DOI: 10.6688/JISE.20220X_38(X).00XX

Hierarchical Temporal Semantic Tree Based Personalized Trajectory Privacy Protection Scheme over LBSNs

PEI-XU XING[1,2], MENG-XING HUANG[1,+], LIANG ZHU[2,+] AND YUAN-YUAN WU[1]

[1] *College of Information and Communication Engineering, Hainan University, Haikou 570228, China*

[2] *College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China*

The evolution of mobile networks, particularly with the advent of 5G and the upcoming 6G technologies, presents substantial concerns regarding privacy protection. Most existing privacy protection methods offer the uniform level of privacy safeguards for targeted users. This is especially true in the context of Location-Based Social Networks (LBSNs), where prevailing methodologies predominantly emphasize the spatial and temporal dimensions of trajectory data, while largely neglecting the semantic information associated with locations. To address these issues, the Hierarchical Temporal Semantic Tree based Personalized Trajectory Privacy Protection (HTST-PTPP) scheme is proposed in this paper. It mainly constructs the hierarchical temporal semantic tree by considering semantic categories and residence time. Then, the privacy requirements of different users are quantified according to the differential privacy model, in order to provide the personalized privacy protection for users. Finally, the 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 wireless communication networks and the rapid development of intelligent mobile terminals, the Location-Based Services (LBS) have been widely used in various fields of life (e.g. travel services, social media, health and motion tracking, etc.) [1-3]. In the Location Based Social Networks (LBSNs), the personalized services are provided to the target users by analyzing the location datasets. Also, the new dimension is added to form the social media with geographical location information (e.g., documents, pictures, audio or video, etc.) and record the historical data about location of users [4]. The locations have become the new object in LBSNs, which promotes the development of various LBS, such as smart city [5], smart transport [6] or disaster rescue [7], etc. It is the bridge connecting the virtual world and the real society, so as to realize the new Online to Offline (O2O) coordination mechanism [8]. Namely, Online users share the real experience of the real society in LBSNs, while offline servers make use of the historical information to analyze the preference and behavior of users. It promotes the research on the correlation between the real activity status and social activity characteristics of users [9-11]. Rich LBS bring great convenience to the lives of users, but also cause the risk of

Received March 14, 2020; revised June 20, 2020; accepted September 28, 2020.
Communicated by Editor Jiann-Liang Chen.

[+] Corresponding author：Huangmx09@163.com; lzhu@zzuli.edu.cn.

leakage of the personal privacy (e.g. identity, location or query information, etc.) [12]. For LBSNs, the users need to upload the real location information to the servers, so as to obtain the personalized service. Thus, the users not only enjoy the great convenience of location services, but also bear the risk of privacy leakage [13-14].

The existing trajectory privacy protection methods can generalize the actual locations to fake locations by considering the spatial and temporal dimensions of trajectory data. However, the methods lack of the unified degree of privacy protection, which results the data utility is reduced. The semantic information refers to the abstract related to the geographical locations. It provides deeper information about the locations, which can be utilized to analyze the preference of target users and provide the personalized privacy protection for users. The trajectory privacy protection considering semantic information can improve the data utility. However, the same semantic category also has different privacy requirements due to the different residence times of users. For example, the semantic type of location A is cinema. The user 1 and user 2 stay in location A for eight hours and two hours respectively. It can be inferred that the user 1 works in location A and the user 2 plays in location A. The two users have different privacy needs. Therefore, it is necessary to consider the 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 the above problems, the Hierarchical Temporal Semantic Tree based Personalized Trajectory Privacy Protection (HTST-PTPP) scheme is proposed in this paper. The contributions can be concluded as follows.

(1) The hierarchical temporal semantic tree of each user is built by considering the semantic category and residence time, so as to acquire the personalized behavior and preference information of users.

(2) The privacy level division strategy is proposed, which can calculate the privacy sensitivity of users for each location by utilizing the TF-IDF algorithm. Also, the privacy level is divided by considering differential privacy model to satisfy the different privacy protection demands of users.

(3) The experimental study is conducted on two real data sets, so as to verify the data availability and privacy protection of the proposed HTST-PTPP scheme.

The remainder of this paper is organized as follows. Section 2 describes the related works of trajectory privacy protection in detail. The HTST-PTPP scheme is presented in Section 3. Section 4 shows the generation process of the hierarchical temporal semantic tree and the personalized trajectory privacy protection. The privacy and usability of our scheme is evaluated in Section 5. Section 6 concludes our solutions and the future works.

2. RELATED WORK

In this section, the related research is conducted in two aspects. Firstly, the development process about trajectory privacy protection is explained. Second, the references about privacy quantification are described.

2.1 Trajectory Privacy Protection

Generalization-based trajectory privacy protection technology [15] 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. [16] proposes the anonymization method including two stages as virtualization and suppression. The virtualization method is used as the alternative mechanism for sensitive attributes, and the suppression method is used as the anonymous mechanism for user trajectories, so that users can resist multiple attacks. Shaham et al. [17] proposes the anonymous method for spatial-temporal trajectory data sets, where the machine learning algorithms are used to cluster trajectories. In addition, a variant of the *k*-means algorithm is proposed to prevent the leakage of over-sensitive datasets.

