Knowledge-Based Systems 94 (2016) 43–59

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression

Elahe Ghasemi Komishani, Mahdi Abadi*, Fatemeh Deldar

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, P.O. Box 14115-194, Tehran, Iran

ARTICLE INFO

Article history:
Received 26 January 2015
Revised 1 November 2015
Accepted 6 November 2015
Available online 14 November 2015

Keywords:
Trajectory data
Privacy preservation
Personalized privacy
Privacy attack
Generalization
Local suppression
Information loss
Disclosure risk

ABSTRACT

Trajectory data often provide useful information that can be used in real-life applications, such as traffic management, Geo-marketing, and location-based advertising. However, a trajectory database may contain detailed information about moving objects and associate them with sensitive attributes, such as disease, job, and income. Therefore, improper publishing of the trajectory database can put the privacy of moving objects at risk, especially when an adversary uses partial trajectory information as its background knowledge. The existing approaches for privacy preservation in trajectory data publishing provide the same privacy protection for all moving objects. The consequence is that some moving objects may be offered insufficient privacy protection, while some others may not require high privacy protection. In this paper, we address this problem and present PPTD, a novel approach for preserving privacy in trajectory data publishing based on the concept of personalized privacy. It aims to strike a balance between the conflicting goals of data utility and data privacy in accordance with the privacy requirements of moving objects. To the best of our knowledge, this is the first paper that combines sensitive attribute generalization and trajectory local suppression to achieve a tailored personalized privacy model for trajectory data publishing. Our experiments on two synthetic trajectory datasets suggest that PPTD is effective for preserving personalized privacy in trajectory data publishing. In particular, PPTD can significantly improve the data utility of anonymized trajectory databases when compared with previous work in the literature.

1. Introduction

With the proliferation of location-aware devices, such as RFID tag readers and GPS mobile phones, it is easy to collect the spatio-temporal traces of moving objects. Collected data, called *trajectory data* or *moving object data*, could be used in real-life applications, such as intelligent transportation systems, city traffic planning, Geo-marketing, location-based advertising, and so on [1].

Trajectory data often contain detailed information about moving objects, and for many applications, these data need to be published with sensitive attributes, such as disease, job, and income. Therefore, there is a growing concern about breaching the privacy of moving objects whose locations are easily monitored and tracked [2].

Most privacy protection principles are to bind the breach of sensitive information [3]. Removing identifying attributes, such as name and social security number, from a trajectory database before the public release, is not effective against privacy attacks, especially when an adversary uses partial trajectory information as its background knowledge.

Example 1. A hospital uses an RFID tagging system for the care of its patients, in which patient information are stored in a central trajectory database. The hospital decides to publish the trajectory database for data mining tasks. Each data record is represented as a tuple (ID, Trajectory, Disease), in which “Trajectory” is a sequence of spatio-temporal pairs. For example, the data record r_3 indicates that the patient with ID#3 has visited locations *d*, *c*, and *a* at timestamps 3, 4, and 7, respectively and has Pancreatitis.

With adequate background knowledge, an adversary could perform three types of privacy attacks on a published trajectory database:

Identity linkage attack. If a trajectory in the trajectory database is very specific, such that not many moving objects can match it, the adversary using some background knowledge may uniquely identify the data record of the target victim and, therefore, its sensitive attribute values [4–6]. For example, suppose the adversary knows that the data record of Alice is in Table 1 and that she has visited locations *c* and *d* at timestamps 4 and 5, respectively. The adversary could associate the data record r_6 with Alice and subsequently identify her disease as Diabetes, because r_6 is the only data record containing the sub-trajectory $\langle c4, d5\rangle$.

* Corresponding author. Tel.: +98 21 82884935; fax: +98 21 82884325.
E-mail address: abadi@modares.ac.ir, mahdi.abadi@gmail.com (M. Abadi).

http://dx.doi.org/10.1016/j.knosys.2015.11.007
0950-7051/© 2015 Elsevier B.V. All rights reserved.

Table 1
A trajectory database.

ID	Trajectory	Disease
1	$\langle b2, d3, c4, f6, a7, e8\rangle$	HIV
2	$\langle c4, f6, a7, e9\rangle$	SARS
3	$\langle d3, c4, a7\rangle$	Pancreatitis
4	$\langle b2, f6, a7, e8\rangle$	HIV
5	$\langle d5, f6, e9\rangle$	Flu
6	$\langle c4, d5, f6\rangle$	Diabetes
7	$\langle b2, f6, e9\rangle$	Cold

Attribute linkage attack. If a sensitive attribute value occurs frequently with some sub-trajectories, the adversary may identify it from these sub-trajectories even though cannot uniquely identify the data record of the target victim [4–6]. For example, suppose the adversary knows that Bob has visited locations *f* and *a* at timestamps 6 and 7, respectively. The adversary could infer that Bob has HIV with confidence $2/3 = 0.67,$ because among the three data records r_1, r_2, and r_4 containing the sub-trajectory $\langle f6, a7\rangle$, two of them have HIV as their sensitive attribute value.

Similarity attack. If some sensitive attribute values that are distinct but semantically similar occur frequently with some sub-trajectories, the adversary may infer sensitive information from these sub-trajectories even though cannot uniquely identify the data record of the target victim. For example, suppose the adversary knows that Carol has visited locations *f* and *e* at timestamps 6 and 9, respectively. The adversary could infer that Carol has Lung Infection with confidence 1.0, because the sensitive attribute values of the data records r_2, r_5, and r_7 containing the sub-trajectory $\langle f6, e9\rangle$ are respectively SARS, Flu, and Cold, all of which are different types of Lung Infection.

Many approaches have been proposed for preserving privacy in trajectory data publishing, but most of them do not consider different privacy requirements of different moving objects. The result is that we may be offering insufficient privacy protection to some moving objects, while providing excessive privacy control over some others, leading to an increase in information loss and disclosure risk. In this paper, we address this problem by focusing on the concept of personalized privacy [7,8], so that a moving object can specify the degree of privacy protection for its sensitive attribute values. More specifically, we acknowledge the emerging trajectory data publishing scenario, in which moving objects have different levels of privacy protection and the trajectory database needs to be published with sensitive attributes. This naturally requires preventing from three identity linkage, attribute linkage, and similarity attacks, which the latter has not been studied in previous work.

The main contributions of this paper are summarized as follows:

- We present PPTD, a novel approach for preserving personalized privacy in trajectory data publishing, which takes into consideration not only the identity and attribute linkage attacks, but also the similarity attack via trajectory data.
- To the best of our knowledge, this is the first study that combines both sensitive attribute generalization and trajectory local suppression to strike a balance between the conflicting goals of data utility and data privacy in accordance with the privacy requirements of moving objects.
- The core of PPTD is the concept of personalized privacy. PPTD is flexible enough, since it decides the minimum amount of necessary generalization and local suppression for satisfying the privacy requirements of each moving object and, hence, reduces the amount of information loss.
- We use the disclosure risk as a metric to measure the privacy breach probability of moving objects and define the affinity coefficient to evaluate the effect that the disclosure risk of a moving object takes when it shares some sub-trajectories with other moving objects.

The rest of the paper is organized as follows: Section 2 reviews some related work and Section 3 gives basic definitions. Section 4 presents the main steps of PPTD and Section 5 discusses its time complexity. The experimental results are reported in Section 6 and, finally, some concluding remarks are given in Section 7.

2. Related work

Most existing work on preserving the privacy of moving objects has been developed in the context of location-based services (LBSs) [9–11], where a trusted server is usually in charge of handling incoming requests and passing them to available service providers. Hence, the main goal is to provide the service as quickly as possible without violating the anonymity of the moving object that is requesting it; therefore, the data might be forgotten once the service is provided. In general, there are two major types of LBS-related privacy [9]: *location privacy* and *query privacy*. Location privacy aims at protecting moving objects' private information that may be disclosed from their locations and query privacy aims at protecting moving objects' private information that may be disclosed from their query terms. One of the most popular metric used for both location and query privacy is *k*-anonymity. It was first introduced in the relational database community to anonymize relational data and then became a popular privacy metric among researchers from other communities. It should be mentioned that a relational database satisfies *k*-anonymity iff each record is indistinguishable from at least $k - 1$ other records with respect to a set of quasi-identifier attributes. In the context of LBSs, a moving object is considered *k*-anonymous iff its location information sent to a service provider is made indistinguishable from that of at least $k - 1$ other moving objects or corresponds to an area where the moving object is indistinguishable from at least $k - 1$ other moving objects also present in that area [10].

As mentioned above, privacy in general has been a major topic in the context of LBSs. However, trajectory data scenarios are not in the focus of LBS privacy research. The reason is that in the context of trajectory data, anonymization is offline and data-centric, while in the context of LBSs, it is online and service-centric. In this paper, we concentrate on the context of trajectory databases, where we have a static database of moving objects' trajectory data associated with sensitive attributes such as disease, job, and income. Our aim is to anonymize the trajectory database with respect to the privacy requirements of moving objects such that both data utility and data privacy are kept high simultaneously.

Currently, there are three main categories of approaches for privacy preservation in trajectory data publishing: (1) clustering-based approaches that apply the concept of *k*-anonymity in relational databases, (2) quasi-identifier approaches that assume an adversary uses some partial knowledge of a trajectory to identify its remaining moving points or sensitive attributes, and finally (3) differentially private approaches that guarantee moving objects are protected under the definition of differential privacy.

Differential privacy [12] is defined as a property of a query answering mechanism, where the database is held by a trusted party that answers statistical queries in a differentially private way. It ensures that query answers are not substantially influenced by the presence or absence of any particular record in the database [13,14]. Recently, some approaches have been proposed that adopt the idea of differential privacy on trajectory data [15,16]. The aim of these approaches is to publish noisy statistics that are effective at supporting specific data analysis tasks, such as count query answering and frequent sequential pattern mining. On the other hand, publishing trajectories with differential privacy guarantees may not be able to provide meaningful data utility. This is due to the uncertainty (e.g., Laplace noise) introduced for achieving differential privacy. In a nutshell, differentially private approaches usually impose a guarantee on the mechanism of publishing trajectory data and so are out of the scope of this paper.

2.1. Clustering

Nergiz et al. [17] adopt the notion of k-anonymity to trajectories and propose a clustering-based approach for trajectory data anonymization. They show that releasing anonymized trajectories may lead to some privacy breaches, and therefore present a randomization-based reconstruction algorithm for releasing anonymized trajectory data. Monreale et al. [18] present a technique for trajectory data anonymization that combines the notions of spatial generalization and k-anonymity. The main idea is to anonymize trajectories by replacing exact locations by approximate ones.

Mahdavifar et al. [19] propose a greedy clustering-based approach in which trajectories are anonymized to some extent proportional to the privacy requirements of their moving objects. Although this approach aims at preserving personalized privacy in trajectory data publishing, but it is not resistance to both attribute linkage and similarity attacks.

Domingo-Ferrer and Trujillo-Rasua [20] present two trajectory anonymization methods, called SwapLocations and ReachLocations, both of which preserve original locations in the sense that the anonymized trajectories contain no perturbed or generalized locations. SwapLocations is based on the microaggregation of trajectories and permutation of locations. It guarantees k-anonymity of trajectories, but does not consider reachability constraints. This may cause that some consecutive locations in anonymized trajectories are not directly reachable in real world, making it easy for the adversary to identify fake trajectories given the road map is publicly available. Moreover, ReachLocations is only based on the permutation of locations and aims at taking reachability constraints into account. It only guarantees k-diversity of locations instead of k-anonymity of trajectories. The reason is that enforcing reachability constraints along with providing k-anonymity would result in a lot of original locations being eliminated [20].

2.2. Quasi-identifier

Terrovitis et al. [21] assume that the adversary uses some partial trajectory information as its background knowledge to infer unknown moving points. Hence, they iteratively eliminate selected moving points from the original trajectory data until a privacy constraint is satisfied. Yarovoy et al. [22] introduce a notion of k-anonymity by defining an attack graph associated with the original trajectory data and its distorted one. They consider timestamps as the quasi-identifiers and present two different algorithms, namely extreme-union and symmetric-anonymization, to build anonymization groups that provably satisfy the k-anonymity requirement.

Mohammed et al. [5] adopt a privacy model called *LKC*-privacy and develop an anonymization framework that employs global suppression to achieve *LKC*-privacy. The general intuition is to ensure that each sub-trajectory with maximum length L in a trajectory database is shared by at least K trajectory data records and the confidence of inferring any sensitive attribute value is not greater than C. Chen et al. [6] present a similar framework that supports both local and global suppressions. The aim is to preserve both instances of moving points and frequent sub-trajectories in a trajectory database.

Ghasemzadeh et al. [23] propose a simple method for achieving anonymity in trajectory databases that only thwarts the identity linkage attack while preserving the information to support effective passenger flow analysis. They first generate a probabilistic flowgraph from a trajectory database and then anonymize the database in such a way that *LK*-privacy is satisfied and any impact on the flowgraph is minimized.

Most of the aforementioned approaches do not consider different privacy requirements of moving objects. Moreover, the majority of them are not resistant to all three identity linkage, attribute linkage, and similarity attacks.

3. Notations and basic definitions

3.1. Trajectory database

A typical location-aware system generates a sequence of spatio-temporal data records of the general form (*id, l, t*), each of which indicates that a moving object having the unique identifier *id* was detected in the location *l* at time *t*.

Definition 1 (Trajectory)**.** Let O be a set of moving objects. The trajectory of a moving object $o_i \in O$, denoted by τ_i, is a sequence of spatio-temporal pairs:

$$\tau_i = \langle (l_i^1, t_i^1), (l_i^2, t_i^2), \ldots, (l_i^m, t_i^m) \rangle, \tag{1}$$

where each $(l_i^k, t_i^k) \in \tau_i$ is called a *moving point* and denoted by p_i^k.

