TRANSACTIONS ON DATA PRIVACY 6 (2013) 87–106

Preserving Privacy for Interesting Location Pattern Mining from Trajectory Data

Shen-Shyang Ho*, **Shuhua Ruan****

*School of Computer Engineering, Nanyang Technological University, Singapore, 639798, Singapore

**College of Computer Science, Sichuan University, Chengdu, Sichuan, 610065, China

E-mail: ssho@ntu.edu.sg, ruanshuhua@scu.edu.cn

Abstract. One main concern for individuals participating in the data collection of personal location history records (i.e., trajectories) is the disclosure of their location and related information when a user queries for statistical or pattern mining results such as frequent locations derived from these records. In this paper, we investigate how one can achieve the privacy goal that the inclusion of his location history in a statistical database with interesting location mining capability does not substantially increase risk to his privacy. In particular, we propose a (ϵ, δ)-differentially private interesting geographic location pattern mining approach motivated by the sample-aggregate framework. The approach uses spatial decomposition to limit the number of stay points within a localized spatial partition and then followed by density-based clustering. The (ϵ, δ)-differential privacy mechanism is based on translation and scaling insensitive Laplace noise distribution modulated by database instance dependent smoothed local sensitivity. Unlike the database independent ϵ-differential privacy mechanism, the output perturbation from a (ϵ, δ)-differential privacy mechanism depends on a lower (local) sensitivity resulting in a better query output accuracy and hence, more useful at a higher privacy level, i.e., smaller ϵ. We demonstrate our (ϵ, δ)-differentially private interesting geographic location discovery approach using the region quadtree spatial decomposition followed by the DBSCAN clustering. Experimental results on the real-world GeoLife dataset are used to show the feasibility of the proposed (ϵ, δ)-differentially private interesting location mining approach.

Keywords. Differential privacy, moving objects data mining, frequent location pattern mining, density-based clustering, spatial decomposition.

1 Introduction

The interesting location discovery task is an important intermediate procedure in many real-world prediction [13] and recommender system [31, 32] applications. Informally, an interesting location can be understood as a geographic region that many people have visited. The interesting location patterns and statistics can be extracted from individual location history records, the so-called *trajectory data*. With the prevalence of location acquisition technology on mobile devices, these data can be easily obtained with permission from the individuals carrying the mobile devices. However, to obtain meaningful patterns, a large and diverse amount of individual location history records have to be collected for analysis or mining. One main concern for individuals participating in such data collection is the disclosure of their location and related information when a user (repeatedly) queries for

the analysis or mining results. The privacy goal for this problem is to ensure that an individual's participation in such a *statistical database* does not substantially increase risk to his privacy. The desired privacy level for such a goal is captured by the measure of *differential privacy* [7]. If one can provide a guarantee that limits privacy risk when one's location history is included into a database, an individual would be more willing to allow his location history to be collected. With the increase number of participations, results mined from the location history database become more accurate with higher confidence.

For a query on a location history statistical database with data mining capabilities, one assumes that only statistics and patterns are returned to users. Without disclosing a person's membership in the location history database or even when the person's location history is *not* in the database, an attack can take place to gain information about the person on the discovered spatial patterns. This is especially true when one has *auxiliary information* [9] not available from the location history database.

An overview of the interesting location discovery task for a trajectory database is shown in Figure 1. In the first stage, stay points are extracted from trajectories (see Section 3.1). In the next stage, interesting locations with their data count and centroid information are mined from the stay point database.

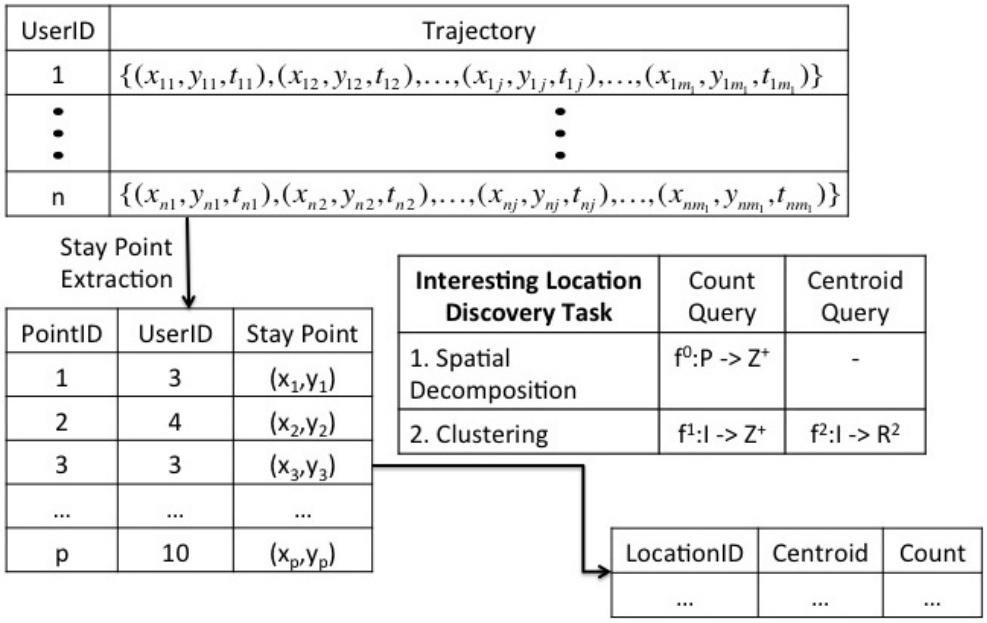

Figure 1: Overview of the Interesting Location Discovery Task for a Trajectory Database.

There are two specific privacy policies that we would like to control: (i) the precision of the discovered locations and (ii) the output counts for queries on these locations. One would like to preserve the privacy for an individual, say John, and also to ensure the usefulness and accuracy for database queries that utilize the discovered location patterns and their stay point counts. Differential privacy ensures that the "ability of an adversary to inflict harm [...] should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset" [9]. Figure 2 shows a non privacy-preserving example when an individual (say John) decides to opt out from the location history database, an

interesting region changes from Region A to Region A'. With this change, an adversary discovers a more specific location (dashed non-shaded region, $A - A'$) within Region A that John visits regularly.

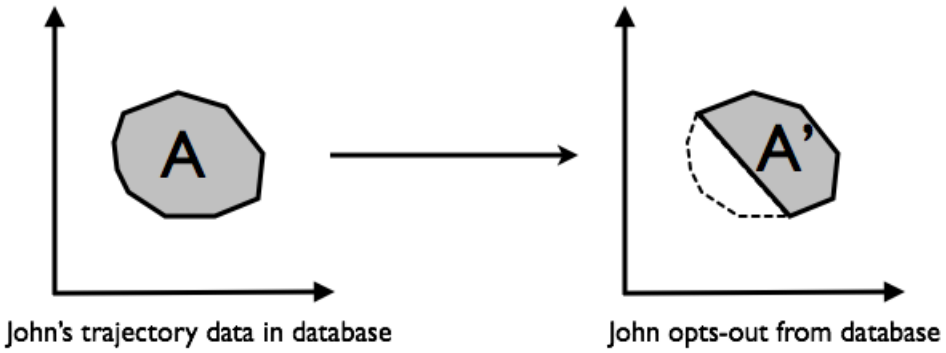

Figure 2: An interesting region changes from Region A to Region A' when John opts out from the location history database.

In this paper, we investigate in detail how one can ensure differential privacy for two spatial queries needed to accomplish the interesting location discovery task with differential privacy, namely:

$$\text{[Data Count in Region]} \qquad f^1 : I \to \mathbf{Z}^+ \tag{1}$$

$$\text{[Centroid for Region]} \qquad f^2 : I \to \mathbf{R}^2 \tag{2}$$

such that I is the set of interesting locations computed from a database D, with each record containing a stay point (latitude, longitude) and a number unique to each individual, $\mathbf{Z}^+$ is the set of positive integers, and $\mathbf{R}^2$ is the 2D space. Besides the above two spatial queries, we apply differential privacy mechanism to the count query

$$\text{[Data Count in Partition]} \qquad f^0 : P \to \mathbf{Z}^+ \tag{3}$$

such that P is the set of partition computed using spatial decomposition such as quadtree, KD-tree, or R-tree, from the database D.

