Article

Not peer-reviewed version

Hierarchical Temporal Semantic Tree Based Trajectory Privacy Protection Scheme over LBSNs

Peixu Xing , Mengxing Huang * , Liang Zhu * , Yuanyuan Wu

Posted Date: 10 October 2024

doi: 10.20944/preprints202410.0727.v1

Keywords: location-based service; trajectory privacy protection; differential privacy; semantic tree; user preference doi:10.20944/preprints202410.0727.v1

Article

Hierarchical Temporal Semantic Tree Based Trajectory Privacy Protection Scheme over LBSNs

Peixu Xing [1,2], Mengxing Huang [1,*], Liang Zhu [2,*] and Yuanyuan Wu [1]

[1] College of Information Science and Technology, Hainan University, Haikou 570228, China
[2] College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China
* Correspondence: Huangmx09@163.com (M.H.); lzhu@zzuli.edu.cn (L.Z.)

Featured Application: The research results can be applied to IOT security protection to protect the privacy of users from being leaked.

Abstract: Most of the existing trajectory privacy protection methods provide a unified degree of privacy protection for data, and most of the methods mainly focus on the spatial and temporal dimensions of trajectory data, with little consideration of the semantic information of the location. To address these issues, we propose a hierarchical temporal semantic tree-based personalized trajectory privacy protection (HTST-PTPP) scheme, which mainly constructs a hierarchical temporal semantic tree by considering semantic categories and residence time. Then, quantifies the privacy protection requirements of each location, divides the privacy level, and uses differential privacy technology to realize personalized trajectory privacy protection. Finally, experiments are carried out on two real data sets. The experimental results show that our HTST-PTPP scheme has good performance in data availability and privacy protection.

Keywords: location-based service; trajectory privacy protection; differential privacy; semantic tree; user preference

1. Introduction

With the rise of the mobile Internet and the rapid development of intelligent mobile terminals, Location-Based Services (LBS) have been widely used in various fields of life, such as travel services, social media positioning, health and motion tracking [1]. Location service providers provide users with more personalized services through a large of collection and analysis of user's location data. Location-based social networks (LBSNs) have three characteristics: First, the new dimension is added to form a user's social media with geographical location information [2] (e.g., documents, pictures, audio, video, etc.), and record the user's historical location data. Secondly, locations have become the new object in LBSNs, which promotes the development of various LBS, including mobile marketing [3], traffic simulation and prediction [4], and disaster rescue [5]. Third, it is the bridge connecting the virtual world and the real society, and realizes a new Online To Offline (O2O) coordination mechanism [6], namely: Online users use LBSNs to share the real experience of the real society, while offline servers use users' historical information to analyze user behavior, and promote the research on the correlation between users' real activity status and social activity characteristics. For example, mobile applications developed by combining LBS and O2O can provide a series of mobile value-added services [7], including: learning, shopping, travel, etc.

Rich LBS bring great convenience to people's lives, but also cause the risk of leakage of users' personal privacy (such as identity, location or query information, etc.) [8]. In LBSNs, users need to upload their real location information to the location-based social network server to obtain personalized service experience. Users not only enjoy the great convenience of location services, but also bear the risk of privacy leakage [9,10].

Although the existing trajectory privacy protection methods show the ability of privacy protection to a certain extent, most of these methods provide a unified degree of privacy protection for data, resulting in some data cannot being fully protected, while other data may not require a high degree of privacy protection. And most of the methods mainly focus on the spatial and temporal dimensions of trajectory data, and rarely consider the semantic information of locations. The semantic information of the location refers to the abstract information related to the geographical location, which provides deeper information about the location. Therefore, it is very important to consider semantic information when protecting the location privacy. However, the same semantic category also has different privacy requirements due to the different residence times of users. For example, the semantic category of location A is cinema. User 1 and user 2 stay in location A for eight hours and two hours respectively. It can be inferred that user 1 works in location A and user 2 plays in location A, so they have different privacy needs. In the process of personalized privacy protection, it is necessary to consider the two factors of semantic category and residence time at the same time, so as to better adapt to the future diversified and personalized location service development needs. To address this problem, a hierarchical temporal semantic tree-based personalized trajectory privacy protection (HTST-PTPP) scheme is proposed in this paper.

The significant contributions brought forward by this paper are as follows:

(1) We constructed a hierarchical temporal semantic tree to realize personalized privacy requirements from the user's perspective by fully considering the semantic category and residence time.

(2) We proposed a personalized privacy level division strategy. By using the TF-IDF algorithm to calculate the privacy sensitivity of each location, the privacy protection requirements of the location are quantified, the privacy level is divided, the privacy budget is reasonably allocated, and the differential privacy technology is used to realize the personalized trajectory privacy protection.

(3) We conducted an extensive experimental study to verify the data availability and privacy protection of the proposed HTST-PTPP scheme on two real data sets.

2. Related Works

The existing trajectory privacy protection methods can be divided into four categories: trajectory generalization, fake trajectories, suppression technology and differential privacy protection.

2.1. Trajectory Privacy Protection

Generalization-based trajectory privacy protection technology [11] is to generalize all sampling points on the trajectory, and achieve the purpose of privacy protection by generalizing them to the corresponding anonymous area. At present, the most commonly used generalization method is *k*-anonymity technology. The basic idea is that any individual has at least the same identifier with *k*-1 records in the published data, so it cannot be distinguished. Hemkumar et al. [12] proposed a new anonymization method, which includes two stages: virtualization and suppression. The virtualization method is used as an alternative mechanism for sensitive attributes, and the suppression method is used as an anonymous mechanism for user trajectories, so that users can resist multiple attacks. Shaham et al. [13] proposed an anonymous method for spatial-temporal trajectory data sets based on machine learning. This method uses machine learning algorithms to cluster trajectories. In addition, a variant of the *k*-means algorithm is proposed to prevent the leakage of over-sensitive data sets.

The core of trajectory privacy protection based on fake trajectory is to generate a series of fake trajectories by using the user's real moving trajectory, and mix these fake trajectories with the original trajectory. Wu et al. [14] proposed a new adaptive trajectory generation algorithm, which considers the influence of historical trajectories on false trajectories. Under the same privacy protection requirements, the method should generate fewer false trajectories to meet, and the generated false position distribution is more uniform, which can meet the stricter privacy protection requirements. Zhang et al. [15] proposed a virtual trajectory privacy protection scheme based on radius constraint, which is used to enhance the trajectory privacy protection of users in mobile social network applications and effectively reduce the exposure risk of single point location and trajectory. Wang et al. [16] proposed a triple real-time trajectory privacy protection mechanism (T-LGEB) based on edge computing and blockchain, in order to protect the trajectory privacy of task participants while ensuring high data availability and real-time data.