The length of τ_i, denoted by $|\tau_i|$, is defined as the number of moving points it contains. A trajectory that contains only the first $k \le |\tau_i|$ moving points of τ_i is denoted by τ_i^k. We define a strict total order relation, $\prec$, between each two moving points $p_i^k = (l_i^k, t_i^k)$ and $p_i^{k'} = (l_i^{k'}, t_i^{k'})$ in τ_i:

$$p_i^k \prec p_i^{k'} \text{ iff } t_i^k < t_i^{k'}. \tag{2}$$

Definition 2 (Joinable trajectories)**.** Two trajectories $\tau_i = \langle (l_i^1, t_i^1), (l_i^2, t_i^2), \ldots, (l_i^m, t_i^m) \rangle$ and $\tau_j = \langle (l_j^1, t_j^1), (l_j^2, t_j^2), \ldots, (l_j^m, t_j^m) \rangle$ are said to be *joinable* iff $\tau_i^{m-1} = \tau_j^{m-1}$ and $t_i^m < t_j^m$. The joined trajectory is denoted by $\tau_i \bowtie \tau_j$:

$$\tau_i \bowtie \tau_j = \langle (l_i^1, t_i^1), (l_i^2, t_i^2), \ldots, (l_i^m, t_i^m), (l_j^m, t_j^m) \rangle. \tag{3}$$

Definition 3 (Sub-trajectory)**.** Let $\tau_i = \langle p_i^1, p_i^2, \ldots, p_i^m \rangle$ and $\tau_j = \langle p_j^1, p_j^2, \ldots, p_j^s \rangle$ be two trajectories. τ_j is said to be a *sub-trajectory* of τ_i, denoted by $\tau_j \sqsubseteq \tau_i$, if there exist integers $1 \le k_1 < \cdots < k_s \le m$ such that

$$p_j^1 = p_i^{k_1}, p_j^2 = p_i^{k_2}, \ldots, p_j^s = p_i^{k_s}. \tag{4}$$

A trajectory database may contain other attributes that are associated with trajectory data. These attributes are divided into two categories: *sensitive* and *insensitive*. If moving objects of a location-aware system are patients, sensitive attribute(s) may be their disease. Formally, a trajectory database contains a set of trajectory data records in the form of

$$r_i = \langle p_i^1, p_i^2, \ldots, p_i^m \rangle : s_i^1, s_i^2, \ldots, s_i^n : a_i^1, a_i^2, \ldots, a_i^q, \tag{5}$$

where $\langle p_i^1, p_i^2, \ldots, p_i^m \rangle$ is the trajectory, s_i^1 to s_i^n are the sensitive attributes values, and a_i^1 to a_i^q are the insensitive attributes values of a moving object. The trajectory in r_i is denoted by $\tau(r_i)$:

$$\tau(r_i) = \langle p_i^1, p_i^2, \ldots, p_i^m \rangle. \tag{6}$$

The values of each sensitive attribute are usually divided into different categories. We can use a taxonomy tree [7] to represent these values and their categories. To illustrate the concept, Fig. 1 shows a simple taxonomy tree for the sensitive attribute Disease that organizes all diseases as its leaves. Each internal node has been uniquely labeled with a name showing the category of diseases in the node's sub-tree.

In the rest of the paper, for simplicity, we assume that the trajectory database contains only a sensitive attribute and each moving object corresponds to only one trajectory data record. In this case, the sensitive attribute value of each trajectory data record r_i is denoted by $s(r_i)$.

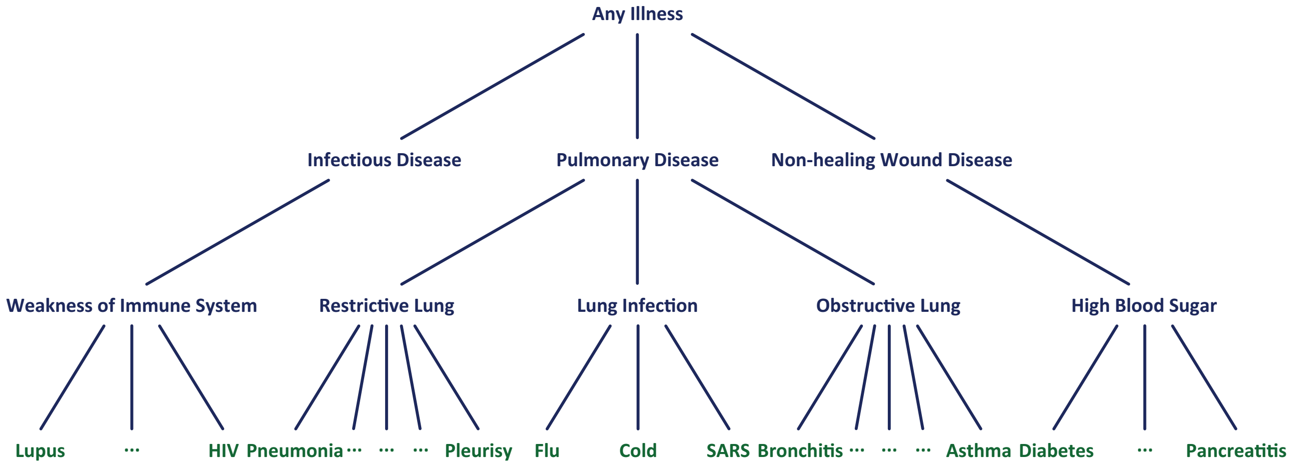

Fig. 1. A taxonomy tree for the sensitive attribute Disease.

Definition 4 (Taxonomy tree)**.** Let S be the set of all possible values of the sensitive attribute. A taxonomy tree for this attribute is a tuple $\Gamma = (V, E, \ell)$, where V is the set of nodes, E is the set of edges, and $\ell: V \rightarrow 2^S$ is a labeling function that assigns a subset of sensitive attribute values in S to each node in V. For a taxonomy tree Γ, $V(\Gamma)$, $E(\Gamma)$, and ℓ_Γ are the set of nodes, the set of edges, and the labeling function, respectively. When the taxonomy tree is clear from the context we drop the subscript Γ in notation and use ℓ instead.

Generally, there are two types of nodes in a taxonomy tree Γ: *internal nodes* and *leaf nodes*. We assume that all leaf nodes of Γ have the same depth and are at level 0. We also assume the root node of Γ is at level h. The level of each node $v_j \in V(\Gamma)$ is denoted by $\iota(v_j)$.

Let $\ell(v_j)$ be the subset of sensitive attribute values assigned to a node $v_j \in V(\Gamma)$, the size of $\ell(v_j)$ is denoted by $|\ell(v_j)|$ and called the *cardinality* of v_j. For example, $\ell(\text{Flu}) = \{\text{Flu}\}$ and $\ell(\text{Lung Infection}) = \{\text{Flu, Cold, SARS}\}$. Thus, $|\ell(\text{Flu})| = 1$ and $|\ell(\text{Lung Infection})| = 3$. It should be noted that the cardinality of all leaf nodes in $\Gamma$is always equal to one.

Definition 5 (Covering node)**.** Given a taxonomy tree Γ, a node $v_j \in V(\Gamma)$ is said to be *strictly covered* by a node $v_k \in V(\Gamma)$ iff $\ell(v_j) \subset \ell(v_k)$. In such a case, v_k is called a *covering node* of v_j. The set of all covering nodes for v_j is denoted by $c(v_j)$.

Example 2. Consider the taxonomy tree in Fig. 1. The level of the leaf nodes, e.g., Flu, Cold, and SARS, is *zero*. The level of the internal nodes Weakness of Immune System, Lung Infection, and High Blood Sugar is *one* and the level of the internal nodes Infectious Disease, Pulmonary Disease, and Non-healing Wound Disease is *two*. Also, Lung Infection and Pulmonary Disease are covering nodes for Flu, Cold, and SARS.

3.2. Privacy level

Different moving objects may have different privacy requirements. Therefore, we assign a privacy level to each moving object to represent its privacy requirements. Let O be a set of moving objects, T be a trajectory database, Γ be a taxonomy tree for the sensitive attribute of T, and $L = \{\vartheta_0, \vartheta_1, \ldots, \vartheta_{h-1}\}$ be a totally ordered set of privacy levels, where h is the level of the root node of Γ. We define $\rho: T \rightarrow O$ to be a total function that assigns each trajectory data record in T to a moving object in O and define $\theta: O \rightarrow L \cup \{\epsilon\}$ to be a total function that assigns each moving object in O to a privacy level in $L \cup \{\epsilon\}$. Therefore, a privacy level is assigned to each trajectory data record in T. It should be noted that if a moving object does not need any privacy protection, its privacy level is defined to be equal to ϵ.

Definition 6 (Guarding node)**.** Given a trajectory database T and a taxonomy tree Γ, a node $v_j \in V(\Gamma)$ is called the *guarding node* for a trajectory data record $r_i \in T$, denoted by $v_j = g(r_i)$, iff $s(r_i) \in \ell(v_j)$ and $\iota(v_j) = \theta(\rho(r_i))$.

Table 2
A trajectory database with the privacy levels of moving objects.

ID	Privacy level	Trajectory	Disease
1	Low	$\langle b2, d3, c4, f6, a7, e8\rangle$	HIV
2	Medium	$\langle c4, f6, a7, e9\rangle$	SARS
3	Low	$\langle d3, c4, a7\rangle$	Pancreatitis
4	High	$\langle b2, f6, a7, e8\rangle$	HIV
5	Medium	$\langle d5, f6, e9\rangle$	Flu
6	Low	$\langle c4, d5, f6\rangle$	Diabetes
7	No Privacy	$\langle b2, f6, e9\rangle$	Cold

The guarding node of a trajectory data record r_i, $g(r_i)$, indicates that the adversary should not be able to associate the moving object $\rho(r_i)$ with any sensitive attribute value in $\ell(g(r_i))$. In general, the higher the level of a guarding node in the taxonomy tree, the higher the privacy protection should be guaranteed. Note that guarding nodes depend entirely on personalized privacy and are not determined by the sensitive attribute values of trajectory data records.

Example 3. Consider the trajectory database in Table 2. Each trajectory data record has one of the three privacy levels of Low, Medium, or High, which are respectively related to the node levels of *zero, one*, or *two* in the taxonomy tree of Fig. 1. Also, one of the trajectory data records does not need any privacy protection, which is indicated by No Privacy. The sensitive attribute value and privacy level of the trajectory data record r_2 are SARS and Medium, respectively. Therefore, the node Lung Infection is the guarding node for r_2. It indicates that the adversary should not be able to infer, with high confidence, that the moving object $\rho(r_2)$ suffers from a disease under Lung Infection in the taxonomy tree.

3.3. Adversary's background knowledge

In the real world, it is very unlikely that the adversary can identify all moving points of a moving object and use them as its background knowledge, because it must spend a significant amount of time and effort to collect each of them. Therefore, we assume that the adversary's background knowledge for a given moving object is bounded by at most δ moving points.

Suppose a trajectory database T is to be published for data mining purposes. Explicit identifiers, e.g., name and ID, have been removed. One recipient, the adversary, employs his relevant background knowledge to identify the trajectory data record or sensitive attribute value of a victim from among those in T.

Given a set of moving objects O, let r_i be the trajectory data record of a victim $\rho(r_i) \in O$. The adversary's background knowledge about this victim, denoted by ξ_i, contains at most δ moving points:

$$\xi_i = \langle p_i^1, p_i^2, \ldots, p_i^l \rangle, \quad l \leq \delta, \tag{7}$$

where δ is the maximum length of the adversary's background knowledge.

Using the background knowledge ξ_i, the adversary can identify a set $T(\xi_i)$ of trajectory data records in T matching ξ_i:

$$T(\xi_i) = \{r_k \in T \mid \xi_i \sqsubseteq \tau(r_k)\}. \tag{8}$$

It should be noted that a trajectory data record $r_k \in T$ matches ξ_i iff ξ_i is a sub-trajectory of the trajectory $\tau(r_k)$. For example, in Table 2, if $\xi_i = \langle d3, c4 \rangle$, then $T(\xi_i) = \{r_1, r_3\}$.

3.4. Privacy attacks

The ability to link background knowledge to the published trajectory data, which enables the adversary to associate moving objects to their sensitive attribute values, is known as *linkage attack*. A linkage attack can be conducted by a sub-trajectory known by the adversary that serves as a quasi-identifier.

Given a trajectory database T and a background knowledge ξ_i, the adversary could utilize $T(\xi_i)$ to perform three types of linkage attacks: identity linkage [4–6], attribute linkage [4–6], and similarity attacks. Example 1 illustrates these attacks.

Definition 7 (Identity linkage attack)**.** Given a trajectory database T, a trajectory data record $r_i \in T$, and a background knowledge $\xi_i \sqsubseteq \tau(r_i)$, if the size of $T(\xi_i)$, denoted by $|T(\xi_i)|$, is small, then the adversary may identify r_i and, therefore, $s(r_i)$.

Definition 8 (Attribute linkage attack)**.** Given a set of moving objects O, a trajectory database T, a trajectory data record $r_i \in T$, and a background knowledge $\xi_i \sqsubseteq \tau(r_i)$, the adversary may identify $s(r_i)$ with confidence $P_c(s(r_i) \mid \xi_i)$:

$$P_c(s(r_i) \mid \xi_i) = |T(s(r_i)) \cap T(\xi_i)| / |T(\xi_i)|, \tag{9}$$

where $T(s(r_i))$ is the set of trajectory data records in T whose sensitive attribute value is equal to $s(r_i)$. $P_c(s(r_i) \mid \xi_i)$ is the percentage of trajectory data records in $T(\xi_i)$ containing $s(r_i)$. The privacy of the moving object $\rho(r_i) \in O$ is at risk if $P_c(s(r_i) \mid \xi_i) > \sigma$, where σ is a parameter specifying the amount of privacy disclosure and is called the *privacy breach threshold*.

Definition 9 (Similarity attack)**.** Given a trajectory database T, a trajectory data record $r_i \in T$, and a background knowledge $\xi_i \sqsubseteq \tau(r_i)$, the adversary may identify $g(r_i)$ iff $s(r_k) \in \ell(g(r_i))$ for all $r_k \in T(\xi_i)$.

3.5. Sensitive attribute generalization

To preserve the privacy of moving objects, we generalize the sensitive attribute value of each trajectory data record, if necessary. This generalization is done such that a desirable balance between information loss and privacy disclosure is achieved.

Definition 10 (Generalized value of sensitive attribute)**.** Given a trajectory data record r_i, let S be the set of all possible sensitive attribute values. A subset $S_i \subseteq S$ is said to be an *SA-generalized value* of r_i and denoted by $\Omega(r_i)$ iff $s(r_i) \in S_i$.

Definition 11 (Twin node)**.** Given a taxonomy tree Γ and a trajectory data record r_i, let $\Omega(r_i)$ be the SA-generalized value of r_i. A node $v_j \in V(\Gamma)$ is said to be a *twin node* for r_i and denoted by $\eta(r_i)$ iff $\ell(v_j) = \Omega(r_i)$.

Definition 12 (Generalization depth)**.** Given a trajectory data record r_i, let $\iota(g(r_i))$ and $\iota(\eta(r_i))$ be the level of the guarding node and the twin node of r_i, respectively. The generalization depth of r_i is denoted by $\zeta(r_i)$ and defined as

$$\zeta(r_i) = \begin{cases} \iota(\eta(r_i)) - \iota(g(r_i)) & \iota(\eta(r_i)) > \iota(g(r_i)), \\ 0 & \text{otherwise}. \end{cases} \tag{10}$$

Table 3
An SA-generalized trajectory database.