The main contributions of the paper are: (i) the first application of (ϵ, δ)-differential privacy mechanism to a spatial data mining task, in particular, the interesting geographic location discovery problem and (ii) the theoretical justification for our (ϵ, δ)-differentially private interesting location pattern mining approach based on the sample-aggregate framework [24] . One distinct advantage of (ϵ, δ)-differential privacy Laplace noise mechanism over the conventional ϵ-differential privacy Laplace noise mechanism is the much lower noise level required to achieve the same ϵ privacy level. This lower noise level, in turn, results in a better query output accuracy. Hence, one obtains a more useful output at similar privacy level (with a negligible privacy loss) with a slight variation in the output distribution.

The rest of the paper is organized as follows. In Section 2, we briefly review privacy issues in data mining, privacy goals, and related work on privacy approaches for spatial and location-based domains. In Section 3, we define the interesting location pattern mining problem for trajectory data, introduce the concept of differential privacy, and the Laplace

noise differential privacy mechanism. In Section 4, we describe and discuss in detail our proposed differentially private pattern mining solution to discover interesting locations. In Section 5, experimental results on the GeoLife dataset [31], consisting of human trajectory collected using mobile devices, are used to show the feasibility of the proposed (ϵ, δ)-differentially private interesting location pattern mining approach.

2 Related Work

There have been many research work on preserving privacy for sensitive information in statistical databases [2] (and the references therein) and privacy preserving knowledge discovery techniques [4] (and the references therein). Decision tree classifier is the most commonly used data mining algorithm that research has been done to explore privacy issues [2, 11, 19]. Agrawal and Srikant [2] addressed the task of developing accurate decision tree using perturbed data sets to preserve data record privacy. Lindell and Pinkas [19] developed theoretically a protocol on the task when two parties perform data mining, in particular running the ID3 decision tree classifier, on the union of their confidential databases without revealing any confidential information to each other. Friedman and Schuster [11] used the Privacy Integrated Queries (PINQ) API [21], an infrastructure for implementing differentially private algorithms, to construct a differentially private ID3 decision tree classifier.

Conventional privacy goals such as data/response perturbation, k-anonymity, and differential privacy have been considered for privacy preserving data mining algorithms for statistical databases. An algorithm achieves privacy protection if it can still perform well (i.e., accurate prediction model) when the records in the database are perturbed [2]. Although such a privacy goal is intuitive from a data mining perspective, it lacks theoretical results to establish a relationship between the model accuracy and the amount of data/response perturbation. An algorithm achieves k-anonymity privacy protection if each record in the database cannot be distinguished from at least $k-1$ records in the database when the algorithm is applied [27]. An algorithm achieves ϵ-differential privacy protection if the outputs returned from the algorithm for any two databases differ on a single record are approximately the same described by a probabilistically bound using an exponential function of ϵ [7] (see Section 3.2).

Privacy is an important issue for both spatial and spatiotemporal data mining due to the challenge of protecting personal location information [3, 28, 14] (and the references therein). k-Anonymity approaches are used frequently to preserve privacy during personal location collection in location-based services [12] or to perform privacy preserving spatial (range or nearest neighbors) queries [18]. Obfuscation or perturbation techniques are used to hide and confuse an adversary by modifying a user's trajectory or location history. To achieve anonymity in publishing trajectory dataset is also a challenging problem. Monreale et al. [22] proposed a method to transform trajectory data based on spatial generalization (i.e., replacing exact locations by their approximations) and k-anonymity justified by a theoretical upper bound to the probability of re-identification. Nergiz et al. [23] proposed a method that ensures k-anonymity and included sampling from the anonymized data to prevent leakage due to the anonymization step. Yarovoy et al. [29] introduced a form of k-anonymity based on spatial generalization for moving objects and proposed two anonymization approaches that achieve their notion of k-anonymity. Recently, Abul et al. [1] also introduced the concept of $(k, \delta)-$anonymity that exploits the inherit location uncertainty of moving objects, represented by radius δ, to reduce the amount of distortion

needed to anonymize data using clustering and spatial perturbation. Due to the difficulty in obtaining global sensitivities needed to achieve $\epsilon-$differential privacy for location-based pattern mining queries, Ho [17] suggested the necessity to handle location-based pattern mining approaches from the (ϵ, δ)-differentially privacy perspective [8].

3 Background

In Section 3.1, we define and describe the interesting location mining problem. In Section 3.2, we formally define differential privacy and describe existing concepts and tools that are needed to construct the differentially private pattern mining approach to discover interesting locations.

3.1 Discovering Interesting Locations

Let $traj_i^{k_i} = \{p_1, p_2, \ldots, p_{k_i}\}$ be a trajectory consisting of k_i measurements. Each $p_j = (x_j, y_j, t_j)$ such that x_j is the latitude, y_j is the longitude, and t_j is a timestamp and $t_j < t_{j+1}$. Following the definition in [31] with a slight variation, we define a **stay point** to be the center (x, y) of a circle region with radius η that a trajectory stays for at least a time period of $\triangle T$ (see Figure 3).

An **individual interesting location** is a region containing more than r stay points for the individual. Given a set of trajectories, $TJ = \{traj_1^{k_1}, traj_2^{k_2}, \ldots, traj_s^{k_s}\}$. An **interesting location** is a region containing more than r' stay points. One can also define an interesting location as a region that satisfies the condition that it is an individual interesting location for at least m individuals if each individual has more than r' stay points in the region.

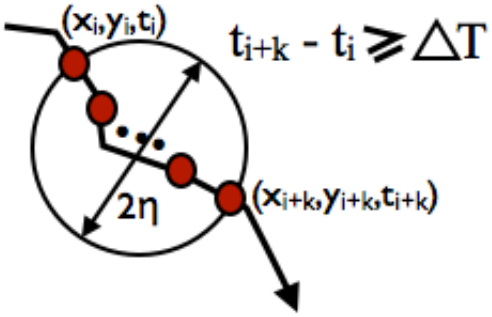

Figure 3: Stay point is the center of the circle region.

While the interesting location discovery problem, under a more complex assumption, is motivated by its application to recommender system [31], it is a main research topic in spatial and spatial-temporal data mining which the objective is to find regions with high object density [10, 14, 15]. One notes that the task of finding interesting locations is similar to the task of finding frequent locations and it is an intermediate step to find correlation between interesting or frequent locations.

3.2 Differential Privacy

Differential privacy ensures that one can extract meaningful knowledge or information from a dataset while privacy is preserved for any individual whether his data is in the dataset or not.

Definition 1. [9] A randomized function K gives $\epsilon-$ differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subset Range(K)$,

$$Pr[K(D_1) \in S] \leq \exp(\epsilon) \times Pr[K(D_2) \in S].$$

The function K satisfies the fact that when one element is removed from the dataset, no output would become significantly more or less likely. From another perspective, ϵ-differential privacy for a pattern mining algorithm can be defined as follows.

Definition 2. A pattern mining algorithm M provides $\epsilon-$differential privacy if for any two datasets D_1 and D_2 that differ in a single entry and for any a,

$$\left|\log \frac{\mu(M(D_1) = a|D_1)}{\mu(M(D_2) = a|D_2)}\right| \leq \epsilon \tag{4}$$

such that $M(D)$ is the random variable that represents the algorithm output and μ denotes the probability density for the algorithm output. Inequality (4) is also called the $\epsilon-$ indistinguishable and ϵ is called the *leakage* [6]. Inequality (4) implies that for the pattern mining algorithm outputs to be indistinguishable probabilistically when there is only one different entry in the dataset, ϵ has to be small, since $\log(1+\epsilon) \approx \epsilon$ for small ϵ. In other words, a high degree of differential privacy is quantified by a small leakage.

