An Enhanced l-Diversity Privacy Preservation of Spatial-Temporal Data

Lin Yao*, Zhenyu Chen†, Haibo Hu‡, Yundong Sun†, and Guowei Wu†

*International School of Information Science & Engineering, Dalian University of Technology, China
†School of Software, Dalian University of Technology, China
‡Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

***Abstract*—The widely application of positioning technology has made collecting the movement of people feasible and therefore plenty of trajectory data have been collected, published, and analyzed in real-life applications. However, maintaining privacy in the published data is a critical problem, because known partial information of an individual can be used to determine the specific record. For example, some trajectory points may leak an individual's sensitive spatial-temporal event, such as "visiting one hospital last week" and "getting some kinds of infectious disease". To prevent record linkage, attribute linkage, and similarity attacks based on the background knowledge of trajectory data and protect the individuals' information, we propose a data privacy preservation with enhanced l-diversity. First, determine those critical spatial-temporal sequences which are more likely to result in 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 EDPP achieves better privacy while still ensuring high utility, compared with existing privacy preservation schemes on trajectory.**

***Keywords*-Spatial Temporal, Sensitive Privacy Preservation, Trajectory Data Publishing**

I. Introduction

The popularity of smart mobile devices with positioning technologies triggers the advent of location-based services, such as "where is the nearest China restaurant." Mobile users must share their current locations or a sequence of past trajectory with the service provider. Therefore, vast amounts of trajectory data are collected with other information. For example, wearable devices 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 [22]. Data miners have also shown great interest in analyzing these data to provide plentiful serves for people. Recent studies [1][7] have shown that tracking the environmental exposure of a person with his daily trajectories helps to improve diagnose. However, the public data may contain sensitive and private information about individuals, such as health status. Therefore, protecting user privacy when publishing data is a significant challenge that goes beyond removing the identity identifier such as the name. The critical problem is that the attacker can infer an individual's other sensitive information through some background knowledge, such as frequently visited locations. The focus of this paper is to preserve individual privacy for publishing trajectory data which is associated with non-sensitive and sensitive attributes.

Table I
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 I [14] shows 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 [19]. *Explicit Identifier (EI)*, such as the name, is used to identify an individual uniquely, which is always removed from the published table in the anonymization step. On the other hand, a single *Quasi-Identifier (QI)* cannot uniquely identify an individual, but a few QIs can be combined to identify him. 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 I. *Non-sensitive attribute* can be known by the public without any privacy concern.

The following types of attacks are mostly considered in current approaches of preserving data privacy, record linkage attack, attribute linkage attack, and similarity attack [14][9]. To illustrate them, we take Table II as an example.

- **Record linkage attack.** An adversary could identify the unique record from the published table according to a certain trajectory sequence of limited length. For example, if an adversary has the background knowledge of Alice's trajectory sequence $d2 \rightarrow e4$. Based on it,

The following publication Yao, L., Chen, Z., Hu, H. et al. Sensitive attribute privacy preservation of trajectory data publishing based on l-diversity. Distrib Parallel Databases 39, 785–811 (2021) is available at https://doi.org/10.1007/s10619-020-07318-7.

Table II
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

the adversary can infer that the 1st record belongs to Alice. Thus, Alice's disease privacy of HIV in Table II is leaked.

- **Attribute linkage attack.** The adversary may not precisely identify the victim's record but could infer his or her sensitive information such as SA from the published data based on a trajectory sequence. For example, an adversary knows that Bob has a trajectory sequence of $c5 \rightarrow c7$. Since only the 2nd and 5th records contain it in Table II, he can infer that Bob must have got Flu.
- **Similarity attack.** The adversary may not precisely identify the victim's record, but could infer his group information from the published data based on the intrinsic relationship between SA values. For example, an adversary knows that Tom has a trajectory sequence of $c7$. Based on Table II, he can infer that Tom may suffer Flu, $Fever$, or $SARS$. By excluding $Fever$, he can speculate that Tom has a probability of $\frac{4}{5}$ to have a lung infection Flu or $SARS$.

These three attacks can cause identity disclosure, attribute disclosure, and similarity disclosure respectively [14]. *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 set of SA values with a high probability.