The core of trajectory privacy protection based on fake trajectory is to generate a series of fake trajectories by using the real moving trajectory of users. The fake trajectories are mixed with the original trajectory. Wu et al. [18] proposes the 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. Also, the distribution of generated false locations is more uniform, so as to meet the stricter privacy protection requirements. Zhang et al. [19] proposes the virtual trajectory privacy protection scheme based on radius constraint, which is used to enhance the trajectory privacy protection of users in the mobile social network. It can effectively reduce the exposure risk of the single location and trajectory. Wang et al. [20] proposes the 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 locations in the trajectory. The method based on suppression is simple to implement. However, it is easy to cause information loss and data availability decrease. Wang et al. [21] proposes the secure trajectory publishing mechanism based on federated analysis, which can achieve stronger data privacy protection locally without sharing the original data. It solves the problem that the trusted server is attacked in the traditional trajectory publishing framework. Chen et al. [22] proposes the local suppression method to achieve customizable trajectory anonymization. It improves the data utility by addressing the problems of high dimension, sparsity and sequence in trajectory anonymization. Hu et al. [23] proposes the security-enhanced data sharing scheme with location privacy preservation. The homomorphic encryption is utilized to aggregate data at the edge of the network to achieve privacy protection of users.

2.2 Privacy Quantification

The purpose of privacy quantification is to transform the location information of a specific user into the numerical datasets, so as to realize the measurement of location privacy and ensure the acquisition of the key information. The common methods of privacy quantification include information entropy, mutual information, decision tree and attack model, etc. Differential privacy is the popular model of privacy protection proposed by Dwork et al. [24] in 2006, which solves the two defects of the traditional model of privacy protection. It does not need to consider the background knowledge of the attacker because of the strict mathematical definition. Wang et al. [25] proposes the scheme for publishing time-related location data based on differential privacy, which makes it impossible for attackers to distinguish the noise trajectory from the original trajectory. Ghane et al. [26] proposes the trajectory generation algorithm, which the location data is modeled as the graph. It retains the information of the real trajectories (e.g. spatial information, temporal information, distance and stay location), in order to effectively improve the computational efficiency and practicability. Based on the technology of location generalization and local differential privacy, Yang et al. [27] proposes the trajectory data perturbation method based on quadtree index. The correlation of the adjacent spatial-temporal nodes of the trajectory is considered to protect the trajectory privacy of users. Zheng et al. [28] proposes the trajectory privacy protection method for the sharing of the semantic-sensitive trajectory data. The data privacy and semantic privacy of users are both protected by considering the balance between privacy and utility. Wu et al. [29] proposes the privacy protection mechanism related to confidentiality that satisfies differential privacy. The trajectory association problem between multiple users is considered, so as to realize the protection of trajectory association between multiple users. Wu et al. [30] proposes the trajectory correlation privacy-preserving mechanism (TCPP) that fulfills differential privacy. The mechanism can protect the trajectory correlation based on the customized privacy budget allocation strategy. Chen et al. [31] proposes the development of the optimal privacy budget allocation algorithm for the transit smart card data. The goal is to publish the non-interactive sanitized trajectory data under the differential privacy definition. Min et al. [32] proposes the semantic adaptive geo-indistinguishability mechanism by adding random noise to users' locations, in order to quantify personalized location privacy.

The above research has made significant strides in the trajectory privacy protection and privacy quantification through various methodologies. It remains a notable gap in adequately considering the semantic information of locations and the individualized preferences of users. Therefore, the research on the personalized location privacy protection is needed to enhance the performance of data availability and privacy protection.

3. OVERVIEW OF SCHEME

This section provides the comprehensive definition of the problem, outlines the scheme design, and specifies the attack hypotheses.

3.1 Problem Definition

Definition 1(Trajectory sequence). The trajectory sequence is the combination of the location points, which connects the single coordinate collected by GPS in chronological order. Each location point p_i is composed of the triple $\langle lon\_p_i, lat\_p_i, t\_p_i \rangle$, where lon_p_i and lat_p_i represent the longitude and latitude, respectively. t_p_i represents the time that the user passes through the location point. The trajectory sequence Tra_p can be expressed as $Tra\_p = p_1 \to p_2 \to \cdots \to p_n$.

Definition 2(Stay-points). The stay-point s_i is the clustering of the original location points, indicating that the user stays in a certain geographical area for a period of time.

Given the distance threshold θ_d and the time threshold θ_t, for a set of consecutive location points $\mathrm{P}=\{p_m,p_{m+1},\cdots,p_n\}$, where $m<k\le n$, $Dist(p_m,p_k)\le\theta_d$ and $Int(p_m,p_n)\ge\theta_t$. Each stay-point s_i consists of a triple $\langle lon\_s_i,lat\_s_i,t\_s_i\rangle$, where lon_s_i and lat_s_i represent the longitude and latitude, respectively. t_s_i represents the length of the time spent at the stay-point s_i.