ID	Privacy level	Trajectory	Disease
1	Low	$\langle b2, d3, c4, f6, a7, e8 \rangle$	Weakness of Immune
2	Medium	$\langle c4, f6, a7, e9 \rangle$	Pulmonary Disease
3	Low	$\langle d3, c4, a7 \rangle$	Pancreatitis
4	High	$\langle b2, f6, a7, e8 \rangle$	Any Illness
5	Medium	$\langle d5, f6, e9 \rangle$	Pulmonary Disease
6	Low	$\langle c4, d5, f6 \rangle$	High Blood Sugar
7	No Privacy	$\langle b2, f6, e9 \rangle$	Cold

We apply the sensitive attribute generalization such that the generalization depth of each trajectory data record is less than or equal to a pre-specified maximum generalization depth ζ_{max}.

Given a set of moving objects O, a trajectory database T, and a trajectory data record $r_i \in T$, the privacy of the moving object $\rho(r_i) \in O$ is breached when the adversary can associate this object with one of the sensitive attribute values in the set $\ell(g(r_i))$, where $g(r_i)$ is the guarding node of r_i. For each trajectory data record $r_k \in T(\xi_i)$, we define $P(\ell(g(r_i)) \mid \ell(\eta(r_k)))$ as the probability of breaching the privacy of $\rho(r_i)$ given the SA-generalized value of r_k:

$$P(\ell(g(r_i)) \mid \ell(\eta(r_k))) = \frac{|\ell(g(r_i)) \cap \ell(\eta(r_k))|}{|\ell(\eta(r_k))|}. \tag{11}$$

We can identify three possible situations for it as

$$P(\ell(g(r_i)) \mid \ell(\eta(r_k))) = \begin{cases} 1 & g(r_i) \in c(\eta(r_k)) \vee g(r_i) = \eta(r_k), \\ \alpha & \eta(r_k) \in c(g(r_i)), \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where α is a real value in the range (0,1). If $g(r_i)$ and $\eta(r_k)$ do not cover each other and are not the same, then $\ell(g(r_i)) \cap \ell(\eta(r_k)) = \emptyset$, and so $P(\ell(g(r_i)) \mid \ell(\eta(r_k))) = 0$. Otherwise, we distinguish three scenarios: (1) $g(r_i)$ is a covering node of $\eta(r_k)$, (2) it is covered by $\eta(r_k)$, or (3) they are the same. In the first and third scenarios, $\ell(g(r_i)) \cap \ell(\eta(r_k)) = \ell(\eta(r_k))$, and therefore $P(\ell(g(r_i)) \mid \ell(\eta(r_k))) = 1$. In the second scenario, $\ell(g(r_i)) \cap \ell(\eta(r_k)) = \ell(g(r_i))$ and $|\ell(g(r_i))| < |\ell(\eta(r_k))|$. Therefore, $P(\ell(g(r_i)) \mid \ell(\eta(r_k)))$ is equal to a real value between 0 and 1.

Having a background knowledge $\xi_i \sqsubseteq \tau(r_i)$, the adversary first identifies $T(\xi_i)$ according to (8) and then tries to infer $\ell(g(r_i))$ using the SA-generalized values of all trajectory data records in $T(\xi_i)$. Hence, the probability of privacy breach for $\rho(r_i)$, and thus r_i, given ξ_i is calculated as

$$P_b(\rho(r_i) \mid \xi_i) = \frac{1}{|T(\xi_i)|} \sum_{r_k \in T(\xi_i)} P(\ell(g(r_i)) \mid \ell(\eta(r_k))), \tag{13}$$

where $g(r_i)$ is the guarding node of r_i and $\eta(r_k)$ is the twin node of a trajectory data record $r_k \in T(\xi_i)$.

Example 4. Consider the SA-generalized trajectory database T in Table 3 with $\zeta_{max} = 1$. Assuming $\xi_i = \langle a7 \rangle$, we obtain $T(\xi_i) = \{r_1, r_2, r_3, r_4\}$. In addition, we know that $g(r_3) = \text{Pancreatitis}$, $\eta(r_1) = \text{Weakness of Immune}$, $\eta(r_2) = \text{Pulmonary Disease}$, $\eta(r_3) = \text{Pancreatitis}$, and $\eta(r_4) = \text{Any Illness}$. Therefore, according to (12) and (13), we have

$$P(\ell(g(r_3)) \mid \ell(\eta(r_1))) = P(\ell(g(r_3)) \mid \ell(\eta(r_2))) = 0,$$
$$P(\ell(g(r_3)) \mid \ell(\eta(r_3))) = 1,$$
$$P(\ell(g(r_3)) \mid \ell(\eta(r_4))) = 0.05.$$

Therefore, $P_b(\rho(r_i) \mid \xi_i) = \frac{1}{4}(0 + 0 + 1 + 0.05) = 0.26$. Note that in this example we assumed that the taxonomy tree has 19 leaf nodes.

3.6. Trajectory local suppression

After the sensitive attribute generalization, the personalized privacy of some moving objects may still be breached. Therefore, we should eliminate a number of moving points from trajectory data records such that we have no any moving object with high privacy breach probability. To this end, we can apply a local or global suppression on trajectory data records.

Let T be a trajectory database. The global suppression eliminates a moving point from all trajectory data records in T if it makes the privacy breach probability of some trajectory data records so high, while the local suppression eliminates the moving point only from trajectory data records with high privacy breach probability and leaves others intact. Hence, the local suppression preserves better data utility in comparison with the global suppression. As a result, we apply it on trajectory data records.

Definition 13 (Critical sub-trajectory)**.** Given a trajectory database T, a non-empty sub-trajectory τ_j is called *critical* iff there is a trajectory data record $r_i \in T$ such that $P_b(\rho(r_i) \mid \tau_j) > \sigma$, where σ is the privacy breach threshold. In this case, r_i is called a *critical trajectory data record*.

Example 5. Consider the SA-generalized trajectory database in Table 3. Given $\delta = 2$ and $\sigma = 0.50$, the sub-trajectory $\tau_j = \langle b2, e8\rangle$ is critical and therefore r_4 is a critical trajectory data record, because $P_b(\rho(r_4) \mid \tau_j) > \sigma$. However, the sub-trajectory $\tau_k = \langle f6, a7, e8\rangle$ is not critical even though $P_b(\rho(r_4) \mid \tau_k) > \sigma$, because $|\tau_k| > \delta$.

Generally, eliminating moving points from critical sub-trajectories in T increases information loss and decreases disclosure risk. Hence, we define a suppression score for each moving point that considers both information loss and disclosure risk in accordance with the privacy levels of moving objects. It guides us to find a sub-optimal balance between data utility and personalized privacy preservation.

Definition 14 (Personalized suppression score)**.** Given a trajectory database T, let $\mathcal{T}_c$ be a set of critical sub-trajectories and $\tau_j \in \mathcal{T}_c$ be a critical sub-trajectory. The personalized suppression score of a moving point $p_j^k \in \tau_j$ with respect to $\mathcal{T}_c$ is denoted by $\varphi(p_j^k, \tau_j, \mathcal{T}_c)$ and calculated as

$$\varphi(p_j^k, \tau_j, \mathcal{T}_c) = \frac{|\mathcal{T}_c(p_j^k)|}{|T(\tau_j)|} \sum_{r_i \in T(\tau_j)} \theta(\rho(r_i)), \tag{14}$$

where $\mathcal{T}_c(p_j^k)$ is the set of sub-trajectories in $\mathcal{T}_c$ containing p_j^k and $T(\tau_j)$ is the set of trajectory data records in T matching τ_j. Also, the personalized suppression score of τ_j with respect to $\mathcal{T}_c$, denoted by $\psi(\tau_j, \mathcal{T}_c)$, is calculated as

$$\psi(\tau_j, \mathcal{T}_c) = \max_{p_j^k \in \tau_j} \varphi(p_j^k, \tau_j, \mathcal{T}_c). \tag{15}$$

4. PPTD

To achieve personalized privacy preservation in trajectory data publishing, we present PPTD, an approach that conducts both the sensitive attribute generalization and the trajectory local suppression of moving points to prevent personalized privacy attacks by the adversary, while preserving as much data utility as possible.

4.1. Sensitive attribute generalization

The aim of the sensitive attribute generalization is to identify critical trajectory data records and generalize their sensitive attribute values with respect to a pre-specified maximum generalization depth ζ_{max}. To the best of our knowledge, none of the existing approaches for privacy preservation in trajectory data publishing apply the sensitive attribute generalization. Although this results in less precise sensitive attribute values but maintains more information about trajectories. For example, if we publish the trajectory database in Table 4, there is a low probability that the adversary could infer that Diabetes is the real disease of the patient with ID#6, despite the fact that we have not eliminated any moving point from r_6.

Table 4
An anonymized trajectory database.

ID	Privacy level	Trajectory	Disease
1	Low	$\langle b2, d3, c4, f6, a7, e8\rangle$	Weakness of Immune
2	Medium	$\langle c4, f6, a7, e9\rangle$	Pulmonary Disease
3	Low	$\langle d3, c4, a7\rangle$	Pancreatitis
4	High	$\langle f6, a7\rangle$	Any Illness
5	Medium	$\langle d5, f6, e9\rangle$	Pulmonary Disease
6	Low	$\langle c4, d5, f6\rangle$	High Blood Sugar
7	No Privacy	$\langle b2, f6, e9\rangle$	Cold

Lemma 1. Given a trajectory database T and two arbitrary trajectory data records $r_{i_1}, r_{i_2} \in T$, let τ_j be a sub-trajectory of both $\tau(r_{i_1})$ and $\tau(r_{i_2})$. If the guarding node of r_{i_1} is a covering node for the guarding node of r_{i_2}, i.e., $\ell(g(r_{i_2})) \subset \ell(g(r_{i_1}))$, then $P_b(\rho(r_{i_2}) \mid \tau_j)$ is less than or equal to $P_b(\rho(r_{i_1}) \mid \tau_j)$ regardless of whether the sensitive attribute generalization is applied or not.

Proof. Let $\eta(r_k)$ be the twin node of a trajectory data record $r_k \in T(\tau_j)$. Because $g(r_{i_1})$ is a covering node for $g(r_{i_2})$, therefore $|\ell(g(r_{i_2})) \cap \ell(\eta(r_k))|$ is less than or equal to $|\ell(g(r_{i_1})) \cap \ell(\eta(r_k))|$ and so, according to (13), $P_b(\rho(r_{i_2}) \mid \tau_j)$ is less than or equal to $P_b(\rho(r_{i_1}) \mid \tau_j)$. □

When searching for critical trajectory data records, we can avoid calculating the privacy breach probabilities of the trajectory data records like r_{i_2} in Lemma 1, because they will be adequately protected against privacy attacks once the privacy of the other trajectory data records is preserved.

Algorithm 1 shows the pseudo-code of SAGTD that identifies critical trajectory data records in a given trajectory database and generalizes their sensitive attribute values. The algorithm takes an original trajectory database T as input and returns an SA-generalized trajectory database $T^{\mathcal{S}}$ as output. Let δ and σ be the background knowledge and privacy breach thresholds, respectively. SAGTD first applies the STR algorithm on T to generate the set $\mathcal{A}^{\delta}$ of sub-trajectories with maximum length δ (Line 1). Then, for each sub-trajectory $\tau_j \in \mathcal{A}^{\delta}$, it finds a subset $\mathcal{B}_j$ of trajectory data records in $T(\tau_j)$, whose guarding node is not covered by the guarding node of any other trajectory data record in $T(\tau_j)$ (Line 3) and makes the set $\mathcal{C}_j$ of critical trajectory data

Algorithm 1 SAGTD.

```
input:
    $T$: Trajectory database
output:
    $T^{\mathcal{S}}$: SA-generalized trajectory database

 1: $\mathcal{A}^{\delta} := \mathrm{STR}(T)$
 2: for each $\tau_j \in \mathcal{A}^{\delta}$ do
 3:    $\mathcal{B}_j := \{r_k \in T(\tau_j) \mid \ell(g(r_k)) \not\subset \ell(g(r_i))$ for all $r_i \in T(\tau_j)\}$
 4:    $\mathcal{C}_j := \{r_k \in \mathcal{B}_j \mid P_b(\rho(r_k) \mid \tau_j) > \sigma\}$
 5:    if $\mathcal{C}_j \neq \emptyset$ then
 6:        $T := (T - \mathcal{C}_j) \cup \mathrm{SAG}(\tau_j, \mathcal{C}_j)$
 7:    end if
 8: end for
 9: $T^{\mathcal{S}} := T$
10: return $T^{\mathcal{S}}$
```

Algorithm 2 STR.

```
Input:
    $T$: Trajectory database
Output:
    $\mathcal{A}^{\delta}$: Set of sub-trajectories
 1: $\mathcal{A}^{\delta} := \emptyset$
 2: $\mathcal{A}_1 := \{\tau_j \mid \tau_j \sqsubseteq \tau(r_i)$ for some $r_i \in T \wedge |\tau_j| = 1\}$
 3: for each $\tau_j \in \mathcal{A}_1$ do
 4:     Compute $T(\tau_j)$ using (8)
 5:     $\mathcal{A}^{\delta} := \mathcal{A}^{\delta} \cup \{\tau_j\}$
 6: end for
 7: $i := 1$
 8: while $i \leq \delta$ and $\mathcal{A}_i \neq \emptyset$ do
 9:     $\mathcal{A}_{i+1} := \emptyset$
10:     for $j := 1$ to $|\mathcal{A}_i|$ do
11:         for $k := j + 1$ to $|\mathcal{A}_i|$ do
12:             if $\tau_j^{i-1} = \tau_k^{i-1}$ and $T(\tau_j) \cap T(\tau_k) \neq \emptyset$ then
13:                 $T(\tau_j \bowtie \tau_k) := T(\tau_j) \cap T(\tau_k)$
14:                 $\mathcal{A}_{i+1} := \mathcal{A}_{i+1} \cup \{\tau_j \bowtie \tau_k\}$
15:                 $\mathcal{A}^{\delta} := \mathcal{A}^{\delta} \cup \{\tau_j \bowtie \tau_k\}$
16:             end if
17:         end for
18:     end for
19:     $i := i + 1$
20: end while
21: return $\mathcal{A}^{\delta}$
```

records in $\mathcal{B}_j$ (Line 4). If $\mathcal{C}_j$ is non-empty, it applies the SAG algorithm to generalize the sensitive attribute values of trajectory data records in $\mathcal{C}_j$ (Lines 5–7).

Example 6. Consider the trajectory database in Table 2 with $\delta = 2$ and $\sigma = 0.50$. The set of critical trajectory data records for the sub-trajectory $\langle b2, a7\rangle$ is $\{r_4\}$.