To prevent the above three kinds of disclosure, some anonymization operations should be taken to modify the original table. The typical anonymization approaches [9] include generalization, suppression, anatomization, permutation, and perturbation. Generalization and suppression replace values of specific attributes with less specific values. Anatomization and permutation de-associate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in a group including QID. In perturbation, the data will be distorted by adding noise, swapping values, or generating synthetic data. Most existing privacy-preserving approaches in publishing trajectory data are mainly based on generalization and suppression, and perturbation privacy [14]. While, generalization and suppression may eliminate a certain number of moving points by replacing some spatial-temporal points with a broader category or wildcard "*", which causes significant loss of data utility. Comparatively, perturbation can protect privacy by distorting the dataset while keeping some statistical properties [9].

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 k-anonymity, l-diversity can provide stronger privacy preservation by guaranteeing l different sensitive attributes in a group [16]. However, it cannot resist attribute linkage attack and similarity attack. To resist the three kinds of attacks, we propose our (l, α, β)-privacy model. With the trajectory sequence being background knowledge, 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. 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 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** [14], **KCL-Local** [5] and **DPTD** [15], our DPPP is superior in terms of data utility ratio and privacy.

The remainder of this paper is organized as follows. In Section II, we discuss the related work. Privacy model is given in Section III. In Section IV, we present the details of our approach. Privacy analysis is given in Section V. Simulations on data utility are presented in Section VI. Finally, we conclude our work in Section VII.

II. RELATED WORK

Most existing privacy preserving approaches in publishing trajectory data are based on generalization and suppression, and perturbation [14].

Generalization and Suppression: Generalization replaces some QI values with a broader category such as a parent value in the taxonomy of an attribute. In [14],

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 [21]. In [8], 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 [24], extreme-union and symmetric anonymization were proposed to build anonymous groups and avoid a moving object being identified through the correlation between anonymization groups. [5] 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 [10]. In [2], k^m-anonymity was proposed to suppress the critical location points chosen from quasi-identifiers to protect against the record linkage attack. In [18], location suppression and trajectory splitting were used to prevent privacy leaks and improve data utility of aggregate query and frequent sequences.

Perturbation: Perturbation aims to protect the privacy with limiting the upper bound of utility loss. Differential privacy is a main form of data perturbation. In cryptography, differential privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. Differential privacy can protect the privacy of individual users under any background knowledge of adversaries. In [3], differential privacy was first adopted to protect the privacy of trajectory data. It was also applied in sequential data by extracting the essential information in the form of variable-length n-grams [4]. Hua et al. proposed a generalization algorithm for differential privacy to merge nodes based on their distances [12]. In [15], a differentially private trajectory data publishing algorithm with a bounded noise generation algorithm was proposed. To solve the privacy of continuous publication in population statistics, a monitoring framework with w-event privacy guarantee was designed. It includes adaptive budget allocation, dynamic grouping and perturbation [20]. In [17], an n-body Laplace framework was proposed to prevent social relations inference through the correlation between trajectories. Nonetheless, differential-based approaches add random and unbounded noises to the original data, which may seriously degrade the utility of released trajectory data. To provide better data utility, we proposed a privacy model to resist background knowledge attacks based on some trajectory sequences by adding or subtracting some trajectory points in the published data [23].

Summary Work: In this work, we aim to 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. Similar to our previous work, we regard the critical sequences which can determine the specific individuals as the attackers' background knowledge and execute addition or subtraction on them in order to eliminate these sequences. Different from our previous work [23], on the one hand, the three parameters, l, α and β, are not pre-defined, but can be adjusted by the data owner based on his privacy requirement. On the other hand, we consider the special case that the addition operation may bring new critical sequences, which has been ignored in our previous work.

III. Privacy Model

In this paper, we focus on publishing trajectory data as in Table I while protecting the privacy of sensitive attribute such as *Disease* against attackers with background knowledge about the trajectory. Each individual may visit different locations at different time. Consequently, a sequence of spatial-temporal records is generated 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,*

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

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.

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 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 $l - diversity$, $\alpha - sensitive - association$ and $\beta - similarity - association$. First, we define $Q = \{q_1, q_2, \cdots, q_n\}$ 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 II. The number of SA values is 3, i.e. $|ASA(f6 \rightarrow e8)| = 3$.