$\epsilon-$differential privacy can be achieved by the addition of random noise whose magnitude is chosen as a function on the largest change a single participant could have on the output to the query function, called the *sensitivity* of the function.

Definition 3. For $f : D \to R^d$, the sensitivity of f is

$$\triangle f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1 \tag{5}$$

for all D_1, D_2 differing in at most one element.

The most common mechanism to handle differential privacy is by Laplace noise perturbation on the outputs [7] described by Theorem 4 below.

Theorem 4. *Let K_f be the privacy mechanism for a query function $f : D \to R^d$ and adds noise with a scaled symmetric exponential distribution with variance σ^2 described by the density function*

$$Pr[K_f(X) = a] \propto \exp\left(-\frac{||f(X) - a||_1}{\sigma}\right)$$

to each component of $f(X)$ with $X \in D$. This density function defines the Laplace distribution, $Lap(\sigma)$. The mechanism K_f gives $\left(\frac{\triangle f}{\sigma}\right)$-differential privacy.

The proof of Theorem 4 shows that the application of Laplace noise satisfies Inequality (4) [6]. To achieve $\epsilon-$differential privacy, one perturbs the query output so that

$$f'(X) = f(X) + N \tag{6}$$

where $N \sim Lap(\sigma)$ with $\sigma = \frac{\triangle f}{\epsilon}$. The sensitivity $\triangle f$ used to specify the σ parameter for the Laplace distribution is a "global sensitivity" [24] as the sensitivity depends on all possible pairs of D_1 and D_2 with one element difference. Since $\epsilon = \frac{\triangle f}{\sigma}$, it is clear that in practice one needs to sacrifice differential privacy (or indistinguishability) by allowing higher leakage if the problem domain has a large sensitivity instead of using a higher σ, affecting the output accuracy, to maintain a smaller leakage.

One important result for the differential privacy concept is that the application of a sequence of differential privacy mechanisms ensures differential privacy for a task, even when a later computation makes use of results from earlier computations [21].

Theorem 5 (Sequential Composition)**.** *Let M_i be a mechanism that provides ϵ_i- differential privacy for $i = 1, \ldots, n$. The sequence of M_i on an input domain X provides $(\sum_i \epsilon_i)-$differential privacy.*

Another important result is that if a sequence of differential privacy mechanisms is applied to a set of disjoint subsets of the input domain, one achieves a better overall privacy guarantee which is the worst privacy guarantee among all the disjoint subsets.

Theorem 6 (Parallel Composition)**.** *[21] Let M_i be a mechanism that provides ϵ_i-differential privacy for $i = 1, \ldots, n$ on arbitrary disjoint subsets D_i of the input domain D. The sequence of M_i on the input domain D provides $(\max_i \epsilon_i)$-differential privacy.*

Due to this database independent nature of global sensitivity, the privacy mechanism can perform very badly in practice [26]. In particular, if the noise perturbation is too much due to large sensitivity, the output responses from the algorithm becomes meaningless. While Sarathy and Muralidhar [26] presented a negative argument challenging the usefulness of the Laplace noise perturbation mechanism for practical applications, we claim that the differential privacy mechanism is useful if one can decompose the problem domain into smaller subproblems and one utilizes database instance dependent sensitivity for query function f. This can be achieved based on the utilization of a "relaxed" definition for differential privacy defined as follows.

Definition 7. [8] A randomized function K is (ϵ, δ)-differentially private if for all datasets D_1 and D_2 differing on at most one element, and all $S \subset Range(K)$

$$Pr[K(D_1) \in S] \leq \exp(\epsilon) \times Pr[K(D_2) \in S] + \delta.$$

This is the differential privacy definition we use for the rest of the paper. If $\delta = 0$, $(\epsilon, 0)$-differential privacy is ϵ-differential privacy. One notes that this definition has no relation to the (ϵ, δ)-probabilistic differential privacy [20] such that δ is used to ensure that the disclosure set has low probability.

(ϵ, δ)-differential privacy allows one to claim the same privacy level as Definition 1 when there is a small amount of privacy loss due to a slight variation in the output distribution for the privacy mechanism K. Moreover, one can use a sample-aggregate framework [24] to construct the privacy-preserving pattern mining algorithm that preserves (ϵ, δ)-differential privacy.

Note that the composition theorems above are also valid for (ϵ, δ)-differential privacy mechanism. They are used in the justification of our (ϵ, δ)-differentially private interesting location pattern mining approach.

4 (ϵ, δ)-Differentially Private Interesting Location Discovery

In Section 4.1, we provide intuition and motivation for our (ϵ, δ)-differentially private interesting location pattern mining algorithm. In Section 4.2, we present and describe the differentially private region quadtree algorithm and interesting location extraction algorithm using DBSCAN clustering in detail. In Section 4.3, we show that our proposed privacy preserving region quadtree algorithm and interesting location extraction algorithm using DBSCAN clustering are (ϵ, δ)-differentially private. Towards that end, privacy is preserved for the proposed interesting location pattern mining approach motivated by the sample-aggregate framework [24].

4.1 Motivation and Intuition

A practical problem for the Laplace noise perturbation privacy mechanism (refer to Equation 6) is the magnitude of the database independent (global) sensitivity, $\triangle f$. For the interesting location discovery task, the criterion to identify an interesting location is the number (or count) of stay points in a specific region. One cannot derive a reasonably useful database independent $\triangle f$ for a stay point database (see Figure 1) for our task. The main reason is that $\triangle f$ for the count record query (f^0 and f^1) *does not equal one* when an individual is added or removed from the database. The individual set to stay point set mapping is a one-to-many relation, i.e., one individual can have more than one stay point. There is no database independent upper bound for $\triangle f$ except the database record size which is practically useless as the noise added to the output of the count query is most likely meaningless. The motivation for our privacy preserving pattern mining framework is a privacy mechanism that one can control and add significantly less noise without revealing information from the database for the interesting location discovery task. The two main ideas we pursued are: (i) privacy mechanism that utilizes database instance dependent local sensitivity, and (ii) spatial decomposition to partition the data domain into disjoint subsets to lower the local sensitivity for f^0, f^1, and f^2 (see (1), (2), and (3), respectively) and to satisfy parallel composition condition (see Theorem 6). In particular, our approach is motivated by the sample-aggregate framework [24] that consists of random partitioning of a database into smaller databases and obtaining a query output by combining query outputs from the smaller databases.

Our (ϵ, δ)-differentially private pattern mining approach consists of two main steps: spatial decomposition and clustering (see Figure 1). For the interesting location discovery task, we specifically use region quadtree [25] together with DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm [10]. Both steps, with the inclusion of (ϵ, δ)-differential privacy mechanisms, will be elaborated in Subsection 4.2. Zheng et al. [32] used a tree-based hierarchical graph (TBHG) to construct a data structure by clustering the stay points at different tree height and connecting the clusters with directed edges, defining the chronological order of the stay points. Here, we only focus on the spatial context of the data and hence we use a standard spatial decomposition such as a quadtree, with DBSCAN clustering to achieve both tree-based density-based spatial objects clustering and to demonstrate the differential privacy mechanism on a tree-based location mining approach. One notes that our proposed privacy preserving pattern mining approach can be directly applied to the tree-based TBHG with privacy guarantee. In particular, one can use any spatial decomposition or clustering approach and the results in Subsection 4.3 still hold true.

For a region quadtree, one needs to specify a criterion to halt the spatial decomposition

based on the number of stay points in a quadrant. In other words, if the number of stay points in a quadrant is less than (say) H, the space will stop splitting. Hence, by performing the spatial decomposition, one bounds the sensitivity $\triangle f$ using H that affects the Laplace noise perturbation as no individual in a quadrant can have a stay point count larger than H.