The pseudo-code of STR is shown in Algorithm 2. The algorithm takes a trajectory database T as input and returns the set $\mathcal{A}^{\delta}$ of sub-trajectories in T whose length is at most δ as output, where δ is the background knowledge threshold. Let $\mathcal{A}_i$ be an ordered set of sub-trajectories of length i. STR first initializes $\mathcal{A}^{\delta}$ to the empty set and $\mathcal{A}_1$ to the set of all sub-trajectories of length one (Lines 1–2). It then computes a subset $T(\tau_j) \subseteq T$ for each sub-trajectory $\tau_j \in \mathcal{A}_1$ and adds τ_j to $\mathcal{A}^{\delta}$ (Lines 3–6). Moreover, in each iteration of nested loops (Lines 10–18), if each two sub-trajectories $\tau_j, \tau_k \in \mathcal{A}_i$ are joinable and the intersection $T(\tau_j)$ and $T(\tau_k)$ is non-empty, it adds the joined sub-trajectory $\tau_j \bowtie \tau_k$ to $\mathcal{A}_{i+1}$ and $\mathcal{A}^{\delta}$ (Lines 12–16). The above steps are repeated until i is greater than δ or $\mathcal{A}_i$ is empty (Lines 8–20).

Example 7. The sets $\mathcal{A}_1$ and $\mathcal{A}_2$ of sub-trajectories of length one and two in Table 2 are respectively as follows:

$$\mathcal{A}_1 = \{b2, d3, c4, d5, f6, a7, e8, e9\}$$
$$\mathcal{A}_2 = \{b2d3, b2c4, b2f6, b2a7, b2e8, b2e9, d3c4, d3f6, d3a7, d3e8, c4d5, c4f6, c4a7, c4e8, c4e9, d5f6, d5e9, f6a7, f6e8, f6e9, a7e8, a7e9\}$$

Lemma 2. If r_i is a critical trajectory data record with respect to a background knowledge $\xi_i \sqsubseteq \tau(r_i)$, i.e., $P_b(\rho(r_i) \,|\, \xi_i) > \sigma$, then after applying the sensitive attribute generalization, $\Omega(r_i)$ must be set to the label of a covering node in the set $c(g(r_i))$, where $\Omega(r_i)$ is the SA-generalized value of r_i.

Proof. Obviously, after applying the sensitive attribute generalization, $P_b(\rho(r_i) \,|\, \xi_i)$ should decrease. Suppose, on the contrary, that $\Omega(r_i)$ is not set to the label of a covering node in the set $c(g(r_i))$. Therefore, $\eta(r_i)$ is covered by $g(r_i)$ and $|\ell(g(r_i)) \cap \ell(\eta(r_i))|$ is equal to $|\ell(\eta(r_i))|$. Thus, according to (13), $P_b(\rho(r_i) \,|\, \xi_i)$ remains the same before and after the sensitive attribute generalization is applied, leading to a contradiction. □

Algorithm 3 SAG.

```
Input:
    $\tau_j$: Trajectory
    $\mathcal{C}_j$: Set of trajectory data records
Output:
    $\mathcal{S}_j$: Set of SA-generalized trajectory data records
 1: $\mathcal{S}_j := \emptyset$
 2: for each $r_i \in \mathcal{C}_j$ do
 3:     if $\ell(\eta(r_i)) \subset \ell(g(r_i))$ or $\ell(\eta(r_i)) = \ell(g(r_i))$ then
 4:         $\Omega(r_i) := \ell(p(g(r_i)))$
 5:         if $P_b(\rho(r_i) \mid \tau_j) \leq \sigma$ then
 6:             $\mathcal{D}_j := \{r_k \in \mathcal{C}_j \mid \ell(g(r_k)) = \ell(g(r_i))\}$
 7:             $\mathcal{C}_j := \mathcal{C}_j - \mathcal{D}_j$
 8:             $\mathcal{S}_j := \mathcal{S}_j \cup \mathcal{D}_j$
 9:         end if
10:     end if
11: end for
12: while $\mathcal{C}_j \neq \emptyset$ do
13:     for each $r_i \in \mathcal{C}_j$ do
14:         if $\iota(\eta(r_i)) - \iota(g(r_i)) \geq \zeta_{max}$ or $\iota(\eta(r_i)) = h$ then
15:             $\mathcal{C}_j := \mathcal{C}_j - \{r_i\}$
16:             $\mathcal{S}_j := \mathcal{S}_j \cup \{r_i\}$
17:         else if $P_b(\rho(r_i) \mid \tau_j) \leq \sigma$ then
18:             $\mathcal{D}_j := \{r_k \in \mathcal{C}_j \mid \ell(g(r_k)) = \ell(g(r_i))\}$
19:             $\mathcal{C}_j := \mathcal{C}_j - \mathcal{D}_j$
20:             $\mathcal{S}_j := \mathcal{S}_j \cup \mathcal{D}_j$
21:         else
22:             $\Omega(r_i) := \ell(p(\eta(r_i)))$
23:         end if
24:     end for
25: end while
26: return $\mathcal{S}_j$
```

The pseudo-code of SAG is shown in Algorithm 3. The algorithm takes a trajectory τ_j and a set $\mathcal{C}_j$ of critical trajectory data records as input and returns a set $\mathcal{S}_j$ of SA-generalized trajectory data records as output. Let σ and ζ_{max} be the privacy breach threshold and the maximum generalization depth, respectively. SAG first initializes $\mathcal{S}_j$ to the empty set (Line 1). Then, for each trajectory data record $r_i \in \mathcal{C}_j$, if the guarding node of r_i, $g(r_i)$, is a covering node for its twin node, $\eta(r_i)$, or these nodes are the same, it sets $\Omega(r_i)$ to $\ell(p(g(r_i)))$, where $p(g(r_i))$ is the parent of $g(r_i)$ (Lines 3–4). After that, SAG checks whether r_i is non-critical with respect to the threshold σ and if so, it selects a subset $\mathcal{D}_j \subseteq \mathcal{C}_j$ of critical trajectory data records whose guarding node is equal to $g(r_i)$, removes all trajectory data records in $\mathcal{D}_j$ from $\mathcal{C}_j$, and adds them to $\mathcal{S}_j$ (Lines 5–9). The rationale behind this is that all trajectory data records in $\mathcal{D}_j$ have the same guarding node as r_i and so they have the same privacy breach probability as r_i. As a result, after generalizing the sensitive attribute of r_i, if the probability of privacy breach for $\rho(r_i)$, $P_b(\rho(r_i) \,|\, \tau_j)$, becomes less than or equal to σ, the trajectory data records in $\mathcal{D}_j$ are no longer critical and thus they must be removed from $\mathcal{C}_j$. Next, SAG repeats the following steps until $\mathcal{C}_j$ is empty (Lines 12–25): for each trajectory data record $r_i \in \mathcal{C}_j$, if the difference between the levels of $\eta(r_i)$ and $g(r_i)$ is greater than or equal to the threshold ζ_{max} or $\eta(r_i)$ is the root node of Γ, it removes r_i from $\mathcal{C}_j$ and adds to $\mathcal{S}_j$ (Lines 14–16). Otherwise if r_i is not a critical trajectory data record, it selects a subset $\mathcal{D}_j \subseteq \mathcal{C}_j$ of critical trajectory data records whose guarding node is equal to $g(r_i)$, removes all trajectory data records in $\mathcal{D}_j$ from $\mathcal{C}_j$, and adds them to $\mathcal{S}_j$ (Lines 17–20).

Algorithm 4 MPSTD.

```
Input:
    $T^{\mathcal{S}}$: SA-generalized trajectory database
    $\mathcal{A}^{\delta}$: Set of sub-trajectories
Output:
    $T^{\mathcal{G}}$: Anonymized trajectory database

1:  $\mathcal{T}_c := \emptyset$
2:  for each $\tau_j \in \mathcal{A}^{\delta}$ do
3:      $\mathcal{B}_j := \{r_k \in T^{\mathcal{S}}(\tau_j) \mid \ell(g(r_k)) \not\subset \ell(g(r_i))$ for all $r_i \in T^{\mathcal{S}}(\tau_j)\}$
4:      $\mathcal{C}_j := \{r_k \in \mathcal{B}_j \mid P_b(\rho(r_k) \mid \tau_j) > \sigma\}$
5:      if $\mathcal{C}_j \neq \emptyset$ then
6:          $\mathcal{T}_c := \mathcal{T}_c \cup \{\tau_j\}$
7:      end if
8:  end for
9:  while $\mathcal{T}_c \neq \emptyset$ do
10:     $\tau_z := \arg\max_{\tau_j \in \mathcal{T}_c} \psi(\tau_j, \mathcal{T}_c)$
11:     $\mathcal{C}_z := \{r_k \in T^{\mathcal{S}}(\tau_z) \mid P_b(\rho(r_k) \mid \tau_z) > \sigma\}$
12:     $p_z^q := \arg\max_{p_z^k \in \tau_z} \varphi(p_z^k, \tau_z, \mathcal{T}_c)$
13:     $\mathcal{D}_z := \emptyset$
14:     while $\mathcal{C}_z \neq \emptyset$ do
15:         $r_i := \arg\max_{r_k \in \mathcal{C}_z} \theta(\rho(r_k))$
16:         $\mathcal{D}_z := \mathcal{D}_z \cup \{r_i\}$
17:         $T^{\mathcal{S}} := T^{\mathcal{S}} - \{r_i\}$
18:         $\tau(r_i) := \tau(r_i) - \langle p_z^q \rangle$
19:         $T^{\mathcal{S}} := T^{\mathcal{S}} \cup \{r_i\}$
20:         $\mathcal{C}_z := \{r_k \in T^{\mathcal{S}}(\tau_z) \mid P_b(\rho(r_k) \mid \tau_z) > \sigma\}$
21:     end while
22:     for each $r_i \in \mathcal{D}_z$ do
23:         $\mathcal{T}_c := \mathcal{T}_c \cup \mathrm{MCST}(T^{\mathcal{S}}, \tau(r_i), p_z^q)$
24:     end for
25:     $\mathcal{T}_c := \mathcal{T}_c - \{\tau_z\}$
26: end while
27: $T^{\mathcal{G}} := T^{\mathcal{S}}$
28: return $T^{\mathcal{G}}$
```

Otherwise, it sets $\Omega(r_i)$ to $\ell(p(\eta(r_i)))$, where $p(\eta(r_i))$ is the parent of $\eta(r_i)$ (Lines 21–22).

4.2. Trajectory local suppression

After the sensitive attribute generalization, some trajectory data records in $T^{\mathcal{S}}$ may still be critical. Therefore, we apply the MPSTD algorithm to identify the remaining critical trajectory data records and to eliminate a number of moving points from them, in such a way that there is no critical trajectory data record in the anonymized trajectory database and the amount of information loss is minimized.

We can identify all personalized privacy attacks by generating all critical sub-trajectories. Therefore, it is sufficient that we remove critical sub-trajectories with respect to personalized privacy, because all possible ways for the identity linkage, attribute linkage, and similarity attacks are destroyed.

Algorithm 4 shows the pseudo-code of MPSTD. The algorithm takes an SA-generalized trajectory database $T^{\mathcal{S}}$ as input and returns an anonymized trajectory database $T^{\mathcal{G}}$ as output. Let δ be the background knowledge threshold, σ be the privacy breach threshold, and $\mathcal{A}^{\delta}$ be the set of sub-trajectories in $T^{\mathcal{S}}$ with maximum length δ. MPSTD first initializes the set $\mathcal{T}_c$ of critical sub-trajectories to the empty set (Line 1). Then, for each sub-trajectory $\tau_j \in \mathcal{A}^{\delta}$, it finds a subset $\mathcal{B}_j$ of trajectory data records in $T^{\mathcal{S}}(\tau_j)$ whose guarding node is not covered by the guarding node of any other trajectory data record in $T^{\mathcal{S}}(\tau_j)$ (Line 3) and makes the set $\mathcal{C}_j$ of critical trajectory data records in $\mathcal{B}_j$ (Line 4). If $\mathcal{C}_j$ is non-empty, MPSTD adds τ_j to $\mathcal{T}_c$ (Lines 5–7) and repeats the following steps until $\mathcal{T}_c$ is empty (Lines 9–26): it finds a sub-trajectory $\tau_z \in \mathcal{T}_c$ that maximizes the personalized suppression score $\psi(\tau_z, \mathcal{T}_c)$ and makes the set $\mathcal{C}_z$ of critical trajectory data records in $T^{\mathcal{S}}(\tau_z)$ (Lines 10–11). It then finds a moving point $p_z^q \in \tau_z$ maximizing the personalized suppression score $\varphi(p_z^q, \tau_z, \mathcal{T}_c)$ (Line 12). It next selects r_i with the maximum privacy level $\theta(\rho(r_i))$ from critical trajectory data records in $\mathcal{C}_z$ and adds r_i to the set $\mathcal{D}_z$ (Lines 15–16). It subsequently eliminates p_z^q from $\tau(r_i)$ and again makes the set $\mathcal{C}_z$ of critical trajectory data records in $T^{\mathcal{S}}(\tau_z)$ (Lines 17–20). These steps are repeated until $\mathcal{C}_z$ is empty (Lines 14–21). Eliminating p_z^q from trajectory data records may result in the generation of new critical sub-trajectories. Hence, MPSTD identifies these sub-trajectories using the MCST algorithm and adds them to $\mathcal{T}_c$ (Lines 22–24).

Algorithm 5 MCST.

```
Input:
    $T^{\mathcal{S}}$: SA-generalized trajectory database
    $\tau_z$: Trajectory
    $p_z^q$: Moving point
Output:
    $\mathcal{T}_c$: Set of sub-trajectories

1:  $\mathcal{T}_c := \emptyset$
2:  $\mathcal{A}_1 := \{\langle p_z^q \rangle\}$
3:  $i := 1$
4:  while $i \leq \delta$ and $\mathcal{A}_i \neq \emptyset$ do
5:      for each $\tau_j \in \mathcal{A}_i$ do
6:          if $(P_b(\rho(r_k) \mid \tau_j) > \sigma)$ for some $r_k \in T^{\mathcal{S}}(\tau_j)$ then
7:              $\mathcal{T}_c := \mathcal{T}_c \cup \{\tau_j\}$
8:          end if
9:      end for
10:     $\mathcal{A}_{i+1} := \{\tau_k \sqsubseteq \tau_z \mid p_z^q \in \tau_k \wedge |\tau_k| = i + 1\}$
11:     $i := i + 1$
12: end while
13: return $\mathcal{T}_c$
```

It should be noted that eliminating a moving point from a trajectory data record by local suppression may generate new critical sub-trajectories. Identifying all of these critical sub-trajectories requires expensive computational cost.

An intuitive way to identify new critical sub-trajectories is to recall MPSTD. However, it is very costly. Instead, we apply the MCST algorithm to reduce the computational cost of identifying all new critical sub-trajectories. It significantly restricts the whole space of sub-trajectories to a very small set of sub-trajectories that are affected by local suppression.

