Image Processing, Electronics and Computers
L. Trajkovic et al. (Eds.)

doi:10.3233/ATDE240514

Personalized Privacy Protection Mechanism Integrating Spatiotemporal Correlation

Yanmei SHEN[a], Rui HUA[b,1], Hui WANG[a], Zihao SHEN[b] and Peiqian LIU[a]
[a] *School of Software, Henan Polytechnic University, Jiaozuo, China*
[b] *School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China*

Abstract. In view of the insufficient data availability in traditional trajectory privacy protection schemes, in order to achieve a balance between privacy security and data usage efficiency. This paper achieves personalization of privacy protection and optimization of data availability by comprehensively considering sensitive locations and their correlations in user trajectories. It can realize personalized protection of sensitive locations with different privacy levels based on the user's preset privacy needs and sensitivity assessment. At the same time, by introducing trajectory candidate sets and cross-correlation functions, a new trajectory publishing mechanism is constructed, which can maintain the spatiotemporal characteristics of trajectory data while ensuring privacy. Through simulation experiments, the effectiveness of the proposed scheme was verified on real data sets. Experimental results show that the method proposed in this article has good results in terms of privacy protection strength and data availability.

Keywords. Trajectory privacy, privacy budget allocation, personalized differential privacy, data publication

1. Introduction

In the information age, with the close integration of mobile networks and positioning technology, location-based services[1] (LBS) have penetrated into people's daily lives through the widespread popularity of mobile devices, such as searching for nearby people in social networks and using navigation functions to reach their destinations[2]. For example, recommendation systems based on points of interest such as Xiaohongshu and Dianping can recommend surrounding hotels, restaurants, cinemas, parking lots, gas stations, etc. to users. Map applications provide navigation paths based on the user's location. Baidu Maps and Amap not only assist users in traveling to unknown locations, they can also update road conditions in real time and guide users to choose the best route[2,3], playing an important role in predicting natural disasters and post-disaster rescue, while accurately locating affected users to implement rescue[4].

At present, a common trajectory privacy protection method is to perform consistent anonymization of trajectories[5]. Taking the specific geographical topology Figure 1 as an example, the sensitive location S (marked with an asterisk) is

[1] Corresponding Author: Rui HUA, 1710264257@qq.com.

generalized into a larger gray area to increase the intensity of privacy protection. When a user moves from one non-sensitive location to another, if the movement passes through a sensitive area, even if the user's specific movements in the sensitive area are not directly reported, the attacker may still arrive at the gray area from point A by analyzing the user[6]. In addition, by analyzing user behavior patterns (for example, a user who moves along a specific path after nine o'clock in the evening is likely to go to a movie theater), combined with additional information such as geographical location restrictions, attackers may be able to more accurately infer the user's specific activities[7]. This may further cause privacy leaks.

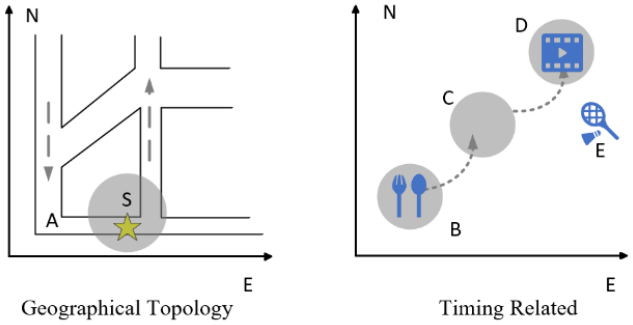


Figure 1. The impact of geographical topology and time series correlation on privacy protection.

The differential privacy model is widely welcomed by domestic and foreign scholars because of its strict mathematical definition and highly customized privacy protection level[8]. In this way, whether any piece of data is added or deleted from the data set, it will only have a minimal impact on the query results, which makes it difficult for an attacker to accurately infer any specific piece of data in the data set even through multiple query results[9]. real data. Allowing data analysts to gain valuable insights into the overall characteristics of the data set without leaking individual data.

