SCTP: Achieving Semantic Correlation Trajectory Privacy-Preserving With Differential Privacy

Haojie Yuan, Lei Wu*, Lijuan Xu, Libo Ban, Hao Wang, *Member, IEEE*, Ye Su, and Weizhi Meng, *Senior Member, IEEE*

***Abstract*—With the rapid proliferation of vehicular technology, location-based services (LBS) have become a crucial component of Internet of Vehicles (IoV) applications. These applications, such as map navigation and health tracking, rely on users' location information to provide services, enabling users to effectively share their locations, access information about nearby activities, and engage in real-time communication. However, the extensive collection and sharing of location data pose serious challenges to the semantic privacy preservation of user locations. To address these challenges in IoV, we propose a Semantic Correlation Trajectory Privacy-preservation mechanism (SCTP). The SCTP combines Hidden Markov Models(HMM) with differential privacy, aiming to protect the semantic privacy of user trajectory locations while maintaining high-quality location services and data usability. Our scheme introduces a trajectory prediction algorithm based on HMM, which dynamically and accurately predicts user trajectories and generates highly available semantically correlated trajectory datasets. Additionally, we design a personalized privacy budget allocation strategy based on semantic frequency. By assigning privacy weights, we significantly improve the usability of trajectory data while protecting data privacy. Theoretical analysis and experimental validation demonstrate that SCTP rigorously adheres to ε-differential privacy standards while exhibiting significant advantages in safeguarding the semantic privacy of user locations.**

***Index Terms*—Differential privacy, semantic correlation preserving, privacy budget, trajectory publishing.**

I. Introduction

With the rapid proliferation of mobile applications and the rise of the IoV, location-based services have become an integral part of modern society. These services offer conveniences such as map navigation and health tracking, significantly enhancing the collection of user behavior data within IoV. Some geographic information providers and third-party research institutions frequently utilize these trajectory data in fields like transportation planning and urban management to predict and address practical issues.

However, user trajectory data reflects more than just users' geographic locations; it also includes multi-layered information such as temporal dimensions and semantic behaviors. In other words, the extensive collection and sharing of location data increasingly threaten the semantic privacy of individuals' locations. Given the high sensitivity of these data, malicious use can severely jeopardize personal information, such as social relationships and points of interest. Although third-party institutions are often regarded as trustworthy partners, they may pose potential security threats. This privacy risk mainly stems from the intrinsic correlations in user behavior, whereby even indirect data analysis may reveal intricate details of individuals' daily lives.

In recent years, trajectory data processing technology has made significant progress in handling data with rich spatiotemporal features, and a wide range of techniques for protecting location privacy have been proposed. These methods include but are not limited to, anonymization [1], data generalization [2], [3], data obfuscation [4], and data perturbation [5]. In particular, Zhao et al. [6] proposed a privacy-preserving scheme for trajectory data based on a prefix tree structure, ingeniously integrating differential privacy technology to support applications such as location-based services. Inspired by this, an effective strategy is to employ differential privacy (DP) to process collected trajectory statistics, thereby releasing a synthetic trajectory dataset [7]–[9]. The synthetic dataset generated by this strategy preserves the statistical features and distribution of the original data, undermines attackers' ability to infer user information through semantic correlation, and remains applicable for transportation planning and urban management while ensuring privacy-preserving.

Although these techniques can protect user privacy by generating seemingly reasonable user location trajectories [10], they mainly focus on processing geographic coordinates and neglect the deeper semantic factors in users' social mobility. This approach may lead to synthetic location trajectories displaying uniform or highly similar anonymized characteristics, inadvertently exposing users' actual behavioral patterns and semantic information [11]. The moving semantics of locations encompass abstract semantic information derived from user movements [12], revealing the associative behaviors between users and specific locations (for example, specific activities of users at a certain place). Moreover, traditional privacy-preserving techniques often struggle to balance privacy-preserving with the retention of data usability. Existing literature [13]–[15] indicates that research on the semantic associations between two trajectories remains sparse. This scarcity is largely due to the complex nature of assessing correlations between trajectories and the significant challenge of tuning the privacy budget to maximize data usability within

Haojie Yuan, Lei Wu, Libo Ban, Hao Wang and Ye Su are with School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (E-mail: yuanhaojie7218@163.com; wulei@sdnu.edu.cn; banlibo2022@163.com; wanghao@sdnu.edu.cn; suye@sdnu.edu.cn).

Lijuan Xu is with Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences) (E-mail: xulj@sdas.org).

Weizhi Meng is with Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Denmark (E-mail: weme@dtu.dk).

* Corresponding author.

the framework of differential privacy techniques.

Protecting privacy through single-trajectory correlation is risky in vehicular technologies and Telematics because an attacker can obtain the target user's trajectory semantics by analyzing the semantic behaviors of the relevant vehicle users. To balance in-vehicle location privacy and quality of service and to address the above challenges, our paper proposes a semantic correlation trajectory privacy-preserving mechanism (SCTP) that combines semantic correlation quantification with personalized privacy budget allocation. The mechanism protects vehicle users' location semantic privacy while providing high-quality location services enhances users' trust in Telematics applications, and promotes the healthy development of digital society.

The main contributions of this article are as follows:

- We propose a Semantic Correlation Trajectory Privacy-preserving mechanism(SCTP). This mechanism aims to protect the privacy of semantic correlation between user trajectories while ensuring the secure distribution of location data and high-quality query responses. With this mechanism, user trajectory data can be privacy-protected while preserving semantic correlation, thus achieving a balance between security and usability in data publishing and querying within IoV applications.
- The prediction algorithm within the SCTP mechanism, based on HMMs, dynamically predicts user trajectories, thereby constructing a dataset of related trajectories. Compared to existing prediction methods, this algorithm effectively reveals and captures the semantic state transitions within trajectories, enhancing the accuracy and robustness of trajectory prediction. Moreover, the model is versatile and adaptable, capable of handling various types of trajectory data processing tasks, including those involving incomplete data and significant noise interference, by inferring internal state probabilities to improve predictive performance.
- We propose a personalized privacy budget allocation strategy, designing a privacy level allocation algorithm based on semantic frequency and introducing the concept of privacy weights, aimed at achieving a personalized and rational distribution of the privacy budget. This strategy effectively balances the trade-off between noise injection errors and prediction accuracy, thus enhancing the usability of trajectory data.
- We demonstrate that the SCTP mechanism strictly satisfies ε-differential privacy and analyze its security and usability. Furthermore, we compare the SCTP mechanism with other relevant mechanisms. Analysis of experimental results highlights the superiority of our mechanism.

The remainder of this paper is structured as follows: Section II explores the related research covered in this paper. Section III provides the necessary background knowledge of the techniques utilized in the SCTP scheme. In Section IV, we outline the problem formulation of the scheme, including the system model and mechanism architecture. Section V offers a detailed description of the construction of each module. Sections VI and VII present the security proof and performance evaluation, respectively. Finally, Section VIII concludes this paper.

II. Related Work

Trajectory location data occupy a central role in a variety of socially and publicly beneficial sectors, such as smart city development and traffic management [16], with much of this data derived from mobile devices tracking individual location behaviors. However, due to the highly sensitive nature of trajectory data, there is an undeniable risk of privacy breaches during data-sharing processes [17].

Recent studies have demonstrated that even when trajectory data is anonymized, privacy-invasive techniques such as trajectory reconstruction, deanonymization attacks, and membership inference still pose a threat, enabling attackers to reconstruct actual trajectory information from processed data [18]. Li et al. [19] empirically analyzes to quantify the extent of location privacy leakage in MSNs, revealing that even the smallest amount of shared location information exposes users' points of interest (POIs). Xiao et al. [20] considers temporal correlation based on HMM and Bayesian theory. These literatures rarely consider the impact of location semantic correlation on trajectory privacy preservation. More and more researchers are beginning to integrate semantic behaviors into location privacy preservation.

To address this set of privacy concerns, the method of trajectory synthesis using differential privacy is seen as a rather forward-looking solution. Much of the previous work on differential privacy in the area of trajectory data analysis has focused on the design of specialized algorithms for specific application scenarios, such as community mining [21], participatory perception [22], and crowdsourcing recommender systems [23]. In contrast, this paper aims to explore more general and flexible strategies, focusing on how to construct and publish a synthetic trajectory dataset that retains similar properties to the original trajectory dataset while complying with differential privacy requirements.