The pseudo-code of MCST is shown in Algorithm 5. The algorithm takes a trajectory database $T^{\mathcal{S}}$, a trajectory τ_z, and a moving point p_z^q as input and returns a set $\mathcal{T}_c$ of new critical sub-trajectories as output. Let δ and σ be the background knowledge and privacy breach thresholds, respectively. MCST first initializes $\mathcal{T}_c$ to the empty set and $\mathcal{A}_1$ to the set $\{\langle p_z^q \rangle\}$ (Lines 1–2). Then for each sub-trajectory $\tau_j \in \mathcal{A}_i$, if τ_j is a critical sub-trajectory with respect to the threshold σ, it adds τ_j to $\mathcal{T}_c$ (Lines 5–9). Finally, it makes $\mathcal{A}_{i+1}$ from sub-trajectories $\tau_k \sqsubseteq \tau_z$ of length $i + 1$ containing the moving point p_z^q (Line 10). The above steps are repeated until i is greater than the threshold δ or $\mathcal{A}_i$ is empty (Lines 4–12).

Theorem 1. MCST is sufficient to identify all new critical sub-trajectories.

Proof. Suppose we are going to identify new critical sub-trajectories in the trajectory database $T^{\mathcal{S}}$ after eliminating a moving point p_z^q from a given trajectory τ_z. For any sub-trajectory τ_j in $T^{\mathcal{S}}$ not containing an instance of p_z^q, $P_b(\rho(r_k) \mid \tau_j)$ is the same before and after eliminating p_z^q, for all trajectory data records $r_k \in T^{\mathcal{S}}(\tau_j)$. Therefore, $\tau_j$cannot be a new critical sub-trajectory. Hence, if there is a new critical sub-trajectory, it must contain p_z^q. Similarly, since p_z^q is eliminated from τ_z, we only need to consider the sub-trajectories of τ_z. Consequently, τ_j is possible to be a new critical sub-trajectory only if $p_z^q \in \tau_j$ and $\tau_j \sqsubseteq \tau_z$. Therefore, MCST is sufficient to identify all new critical sub-trajectories. □

Example 8. Consider the SA-generalized trajectory database in Table 3 with $\delta = 2$ and $\sigma = 0.50$. After eliminating the moving point $e8$ from r_4, we only need to check the sub-trajectories $\langle e8\rangle$, $\langle b2, e8\rangle$, $\langle f6, e8\rangle$, and $\langle a7, e8\rangle$. Since all the sub-trajectories are not critical, no new critical sub-trajectory is generated with eliminating $e8$.

Lemma 3. An anonymized trajectory database $T^{\mathcal{G}}$ satisfies personalized privacy iff it contains no critical sub-trajectories.

Proof. Suppose $T^{\mathcal{G}}$ does not satisfy personalized privacy even though it contains no critical sub-trajectories. Hence, the adversary can perform at least one of the identity linkage, attribute linkage, or similarity attacks. As a result, according to Definitions 7–9, there is a background knowledge ξ_i such that $P_b(\rho(r_i)|\xi_i) > \sigma$ and thus ξ_i is a critical sub-trajectory, which contradicts the initial assumption. Therefore, $T^{\mathcal{G}}$ must satisfy personalized privacy. □

Our approach provides personalized privacy preservation in trajectory data publishing by combining both sensitive attribute generalization and trajectory local suppression. Therefore, the adversary cannot perform the personalized privacy attacks on the anonymized trajectory database.

Theorem 2. An anonymized trajectory database $T^{\mathcal{G}}$ is resistant to all three identity linkage, attribute linkage, and similarity attacks.

Proof. Let $r_i \in T^{\mathcal{G}}$ be a trajectory data record. Since $T^{\mathcal{G}}$ has been made anonymous; therefore, $P_b(\rho(r_i)|\xi_i) \leq \sigma$ for all background knowledge ξ_i with maximum length δ. Thus, according to (13), the adversary cannot correctly identify $g(r_i)$, and subsequently, $s(r_i)$ with confidence greater than σ, even though the size of $T(\xi_i)$ is small. Therefore, we conclude that $T^{\mathcal{G}}$ is resistant to all three identity linkage, attribute linkage, and similarity attacks. □

Example 9. Consider Table 4 that is an anonymized trajectory database from Table 1, satisfying personalized privacy with $\delta = 2$ and $\sigma = 0.50$. As discussed in Example 1, the adversary can perform all three privacy attacks on Table 1, while Table 4is resistant to all of them. As an example, given the background knowledge $\langle c4, d5\rangle$, $\langle f6, a7\rangle$, and $\langle f6, e9\rangle$, the adversary can infer Alice, Bob, and Carol have Diabetes, HIV, and Lung Infection with the confidences 0.33, 0.13, and 0.49, respectively, which are less than or equal to σ.

5. Complexity analysis

As mentioned before, PPTD consists of two main steps: sensitive attribute generalization and trajectory local suppression. Given a trajectory database T, we first apply STR in order to generate the set $\mathcal{A}^\delta$ of sub-trajectories in T whose length is at most δ, where δ is the background knowledge threshold. We then identify critical trajectory data records for each sub-trajectory in $\mathcal{A}^\delta$ and apply SAG to generalize their sensitive attribute values. The number of sub-trajectories in $\mathcal{A}^\delta$ is $O(n^\delta)$, where n is the number of distinct moving points in T. Hence, the worst-case time complexity of STR is $O(n \cdot |T| + n^\delta)$, where $|T|$ is the number of trajectory data records in T. Moreover, the worst-case time complexity of SAG is $O(\zeta_{max} \cdot |T|^2)$, where ζ_{max} is the maximum generalization depth, and since ζ_{max} is small, it becomes $O(|T|^2)$. Therefore, the time complexity of sensitive attribute generalization is bounded by $O(n^\delta \cdot |T|^2)$. Subsequently, we make a set $\mathcal{T}_c$ of remaining critical sub-trajectories in T and compute the personalized suppression score of each moving point in $\mathcal{T}_c$. We next eliminate the moving point with maximum personalized suppression score from some critical trajectory data records and apply MCST in order to update $\mathcal{T}_c$ with new critical sub-trajectories. In the worst-case, MCST has a time complexity of $O(n^\delta \cdot |T|)$. Thus, the time complexity of trajectory local suppression is bounded by $O(n^{2\delta} \cdot |T|^2)$. By incorporating both steps, the time complexity of PPTD becomes $O(n^{2\delta} \cdot |T|^2)$.

6. Experiments and analysis

In this section, we evaluate the performance of PPTD in terms of *information loss, disclosure risk*, and *query error*. We also compare it with related work in the literature.

6.1. Experimental results

We used two different trajectory datasets in our experiments: City80K and Metro100K. City80K [5,6] simulates the routes (i.e., trajectories) of 80,000 citizens in a metropolitan area with 26 city blocks in 24 hours and Metro100K [5] simulates the routes of 100,000 passengers in the Montréal subway transit system with 65 stations in 60 minutes, forming 3,900 dimensions. In both datasets, each trajectory data record corresponds to the route of one citizen (or passenger) and contains a sensitive attribute with five possible values.

We randomly assigned each trajectory data record to one of five privacy levels No Privacy, Low, Medium, High, or Very High, so that trajectory data records with lower privacy levels are more than those with higher privacy levels. We also generated a taxonomy tree of depth 6 with 108 leaf nodes for the sensitive attribute.

We conducted all experiments on a PC with an Intel Core i7 3.6 GHz CPU and 16 GB RAM.

6.1.1. Information loss

The aim of PPTD is to maintain an anonymized trajectory database as close to its original trajectory database as possible. Hence, we evaluate the information loss for different privacy levels due to the sensitive attribute generalization and the trajectory local suppression. Obviously, the lower the information loss is, the better the quality of the anonymized trajectory database will be.

Definition 15 (Sensitive attribute information loss)**.** Given an anonymized trajectory database $T^{\mathcal{G}}$ and a taxonomy tree Γ, the sensitive attribute information loss of any trajectory data record $r_i \in T^{\mathcal{G}}$ with respect to Γ is denoted by $\mathcal{I}_s(r_i)$ and defined as

$$\mathcal{I}_s(r_i) = \frac{|\Omega(r_i)| - 1}{|\mathcal{L}|}, \tag{16}$$

where $\Omega(r_i) = \ell(\eta(r_i))$ is the label of the twin node (or the SA-generalized value) of r_i and $\mathcal{L} = \ell(\upsilon_0(\Gamma))$ is the label of the root node (or the set of all sensitive attribute values assigned to the leaf nodes) of Γ. Note that $\upsilon_0(\Gamma)$ denotes the root node of Γ.

The amount of eliminated moving points is a general measure of the usefulness of anonymized trajectory data for a wide range of trajectory data mining tasks [24,25]. Hence, we define the trajectory information loss as a metric to illustrate the preservation of moving points in a trajectory database.

Definition 16 (Trajectory information loss)**.** Given an anonymized trajectory database $T^{\mathcal{G}}$ and its original trajectory database T, let $o: T^{\mathcal{G}} \to T$ be a function that maps trajectory data records in $T^{\mathcal{G}}$ to their corresponding trajectory data records in T. The trajectory information loss of any trajectory data record $r_i \in T^{\mathcal{G}}$ is denoted by $\mathcal{I}_\tau(r_i)$ and defined as

$$\mathcal{I}_\tau(r_i) = \frac{|\tau(o(r_i))| - |\tau(r_i)|}{|\tau(o(r_i))|}, \tag{17}$$

where $|\cdot|$ is the length of a trajectory.

Table 5
Effect of δ and σ on the average sensitive attribute information loss in percent for $\zeta_{max} = 1$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	0.0060	0.0060	0.0060	0.0059	0.0059	0.0076	0.0060	0.0060	0.0059	0.0059
Medium	0.0100	0.0100	0.0092	0.0099	0.0099	0.0871	0.0236	0.0137	0.0099	0.0099
High	0.0291	0.0224	0.0230	0.0234	0.0234	1.5186	0.5618	0.1850	0.0355	0.0263
Very High	25.3947	12.9460	8.6903	0.0787	0.0787	26.5323	14.1489	9.6481	1.1458	0.0962

Table 6
Effect of δ and σ on the average sensitive attribute information loss in percent for $\zeta_{max} = 2$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	0.0148	0.0148	0.0086	0.0082	0.0082	0.0164	0.0148	0.0086	0.0082	0.0082
Medium	0.0214	0.0214	0.0203	0.0165	0.0169	0.1639	0.0463	0.0260	0.0165	0.0169
High	0.0812	0.0707	0.0722	0.0575	0.0560	4.1551	1.5462	0.3566	0.0789	0.0650
Very High	73.7724	39.5759	17.8198	0.1867	0.1777	78.3781	42.2227	25.7104	2.0823	0.2042

Table 7
Effect of δ and σ on the average sensitive attribute information loss in percent for $\zeta_{max} = 1$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	0.0111	0.0037	0.0015	0.0007	0.0004	0.0112	0.0038	0.0015	0.0007	0.0004
Medium	0.0931	0.0382	0.0174	0.0056	0.0038	0.0931	0.0382	0.0174	0.0056	0.0038
High	1.0293	0.4762	0.1878	0.0491	0.0201	1.0293	0.4768	0.1878	0.0491	0.0201
Very High	27.3991	14.0639	8.3277	0.6340	0.1003	27.3991	14.0639	8.3277	0.6340	0.1003

Table 8
Effect of δ and σ on the average sensitive attribute information loss in percent for $\zeta_{max} = 2$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	0.0159	0.0051	0.0021	0.0008	0.0005	0.0164	0.0052	0.0022	0.0008	0.0005
Medium	0.1702	0.0661	0.0269	0.0087	0.0051	0.1732	0.0676	0.0269	0.0087	0.0051
High	2.8581	1.2189	0.3710	0.0985	0.0311	2.8909	1.2337	0.3729	0.0985	0.0311
Very High	81.2970	41.9332	17.8724	1.1683	0.1519	81.3738	42.0042	17.8797	1.1683	0.1519

Table 9
Effect of δ and σ on the average trajectory information loss in percent for $\zeta_{max} = 1$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	73.0853	0.2756	0.0236	0.0096	0.0090	86.4799	3.4664	0.0236	0.0096	0.0090
Medium	99.5359	21.7531	0.2020	0.0238	0.0238	99.9733	29.9825	1.5453	0.0523	0.0293
High	99.9996	40.3191	1.3688	0.0349	0.0263	99.9993	43.8242	6.1473	0.1771	0.0321
Very High	99.9698	40.0844	4.7881	0.0358	0.0307	99.9993	46.7650	8.5228	0.2818	0.0341

We can use the maximum generalization depth ζ_{max} to find a balance between the sensitive attribute and trajectory information losses.

Tables 5–8 show the effect of δ and σ on the average sensitive attribute information loss of trajectory data records in City80K and Metro100K, for $\zeta_{max} = 1$ and $\zeta_{max} = 2$, where δ and σ are, respectively, the background knowledge and privacy breach thresholds. In general, with increasing δ and decreasing σ, the average sensitive attribute information loss slightly increases due to generalizing more sensitive attribute values resulting from the increase in the number of critical trajectory data records.

Fig. 2 shows the effect of ζ_{max} on the average sensitive attribute information loss of trajectory data records in City80K and Metro100K, for $\sigma = 0.6$ and $\delta = 3$. Clearly, trajectory data records with higher privacy levels have more sensitive attribute information loss. This is because their twin nodes are closer to the root node of the taxonomy tree and so are being assigned more sensitive attribute values.