Gursoy et al.[10] proposed the DP-Star system framework. Initially, DP-Star utilizes an advanced normalization algorithm, which perceives spatial density to add differential privacy noise conforming to spatial density requirements. Sun et al.[11] proposed a scheme for synthetic private trajectories and real trajectories (SPRT) to address the issue with existing methods focusing on synthesizing trajectories that preserve summary-level statistical data, which results in the loss of individual-level movement patterns. The key idea is to integrate the public geographical structure of the target area into the synthetic process of private trajectories. Ghane et al.[12] argue that simply representing trajectories as a sequence of location points often leads to highly inaccurate query answers, as it ignores the order dependency among locations in the trajectory, violating the consistency of trajectory data. Moreover, as trajectories are often unevenly distributed across urban areas, uniformly adding noise typically results in poor data usability. Subsequently, not only for a specific set of queries, but these regions and their densities are also used to predict the distribution of trajectories in the query space, thus ensuring high accuracy of the query set.

Although some trajectory privacy protection strategies protect sensitive points on user paths and directly expose non-sensitive locations, these strategies do not adequately consider the role of road networks in privacy protection. Facing these challenges, an effective solution is to develop more refined privacy protection

mechanisms. Additionally, new noise addition strategies should be explored, which can generate noise sequences that are more consistent with the logic and temporal sequence of the original trajectories, thereby enhancing the effectiveness of privacy protection and the utility of the data.

- The study identifies user stay points and preset locations through trajectory analysis. Using the PCPL algorithm, it evaluates and filters location data to identify sensitive locations necessitating enhanced protection against information leakage. Additionally, it assesses connected locations to gauge privacy risks in surrounding non-sensitive areas, designating high-risk ones as sub-sensitive locations. This approach expands protection boundaries by considering location interconnections and potential privacy threats.
- Perform appropriate perturbation processing on sensitive locations in the original trajectory. By integrating the CTS algorithm with spatiotemporal correlation, the temporal cross-correlation of the published trajectory sequence is limited, ensuring that the published trajectory is logically and temporally consistent and indistinguishable from the original trajectory and noise sequence, thereby improving privacy protection and ensuring data security. The spatiotemporal correlation constraints are also preserved.

2. Related Knowledge

Definition 1 Trajectory Undirected Graph: The trajectory area undirected graph consists of nodes, the edges between nodes, and the weights of the edges. Each node represents a specific area on the map and is identified by an area number. If any two areas on the map are directly adjacent, an edge is formed in the graph between the nodes corresponding to these two areas: the weight of the edge represents the Manhattan distance between the centroids of the two areas.
Here's the translation of the text:

Definition 2 ε -Differential Privacy: A random algorithm A satisfies ε -differential privacy if for any two datasets D_1 and D_2 that differ in only one element, and for all possible output sets S of A, it holds that:

$$Pr[A(D_1) \in S] \leq e^{\varepsilon} \times Pr[A(D_2) \in S] \tag{1}$$

ε is a non-negative real number known as the privacy budget. The smaller its value, the stronger the ability to protect privacy. $Pr[A(D) \in S]$ represents the probability that the output of algorithm A falls into set S when the input dataset is D.

Definition 3 Laplace Mechanism: A function $f : D \to R^k$, where D is the domain of all datasets, and R^k is a K-dimensional real space. The global sensitivity Δf of the function f is defined as:

$$\Delta f = \max_{D_1, D_2} \left\| f(D_1) - f(D_2) \right\| \tag{2}$$

for any two datasets D_1 and D_2 that differ by at most one element, given a privacy budget $\varepsilon > 0$, the Laplace mechanism provides ε -differential privacy protection by adding Laplace noise related to Δf and ε to the output of function f. For any dataset D, the output of the Laplace mechanism is:

$$M(D, f, \varepsilon) = f(D) + (Y_1, Y_2, ..., Y_k) \tag{3}$$

3. System Architecture of Personalized Privacy Protection Mechanism Integrating Spatiotemporal Correlation

This paper proposes a personalized privacy protection mechanism that integrates spatiotemporal correlation. The specific process is as follows in Figure 2:

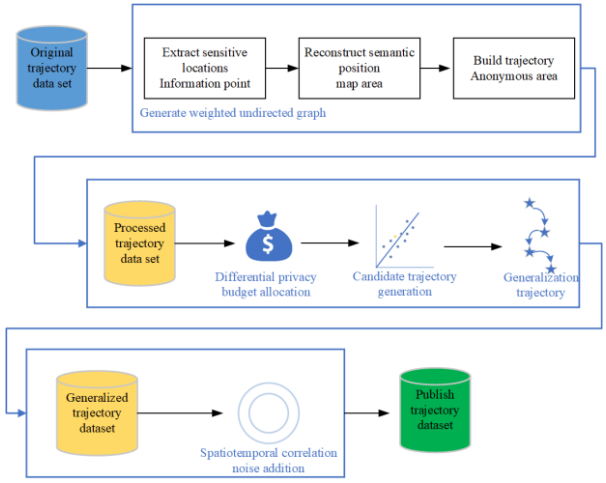


Figure 2. Personalized trajectory privacy protection mechanism process integrating spatiotemporal correlation.