Considering the preservation of semantic information of published trajectories. Chen et al. [24] mapped trajectories to a prefix tree structure, where each node represents a location, and released the prefix tree with noise; Yin et al. [25] constructed prefix trees using semantic nodes, with each node representing a specific location category; Han et al. [26] efficiently merged spatial regions using Hilbert curves; He et al. [27] proposed the DPT scheme that discretizes trajectories through a hierarchical referencing system and builds a prefix tree structure, generating synthetic trajectories that meet differential privacy requirements through random walks in Markov chains. The recent AdaTrace scheme [19] trains a first-order Markov chain model and other key features based on the results of trajectory discretization, generating data through random walks. Despite improved synthesis efficiency, it is limited by the insufficient information carried by the first-order Markov chain model, making it difficult to generate high-quality trajectory data. Furthermore, Wang et al. [28] developed the PrivTrace algorithm, which adaptively combines first-order and second-order Markov models to predict location trajectories, balancing the usability of trajectory data with privacy protection. In addition, Ghane et al. [29] modeled trajectories using a graphical generative model to capture

the statistical characteristics of moving entities and generate synthetic trajectories that meet the criteria.

Meanwhile, recognizing the limitations of existing schemes in maintaining semantic information of trajectories, Zheng et al. [17] transformed trajectories into prefix tree structures, combining the semantic context of locations with their frequency of occurrence to assess the sensitivity of trajectory positions. Furthermore, Du et al. [30] designed a Hierarchical Graphical Model (HGM), capturing the semantic features of trajectories to generate sets of trajectories that meet differential privacy standards. For the maintenance of temporal correlations within trajectories, Wang et al. [31] explored the secure release of trajectories in a crowdsourcing environment, proposing the RescueDP mechanism, which accurately predicts the location for each time segment based on the temporal correlations between locations and dynamically adjusts the privacy budget to maintain temporal consistency of trajectories. Furthermore, Ou et al. [13] developed a trajectory release mechanism based on differential privacy, which uses a hidden Markov model to build candidate sets and measures the similarity to ensure the protection of correlations among multiple users' trajectories.

However, although current research has incorporated user location and semantic information in the construction of transition models and generation of predicted trajectories, it has yet to adequately integrate location transitions and semantic similarities to optimize trajectory synthesis. In particular, there is still a need for more extensive research on the privacy preservation of semantic associations between multiple user trajectories. Given the potentially severe consequences of deep privacy breaches, such as those involving social relationships, enhancing research on such privacy protections is both urgent and critical.

III. PRELIMINARIES

A. Differential Privacy

Definition 1. *(ε-differential privacy [31]):* Suppose there exists a randomized algorithm A, given a dataset D and its neighboring dataset D', if the ratio of probabilities of the possible output range R when algorithm A is applied to the dataset and its neighboring dataset satisfies:

$$\frac{\Pr(A(D) \in R)}{\Pr(A(D') \in R)} \leq \exp(\varepsilon). \tag{1}$$

Then Algorithm A satisfies ε-differential privacy, where the parameter ε is the privacy budget of differential privacy, which is designed to measure the degree of privacy-preserving of Algorithm A.

Definition 2. *(Global sensitivity [31]):* Suppose there exists a set containing multiple query functions. Given adjacent datasets D and D', for any query function Q, its global sensitivity Δf is defined as the maximum difference between the results obtained from the adjacent datasets.

$$\Delta f = \max_{D,D'} \|Q(D) - Q(D')\|_1 \tag{2}$$

where $\|\|_1$ is referred to as the L_1-norm.

B. Hidden Markov Model

The Hidden Markov Model (HMM) [32] is capable of internally simulating and quantifying the temporal associations between unseen hidden states along a trajectory, enhancing the accuracy of future location predictions without the need for external manual adjustment of errors. HMM is represented by the triplet $\mu = (A, B, \pi)$. $\mathbf{A}$ denotes the state transition matrix where $a_{ij} = P(X_t = x_j \mid X_{t-1} = x_i)$, indicating the probability of transitioning from state x_i at time $t-1$ to state x_j at time t. $\mathbf{B}$ represents the emission probability matrix where $b_i(k) = P(O_t = o_k \mid X_t = x_i)$, representing the probability of state x_i producing observable output o_k. π is the initial state probability vector where $\pi_i = P(X_1 = x_i)$, representing the probability of starting in state x_i at time $t = 1$. The HMM possesses several key features:

1) **Markov Property:** Within the HMM framework, the probability of a current state occurring is influenced solely by its immediate predecessor, independent of any other states.
2) **Hidden Markov Chain:** The model encompasses an intrinsic hidden Markov chain, which is responsible for describing the temporal evolution of these hidden states.
3) **Observation Generation Mechanism:** The observed data in the model are generated by the hidden states according to specific probability distributions. The probability distribution for generating particular observational data varies across different hidden states.
4) **Parameter Estimation:** The parameters of the model are determined either through maximum likelihood estimation methods or Bayesian approaches.
5) **Sequence Prediction and Interpretation:** HMM is utilized for predicting future observational sequences and also for understanding and classifying current observational sequences.

C. Trajectory Database

A trajectory database D contains numerous trajectories of moving users, wherein each trajectory T represents a data record of the trajectory database D. The range of position nodes in the trajectory database D is

$$\mathrm{T} = \{\mathrm{loc}_1, \mathrm{loc}_2, \ldots, \mathrm{loc}_t\} \tag{3}$$

In the trajectory database D, the position node loc at time i is represented by coordinates, where $\mathrm{loc}_i = (x_i, y_i)$, and both latitude and longitude values are stored as spatial information in the database.

The location points in the trajectory database are broadly categorized into two types: stay points and pass-by points. Stay points are geolocation markers where a mobile entity stays at a location for a certain period, while pass-by points are trajectory segments connecting two consecutive stay points. To fully realize the semantic meaning of stay points, we adopt a mapping strategy where each stay point corresponds to multiple specific semantic information. This is mainly because stay points are more closely related to the actual location of an individual's activities, which is crucial for accurately inferring an individual's daily life pattern.

TABLE I
TABLE OF NOTATIONS

Notation	Definition
D_h	The historical trajectory database
D'	The filtered historical trajectory database
$\tilde{D}$	The predicted trajectory database
loc_i	The location of the user at time point i
T_a	$User_a$'s original trajectory
$T_{a'}$	$User_a$'s publishing trajectory
T_k	A trajectory in the historical trajectory database
$\overrightarrow{\text{Sem}}_p$	The semantic feature vector at location p.
θ, γ	The semantic similarity threshold
V_{ai}	The semantic feature vector of user a at point i
ξ_i	User predefined semantic location sensitivity
SL_s	User's semantic sensitivity set at location s
PL_i	The privacy level of location i
PW_i	The privacy weight of location i

D. Location Semantic Information

The semantic implications of a location are closely associated with the behavior patterns of active users and the specific category of the location. By comparing users' trajectory data with the functional zones on the map, we can infer the users' behavioral intentions [33]. This characteristic, which describes the functional division of locations, is referred to as the semantic information of a location.

Location Semantic Information: The semantic attributes of a location can be precisely characterized by its spatial position through latitude and longitude coordinates, a process analogous to the mapping and positioning mechanism in Geographic Information Systems (GIS). By conducting a thorough analysis of an individual's daily movement trajectories, we can systematically extract and construct a set of semantic features for a location. From the constructed foundational attributes, m key features that a location possesses are selected to form a semantic feature vector, which serves to represent the semantic properties of that location.

The spatial position of each node loc_i in the trajectory database is represented by its coordinates $loc_i = (x_i, y_i)$, which facilitates an accurate portrayal of its geographical location. For the semantic characterization of each location, a semantic feature vector $\overrightarrow{\text{Sem}}_i$ is constructed, consisting of m key semantic features $\overrightarrow{\text{Sem}}_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$ associated with the location node loc_i. Each semantic feature s_{ij} is a binary value that reflects the presence (1) or absence (0) of a particular semantic attribute at the given location. The feature vectors corresponding to all positional points within a single trajectory collectively constitute the semantic feature matrix:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \end{bmatrix} \tag{4}$$

where rows represent positional points and columns represent semantic attributes.

Semantic Frequency: Semantic frequency refers to the probability that a location of a given semantic type is visited. This can be calculated by counting the number of visits to locations of a specific semantic type and then dividing by the total number of visits. Assume there is a dataset of trajectory data that includes various types of semantics, such as restaurants, shops, and parks. By counting the number of times locations of a particular semantic type are visited in the historical trajectory data, we can determine the visiting frequency for different semantic types. The method for calculating the frequency of semantic information is as follows:

$$\text{freq}_j = \frac{N_j}{\sum_{j=1}^{|\text{Geoset}|} N_j} \tag{5}$$

where N_j represents the number of occurrences of semantic type s_j, and |Geoset| denotes the set of geographic locations within the entire region, the size of which is equivalent to the number of locations visited in the trajectory dataset for the region.

