to calibrate the Laplace noise of the total number of visits per geographic region. Then a density-based spatial clustering (DBSCAN) [36] is used to identify the points-of-interests' encompassing regions. Finally Laplace noise is used to perturb the centroids of the regions of interests with a local sensitivity that is based on the maximal difference of all points in the encompassing region.

Another approach to reduce the global sensitivity is by limiting the contribution of a single user visits to a given location. It can be achieved by sampling the location data using a threshold l such that only l visits are retained per user as proposed in [37]. This threshold is then used as the upper bound of the global sensitivity.

Discussion: Clustering techniques are common in mining for points-of-interests, however it is difficult to assess the clustering quality due to the difficulty in obtaining global sensitivities. In fact, limiting the number of location visits per user considered to reduce the sensitivity has the disadvantage of distorting large counts. As a consequence, impacting the accuracy of the identified points-of-interests.

V. Future Directions

The proposed methods in Section III guarantee the privacy of location data collection from different perspectives, while the solutions presented in Section IV protect the user's privacy by ensuring that the published statistical data does not depend on the presence or absence of an individual user in the dataset. A summary of these solutions is presented in Table I. Based on the challenges introduced above, we identify two areas that require further attention: the utility and privacy tradeoff and the resiliency of present location collection methods to privacy attacks.

A. Utility and Privacy Tradeoff

The definition of utility and privacy depend on the objective of applying differential privacy: protecting the user's location, such as in the case of location collection, or the presence/absence of the user in the dataset, such as the case when sharing location aggregates.

1) Utility in location collection methods: Laplace mechanism when used on count queries where one user corresponds to one record enables the achievement of high level of privacy while preserving utility. This is because the low sensitivity of count queries (the contribution of a singe user changes the outcome of the query by a maximum of 1 between neighbouring datasets). When protecting a user's location, the utility of distance-based solutions is evaluated as the distance between the reported location and the true location. However, since the sensitivity used to induce noise in the reported location depends on the distance metric, the reported location may result in utility loss as mentioned in Challenge **C1**. For example, if a user is using her obfuscated location to request the closest shopping mall within a radius from her true location, the resulting shopping may be far away from her. Therefore, even though her location privacy is preserved, utility is sacrificed.

Furthermore, distance-based solutions treat space in a uniform way, regardless of the privacy needs difference between imposing the addition of the same amount of noise everywhere on the location space which results in utility loss **C2**. [38] defines an elastic indistinguishability metric that captures the different degrees of density of each area on the map, and a randomization algorithm that adapts the level of noise while achieving the same degree of privacy everywhere.

Moreover, the distance-based and randomizes response methods guarantee indistinguishability during a single location collection. However, when these methods are applied in a continuous setting, privacy declines exponentially **C4**. Dwork et al. [39] propose two different privacy goals for continuous

	Method		Description	Advantages	Challenges
Privacy protection of location data collection	Distance-based methods [14] , [15], [16], [17], [18], [19], [20]		An adversary cannot distinguish between any two originating locations within a radius based on the reported location. The indistinguishability between any two users locations is proportional to the distance between these locations.	Intuitive and easy to implement	**C1** Utility is compromised when using Laplace noise calibrated with distance, **C2** when treating the location space uniformly regardless of the population density. **C3** Vulnerable to privacy attacks that use the temporal correlation between consecutive locations
	Temporal correlation methods [21], [22]		Use the set of all probable locations (where the user might appear) between consecutive timestamps to hide the true location. This is achieved by making the probable locations indistinguishable by an adversary.	Takes into account the temporal correlation between consecutive locations	**C4** Temporal correlation between any two locations must be modelled, which may not scale when the number of locations is large
	Random response methods [25]		Use the randomized response algorithm to make any two originating locations from a fixed set of locations based on the reported location.	Intuitive and easy to implement	**C5** Privacy decreases exponentially with repeated data collection
Privacy protection of location data analysis	Aggregation	Grid partitioning [26], [27], [28]	Divide the location space using a uniform grid, then induce noise to the counts of each cell	Intuitive and easy to implement	**C6** Defining the guidelines to choose the right grid size, tree structure and height per dataset
		Private Spatial Decomposition [29]	Recursively divide the location using a tree structure, then induce noise to the count at each node of the tree	Assures utility by providing different levels of division granularity	
	Mining of Points-of-interests [32], [33], [34], [35], [37]		Limits the number of visits per user considered when calculating the sensitivity of the popularity of a location	Induce less noise to the number of visits per user to each location to assure utility of the points-of-interests identification	**C7** Assessing the utility of the clustering results is difficult since setting a threshold of visits per user may distort larger values

