Distributed and Parallel Databases (2021) 39:785–811
https://doi.org/10.1007/s10619-020-07318-7

Sensitive attribute privacy preservation of trajectory data publishing based on l-diversity

Lin Yao[1,2] · Zhenyu Chen[3] · Haibo Hu[4] · Guowei Wu[5] · Bin Wu[6]

Accepted: 5 November 2020 / Published online: 17 November 2020

Abstract
The widely application of positioning technology has made collecting the movement of people feasible for knowledge-based decision. Data in its original form often contain sensitive attributes and publishing such data will leak individuals' privacy. Especially, a privacy threat occurs when an attacker can link a record to a specific individual based on some known partial information. Therefore, maintaining privacy in the published data is a critical problem. To prevent record linkage, attribute linkage, and similarity attacks based on the background knowledge of trajectory data, we propose a data privacy preservation with enhanced *l*-diversity. First, we determine those critical spatial-temporal sequences which are more likely to cause privacy leakage. Then, we perturb these sequences by adding or deleting some spatial-temporal points while ensuring the published data satisfy our (L, α, β)-privacy, an enhanced privacy model from *l*-diversity. Our experiments on both synthetic and real-life datasets suggest that our proposed scheme can achieve better privacy while still ensuring high utility, compared with existing privacy preservation schemes on trajectory.

Keywords Sensitive attribute · Privacy preservation · Trajectory data publishing

1 Introduction

The popularity of smart mobile devices with positioning technologies triggers the collection of location information by suppliers, corporations, individuals etc. for knowledge-based decision making. Therefore, vast amounts of trajectory data are collected with other information. Data miners have also shown great interest in analyzing these data to provide plentiful serves for people. For example, recent studies [1, 2] have shown that tracking the environmental exposure of a person with his daily trajectories helps to improve diagnose. Therefore, wearable devices

✉ Guowei Wu
wgwdut@dlut.edu.cn

Extended author information available on the last page of the article

Table 1 Original table

ID.	Name	Trajectory	Disease	···
1	Alice	$a1 \rightarrow d2 \rightarrow b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	HIV	···
2	Bob	$d2 \rightarrow c5 \rightarrow f6 \rightarrow c7 \rightarrow e9$	Flu	···
3	Caesar	$b3 \rightarrow f6 \rightarrow c7 \rightarrow e8$	SARS	···
4	Daniel	$b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	Fever	···
5	Eden	$a1 \rightarrow d2 \rightarrow c5 \rightarrow f6 \rightarrow c7$	Flu	···
6	Freeman	$c5 \rightarrow f6 \rightarrow e9$	SARS	···
7	Georgia	$f6 \rightarrow c7 \rightarrow e8$	Fever	···
8	Hugo	$a1 \rightarrow c2 \rightarrow b3 \rightarrow c7 \rightarrow e9$	SARS	···
9	Ishtar	$e4 \rightarrow f6 \rightarrow e8$	Fever	···

Table 2 Table without explicit identifier

Trajectory	Disease
$a1 \rightarrow d2 \rightarrow b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	HIV
$d2 \rightarrow c5 \rightarrow f6 \rightarrow c7 \rightarrow e9$	Flu
$b3 \rightarrow f6 \rightarrow c7 \rightarrow e8$	SARS
$b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	Fever
$a1 \rightarrow d2 \rightarrow c5 \rightarrow f6 \rightarrow c7$	Flu
$c5 \rightarrow f6 \rightarrow e9$	SARS
$f6 \rightarrow c7 \rightarrow e8$	Fever
$a1 \rightarrow c2 \rightarrow b3 \rightarrow c7 \rightarrow e9$	SARS
$e4 \rightarrow f6 \rightarrow e8$	Fever

have been generating tremendous amounts of location-rich, real-time, and high-frequency sensing data with the physical symptoms for remote monitoring on patients of common chronic diseases including diabetes, asthma, depression [3]. However, the original data may contain sensitive information about individuals such as health status. Let's take Table 1 to illustrate it.

Table 1 [4] is an original table without omitting any attribute. In this table, there are four typical types of attributes: explicit identifier, quasi-identifiers, sensitive attribute, and non-sensitive attribute [5]. *Explicit Identifier* (*EI*), such as the name, is used to identify an individual uniquely, which is always removed from the published table shown in Table 2. On the other hand, single *Quasi-Identifier* (*QI*) cannot uniquely identify a specific individual, but a few *QIs* can be combined to achieve it. In this paper, our focused *QI* is *Trajectory*, which consists of a set of spatial-temporal trajectory points, each with a location and a time stamp. *Sensitive Attribute* (*SA*) contains private information of users, such as *Disease* in Table 1. *Non-sensitive attribute* can be known by the public without any privacy concern.

If the attacker has limited background knowledge of a certain trajectory sequence, the following three attacks are mostly considered in current approaches, record linkage attack, attribute linkage attack and similarity attack [4, 6]:

- *Record linkage attack*. An adversary could identify the unique record of the victim from the published table according to a certain trajectory sequence whose length is no more than m. For example, when an adversary gets the background knowledge of Alice's trajectory sequence $d2 \rightarrow e4$, the adversary can infer that the 1st record belongs to Alice in Table 2. As a result, Alice's record in Table 2 is leaked.
- *Attribute linkage attack*. An adversary could infer the sensitive attribute of the victim from the published table according to a certain trajectory sequence whose length is no more than m. or example, when an adversary gets the background knowledge of Bob's trajectory sequence $c5 \rightarrow c7$, the adversary can infer that Bob's record is either the 2nd or the 5th in Table 2. Because the two records have the same disease Flu, the adversary can infer that Bob has the Flu.
- *Similarity attack*. An adversary could infer the sensitive attribute category of the victim from the published table according to a certain trajectory sequence whose length is no more than m and a sematic dictionary which contains the sematic relevance among sensitive attributes. For example, when an adversary gets the background knowledge of Tom's trajectory sequence $c7$, the adversary can infer that Tom may suffer *Flu*, *Fever*, or *SARS* in Table 2. Based on the sematic dictionary that Flu and SARAS both belong to lung infections, the adversary can learn that the probability of Tom's lung infection is $\frac{4}{5}$.

These attacks generally cause identity disclosure, attribute disclosure, and similarity disclosure [4]. *Identity disclosure* refers to re-identifying a target user from some background knowledge. *Attribute disclosure* occurs when some *QI* values can link to a specific *SA* value with a high probability. *Similarity disclosure* happens when some similar *QI* values can link to a type of *SA* values with a high probability. To prevent the above three kinds of disclosure cause by the background knowledge attack, where an adversary has some prior knowledge (or auxiliary information) about the target of his attack, some anonymization operations should be taken to modify the original table. The typical anonymization approaches in publishing trajectory data include generalization, suppression, and perturbation [4, 6]. Generalization and suppression aim to replace values of specific attributes with less specific values. For trajectory data, generalization and suppression may eliminate a certain number of moving points by replacing some spatial–temporal points with a broader category or wildcard "*". In perturbation, the data will be distorted by adding noise, swapping values, or generating synthetic data. Comparatively, perturbation can protect privacy by distorting the dataset while keeping some statistical properties [6]. Generalization and suppression causes significant loss of data utility.

To protect user privacy while ensuring data utility, we propose an Enhanced l-diversity Data Privacy Preservation for publishing trajectory data (called EDPP). Compared with t-closeness, k-anonymity and l-diversity can resist identity disclosure [7]. Compared with k-anonymity, l-diversity can provide stronger privacy preservation by guaranteeing l different sensitive attributes in a group [8]. To resist attribute disclosure, and similarity disclosure, we propose our (l, α, β)-privacy model, where l-diversity ensures that each trajectory sequence matches more than l types of *SA* values in the published table, α-privacy ensures that the probability of

determining each *SA* value is not greater than α and β-privacy guarantees that the probability that an attacker obtains similar *SA* values is not larger than β. To summarize, this paper has the following contributions:

- We propose our (l, α, β)-privacy model to resist the attacks based on background knowledge including the record linkage, attribute linkage and similarity attacks without changing any sensitive attribute. The three parameters, l, α and β, which are used to prevent identity closure, privacy closure and similarity closure respectively, can be set based on the requirements of data owners.
- We design a novel perturbation approach by executing addition or subtraction operation on the chosen critical sequences based on which the attacker can infer some sensitive information of an individual. Compared with generalization and suppression, perturbation can keep the statistical property of the original trajectory data.
- Privacy analysis prove that our EDPP scheme can meet l, α and β privacy requirements of our model.
- We evaluate the performance through extensive simulations based on a real-world data set. Compared with **PPTD** [4], **KCL-Local** [9] and **DPTD** [10], our EDPP is superior in terms of data utility ratio and privacy.