Tables 9–12 show the effect of δ and σ on the average trajectory information loss of trajectory data records in City80K and Metro100K, for $\zeta_{max} = 1$ and $\zeta_{max} = 2$. On the whole, with decreasing $\delta$and increasing σ, the average trajectory information loss slightly decreases due to the decrease in the number of critical sub-trajectories

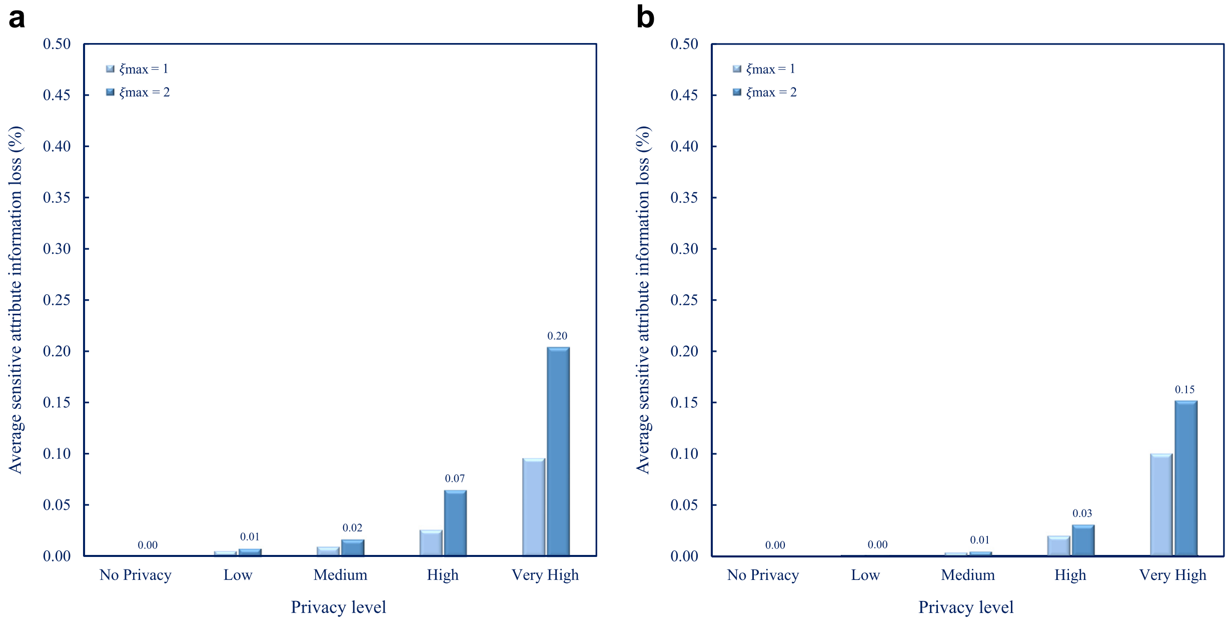

Fig. 2. Effect of ζ_{max} on the average sensitive attribute information loss in percent for $\sigma = 0.6$, $\delta = 3$, (a) City80K, and (b) Metro100K.

Table 10
Effect of δ and σ on the average trajectory information loss in percent for $\zeta_{max} = 2$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	74.8666	0.0227	0.0150	0.0094	0.0028	81.6327	1.7729	0.0155	0.0094	0.0028
Medium	99.5200	20.8296	0.0561	0.0170	0.0162	99.6832	28.8251	1.4126	0.0464	0.0217
High	99.8946	40.1283	0.8023	0.0255	0.0018	99.9219	43.4581	5.9018	0.1541	0.0077
Very High	99.8977	39.9509	2.1091	0.0269	0.0081	99.9280	45.2966	7.2227	0.2062	0.0108

Table 11
Effect of δ and σ on the average trajectory information loss in percent for $\zeta_{max} = 1$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	79.2527	0.2990	0.0264	0.0055	0.0023	87.0525	0.7323	0.0506	0.0068	0.0023
Medium	99.4738	21.9078	0.4983	0.0436	0.0090	99.1222	23.8958	0.7405	0.0539	0.0090
High	99.9863	39.8781	2.3720	0.0989	0.0137	99.9112	41.0811	3.2008	0.1291	0.0137
Very High	99.9894	42.2631	5.3265	0.1160	0.0097	99.9905	44.4119	6.2747	0.1533	0.0097

Table 12
Effect of δ and σ on the average trajectory information loss in percent for $\zeta_{max} = 2$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Low	79.6665	0.2753	0.0237	0.0055	0.0009	80.1432	0.7009	0.0456	0.0059	0.0009
Medium	99.4067	21.5196	0.4589	0.0387	0.0041	99.1456	23.6810	0.6716	0.0456	0.0041
High	99.9829	39.5783	1.9749	0.0743	0.0089	99.9237	40.9575	2.8208	0.0982	0.0103
Very High	99.9811	41.9730	2.8563	0.0884	0.0083	99.9915	44.1305	3.9980	0.1105	0.0097

resulting from the decrease in the number of critical trajectory data records.

Fig. 3 shows the effect of ζ_{max} on the average trajectory information loss of trajectory data records in City80K and Metro100K, for $\sigma = 0.4$ and $\delta = 3$. Clearly, trajectory data records with higher privacy levels and lower ζ_{max}have more trajectory information loss. Since they need more privacy protection and, in lower ζ_{max}, sensitive attribute values are less generalized; therefore, more moving points are eliminated from them by the trajectory local suppression.

From the above experiments, we conclude that $\sigma = 0.4$ and $\zeta_{max} = 1$ are good choices to balance between the sensitive attribute and trajectory information losses, because the adversary with the background knowledge threshold $\delta = 3$ cannot breach the privacy of moving objects with a probability of more than 40%, while the

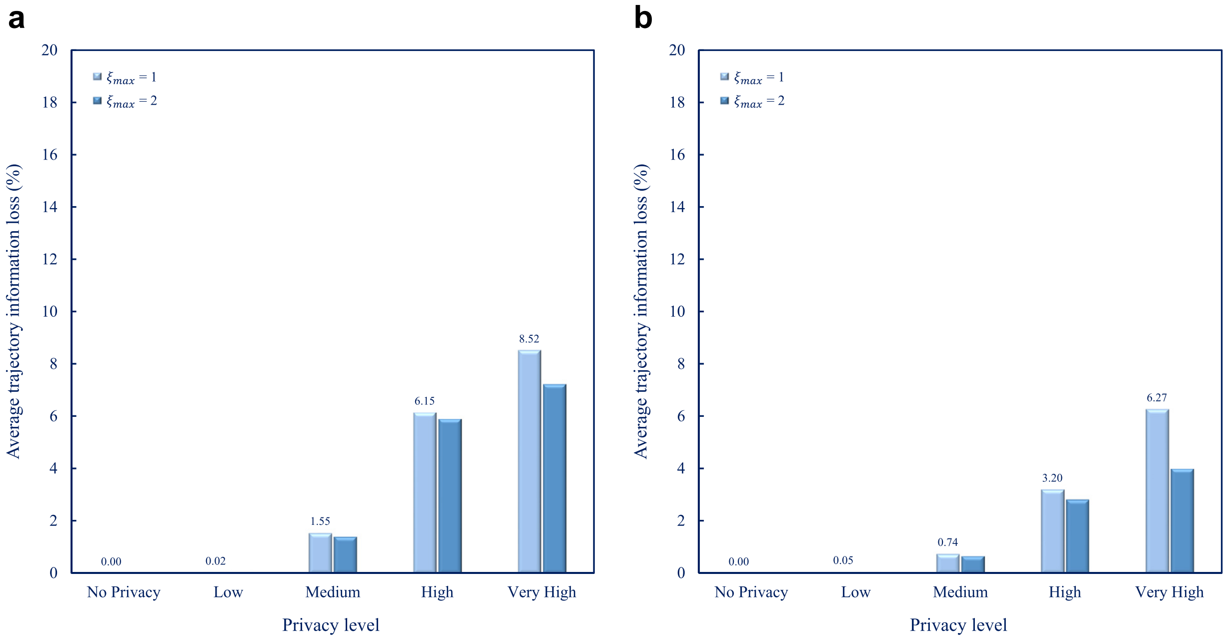

Fig. 3. Effect of ζ_{max} on the average trajectory information loss in percent for $\sigma = 0.4$, $\delta = 3$, (a) City80K, and (b) Metro100K.

average sensitive attribute and trajectory information losses for trajectory data records with the highest privacy level in City80K and Metro100K are at most 9.65% and 8.52%, respectively.

6.1.2. Disclosure risk

We use the disclosure risk as a metric to measure the privacy breach probability of moving objects. Given an anonymized trajectory database $T^{\mathcal{G}}$ and its original trajectory database T, let r_i be a trajectory data record in $T^{\mathcal{G}}$ and $\xi_i \sqsubseteq \tau(o(r_i))$ be the adversary's background knowledge, where $o(r_i)$ is a trajectory data record in T that corresponds to r_i. The disclosure probability of the sensitive attribute value $s(o(r_i))$ given ξ_i is calculated as

$$P(s(o(r_i)) \mid \xi_i) = \begin{cases} \frac{1}{|T^{\mathcal{G}}(\xi_i)|} \sum_{r_k \in T^{\mathcal{G}}(\xi_i)} P(s(o(r_i)) \mid \Omega(r_k)) & \xi_i \sqsubseteq \tau(r_i), \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

where $P(s(o(r_i)) \mid \Omega(r_k))$ is the disclosure probability of $s(o(r_i))$ given the SA-generalized value $\Omega(r_k)$ of a trajectory data record $r_k \in T^{\mathcal{G}}(\xi_i)$:

$$P(s(o(r_i)) \mid \Omega(r_k)) = \begin{cases} \frac{1}{|\Omega(r_k)|} & s(o(r_i)) \in \Omega(r_k), \\ 0 & \text{otherwise}. \end{cases} \tag{19}$$

The adversary may use any sequence of moving points with length not greater than δ as its background knowledge to perform a privacy attack. Therefore, the disclosure probability of $s(o(r_i))$ should be calculated for different lengths of ξ_i.

Definition 17 (Disclosure risk)**.** Given an anonymized trajectory database $T^{\mathcal{G}}$, the disclosure risk of any trajectory data record $r_i \in T^{\mathcal{G}}$ is denoted by $\mathcal{R}(r_i)$ and defined as

$$\mathcal{R}(r_i) = \frac{1}{|\mathcal{K}_i|} \sum_{\xi_i \in \mathcal{K}_i} P(s(o(r_i)) \mid \xi_i), \tag{20}$$

where

$$\mathcal{K}_i = \{\xi_i \mid \xi_i \sqsubseteq \tau(o(r_i)) \wedge |\xi_i| \le \delta\}. \tag{21}$$

Tables 13–16 show the effect of δ and σ on the average disclosure risk of trajectory data records in City80K and Metro100K, for $\zeta_{max} = 1$ and $\zeta_{max} = 2$. The results suggest that with decreasing σ, the average disclosure risk decreases. This is because more sensitive attribute values are generalized and more moving points are eliminated from trajectory data records. For $\zeta_{max} = 1$ and $\zeta_{max} = 2$, the average disclosure risk is almost the same. Because, as shown in Figs. 2 and 3, in $\zeta_{max} = 2$, sensitive attribute values are more generalized and, in $\zeta_{max} = 1$, moving points are more eliminated. Therefore, we conclude that ζ_{max} has no effect on the disclosure risk of trajectory data records and can only be used to balance between the sensitive attribute and trajectory information losses.

The anonymization, i.e., sensitive attribute generalization and/or trajectory local suppression, of a trajectory data record may affect the disclosure risk of other trajectory data records in addition to its own disclosure risk.

Example 10. Consider the trajectory data record r_1 in Table 2. Given the background knowledge $\langle a7, e8\rangle$, the probability of disclosure of $s(r_1)$ is 1. If we generalize the sensitive attribute value of the trajectory data record r_4 to Any Illness, then the probability of disclosure of $s(r_1)$ will be 0.53. Therefore, we conclude that anonymizing a trajectory data record may affect the disclosure probability and thus the disclosure risk of other trajectory data records.

We define the *affinity coefficient* to evaluate the effect that the disclosure risk of a given trajectory data record takes away from the anonymization of other trajectory data records relative to the information loss of that trajectory data record.

Definition 18 (Affinity coefficient)**.** Given an anonymized trajectory database $T^{\mathcal{G}}$, the affinity coefficient of any trajectory data record $r_i \in T^{\mathcal{G}}$ is denoted by $\mathcal{C}(r_i)$ and defined as

$$\mathcal{C}(r_i) = \begin{cases} \frac{\mathcal{O}(r_i) - \mathcal{R}(r_i)}{0.01 + \omega \cdot \mathcal{I}_s(r_i) + (1-\omega) \cdot \mathcal{I}_\tau(r_i)} & \xi_i \sqsubseteq \tau(r_i), \\ 0 & \text{otherwise}. \end{cases} \tag{22}$$

where ω is a weighting parameter and $\mathcal{O}(r_i)$ is the disclosure probability of $s(o(r_i))$ given that the adversary uniquely identifies r_i and so its SA-generalized value $\Omega(r_i)$:

$$\mathcal{O}(r_i) = P(s(o(r_i)) \mid \Omega(r_i)). \tag{23}$$

Note that, in (22), we have added 0.01 to the denominator to avoid dividing by zero.

Table 13
Effect of δ and σ on the average disclosure risk in percent for $\zeta_{max} = 1$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	20.3129	20.1744	19.7605	20.1809	20.1808	20.3847	20.4115	20.0358	20.2665	20.3592
Low	3.1805	19.9533	19.6438	20.0839	20.0842	2.6817	19.3127	19.9007	20.1586	20.2544
Medium	0.0540	16.4161	19.5912	20.0346	20.0346	0.0244	14.1736	19.3483	20.0884	20.1885
High	0.0000	14.1543	19.2180	20.0211	20.0225	0.0015	12.4787	17.7954	20.0171	20.1758
Very High	0.0002	14.2124	18.1648	20.0244	20.0248	0.0007	11.4155	17.1522	19.9630	20.1886

Table 14
Effect of δ and σ on the average disclosure risk in percent for $\zeta_{max} = 2$ (City80K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	20.2980	20.1570	19.6708	20.1798	20.1775	20.4427	20.3859	19.9799	20.2621	20.3573
Low	2.9806	19.9889	19.5529	20.0740	20.0758	1.7299	19.3607	19.8435	20.1450	20.2468
Medium	0.0406	16.5033	19.5228	20.0296	20.0290	0.0136	14.3839	19.3072	20.0799	20.1836
High	0.0015	14.1524	19.2622	20.0165	20.0192	0.0015	12.4787	17.7850	20.0161	20.1727
Very High	0.0006	14.2027	18.8544	20.0224	20.0233	0.0007	11.4057	17.4207	19.9889	20.1870

Table 15
Effect of δ and σ on the average disclosure risk in percent for $\zeta_{max} = 1$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	20.5484	20.1232	19.8232	20.1562	20.1935	20.8821	20.3086	19.9917	20.2840	20.3190
Low	2.0609	19.9435	19.8027	20.1552	20.1950	2.5609	19.9115	19.9512	20.2755	20.3159
Medium	0.0375	16.0863	19.6463	20.1392	20.1912	0.0445	15.4275	19.6794	20.2491	20.3100
High	0.0008	13.8918	19.1033	20.1186	20.1899	0.0009	13.4385	18.8567	20.2086	20.3055
Very High	0.0006	13.2771	18.2866	20.1107	20.1934	0.0006	12.4974	17.9419	20.1911	20.3083

Table 16
Effect of δ and σ on the average disclosure risk in percent for $\zeta_{max} = 2$ (Metro100K).