Definition 3. *($\alpha - sensitive - association$) T^* satisfies $\alpha - sensitive - association$ 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 II, we can get $Pr[ASA(Bob)] = Pr[Flu] = \frac{1}{2}$ and $Pr[ASA(Freeman)] = Pr[SARS] = \frac{1}{2}$.

Definition 4. *($\beta-similarity-association$) All the records can be divided into k groups $T = \{g_1, g_2, \cdots, g_k\}$ according to the SA value type, where g_j represents the j-th group. T^* satisfies $\beta - similarity - association$ 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 II are divided into two groups: {{1,4,7,9},{2,3,5,6,8}} Given a trajectory sequence $d2$, we can get $Pr[Alice \in g_1] = \frac{1}{2}$ and $Pr[Eden \in g_2] = \frac{1}{2}$.

IV. 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 III.

Table III
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.
$SUIAD$	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.

A. 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 I.
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 $\alpha - sensitive - association$ and $\beta - similarity - association$. Similarly, we make all the sequences of length no more than m satisfy α or β by adding points.

B. 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.

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

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

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

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

$$Max(\frac{c_f^1}{c_t^1}, \frac{c_f^2}{c_t^2}, \cdots, \frac{c_f^n}{c_t^n}) \leq \alpha. \quad (3)$$

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 $\beta - similarity - association$, each c_g^i should satisfy

$$Max(\frac{c_g^1}{c_t^1}, \frac{c_g^2}{c_t^2}, \cdots, \frac{c_g^n}{c_t^n}) \leq \beta. \quad (4)$$

C. Detailed Algorithms

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

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*

$$|ASA(q)| < l \wedge |ASA(q_i)| \geq l, \quad (5)$$

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 $|ASA(q)| \geq l$ does not hold, q is added into a list called QNL and is given a false mark, i.e. $l - diversity$ is not satisfied. Else if $\alpha - sensitive - association$ or $\beta - similarity - association$ is not satisfied, q is added into a list called $QNAB$.

Step 3: We find a sequence q of the shortest length in QNL whose mark is false and then move q into a list QCQ if QCQ does not have any subsequence of q. Else, we set the mark of q to be true.

Step 4: Step 3 is repeated until the mark of all sequences in QNL is true. Then, we get the set of critical sequences in QCQ.

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. We also 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.

In this step, the key issue is how to determine the newly generated critical sequences in QNL. If we adopt **Steps 3** and **Step 4** of last section, all the sequences in QNL should be considered. In fact, only the sequences deleting p in QNL will be affected. To reduce the computational overhead, we only execute **Steps 3** and **Step 4** of last section for those sequences deleting p in QNL.

Step 3: Step 1 and **Step** 2 are repeated until both QNL and QCQ are empty.

Every time **Step** 1 to **Step** 3 are executed, the total number of sequences in QNL and QCQ 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 $QNAB$.

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 I, 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:

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

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

$H^s_{T^*}(q)$ 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. 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.

V. 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.

A. 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.

B. 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 **Definition 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 $\beta-similarity-association$. 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 $Pr[ASA(r)] \leq \frac{c_f^i}{c_t^i} \leq Max(\frac{c_f^1}{c_t^1}, \frac{c_f^2}{c_t^2}, \cdots, \frac{c_f^n}{c_t^n}) \leq \alpha$, 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 $Pr[r \in g_j] \leq \frac{c_g^i}{c_t^i} \leq Max(\frac{c_g^1}{c_t^1}, \frac{c_g^2}{c_t^2}, \cdots, \frac{c_g^n}{c_t^n}) \leq \beta$, which implies that the risk that the probability of similarity disclosure is no more than β.

VI. Performance Evaluation

Setup: We implement our DPPP algorithm in Python. We conduct all experiments on a PC with an Intel Core i7 2.5GHz CPU and 8 GB RAM.

Dataset: To evaluate the performance of our DPPP, we use a real-world dataset that joins the **Foursquare** dataset and **MIMIC-III** dataset. **Foursquare** dataset [6] is a real-world trajectory dataset containing the routes of 140,000 users in a certain area with 92 venues in 24 hours, forming 2,208 dimensions. **MIMIC-III** [13] 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. We compare our DPPP with **PPTD** [14], **KCL-Local** [5] and **DPTD** [15].

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.

A. Information Loss

The aim of DPPP 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.

1) Effect of l: l varies from 3 to 8 for different combinations of parameters α, β, and m. Table IV 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.