Definition 3(Semantic label). The semantic label of each stay-point usually includes the geographic information and the semantic information. Geographic information refers to the longitude and latitude of a certain location point. Semantic information usually includes semantic categories, such as schools, hospitals or commercial areas. The semantic label of stay-points s_i is composed of the triple $\langle lon\_s_i,lat\_s_i,type\_s_i\rangle$. $type\_s_i$ represents the semantic type of each location.

Definition 4(Hierarchical temporal semantic tree [33]). According to the semantic type $type\_s_i$ and residence time t_s_i, the hierarchical temporal semantic tree is established, which is represented as a set of $G=(V,E,f)$. V records the set of nodes, which represents the type of the semantic attributes for the different granularities. E represents the set of edges, which represents the relationship between two nodes. f is the label function, which can be used to assign semantic attributes to each node V_i in V.

3.2 Scheme Design

As shown in Fig. 1, the HTST-TPP scheme is mainly divided into mobile client and server. The workflow is composed of three stages, which are 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.

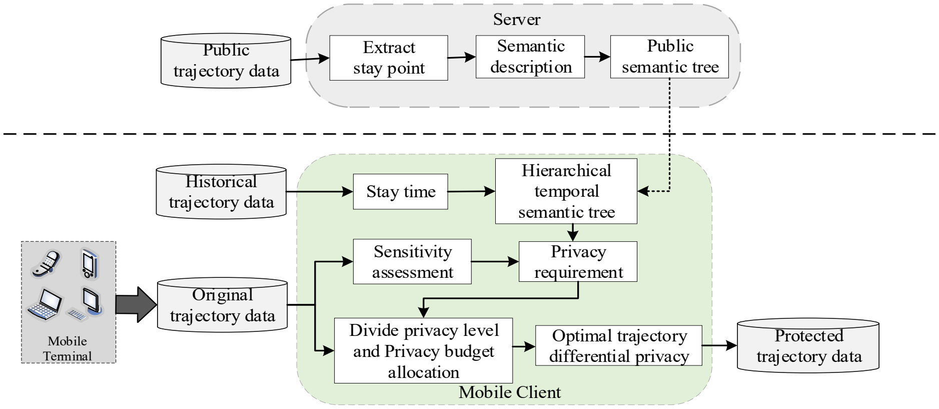

Fig. 1. The workflow of HTST-PTPP scheme.

(1) Constructing hierarchical time semantic tree

The server undertakes the processing of public trajectory data to extract the stay-points within the trajectory. Subsequently, it employs data fusion techniques to annotate the extracted stay-points with semantic information, thereby facilitating the construction of the common semantic tree. In parallel, the mobile client analyzes the historical trajectory sequence of individual users, so as to construct the personalized hierarchical time semantic tree. The personalized tree is developed based on the duration of residence at each stay-point. It is aligned with the common semantic tree retrieved from the server.

(2) Evaluating location privacy protection demands

The mobile client utilizes GPS trajectory data collected from the mobile device as the primary datasets for target user. It extracts the stay-points and the corresponding semantic label from the original datasets. Subsequently, the semantic sensitivity of each location is assessed by using the model of Term Frequency-Inverse Document Frequency (TF-IDF). The assessment is further enhanced by quantifying the privacy protection demands associated with each location, so as to incorporate the hierarchical time semantic tree to provide a comprehensive analysis.

(3) Personalized trajectory privacy protection

According to the identified privacy protection demands for each location, the study classifies various levels of privacy risk. The rational allocation of the privacy budget is established. Also, the techniques of differential privacy are implemented to ensure personalized privacy protection for the trajectory sequence.

3.3 Attack Hypothesis

This study assumes that the Location-Based Service (LBS) server is honest and curious. It might act as the attacker who has the capability to obtain the sensitive information of the target users. This paper focuses on the probabilistic distribution attacks and the location semantic attacks. The probabilistic distribution attacks mainly involve analyzing the probability distribution of location data. The attacker can infer the actual location of the target user. The location semantic mainly analyze the semantic information of the location to uncover the behavioral preferences of the target user, thereby inferring the places where the user is likely to go next.

To protect the privacy of users, each user needs to preprocess the actual trajectory data on the mobile client before uploading it to the server. Therefore, the LBS server cannot receive the real trajectory information of the target user, so as to prevent the leakage of the user's sensitive information.

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-points. Based on our previous research [38], the stay-point refers to a period of time in the specific geographical area. It usually indicates that the user has carried out some meaningful activities in the area.

Fig. 2 shows the flow of stay-point extraction and semantic information labeling. There are two spaces for the historical trajectory of users, one is geographic space and another one is semantic space. Firstly, the original points p_i are clustered into stay-points s_i according to the characteristics of geographical location in the geographical space. Then, in the semantic space, the feature vector of the stay area is calculated, and the stay-points are divided into different semantic types C_i. Finally, the feature vectors of the stay-points in the semantic space are clustered, so that each stay area has certain semantic information. It can be used for the following temporal hierarchical semantic tree construction.