3.1. Construction of Sensitive Location Point Collection

In the field of trajectory privacy protection, sensitive location points refer to geographical location information that individuals do not want or should not be easily obtained or inferred[13]. These locations may require special protection due to personal privacy, security or other sensitive factors, such as personal addresses, workplaces, medical institutions, religious sites, sensitive government agencies, etc. The purpose of protecting these sensitive locations is to prevent personal privacy leaks and maintain personal security and privacy rights. Therefore, identifying and protecting sensitive locations is an important aspect of trajectory privacy protection, aiming to reduce potential threats to personal privacy through technical means. Users can manually set these sensitive locations according to their privacy protection needs, or determine them through system presets. Can be expressed as $SL = \{SL_1, SL_2 ..., SL_n\}$.

Non-sensitive location points refer to those location points that do not involve the user's personal privacy information, and the user's information at these locations is not

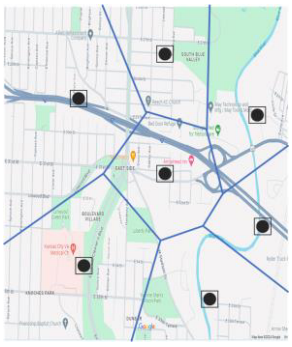
considered to require special protection. When a user is in such a location, their location information is considered safe to disclose. These locations usually include public spaces and places frequently visited by the public, such as parks, squares, etc. These locations constitute a collection of non-sensitive locations. In current privacy protection practices, non-sensitive location points are usually not included in the scope of privacy protection. Their information can be released directly without special privacy treatment. It can be expressed as: $NSL = \{NSL_1, NSL_2 ..., NSL_n\}$.

Logically unreachable location points mainly refer to location points that are logically unavailable or deviate greatly from the location where users often stay, such as lakes, deserts, etc. The set composed of these location points is called a logically unreachable point location set. Expressed as NA .

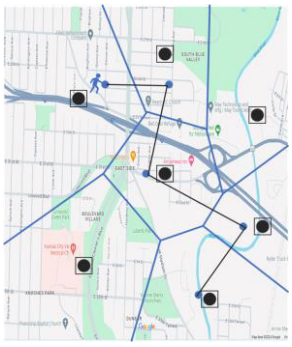
3.2. Weighted Undirected Graph Based on Semantic Position

However, anonymizing only a pre-defined set of sensitive locations without considering the geographical topology may make other location points in the trajectory also a risk of privacy leakage. Therefore, the overall connectivity of location points needs to be considered, and a certain sensitivity is also assigned to semantic locations close to sensitive locations. In real life, the time required to travel to and from the same two locations at different times will be very different, such as walking, cycling, driving, etc. If only distance is used as a measurement criterion, even if users adopt different travel modes, the sensitivity of their allocation is fixed, which may have an impact on privacy protection and data availability. This study comprehensively considers the overall connectivity between geographical points and combines the user's travel mode to perform service similarity queries on the map, merge location sets with the same similarity, the calculation rules are as follows:

$$d(p,q) = |x_1 - x_2| + |y_1 - y_2| \tag{4}$$



(a) User-defined semantic map



(b) User-defined semantic map trajectory information

Figure 3. Semantic map of user-defined sensitive locations.

Generally speaking, the area where the user is located can be defined by irregular polygons composed of locations. After completing the segmentation of each area, a semantic map is formed, as shown in Figure 3(a). According to the user's preset

sensitive location and the privacy level corresponding to each location, with the sensitive location as the center, it is converted into a location-weighted undirected graph, which can be expressed as: represents the location point, is the edge of the location point, Represents the edge weight. Because the basic property of an undirected graph is that the distance from each point in the area to the generating point is smaller than the distance to other generating points, it can be used to calculate the privacy level of all locations in the area. Figure 3(b) shows the user's trajectory information in the semantic map of customized sensitive locations.

3.3. Personalized Privacy Level Classification PCPL Algorithm

Algorithm 1 PCPL algorithm

Input: Map semantic location division $M=\{SL,NSL,NA\}$, Positional weighted undirected graph $PG=\{P,E,W\}$, Default set of sensitive locations and corresponding set of privacy levels S, Preset travel mode corresponding speed V, Privacy level threshold θ.