The remainder of this paper is organized as follows. In Sect. 2, we discuss the related work. Privacy model is given in Sect. 3. In Sect. 4, we present the details of our approach. Privacy analysis is given in Sect. 5. Simulations on data utility are presented in Sect. 6. Finally, we conclude our work in Sect. 7.

2 Related work

Different from those studies which have investigated re-identification attack or semantic attack, re-identifying an individual or inferring semantic information of the victim's visited locations based on the published trajectory dataset, we aim to prevent the attacks based on background knowledge and protect the privacy of an individual's sensitive attribute such as disease linked by the frequently visited locations. In this section, we only discuss works that are related to our approach.

2.1 Generalization and suppression

Generalization replaces some QI values with a broader category such as a parent value in the taxonomy of an attribute. In [4], sensitive attribute generalization and trajectory local suppression were combined to achieve a tailored personalized privacy model for trajectory data publication. In [11], an effective generalization method was proposed to achieve $k^{\tau,\epsilon}$-anonymity in spatiotemporal trajectory data. Combining suppression and generalization, the dynamic trajectory releasing method based on adaptive clustering was designed to achieve k-anonymity in [12]. In [13], a new approach that uses frequent path to construct k-anonymity was proposed. In

the suppression method, a certain number of moving points are eliminated from trajectory data. In [14], extreme-union and symmetric anonymization were proposed to build anonymous groups and avoid a moving object being identified through the correlation between anonymization groups. [9] was the first paper to adopt suppression to prevent record linkage and attribute linkage attacks. To thwart identity record linkage, passenger flow graph was first extracted from the raw trajectory data to satisfy the *LK*-privacy model [15]. In [16], k^m-anonymity was proposed to suppress the critical location points chosen from quasi-identifiers to protect against the record linkage attack. In [17], location suppression and trajectory splitting were used to prevent privacy leaks and improve data utility of aggregate query and frequent sequences.

2.2 Perturbation

Perturbation aims to protect the privacy with limiting the upper bound of utility loss. Recently, differential privacy has become a main form of data perturbation [18]. Differential privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. In [19], differential privacy was first adopted to protect the privacy of trajectory data. Different from the traditional method that privacy was achieved by perturbation of the result [19], sampling and interpolation were combined to achieve differential privacy [20]. Differentially private synthetic trajectory was first proposed in [21]. The original database was built as a prefix tree, where trajectories are grouped based on the length of the matching location subsequences. Then, spatial generalization was combined to protect the trajectory privacy at each tree layer. To solve the problem that frequent sequential patterns can be identified in [21], differential privacy was applied in sequential data by extracting the essential information in the form of variable-length *n*-grams [21]. In [22], a model-based prefix tree was also constructed and a candidate set of substring patterns were determined. Then, the frequency of the substring patterns was further refined to transform the original data. The problem of constructing a differentially private synopsis for two-dimensional dataset was tackled in [23], where the uniform-grid approach as the partition granularity was applied to balance the noise error and the non-uniformity error. Based on the work [21], a prediction suffix tree model of trajectory micro-data was proposed to automatically adapt the tree height to the data [24] and multiple prefix trees corresponding to different spatial resolutions were proposed to ensure strong privacy protection in the form of e-differential privacy [25]. Hua et al. proposed a generalization algorithm for differential privacy to merge nodes based on their distances [26]. To solve the problem of random and unbounded noises [26], Li et al. proposed a novel differentially private algorithm with a bounded noise generation [10]. To solve the privacy of continuous publication in population statistics, a monitoring framework with *w*-event privacy guarantee was designed [27] including adaptive budget allocation, dynamic grouping and perturbation. In [28], an *n*-body Laplace framework was proposed to prevent social relations inference through the correlation between trajectories. A methodical framework for publishing trajectory data with differential

privacy guarantee as well as high utility preservation was designed by automatically splitting the privacy budget among the different trajectory sequences [29].

2.3 Summary work

As introduced before, perturbation can protect privacy by distorting the dataset while keeping some statistical properties compared with generalization and suppression, which causes less loss of data utility. We prefer perturbation technique to design our privacy preservation scheme. In our previous work, we have proposed a privacy model based on perturbation to resist attacks based on the critical trajectory sequences [30]. To the best of our knowledge, we are the first that proposes an perturbation approach to protect the sensitive attribute of the published trajectory data. However, our previous work has ignored the special case that adding points on the critical sequences may bring new critical sequences. Besides, the data owner can not set the privacy parameters flexibly based on his privacy requirement. To solve the above problems, we propose a privacy model called (l, α, β)-privacy model to resist the record linkage, attribute linkage and similarity attacks without changing any sensitive attribute and further prevent identity closure, privacy closure and similarity closure.

3 Privacy model

In this paper, we focus on publishing trajectory data as in Table 1 while protecting the privacy of sensitive attribute such as *Disease* against attackers with background knowledge about the trajectory. In Table 1, each record corresponds to one individual and contains an identifier as well as a set of geo-referenced and time-stamped elements or spatiotemporal points [18]. These spatiotemporal points constitute an individual's trajectory as one kind of quasi-identifier. Therefore, each trajectory is a sequence of geographical positions of each monitored individual over time in the form (ID, loc, t), where ID represents the owner' s unique identifier and loc represents the owner' s location and t represents a time stamp. The set of locations are arranged in the chronological order to form a trajectory L_t which is defined as follows:

Definition 1 (*Trajectory*) A trajectory L_t is defined as a sequence of spatiotemporal points,

$$L_t = (loc_1, t_1) \rightarrow (loc_2, t_2) \rightarrow \cdots \rightarrow (loc_n, t_n). \tag{1}$$

where n is the length of trajectory, t_i is the time stamp and loc_i represents the owner' s location at t_i.

A trajectory sequence is a non-empty subset of a trajectory, and the length of the sequence is the number of spatiotemporal points contained in the sequence.

In this paper, we mainly consider record linkage attack, attribute linkage attack and similarity attack based on the background knowledge [4]. Generally, background knowledge is a part of the victim's information such as a sequence of spatiotemporal points in this paper. How different attackers can get the background knowledge is not considered in our scheme. We only need to consider the maximum background knowledge for all adversaries to design our preservation approach. The maximum background knowledge represents the maximum length of the trajectory sequence m in this work, which can ensure that all adversaries launch attacks within the range of m.

To resist record linkage attack, attribute linkage attack and similarity attack based on the trajectory sequence, we define our (l, α, β)-privacy model in this paper. l-Diversity ensures that each trajectory sequence whose length is no more than m matches more than l types of SA values in the published table. α-Privacy ensures that the probability of determining each SA value is not greater than α. β-Privacy guarantees that the probability of obtaining similar SA values is not larger than β. Given the original trajectory table T and three privacy parameters l, α and β, our goal is to anonymize T into T^* that satisfies (l, α, β)-privacy model if each record in T^* simultaneously satisfies l-diversity, α-sensitive-association and *β-similarity-association*. First, we define $Q = \{q_1, q_2, \ldots, q_n\}$ as the sequence set of an attacker's background knowledge. For each $q_i \in Q$, we have $q_i \in T^* \wedge |q_i| \leq m$, where m is the sequence upper limit of the attacker's background knowledge. For each $q_i \notin Q$, we have $\neg(q_i \in T^* \wedge |q_i| \leq m)$.