Privacy level	$\delta = 2$					$\delta = 3$				
	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$
No Privacy	20.5444	20.1092	19.7642	20.1535	20.1926	20.7131	20.2977	19.9311	20.2802	20.3178
Low	1.9925	19.9394	19.7467	20.1527	20.1948	2.1884	19.9136	19.8965	20.2725	20.3161
Medium	0.0399	16.1234	19.6030	20.1383	20.1922	0.0571	15.4413	19.6478	20.2495	20.3115
High	0.0009	13.8972	19.1537	20.1250	20.1911	0.0007	13.4431	18.9234	20.2199	20.3072
Very High	0.0010	13.3139	18.8731	20.1179	20.1933	0.0003	12.5624	18.5255	20.2070	20.3080

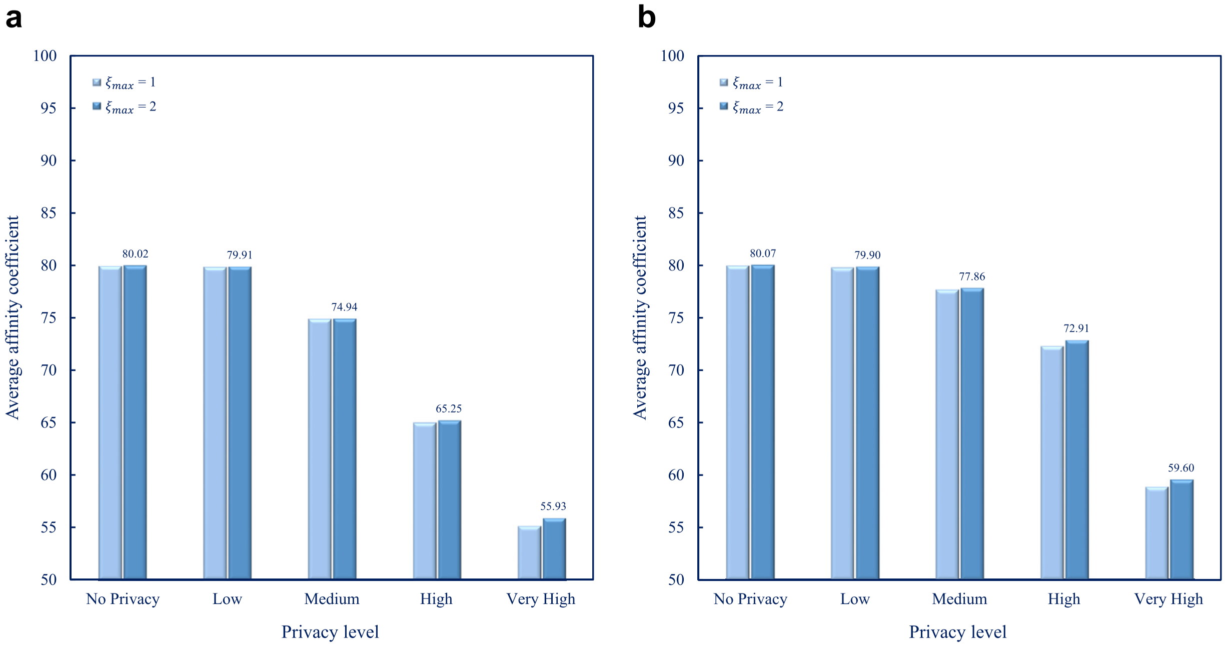

Fig. 4. Effect of ζ_{max} on the average affinity coefficient for $\sigma = 0.4$, $\delta = 3$, (a) City80K, and (b) Metro100K.

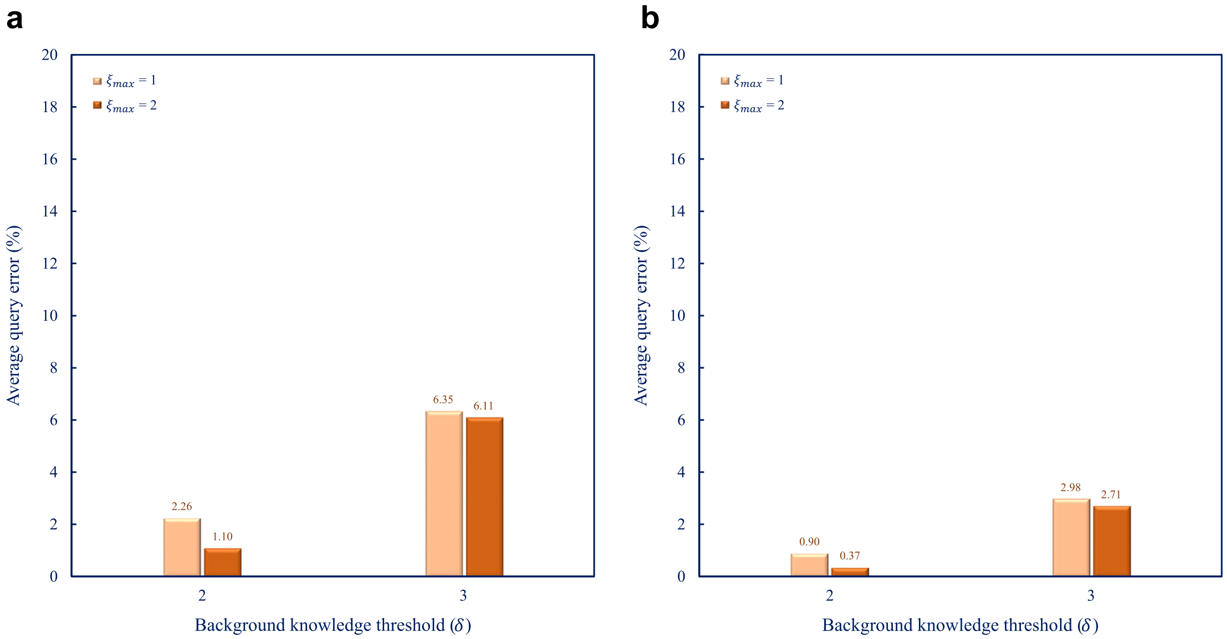

Fig. 5. Effect of δ and ζ_{max} on the average error of count queries for $\sigma = 0.4$, (a) universal, and (b) existential (City80K).

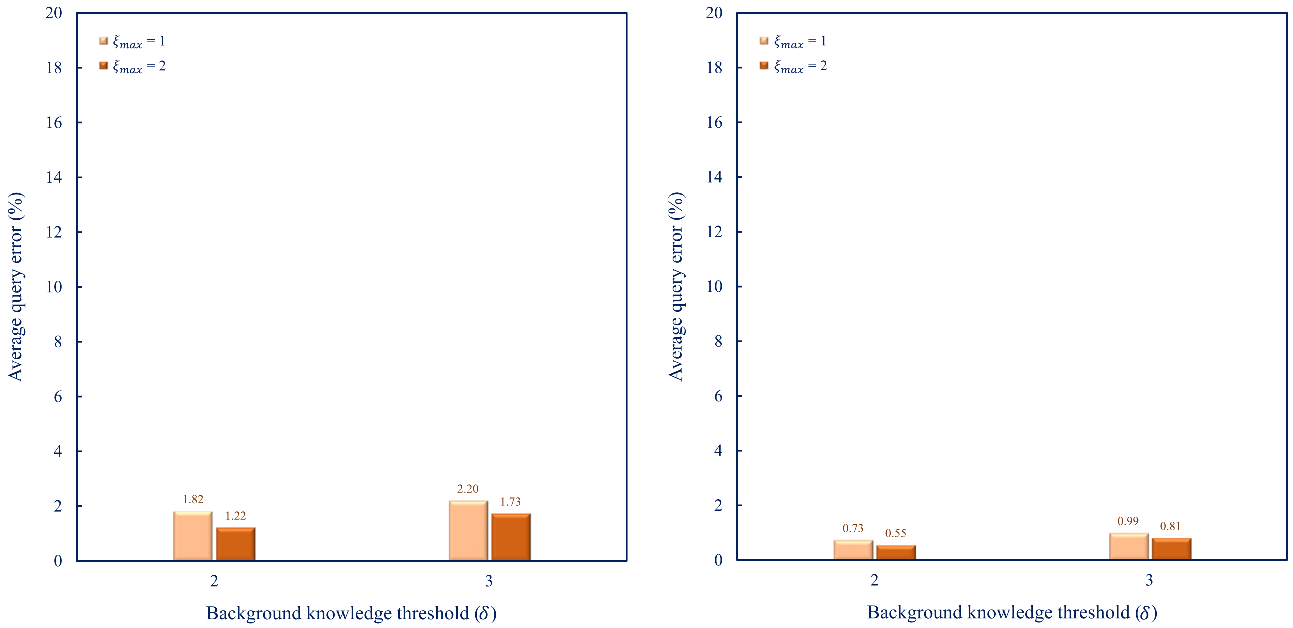

Fig. 6. Effect of δ and ζ_{max} on the average error of count queries for $\sigma = 0.4$, (a) universal, and (b) existential (Metro100K).

Fig. 4 shows the effect of ζ_{max} on the average affinity coefficient of trajectory data records in City80K and Metro100K, for $\omega = 0.5$, $\sigma = 0.4$, and $\delta = 3$. In general, trajectory data records with the privacy levels of No Privacy and Very High have the highest and lowest affinity coefficients, respectively. However, from Tables 5–12, we can see that trajectory data records with the privacy level of No Privacy do not have any information loss and trajectory data records with the privacy level of Very High have the most information loss. Therefore, we conclude that the anonymization of trajectory data records with higher privacy levels significantly reduces the disclosure risk of those with lower privacy levels.

6.1.3. Query error

One of the main goals of data publishing is to allow users to run different queries on the published data in order to find useful information they are seeking. Therefore, another way to measure utility is to compare the answers to queries on both the original and anonymized data. Intuitively, when the answers on both data are similar for a large and diverse number of queries, the anonymized data can be regarded as preserving the utility of the original data. Count queries, as one of the most popular queries on trajectory data, are used in many different data mining tasks. Hence, we define two types of count queries on trajectory data: *universal count query* and *existential count query*.

Definition 19 (Universal count query)**.** Given a trajectory database T and a sub-trajectory τ_j, the universal count query for τ_j on T, denoted by $Q_u(\tau_j, T)$, is defined as the number of trajectory data records in T for which τ_j is a sub-trajectory:

$$Q_u(\tau_j, T) = |\{r_i \in T \mid \tau_j \sqsubseteq \tau(r_i)\}|. \tag{24}$$

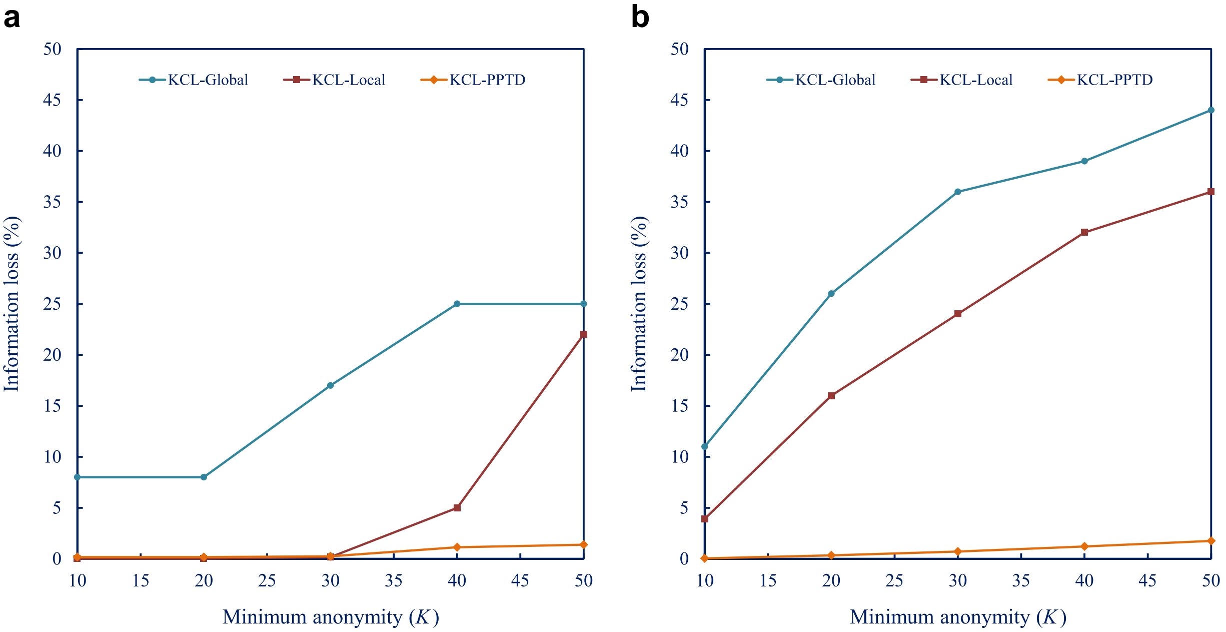

Fig. 7. Effect of K on information loss of KCL-Global [4,5], KCL-Local [6], and KCL-PPTD for $C = 0.6$, $L = 3$, (a) City80K, and (b) Metro100K.

Definition 20 (Existential count query)**.** Given a trajectory database T and a sub-trajectory τ_j, the existential count query for τ_j on T, denoted by $Q_e(\tau_j, T)$, is defined as the number of trajectory data records in T for which at least one sub-trajectory of τ_j is a sub-trajectory:

$$Q_e(\tau_j, T) = |\{r_i \in T \mid \exists \tau_k \sqsubseteq \tau_j : \tau_k \sqsubseteq \tau(r_i)\}|. \tag{25}$$

Given an original trajectory database T and a sub-trajectory $\tau_j \in T$, we measure the utility of both universal and existential count queries for τ_j on an anonymized trajectory database $T^{\mathcal{G}}$ by their errors with respect to the true results on T, which are calculated as

$$\begin{aligned} \mathcal{E}_u(\tau_j) &= \frac{|Q_u(\tau_j, T) - Q_u(\tau_j, T^{\mathcal{G}})|}{|Q_u(\tau_j, T)|}, \\ \mathcal{E}_e(\tau_j) &= \frac{|Q_e(\tau_j, T) - Q_e(\tau_j, T^{\mathcal{G}})|}{|Q_e(\tau_j, T)|}. \end{aligned} \tag{26}$$

For our experiments, we randomly selected 1000 sub-trajectories of different sizes from trajectory data records in City80K and Metro100K, and calculated the average error of the universal and existential count queries for them. Figs. 5 and 6 show the obtained results for $\sigma = 0.4$. Obviously, with increasing $\delta$and decreasing ζ_{max}, the average error of both universal and existential count queries increase. The reason is that an increase in δ has a direct positive impact on the number of critical sub-trajectories and this direct impact is strengthened by a decrease in ζ_{max}, leading to more moving points to be eliminated from trajectory data records.

6.2. Comparison

We cannot directly compare our approach with previous related work [4–6] on privacy preservation in trajectory data publishing, because none of them consider the personalized privacy requirements and the sensitive attribute generalization. Instead, we consider equal conditions with KCL-Global and KCL-Local described in [4–6] and present a new variant of our approach, called KCL-PPTD. KCL-PPTD is similar to PPTD, with the difference that it does not use the sensitive attribute generalization and applies the k-anonymity to prevent the identity linkage attack. However, KCL-Global, KCL-Local, and KCL-PPTD are not resistant to the similarity attack. Note that C and L are equivalent to σ and δ in PPTD, respectively.