Semantic Correlation: We analyze the spatial relationships and semantic similarities between different locations using the feature vectors Sem_p and Sem_q, corresponding to locations p and q with coordinates (x_p, y_p) and (x_q, y_q) respectively. One commonly used metric is the Jaccard index, which evaluates the degree of similarity by comparing the intersection and union of the feature sets of the two locations. This metric, however, may oversimplify the relationships by focusing only on the presence or absence of features without considering their distribution or diversity. Conversely, cosine similarity provides a more nuanced assessment by evaluating the orientation agreement between the two feature vectors. The calculation method is as follows:

$$\begin{aligned} \text{SemSim}(\overrightarrow{\text{Sem}}_p, \overrightarrow{\text{Sem}}_q) &= \frac{\langle \overrightarrow{\text{Sem}}_p, \overrightarrow{\text{Sem}}_q \rangle}{|\overrightarrow{\text{Sem}}_p||\overrightarrow{\text{Sem}}_q|} \\ &= \frac{\sum_{p=1}^{m} \sum_{q=1}^{m} s_p s_q}{\sqrt{\sum_{p=1}^{m} s_p^2} \sqrt{\sum_{q=1}^{m} s_q^2}} \end{aligned} \tag{6}$$

Based on the relationship between vectors, if two vectors are in the same direction (i.e., co-directional), the angle between them is zero, resulting in a cosine value of 1, indicating the highest degree of similarity between these vectors. Therefore, the semantic similarity of locations is directly proportional to the cosine value between their semantic vectors. The larger the cosine value between the vectors, the stronger their similarity and the smaller their differences; conversely, the smaller the cosine value, the lower their similarity and the greater their differences. Typically, semantic similarity is denoted by $\text{SemSim}(\overrightarrow{\text{Sem}}_p, \overrightarrow{\text{Sem}}_q)$, with values ranging from 0 to 1. A higher value indicates greater semantic similarity between two locations.

IV. System Overview

A. System Model

Our scheme proposes a new mechanism called SCTP to deal with the problem of semantic correlation privacy publishing of trajectories among users. We describe the design principle

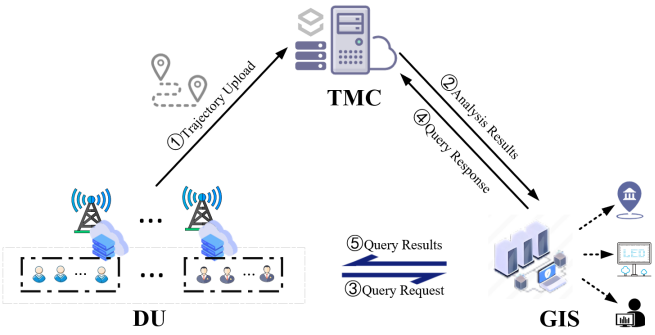

Fig. 1: System model

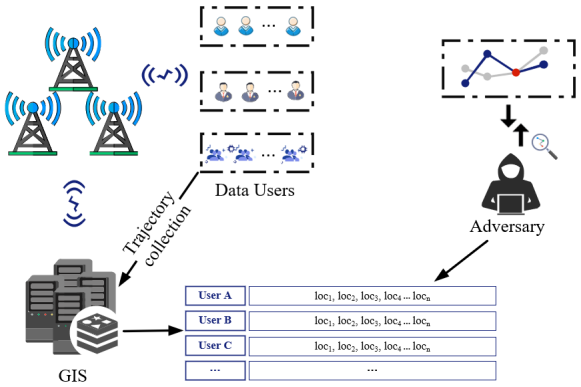

Fig. 2: Security model

and working mechanism of the SCTP mechanism in detail using the example of users enjoying location-based perceptual services (such as navigation guidance and social check-in). As shown in Fig. 1, the system model consists of three core entities: the Data Users(DU), the Trajectory Management Center (TMC), and the Geographic Information Server (GIS). The details are as follows.

*1) **Data Users (DU):*** When data users utilize certain social applications, they upload their trajectory information to trusted TMC. In practice, trajectory data is continuously uploaded, with applications automatically collecting and uploading location coordinates at predefined time intervals or when users move a certain distance, resulting in the upload of a complete trajectory each time. Through these applications, users can discover nearby amenities such as restaurants, gyms, or hospitals. Furthermore, these location data can also be utilized for personalized recommendations, route planning, or social interactions.

*2) **Trajectory Management Center (TMC):*** The TMC serves as a crucial intermediary server situated between users and the GIS. Its primary goal is to establish a secure and trustworthy data exchange environment to safeguard user privacy against attacks from untrusted third-party servers.

The primary responsibility of the TMC is to collect users' raw trajectory data and process their query requests, securely releasing users' sensitive trajectory information using advanced privacy-preserving techniques. In doing so, TMC maps each geographical location point to a semantic feature vector $\overrightarrow{\text{Sem}_i}$, integrating the semantic information of the location. This implies that within the trajectory dataset, each location point is labeled with a set of feature tags that explicitly denote the semantic category of the place, such as hospital, school, cinema, and so forth. Furthermore, TMC conducts preprocessing of the collected location trajectory data. Next, for every trajectory uploaded by users, TMC performs semantic relevance measurement and synthesizes based on HMM and differential privacy. If the semantic correlation measurement value falls within a predefined threshold, TMC will require reselection to protect the semantic correlation privacy between different trajectories. Finally, TMC responds to requests from the GIS by releasing trajectory data that has undergone semantic correlation privacy-preserving processing.

In practical applications, the TMC may be a data center managed by a professional team or a service institution regulated by the government or relevant industry associations. It plays a vital role in protecting user privacy and facilitating the secure exchange of data, thereby providing users with a safe and reliable location service environment.

*3) **Geographic Information Server (GIS):*** The GIS is a specialized server dedicated to storing, managing, and disseminating various types of geographic information data. It plays a crucial role in receiving user query requests, obtaining query results from the TMC, and responding to user needs. As a provider of location services, GIS works closely with TMC to offer accurate and timely location information services while ensuring the effective protection of user privacy. This collaborative working mechanism not only enhances the security and usability of location services but also provides users with a safer and more comfortable experience.The primary responsibilities of GIS include the storage, management, and publication of various geographic trajectories and related location semantic information data. GIS is capable of receiving user requests and retrieving query results from TMC to meet user demands.

B. Security Model

The objective of the SCTP is to safeguard the privacy of semantic correlations among different users' trajectory data and their location trajectories. In this security model, we assume that the TMC and DU are honest and trustworthy; the TMC faithfully executes our data processing scheme, whereas the GIS is considered honest but curious. Briefly, being honest yet curious means that participants strictly adhere to the protocol's execution, but at the same time, they might possess curiosity about users' sensitive information, attempting to analyze and mine users' semantic data using the background knowledge they have or through differential attack techniques to infer personal sensitive information.

The security model for SCTP is shown in Fig. 2, the GIS is considered the most potentially threatening adversary, as it has access to users' location service requests and the capability to analyze sensitive semantic information revealed in users' location trajectories. Consequently, the design of the SCTP must address how to effectively counteract various privacy in-

fringements and data mining activities that GIS might initiate, ensuring that users' location privacy and personal data security are adequately protected.

C. Mechanism Architecture

Algorithm 1 SCTP

Input: T_a, dataset D, threshold θ, privacy budget ε
Output: T'_a

for T_k in D **do**
  Compute semantic correlation SemSim(T_a, T_k)
  if SemSim$(T_a, T_k) > \theta$ **then**
    for each point $i = 1$ to n in T_a **do**
      Execute Hidden Markov Model for Trajectory Prediction at point i
      Obtain predicted trajectory datase $\tilde{D}$
    end for
  end if
end for
Filter trajectories with correlation below θ from the predicted trajectory dataset $\tilde{D}$
Obtain Semantic Disparity Predicted Trajectory Database $SDPT$
Construct the predicted trajectory protection set $SDPT + T_a$
Personalized allocation of privacy budget ε based on semantic frequency
for each point t in T_a **do**
  $T'_a = T_a + Lap(b)$
end for
return T'_a

To implement the SCTP mechanism, we designed Algorithm 1 to process the semantic correlation of a new user's trajectory before privacy publishing. First, we judge the semantic similarity of trajectories between different users. If their similarity exceeds a predefined threshold θ, we need to filter the obtained set D' for trajectory privacy preservation. For the trajectories to be protected, the algorithm first utilizes the Hidden Markov Model for positional trajectory prediction to get the set of predicted trajectories $\tilde{D}$. Subsequently, based on the semantic similarity between the predicted trajectories and D', the predicted trajectories whose trajectory semantic similarity is lower than a threshold value of γ are filtered out and retained to obtain the set of semantically different predicted trajectories $SDPT$. Next, considering the sensitivity of locations and user preferences, a personalized privacy budget is assigned to each location based on the semantic correlates of the area in the trajectory. Finally, we add Laplace noise to the sequence of trajectory locations to be published for secure trajectory data publishing.