Definition 2 (*l-diversity*) T^* satisfies l-diversity if the number of different SA values in $ASA(q_i)$ satisfies $|ASA(q_i)| \geq l$, where q_i represents a trajectory sequence in Q, and $ASA(q)$ represents all the SA values associated with q.

For example, based on the knowledge of $f6 \rightarrow e8$, $ASA(f6 \rightarrow e8)$ ={HIV, SARS, FEVER} can hold in Table 2. The number of SA values is 3, i.e. $|ASA(f6 \rightarrow e8)| = 3$.

Definition 3 (*α-sensitive-association*) T^* satisfies α-sensitive-association if the probability of inferring the right SA of a record r satisfies $Pr[ASA(r)] \leq \alpha$ with the background knowledge $\forall q_i \in Q$.

For example, an adversary has known that Bob and Freeman possess the trajectory sequence $f6 \rightarrow e9$. From Table 2, we can get $Pr[ASA(Bob)] = Pr[Flu] = \frac{1}{2}$ and $Pr[ASA(Freeman)] = Pr[SARS] = \frac{1}{2}$.

Definition 4 (*β-similarity-association*) All the records can be divided into k groups $T = \{g_1, g_2, \ldots, g_k\}$ according to the SA value type, where g_j represents the jth group. T^* satisfies *β-similarity-association* if the probability of inferring the right group g_j of a record r satisfies $Pr[r \in g_j] \leq \beta$ for $0 \leq \beta \leq 1$ with the background knowledge $\forall q_i \in Q$.

For example, the records in Table 2 are divided into two groups: {{1,4,7,9},{2,3,5,6,8}}. Since Fever is the typical symptom of HIV, HIV and Fever belong to the same disease type. Flu and SARAS both belong to the lung infection, so they are considered as the same disease type. Given a trajectory sequence $d2$, we can get $Pr[Alice \in g_1] = \frac{1}{2}$ and $Pr[Eden \in g_2] = \frac{1}{2}$.

4 Enhanced l-diversity data privacy preservation (EDPP)

Our main research goal is to protect the *SA* privacy while retaining the utility of published data. In this section, we first introduce our basic framework and then elaborate the details of EDPP. Major notations used in this section are listed in Table 3.

4.1 Overview

Our EDPP scheme includes two processes: (1) determining the critical sequences for a given length of trajectory segment, and (2) performing the anonymization operation. A critical sequence is a part of trajectory which meets the predefined length but the matched *SA* values do not meet the (l, α, β)-privacy model. The anonymization operation aims to make each *SA* value satisfy (l, α, β)-privacy model by adding or deleting moving points in each sequence. EDPP includes the following procedures:

(1) Explicit Identifier (EI) is first removed from the original table to generate Table 1.
(2) To determine critical sequences, we find all possible sequences of length no more than m whose *SA* values do not satisfy (l, α, β)-privacy model.
(3) By adding or subtracting points in each sequence obtained from Step (2), we either make the corresponding *SA* values of this sequence satisfy l-diversity or eliminate this sequence.
(4) By adding trajectory points in each sequence obtained from Step (2), we make the corresponding *SA* value of each sequence satisfy $\alpha - sensitive - association$

Table 3 Notations

Notations	Description
m	Maximum sequence length of adversary knowledge
QNL	Set of sequences that do not satisfy l-diversity
QCQ	Set of critical sequences
$QNAB$	Set of sequences that do not satisfy α or β
$T(q)$	Records including q in T
$ASA(q)$	Set of SA values associated with q in T
SU/AD	Set of sequences that are subtracted or added in QNL
max_α	# records whose SA value has the most records in $T(q)$
max_β	# records whose category has the most records in $T(q)$
$PriGain(q)$	Tradeoff metric of q between privacy and utility loss

and *β-similarity-association*. Similarly, we make all the sequences of length no more than m satisfy α or β by adding points.

4.2 Privacy requirements

As mentioned before, our (l, α, β)-privacy model can guarantee the published data T^* satisfies l, α and β privacy requirements to resist record linkage attack, attribute linkage attack and similarity attack. In this subsection, we aim to give the definitions of l, α and β requirements.

4.2.1 *l* Requirement

Based on any trajectory sequence $q_i \in Q$, the inferred total number of distinct *SA* values $|ASA(q_i)|$ is larger than l.

We define c_s^i as the inferred total number of distinct *SA* values based on q_i. We can get the probability of inferring the target individual's record r, $Pr[r]$, must be smaller than the inverse of c_s^i,

$$Pr[r] \leq \frac{1}{c_s^i} \quad s.t. \quad q_i \subset tra(r)$$

. To satisfy l-diversity, c_s^i should satisfy

$$Max\left(\frac{1}{c_s^1}, \frac{1}{c_s^2}, \ldots, \frac{1}{c_s^n}\right) \leq \frac{1}{l}, \tag{2}$$

where the function *Max* always returns the biggest value among the elements.

4.2.2 α Requirement

For each trajectory sequence $q_i \in Q$, the probability of inferring the target individual's *SA* in a specific record, $Pr[ASA(r)]$, is less than α.

We define c_f^i as the maximum number of the same *SA* values and c_t^i as the number of inferred records based on q_i. We can get that the probability of inferring the right *SA* value, $Pr[ASA(r)]$, is less than the ratio between c_f^i and c_t^i,

$$Pr[ASA(r)] \leq \frac{c_f^i}{c_t^i}.$$

To satisfy $\alpha - sensitive - association$, each c_f^i should satisfy

$$Max\left(\frac{c_f^1}{c_t^1}, \frac{c_f^2}{c_t^2}, \ldots, \frac{c_f^n}{c_t^n}\right) \leq \alpha. \tag{3}$$

4.2.3 β Requirement

For each trajectory sequence $q_i \in Q$, the probability of inferring the right group g_j which the target individual's record r belongs to, $Pr[r \in g_j]$, is smaller than β.

We define c_g^i as the maximum number of the same type of SA values inferred according to q_i. We can get the probability of inferring the right group of r, $Pr[r \in g_j]$, must satisfy

$$Pr[r \in g_j] \leq \frac{c_g^i}{c_t^i}.$$

To satisfy *β-similarity-association*, each c_g^i should satisfy

$$Max\left(\frac{c_g^1}{c_t^1}, \frac{c_g^2}{c_t^2}, \ldots, \frac{c_g^n}{c_t^n}\right) \leq \beta. \tag{4}$$

4.3 Detailed algorithms

In what follows, we give the detailed algorithm for each step in the above EDPP scheme.

4.3.1 Determining the critical sequences

Recall that m is the upper bound of the attacker's background knowledge on the trajectory sequence, our goal is to identify all the critical sequences of length m in T. Critical sequence is defined as follows:

Definition 5 (*Critical sequence*) A trajectory sequence q is a critical sequence if and only if it satisfies

$$|ASA(q)| < l \wedge |ASA(q_i)| \geq l, \tag{5}$$

where q_i is a subsequence of q with $\forall q_i \subset q$.

Based on the above definition, we can get the two assertions:

Assertion 1 For an anonymized table T^*, it satisfies l-diversity requirement if and only if it satisfies

$$CS(q) \rightarrow |q| > m \quad s.t. \forall q \in T^*,$$

where $CS(q)$ represents that q is a critical sequence.

Proof. Let T^* satisfy $CS(q) \rightarrow |q| > m$ with $\forall q \in T^*$ and q be a sequence in T^* with $|q| \leq m$. Based on Definition 5, q is obviously not a critical sequence. Then,

we can get $ASA(q) \geq l$ according to Definition 5. In this case, T^* satisfies l-diversity according to Definition 2.

Conversely, let q be a critical sequence in T^* with $|q| \leq m$. We can get T^* does not satisfy the l-diversity requirement according to Definition 2.

Assertion 2 For a critical sequence q, it is no longer a critical sequence after eliminating a spatial–temporal point p with $p \in q$.

Proof. Let q be a critical sequence and p a spatial–temporal point in q. After eliminating p from the original sequence q, we can get a new sequence q_i with $q_i \subset q$. Obviously, we can have $|ASA(q_i)| \geq l$. Based on Definition 5, q_i is not a critical sequence.