Output: The privacy level corresponding to each location point pl

```
Select the head element in SL:α
While Seneitive location α ≠ Null do
    Get the neighborSet for sensitive location α
    for each g∈ neighborSet do
        newpl = g.pl
        if newpl < θ then continue
            if g ∈ SL then
                g.pl = max(g.pl, newpl)
            else g.pl = newpl
        end if
        SL.add(g)
    end for
    Select the next element in the SL:α.next
end while
return pl
```

The set of location points with privacy level near the sensitive location point α is defined as *neighborSet*, its size is equal to the degree of the location point α in the location weighted undirected graph, the corresponding privacy level is *pl*, where the value range is $[0,1]$, the larger the value, the location The higher the privacy level. The specific execution process of the algorithm is shown in Algorithm 1.

Calculating the privacy level of other locations based on the user's sensitive location requires taking the user's travel speed and time into consideration. *v* is the preset average speed of different travel modes for users. At t_i, where $i\in[1,w]$, the time weight value of the sensitive location preset by the user is expressed by $weight_{t_i(g)}=\{weight_{t_1},weight_{t_2},...,weight_{t_n}\}$, where $weight_{t_i(g)}\in[0,1]$, the privacy

sensitivity assigned to the location point can g be calculated by Equation 5. In the PCPL algorithm, the threshold parameter θ plays a certain regulatory role.

$$g.pl = weight_{t_i(g)} \times \frac{1 \Big/ g.\frac{dis}{v}}{\sum_{g' \in neighborSet} 1/g'.dis} \times (a.pl) \tag{5}$$

3.4. Improved Differential Privacy Protection Model

In the traditional ε -differential privacy model, strategies to achieve privacy protection usually add noise to the data set so that query responses obtained from any two similar data sets are statistically indistinguishable. However, this approach typically applies a uniform noise intensity to the data traces, a generalization that is not optimal in all situations, especially when processing data that contains geolocation information. The need for differential privacy processing of spatial data emphasizes that adding an equal amount of noise to each geographical location may cause undesirable effects, because different locations may have different needs for privacy protection. In order to solve this problem, this chapter proposes an optimized (r, ε) -Differential privacy protection model, which allows the intensity of injected noise to be dynamically adjusted based on the privacy model parameter r set by the user or system, as well as the sensitivity of each location point, thereby improving privacy protection efficiency while ensuring better data availability.

When a location is to be published, it must satisfy the location's privacy level weight pl , which is inversely proportional to the differential privacy parameter ε assigned to the location:

$$pl \times \varepsilon = \gamma \tag{6}$$

Within a certain period of time, the higher the weight of the privacy level, the smaller the privacy budget allocated to the location, which indicates a stronger degree of privacy protection at the location. Conversely, positions with lower weights will receive larger privacy budgets. Based on the privacy level weight of each location, the privacy budget of each location on the user's trajectory can be calculated. r is a parameter of the differential privacy protection model, used to control the intensity of privacy protection.

In the field of location privacy protection, considering the possible temporal correlation between points on the trajectory, simply adding independent Laplace noise to each point may not be enough to prevent attackers from using background knowledge to remove noise through filtering, thereby Identify the user's real trajectory. This is because if the noise is independent of each other and the particularity of the trajectory information is not taken into account, their overall effect may be weakened by some statistical methods only by the addition of a single noise, resulting in a reduction in the intensity of privacy protection.

Algorithm 2 CTS algorithm

Input: Map semantic location division $M=\{SL, NSL, NA\}$, real location l, the privacy level corresponding to each location point pl

Output: candidate trajectory set T_{set}

Calculate the pl of each location
$T_{set} \leftarrow$ select $x-1$ location of the $pl>0$ closest to l
$l_d \leftarrow$ **Random** select one location from T_{set}
$T_{Lset} \leftarrow$ **Random** select $2k$ locations of the $pl>0$
intialise $T_{set}=\{d_1, d_2, ..., d_k\}$
While $n<m$ **do**
  $T'_{Lset} \leftarrow$ **Random** select $k-1$ locations from T_{Lset}
  $T'_{set} \leftarrow T'_{Lset}+l_d$
  if $\max_{D'_{set}}\left\{\sum_{i\neq j}^{k} Dis\left(d_i, d_j\right)\right\} > \max_{D_{set}}\left\{\sum_{i\neq j}^{k} Dis\left(d_i, d_j\right)\right\}$
    $T_{set} = T'_{set}$
  end if
  $n++$
end while
return T_{set}