The trajectory privacy protection based on suppression is to protect the trajectory privacy directly by removing or hiding some sensitive positions in the trajectory. This method is simple to implement, but it is easy to cause information loss and reduce the availability of data. Wang et al. [17] found that the traditional trajectory publication framework in which a trusted server has access to the raw data from mobile clients. The clients call for much stronger data privacy preservation locally without sharing their raw data. Therefore, a federated analytics-based secure trajectory publication (FASTPub) mechanism is proposed to solve this kind of problem. Chen et al. [18] proposed a local suppression method to achieve customizable trajectory data anonymization. This method overcomes the challenges of high dimension, sparsity and sequence in trajectory data anonymization, and significantly improves data utility.

2.2. Privacy Quantification

The purpose of privacy quantification is to realize the measurement of geolocation privacy while ensuring the acquisition of key information by transforming the geolocation information of a specific user into a numerical data set. Commonly used quantitative privacy metrics include information entropy, mutual information, decision tree and attack model, etc. Differential privacy is a privacy protection model proposed by Dwork et al. [19] in 2006, which solves the two defects of the traditional privacy protection model. It does not need to consider the background knowledge of the attacker and has a strict mathematical definition. Wang et al. [20] introduced the concept of sequence indistinguishability and proposed a scheme for publishing related time series data based on differential privacy, which makes it impossible for attackers to distinguish the noise sequence from the original sequence. Ghane et al. [21] proposed a new trajectory generation algorithm. The algorithm models the data as a graph, and retains the spatial and temporal information of the trajectory and the distance and stay position between the real trajectories, which effectively improves the computational efficiency and practicability. Based on location generalization and local differential privacy technology, Yang et al. [22] proposed a trajectory data perturbation method based on quadtree index, which considers the correlation between adjacent spatial-temporal nodes of the trajectory while protecting the user's trajectory privacy. Zheng et al. [23] proposed a semantic-sensitive privacy-preserving location trajectory data sharing technology. This method protects both user data privacy and semantic privacy, and achieves a balance between privacy and utility. Wu et al. [24] proposed a privacy protection mechanism related to confidentiality that satisfies differential privacy. This method considers the trajectory association problem between multiple users and realizes the protection of trajectory association between multiple users. Wu et al. [25] proposed a trajectory correlation privacy-preserving mechanism (TCPP) that fulfills differential privacy. The mechanism can preserve the trajectory correlation based on a customized privacy budget allocation strategy. Chen et al. [26] proposed the development of an optimal privacy budget allocation algorithm for the transit smart card data. The goal is to publish the non-interactive sanitized trajectory data under a differential privacy definition.

3. Overview of Scheme

In this section, we give the problem definition, scheme design and attack model.

3.1. Problem Definition

Definition 1(Original trajectory data). The original trajectory is a coordinate sequence that connects the single position coordinates collected by GPS in chronological order. Each location point l is composed of a triple $\langle lon, lat, t\rangle$, where lon and lat represent the longitude and latitude of the location point l, respectively, and t represents the time that the user passes through the location point. The trajectory sequence T can be expressed as $T = l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_n$.

Definition 2(Stay points). The stay point S is the clustering of the original location points, indicating that the user stays in a certain geographical area for a period of time. Given a distance threshold θ_d and a time threshold θ_t, for a set of consecutive position points $L=\left(l_m,l_{m+1},\cdots,l_n\right)$, where $m<k\leq n$, $Dist\left(l_m,l_k\right)\leq\theta_d$ and $Int\left(l_m,l_n\right)\geq\theta_t$. Each stay point S consists of a triple $\left\langle lon,lat,s_t\right\rangle$, where lon and lat represent the longitude and latitude of the stay point S, respectively, and s_t represents the length of time spent at the stay point.

Definition 3(Location semantic). Location information usually includes geographic information and semantic information. Geographic information refers to the longitude and latitude of a certain location point. Location semantic information usually includes semantic categories, such as schools, hospitals, and commercial areas. The location semantic LS is composed of a triple $\left\langle lon,lat,type\right\rangle$, and $type$ represents the semantic category.

Definition 4(Hierarchical temporal semantic tree [27]). According to the semantic and temporal features, a hierarchical temporal semantic tree is established, which is represented as a set of $G=\left(V,E,f\right)$. V records the set of nodes, and its value represents the category of location semantic attributes of different granularities. E represents the set of edges, and its value represents the relationship between two nodes. f is a label function used to assign semantic attributes to each node V_i in V .

Definition 5(Semantic sensitivity). Semantic sensitivity refers to the sensitivity of the semantic type of the location. The location semantic sensitivity $SS=\left(w_1,w_2,\cdots,w_k\right)$, w_i represents the weight of the *i*-th semantic category in the user's trajectory.

Definition 6(ε-geo-indistinguishability [28]). According to the definition of differential privacy, location differential privacy can be defined as follows: If a location privacy protection mechanism K satisfies ε-geo-indistinguishability, if and only if:

$$K\left(l\right)\left(l^*\right)\leq e^{\varepsilon}K\left(l^{'}\right)\left(l^*\right), \tag{1}$$

where, $l,l^{'}\in L$ and $l^*\in L^*$.

3.2. Scheme Design

As shown in Figure 1, the HTST-PTPP scheme is mainly divided into two parts: mobile client and server. The workflow is mainly composed of three parts: constructing hierarchical time semantic tree, evaluating location privacy protection requirements and personalized trajectory privacy protection. The detailed explanation of the three stages is as follows.

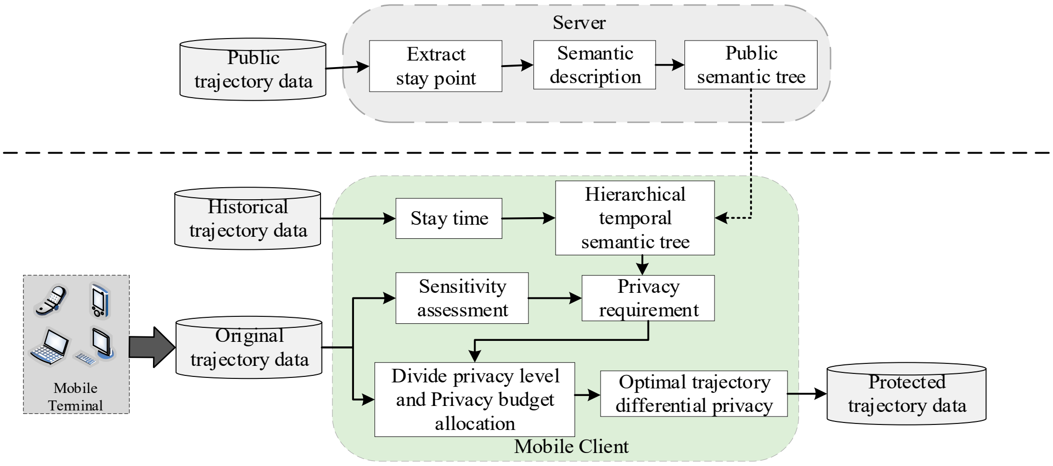


Figure 1. The workflow of HTST-PTPP scheme.