According to the two assertions, we can anonymize T into T^* to satisfy l-diversity requirement by eliminating all critical sequences of length no more than m. The following steps are used to determine the critical sequences:

Step 1: First, we obtain all the sequences of length no more than m from T.

Step 2: For each sequence q, if α requirement or β requirement is not satisfied, q is added into a list called $QNAB$.

Step 3: Then, we treat these sequences as vertices. If two sequences q_1 and q_2 satisfy $||q_1| - |q_2|| = 1 \wedge (q_1 \subset q_2 \vee q_2 \subset q_1)$, we will add an edge between vertices v_1 and v_2. By repeating this step, we can get a m-partite graph G, where the whole vertex set can be partitioned into m subsets according to the sequence length from 1 to m. At last, we can get a layered graph according to the sequence length, where the sequence length of each layer is the same, and the smaller length is in the upper layer. Figure 1 is an example of 3-partite graph.

Step 4: For each sequence q in G, q is deleted from the top layer to the end, if $|ASA(q)| \geq l$ holds.

Step 5: Step 4 is repeated until each sequence $q \in G$ in the top layer does not satisfy $|ASA(q)| \geq l$. Figure 2 shows an example after a sequence is deleted from the top layer in Fig. 1.

Step 6: For each sequence $q \in G$, q is added into a list called QNL if it is not in the top layer. Else, q is inserted into a list called QCQ.

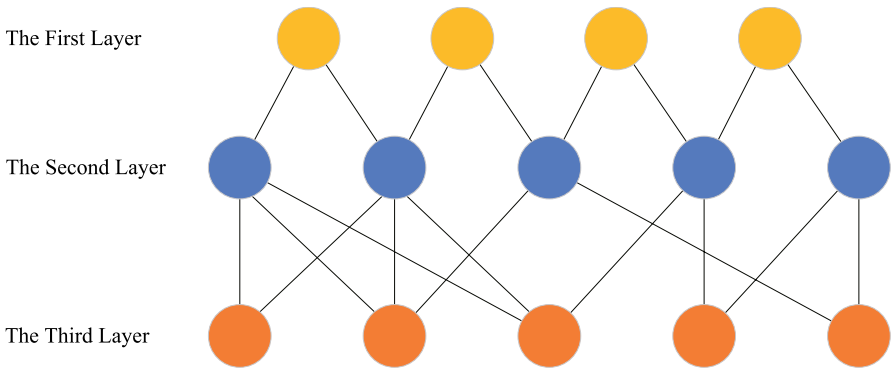


Fig. 1 An example of 3-partite graph

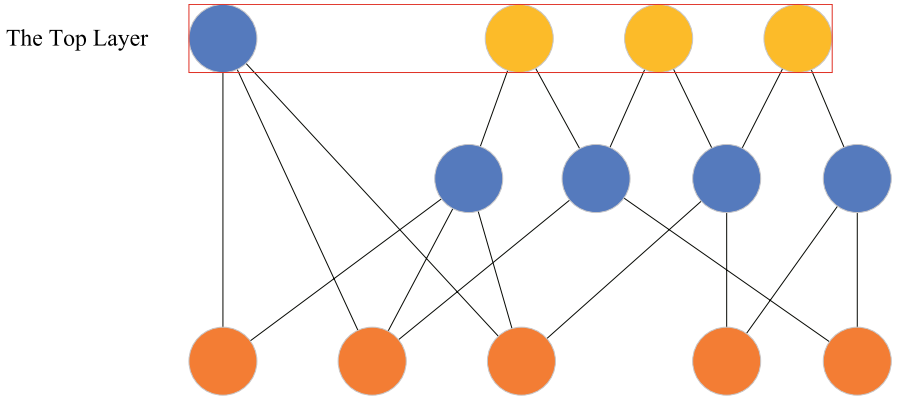


Fig. 2 An example of top layer

4.3.2 Anonymization for *l*-diversity

To achieve *l*-diversity better, we try to eliminate a common spatial–temporal point from sequences in *QCQ*. Therefore, we should make statistics of each point in all sequences of *QCQ* and determine which point should be deleted.

Step 1: We make statistics on the spatial–temporal points in all the sequences of *QCQ* and get a rank list of these points based on their occurrence frequency. Then, we eliminate the point p ranking the first from sequences including p in *QCQ*.

Step 2: Last step can ensure that newly generated sequences in *QCQ* are not critical ones and are removed from *QCQ*. In this step, we should delete p from the sequences including it in *QNL*, where the generated critical sequences are moved to *QCQ* and the non-critical sequences satisfying l requirement are removed from *QNL*. To achieve it, we rely on G to determine the newly generated critical sequences in *QNL*. First, we delete p in G. Then, we execute **Step 5** of last section and make the newly generated top layer contain all critical sequences. Finally, we update *QNL* and *QCQ* according to the **Step 6** of last section.

Step 3: Step 1 and Step 2 are repeated until G is empty.

After Step 1 to Step 3 are executed every round, the total number of sequences in G will decrease. Consequently, our algorithm is strictly convergent no matter what l is.

Step 4: If α requirement or β requirement is not satisfied, q will be added into *QNAB*.

4.3.3 Anonymization for α and β requirements

Before publishing T^*, we adopt addition operation to achieve α requirement and β requirement on those sequences who satisfy *l*-diversity. For a sequence q in *QNAB*, the steps of addition operation are as follows:

First, we choose the records whose *SA* values do not belong to $ASA(q)$ to execute addition. In order to insert a trajectory point at a time stamp, we must ensure that no point in the selected record is associated with the time already, as a person cannot appear in two different places at the same time. Otherwise, the record cannot be modified will not be chosen. Besides, adding a new point in a record may produce more than one new sequence with a limited length of m. Consequently, we must strictly choose the records that generate new critical sequences belonging to Q after addition operation.

Then, we sort the chosen records in descending order of Longest Common Subsequence (*LCS*). *LCS* is a sequence of points common to q and a chosen record. For example, the *LCS* of a sequence $a1 \rightarrow d2 \rightarrow b3$ and a record $a1 \rightarrow d2 \rightarrow c5 \rightarrow f6 \rightarrow c7$ is $a1 \rightarrow d2$.

Step 1: For each q, we first pick up some records to execute the addition operation. To satisfy α requirement and β requirement, a record satisfying the following two conditions will be chosen: (1) Its *SA* value is not associated with the one which has the maximum number of records, max_α, in $T(q)$; and (2) It does not belong to the category which possesses the maximum number of records, max_β, in $T(q)$. These two conditions ensure that the worst-case meets α requirement and β requirement. For example, a sequence $f6 \rightarrow e8$ has five corresponding records in Table 1, the 1st, 3rd, 4th, 7th and 9th ones. The corresponding *SA* values are *HIV*, *SARS*, *Fever*, *Fever* and *Fever*. *Fever* possesses the maximum number of records. If we set α to 50%, we should select another record, such as the 2nd one, to construct q to reduce the probability of inferring *Fever*. After adding $e8$ in the 2nd record, the probability is 50%. Similarly, we prefer the records not belonging to the category which possesses the maximum number of records.

Furthermore, all the chosen records will be sorted in a descending order of *LCS* between q and itself.

Step 2: For each q, we compute num_p, the number of records which need the addition operation to satisfy α requirement, and num_g, the number of records to be added to satisfy β requirement. We use $\max(num_p, num_g)$ to represent the maximum of num_p and num_g.