2) Effect of α: α varies from 0.1 to 0.5 for different combinations of l, β, and m. Table V 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 V 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.

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 VI shows the information loss increases slowly with the decrease of β and increase of m.

4) Effect of K': K' varies from 50 to 130 with a set of random parameters $l = 3$, $\alpha = 0.4$, and $\beta = 0.5$. Fig.1 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'.

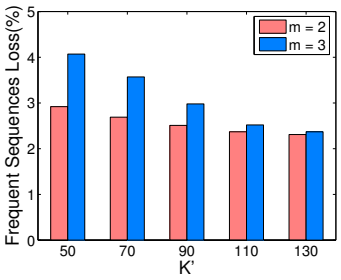


Figure 1. Frequent sequences loss vs. K' ($l = 3, \alpha = 0.4, \beta = 0.5$)

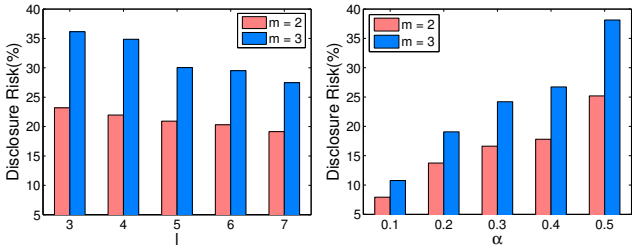
B. Disclosure Risk

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

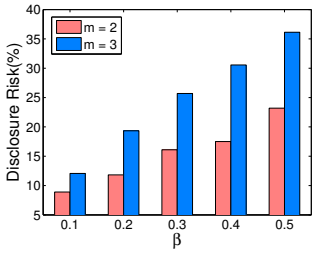
$$P_{dis}(q) = max(\frac{1}{|ASA(q)|}, \frac{max_\alpha}{|T(q)|}, \frac{max_\beta}{|T(q)|}),$$

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. Fig.2 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.



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



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

Figure 2. Disclosure risk

C. Comparison

We also compare our DPPP with **KCL-Local**, **PPTD** and **DPTD** on trajectory information loss, frequent sequences

Table IV
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	4.19	4.40	5.15	5.67	6.00	6.27	4.50	4.75	5.45	5.82	6.03	6.32	4.77	5.03	5.79	6.08	6.37	6.65
	100K	4.15	4.34	5.08	5.53	5.79	6.12	4.61	4.69	5.37	5.71	5.97	6.24	4.75	4.96	5.54	6.07	6.25	6.57
	140K	4.50	4.69	5.39	5.67	6.00	6.25	4.52	4.76	5.50	5.93	6.09	6.42	4.56	4.93	5.65	6.16	6.52	6.74
FSL	50K	2.73	2.84	3.12	3.51	3.62	4.39	2.84	2.99	3.20	3.44	3.69	4.53	3.19	3.27	3.52	3.60	3.97	4.59
	100K	2.63	2.74	3.12	3.52	3.72	4.39	2.79	2.88	3.17	3.35	3.58	4.44	3.07	3.21	3.46	3.60	4.11	4.45
	140K	2.75	2.89	3.14	3.34	3.65	4.25	2.76	2.94	3.12	3.43	3.73	4.33	3.06	3.17	3.33	3.49	3.92	4.49

Table V
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	8.59	5.65	5.41	4.77	4.19	8.98	5.80	5.43	5.01	4.50	10.04	7.21	6.44	5.38	4.77
	100K	8.26	5.51	4.97	4.56	4.15	8.88	5.75	5.35	5.00	4.61	10.74	7.61	6.32	5.09	4.75
	140K	8.47	5.84	5.14	4.82	4.50	8.91	5.73	5.27	4.81	4.52	10.39	7.10	6.03	5.01	4.56
FSL	50K	10.30	7.30	5.03	3.10	2.73	10.42	7.35	5.37	3.74	2.84	10.77	7.46	5.91	4.57	3.19
	100K	10.20	7.31	5.24	3.51	2.63	10.47	7.43	5.57	4.03	2.79	11.32	8.00	5.71	4.29	3.07
	140K	10.42	7.14	4.83	2.92	2.75	10.48	7.10	5.97	4.07	2.76	11.35	7.58	6.23	4.32	3.06