In fact, the threshold H for the spatial decomposition step is a database independent global sensitivity for f^0 and f^1 (see Lemma 12 in Subsection 4.3). The only problem is that H can be chosen to be so large (i.e., an interesting location has to have a reasonable large number of stay points) that the Laplace noise perturbation affects the utility of the output results. On the other hand, if we choose a relatively small H (say, less than 100 for a database size in the order of 10^7), we will need to perform a merge operation for interesting locations in adjacent regions based on the stay point counts. We will need an additional interesting locations threshold H' ($> H$) for the merging operation and additional privacy preserving steps.

As the focus of the paper is on the general differential privacy framework for the interesting location discovery task and how the sensitivity for query functions can be controlled so that a privacy preserving location pattern mining algorithm returns reasonable outputs in practice, we use a reasonably large H in our experiments and do not consider the merging step. As described earlier, spatial decomposition partitions the spatial domain into smaller disjoint sets, i.e., one handles many independent interesting location discovery tasks for smaller spatial regions. Moreover, the database dependent sensitivities for query function f^0, f^1, and f^2 are lower for each smaller disjoint set as the number of stay points for each individual is now smaller.

4.2 Algorithms

Differentially Private Region QuadTree-based Spatial Decomposition

There are two types of quadtree [25] (and the references therein): point quadtree and trie-based quadtree. We utilize the second type which is a decomposition on the space that the input data is drawn. In particular, we use the region quadtree representing the 2-D space by partitioning it into four quadrants. The space decomposition continues till a predefined tree height is reached. Or, spatial decomposition stops for a quadrant when the point counts in that quadrant is less than a global threshold.

A region quadtree spatial decomposition is used to partition the geographic domain for the interesting location discovery task to ensure a reasonable sensitivity for the count queries f^0 and f^1 so that one achieves both the differential privacy goal and accurate output results in practice. Other spatial decomposition can also be used to partition the data domain. Note that we perform the tree construction when the complete input dataset becomes available. Hence, we do not require data insertion into and deletion from the quadtree.

The input variables for Algorithm 1 (BuildDPQuadTree) are: (i) the stay point database S consisting of the stay points and their corresponding user IDs extracted from the trajectory database (see Figure 1), (ii) the spatial extent of the domain, and (iii) a threshold value T. In Line 1, **NoisyCount** takes in the set of stay points records for a partition and returns stay point count S' perturbed by the Laplace noise. **NoisyCount** corresponds to the query function f^0 (Equation 3) with the (ϵ, δ)-differential privacy mechanism (see Lemma 11 in Subsection 4.3) applied to it. The perturbed count S' decides (Line 2–3) whether to halt without returning stay points when there are too few points in the spatial partition (Line 14–15), to continue splitting the space when there are more than H points (Line 7–11), or

Input: S, stay point database; R, spatial region; T, threshold
Output: P, set of spatial partitions; Sp, set of sets of stay points (with User IDs) for corresponding partitions in P
Global variables: $P = \{\}, Sp = \{\}, H = 3T$;
Procedure BuildDPQuadTree(S, R, T)

```
1:  S' = NoisyCount_{ε,δ}(S) ;
2:  if S' > T then
3:    if S' ≤ H then
4:      P = P ∪ {R}; Sp = Sp ∪ {S};
5:      return
6:    else
7:      Split spatial region R into 4 equal
        quadrants: R_nw, R_ne, R_sw, R_se with
        corresponding point sets S_nw, S_ne, S_sw,
        S_se;
8:      BuildDPQuadTree(S_nw, R_nw, T) ;
9:      BuildDPQuadTree(S_ne, R_ne, T) ;
10:     BuildDPQuadTree(S_sw, R_sw, T) ;
11:     BuildDPQuadTree(S_se, R_se, T) ;
12:   end if
13: else
14:   P = P; Sp = Sp;
15:   return
16: end if
```

Algorithm 1: Differentially private region quadtree-based spatial decomposition (BuildDPQuadTree).

return the stay points of the quadrant when there are between T and H points (Line 4–5). The outputs are the spatial partitions and their corresponding database subsets consisting of stay points with their user IDs.

We set the upper bound for perturbed stay point counts in a returned partition to be $H = 3T$. A quadrant partition stops splitting when its perturbed stay point count is less than T (Line 13-15). If the perturbed stay point count for a partition is more than T and less than H (Line 3), a quadrant partition and its perturbed stay point count is returned. From Algorithm 1, one sees that the region quadtree is constructed recursively. Figure 4 shows the partitions containing the interesting locations for a particular synthetic dataset using the BuildDPQuadTree algorithm.

One notes the importance of storing not only the stay points but also their corresponding user IDs which are needed to compute the local sensitivity (see Definition 8 and Lemma 13 in Subsection 4.3) used by **NoisyCount**. Moreover, one notes that our approach is similar to a recent work on differentially private spatial decomposition [5].

Differentially Private Interesting Location Discovery using DBSCAN

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm [10] is used in our interesting location discovery task. Since location pattern mining works on a 2-D space, the Euclidean distance is used. The DBSCAN algorithm does

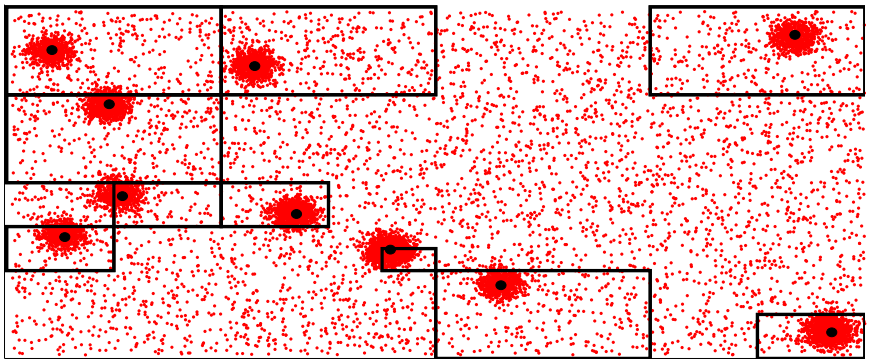

Figure 4: Quadtree partitions for regions where the interesting locations (ten data blobs) are likely to be based on the BuildDPQuadTree algorithm.

not require the number of clusters to be specified and there are only two parameters, γ and $MinPts$. Also, the algorithm can find arbitrary shaped clusters and points that are not in the clusters are labeled as noise.

The input variables for Algorithm 2 (DP-ILD) are (i) the set of the stay point/userID data subsets from Algorithm 1, (ii) threshold r', and (iii) DBSCAN parameters $(MinPts, \gamma)$. For each stay point/userID subset, the DBSCAN density-based clustering algorithm [10] is used to extract the likely interesting regions (Line 4). Similar to Algorithm 1, **NoisyCount** takes in the set of stay point/userID records representing a cluster and returns the perturbed stay point count Ct' (Line 8) and it corresponds to the query function f^1 (Equation 1) with the (ϵ, δ)-differential privacy mechanism (Lemma 11 in Subsection 4.3) applied to it. If the perturbed count Cts' is greater than r', then the region R_{jl} is labeled as an "interesting location" with Cts' stay points. **NoisyCentroid** takes in the set of stay point/userID records representing the cluster and returns the perturbed centroid Cg' (Line 10) and it corresponds to the query function f^2 (Equation 2) with the (ϵ, δ)-differential privacy mechanism (Lemma 11 in Subsection 4.3) applied to it. The local sensitivities used for **NoisyCount** and **NoisyCentroid** are found in Lemma 13 and Lemma 15, respectively. The perturbed centroid Cg' represents an interesting region R_{jl} defined by its convex hull. The outputs for Algorithm 2 are the privacy preserved interesting locations and their corresponding stay point counts. In Figure 5, the left and right diagrams show the original interesting locations and the outputs from DP-ILD, respectively, given inputs shown in Figure 4 .

Since the local sensitivities for f^1 and f^2 require the knowledge of the set of likely interesting locations, we have to completely identify these locations (Line 2–5) before applying the privacy mechanism.