According to the first $\max(num_p, num_g)$ chosen records, we compute the metric *PriGain* to get a balance between privacy protection and utility loss. $PriGain(q)$ is defined as follows:

$$PriGain(q) = \frac{\lambda \Delta H^s(q) + (1-\lambda)\Delta H^c(q)}{W(q)} \qquad (\lambda \in [0,1])$$

where $H^s_{T^*}(q)$ and $H^s_T(q)$ represent the entropy of *SA* values in $T^*(q)$ and $T(q)$ respectively. $\Delta H^s(q)$ represents the entropy difference. $H^c_{T^*}(q)$ and $H^c_T(q)$ represent the entropy of categories in $T^*(q)$ and $T(q)$ respectively. $\Delta H^c(q)$ represents the difference in category entropy. k is the number of categories. λ is a weight constant representing the impact factor of $\Delta H^s(q)$. Bigger $\lambda \Delta H^s(q) + (1-\lambda)\Delta H^c(q)$ brings more privacy protection.

$$\begin{aligned}\Delta H^s(q) &= H^s_{T^*}(q) - H^s_T(q)\\ &= \sum_{i=1}^{|ASA(q)|} p_i \log p_i - \sum_{i=1}^{|ASA(q)|} p_i^* \log p_i^*\\ \Delta H^c(q) &= H^c_{T^*}(q) - H^c_T(q)\\ &= \sum_{i=1}^{k} p_i \log p_i - \sum_{i=1}^{k} p_i^* \log p_i^*\end{aligned}$$

The utility loss $W(q)$ after anonymization is defined as follows:

$$W(q) = \sum_{i=1}^{|q|} w_i num_i,$$

where num_i represents the number of times that the i-th point needs to be added, and w_i is the weight value of the i-th point. w_i is defined as reciprocal of the number of the i-th point in all the critical sequences of $QNAB$. If one point occurs more frequently, it means the point is required by more sequences to add to meet their privacy requirements. So, its addition may benefit more sequences, and fewer overall points need to be added to make the table meet the privacy requirement. As an example, we have the sequences $a1 \rightarrow b3$, $a1 \rightarrow c5$ and $a1 \rightarrow e4$. To process the 1st sequence, $a1$ may be added into several records. This may make some records contain $a1 \rightarrow c5$ or $a1 \rightarrow e4$, which avoids modifying more records specific for the two sequences. Thus, adding $a1$ can bring more usability and cause lower utility loss.

Finally, q is put into a list in which the elements are sorted in descending order of $PriGain$.

Step 3: In this step, we aim to add points in the above selected records to achieve α requirement and β requirement. We choose a sequence from the list generated in Step 1 to add points to form q until $\max(num_p, num_g)$ records have been processed. During this process, we will not add points into a record if the number of records which possess the same SA value is up to max_α or the number of records associated with a category is up to max_β. Then, q is moved from $QNAB$. If any revised record cannot be further modified to construct a new record for next sequence(s), it will be deleted from the candidate record list of the corresponding sequences, and a new candidate needs to be selected as done in Step 1. For example, $e5$ has been added into one record for the 1st sequence. This record cannot be used by another sequence if a different location needs to be attached, with the time stamp 5. The above process is repeated until none is left in the list.

Step 4: Eventually, we get the anonymous data T^* satisfying (l, α, β)-privacy model.

4.4 Instruction on parameters setting

As discussed in the above sections, the data owners can set three parameters, l, α, and β, according to their privacy requirements. In this subsection, we aim to give some instructions on how to set the parameter reasonably.

l is used to resist record linkage attack. A bigger l represents a smaller probability of successfully launching the record linkage attack. α parameter is used to resist attribute linkage attack. The less α, the smaller the probability of successfully launching the attribute linkage attack. β parameter is used to resist similarity attack. The less β, the smaller the probability of successfully launching the similarity attack.

The more privacy, the less data utility. When the data owner pays more attention on privacy than data utility, he can set a bigger l with smaller α and β. On the contrary, if data utility is more concerned, a smaller l with bigger α and β is optimal. As a result, different owners can set l, α, and β within their tolerance.

5 Privacy analysis

In this section, we prove that our EDPP can both satisfy three privacy requirements of our (l, α, β)-privacy model and resist the corresponding attack. These three parameters can be set based on the data owner's privacy requirement.

5.1 Privacy proof for *l*-diversity

We divide sequences of length no more than m into two types in the original table T. One type of sequences without satisfying l requirement are put into QNL to execute the subtraction operation and critical sequences of length no more than m should be eliminated. The second type of sequences can satisfy l requirement. After our anonymization approach, there is no critical sequence of length more than m in T^*. According to **Assertion 1**, T^* can satisfy l-diversity.

For record linkage attack, the attacker aims to infer the accurate record of the target individual(e.g., Alice) based on the trajectory sequence q_i with $|q_i| \leq m$. l-diversity guarantees that at least l different records include q_i (i.e. $|ASA(q_i) \geq l|$). Then, the probability of inferring Alice's record is less than $\frac{1}{l}$, i.e. the probability of identity closure is less than $\frac{1}{l}$.

As a conclusion, our EDPP scheme can satisfy l privacy requirement and resist record linkage attack.

5.2 Privacy for α-sensitive-association and β-similarity-association

To satisfy α-sensitive-association and β-*similarity-association*, we perform addition for num_p and num_g records including q of length no more than m based on Definitions 2 and 3.

To simplify our algorithm, the $\max(num_p, num_g)$ records are selected to construct q. Because max_α and max_β are constant, the following equations will hold,

$$\frac{max_\alpha}{|T(q)| + \max(num_p, num_g)} \leq \frac{max_\alpha}{|T(q)| + Num_p} \leq \alpha$$

and

$$\frac{max_\beta}{|T(q)| + \max(num_p, num_g)} \leq \frac{max_\beta}{|T(q)| + Num_g} \leq \beta,$$

where the equations can prove that all the sequences of length no more than m in T^* can satisfy both $\alpha - sensitive - association$ and β-$similarity$-$association$. For attribute linkage attack, the attacker aims to infer the sensitive information of the target individual(e.g., Alice) based on the trajectory sequence q with $|q_i| \leq m$. The probability of inferring Alice's SA value of record r, $Pr[ASA(r)]$, is no more than $\frac{c_f^i}{c_t^i}$. Based on α requirement, we have $Pr[ASA(r)] \leq \frac{c_f^i}{c_t^i} \leq Max(\frac{c_f^1}{c_t^1}, \frac{c_f^2}{c_t^2}, \ldots, \frac{c_f^n}{c_t^n}) \leq \alpha$, which implies that the probability of attribute disclosure is no more than α.

For similarity attack, the attacker aims to infer the accurate group of the target individual(e.g., Alice) based on the background knowledge of a trajectory sequence q_i with $|q_i| \leq m$. The probability of inferring the right group g_j of Alice's record r, $Pr[r \in g_j]$, is no more than $\frac{c_g^i}{c_t^i}$. Based on β requirement, we have $Pr[r \in g_j] \leq \frac{c_g^i}{c_t^i} \leq Max(\frac{c_g^1}{c_t^1}, \frac{c_g^2}{c_t^2}, \ldots, \frac{c_g^n}{c_t^n}) \leq \beta$, which implies that the risk that the probability of similarity disclosure is no more than β.

6 Performance evaluation

Setup: We implement our EDPP algorithm in Java. We conduct all experiments on a Mac PC with an Intel Core i5 2.3GHz CPU and 8 GB RAM.

Dataset: To evaluate the performance of our EDPP, we use a real-world dataset that joins the **Foursquare** dataset and **MIMIC-III** dataset. **Foursquare** dataset [31] is a real-world trajectory dataset containing the routes of 140,000 users in a certain area with 92 venues every one hour, forming 2,208 dimensions. **MIMIC-III** [32] is a freely accessible critical care database. The SA is $Disease$ which contains 36 possible values and 9 of them are considered as sensitive values. The SA values are divided into 6 categories, one of which is private. Similarly, we match the diseases in **MIMIC-III** with the trajectory in **Foursquare** in a uniform distribution [4]. We compare our EDPP with **PPTD** [4], **KCL-Local** [9] and **DPTD** [10].

KCL-Local adopts local suppression to achieve the privacy of sensitive information by anonymizing the trajectory data. $(k, C)_m$-privacy model is proposed to adopt k-anonymity to prevent record linkage attack, where C is the confidence threshold to resist attribute linkage attack and the probability of each SA value is not greater than C. In **PPTD**, the sensitive attribute generalization and trajectory local suppression are combined to achieve a tailored personalized privacy model for the publication of trajectory data. In **DPTD**, a novel differentially private trajectory data publishing algorithm is proposed with bounded Laplace noise generation, and trajectory points are merged based on trajectory distances.