KCL-Global and KCL-Local use Metro100K [5] and City80K [5,6] as the trajectory dataset and consider one of five possible values of its sensitive attribute as sensitive and the others as non-sensitive, which in KCL-PPTD, they correspond to sensitive attribute values with the privacy levels of Low and No Privacy, respectively. Therefore, approximately 80 percent of the trajectory data records in City80K and Metro100K do not need any privacy protection. In the following experiments, we show that KCL-PPTD would significantly lower the information loss of trajectory data.

For the purpose of fair comparison, we use the same information loss metric as that defined in [6], to measure the percentage of moving points that are lost due to suppressions:

$$\mathcal{I}_\tau = \frac{N(T) - N(T^{\mathcal{G}})}{N(T)}, \tag{27}$$

where $N(T)$ and $N(T^{\mathcal{G}})$ are the numbers of moving points in the original and anonymized trajectory databases T and $T^{\mathcal{G}}$, respectively.

6.2.1. Effect of K

We vary the parameter K from 10 to 50 while fixing $C = 0.6$ and $L = 3$ (which are equivalent to taking $\sigma = 0.6$ and $\delta = 3$ in PPTD), on both Metro100K and City80K to compare the effect of K on KCL-Global [4,5], KCL-Local [6], and KCL-PPTD, the results of which are demonstrated in Fig. 7. Clearly, KCL-PPTD can significantly reduce the information loss for higher values of K.

6.2.2. Effect of C

Fig. 8 shows the effect of C on the information loss of KCL-Global [4,5], KCL-Local [6], and KCL-PPTD, while fixing $K = 30$ and $L = 3$. When $C$is small, the information loss is high for KCL-Global and KCL-Local. However, for different values of C, KCL-PPTD results in substantially low information loss.

As a result, KCL-PPTD totally has low information loss. This is because it eliminates critical moving points only from critical trajectory data records, while KCL-Global [4,5] and KCL-Local [6] may eliminate critical moving points from non-critical trajectory data records in addition to critical ones.

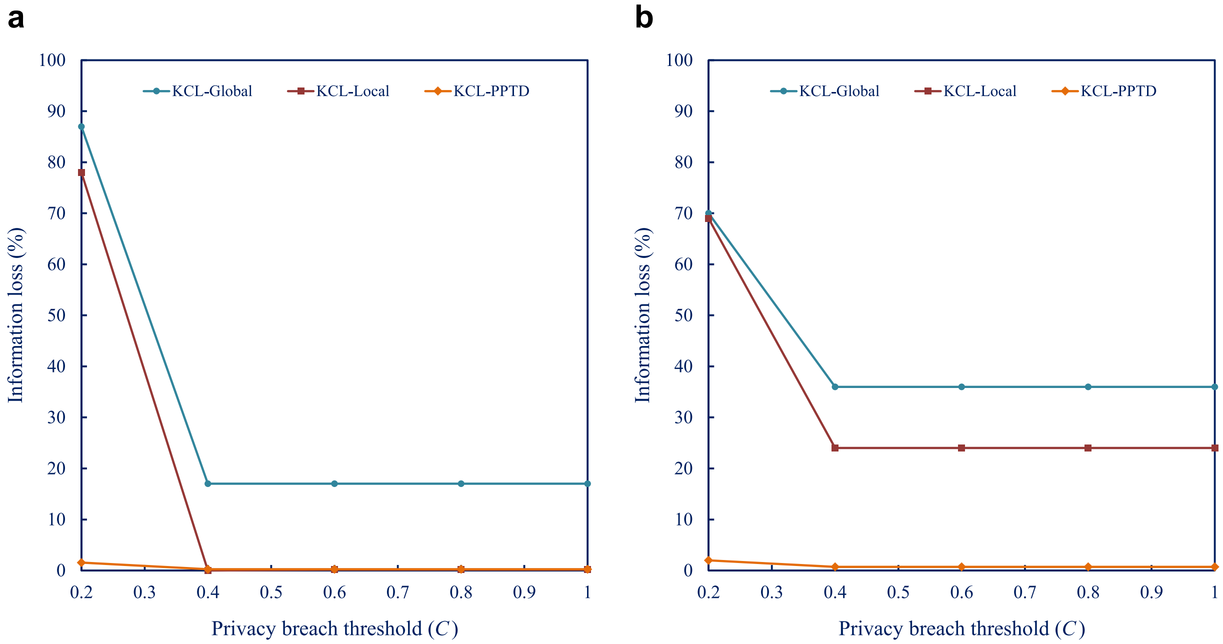

Fig. 8. Effect of C on information loss of KCL-Global [4,5], KCL-Local [6], and KCL-PPTD for $K = 30$, $L = 3$, (a) City80K, and (b) Metro100K.

Table 17
Comparison of PPTD with other privacy preserving approaches.

Approach	Category	Prior privacy model	Personalized privacy	Attack resistance		
				Identity linkage	Attribute linkage	Similarity
Nergiz [17]	Clustering	✓	–	✓	–	–
Monreale [18]	Clustering	✓	–	✓	–	–
Mahdavifar [19]	Clustering	✓	✓	✓	–	–
Domingo-Ferrer [20]	Clustering	✓	–	✓	–	–
Terrovitis [21]	Quasi-identifier	–	–	✓	–	–
Yarovoy [22]	Quasi-identifier	✓	–	✓	–	–
Mohammed [5]	Quasi-identifier	✓	–	✓	✓	–
Chen [6]	Quasi-identifier	✓	–	✓	✓	–
Ghasemzadeh [23]	Quasi-identifier	✓	–	✓	–	–
KCL-PPTD	Quasi-identifier	✓	–	✓	✓	–
PPTD	Quasi-identifier	–	✓	✓	✓	✓

7. Conclusion and further study

Trajectory data are becoming more popular due to the rapid development of mobile devices and the widespread use of location-based services. They may be associated with sensitive attributes, such as disease, job, and income. Therefore, improper publishing of trajectory data may breach the privacy of moving objects whose locations are easily monitored and tracked.

Most previous work for preserving privacy in trajectory data publishing consider the same level of privacy protection for all moving objects. The result is that some moving objects with high privacy requirements may be offered low privacy protection, and vice versa, leading to an increase in information loss and disclosure risk. We addressed this problem by focusing on the concept of personalized privacy and presented PPTD, a novel approach for preserving privacy in trajectory data publishing that combines both sensitive attribute generalization and trajectory local suppression to achieve a balance between data utility and data privacy in accordance with the privacy requirements of moving objects.

PPTD not only is able to provide personalized privacy protection in trajectory data publishing, but also it is resistant to all three identity linkage, attribute linkage, and similarity attacks. It is critical to prevent these attacks in trajectory data publishing because more and more trajectory data mining tasks will resort to both trajectory data and sensitive attributes [6]. Table 17 compares the attack resistance of PPTD with that of other privacy preserving approaches previously reported in the literature. For the sake of simplicity, we refer to each approach by the first author's surname. For each approach, we specify whether or not it follows a certain privacy model and so offers prior privacy guarantees. From the table, we can observe that only PPTD is resistant to the similarity attack.

We used two trajectory datasets, namely City80K and Metro100K, and randomly assigned each of trajectory data records in the datasets to one of five privacy levels No Privacy, Low, Medium, High, or Very High. We also generated a taxonomy tree of depth 6 with 108 leaf nodes for the sensitive attribute Disease. We next evaluated the performance of PPTD in terms of sensitive attribute information loss, trajectory information loss, disclosure risk, affinity coefficient, and query error for different values of parameters. The experimental results showed that trajectory data records with the privacy level of Very High have the most information loss and the least disclosure risk. In addition, the anonymization of trajectory data records with higher privacy levels significantly reduces the disclosure risk of those with lower privacy levels. We finally considered equal conditions with KCL-Global [4,5] and KCL-Local [6] and presented a new variant of PPTD, called KCL-PPTD. It should be noted that KCL-Global, KCL-Local, and KCL-PPTD are not resistant to the similarity attack. The experimental results suggested KCL-PPTD totally has low information loss. This is because it eliminates critical moving points only from critical trajectory data records, while KCL-Global [4,5] and KCL-Local [6] may eliminate critical moving points from non-critical trajectory data records in addition to critical ones.

Throughout the paper, we assumed that the given trajectory database contains only a sensitive attribute. As a future work, we are interested in extending PPTD for trajectory databases with multiple sensitive attributes. This enables PPTD to be applied to more realistic and complex scenarios of moving objects.

References

[1] E. Kaplan, T.B. Pedersen, E. Savaş, Y. Saygın, Discovering private trajectories using background information, Data Knowl. Eng. 69 (7) (2010) 723–736, doi:10.1016/j.datak.2010.02.008.

[2] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: A survey of recent developments, ACM Comput. Surv. 42 (4) (2010) 1–53, doi:10.1145/1749603.1749605.

[3] A.H.M. Sarowar Sattar, J. Li, X. Ding, J. Liu, M. Vincent, A general framework for privacy preserving data publishing, Knowl. Based Syst. 54 (2013) 276–287, doi:10.1016/j.knosys.2013.09.022.

[4] B.C.M. Fung, M. Cao, B.C. Desai, H. Xu, Privacy protection for RFID data, in: Proceedings of the 2009 ACM Symposium on Applied Computing, in: SAC '09, ACM, New York, NY, USA, 2009, pp. 1528–1535, doi:10.1145/1529282.1529626.

[5] N. Mohammed, B.C.M. Fung, M. Debbabi, Preserving Privacy and Utility in RFID Data Publishing, Technical Report. 6850, Concordia University, Montreal, QC, Canada, 2010.

[6] R. Chen, B.C.M. Fung, N. Mohammed, B.C. Desai, K. Wang, Privacy-preserving trajectory data publishing by local suppression, Inf. Sci. 231 (2013) 83–97, doi:10.1016/j.ins.2011.07.035.

[7] X. Xiao, Y. Tao, Personalized privacy preservation, in: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, in: SIGMOD '06, ACM, New York, NY, USA, 2006, pp. 229–240, doi:10.1145/1142473.1142500.

[8] E. Ghasemi Komishani, M. Abadi, A generalization-based approach for personalized privacy preservation in trajectory data publishing, in: Proceedings of the 2012 6th International Symposium on Telecommunications, in: IST '12, 2012, pp. 1129–1135, doi:10.1109/ISTEL.2012.6483156.

[9] K.G. Shin, X. Ju, Z. Chen, X. Hu, Privacy protection for users of location-based services, IEEE Wirel. Commun. 19 (1) (2012) 30–39, doi:10.1109/MWC.2012.6155874.

[10] X. Pan, J. Xu, X. Meng, Protecting location privacy against location-dependent attacks in mobile services, IEEE Trans. Knowl. Data Eng. 24 (8) (2012) 1506–1519, doi:10.1109/TKDE.2011.105.

[11] M. Wernke, P. Skvortsov, F. Dürr, K. Rothermel, A classification of location privacy attacks and approaches, Personal Ubiquitous Comput. 18 (1) (2014) 163–175, doi:10.1007/s00779-012-0633-z.

[12] C. Dwork, Differential privacy, in: M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (Eds.), Automata, Languages and Programming, *Lecture Notes in Computer Science*, vol. 4052, Springer, Berlin, Heidelberg, Germany, 2006, pp. 1–12, doi:10.1007/11787006_1.

[13] N. Niknami, M. Abadi, F. Deldar, SpatialPDP: A personalized differentially private mechanism for range counting queries over spatial databases, in: Proceedings of the 2014 4th International Conference on Computer and Knowledge Engineering, in: ICCKE'14, 2014, pp. 709–715, doi:10.1109/ICCKE.2014.6993414.

[14] J. Domingo-Ferrer, J. Soria-Comas, From t-closeness to differential privacy and vice versa in data anonymization, Knowl.Based Syst. 74 (2015) 151–158, doi:10.1016/j.knosys.2014.11.011.

[15] R. Chen, B.C.M. Fung, B.C. Desai, N.M. Sossou, 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, in: KDD '12, ACM, New York, NY, USA, 2012, pp. 213–221, doi:10.1145/2339530.2339564.

[16] S.-S. Ho, S. Ruan, Preserving privacy for interesting location pattern mining from trajectory data, Trans. Data Priv. 6 (1) (2013) 87–106.

[17] M.E. Nergiz, M. Atzori, Y. Saygın, B. Güç, Towards trajectory anonymization: A generalization-based approach, Trans. Data Priv. 2 (1) (2009) 47–75.

[18] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, S. Wrobel, Movement data anonymity through generalization, Trans. Data Priv. 3 (2) (2010) 91–121.

[19] S. Mahdavifar, M. Abadi, M. Kahani, H. Mahdikhani, A clustering-based approach for personalized privacy preserving publication of moving object trajectory data, in: L. Xu, E. Bertino, Y. Mu (Eds.), Network and System Security, *Lecture Notes in Computer Science*, volume 7645, Springer, Berlin, Heidelberg, Germany, 2012, pp. 149–165, doi:10.1007/978-3-642-34601-9_12.

[20] J. Domingo-Ferrer, R. Trujillo-Rasua, Microaggregation- and permutation-based anonymization of movement data, Inf. Sci. 208 (2012) 55–80, doi:10.1016/j.ins.2012.04.015.

[21] M. Terrovitis, N. Mamoulis, Privacy preservation in the publication of trajectories, in: Proceedings of the 9th International Conference on Mobile Data Management, in: MDM '08, IEEE Computer Society, Washington, DC, USA, 2008, pp. 65–72, doi:10.1109/MDM.2008.29.

[22] R. Yarovoy, F. Bonchi, L.V.S. Lakshmanan, W.H. Wang, Anonymizing moving objects: How to hide a MOB in a crowd? in: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, in: EDBT '09, ACM, New York, NY, USA, 2009, pp. 72–83, doi:10.1145/1516360.1516370.

[23] M. Ghasemzadeh, B.C.M. Fung, R. Chen, A. Awasthi, Anonymizing trajectory data for passenger flow analysis, Transport. Res. Part C Emerg. Technol. 39 (2014) 63–79, doi:10.1016/j.trc.2013.12.003.

[24] J.-G. Lee, J. Han, K.-Y. Whang, Trajectory clustering: A partition-and-group framework, in: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, in: SIGMOD '07, ACM, New York, NY, USA, 2007, pp. 593–604, doi:10.1145/1247480.1247546.

[25] J.-G. Lee, J. Han, X. Li, H. Gonzalez, TraClass: Trajectory classification using hierarchical region-based and trajectory-based clustering, Proc. VLDB Endow. 1 (1) (2008) 1081–1094, doi:10.14778/1453856.1453972.