4.3 Theoretical Justifications

Definition 8. [24] For $f : D \to R^d$ and $x \in D$, the *local sensitivity* of f at $x \in D$ is

$$\triangle f_{LS}(x) = \max_{y \in D'} ||f(x) - f(y)||_1 \tag{7}$$

Input: $Sp = \{S_1, \ldots, S_k\}$, set of stay point/user ID sets for corresponding partitions in P; threshold r'; DBSCAN parameters: $MinPts$, γ.
Output: I, set of interesting locations (region centroids); C, set of corresponding counts for I.
Procedure DP-ILD(Sp, r', $MinPts$, γ)
1: $I = \{\}; C = \{\}; Cts' = 0; Cg' = \mathbf{0}$;
2: **for** $i = 1$ to k **do**
3: $//CL_i = \{R_{i1}, \ldots, R_{il_i}\}$ such that R_{ij} is the set of points of the form $\{(x_m, y_m), userID\}$ in cluster $j, j = 1, \cdots, l_i$.
4: $CL_i = \mathbf{DBSCAN}(S_i, Minpts, \gamma)$;
5: **end for**
6: **for** $j = 1$ to k **do**
7: **for** $l = 1$ to $|CL_j|$ **do**
8: Count $Cts' = \mathbf{NoisyCount}_{\epsilon,\delta}(R_{jl})$;
9: **if** $Cts' > r'$ **then**
10: $Cg' = \mathbf{NoisyCentroid}_{\epsilon,\delta}(R_{jl})$
11: $I = I \cup \{Cg'\}$;
12: $C = C \cup \{Cts'\}$;
13: **end if**
14: $Cts' = 0; Cg' = \mathbf{0}$;
15: **end for**
16: **end for**

Algorithm 2: Differentially private interesting location discovery (DP-ILD).

for all $y \in D'$ differing in at most one element from $x \in D$.

According to [24], the noise function has to be insensitive in the database space. In other words, a small difference between two databases cannot induce a spike or sharp dip in the noise added. [24] introduced the concepts of $\beta-$*smooth sensitivity* and $\beta-$*smooth upper bound* for local sensitivity.

Definition 9. [24] For $\beta > 0$, the β-smooth sensitivity of f for $x \in D$ is

$$\triangle f_\beta(x) = \max_{y \in D}(\triangle f_{LS}(y) \cdot e^{-\beta d(x,y)}). \tag{8}$$

such that $d(x, y)$ is the number of elements that x and y differs.

One notes that it is difficult to compute $\triangle f_\beta(x)$ as one has to find local sensitivity for all databases in D.

Definition 10. [24] For $\beta > 0$, a function $S : D \to \mathbf{R}^+$ is a β-smooth upper bound on $\triangle f_{LS}$, local sensitivity of f, if it satisfies the following requirements:

$$\begin{aligned} \forall x \in D: &\quad S(x) \geq \triangle f_{LS}(x) \\ \forall x, y \in D: &\quad S(x) \leq e^\beta S(y) \end{aligned}$$

such that x and y differing on one element.

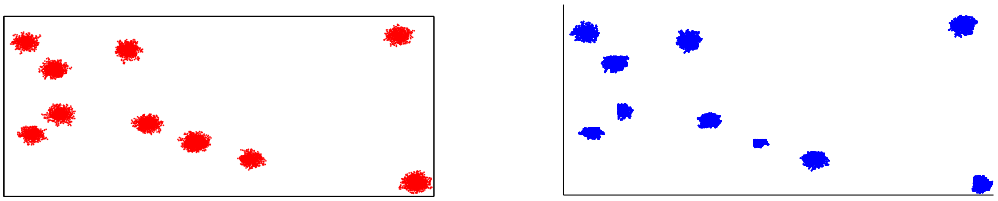

Figure 5: Interesting location outputs from DP-ILD (right) and outputs when no privacy preserving mechanism is used (left).

The noise added to the output is weighted by the smooth upper bound function S on the local sensitivity for all databases in D such that $\ln S(\cdot)$ has a global sensitivity bounded by β. Note that $\triangle f_\beta$ is a $\beta-$smooth upper bound on $\triangle f_{LS}$ and the global sensitivity $\triangle f$ is, in general, a relax/conservative $\beta-$smooth upper bound. Moreover, the noise distribution $\mathcal{N}$ on R^d used (e.g., Laplace distribution) should not change much under translation (T) and scaling (S) [24], i.e., when $N_1, N_2 \sim \mathcal{N}$,

T: $P(N_1 \in \mathcal{U}) \leq e^{\frac{\epsilon}{2}} \cdot P(N_2 \in \mathcal{U} + \mathcal{T}) + \frac{\delta}{2}$

S: $P(N_1 \in \mathcal{U}) \leq e^{\frac{\epsilon}{2}} \cdot P(N_2 \in e^{\lambda} \cdot \mathcal{U}) + \frac{\delta}{2}$

for all $||\mathcal{T}|| \leq \alpha = \alpha(\epsilon, \delta)$ and $|\lambda| \leq \beta = \beta(\epsilon, \delta)$, all subsets $\mathcal{U} \subseteq R^d$, $\epsilon > 0$, and $\delta > 0$. A noise distribution $\mathcal{N}$ is (α, β)-admissible if it satisfies the above two conditions.

Now, we can achieve $(\epsilon, \delta)-$differential privacy using the $(\alpha, \beta)-$admissible Laplace noise as follows.

Lemma 11. *For $f : D \to R^d$ with $\triangle f_\beta$, a β-smooth upper bound on $\triangle f_{LS}$, local sensitivity of f. Given the noise random variable $N \sim Lap(1) = \frac{1}{2} \cdot e^{-|z|}$, $\alpha = \frac{\epsilon}{2}$, and $\beta = \frac{\epsilon}{2} \ln(\frac{1}{\delta})$ with $\delta > 0$, the privacy mechanism $K(x) = f(x) + \frac{S}{\alpha} \cdot N = f(x) + \frac{\triangle f_\beta}{\alpha} \cdot N$ is (ϵ, δ)-differentially private.*

The Lemma follows from Lemma 2.5 in [24] that provides the condition for (ϵ, δ)-indistinguishable for $\triangle f_{LS}$ and local sensitivity of f, together with Example 3 in [24] describing a (α, β)-admissible Laplace distribution with location parameter $\mu = 0$ and scaling parameter $\sigma = 1$ that satisfies Definition 7 with $\delta > 0$.

From Lemma 11, one observes that the differential privacy level ϵ and the local sensitivity $\triangle f_{LS}$ are no longer used to parametrize the Laplace distribution (6). Instead, they are used to weigh the noise generated from a standard Laplace distribution with mean zero and variance one.

Lemma 12. *$\triangle f^0_{GS}$ and $\triangle f^1_{GS}$, the global sensitivity for f^0 and f^1, is the count threshold parameter, $3T$, in the BuildDPQuadTree algorithm.*

$\triangle f^0_{GS} = \max_{D_1, D_2} |f^{(0)}(D_1) - f^{(0)}(D_2)|$ is the count difference in a partition when an arbitrary individual, together with his trajectory and derived stay points, is removed or added to either D_1 or D_2. The worst case situation occurs when there are maximum number of stay points in the partition, i.e. 3T, and they are owned by the removed/added individual. This worst case situation is independent of a database but depends on the BuildDPQuadTree algorithm stay point count upper bound threshold $3T$. $\triangle f^1_{GS} = 3T$ follows from the earlier argument on the counts in a partition with an addition observation that all the stay points in the partition forms an interesting location in a worst case scenario.