Table VI
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	6.04	5.15	4.79	4.41	4.19	6.33	5.57	5.10	4.93	4.50	6.53	5.70	5.43	5.02	4.77
	100K	5.84	5.02	4.63	4.34	4.15	6.20	5.46	5.17	5.02	4.61	6.41	5.85	5.39	4.89	4.75
	140K	6.25	5.42	5.08	4.81	4.50	6.09	5.53	5.20	4.76	4.52	6.49	5.88	5.50	4.89	4.56
FSL	50K	6.51	5.12	3.83	2.94	2.73	6.84	5.29	3.95	3.14	2.84	7.23	5.69	4.10	3.75	3.19
	100K	6.68	5.10	3.71	2.73	2.63	6.79	5.28	3.97	3.35	2.79	7.10	5.72	4.10	3.63	3.07
	140K	6.47	5.55	3.57	2.80	2.75	6.76	5.45	3.90	3.43	2.76	7.03	5.70	4.05	3.63	3.06

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 DPPP to implement $(k, C)_m$-privacy model as used in KCL-Local, called **DPPP-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:

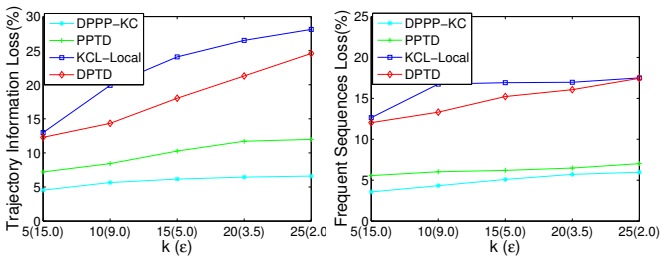
$$P_{dis}(q) = max(\frac{1}{|ASA(q)|}, \frac{max_\alpha}{|T(q)|})$$

and

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

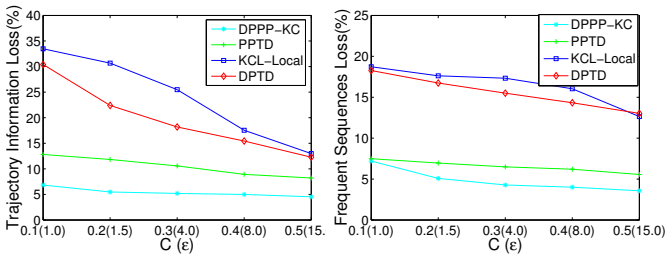
where $P_{dis}(k, C)$ represents the disclosure probability under different k and C, and $P_{dis}(\epsilon)$ represents the disclosure probability under different ϵ which is determined according to the disclosure risk level. $\frac{1}{|ASA(q)|}$ and $\frac{max_\alpha}{|T(q)|}$ represent the probability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

1) Effect of k: k varies from 5 to 25 with $C = 0.5$, $m = 3$ and $K' = 50$ under 140K records. Fig.3 shows both kinds of loss increases with k because more sequences not satisfying k-anonymity causes the higher information loss. Our DPPP-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.3, 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.



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

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



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

Figure 4. Information loss vs. C ($k = 5, m = 3, K' = 50$)

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.4, 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, DPPP-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.

Compared with **KCL-Local**, **PPTD**, and **DPTD** the trajectory information loss of DPPP can be improved by up to 71.57%, 42.77% and 67.6% respectively and the frequent sequences loss can be improved by up to 69.91%, 35.79% and 60.59% respectively.

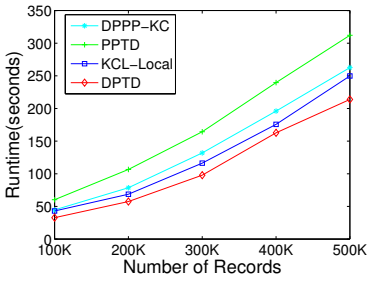


Figure 5. Run Time vs. records ($k = 20, C = 0.4$)

3) Run Time: Fig.5 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 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 DPPP-KC, it takes much time to determine the critical sequences.

VII. Conclusion

We design and implement an anonymous technique named DPPP 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 DPPP can provide higher data utility compared to peer schemes. Our privacy analysis shows that DPPP 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.