D. Design Goals

In our scheme, we aim to explore a mechanism for publishing highly usable trajectory data while preserving the semantic correlation and privacy of multi-user trajectories. The details are as follows:

- **Semantic Correlation Preservation:** Ensuring the protection of individual trajectory privacy and the preservation of semantic correlations between individuals is the fundamental requirement of our proposed scheme. By handling the semantic relationships among user trajectories based on trajectory similarity and semantic differences in predicted trajectories, this requirement is guaranteed through the allocation of personalized privacy budgets for different locations, achieving ε-differential privacy.
- **Enhancing Usability:** Our scheme employs the HMM to capture the intrinsic structures and temporal characteristics of trajectories, thereby enhancing the usability of individual trajectories. Furthermore, we filter and construct adjacent trajectory datasets based on the calculated semantic correlation coefficients before publishing, ensuring the high usability of the released trajectory data.

V. MODULES DESIGN

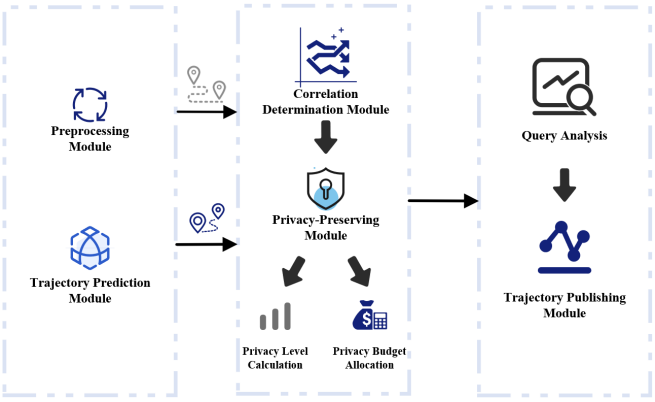

Fig. 3: Module Composition

In this section, we introduce an innovative approach for privacy protection focused on the semantic relevance of location data. Initially, we construct a privacy candidate set based on a single trajectory. Subsequently, utilizing this candidate set, we transform individual location data to effectively guard against potential privacy leaks arising from semantic associations among multiple users. Ultimately, the trajectory release mechanism we have designed allows for the secure publication of privacy-protected location information without disclosing any semantic relevance associated with the trajectories.The module composition of SCTP is shown in Fig. 3.

A. Preprocessing Module

Calculation of semantic similarity: Each trajectory can be represented by a semantic sequence based on the transfer between locations, so the calculation of semantic distance requires that both trajectories are of equal length and there is a one-to-one correspondence between the positions. Based on the semantic feature vectors corresponding to each location point, we define the semantic correlation coefficient as a measure of semantic similarity between two trajectories:

$$\text{SemSim}(V_a, V_b) = \frac{1}{n}\sum_{i=1}^{n}\frac{V_{ai}\cdot V_{bi}}{|V_{ai}||V_{bi}|} \tag{7}$$

The intermediate V_{ai} and V_{bi} respectively represent a position in the trajectory points V_a and V_b. First, it must be ensured that V_{ai} and V_{bi} can form an n-dimensional semantic feature vector: $V_{ai} = [s_{a1}, s_{a2}, \ldots, s_{an}]$.

If $\text{SemSim}(V_a, V_b) > \theta$, the semantic association between the two trajectories is utilized to compute the intensity of semantic association between each trajectory in the dataset D_h and the new user U_{new}'s trajectories above and below. Once this exceeds the threshold θ, the trajectory is filtered out from the TMC trajectory dataset D_h and ultimately forms the effective trajectory dataset D'.

B. Trajectory prediction module

Algorithm 2 Trajectory Prediction Algorithm

Input: historical trajectory dataset D; user trajectory T_a, threshold θ, privacy budget ε
Output: predicted trajectories $\tilde{D}$

```
Model initialization:
HMM training on D to estimate parameters A, B, π
Trajectory decoding:
Decode T to obtain the most probable sequence of hidden states
Trajectory prediction:
Initialization
for each position l_i in T do
    Predict the next position using HMM
    Add the predicted position to T'
end for
return D̃
```

The trajectory prediction algorithm of SCTP is shown in Algorithm 2. For the trajectories T uploaded by users, we employ a Hidden Markov Model (HMM) to dynamically predict user trajectories. This model proficiently forecasts the likely sequence of trajectories by conducting an in-depth analysis of the historical user trajectory dataset. It accomplishes this by learning and mastering the transition laws between hidden states and the corresponding observation probability distribution characteristics of each state. When dynamically predicting each specific temporal node on the trajectory, the system ascertains the most probable next positional state based on the observed state at the current node and the pre-calculated model parameters, particularly the state transition matrix. The Hidden Markov Model's intrinsic capability to model and quantify the time-series correlations between hidden states, which are not observable on the trajectory, significantly enhances the accuracy of future position predictions without the necessity for external manual error adjustment. Practically, the model iteratively conducts n prediction calculations according to this mechanism, subsequently generating a set of predicted trajectories for new user trajectories.

C. Correlation Determination Module

In this phase, we will apply the obtained trajectory prediction set to construct a data subset of neighboring trajectories. First, for the predicted trajectory collection and the subset of trajectories with high semantic similarity to the target original trajectories that have been rigorously filtered from the historical trajectory database, we will further implement an in-depth semantic relatedness assessment aiming at quantifying and confirming the degree of correlation between the two at the semantic level.

In this scheme, we define that the absolute value of semantic similarity less than or equal to γ indicates a weak correlation; greater than γ indicates a strong correlation. Therefore, we need to filter out the trajectory dataset with the absolute value of correlation $\text{SemSim} \leq \gamma$ and further process it to protect its privacy. This threshold can be set freely according to the actual situation, for example, when $\gamma = 0.4$, if the absolute value of semantic similarity between two trajectories is greater than γ, it indicates that there exists a large semantic correlation between them, and we will carry out the preliminary screening of them, and then carry out the processing of privacy protection for trajectories that meet the requirements.

Among the predicted trajectories, the trajectories with semantic correlation $\text{SemSim} \leq \gamma$ between the trajectories are selected for privacy-preserving and constitute the predicted trajectory privacy-preserving set.

The trajectories in the predicted trajectory dataset $\tilde{D}$ based on the new user U_{new} will be measured for semantic relevance with each user U' in D', if $\text{SemSim} \leq \gamma$, the predicted trajectory of the user will be put into Semantic Disparity Predicted Trajectory Set, otherwise, the trajectories with $\text{SemSim} \leq \gamma$ will be continued to be selected from the predicted trajectories to be judged for semantic relevance, then the semantic disparity predicted trajectory set of the user U_{new}, Semantic Disparity Predicted Trajectory Set (SDPT) can be represented as:

$$SDPS_n = \{sdps_n \in (SDPT - SemSim_\gamma)\} \tag{8}$$

The trajectory of a new user can be denoted as T_{new}, and the predicted trajectory privacy-preserving set is $PTPP_n$: $PTPP_n = SDPT_n + T_{new}$. Only one trajectory is different between this trajectory dataset and the set of semantically differentiated predicted trajectories, and thus $PTPP_n$ and $SDPT_n$ can constitute neighboring datasets.

D. Privacy-Preserving Module

We propose a comprehensive allocation scheme for privacy protection levels concerning location points, which takes into account various factors. This scheme comprehensively considers factors such as semantic point frequency information and semantic location sensitivity, thereby assigning appropriate privacy protection levels to each location point.

For each location, we utilize the frequency of semantic information appearing in its vicinity as the basis for weighting. The weight of a location is calculated by weighting the frequency of semantic information surrounding it. For instance, if sensitive information (such as hospitals or homes) appears frequently near a location, the weight of that location should be higher, indicating a need for stronger privacy protection.

Assuming that the semantic location sensitivity, predefined by users, is initialized to a value between 0 and 1, the set

representing the semantic location sensitivity of user trajectories can be denoted as $SL_s = [\xi_1, \xi_2, \ldots, \xi_m]$. We define the semantic frequency of a location point, denoted as $freq_j$, as the proportion of this semantic information within the entire set of semantic information in the trajectory database, as previously defined in Equation (5).

Assuming that the new user's trajectory during a period t is represented as $T_{\text{new}} = \{\text{loc}_1, \text{loc}_2, \ldots, \text{loc}_t\}$, where the semantic information for each location can be denoted as $\overrightarrow{sem}_{\text{new}} = [s_1, s_2, \ldots, s_m]$, the calculation method for the privacy level $PL(q)$ possessed by each location $q(q \in [1, t])$ in this trajectory is as follows:

$$PL(q) = \sum_{j=1}^{m} s_j \cdot \xi_j \cdot freq_j \tag{9}$$

Finally, it is necessary to normalize the privacy level weights of all the user's locations to ensure that their sum equals 1. This ensures that the resulting location weights accurately reflect the relative importance of various semantic information. Through this method, we can compute the weight of each location, which can more accurately reflect the prevalence of semantic information surrounding the location, thereby assessing the intensity of privacy protection needs for each location.