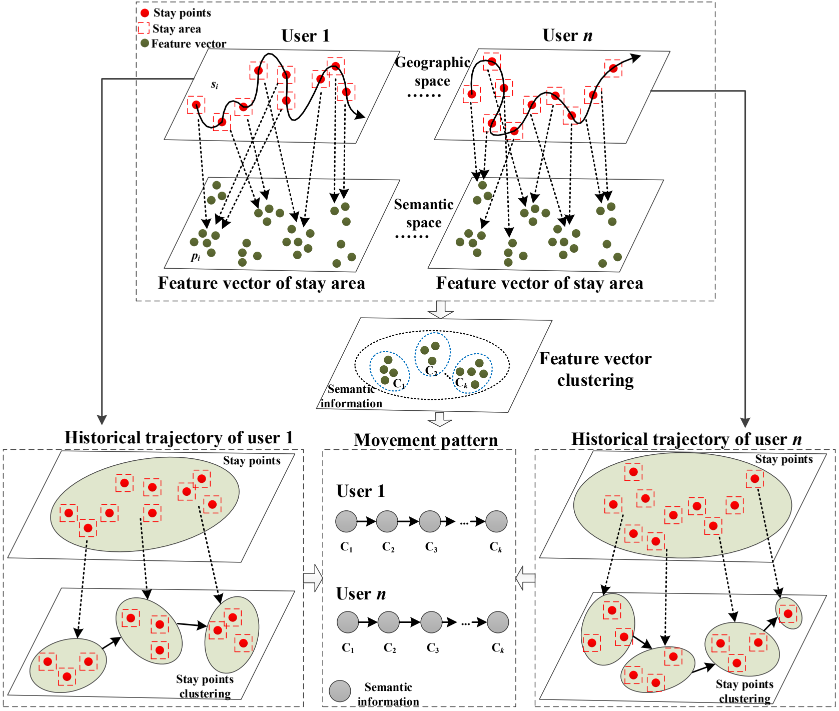

Fig. 2. The flow of stay-points extraction and semantic information labeling.

The latitude and longitude information of the stay-point is calculated by Formulas (1) and (2).

$$s_i(lon) = \sum_{k=m}^{n} \frac{p_k(lon)}{|Tra\_p|}. \tag{1}$$

$$s_i(lat) = \sum_{k=m}^{n} \frac{p_k(lat)}{|Tra\_p|}. \tag{2}$$

The trajectory data composed of the extracted stay-points can be expressed as $Tra\_s = s_1 \to s_2 \to \cdots s_n$.

In this paper, the stay-points in the user trajectory are called as locations. The POI data sets describe the geographical location covered by each semantic type. By fusing each location with the POI datasets, each geographical location can be labeled with its semantic category. 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, the POI attribute table is set up. The latitude and longitude information and semantic category are stored in the 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 duration that the user stays at a certain location can be calculated based on the arrival and departure times of the stay-points.

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 location points, the stay-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 historical stay-points and the cluster obtained by the common granularity are used as the nodes of the middle layer. 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 the more detailed time hierarchy. Finally, through the above steps, the hierarchical temporal semantic tree with the specific number of layers is generated.

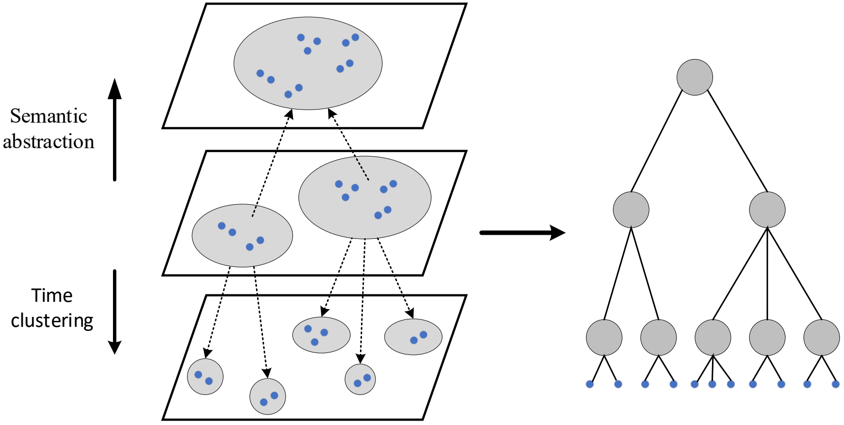

Fig. 3. The flow of stay-points extraction and semantic information labeling.

Fig. 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.

Algorithm 1. Public semantic tree generation algorithm

```
Input: Public trajectory data $Tra\_p = \{p_1, p_2, ..., p_n\}$, Semantic information $C_i$
Output: Public semantic tree $C$
    Initialize the set of stay-points $S \leftarrow \varnothing$, the public semantic tree $C \leftarrow \varnothing$;
    Define $i \leftarrow 0$, $len \leftarrow |Tra\_p|$;
    while $i < len$ do
        $j \leftarrow i+1$, $flag \leftarrow 0$;
```

```
5:      while $i < len$ do
6:          if $D(p_i, p_j) < \theta_d$ then
7:              if $T(p_i, p_{j-1}) > \theta_t$ then
8:                  $s_j(x) \leftarrow \sum_{k=i}^{j-1} \frac{p_k(x)}{j-i}$, $s_j(y) \leftarrow \sum_{k=i}^{j-1} \frac{p_k(y)}{j-i}$;
9:                  put $s_j$ into the set $\mathcal{S}$, $i \leftarrow j$, $flag \leftarrow 1$;
10:                 compute the semantic type of $s_j$;
11:             end if
12:             break;
13:         end if
14:         $j \leftarrow j+1$;
15:     end while
16:     if $flag != 1$ then
17:     $i \leftarrow i+1$;
18:     end if
19: end while
20: Cluster the stay-points with the same semantic type;
21: Select the root node $C_0$;
22: Generate the hierarchical public semantic tree $C$ according to semantic type;
23: Return $C$
```