6.1 Information loss

The aim of EDPP is to implement the privacy of published data while preserving the data utility. We use information loss to evaluate the utility. In this section, the following metrics are used to evaluate it:

– Trajectory information loss (TIL), the loss rate of the original trajectory data, is defined as

$$\frac{|N(T^*) - N(T)| + |N(T) - N(T^*)|}{|N(T)|},$$

where $N(T^*)$ and $N(T)$ are the sets of trajectory points in T^* and T.

– Frequent sequences loss (FSL), the loss rate of the frequent trajectory sequences, is defined as

$$\frac{|F(T^*) - F(T)| + |F(T) - F(T^*)|}{|F(T)|},$$

where $F(T^*)$ and $F(T)$ are the sets of the frequent items in T^* and T.

We validate the effectiveness of our anonymization algorithm in terms of l, α and β. In this set of experiments, we define $K' = 50$ as the threshold of the frequent sequences and do experiments for the three random number of records, 50K, 100K and 140K.

6.1.1 Effect of *l*

l varies from 3 to 8 for different combinations of parameters α, β, and m. Table 4 shows that the trajectory information loss and frequent sequences loss increase slowly with l, because the substraction or addition operation aims to minimize the number of changed points in order to satisfy l-diversity, which makes the information loss not increase much. In addition, both types of loss increase with m. However, when the number of records change from 50K, 100K to 150K, both types of loss stay relatively stable.

6.1.2 Effect of α

α varies from 0.1 to 0.5 for different combinations of l, β, and m. Table 5 shows that the information loss increases with the decrease of α, because more sequences do not satisfy α-sensitive-association. As discussed before, we select records based on *LCS* and add points based on *PriGain*, which can reduce the number of points to be added. As such, the information loss increases slowly. In addition, Table 5 shows the information loss increases with m, while both types of loss have relatively stable values as the number of records change from 50K, 100K to 150K.

Table 4 Effect of l and m on the information loss in percent ($\alpha = 0.5, \beta = 0.5$)

Metric	Dataset	m = 2						m = 3						m = 4					
		l = 3	l = 4	l = 5	l = 6	l = 7	l = 8	l = 3	l = 4	l = 5	l = 6	l = 7	l = 8	l = 3	l = 4	l = 5	l = 6	l = 7	l = 8
TIL	50K	3.51	4.23	4.92	5.15	5.41	5.86	3.97	4.13	5.22	5.32	5.44	5.45	4.35	4.66	5.41	5.76	5.83	6.37
	100K	3.80	4.27	4.52	5.04	5.18	6.13	4.21	4.44	4.65	5.23	5.53	5.57	4.42	4.79	4.98	5.85	6.03	6.42
	140K	3.82	3.83	4.74	5.27	5.63	6.26	4.08	4.64	4.68	5.32	5.51	5.82	4.27	4.63	4.98	5.85	6.10	6.38
FSL	50K	2.28	2.52	2.79	3.13	3.48	3.80	2.42	2.76	2.91	3.24	3.80	4.03	2.55	3.04	3.18	3.44	3.95	4.15
	100K	2.31	2.49	2.81	3.14	3.52	3.89	2.47	2.65	2.92	3.26	3.81	4.06	2.54	2.83	3.14	3.45	3.94	4.17
	140K	2.33	2.50	2.82	3.17	3.58	3.91	2.48	2.64	2.92	3.25	3.82	4.02	2.56	2.99	3.17	3.57	3.97	4.18

Table 5 Effect of α and m on the information loss in percent ($l = 3, \beta = 0.5$)

Metric	Dataset	m = 2					m = 3					m = 4				
		α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5
TIL	50K	8.05	5.62	5.26	4.33	3.97	8.11	5.79	5.06	4.62	3.96	10.02	7.36	6.19	4.97	4.38
	100K	8.16	5.46	5.22	4.53	3.87	8.60	5.47	5.07	4.54	4.03	10.21	7.26	6.30	5.02	4.54
	140K	8.01	5.44	5.24	4.65	4.03	8.64	5.47	5.10	4.67	3.99	10.28	7.11	6.33	5.06	4.49
FSL	50K	9.89	7.11	4.03	2.84	2.45	9.55	6.83	5.03	3.66	2.78	10.83	7.24	5.30	4.37	2.41
	100K	9.63	6.86	5.14	3.18	2.33	9.55	7.32	5.57	3.57	2.67	10.74	7.78	5.52	3.87	2.71
	140K	10.01	6.99	4.88	2.78	2.34	9.73	6.90	5.15	3.71	2.77	10.43	7.87	5.40	3.82	2.55

6.1.3 Effect of β

Under different number of records, for selected parameters l, α, and m, we vary β from 0.1 to 0.5. Similar to the effect of α, Table 6 shows the information loss increases slowly with the decrease of β and increase of m.

6.1.4 Effect of K'

K' varies from 50 to 130 with a set of random parameters $l = 3$, $\alpha = 0.4$, and $\beta = 0.5$. Figure 3 shows the frequent sequences loss decreases with the increase of K', because the number of frequent sequences not satisfying (l, α, β) begins to drop with the increase of K'.

6.2 Disclosure risk

We use the disclosure risk as a metric to measure the probability of privacy breach for each sequence q:

$$P_{dis}(q) = max\left(\frac{1}{|ASA(q)|}, \frac{max_\alpha}{|T(q)|}, \frac{max_\beta}{|T(q)|}\right),$$

where $\frac{1}{|ASA(q)|}$, $\frac{max_\alpha}{|T(q)|}$, and $\frac{max_\beta}{|T(q)|}$ represent the probability of identity disclosure, that of attribute disclosure, and that of similarity disclosure, respectively.

We randomly select 50K sub-trajectories of length no more than m from the anonymous database, and calculate the probability of privacy disclosure for these sequences. Figure 4 shows that the average disclosure probability decreases with the increase of l and decrease of α or β, because the privacy requirements become higher. Moreover, the average disclosure probability increases with m.

6.3 Comparison

We also compare our EDPP with **KCL-Local**, **PPTD** and **DPTD** on trajectory information loss, frequent sequences loss and run time. Since these schemes adopt different privacy models, we cannot directly compare them. To have a fair comparison, we modify our algorithm EDPP to implement $(k, C)_m$-privacy model as used in KCL-Local, called **EDPP-KC**. ϵ used in the differential privacy method DPTD is assigned as follows to keep the disclosure risk at the same level as that of other three schemes:

$$P_{dis}(q) = max\left(\frac{1}{|ASA(q)|}, \frac{max_\alpha}{|T(q)|}\right)$$

Table 6 Effect of β and m on the information loss in percent ($l = 3, \alpha = 0.5$)

Metric	Dataset	m = 2					m = 3					m = 4				
		β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5
TIL	50K	5.76	5.08	4.60	4.47	4.33	5.95	5.17	4.84	4.68	4.33	6.12	5.19	5.01	4.84	4.44
	100K	5.96	5.14	4.70	4.64	4.52	6.13	5.22	4.97	4.69	4.13	6.05	5.44	5.08	4.92	4.32
	140K	5.89	5.11	4.73	4.66	4.43	6.07	5.30	4.85	4.60	4.34	6.21	5.37	5.01	4.92	4.48
FSL	50K	6.32	4.84	3.28	2.99	2.36	6.50	5.21	3.68	3.26	2.79	6.88	5.53	3.86	3.49	3.01
	100K	6.34	4.72	3.34	2.98	2.38	6.61	5.38	3.63	3.31	2.68	6.83	5.49	3.81	3.47	2.98
	140K	6.41	4.81	3.42	3.02	2.41	6.59	5.16	3.74	3.32	2.77	6.92	5.41	3.84	3.54	3.01