TABLE I: Summary of the reviewed work

settings, namely event-level that hides the presence or absence of a single user event and user-level that hides the presence of a user with all her data on the timeline. New possibilities for combination of these methods need to be explored.

2) Utility in location analysis methods: Tuning the parameters for spatial decomposition, such as the number of cells in the grid partitioning or the tree height in the private spatial decomposition has the advantage of allowing the proposed solutions to be applied in different settings. Yet, this raises the challenge (**C6**) of defining the guidelines to choose these parameters for optimal utility. A comparison between different tree structure is done in [30] that shows the link with between data sparsity and the tree structure choice. However, the study is limited to recommender systems. We believe that further investigation is necessary in defining guidelines to choosing these tuning parameters.

B. Privacy Attacks Resiliency

The distance-based and temporal correlation methods presented above make use of the characteristics of location data to create new methods to achieve differential privacy. However these variations may give rise to novel privacy attacks that cause privacy leakage (challenge **C3**).

For example, geo-indistinguishability [15] works well in the case of a sporadic use. However, the correlation between the locations may lead to privacy loss under repeated use. An attack presented in [40] takes advantage of this weakness to successfully run an inference attack where the user's points-of-interests were predicted with reasonable precision (63% of the points of interests were identified) based on the reported perturbed locations.

The vulnerability of geo-indistinguishability [15] to correlation analysis attacks is shown in [41] where an attacker uses a regression algorithm to estimate the users true location based on the previously obfuscated locations. It shows that the prediction accuracy, estimated as the distance between the predicted location and the true location, improves as the prediction duration increases.

We believe that similar evaluation of privacy attacks need to be investigated in the other differential privacy variations that release noisy locations.

VI. Conclusion

This paper presents an overview of differential privacy advances in big location data. We mapped the differential privacy solutions to the data flow as follows: (i) location collection, including solutions that make use of the spatial characteristics and temporal characteristics of location data, (ii) privacy-preserving frameworks that design solutions for location aggregation and mining by combining the traditional differential privacy mechanisms (i.e., Laplace, Exponential) with partitioning and clustering techniques to tackle the high sensitivity challenge in location data. Although differential privacy is originally designed for statistical datasets, the subsequent literature proves its power and versatility in guaranteeing

privacy in novel settings, such as location obfuscation. However, there are research directions that need to be explored, such as the impact of the high sensitivity on the utility and the robustness of the novel solutions w.r.t location privacy attacks.

REFERENCES

[1] A. S. Cheung, "Location privacy: The challenges of mobile service devices," *Computer Law & Security Review*, vol. 30, no. 1, pp. 41–54, 2014.

[2] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st international conference on Mobile systems, applications and services*. ACM, 2003, pp. 31–42.

[3] D. E. Bakken, R. Rarameswaran, D. M. Blough, A. A. Franz, and T. J. Palmer, "Data obfuscation: Anonymity and desensitization of usable data sets," *IEEE Security & Privacy*, vol. 2, no. 6, pp. 34–41, 2004.

[4] E. Kushilevitz and R. Ostrovsky, "Replication is not needed: Single database, computationally-private information retrieval," in *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*. IEEE, 1997, pp. 364–373.

[5] J. Krumm, "Inference attacks on location tracks," in *International Conference on Pervasive Computing*. Springer, 2007, pp. 127–143.

[6] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: optimal strategy against localization attacks," in *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012, pp. 617–627.

[7] M. S. Islam, M. Kuzu, and M. Kantarcioglu, "Access pattern disclosure on searchable encryption: Ramification, attack and mitigation." in *Ndss*, vol. 20, 2012, p. 12.