Algorithm 1 shows the generation process of the public semantic tree, where the lines 1-19 are to extract the stay-points from the public trajectory sequence Tra_p and computer the semantic type of each stay-point by fusing the POI dataset. In line 20, stay-points are clustered into different clusters according to semantic categories and common granularity. Lines 21-22 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 the public semantic tree.

Algorithm 2. Hierarchical temporal semantic tree generation algorithm

```
Input: Historical trajectory of user $i$ $HT^i$, Public semantic tree $C$
Output: Hierarchical temporal semantic tree of user $i$ $G^i$
Generate the set of stay-points $\mathcal{S}^i$ according to Algorithm 1;
Count the residence time $t\_s_j$ of each stay-point;
Mark the semantic information $C_i$ of each stay-point;
$\kappa \leftarrow$ cluster the stay-points ($\mathcal{S}^i$, $C$);
$G_0^i \leftarrow Root(\kappa)$, $m = 0$;
$G \leftarrow$ abstract semantic categories ($G_0^i$, $C$);
$m = h(G)$;
while $m < h$ do
    for $n$ in range($G_m^i$) do
        if $g_{m,n} != \varnothing$ then
            $G_{m+1}^i = \Pi(g_{m,n}, t\_s_j)$;
```

```
12:                 end if
13:         end for
14: end while
15: Return $G^i$
```

Algorithm 2 shows the generation process of hierarchical temporal semantic tree, where rows 1-3 are similar to the generation process of public semantic tree, indicating that the stay-point and semantic annotation operation are extracted from the historical trajectory of user *i*. The residence time of the user's stay-point is counted by the time to leave the stay area minus the time to enter the stay area. In line 4, the different clusters obtained by the stay-point clustering are used as the middle layer according to the semantic category and the common granularity. In lines 5-7, it represents the abstraction of semantic categories from the middle layer up in combination with the common semantic tree. In lines 8-14, it represents that the clustering operation is performed from the middle layer down according to the residence time of the stay-point. Thereby, the hierarchical temporal semantic tree of user *i* is built.

4.2 Evaluating Location Privacy Protection Demands

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 the 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. Thus, the user may have the stronger demand for privacy protection of the location. The definition of semantic sensitivity is as following.

Definition 5(Semantic sensitivity). Semantic sensitivity refers to the sensitivity of the semantic type of each stay-point. It can be represented as $Sen\_s_i = (w_1, w_2, \cdots, w_k)$, w_k represents the weight of the *k*-th semantic category in the user's trajectory.

Refer to the research work of reference [36], the TF-IDF model is utilized to measure the privacy sensitivity of users for each 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. The TF represents the word frequency, that is, the frequency of a semantic category in the user's personal trajectory data set. The IDF represents the inverse document frequency, that is, the frequency of the semantic category in other user trajectories. There is no trajectory data of other users because the process is performed on the user's mobile client. 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. There is also a positive correlation between the demand for the semantic attributes and the number of semantic categories in the POI datasets. TF and IDF can be calculated by Formulas (3) and (4).

$$TF_i = \frac{n_{iu}}{N_u}, \tag{3}$$

$$IDF_i = \log\frac{|N|}{|n_i|}, \tag{4}$$

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 stay-points in the trajectory data of the user, and $|N|$ represents the total number of POIs.

The semantic sensitivity of the locations in the trajectory sequence of target user can be calculated as:

$$w_i = TF_i \times IDF_i = \frac{n_{iu}}{N_u} \times \log \frac{|N|}{|n_i|}. \tag{5}$$

In order to establish the connection between the semantic privacy sensitivity of the location and the hierarchical temporal semantic tree, the semantic privacy sensitivity Sen_s_i is mapped to the temporal semantic tree for backtracking to the corresponding layers. In order to fully express the user's personal requirements about location privacy, the user is predefined in the customized range [O_p , O_r]($1 \le O_p \le O_r \le h$, $O_p, O_r \in \mathrm{P}^*$), where *h* is the height of the semantic tree. The backtracking layer Ψ can be calculated by the formula (6).

$$\Psi = \left(1 - \frac{w_j}{W}\right) O_p + \frac{w_j}{W} O_r, \tag{6}$$

where, $W = \sum_k w_k$.