The CTS algorithm selects K locations in the virtual location candidate set. The distance between pairs of these locations, including the sum of offset locations, is the largest, which can improve the virtual location. degree of dispersion. Using offset positions to generate anonymous sets can reduce the possibility of an attacker inferring the actual location and improve the privacy protection effect of the algorithm. As shown in Equation 7:

$$T_{set} = \arg\max\left\{\sum_{i\neq j}^{k} Dis\left(d_i, d_j\right)\right\} \tag{7}$$

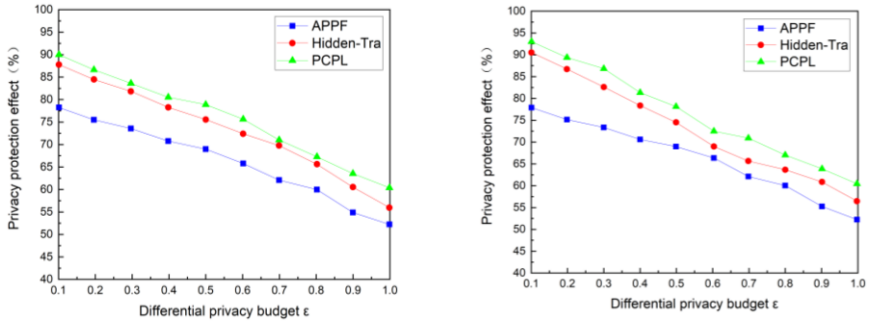
where $Dis\left(d_i, d_j\right)$ represents the distance between any two virtual positions. The greater the sum of the distances of the selected K virtual positions, the more dispersed the distances between the virtual positions are.

4. Experimental Results and Analysis

This article uses two real data sets, Geolife and TaxiService, collected by MSRA to map the accuracy and dimensions of each GPS data to state coordinates. Since the collected trajectory data set does not contain specific personalized location privacy information, some data are randomly extracted from the data set and preset as user-specified sensitive location points.

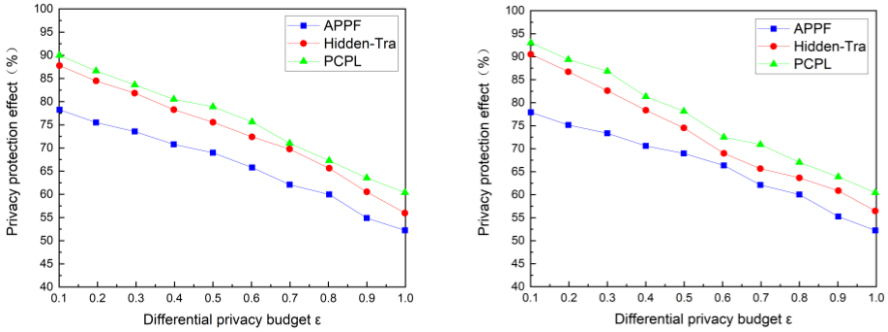
In order to compare and verify the effectiveness of the method proposed in this chapter, comparative experiments were conducted by comparing the PCPL method with the APPF method[14]and the Hidden-Tra method[15]. According to the previous analysis, under the same privacy budget, data availability will decrease as the sensitive radius increases, because as the sensitive radius increases, the privacy requirements of each location point in the trajectory within the sensitive radius circle are basically the same.

As can be seen from Figure 4, as the privacy budget parameters increase, the privacy protection effects of all three methods show a downward trend. This is expected since higher ε values mean less noise is added to the data, thus reducing the strength of privacy protection. Among the three methods examined, the APPF method shows the lowest privacy protection effect, especially when the ε value increases, its privacy protection effect decreases the fastest. This is because APPF does not involve personalized privacy protection requirements.



(a) Geolife (b) TaxiService

Figure 4. Privacy protection effect evaluation.



(a) Geolife (b) TaxiService

Figure 5. Data availability assessment.

In contrast, when the Hidden-Tra method has a low ε value, the privacy protection effect of the Hidden-Tra method is better than APPF, but as ε increases, its privacy protection effect also declines rapidly. However, the Hidden-Tra method fails to take into account the time series correlation between the user's original trajectory and the added noise sequence, which may allow an attacker to easily identify the user's actual trajectory by filtering the noise. By allocating different privacy budgets to sensitive locations and their directly connected location points according to user needs, in

general, the PCPL algorithm has better performance in protecting user privacy, especially when strong privacy protection is provided.