Lemma 13. *$\triangle f^0_{LS} = \max_{u\in U} \max_{p\in P_{u\setminus U}} |o_{pu} - n_{pu}|$ is the local sensitivity for f^0, such that $P_{u\setminus U}$ is the set of partition when an individual u and all its corresponding stay points are removed or added to a database, and o_{pu} and n_{pu} are the number of stay points in such a partition p before and after the removal or addition of stay points for an individual u. Similarly, $\triangle f^1_{LS} = \max_{u\in U} \max_{i\in I_{u\setminus U}} |o_{iu} - n_{iu}|$ is the local sensitivity for f^1, such that $I_{u\setminus U}$ is the set of interesting location when an individual u and all its corresponding stay points are removed or added to a database, and o_{iu} and n_{iu} are the number of stay points in such an interesting location i before and after the removal or addition of stay points for an individual u.*

Both local sensitivities follow from Definition 8.

The local sensitivity for f^1 defined in Lemma 13 does not disclose the interesting location that has the largest change or the individual who contributes to the local sensitivity as long as there are stay points from more than one individual in all the likely interesting locations. Hence, even if the local sensitivities are known by an adversary, the individual's location privacy is preserved.

Theorem 14. *BuildDPQuadTree spatial decomposition algorithm provides (ϵ, δ)-differential privacy for $\delta \geq 0$.*

Using the β-smooth sensitivity in Definition 9 and the local sensitivity in Lemma 13, we can construct a $\beta-$smooth upper bound for $\triangle f^0_{LS}$. Using this, we can construct a privacy mechanism in Step 1 of Algorithm 1 based on Lemma 11 that is (ϵ, δ)-differentially private. Based on Theorem 5 (Sequential Composition) and 6 (Parallel Composition), BuildDPQuadTree provides $\max_p(\sum^{h_p}(\epsilon, \delta))$-differential privacy at a fixed non-negative δ, where $p \in P$, the set of partitions and h_p is the depth of the tree when the algorithm returns the final count for a disjoint partition p.

Lemma 15. *$\triangle f^2_{LS} = \max_{u\in U} \max_{i\in I_{u\setminus U}} |cg^o_{iu} - cg^n_{iu}|$ is the local sensitivity for f^2, such that $I_{u\setminus U}$ is the set of interesting locations when an individual u and all its corresponding stay points are removed or added to a database, and cg^o_{iu} and cg^n_{iu} are the centroids for an interesting location i before and after the removal or addition of an individual u.*

$\triangle f^2_{LS}$ follows from Definition 8. One notes that a meaningful global sensitivity for f^2 without knowing the database before hand is not possible. To reduce computational cost, we use the (somehow weak) upper bound

$$\max_{i\in I} \left(\frac{\text{longest distance between 2 points in } i}{2} \right) \tag{9}$$

such that I is the set of interesting locations, to compute $\triangle f^2_{LS}$.

Theorem 16. *DP-ILD interesting location discovery algorithm provides (ϵ, δ)-differential privacy for $\delta > 0$.*

Using the β-smooth sensitivity in Definition 9 and the local sensitivity in Lemma 13 and Lemma 15, we can construct a $\beta-$smooth upper bound for $\triangle f^1_{LS}$ and $\triangle f^2_{LS}$. Using the two upper bounds, we can construct privacy mechanisms in Step 8 and 10 of Algorithm 2 based on Lemma 11 that are (ϵ_{f^1}, δ)-differentially private and (ϵ_{f^2}, δ)-differentially private for some fixed positive δ. Since a centroid computation f^2 (step 10 in Algorithm 2) follows from f^1 and also considering the $\epsilon_{f^0}-$differentially private spatial decomposition at the same δ, f^2 is $(\epsilon_{f^0} + \epsilon_{f^1} + \epsilon_{f^2}, \delta)$-differentially private based on Theorem 5 (Sequential Composition). Since f^1 (step 8 in Algorithm 2) and f^2 are applied to all independent likely

interesting locations in all the partitions, DP-ILD is $(\epsilon_{f^0} + \epsilon_{f^1} + \epsilon_{f^2}, \delta)$-differentially private based on Theorem 6 (Parallel Composition).

One notes from Theorem 14 that since global sensitivity is available for f^0, δ can equal zero. On the other hand, since global sensitivity is not possible for f^2, δ is strictly greater than zero in Theorem 16.

5 Experimental Results

Our preliminary results (e.g. pattern mining performance and effect of privacy budget allocation between spatial decomposition and clustering) [16] performed on synthetic datasets using an approximated version of our differentially private interesting location pattern mining algorithm demonstrated the feasibility of using local sensitivity on a sample-aggregate approach for the interesting location pattern mining task.

In this section, we present additional experimental results on the Geolife dataset [30, 31] to study the characteristics of our proposed algorithm. In Section 5.1, we describe the performance measures used in our study. In Section 5.2, we briefly describe the Geolife dataset. In Section 5.3, comparison results are presented and discussed.

5.1 Performance Measures

Let $I_0 = \{I_1, I_2, \ldots, I_m\}$ and $I_{(\epsilon,\delta)} = \{I_{1\epsilon}, I_{2\epsilon}, \ldots, I_{n\epsilon}\}$ be the two sets of regions [partitions in spatial decomposition or interesting locations from DBSCAN clustering] without and with the application of any privacy preserving mechanism, respectively. Also, m may not equal n. Let $\mathcal{I}$ be the set of regions that are similar in I_0 and $I_{(\epsilon,\delta)}$. Here, two regions are similar if their intersection is non-empty. Let $C_0 = \{c_1, c_2, \ldots, c_{|\mathcal{I}|}\}$ and $C_{(\epsilon,\delta)} = \{c_{1\epsilon}, c_{2\epsilon}, \ldots, c_{|\mathcal{I}|\epsilon}\}$ be the two sets containing corresponding stay point counts for regions in $\mathcal{I}$ without and with the application of the privacy preserving mechanism, respectively. Let $Cg_0 = \{cg_1, cg_2, \ldots, cg_{|\mathcal{I}|}\}$ and $Cg_{(\epsilon,\delta)} = \{cg_{1\epsilon}, cg_{2\epsilon}, \ldots, cg_{|\mathcal{I}|\epsilon}\}$ be the two sets containing corresponding interesting location centroids in $\mathcal{I}$ without and with the application of the privacy preserving mechanism, respectively. The evaluation criteria for output accuracy are defined as follows.

$$\text{Mean Count Deviation (MCtD)} = \frac{\sum_{x\in\mathcal{I}} |c_i - c_{i\epsilon}|}{|\mathcal{I}|}$$

$$\text{Mean Centroid Deviation (MCgD)} = \frac{\sum_{x\in\mathcal{I}} ||cg_i - cg_{i\epsilon}||}{|\mathcal{I}|}$$

MCtD and MCgD measure the effects of the privacy mechanism on the number of stay points and the interesting location centroid, respectively, compared to the baseline when no privacy mechanism is applied.

5.2 Geolife GPS Trajectory Data

The real world dataset used in our experiment is the GPS trajectory dataset [30, 31], published in September 2010, consists of data from 165 users for a period of over two years (from April 2007 to August 2009). The dataset recorded a broad range of users' outdoor movements, including not only routines like going home and going to work, but also some entertainments and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. We note that 70502 stay points extracted from the Geolife trajectory dataset based on the stay point definition (see Section 3.1).