However, in traditional privacy budget allocation models, when the privacy budget ε is small, it indicates a higher level of privacy protection, which is inversely proportional to the privacy level. That is, in differential privacy, the privacy budget and the level of privacy protection are negatively correlated. Thus, the privacy budget allocation formula can be derived as follows:

$$\varepsilon_i = \frac{1/PL(q)}{1/\sum_{i=1}^{t} PL(q)}\varepsilon \tag{10}$$

Our research framework is built upon the published set of predictive trajectory privacy protection $PTPP_n$, aiming to provide semantic relevance protection for user trajectory data. Thus, the sensitivity of the query function is reflected in the maximum potential change it may induce in individual trajectory information. We conduct multiple queries, and the absolute difference in sensitivity values increases with each query. For example, when querying for nearby restaurants, assuming the number of queries is t, then the sensitivity of the queries can be represented as $\Delta f = t$. Therefore, the sensitivity of the query function in the $SCTP$ mechanism depends on the variation in semantic information within individual trajectory data. This variation influences the maximum potential change in the query function.

E. Trajectory Publishing Module

We propose a trajectory release mechanism that allocates privacy budgets based on the aforementioned computational results. This mechanism adds noise to trajectory data to securely release the trajectory data while preserving the semantic relevance of trajectories. The detailed process is shown in Algorithm 3.

Definition 3. *(Trajectory Noise Addition Mechanism):* We augment each location, loc_i, in the user's trajectory T_n with noise drawn from a Laplace distribution, resulting in trajectory T'_n. Thus, for any function f with a global sensitivity of Δf, algorithm F satisfies ε-differential privacy:

$$F(T_{loc_i}) = f(T_{loc_i}) + < lap(b_1), ...lap(b_n) > \tag{11}$$

where loc_i represents the user's actual location, while b is computed from a global sensitivity of Δf and a privacy budget ε, denoted as $b = \Delta f/\varepsilon$, with $lap(b)$ following a Laplace distribution. Its probability density function can be expressed as:

$$\Pr(x) = \frac{1}{2b} exp(-\frac{|loc_i|}{b}) \tag{12}$$

Algorithm 3 Trajectory Publication Algorithm

Input: Personalized privacy budget ε, Individual's true trajectory T_a, Semantic Difference Prediction Set $SDPS_n$
Output: Individual privacy trajectory T'_a

```
if individual trajectories need to be published then
    The range of T'_a's publication is SDPS_n
    for Each location i in the trajectory T'_a do
        Calculate the privacy weight
        PW_i = PL_i / sum_{i=1}^{n} PL'_i
        Calculate the privacy budget for loc_i
    end for
end if
return T'_a = T_a + lap(b_a).
```

VI. Privacy Analysis

A. Security Analysis

To protect the semantic correlation between trajectories, this approach requires two aspects of analysis: first, the privacy protection of individual trajectories, and second, the protection of semantic correlation between different users' trajectories. Firstly, from the perspective of individual trajectories, our approach employs differential privacy to meet the privacy-preserving requirements.

Theorem 1. *The SCTP mechanism satisfies ε-differential privacy during the noise addition stage.*

The fundamental concept of the Laplace mechanism involves the addition of noise that conforms to the Laplace distribution of the original data. This ensures that the query results, after noise addition, comply with the constraints of differential privacy. The proof proceeds as follows:

Assume two adjacent datasets, Dataset SD and Dataset SD', differ only by a single trajectory, T_{real}. Let $A(SD)$ denote the initiation of corresponding query requests on these datasets. Algorithm A, which adds independent noise to the output of function f, results in query output T'. The quantity of noise added at each position is determined by the privacy budget ε, which is derived from the assigned privacy level, and by the global sensitivity Δf. Specifically, in the worst-case scenario, adding or removing a single trajectory affects a solitary query function by no more than 1. Therefore,

the procedure of introducing noise $Lap(1/\varepsilon_i)$ that follows a Laplace distribution at each position complies with ϵ-differential privacy.

From the probability density function of the Laplace mechanism, we can derive the following:

$$\begin{aligned}\frac{\Pr[A(SD)=T']}{\Pr[A(SD')=T']} &= \frac{\Pr[f(SD)+\mathrm{Lap}(b)=T']}{\Pr[f(SD')+\mathrm{Lap}(b)=T']} \\ &= \prod_{i=1}^{k} \frac{e^{-\varepsilon(|T'-f(SD)_i|/\Delta f)}}{e^{-\varepsilon(|T'-f(SD')_i|/\Delta f)}} \\ &= \prod_{i=1}^{k} e^{\left(\frac{\varepsilon|T'-f(SD)_i|-\varepsilon|T'-f(SD')_i|}{\Delta f}\right)} \\ &\le \prod_{i=1}^{k} e^{\left(\frac{\varepsilon|f(SD')_i-f(SD)_i|}{\Delta f}\right)} \\ &= e^{\left(\frac{\varepsilon\|f(SD')-f(SD)\|_1}{\Delta f}\right)} \\ &\le e^{\epsilon}\end{aligned}$$

In addition, regarding the semantic correlations among trajectories of different users, we employed a hidden Markov model to regenerate the original trajectories and processed them for semantic correlation. This ensures that the published trajectories, while similar to the original, have a low semantic correlation with those of other users, thus protecting user privacy. The formal analysis is as follows:

Assume that the user's trajectory is denoted by T_n, the trajectory of another user with high semantic similarity is represented by T'_n, and the published trajectory is indicated by T'.

$$\frac{\Pr(T'|T'_n)}{\Pr(T'|T_n)}e^{-\epsilon} \le \frac{\Pr(T'|T'_n)}{\Pr(T'|T_n)} \le e^{\epsilon}\frac{\Pr(T'|T'_n)}{\Pr(T'|T_n)} \tag{13}$$

From the analysis above, it can be concluded that the adversary cannot distinguish T_n and T'_n. Therefore, our scheme protects the privacy of semantic correlation among different users' trajectories.

B. Usability Analysis

To evaluate the effectiveness of the perturbed trajectory, we use the expected distance between each position in the original trajectory T_n and each position in its true released trajectory T'_n as the evaluation metric. It is formally represented as follows:

$$Usability = U(\sum_{T_n,T_n} \sqrt{|Dis(T_n,T'_n)|^2}) \tag{14}$$

The distance function Dis can be the Euclidean distance or the Dynamic Time Warping (DTW) distance, among others. Taking the Euclidean distance as an example, we can further obtain a value that comprehensively reflects the degree of deviation between the perturbed trajectory and the real trajectory, thereby effectively evaluating the validity of the perturbed trajectory.

To explore the influencing factors of data usability, we analyze two key dimensions: firstly, the spatial distance between the perturbed position and the original position, an increase in this distance will lead to a decrease in data usability; secondly, the total length of the trajectory data, as the length of the trajectory increases, the number of semantic points involved increases, and to maintain a higher level of accuracy, it is necessary to divide it into more detailed divisions, which will likewise reduce the availability of the data.

In the SCTP mechanism in this paper, we process continuous geographic data based on differential privacy by controlling the probability ratio e^{ϵ} between any two different real locations corresponding to the same published location, which is a flexible balance between privacy-preserving and data usability. This mechanism has greater advantages for processing continuous geographic location data, and through the application of differential privacy, it not only strengthens the protection of the user's geographic location, but also preserves the usability of the trajectory data, which provides a theoretically reliable and practically effective solution for semantic correlation privacy preservation of trajectories, and demonstrates the significant advantages and wide applicability of our scheme.

VII. Experimental evaluation

Environment Setup. To evaluate the performance of the location privacy-preserving method proposed in this paper, the algorithm was extensively tested in terms of data usability, privacy-preserving level, and algorithm runtime. The experiments were implemented using Python 3.9, utilizing the Taxi dataset [34], Gowalla dataset [35], and Geolife dataset [36]. The experimental environment in our scheme is PyCharm. The hardware setup consists of a 4.7GHz Core(TM) i7-12700H processor, 16.0GB of RAM, and a laptop running Windows 11. In this section, we evaluate the performance of our scheme and compare it with AdaTrace [19], DPT [27], and DPLQ [37]. Given the stochastic nature of all the generators used, to ensure data reliability, we performed each experiment five times and averaged the results.