[8] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[9] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," in *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. sn, 2007.

[10] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008.

[11] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[12] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, 2007, pp. 94–103.

[13] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

[14] E. ElSalamouny and S. Gambs, "Differential privacy models for location-based services," *Transactions on Data Privacy*, vol. 9, no. 1, pp. 15–48, 2016.

[15] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013, pp. 901–914.

[16] D. Quan, L. Yin, and Y. Guo, "Enhancing the trajectory privacy with laplace mechanism," in *Trustcom/BigDataSE/ISPA, 2015 IEEE*, vol. 1, Aug 2015, pp. 1218–1223.

[17] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 251–262.

[18] X. Ma, J. Ma, H. Li, Q. Jiang, and S. Gao, "Agent: an adaptive geo-indistinguishable mechanism for continuous location-based service," *Peer-to-Peer Networking and Applications*, pp. 1–13, 2017.

[19] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "A predictive differentially-private mechanism for mobility traces," in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2014, pp. 21–41.

[20] ——, "Location privacy via geo-indistinguishability," *ACM SIGLOG News*, vol. 2, no. 3, pp. 46–69, 2015.

[21] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1298–1309.

[22] Y. Xiao, L. Xiong, S. Zhang, and Y. Cao, "Loclok: location cloaking with differential privacy via hidden markov model," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1901–1904, 2017.

[23] S. Wang, R. Sinnott, and S. Nepal, "Protecting the location privacy of mobile social media users," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1143–1150.

[24] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[25] J. W. Kim, D.-H. Kim, and B. Jang, "Application of local differential privacy to collection of indoor positioning data," *IEEE Access*, 2018.

[26] N. Li, W. Yang, and W. Qardaji, "Differentially private grids for geospatial data," in *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013)*, ser. ICDE '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 757–768. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2013.6544872

[27] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 919–930, 2014.

[28] H. To, G. Ghinita, L. Fan, and C. Shahabi, "Differentially private location protection for worker datasets in spatial crowdsourcing," *IEEE Transactions on Mobile Computing*, vol. 16, no. 4, pp. 934–949, 2017.

[29] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ser. ICDE '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 20–31. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2012.16

[30] J. D. Zhang, G. Ghinita, and C. Y. Chow, "Differentially private location recommendations in geosocial networks," in *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, vol. 1. IEEE, 2014, pp. 59–68.

[31] S.-S. Ho, "Preserving privacy for moving objects data mining," in *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE, 2012, pp. 135–137.

[32] S.-S. Ho and S. Ruan, "Differential privacy for location pattern mining," in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, ser. SPRINGL '11. New York, NY, USA: ACM, 2011, pp. 17–24. [Online]. Available: http://doi.acm.org/10.1145/2071880.2071884

[33] ——, "Preserving privacy for interesting location pattern mining from trajectory data," *Trans. Data Privacy*, vol. 6, no. 1, pp. 87–106, Apr. 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2612156.2612161

[34] S. Wang and R. O. Sinnott, "Supporting geospatial privacy-preserving data mining of social media," *Social Network Analysis and Mining*, vol. 6, no. 1, p. 109, 2016.

[35] ——, "Protecting personal trajectories of social media users through differential privacy," *Computers & Security*, vol. 67, pp. 142–163, 2017.

[36] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[37] G. Acs and C. Castelluccia, "A case study: privacy preserving release of spatio-temporal density in paris," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1679–1688.

[38] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Constructing elastic distinguishability metrics for location privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 2, pp. 156–170, 2015.

[39] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 2010, pp. 715–724.

[40] V. Primault, S. B. Mokhtar, C. Lauradoux, and L. Brunie, "Differentially private location privacy in practice," *Proceedings of the Third Workshop on Mobile Security Technologies*, vol. abs/1410.7744, 2014.

[41] R. Al-Dhubhani and J. Cazalas, "Correlation analysis for geo-indistinguishability based continuous lbs queries," in *Anti-Cyber Crimes (ICACC), 2017 2nd International Conference on*. IEEE, 2017, pp. 203–208.