(1) Constructing hierarchical time semantic tree

The server processes the public trajectory data, extracts the stay points in the trajectory data, and uses the data fusion technology to mark the semantic information of the generated stay points, thereby constructing a common semantic tree. The mobile client analyzes the user's historical trajectory data, and constructs a personalized hierarchical time semantic tree for different users according to the residence time of the stay point and the common semantic tree downloaded from the server.

(2) Evaluating location privacy protection requirements

The mobile client uses the GPS trajectory data collected by the mobile device as the original data of the user, and extracts the stay points and semantic information markers from the original data. After that, by using the TF-IDF model, the semantic sensitivity of the location is calculated, and the privacy protection requirements of the location are quantified by combining the hierarchical time semantic tree.

(3) Personalized trajectory privacy protection

According to the privacy protection requirements of the location, different privacy risk levels are divided, the privacy budget is reasonably allocated, and differential privacy technology is used to achieve personalized trajectory privacy protection.

3.3. Attack Hypothesis

This paper assumes that the LBS server is an attacker who can obtain certain background knowledge. Probability distribution attack and location semantic attack are the two most common attacks. Probability distribution attack involves inferring the probability distribution of the user's location data to reveal the user's real location. The location semantic attack further explores the privacy of users by analyzing the semantic meaning of location information. In order to ensure the user's privacy information, each user processes the real trajectory on the mobile client before sending it to the server. Therefore, the LBS server does not obtain the user's real trajectory information to ensure that the user's trajectory privacy is fully protected.

4. Models and Algorithms

4.1. Constructing Hierarchical Time Semantic Tree

Before analyzing the trajectory data, it is necessary to process the data and extract the stay point. A stay point refers to a period of time in a specific geographical area, which usually indicates that the user has carried out some meaningful activities in the area. Figure 2 shows the extraction of stay points from trajectory data.

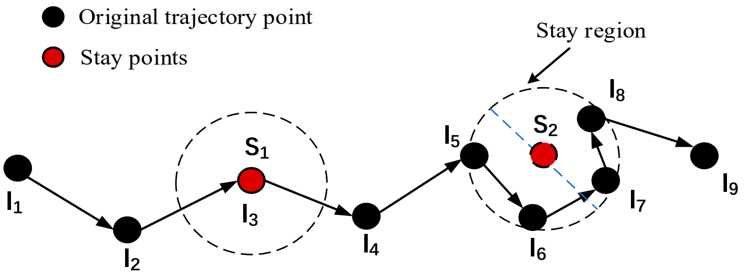


Figure 2. Trajectory point and Stay point.

We can observe that there are two situations, one is that the stay point S_1 appears at the original location point l_1, which is because the user remains stationary for a period of time, and the other is that the user is wandering in a certain geographical area, such as the stay point S_2. The latitude and longitude information of the stay point is calculated by Formula (2) and Formula (3):

$$S_i(lon)=\sum_{k=m}^{n}\frac{l_k(lon)}{|L|}, \tag{2}$$

$$S_i(lat)=\sum_{k=m}^{n}\frac{l_k(lat)}{|L|}. \tag{3}$$

The trajectory data composed of the extracted stay points can be expressed as $T\_S=S_1\rightarrow S_2\rightarrow\cdots S_n$.

We call the stay points in the user trajectory as 'locations'. The POI data set describes the geographical location covered by each semantic type. By fusing each location with the POI data set, we can effectively label each geographical location with its semantic category. In this paper, both the server and the mobile client need to annotate the semantic category of the stay points. On the server side, the distance between each stay point and the nearby POI is calculated, and the semantic category of the closest POI is used to describe the stop point. In the mobile client, a POI attribute table is set up, and the latitude and longitude information and semantic category are stored in this table and saved in the mobile client. The POI attribute table is used to label the semantic category of the stay point. In addition, the stay point of the mobile client needs to calculate the length of the stay based on the arrival time and the departure time.

The hierarchical temporal semantic tree is constructed based on the dwell time of the stay point and combined with the public semantic tree downloaded from the server. The construction of the public semantic tree is completed on the server side. According to the semantic category and common granularity of the public trajectory stop points, these stop points are divided into different clusters and used as the underlying nodes, such as: universities, hospitals, playgrounds, shopping centers, etc. The semantic categories of the underlying nodes are abstracted and summarized to form the upper nodes. This operation is iterated from the bottom up until the root node is obtained, thereby constructing a common semantic tree. The construction of hierarchical temporal semantic tree needs to build an intermediate layer first, and then expand up and down. The semantic category of the user's historical trajectory stay point and the cluster obtained by the common granularity are used as the nodes of the middle layer, and the common semantic tree is combined to abstract and summarize the semantics upward until the root node is found. According to the residence time, the clustering classification is performed iteratively from the middle layer downward, which means that the nodes with similar residence time are aggregated together to form a more detailed time hierarchy. Finally, through the above steps, a hierarchical temporal semantic tree with a specific number of layers is generated.

Figure 3 shows the construction process of hierarchical temporal semantic tree. The hierarchical temporal semantic tree includes leaf nodes and internal nodes. The leaf nodes are all visited POI nodes. The internal nodes represent the semantics of the generated cluster and generalized movement. The height of the leaf nodes is 0.