A. Data Usability Analysis

By quantifying the similarity between the output of a query function Q before and after introducing noise, the potential impact of the privacy protection algorithm on data usability can be effectively revealed. To facilitate the analysis of the impact of changes in the privacy budget on the SCTP mechanism, we assume that the number of semantic categories for the user's position is 10 and the position privacy level (PL) is 0.7. From Fig. 4(a), it can be clearly observed that for the four different algorithms, the corresponding data usability values all show a decreasing trend as ϵ increases. The reason for this is that an increase in ϵ means adding less noise, which directly leads to an improvement in data usability. Given that the semantic set covered by the Taxi dataset is richer than that of the Geolife dataset, this characteristic difference is reflected in the corresponding metric values shown in Fig. 4(b) as an increasing trend. Similarly, compared to the Gollawa dataset, the Geolife dataset contains a wider range of semantics. This characteristic is consistently demonstrated in the results shown

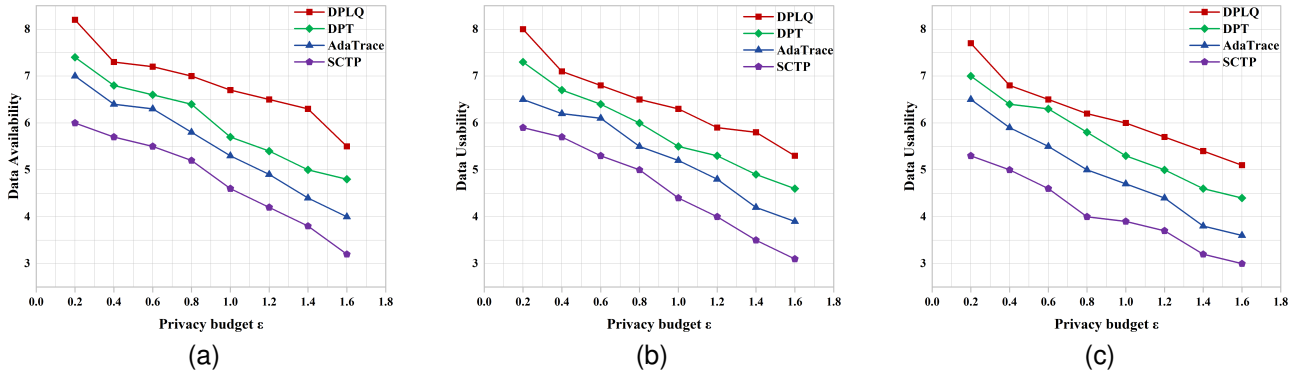

Fig. 4: Comparison of Data Usability under different ε. (a) Taxi dataset. (b) Geolife dataset. (c) Gollawa dataset.

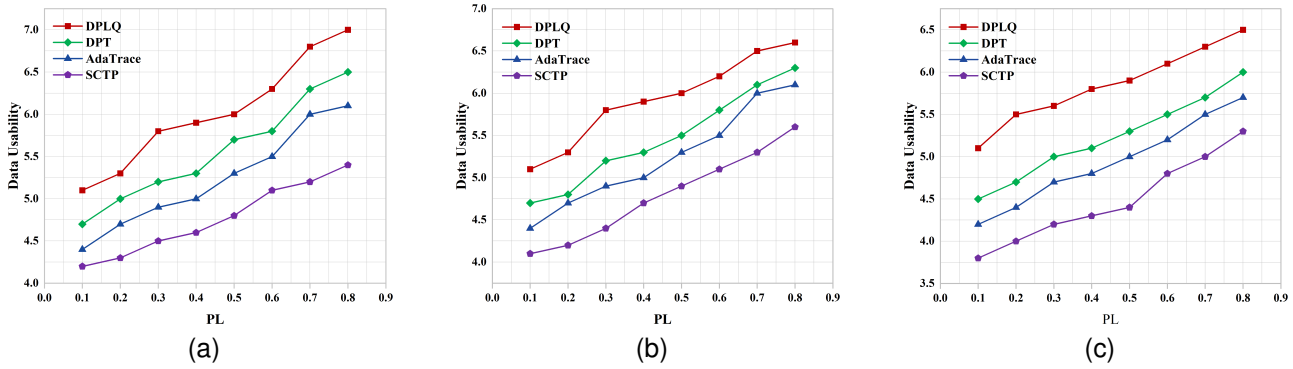

Fig. 5: Comparison of Data Usability under different PL. (a) Taxi dataset. (b) Geolife dataset. (c) Gollawa dataset.

in Fig. 4(c), where the corresponding metric values also exhibit a similar increasing trend.

As the number of trajectory locations increases, the number of key semantic points to be protected also increases, leading to a corresponding increase in the required privacy budget. Therefore, the discussion of the impact of trajectory position changes on the scheme's usability is omitted here. We are dedicated to analyzing the impact of changes in user privacy levels (*PL*) on data usability across different dataset environments, and we present our experimental results in Fig. 5.

We found that as the user privacy level increases, following the design principles of the SCTP mechanism, it becomes necessary to set a lower privacy budget to ensure the privacy protection of highly sensitive semantic information. This process introduces more noise into the original data, potentially reducing data usability. Fig. 5(a) first shows that as the privacy level of positions within a user's trajectory increases, the overall privacy weight calculated based on the semantic sensitivity of positions also rises. This implies that to maintain the same level of privacy protection, more privacy budget needs to be allocated, while more noise is added to obscure location information, which in turn affects data usability.

Experiments found that on different trajectory datasets, due to the varied distribution of semantic points within each dataset, the privacy weights of individual positions vary. For example, the Taxi dataset contains relatively more semantic points, hence the loss of data usability is higher compared to the Geolife and Gowalla datasets, as reflected in Fig. 5(b). Similarly, the Geolife dataset has more semantic points than the Gowalla dataset, and therefore its data usability decreases more significantly, as shown in Fig. 5(c).

In summary, our scheme successfully constructs a privacy protection mechanism that can dynamically adjust the allocation of privacy budgets and adapt to the characteristics of different datasets. The experimental results demonstrate that the SCTP scheme can effectively reflect and respond to the changing privacy protection needs of trajectory data, indicating that this scheme has a certain degree of practicality and effectiveness when dealing with large-scale trajectory data privacy protection issues.

B. Degree of Privacy Preserving

When comparing the SCTP method proposed in this paper with the AdaTrace, DPT, and DPLQ privacy protection algorithms in terms of privacy protection performance, the experimental results are as shown in Fig. 6. The X-axis represents different values of the privacy budget parameter ϵ, while the Y-axis quantifies the level of privacy protection provided by each algorithm, with the results expressed as the probability ratio between two different real locations corresponding to the same published location.

As seen in Fig. 6, the privacy protection effectiveness of the four algorithms exhibits a decreasing trend as the privacy

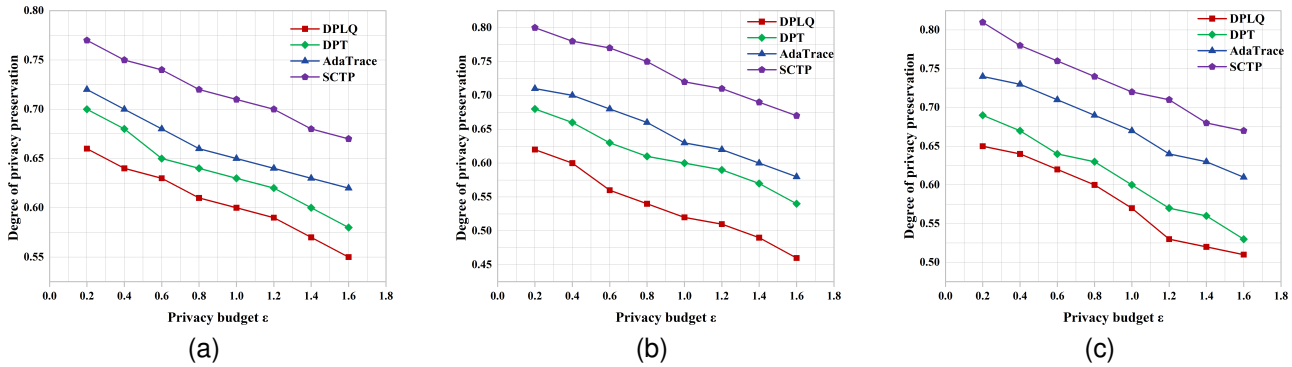

Fig. 5: Comparison of the degree of privacy preservation under different ε. (a) Taxi. (b) Geolife. (c) Gollawa.