The data availability under different differential privacy budget parameter values can be seen in Figure 5. As can be seen from the Figure 5, the data availability of all three methods increases with increasing values. The specific analysis reasons are as follows: Although the APPF method considers differentiated noise addition to different position points on the trajectory, it does not fully consider the distance between the position points and the time threshold for position sensitivity when calculating the privacy levels of different position points. The impact and the difference in the average speed of users in different travel modes lead to an insufficiently detailed classification of privacy levels. This rougher classification of privacy levels may lead to insufficient privacy protection or low data availability in some cases.

5. Summary and Outlook

In the practice of trajectory privacy protection, only protecting location points identified as sensitive and directly disclosing non-sensitive locations may not be able to effectively prevent attackers from destroying the privacy protection effect by analyzing the correlation between locations. The method proposed in this chapter provides personalized privacy protection by evaluating the sensitivity differences between different location points, allocating appropriate privacy budgets to each location and adding differentiated noise, the original trajectory sequence and the noise sequence, the basic requirements for publishing trajectories that meet the mutual correlation constraints are defined, thereby enhancing the effect of privacy protection and ensure data availability. Through this meticulous privacy protection strategy, users can be provided with more secure and reliable personalized privacy protection in complex actual road network environments. In the future, we will optimize the running time of the algorithm and focus on solving the problem of privacy leakage caused by semantic trajectories in trajectory privacy protection.

References

[1] Özdal Oktay S, Heitmann S, Kray C. Linking location privacy, digital sovereignty and location-based services: a meta review. Journal of Location Based Services, 2024, 18(1): 1-52.
[2] Jain P, Raskhodnikova S, Sivakumar S, et al. The price of differential privacy under continual observation//International Conference on Machine Learning. PMLR, 2023: 14654-14678.
[3] Chen X, Xu J, Zhou R, et al. Trajvae: A Variational Autoencoder Model for Trajectory Generation. Neurocomputing, 2021, 428: 332-339.
[4] Cunningham T, Cormode G, Ferhatosmanoglu H, et al. Real-World Trajectory Sharing with Local Differential Privacy. Proceedings of the VLDB Endowment, 2021, 14(11): 2283-2295.
[5] Rahman A, Hasan K, Kundu D, et al. On the ICN-IoT with federated learning integration of communication: Concepts, security-privacy issues, applications, and future perspectives. Future Generation Computer Systems, 2023, 138: 61-88.
[6] Claridades A R C, Kim M, Lee J. Developing a model to express spatial relationships on omnidirectional images for indoor space representation to provide location-based services. ISPRS International Journal of Geo-Information, 2023, 12(3): 101.
[7] Hopkins S B, Kamath G, Majid M, et al. Robustness implies privacy in statistical estimation//Proceedings of the 55th Annual ACM Symposium on Theory of Computing. 2023: 497-506.
[8] Orabi M, Al Aghbari Z, Kamel I. FogLBS: Utilizing fog computing for providing mobile Location-Based Services to mobile customers. Pervasive and Mobile Computing, 2023, 94: 101832.

[9] Zanella-Béguelin S, Wutschitz L, Tople S, et al. Bayesian estimation of differential privacy//International Conference on Machine Learning. PMLR, 2023: 40624-40636.
[10] Gursoy M E, Liu L, Truex S, et al. Differentially Private and Utility Preserving Publication of Trajectory Data. IEEE Transactions on Mobile Computing, 2018, 18(10): 2315-2329.
[11] Sun X, Ye Q, Hu H, et al. Synthesizing Realistic Trajectory Data with Differential Privacy. IEEE Transactions on Intelligent Transportation Systems, 2023.
[12] Ghane S, Kulik L, Ramamoharao K. A Differentially Private Algorithm for Range Queries on Trajectories. Knowledge and Information Systems, 2021, 63(2): 277-303.
[13] Huang H, Cheng Y, Dong W, et al. Context modeling and processing in Location Based Services: research challenges and opportunities. Journal of Location Based Services, 2024: 1-27.
[14] Ye A, Zhang Q, Diao Y, et al. A Semantic-Based Approach for Privacy-Preserving in Trajectory Publishing. IEEE Access, 2020, 8: 184965-184975.
[15] Jia J, Qin H. Dynamic trajectory anonymity algorithm based on genetic algorithm. Computer Engineering and Science, 2021, 43(1): 142-150.