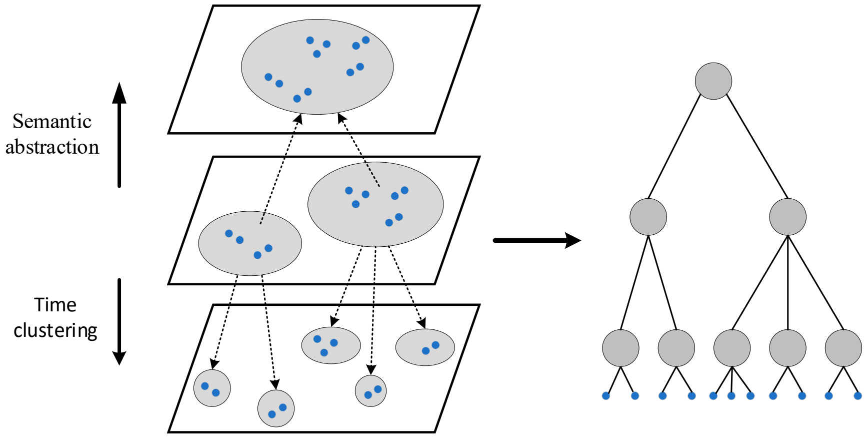


Figure 3. Construction of hierarchical temporal semantic tree.

Algorithm 1. Construction of common semantic tree algorithm

```
Input: Public trajectory data $GT$, POI dataset $P$
Output: Public semantic tree $C$
    1: stay_points=extract_stay_points ($GT$)
    2: stay_semantic_categories =match (stay_points, $P$)
    3: clusters = cluster_stay_points (stay_points, stay_semantic_categories)
    4: $C_0$ =Root (clusters)
    5: $C$ = abstract_semantic_categories ($C_0$)
    6: return $C$
```

Algorithm 1 shows the construction process of the common semantic tree, where the first row is to extract the stay points from the common trajectory data; in the second line, the extracted stay points are fused with the POI dataset, and the semantic category of each stay point is marked by calculation. In line 3, stay points are clustered into different clusters according to semantic categories and common granularity. Lines 4-5 represent that different clusters obtained by clustering the stay points are used as the bottom of the common semantic tree, and the semantic categories are abstracted upward to form a common semantic tree.

Algorithm 2. Hierarchical temporal semantic tree generation algorithm

```
Input: User history trajectory data $HT$, POI dataset $P$, Public semantic tree $C$
Output: Hierarchical temporal semantic tree $G$
    1: stay_points=extract_stay_points ($HT$)
    2: stay_time=calculate_stay_time (stay_points)
    3: stay_semantic_categories =match (stay_points, $P$)
    4: clusters = cluster_stay_points (stay_points, stay_semantic_categories)
    5: $G_0$ =Root (clusters)， $i = 0$
    6: $G$ = abstract_semantic_categories ($G_0$, $C$)
    7: $i = h(G)$
    8: while $i < h$ do
```

```
9:      for $j$ in range ($G_i$) do
10:         if $g_{i,j}$ $!=\varnothing$ then
11:             $G_{i+1}$ = cluster_stay_time ($g_{i,j}$,stay_time)
12:     $i+=1$
13: return $G$
```

Algorithm 2 is the construction process of hierarchical temporal semantic tree, where rows 1-3 are similar to the construction process of common semantic tree, indicating that the stay point and semantic annotation operation are extracted from the user 's historical trajectory data set, and the stay time of the user 's stay point is calculated. In the fourth line, according to the semantic category and the common granularity, the different clusters obtained by the stay point clustering are used as the middle layer ; lines 5-7 represent the abstraction of semantic categories from the middle layer up in combination with the common semantic tree ; lines 8-12 represent that the clustering operation is performed from the middle layer down according to the residence time of the stay point, thereby forming a hierarchical temporal semantic tree.

4.2. Evaluating Location Privacy Protection Requirements

The semantic sensitivity of the location is calculated based on the number of occurrences of the semantic category of the location in the user's trajectory. When a semantic category appears frequently in the user's trajectory and less frequently in other user's trajectories, it indicates that the user is more sensitive to the semantic category of the location, so the user may have a stronger demand for privacy protection of the location.

We use the TF-IDF model to measure the privacy sensitivity of each user's location. The TF-IDF model is a weighting technique widely used in information retrieval and text mining to evaluate the importance of a word in a single document in a file set or a word library, where TF represents the word frequency, that is, the frequency of a semantic category in the user's personal trajectory data set, and IDF represents the inverse document frequency, that is, the frequency of the semantic category in other user trajectories. This process is performed on the user's mobile client, so there is no trajectory data of other users. Based on this situation, a hypothesis is proposed, that is, there is a positive correlation between the access of public users to semantic categories and their demand for semantic attributes, and there is also a positive correlation between the demand for semantic attributes and the number of semantic categories in the POI library. TF and IDF can be calculated by Formula (4) and Formula (5).

$$TF_i = \frac{n_{iu}}{N_u}, \tag{4}$$

$$IDF_i = \log\frac{|N|}{|n_i|}, \tag{5}$$

where, n_{iu} represents the number of the *i*-th semantic category in the user trajectory data, $|n_i|$ represents the total number of the *i*-th semantic category in the POI database, N_u represents the number of stop points in the user trajectory data, and $|N|$ represents the total number of POIs.

The semantic sensitivity of the location in the user trajectory data can be calculated as:

$$w_i = TF_i \times IDF_i = \frac{n_{iu}}{N_u} \times \log\frac{|N|}{|n_i|}. \tag{6}$$

In order to establish a connection between the semantic privacy sensitivity of the location and the hierarchical temporal semantic tree, the semantic privacy sensitivity W is mapped to the temporal semantic tree for backtracking to the corresponding layers. The backtracking layer bt can be calculated by the following formula:

$$bt = \left(1 - \frac{w_j}{W}\right) O_l + \frac{w_j}{W} O_r, \tag{7}$$

where, $W = \sum_k w_k$.

The backtracking layer bt can be regarded as the user's demand for location privacy protection. The smaller the bt value, the lower the user's privacy protection demand for the location, while the larger the bt value, the higher the privacy protection demand. This is because the larger number of backtracking layers means that the distance between the root node and the root node is closer, and the corresponding semantic categories are more generalized.

4.3. Personalized Trajectory Privacy Protection

Trajectory data is usually generated by combining location data in chronological order, so trajectory privacy protection is mainly achieved through location privacy protection. In personalized trajectory privacy protection, according to the privacy protection requirements of each location point in the trajectory, different degrees of privacy protection are provided for each location point, so as to realize personalized protection. In order to achieve this goal, the user's privacy requirements need to be divided into different privacy levels, and the corresponding privacy budget is allocated for each level.