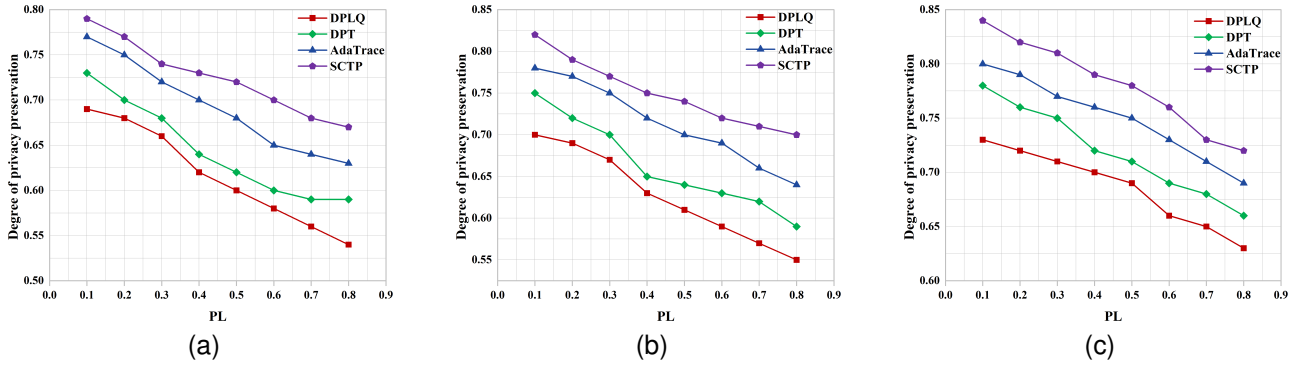

Fig. 6: Comparison of the degree of privacy preservation under different PL. (a) Taxi. (b) Geolife. (c) Gollawa.

budget parameter ϵ increases. This is because ϵ is a key parameter for measuring the risk of privacy leakage, and the larger its value, the less perturbation noise is injected into the location data, which directly weakens the privacy protection effect.

However, it is noteworthy that the SCTP scheme proposed in this paper integrates a personalized privacy allocation scheme based on semantic frequency to enhance the security of location information. Even under the general rule of a decline in overall privacy protection with increasing ϵ, this scheme still demonstrates higher privacy protection effectiveness compared to similar algorithms, thanks to its personalized privacy budget allocation strategy. In summary, even with larger ϵ values, the SCTP scheme may still offer a higher level of privacy protection than other algorithms.

Furthermore, we systematically investigated the effect of user position privacy level (PL) on various privacy-preserving schemes across different datasets and visualized the results in Fig. 7. As shown in Fig. 7(a), the level of privacy protection decreases as the user position privacy level (PL) increases.

This is because, according to the basic principles of differential privacy, a higher privacy level requires a larger privacy budget, which consequently weakens the degree of privacy protection. It is noteworthy that our scheme demonstrates good adaptability under this variable change.

It is important to note that among the three datasets, the semantic region division is the most detailed in the Taxi dataset, followed by the Geolife dataset, and the least detailed in the Gowalla dataset. Therefore, as the level of detail in the semantic information of the datasets increases, the privacy protection cost required to achieve the same level of privacy protection also increases, resulting in a relatively lower level of privacy protection. This pattern is confirmed in the results shown in Fig. 7(b) and Fig. 7(c). Even so, our scheme still demonstrates robust and relatively efficient privacy protection across the datasets.

C. Runtime Analysis

In terms of algorithm runtime, our scheme specifically considers the impact of the number of semantic categories in user trajectory positions and the position privacy level (PL) on the runtime of the scheme and compares our approach with other schemes. The results, as shown in Fig. 8(a), indicate that under the condition of maintaining a constant position privacy level of 0.7, as the number of position semantic categories increases, the scale of semantic points that the scheme needs to traverse also expands, leading to an increase in time costs. The comparison results reveal that the scheme employing the SCTP mechanism uses an efficient prediction and noise addition mechanism, thereby showing more robust performance in terms of the growth rate of runtime, especially demonstrating superior time efficiency when handling a large volume of geographic location data.

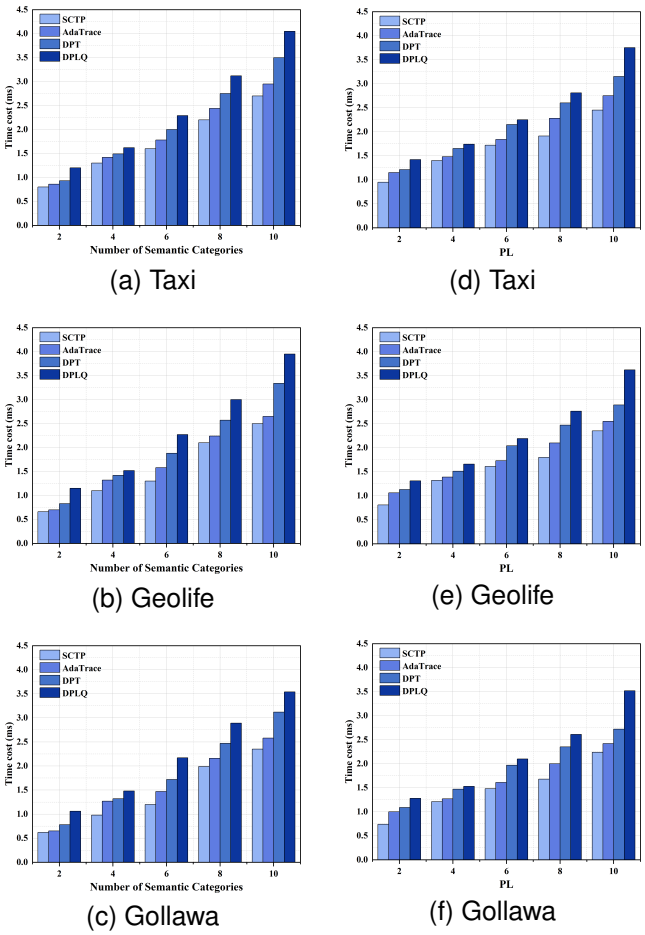

Fig. 7: The effect of semantic categories.

Fig. 8: The effect of semantic categories.

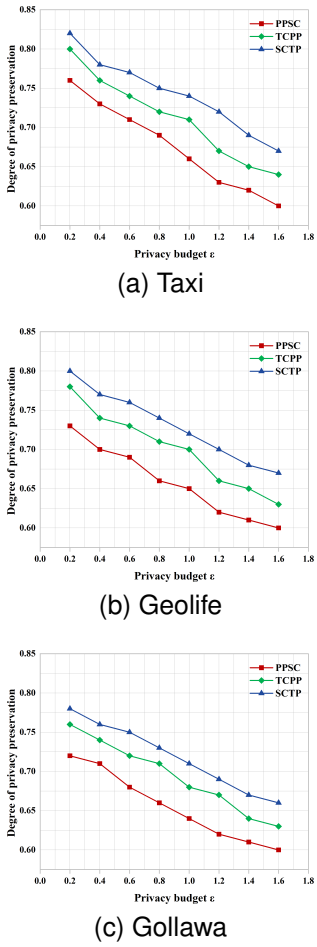

Fig. 9: The effect of ε.

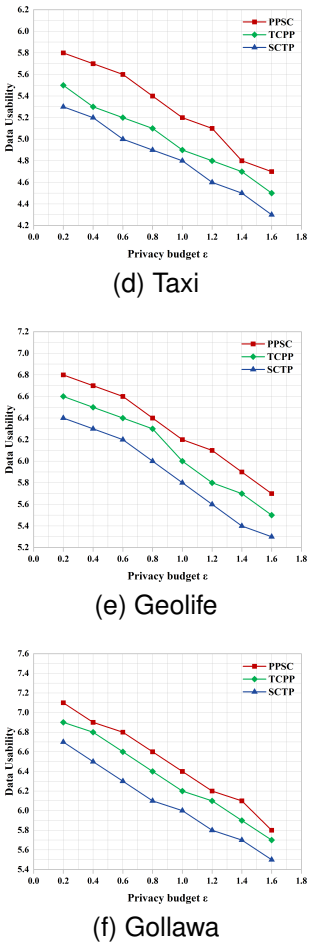

Fig. 10: The effect of semantic categories.

Additionally, this scheme further explores the impact of the user position privacy level (PL) on the scheme's runtime across multiple datasets. As shown in Fig. 9(a), the experimental results indicate that, with the number of position semantic categories fixed at 10, increasing the position sensitivity results in a corresponding increase in the required amount of noise, which in turn leads to an increase in the time cost of the scheme. The SCTP mechanism demonstrates stable or better performance across different datasets, particularly in datasets with larger semantic sets, where its performance is especially notable.

For the results shown in Fig. 9(b) and Fig. 9(c), the trends are similar to those in Fig. 9(a), but their runtime is shorter because the Taxi dataset has finer-grained divisions. This indicates that the SCTP mechanism exhibits significant applicability and efficiency when applied to different datasets.

D. Degree of Semantic Correlation Preserving

To thoroughly evaluate and demonstrate the superior performance of our scheme in protecting the semantic correlation privacy between the trajectories of two different users, we will conduct a comparative study between our scheme and the PPSC and TCPP schemes mentioned in [15], [38]. The study focuses on privacy-preserving level and data usability, with simulation testing and in-depth analysis conducted on both aspects.