Acknowledgment

This work is supported by National Key Research and Development Project of China No.2017YFC0704100. This research is also sponsored in part by the National Natural Science Foundation of China (contract/grant numbers: 61872053, 61572413 and U1636205) and Research Grants Council, Hong Kong SAR, China, under projects 12200914, 15238116, 15222118, and C1008-16G.

References

[1] S. Amendola, R. Lodato, S. Manzari, and C. Occhiuzzi, "Rfid technology for iot-based personal healthcare in smart spaces," *Internet of Things Journal IEEE*, vol. 1, no. 2, pp. 144–152, 2014.

[2] F. T. Brito, A. C. A. Neto, C. F. Costa, A. L. Mendonça, and J. C. Machado, "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*. ACM, 2015, p. 5.

[3] R. Chen, B. Fung, and B. C. Desai, "Differentially private trajectory data publication," *arXiv preprint arXiv:1112.2020*, 2011.

[4] R. Chen, B. C. M. Fung, and B. C. Desai, "Differentially private transit data publication: a case study on the montreal transportation system," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 213–221.

[5] R. Chen, B. C. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Information Sciences*, vol. 231, pp. 83–97, 2013.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility:user movement in location-based social networks," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August*, 2011, pp. 1082–1090.

[7] A. M. Davis, A. V. Perruccio, S. Ibrahim, S. Hogg-Johnson, R. Wong, D. L. Streiner, D. E. Beaton, P. C?t, M. A. Gignac, and J. Flannery, "The trajectory of recovery and the inter-relationships of symptoms, activity and participation in the first year following total hip and knee replacement," *Osteoarthritis and Cartilage*, vol. 19, no. 12, p. 1413, 2011.

[8] Y. Dong and D. Pi, "Novel privacy-preserving algorithm based on frequent path for trajectory data publishing," *Knowledge-Based Systems*, vol. 148, pp. 55–65, 2018.

[9] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010.

[10] M. Ghasemzadeh, B. C. M. Fung, R. Chen, and A. Awasthi, "Anonymizing trajectory data for passenger flow analysis," *Transportation Research Part C Emerging Technologies*, vol. 39, no. 2, pp. 63–79, 2014.

[11] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE.* IEEE, 2017, pp. 1–9.

[12] J. Hua, Y. Gao, and S. Zhong, "Differentially private publication of general time-serial trajectory data," in *Computer Communications*, 2015, pp. 549–557.

[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[14] E. G. Komishani, M. Abadi, and F. Deldar, "Pptd: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression," *Knowledge-Based Systems*, vol. 94, pp. 43–59, 2016.

[15] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Information Sciences*, vol. 400, pp. 1–13, 2017.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L -diversity: Privacy beyond k -anonymity," *ACM Transactions on Knowledge Discovery From Data*, vol. 1, no. 1, pp. 1–12, 2007.

[17] L. Ou, Z. Qin, S. Liao, Y. Hong, and X. Jia, "Releasing correlated trajectories: Towards high utility and optimal differential privacy," *IEEE Transactions on Dependable and Secure Computing*, 2018.

[18] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos, "Local suppression and splitting techniques for privacy preserving publication of trajectories," *IEEE Trans. Knowl. Data Eng*, vol. 29, no. 7, pp. 1466–1479, 2017.

[19] N. Victor, D. Lopez, and J. H. Abawajy, "Privacy models for big data: a survey," *International Journal of Big Data Intelligence*, vol. 3, no. 1, pp. 61–75, 2016.

[20] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 591–606, 2018.

[21] Y. Xin, Z. Q. Xie, and J. Yang, "The privacy preserving method for dynamic trajectory releasing based on adaptive clustering," *Information Sciences*, vol. 378, pp. 131–143, 2017.

[22] C. Xu, W. Zheng, P. Hui, K. Zhang, and H. Liu, "Hygeia: A practical and tailored data collection platform for mobile health," in *IEEE Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE Intl Conf on Autonomic and Trusted Computing and 2015 IEEE Intl Conf on Scalable Computing and Communications and ITS Associated Workshops*, 2015, pp. 20–27.

[23] L. Yao, X. Wang, X. Wang, H. Hu, and G. Wu, "Publishing sensitive trajectory data under enhanced l-diversity model," pp. 160–169, 2019.

[24] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: how to hide a mob in a crowd?" in *EDBT 2009, International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24-26, 2009, Proceedings*, 2009, pp. 72–83.