Firstly, according to the height h of the hierarchical temporal semantic tree, it is divided into four privacy levels. Then, find the privacy level according to the privacy protection requirements of each location. Table 1 shows the privacy level division and privacy budget allocation. The degree of privacy protection of differential privacy is mainly reflected by the size of privacy budget ε , and $\varepsilon > 0$. The smaller ε , the higher the degree of privacy protection. The degree of privacy protection required for the four privacy levels divided in this paper is increasing in turn. The privacy level $pl = \{le_1, le_2, le_3, le_4\}$ and the privacy budget $\varepsilon = \{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$ are defined, among which ε_1 to ε_4 are decreasing in turn.

Table 1. Privacy Level and Privacy Budget.

Privacy level (le)	Privacy protection requirements	Privacy budget
Privacy level 1- le_1	(0, $\frac{h}{4}$]	ε_1
Privacy level 2- le_2	($\frac{h}{4}$, $\frac{h}{2}$]	ε_2
Privacy level 3- le_3	($\frac{h}{2}$, $\frac{3h}{4}$]	ε_3
Privacy level 4- le_4	($\frac{3h}{4}$, h]	ε_4

In the pursuit of personalized trajectory privacy protection, attention should also be paid to data utility. It is necessary to comprehensively consider privacy requirements and data utility to ensure the availability of trajectory data while protecting user privacy to the greatest extent. In this paper, differential privacy technology is used to achieve privacy protection. Therefore, each location point in the trajectory data should meet the requirements of location differential privacy. It is also necessary to find a mechanism to minimize data quality loss. For mechanism $K : L \to O\left(L^*\right)$, where L denotes the set of the original position, and $O\left(L^*\right)$ denotes the set of the perturbation probability distribution on the position L. The input of the mechanism K is the original position l, and the output is the disturbance position l^*. K is expressed as a probability matrix, and k_{ll^*} is the probability from position l to position l^*.

Definition 7(Quality Loss). Given a priori probability π and a quality measure, the quality loss caused by mechanism K can be calculated as:

$$QL\left(K,\pi,d\right)=\sum_{l,l^*}\pi\left(l\right)k_{ll^*}d\left(l\cdot l^*\right), \tag{8}$$

where $d\left(l\cdot l^*\right)$ denotes the quality measure from position l to perturbation position l^*, and the prior probability $\pi\left(l\right)$ is the ratio of the number of occurrences of position l in the whole trajectory T_i to all records.

The linear optimization problem can be used to solve the minimum mass loss, and the following conditions need to be met:

$$\text{Min:} \quad \sum_{l,l^*\in L}\pi\left(l\right)k_{ll^*}d\left(l\cdot l^*\right) \tag{9}$$

$$\text{s.t.:} \quad k_{ll^*}\le e^{\varepsilon}k_{l'l^*} \qquad l,l',l^*\in L \tag{10}$$

$$\sum_{l^*\in L}k_{ll^*}=1 \qquad l\in L \tag{11}$$

$$k_{ll^*}\ge 0 \qquad l,l^*\in L \tag{12}$$

5. Performance Evaluation

5.1. Datasets and Experimental Setup

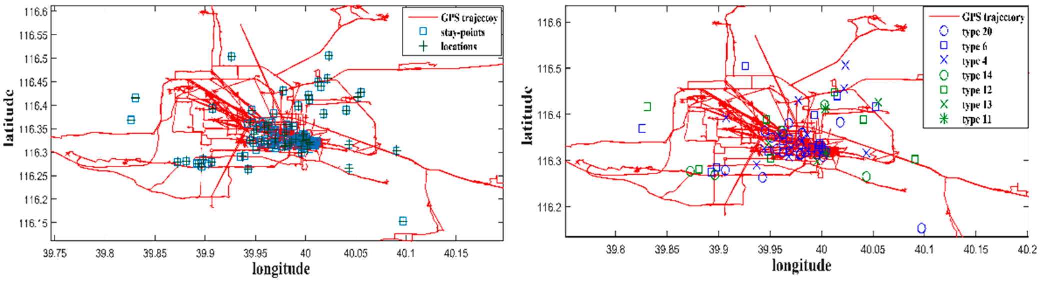
This paper uses Python to analyze trajectory data and verify the performance of the proposed HTST-PTPP method. This paper still uses the GeoLife [29] data set of Microsoft Asia Research Institute and the Beijing POI data set, and uses the GPS trajectory data set of the GeoLife project as the original trajectory data of the user. The GeoLife dataset collected trajectory data of 182 users from 2007 to 2012, including a total of 17,621 trajectories, covering a distance of more than 1.2 million kilometers and a total time of more than 48,000 hours. The dataset contains a variety of different types of trajectory data, from the user 's daily itinerary to the user's personalized activities. The Beijing POI dataset records the location information of most interest points in Beijing, including latitude and longitude coordinates, names, categories, etc. As shown in Table 2, the original POI dataset is divided into 20 different categories to label its semantic information for the stop point in the subsequent steps.

Table 2. Service types of Beijing POI dataset.

Type	Service	Type	Service
1	Food and beverage service	11	Motorcycle service
2	Road ancillary	12	Auto service
3	Name address	13	Vehicle repair

4	Scenic spot	14	Car sales
5	Public facilities	15	Commercial housing
6	Companies	16	Life service
7	Shopping service	17	Sports leisure
8	Traffic facilities	18	Health care
9	Financial insurance	19	Government agencies
10	Science and education	20	Accommodation services

Since the experiment only considers GPS trajectory data in Beijing, all data with latitude ranging from 115.4 to 117.6 and latitude ranging from 39.4 to 41.1 are selected from the GeoLife dataset. Then, the original trajectory is extracted, and the time threshold is set to 15 minutes, and the distance threshold is 200 meters. Finally, the Beijing POI data set is used to refer to the previous study [30] to label the semantic category for each stay point. Figure 4 shows the extraction of stay points and semantic category labeling of one user.



(a) stay points extracting (b) semantic category labeling

Figure 4. The example of stay points extracting and semantic categories labeling.

The GPS trajectory dataset is divided into training set and testing set. Training set includes 80% location data of the trajectory dataset, which can be used to construct a global prior. It can be taken as the average of the individual prior probabilities of all users visiting the area. Then, the obtained average value is used in the post-mapping mechanism to obtain the location of the optimal service quality loss. Testing set includes 20% location data of the trajectory dataset, which can be used to evaluate the mechanisms. It constructs a user-specific prior for at least 20 users and measures the service quality loss of the mechanism when users use their own prior.

5.2. Experimental Results and Performance Analysis

This paper verifies the performance of the proposed HTST-PTPP method from two aspects of data availability and privacy protection, and compares it with the PLDP-TD [31] method and the TPP-POIs [27] method respectively.