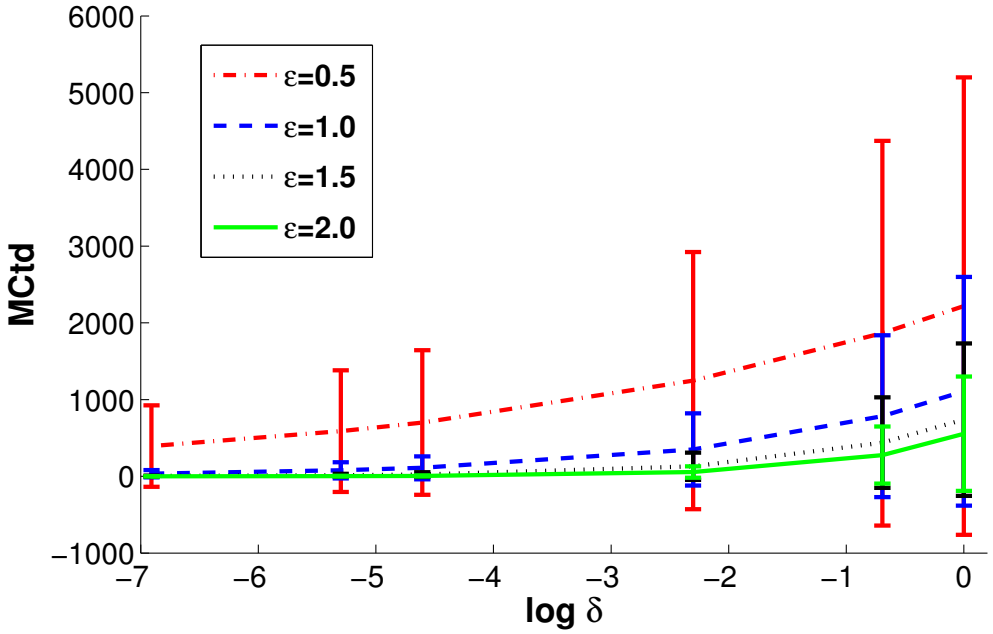

Figure 6: The effect of δ on the (ϵ, δ)-differentially private count query f^0 used in BuildDPQuadTree algorithm for various ϵ.

5.3 Results and Discussions

We use the BuildDPQuadTree (i) to understand the effect of ϵ and δ on the output count accuracy and (ii) to compare the ϵ-differential privacy Laplace noise mechanism (6) and the (ϵ, δ)-differential privacy weighted Laplace noise mechanism (Lemma 11).

The threshold value T for BuidDPQuadTree is set to 500 and hence the upper bound on the number of stay points in a partition is $H = 3T$. We use $\epsilon = 0.1, 0.5, 1.0, 1.5$, and 2.0 and vary δ from 0.001 to 1. Figure 6 shows that smaller δ results in smaller deviation (i.e., smaller $MCtd$) in count output from f^0 for all ϵ values. In Figure 6, the curve for $\epsilon = 0.1$ is not shown as it has a much higher $MCtd$ and variance even for a very small δ. This shows that it is extremely difficult to achieve low differential privacy level even when δ is very small for the count query f^0.

Next, we compare the effect of the output query of f^0 when (ϵ, δ)-differentially private weighted Laplace noise mechanism (Lemma 11) and the ϵ-differential private Laplace noise mechanism (Theorem 4) on BuildDPQuadTree algorithm. We set δ to 0.001. Moreover, we include a pseudo ϵ-differential private Laplace noise mechanism that uses local sensitivity (Definition 8) instead of global sensitivity (Definition 5). Figure 7 shows that (ϵ, δ)-differential privacy mechanism outperforms the ϵ-differential privacy in terms of the mean output count deviation for various ϵ. The inserted graph in Figure 7 magnifies the performance curves for the three mechanisms at around $\epsilon = 1$ to demonstrate the low mean deviation of output count and variance for the (ϵ, δ)-differential privacy mechanism compare to the other two mechanisms when ϵ is greater than 0.5. One observes that even though the pseudo ϵ-differential privacy mechanism that used local sensitivity performs best at

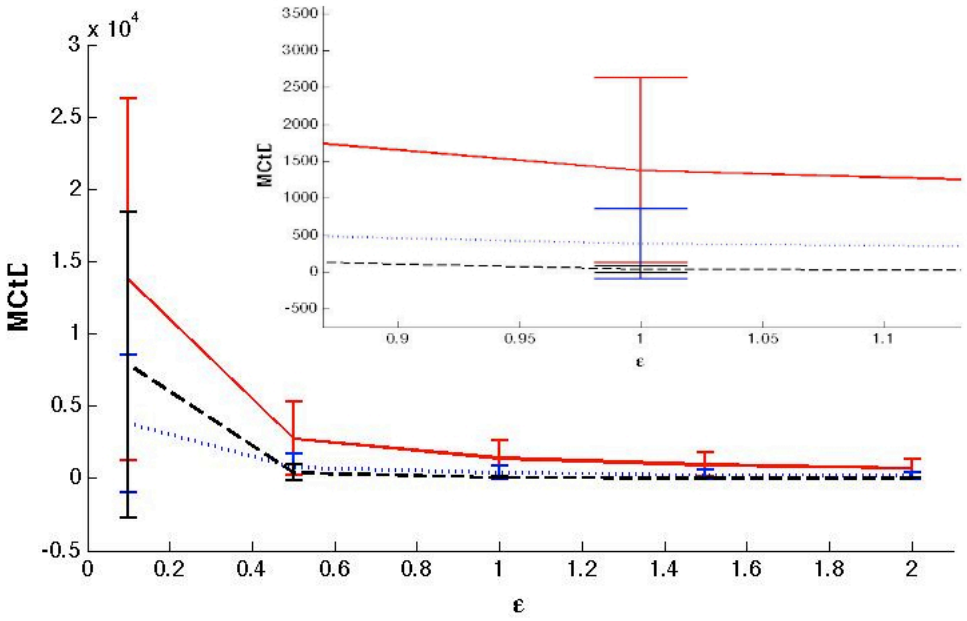

Figure 7: Comparison of the three differential privacy (DP) mechanisms: ϵ-DP (solid line), pseudo ϵ-DP (dotted line), $(\epsilon, \delta = 0.001)$-DP (dashed line) on the count query f^0 for the BuildDPQuadTree algorithm.

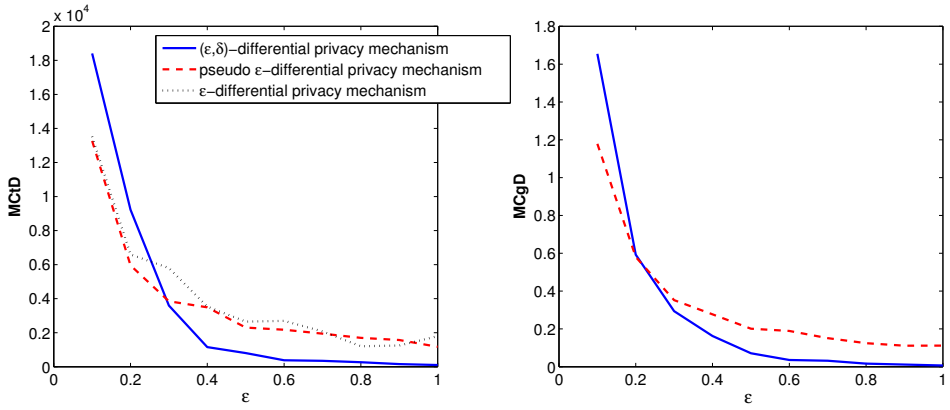

Figure 8: Effect of applying the three differential privacy mechanisms on f^1 and f^2 on the MCtD and MCgD of the DP-ILD algorithm.

$\epsilon = 0.1$, there is no theoretical justification that it preserves differential privacy at ϵ privacy level.

From Figure 8, we observe that the DP-ILD algorithm using the (ϵ, δ)-differential privacy mechanism returns more accurate count and centroid output when ϵ is greater than 0.3. One observes that there is no curve for the application of ϵ-differential mechanism on the right graph as there is no non-trivial global sensitivity for the centroid query function f^2.