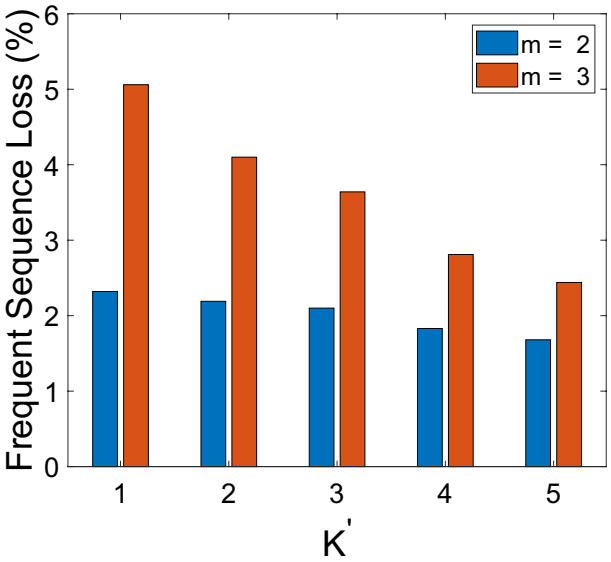
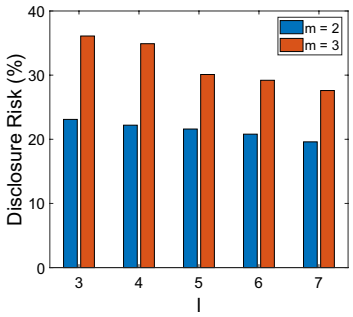
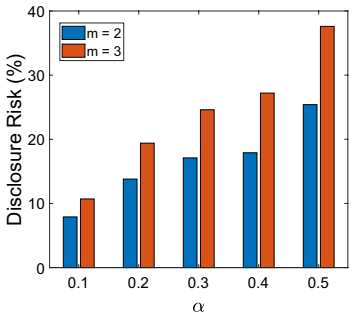


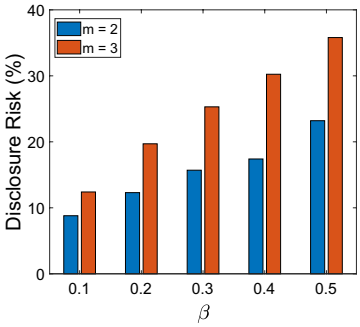
Fig. 3 Frequent sequences loss vs. K' ($l = 3, \alpha = 0.4, \beta = 0.5$)



(a) Disclosure risk vs. l ($\alpha, \beta = 0.5$)



(b) Disclosure risk vs.$\alpha(l = 3, \beta = 0.5)$



(c) Disclosure risk vs.$\beta(l = 3, \alpha = 0.5)$

Fig. 4 Disclosure risk

and

$$P_{dis}(k, C) = P_{dis}(\epsilon),$$

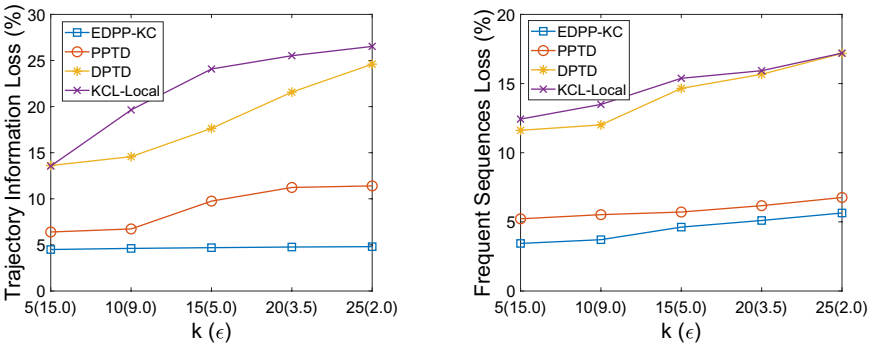
where $P_{dis}(k, C)$ represents the disclosure probability under different k and C, and $P_{dis}(\epsilon)$ represents the disclosure probability under different ϵ which is determined according to the disclosure risk level. $\frac{1}{|ASA(q)|}$ and $\frac{max_{\alpha}}{|T(q)|}$ represent the probability of identity disclosure and attribute disclosure respectively.

6.3.1 Effect of *k*

k varies from 5 to 25 with $C = 0.5$, $m = 3$ and $K' = 50$ under 140K records. Figure 5 shows both kinds of loss increases with k because more sequences not satisfying k-anonymity causes the higher information loss. Our EDPP-KC has the best performance because we aim to minimize the number of the changed points. KCL-Local has the worst performance loss because too much moving points are eliminated from the trajectory data in the global suppression. DPTD generates Laplace noise to achieve differential privacy. As ϵ decreases in Fig. 5, DPTD can get better privacy. However, the larger noise causes more trajectory information and frequent sequences loss than PPTD. PPTD only handles the sensitive records which may cause the privacy disclosure, thus PPTD has a lower information loss than DPTD.

6.3.2 Effect of *C*

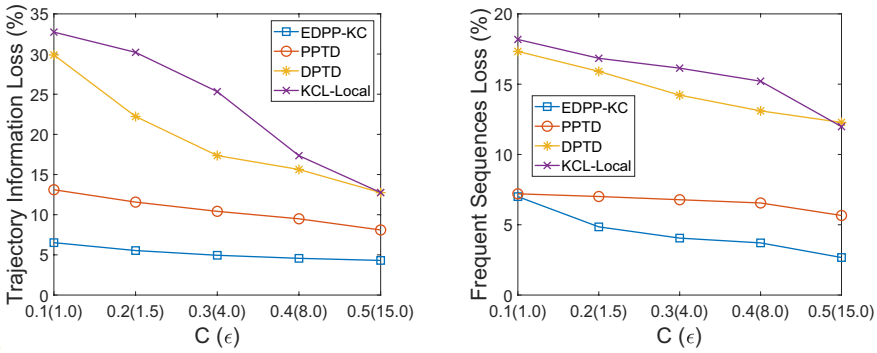
C varies from 0.1 to 0.5 with $k = 5$, $m = 3$ and $K' = 50$ under 140K records. In Fig. 6, both types of information loss decreases with the increase of C because fewer sequences do not satisfy the confidence threshold C, making the loss lower. Similar to the above discussion, EDPP-KC has the best performance. KCL-Local possesses the worst performance. As ϵ decreases, trajectory information loss and frequent sequences loss of DPTD become greater, which is slight better than KCL-Local.



(a) Trajectory information loss vs. k

(b) Frequent sequences loss vs. k

Fig. 5 Information loss vs. k ($C = 0.5, m = 3, K' = 50$)



(a) Trajectory information loss vs. C (b) Frequent sequences loss vs. C

Fig. 6 Information loss vs. C ($k = 5, m = 3, K' = 50$)

Compared with **KCL-Local**, **PPTD**, and **DPTD** the trajectory information loss of EDPP can be improved by up to 76.90%, 48.17% and 72.86% respectively and the frequent sequences loss can be improved by up to 71.03%, 28.99% and 69.32% respectively.

6.3.3 Run time

Figure 7 shows the run time increases with the number of records. With the simplicity of generating Laplace noise, DPTD has the lowest run time. DPTD spends most of its time on constraint inference to guarantee the data utility. KCL-Local also has

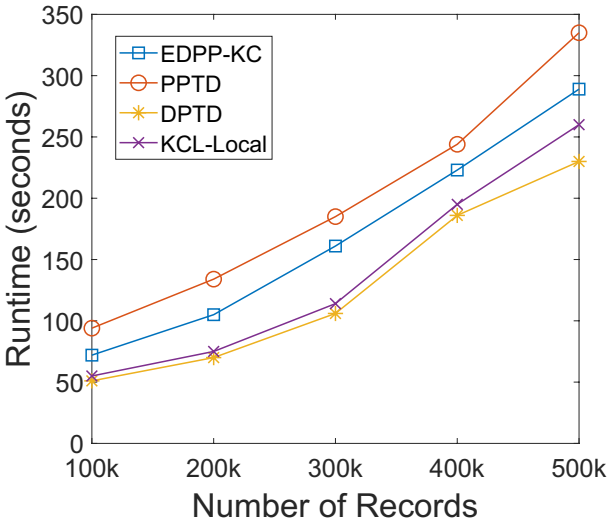


Fig. 7 Run time vs. records ($k = 20, C = 0.4$)

the good performance on run time because only suppression is adopted. In PPTD, the sensitive attribute generalization and trajectory local suppression are combined to achieve the privacy, which causes the most run time. In EDPP-KC, it takes much time to determine the critical sequences.