The PLDP-TD method is a personalized noise trajectory tree structure for personal privacy. This method assumes that each location point on the map has different privacy preferences, but the privacy preference of its location is determined by the location itself and has nothing to do with the individual's behavior. In other words, the privacy preference of the same location is the same for different users. The TPP-POIs method is a privacy protection method based on trajectory reconstruction. This method first labels the semantic attributes of all sampling points on the trajectory and establishes a corresponding classification tree, and then extracts sensitive stay points. Different strategies are used to select POIs to replace different types of sensitive points to complete trajectory reconstruction.

(1) Quality loss

In terms of data availability, Quality Loss is used as its evaluation indicator. Quality loss is proposed by Reference [32] and described in detail in Definition 7. Quality loss measures the degree of interference between the output trajectory and the input trajectory. The greater the mass loss, the greater the difference between the disturbance trajectory data and the original trajectory data, and the lower the data availability.

Figure 5 shows the performance comparison of the HTST-PTPP method proposed in this paper with the PLDP-TD method and the TPP-POIs method in terms of quality loss. It can be seen from the figure that as the privacy budget increases, the quality loss of the three methods decreases, mainly because the privacy protection strength decreases, thereby increasing the similarity between the generated perturbation position and the original position. However, the HTST-PTPP method has less quality loss than the other two methods. The main reason is that the HTST-PTPP method not only considers the semantic information of the location and constructs a common semantic tree, but also constructs a hierarchical temporal semantic tree for each user according to the personalized needs of the user. Through hierarchical privacy protection, the perturbation location is generated for each original location, which effectively reduces the quality loss in the process of trajectory privacy protection. The TPP-POIs method considers the semantic information of the location, adds semantic annotations to each location, and selects the disturbance location by the classified location. The PLDP-TD method lacks the consideration of the semantic information of the location, so the TPP-POIs method has less quality loss than the PLDP-TD method.

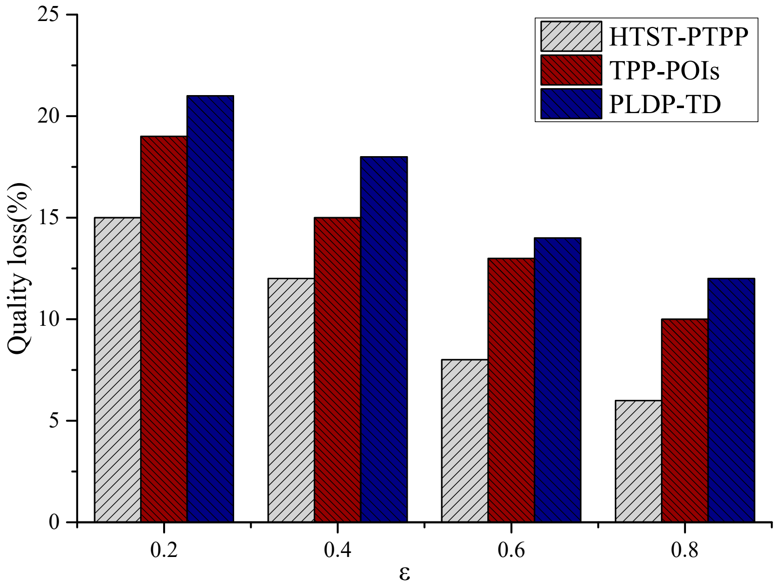


Figure 5. The impact of privacy budget on quality loss.

(2) Data utility

Suppose that T_i and T_i^* are the actual trajectory sequence and privacy-preserved trajectory sequence of user *i*, respectively. The data utility of a query Q can be computed as:

$$DU = \frac{\max\{Q(T), s\}}{\left|Q(T^*) - Q(T)\right|}, \tag{13}$$

where, s is negligible.

Figure 6 shows the performance comparison of the HTST-PTPP method proposed in this paper with the PLDP-TD method and the TPP-POIs method in terms of data utility. It can be seen from the figure that as the privacy budget increases, the data utility of the three methods increases, mainly because the quality loss with privacy protection decreases, and the similarity between the generated perturbation position and the original position increases, thereby enhancing the availability of data. However, the HTST-PTPP method has higher data utility than the other two methods. The main reason is that the hierarchical temporal semantic tree is considered by the HTST-PTPP method, which can give a more precise description of user preferences for personalized privacy protection. The TPP-

POIs method only considers the semantic information of locations and lack of consideration of time. It leads to a decrease in accuracy when describing user preferences. Because the PLDP-TD method dose not consider the semantic information of each location, the data utility is the lowest compared to the other two methods.

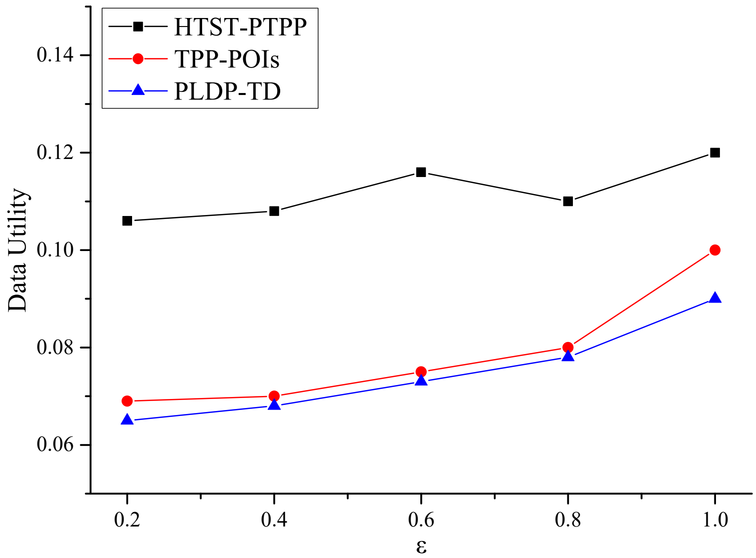


Figure 6. The impact of privacy budget on data utility.

(3) Privacy protection

In terms of privacy protection, Adversarial Error is used as its evaluation index. The adversarial error is proposed by Reference [33], which measures the expected error of the adversary in the inference attack. It can be calculated as:

$$AdvError(\pi,K,H,d)=\sum_{l,l',l^*\in P}\pi(l)k_{ll^*}h\left(l'|l^*\right)d(l,l^*), \tag{14}$$

where, $h\left(l'|l^*\right)$ denotes the probability of mapping l^* to l'.

The adversarial error represents the expected distortion in the reconstruction event, which means that the greater the adversarial error, the greater the difference between the attacker's predicted position and the real position, and the stronger the privacy protection.