The backtracking layer Ψ can be regarded as the user's demand for location privacy protection. The smaller the Ψ value, the lower the user's privacy protection demand for the location, while the larger the Ψ 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

Definition 6(ε -geo-indistinguishability [34]). According to the definition of differential privacy [19], location differential privacy can be defined as follows: The location privacy protection mechanism M satisfies ε -geo-indistinguishability, if and only if:

$$\mathrm{M}(p)(p^*) \le e^{\varepsilon} \mathrm{M}(p')(p^*), \tag{7}$$

where, $p, p' \in \mathrm{P}$, p' is another position different from p. p^* is the disturbance position corresponding to p, and $p^* \in \mathrm{P}^*$.

For mechanism $\mathrm{M} : \mathrm{P} \to O(\mathrm{P}^*)$, where P denotes the set of the original position. $O(\mathrm{P}^*)$ denotes the set of the perturbation probability distribution on the position P. The input of the mechanism M is the original position p. The output is the disturbance position p^*. Π is expressed as the probability matrix. π_{p,p^*} is the probability from position p to position p^*.

Trajectory data is usually generated by combining the location points in chronological order. The trajectory privacy protection is mainly achieved through location privacy protection. In order to realize the personalized trajectory privacy protection, the different degrees of privacy protection are provided for each location point according to the privacy protection demands of each location point in the trajectory sequence. The privacy demands of target user need to be divided into different privacy levels. The corresponding privacy budget is allocated for each level.

Firstly, the privacy protection is divided into four levels according to the height h of the hierarchical temporal semantic tree. Then, the corresponding privacy budget of each location is generated. Table 1 shows the privacy level division and privacy budget allocation for the trajectory privacy protection. 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 is increasing in turn. Finally, 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 according to the privacy demands, among which ε_1 to ε_4 are decreasing in turn.

Table 1. Privacy Level and Privacy Budget.

Privacy level (*le*)	**Privacy demands**	**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.

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

$$QL(\mathrm{M}, \theta, d) = \sum_{p, p^* \in \mathrm{P}} \theta(p) \cdot \pi_{p,p*} \cdot d(p \cdot p^*), \tag{8}$$

where $d(p \cdot p^*)$ denotes the quality measure from position p to perturbation position p^*, and the prior probability $\theta(p)$ is the ratio of the number of occurrences of position p in the whole trajectory sequence Tra_i to all records.

The linear optimization problem can be used to solve the minimum of quality loss. It should meet the following conditions.

$$\textbf{Min}: \sum_{p,p^*\in \mathrm{P}} \theta(p)\cdot \pi_{p,p*}\cdot d\left(p\cdot p^*\right) \tag{9}$$

$$\textbf{s.t.}\ \ \pi_{p,p^*} \le e^{\varepsilon}\pi_{p',p^*} \qquad p,p^{'}\in \mathrm{P}, p^*\in \mathrm{P}^* \tag{10}$$

$$\sum_{p^*\in \mathrm{P}} \pi_{p,p^*} = 1 \qquad p\in \mathrm{P} \tag{11}$$

$$\pi_{p,p^*} \ge 0 \qquad p\in \mathrm{P}, p^*\in \mathrm{P}^* \tag{12}$$

5. ERFORMANCE EVALUATION

5.1 Datasets and Experimental Setup

In this section, Python is utilized to analyze the trajectory sequence and verify the performance of the proposed HTST-PTPP method. The GeoLife datasets of Microsoft Asia Research Institute [37] and the Beijing POI datasets are used as the original trajectory sequence of target user. The GeoLife datasets 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 datasets contain a variety of different location points, from the daily itinerary to the personalized activities of users. The Beijing POI datasets record the location information of most POIs in Beijing, including latitude and longitude coordinates, names, types, etc. As shown in Table 2, the original POI datasets are divided into 20 different types to label the semantic information for the stay-points in the subsequent steps.

Table 2. Service types of Beijing POI dataset [39].

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. After removing the invalid data, the number of actual and useful users is 51 and the number of useful trajectories is approximately reduced to half of the total number of trajectories. 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 to label the semantic category for each stay-point. Fig. 4 and Fig. 5 show the extraction of stay-points and semantic category labeling of one user, respectively.

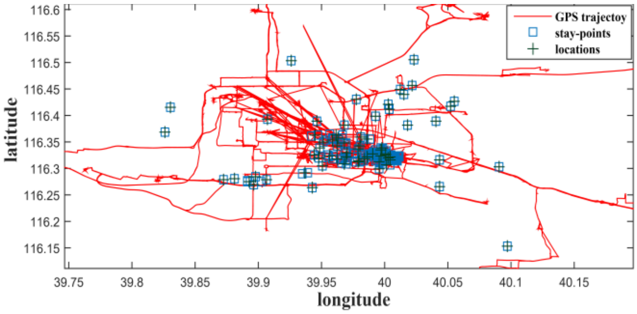

Fig. 4. The example of stay-points extracting.

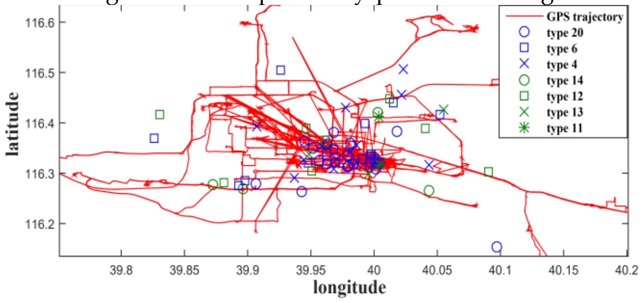

Fig. 5. The example of semantic categories labeling.