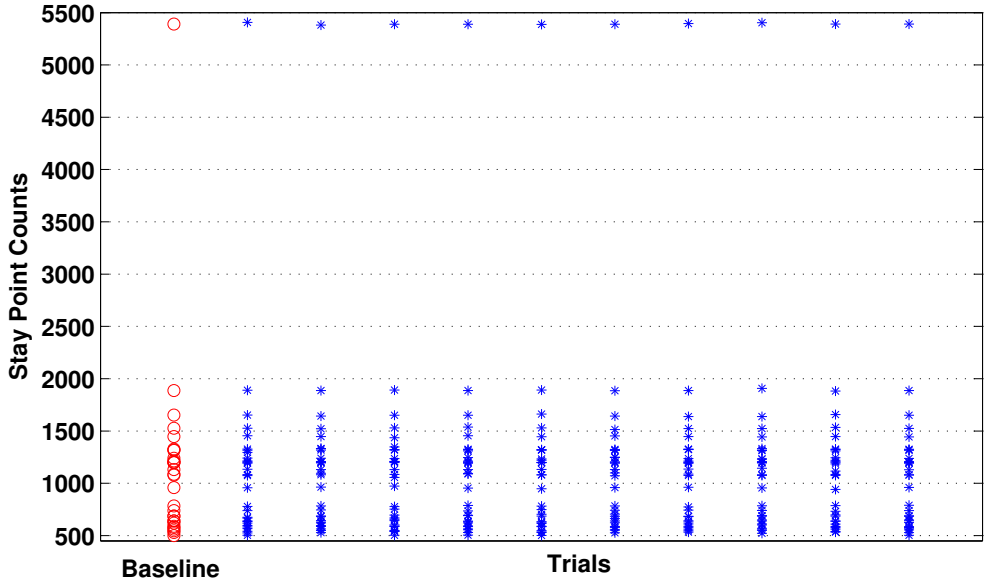

Figure 9: Comparing the baseline (no privacy preserving mechanism used) stay point counts and the stay point counts from the application of the differentially private interesting location pattern mining approach on thirty interesting locations obtained from the Geolife dataset. The experiment was repeated ten times.

Using $\epsilon = 1$ for **NoisyCount** in BuildDPTree (Algorithm 1) and $\epsilon = 2$ for both **NoisyCount** and **NoisyCentroid** in DP-ILD (Algorithm 2) and fixed $\delta = 0.001$, $MinPts = 60$, $\gamma = 0.001$, and $r' = 500$, we performed ten trials on the whole Geolife dataset. There are a total of 30 interesting locations based on our chosen parameters. In Figure 9, we see that the stay point counts for the interesting locations are very similar (but not identical) for each of the ten trials as compared to the baseline using the privacy level that we defined. One notes that to achieve practical privacy preserving objective and utility, one needs to choose the privacy level (ϵ) and mining algorithm parameters (e.g. threshold r' and DBSCAN parameters) based on practical considerations such as the level of tolerance to output accuracy .

6 Conclusions

In this paper, we investigate how one achieves the privacy goal that the inclusion of his location history in a statistical database with interesting location mining capability does not substantially increase risk to his privacy. In particular, we propose a (ϵ, δ)-differentially private approach motivated by the sample-aggregate framework for pattern mining interesting geographic locations using spatial decomposition to preprocess the location points followed by density-based clustering. Preliminary study [16] on an approximated version of the algorithm has demonstrated its feasibility. Here, we further justify our approach and provide additional supportive empirical results using the Geolife dataset.

7 Acknowledgments

This research work was partially carried out at the University of Maryland Institute for Advanced Computer Studies (UMIACS). The second author is supported by grants from Sichuan University, China when she visited the University of Maryland, College Park, in 2011.

References

[1] Osman Abul, Francesco Bonchi, Mirco Nanni (2010), Anonymization of moving objects databases by clustering and perturbation, *Information Systems*, Vol. 35, No. 8, 884-910.

[2] Rakesh Agrawal and Ramakrishnan Srikant, Privacy-preserving data mining (2000), *SIGMOD Rec.*, Vol. 29, No. 2, 439–450.

[3] F. Bonchi, Y. Saygin, V. S. Verykios, M. Atzori, A. Gkoulalas-Divanis, S. V. Kaya, and E. Savas (2008), Privacy in Spatiotemporal Data Mining, in *Mobility, Data Mining and Privacy*, Springer, 297–334.

[4] Francesco Bonchi and Elena Ferrari (2011), Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques, CRC Press.

[5] Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, Entong Shen, Ting Yu (2012), Differentially Private Spatial Decompositions. ICDE, 20–31.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (2006), Calibrating Noise to Sensitivity in Private Data Analysis, *Third Theory of Cryptography Conference*.

[7] Cynthia Dwork (2006), Differential Privacy, *ICALP*, 1-12.

[8] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil P. Vadhan (2009), On the complexity of differentially private data release: efficient algorithms and hardness results. *STOC*, 381–390.

[9] Cynthia Dwork (2011), A Firm Foundation for Private Data Analysis, *Communications of the ACM*.

[10] Martin Ester, Hans-Peter Kriegel, Jrg Sander and Xiaowei Xu (1996), A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *KDD*, 226-231.

[11] Arik Friedman and Assaf Schuster (2010), Data mining with differential privacy, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 493–502.

[12] Bugra Gedik and Ling Liu (2005), Location Privacy in Mobile Systems: A Personalized Anonymization Model, *ICDCS*,620-629.

[13] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi (2007), Trajectory pattern mining, *KDD*, 330–339.

[14] G. Gidofalvi, X. Huang, and T. Bach Pedersen (2011), Probabilistic Grid-Based Approaches for Privacy Preserving Data Mining on Moving Object Trajectories, In *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*, CRC Press, 183– 210.

[15] Marios Hadjieleftheriou, George Kollios, Dimitrios Gunopulos, and Vassilis J. Tsotras (2003), On-Line Discovery of Dense Areas in Spatio-temporal Databases, *SSTD*, 306-324.

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

[17] Shen-Shyang Ho (2012), Preserving privacy for moving objects data mining, IEEE International Conference on Intelligence and Security (ISI), 135-137.

[18] Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias (2007), Preventing Location-Based Identity Inference in Anonymous Spatial Queries, *IEEE Trans. Knowl. Data Eng*, Vol. 19, No. 12, 1719–1733.

[19] Yehuda Lindell, Benny Pinkas (2000), Privacy Preserving Data Mining, *CRYPTO 2000*, 36–54.

[20] Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, Lars Vilhuber: Privacy: Theory meets Practice on the Map. ICDE 2008: 277-286

[21] Frank McSherry (2010), Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Comm. ACM*, Vol. 53, No. 9, 89–97.

[22] Anna Monreale, Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel (2010), Movement Data Anonymity through Generalization, *Trans. Data Privacy*, vol. 3, no. 2, 91–121.

[23] Mehmet Ercan Nergiz, Maurizio Atzori, Yucel Saygin, and Baris Guc (2009), Towards Trajectory Anonymization: a Generalization-Based Approach, *Transactions on Data Privacy*, Vol. 2, No. 1, 47–75.

[24] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2007), Smooth sensitivity and sampling in private data analysis, *STOC*, 75–84.

[25] Hanan Samet (2006), Foundations of Multidimensional and Metric Data Structures, Morgan Kaufmann.

[26] Rathindra Sarathy and Krishnamurty Muralidhar (2011), Evaluating Laplace noise addition to satisfy differential privacy for numeric data, *Trans. Data Privacy*, vol. 4, 1-17.

[27] Latanya Sweeney (2002), k-Anonymity: A model for protecting privacy, *International Journal on Uncertainty, Fuzziness, and Knowledge-based*, Vol. 10, No. 5, 557–570.

[28] V. S. Verykios, M. L. Damiani, and A. Gkoulalas-Divanis (2008), Privacy and Security in Spatiotemporal Data and Trajectories, in *Mobility, Data Mining and Privacy*, Springer, 213–240.

[29] Roman Yarovoy, Francesco Bonchi, Laks V. S. Lakshmanan, and Wendy Hui Wang (2009), Anonymizing moving objects: how to hide a MOB in a crowd?, *EDBT*, 72-83.

[30] Yu Zheng, Quannan Li, Yukun Chen, and Xing Xie (2008), Understanding Mobility Based on GPS Data. *Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008)*, 312–321.

[31] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma (2009), Mining interesting locations and travel sequences from GPS trajectories. *Proc. International conference on World Wild Web*, 791-800.

[32] Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma (2011), Recommending Friends and Locations based on Individual Location History, *ACM Trans. on the Web*, vol. 5, no. 1, 5.