Figure 7 shows the performance comparison of the HTST-PTPP method proposed in this paper with the PLDP-TD method and the TPP-POIs method in terms of adversarial errors. From the graph, it can be seen that with the increase of privacy budget ε , the adversarial error of the three methods decreases. This is because the privacy protection intensity decreases with the increase of privacy budget ε value, and the corresponding need to add less noise, which leads to the decrease of the adversarial error. When the privacy budget ε is the same, the HTST-PTPP method has a large adversarial error compared with the other two methods. The main reason is that the HTST-PTPP method not only considers the semantic information of the location, but also considers the user 's residence time, providing users with a more personalized privacy requirement. Although the TPP-POIs method also considers the semantic information of the location, it lacks the consideration of the time factor, so the adversarial error is lower than the HTST-PTPP method. The PLDP-TD method only considers the privacy preference of the location point, ignoring the influence of semantic factors and time factors. Therefore, the HTST-PTPP method has better privacy protection.

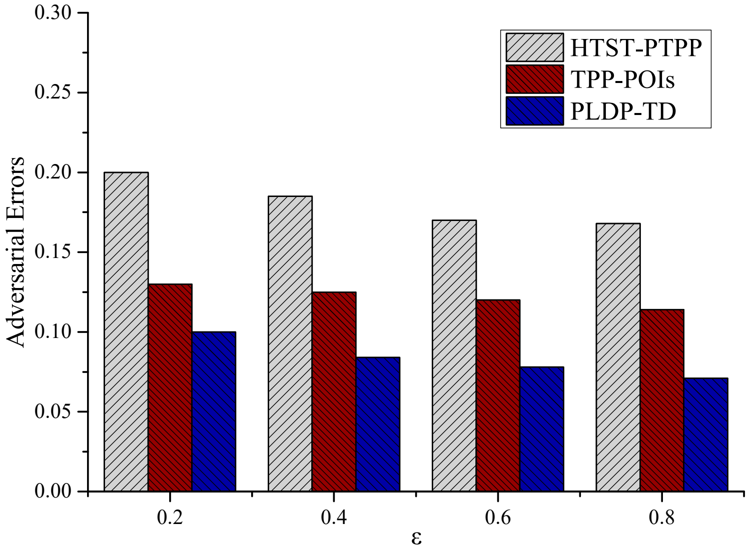


Figure 7. The impact of privacy budget on adversarial errors.

(4) Execution time

In this part, the execution time is also taken as an index to compare the performance of three methods.

In general, it can be seen from Figure 8 that the execution time of the HTST-PTPP method is lower than the other two methods. This is because the hierarchical temporal semantic tree is constructed for each user, which can shorten the execution time of the algorithm and improve the query efficiency. What's more, the execution time of HTST-PTPP method drops dramatically as the privacy strength decreases. The TPP-POIs method has the lower execution time than the PLDP-TD method because the semantic information of each location is not considered by TPP-POIs method when protecting the location privacy of users.

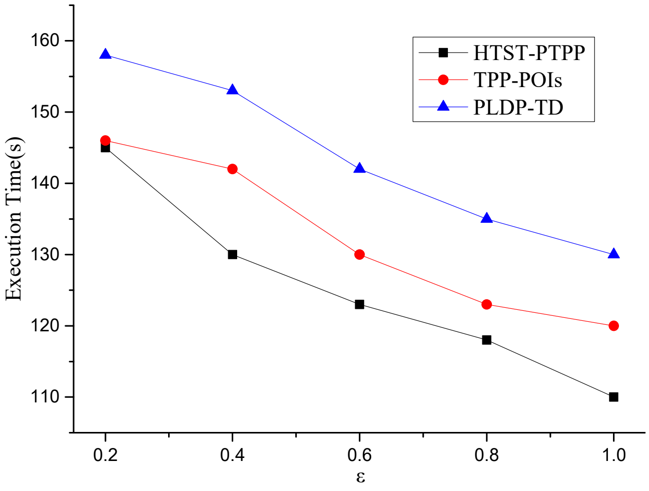


Figure 8. The impact of privacy budget on execution time.

6. Conclusions and Future Work

This paper mainly studies the problem of personalized trajectory privacy protection, and proposes a personalized trajectory privacy protection scheme based on hierarchical temporal semantic tree (HTST-PTPP). Firstly, by considering the semantic category and residence time of the user 's historical trajectory stay point, and combining with the common semantic tree, a personalized hierarchical time semantic tree is constructed. Then, the TF-IDF algorithm is used to calculate the semantic sensitivity of each location in the user 's trajectory, and then the privacy requirements of each location are quantified. Then, the privacy level is divided, the privacy budget is reasonably allocated, and the differential privacy technology is used to realize personalized trajectory privacy protection. Finally, two real data sets are used to verify the proposed HTST-PTPP method. The results show that compared with the previous methods, this method has better performance in data availability and privacy protection. Looking ahead, we will try to combine differential privacy with other privacy protection technologies to achieve higher security and flexibility.