To better validate the data utility, the GPS trajectory dataset is divided into the training set and the test set. Of these, the training set accounts for 80%. It can be considered as the average of the individual prior probabilities of all users visiting the area to be used to build the global prior. The post-mapping mechanism uses the obtained average value to obtain the location of the optimal service quality loss. The test set is 20%. It can build user-specific prior information for at least 20 users. It is used to evaluate the proposed HTST-PTPP scheme and measure the quality loss.

5.2 Experimental Results and Performance Analysis

The performance of the proposed HTST-PTPP method is verified from two aspects of data availability and privacy protection. It is compared with the PLDP-TD [40] method and the TPP-POIs [33] 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. However, the privacy preference of its location is determined by the location itself and has nothing to do with the behavior of target user. 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 [41] 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 sequence and the original trajectory sequence, and the lower the data availability.

Fig. 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 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.

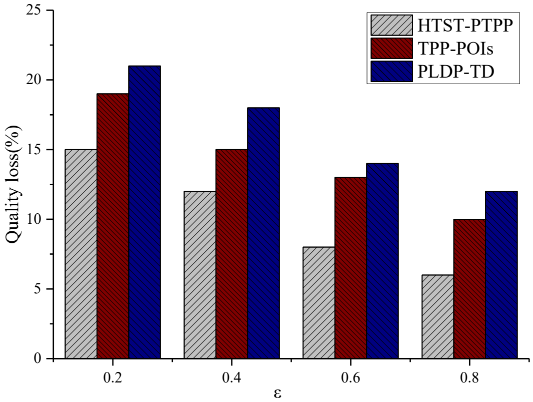

Fig. 6. The impact of privacy budget on quality loss.

(2) Data utility

Suppose that Tra_i and Tra_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:

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

where, τ is negligible.

Fig. 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 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 fails to consider the semantic information of each location, the data utility is the lowest compared to the other two methods.

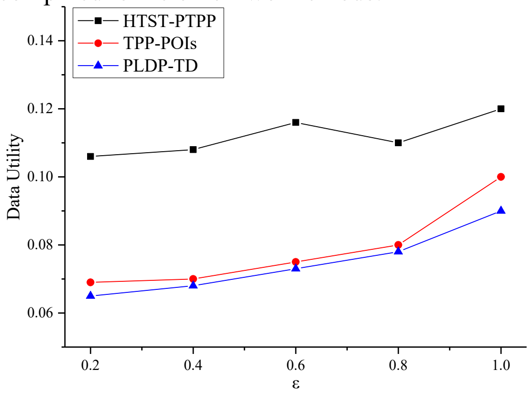

Fig. 7. The impact of privacy budget on data utility.

(3) Privacy protection

In terms of privacy protection, Adversarial Error is used as its evaluation index. The adversarial error is proposed by Reference [42], which measures the expected error of the adversary in the inference attack. It can be calculated as:

$$AdvError(\theta, \pi, H, d) = \sum_{p, p', p^* \in \mathrm{P}} \theta(p) \pi_{p,p^*} \gamma\left(p' \mid p^*\right) d(p, p^*) \tag{14}$$

where, $\gamma\left(p' \mid p^*\right)$ denotes the probability of mapping p^* to p'.

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 predicted position of attackers and the real position, and the stronger the privacy protection.

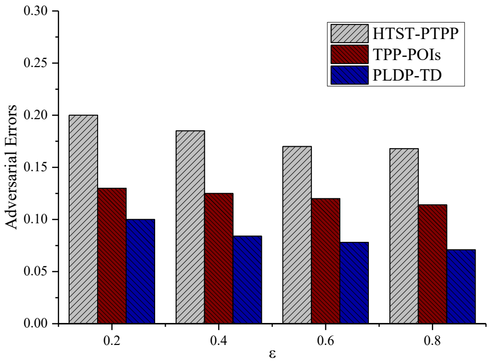

Fig. 8. The impact of privacy budget on adversarial errors.

Fig. 8 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. It can be seen that the adversarial error of the three methods decreases with the increase of privacy budget ε. This is because the privacy protection intensity decreases with the increase of privacy budget ε. The corresponding noise is less, which leads to the decrease of the adversarial error. The HTST-PTPP method has the larger adversarial error compared than the other two methods when the privacy budget ε is same. The main reason is that the HTST-PTPP method not only considers the semantic information of the location, but also considers the residence time of users. It provides users with a more personalized privacy requirement. Although the TPP-POIs method considers the semantic information of the location, it lacks the consideration of the time factor. The adversarial error of TPP-POIs method is lower than the HTST-PTPP method. The PLDP-TD method only considers the privacy preference of the location point. It ignores the influence of semantic factors and time factors. Therefore, the HTST-PTPP method has the better performance of privacy protection than the other two methods.