7 Conclusion

We design and implement an anonymous technique named EDPP to protect the sensitive attribute during the publication of trajectory data. To resist record linkage, attribute linkage and similarity attack based on the background knowledge of critical sequences, we adopt perturbation to process these sequences by adding or deleting some moving points so that the published data satisfy our (l, α, β)-privacy model. Our performance studies based on a comprehensive set of real-world data demonstrate that EDPP can provide higher data utility compared to peer schemes. Our privacy analysis shows that EDPP can provide better privacy for the sensitive attribute. In the future work, we will optimize our algorithm to handle extremely large trajectory dataset with the aid of indexing and pruning.

Acknowledgements This work is supported by the National Key R&D Program of China (Grant No. 2017YFC0704200). This research is also sponsored by the National Natural Science Foundation of China (Grant Nos. 61872053, 61572413 and U1636205) and Research Grants Council, Hong Kong SAR, China (Grant Nos. 15238116, 15222118, 15218919 and C1008-16G), the Open Project of the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences (2020-ZD-04), and the Key-Area Research and Development Program of Guangdong Province (2019B010136001); the Science and Technology Planning Project of Guangdong Province (LZC0023).

References

1. Amendola, S., Lodato, R., Manzari, S., Occhiuzzi, C.: RFID technology for IoT-based personal healthcare in smart spaces. IEEE Internet of Things J. **1**(2), 144–152 (2014)
2. Davis, A.M., Perruccio, A.V., Ibrahim, S., Hogg-Johnson, S., Wong, R., Streiner, D.L., Beaton, D.E., Cote, P., Gignac, M.A., Flannery, J.: The trajectory of recovery and the inter-relationships of symptoms, activity and participation in the first year following total hip and knee replacement. Osteoarthr. Cartil. **19**(12), 1413 (2011)
3. Xu, C., Zheng, W., Hui, P., Zhang, K., Liu, H.: Hygeia: a practical and tailored data collection platform for mobile health. In: IEEE International Conference on Ubiquitous Intelligence and Computing, pp. 20–27 (2015)
4. Komishani, E.G., Abadi, M., Deldar, F.: Pptd: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowl. Based Syst. **94**, 43–59 (2016)
5. Victor, N., Lopez, D., Abawajy, J.H.: Privacy models for big data: a survey. Int. J. Big Data Intell. **3**(1), 61–75 (2016)
6. Fung, B., Wang, K., Chen, R., Philip, S.Y.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. (CSUR) **42**(4), 14 (2010)
7. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: International Conference on Data Engineering, pp. 106–115 (2007)
8. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L -diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. From Data **1**(1), 1–12 (2007)
9. Chen, R., Fung, B.C.M., Mohammed, N., Desai, B.C., Wang, K.: Privacy-preserving trajectory data publishing by local suppression. Inf. Sci. **231**, 83–97 (2013)

10. Li, M., Zhu, L., Zhang, Z., Rixin, X.: Achieving differential privacy of trajectory data publishing in participatory sensing. Inf. Sci. **400**, 1–13 (2017)
11. Gramaglia, M., Fiore, M., Tarable, A., Banchs, A.: Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In: INFOCOM 2017-IEEE Conference on Computer Communications, IEEE, pp. 1–9. IEEE (2017)
12. Xin, Y., Xie, Z.Q., Yang, J.: The privacy preserving method for dynamic trajectory releasing based on adaptive clustering. Inf. Sci. **378**, 131–143 (2017)
13. Dong, Y., Pi, D.: Novel privacy-preserving algorithm based on frequent path for trajectory data publishing. Knowl. Based Syst. **148**, 55–65 (2018)
14. Yarovoy, R., Bonchi, F., Lakshmanan, L.V. S., Wang, W.H.: Anonymizing moving objects: how to hide a mob in a crowd? In: International Conference on Extending Database Technology, pp. 72–83 (2009)
15. Ghasemzadeh, M., Fung, B.C.M., Chen, R., Awasthi, A.: Anonymizing trajectory data for passenger flow analysis. Transport. Res. C **39**(2), 63–79 (2014)
16. Brito, Felipe T., Neto, Antônio C Araújo, Costa, Camila F., Mendonça, André L.C., Machado, Javam C.: A distributed approach for privacy preservation in the publication of trajectory data. In: Proceedings of the 2nd Workshop on Privacy in Geographic Information Collection and Analysis, p. 5. ACM (2015)
17. Terrovitis, M., Poulis, G., Mamoulis, N., Skiadopoulos, S.: Local suppression and splitting techniques for privacy preserving publication of trajectories. IEEE Trans. Knowl. Data Eng. **29**(7), 1466–1479 (2017)
18. Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D.L., Aivodji, U.M., Olivier, B., Quertier, T., Stanica, R.: Privacy in trajectory micro-data publishing: a survey and security. arXiv: Cryptography (2019)
19. Chen, R., Fung, B., Desai, B.C.: Differentially private trajectory data publication. arXiv preprint. arXiv:1112.2020 (2011)
20. Shao, D., Jiang, K., Kister, T., Bressan, S., Tan, K.: Publishing trajectory with differential privacy: a priori vs. a posteriori sampling mechanisms. In: Proceedings of the 24th International Conference on Database and Expert Systems Applications, pp. 357–365 (2013)
21. Chen, R., Fung, B.C.M., Desai, B.C., Sossou, N.M.: Differentially private transit data publication: a case study on the Montreal transportation system. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 213–221 (2012)
22. Bonomi, L., Xiong, L.: A two-phase algorithm for mining sequential patterns with differential privacy. In: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, pp. 269–278 (2013)
23. Qardaji, W., Yang, W., Li, N.: Differentially private grids for geospatial data. In: IEEE 29th International Conference on Data Engineering, pp. 757–768 (2013)
24. Zhang, J., Xiao, X., Xie, X.: PrivTree: a differentially private algorithm for hierarchical decompositions. In: Proceedings of the 2016 International Conference on Management of Data, pp. 155–170 (2016)
25. He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C.M., Srivastava, D.: Dpt: differentially private trajectory synthesis using hierarchical reference systems. In: Proceedings of VLDB, pp. 1154–1165 (2015)
26. Hua, J., Gao, Y., Zhong, S.: Differentially private publication of general time-serial trajectory data. In: Computer Communications, pp. 549–557 (2015)
27. Wang, Q., Zhang, Y., Xiao, L., Wang, Z., Qin, Z., Ren, K.: Real-time and spatio-temporal crowdsourced social network data publishing with differential privacy. IEEE Trans. Depend. Secure Comput. **15**(4), 591–606 (2018)
28. Ou, L., Qin, Z., Liao, S., Hong, Y., Jia, X.: Releasing correlated trajectories: towards high utility and optimal differential privacy. IEEE Trans. Depend. Secure Comput. **14**, 9–21 (2018)
29. Gursoy, M.E., Liu, L., Truex, S., Lei, Y.: Differentially private and utility preserving publication of trajectory data. IEEE Trans. Mob. Comput. **18**(10), 2315–2329 (2019)
30. Yao, L., Wang, X., Wang, X., Hu, H., Wu, G.: Publishing sensitive trajectory data under enhanced l-diversity model. In: The 20th IEEE International Conference on Mobile Data Management, pp. 160–169 (2019)
31. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, pp. 1082–1090 (2011)

32. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)

Affiliations

Lin Yao[1,2] · Zhenyu Chen[3] · Haibo Hu[4] · Guowei Wu[5] · Bin Wu[6]

Lin Yao
yaolin@dlut.edu.cn

Zhenyu Chen
liuluoqianqiu@126.com

Haibo Hu
haibo.hu@polyu.edu.hk

1 International School of Information Science & Engineering, Dalian University of Technology, Dalian, China

2 Peng Cheng Laboratory, Cyberspace Security Research Center, Shenzhen 518057, China

3 School of Software, Dalian University of Technology, Dalian, China

4 Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR, China

5 School of Software, Dalian University of Technology, Dalian, China

6 State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China