Author Contributions: Conceptualization, L.Z.; formal analysis, P.X.; funding acquisition, L.Z.; methodology, P.X. and L.Z.; project administration, M.H.; resources, M.H.; software, Y.W.; supervision, P.X.; validation, P.X. and M.H.; writing—original draft, L.Z.; writing—review and editing, L.Z. and P.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Project of Hainan Province under Grant No. ZDYF2024SHFZ264.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of the study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jiang, H., Pei, J., Yu, D., Yu, J., Gong, B., Cheng, X. Applications of Differential Privacy in Social Network Analysis: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, **2023**, 35, 108-127.
2. Zhu, H., Liu, W., Yin, J., Zheng, L., Huang, X., Xu, J., Lee, W. Continuous Geo-Social Group Monitoring in Dynamic LBSNs. *IEEE Transactions on Knowledge and Data Engineering*, **2023**, 35, 7815-7828.
3. Liu, Z., Zhang, H., Ouyang, G., Chen, J., Wu, K. Data-Driven Pick-Up Location Recommendation for Ride-Hailing Services. *IEEE Transactions on Mobile Computing*, **2024**, 23, 1001-1015.
4. Lee, W., Tseng, S., Shieh, J., et al. Discovering Traffic Bottlenecks in an Urban Network by Spatiotemporal Data Mining on Location-Based Services. *IEEE Transactions on Intelligent Transportation Systems*, **2011**, 12, 1047-1056.
5. Wang, H., Wang, C., Zhou, K., Liu, D., Zhang, X., Cheng, H. TEBChain: A Trusted and Efficient Blockchain-Based Data Sharing Scheme in UAV-Assisted IoV for Disaster Rescue. *IEEE Transactions on Network and Service Management*, **2024**, 21, 4119-4130.
6. Xue, X., Huangfu, S., Zhang, L., Wang, S. Research on Escaping the Big-Data Traps in O2O Service Recommendation Strategy. *IEEE Transactions on Big Data*, **2021**, 7, 199-213.
7. Xu, C., Ding, Y., Chen, C., Ding, Y., Zhou, W., Wen S. Personalized Location Privacy Protection for Location-Based Services in Vehicular Networks. *IEEE Transactions on Intelligent Transportation Systems*, **2023**, 24, 1163-1177.
8. Ardagna, C., Cremonini, M., De Capitani di Vimercati, S., Samarati, P. An Obfuscation-Based Approach for Protecting Location Privacy. *IEEE Transactions on Dependable and Secure Computing*, **2011**, 8, 13-27.
9. Zhang, T., Zhu, T., Liu, R., Zhou, W. Correlated data in differential privacy: Definition and analysis. *Concurrency and Computation: Practice and Experience*, **2022**, 34(16).
10. Wang, T., Zheng, Z., Rehmani, M. H., Yao, S., Huo, Z. Privacy Preservation in Big Data from the Communication Perspective—A Survey. *IEEE Communications Surveys and Tutorials*, **2019**, 21, 753-778.
11. Gao, S., Ma, J., Shi, W., Zhan, G., Sun, C. TrPF: A Trajectory Privacy-Preserving Framework for Participatory Sensing. *IEEE Transactions on Information Forensics and Security*, **2013**, 8, 874-887.
12. Hemkumar, D., Ravichandra, S., Somayajulu, D. V. L. N. Impact of prior knowledge on privacy leakage in trajectory data publishing. *Engineering Science and Technology an International Journal*, **2020**, 23, 1291-1300.
13. Shaham, S., Ding, M., Liu, B., Dang, S., Lin, Z., Li, J. Privacy Preserving Location Data Publishing: A Machine Learning Approach. *IEEE Transactions on Knowledge and Data Engineering*, **2021**, 33, 3270-3283.
14. Wu, X., Sun, G. A Novel Dummy-Based Mechanism to Protect Privacy on Trajectories. *IEEE International Conference on Data Mining Workshops*, **2015**, 1120-1125.
15. Zhang, J., Wang, X., Yuan, Y., Ni, L. RcDT: Privacy Preservation Based on R-Constrained Dummy Trajectory in Mobile Social Networks. *IEEE Access*, **2019**, 7, 90476-90486.
16. Wang, W., Wang, Y., Duan, P., Liu, T., Tong, X., Cai, Z. A Triple Real-Time Trajectory Privacy Protection Mechanism Based on Edge Computing and Blockchain in Mobile Crowdsourcing. *IEEE Transactions on Mobile Computing*, **2023**, 22, 5625-5642.
17. Wang, Z., Zhu, Y., Wang, D., Han, Z. Secure Trajectory Publication in Untrusted Environments: A Federated Analytics Approach. *IEEE Transactions on Mobile Computing*, **2023**, 22, 6742-6754.

18. Chen, R., Fung, B. C. M., Mohammed, N., Desai, B. C., & Wang, K. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, **2013**, 231, 83-97.
19. Dwork, C. Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, **2008**, 1–19.
20. Wang, H., Xu, Z. CTS-DP: Publishing correlated time-series data via differential privacy. *Knowledge-Based Systems*, **2017**, 122, 167-179.
21. Ghane, S., Kulik, L., Ramamohanarao, K. TGM: A Generative Mechanism for Publishing Trajectories with Differential Privacy. *IEEE Internet of Things Journal*, **2020**, 7, 2611-2621.
22. Yang, Z., Wang, R., Wu, D., Wang, H., Song, H., Ma, X. Local Trajectory Privacy Protection in 5G Enabled Industrial Intelligent Logistics. *IEEE Transactions on Industrial Informatics*, **2022**, 18, 2868-2876.
23. Zheng, Z., Li, Z., Jiang, H., Zhang, L. Y., Tu, D. Semantic-Aware Privacy-Preserving Online Location Trajectory Data Sharing. *IEEE Transactions on Information Forensics and Security*, **2022**, 17, 2256–2271.
24. Wu, L., Qin, C., Xu, Z., Guan, Y., Lu, R. TCPP: Achieving Privacy-Preserving Trajectory Correlation with Differential Privacy. *IEEE Transactions on Information Forensics and Security*, **2023**, 18, 4006–4020.
25. Wu, L., Qin, C., Xu, Z., Guan, Y., Lu, R. TCPP: Achieving Privacy-Preserving Trajectory Correlation with Differential Privacy. *IEEE Transactions on Information Forensics and Security*, **2023**, 18, 4006-4020.
26. Chen, C., Hu, X., Li, Y., Tang, Q. Optimization of Privacy Budget Allocation in Differential Privacy-Based Public Transit Trajectory Data Publishing for Smart Mobility Applications. *IEEE Transactions on Intelligent Transportation Systems*, **2023**, 24, 15158-15168.
27. Dai, Y., Shao, J., Wei, C., Zhang, D., Shen, H. T. Personalized semantic trajectory privacy preservation through trajectory reconstruction. *World Wide Web*, **2018**, 21, 875–914.
28. Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., Palamidessi, C. Geo-indistinguishability: Differential privacy for location-based systems. *ACM Conference on Computer and Communications Security*, **2013**, 901-914.
29. Zheng, Y., Xie, X., Ma, W. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data(base) Engineering Bulletin*, **2010**, 33, 32-39.
30. Zhu, L., Xu, C., Guan, J., Zhang, H. SEM-PPA: A semantical pattern and preference-aware service mining method for personalized point of interest recommendation. *Journal of Network and Computer Applications*, **2017**, 82, 35-46.
31. Deldar, F., Abadi, M. PLDP-TD: Personalized-location differentially private data analysis on trajectory databases. *Pervasive and Mobile Computing*, **2018**, 49, 1-22.
32. Levina, E., Bickel, P. The Earth Mover's distance is the Mallows distance: Some insights from statistics. *IEEE International Conference on Computer Vision*, **2001**, 2, 251.
33. Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J. P., le Boudec, J. Y. Protecting location privacy: Optimal strategy against localization attacks. *ACM Conference on Computer and Communications Security*, **2012**, 617-627.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.