(4) Execution time

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

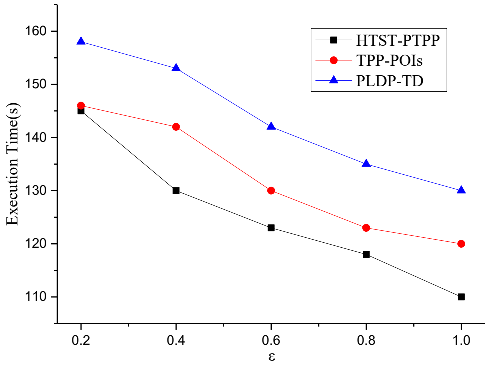

Fig. 9. The impact of privacy budget on execution time.

In general, it can be seen from Fig. 9 that the execution time of the HTST-PTPP method is lower than the other two methods. This is because that the HTST-PTPP method constructs the hierarchical temporal semantic tree 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 considers the semantic information of locations. It can more easily mine the preference characters of users than the PLDP-TD method, which only considers the geographical information of locations. Thus, the TPP-POIs method has the lower execution time than the PLDP-TD method.

6. CONCLUSIONS

This paper mainly studies the problem of personalized trajectory privacy protection, and proposes a Hierarchical Temporal Semantic Tree based Personalized Trajectory Privacy Protection (HTST-PTPP) scheme. 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 trajectory sequence of users. The privacy requirements of each location are quantified. The privacy level is divided, so as to provide the personalized trajectory privacy protection for users by utilizing the differential privacy technology. Finally, the real datasets 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. Also, we will continue this research by considering multiple different datasets (e.g. Gowalla, Foursquare, etc.), so as to well verify the performance of the proposed HTST-PTPP scheme.

ACKNOWLEDGMENTS

This research was funded by the Key R&D Project of Hainan Province (ZDYF2024SHFZ264), and the National Natural Science Foundation of China (61902361).

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. Xiao, H., Xu, C., Feng, Z., Ding, R., Yang, S., Zhong, L., Liang, J., Muntean, G. M. A Transcoding-Enabled 360° VR Video Caching and Delivery Framework for Edge-Enhanced Next-Generation Wireless Networks. IEEE Journal on Selected Areas in Communications, 2022, 40, 1615-1631.
3. Xiao, H., Zhuang, Y., Xu, C., Wang, W., Zhang, H., Ding, R., Cao, T., Zhong, L., Muntean, G. M., Transcoding-Enabled Cloud-Edge-Terminal Collaborative Video Caching in Heterogeneous IoT Networks: An Online Learning Approach with Time-Varying Information. IEEE Internet of Things Journal, 2024, 11, 296-310.
4. 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.

5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. Xiao, H., Xu, C., Ma, Y., Yang, S., Zhong, L., Muntean, G. M. Edge Intelligence: A Computational Task Offloading Scheme for Dependent IoT Application. IEEE Transactions on Wireless Communications, 2022, 21, 7222-7237.
11. Xiao, H., Huang, Z., Xu, Z., Yang, S., Wang, W., Zhong, L. Task-Driven Cooperative Internet of Robotic Things Crowdsourcing: From the Perspective of Hierarchical Game Theoretic. IEEE Internet of Things Journal, 2024, 11, 32350–32362.
12. 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.
13. 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).
14. 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.
15. 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.
16. 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.
17. 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.
18. Wu, X., Sun, G. A Novel Dummy-Based Mechanism to Protect Privacy on Trajectories. IEEE International Conference on Data Mining Workshops, 2015, 1120-1125.
19. 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.
20. 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.
21. 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.
22. 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.

23. Hu, P., Chu, X., Zuo, K., Ni, T., Xie, D., Shen, Z. Security-Enhanced Data Sharing Scheme with Location Privacy Preservation for Internet of Vehicles. IEEE Transactions on Vehicular Technology, 2024, 73, 13751-13764.
24. Dwork, C. Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 2008, 1–19.
25. Wang, H., Xu, Z. CTS-DP: Publishing correlated time-series data via differential privacy. Knowledge-Based Systems, 2017, 122, 167-179.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. Min, M., Zhu, H., Li, S., Zhang, H., Xiao, L., Pan, M. Semantic Adaptive Geo-Indistinguishability for Location Privacy Protection in Mobile Networks. IEEE Transactions on Vehicular Technology, 2024, 73, 9193-9198.
33. 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.
34. 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.
35. Qiu, G., Guo, D., Shen, Y., Tang G., Chen, S. Mobile Semantic-Aware Trajectory for Personalized Location Privacy Preservation. IEEE Internet of Things Journal, 2021, 8, 16165-16180.
36. He, Q., Chang, K., Lim, E. P., et al. Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32, 1795.
37. 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.
38. 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.
39. Zhu, L., Liu, X., Jing, Z., Yu, L., Cai, Z., Zhang, J. Knowledge-Driven Location Privacy Preserving Scheme for Location-Based Social Networks. Electronics, 2023, 12, 1-16.
40. Deldar, F., Abadi, M. PLDP-TD: Personalized-location differentially private data analysis on trajectory databases. Pervasive and Mobile Computing, 2018, 49, 1-22.
41. 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.
42. 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.