When considering the level of semantic correlation of privacy protection between users' trajectories, we will focus on the privacy budget parameter ϵ and comparatively analyze how the privacy protection capability of our scheme changes under different privacy budgets compared to the PPSC and TCPP schemes. Our scheme considers the actual impact of user-defined privacy budget values on the privacy protection level across datasets of different sizes, with the experimental results shown in Fig. 10. According to differential privacy theory, as the privacy budget increases, the level of privacy protection decreases. Furthermore, our scheme demonstrates excellent performance, as it employs a personalized privacy budget approach that accounts for factors such as trajectory semantic frequency and position.

In parts (a), (b), and (c) of Fig. 10, although the overall trend is consistent, there are still differences in the level of privacy protection. These differences arise because the semantic content richness within the Taxi, Geolife, and Gowalla datasets decreases sequentially, which influences the specific manifestation of the privacy protection effect.

Regarding the exploration of the semantic correlation between the trajectories of two users in terms of data usability,

we will also examine and analyze the data usability performance of our scheme compared to the PPSC and TCPP schemes under different privacy budget settings. As shown in Fig. 11(a), Fig. 11(b), and Fig. 11(c), when the privacy budget increases, the evaluation metric for data usability also increases, which means that data usability gradually decreases with increasing ϵ.

Our scheme precisely adjusts the privacy budget by controlling the relationship between the distance from the actual position to the perturbed position and their corresponding probability ratio, effectively reducing the positional deviation and achieving a personalized privacy budget allocation. Consequently, our scheme significantly maintains efficient data usability.

VIII. Conclusion

In our scheme, we have proposed a Semantic Correlation Trajectory Privacy-preserving (SCTP) mechanism based on differential privacy, aiming to address the challenges of location services and user semantic privacy-preserving in IoV. By combining HMM and differential privacy, the SCTP mechanism ensures the semantic privacy of user trajectories while maintaining high-quality location services and data usability. Our scheme includes a trajectory prediction algorithm, which can dynamically and accurately predict vehicle trajectories and generate semantically relevant and highly available trajectory datasets. Additionally, a semantic frequency-based personalized privacy budget allocation strategy is designed, which achieves reasonable privacy budget allocation by setting privacy weights. Theoretical analysis and experimental validation demonstrate that our mechanism strictly satisfies ε-differential privacy and shows significant advantages in protecting the semantic privacy of user trajectories. Future research directions include introducing dynamic adjustment strategies for spatiotemporal attributes and random perturbation elements and deepening the privacy protection mechanisms for the semantic correlation of trajectory data.

Acknowledgements

This work was supported in part by the Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No.2023ZD021, in part by the National Natural Science Foundation of China under Grant 62071280, 62302280.

References

[1] W. X. Zhao, N. Zhou, W. Zhang, J.-R. Wen, S. Wang, and E. Y. Chang, "A probabilistic lifestyle-based trajectory model for social strength inference from human trajectory data," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 1, pp. 1–28, 2016.

[2] M. Gramaglia, M. Fiore, A. Furno, and R. Stanica, "Glove: towards privacy-preserving publishing of record-level-truthful mobile phone trajectories," *ACM/IMS Transactions on Data Science (TDS)*, vol. 2, no. 3, pp. 1–36, 2021.

[3] G. Qiu, D. Guo, Y. Shen, G. Tang, and S. Chen, "Mobile semantic-aware trajectory for personalized location privacy preservation," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 16 165–16 180, 2020.

[4] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1289–1297.

[5] C. Song and A. Raghunathan, "Information leakage in embedding models," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020, pp. 377–390.

[6] X. Zhao, D. Pi, and J. Chen, "Novel trajectory privacy-preserving method based on prefix tree using differential privacy," *Knowledge-Based Systems*, vol. 198, p. 105940, 2020.

[7] R. Tan, Y. Tao, W. Si, and Y.-Y. Zhang, "Privacy preserving semantic trajectory data publishing for mobile location-based services," *Wireless Networks*, vol. 26, pp. 5551–5560, 2020.

[8] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, pp. 1–29, 2011.

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

[10] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 546–563.

[11] Z. Riaz, F. Dürr, and K. Rothermel, "Understanding vulnerabilities of location privacy mechanisms against mobility prediction attacks," in *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2017, pp. 252–261.

[12] H. Cao, F. Xu, J. Sankaranarayanan, Y. Li, and H. Samet, "Habit2vec: Trajectory semantic embedding for living pattern recognition in population," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1096–1108, 2019.

[13] L. Ou, Z. Qin, Y. Liu, H. Yin, Y. Hu, and H. Chen, "Multi-user location correlation protection with differential privacy," in *2016 IEEE 22nd international conference on parallel and distributed systems (ICPADS)*. IEEE, 2016, pp. 422–429.

[14] L. Ou, Z. Qin, S. Liao, Y. Hong, and X. Jia, "Releasing correlated trajectories: Towards high utility and optimal differential privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 5, pp. 1109–1123, 2018.

[15] L. Wu, C. Qin, Z. Xu, Y. Guan, and R. Lu, "Tcpp: Achieving privacy-preserving trajectory correlation with differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4006–4020, 2023.

[16] Z. Zheng, Z. Li, H. Jiang, L. Y. Zhang, and D. Tu, "Semantic-aware privacy-preserving online location trajectory data sharing," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2256–2271, 2022.

[17] G. Qiu, Y. Shen, K. Cheng, L. Liu, and S. Zeng, "Mobility-aware privacy-preserving mobile crowdsourcing," *Sensors*, vol. 21, no. 7, p. 2474, 2021.

[18] S. Chang, C. Li, H. Zhu, T. Lu, and Q. Li, "Revealing privacy vulnerabilities of anonymous trajectories," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12 061–12 071, 2018.

[19] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 646–660, 2016.

[20] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1298–1309.

[21] C. Xu, L. Zhu, Y. Liu, J. Guan, and S. Yu, "Dp-ltod: Differential privacy latent trajectory community discovering services over location-based social networks," *IEEE Transactions on Services Computing*, vol. 14, no. 4, pp. 1068–1083, 2018.

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

[23] F. Jin, W. Hua, M. Francia, P. Chao, M. E. Orlowska, and X. Zhou, "A survey and experimental study on privacy-preserving trajectory data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5577–5596, 2022.

[24] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: a case study on the montreal transportation system," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 213–221.

[25] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3628–3636, 2018.

[26] Q. Han, Z. Xiong, K. Zhang *et al.*, "Research on trajectory data releasing method via differential privacy based on spatial partition," *Security and Communication Networks*, vol. 2018, 2018.
[27] X. He, G. Cormode, A. Machanavajjhala, C. Procopiuc, and D. Srivastava, "Dpt: differentially private trajectory synthesis using hierarchical reference systems," *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1154–1165, 2015.
[28] H. Wang, Z. Zhang, T. Wang, S. He, M. Backes, J. Chen, and Y. Zhang, "Privtrace: Differentially private trajectory synthesis by adaptive markov models," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, 2023, pp. 1649–1666.
[29] S. Ghane, L. Kulik, and K. Ramamohanarao, "Tgm: A generative mechanism for publishing trajectories with differential privacy," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2611–2621, 2020.
[30] X. Du, H. Zhu, Y. Zheng, R. Lu, F. Wang, and H. Li, "A semantic-preserving scheme to trajectory synthesis using differential privacy," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13 784–13 797, 2023.
[31] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, "Differentially private data publishing and analysis: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.
[32] S. Qiu, D. Pi, Y. Wang, and Y. Liu, "Novel trajectory privacy protection method against prediction attacks," *Expert Systems with Applications*, vol. 213, p. 118870, 2023.
[33] J. Ni, X. Lin, and X. Shen, "Privacy-preserving data forwarding in vanets: A personal-social behavior based approach," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.
[34] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393–1402, 2013.
[35] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1082–1090.
[36] Y. Zheng, X. Xie, W.-Y. Ma *et al.*, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.
[37] Q. Zhang, X. Zhang, M. Wang, and X. Li, "Dplq: Location-based service privacy protection scheme based on differential privacy," *IET information security*, vol. 15, no. 6, pp. 442–456, 2021.
[38] K. Dou and J. Liu, "Differential privacy trajectory protection method based on spatiotemporal correlation," in *Data Science: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2020, Taiyuan, China, September 18-21, 2020, Proceedings, Part II 6*. Springer, 2020, pp. 151–166.

Haojie Yuan received a B.S. degree in Computer Science and Technology from the School of Information Science and Engineering at Shandong Normal University in 2022. He is currently pursuing an M.S. degree in Computer Science and Technology at Shandong Normal University. His research interests include privacy preservation and differential privacy.

Lei Wu received his Ph.D. degree in applied mathematics from School of Mathematics, Shandong University in 2009. He is currently a professor with the School of Information Science and Engineering, Shandong Normal University, China. His research interests include applied cryptography, privacy preservation, and cloud computing security.

Lijuan Xu received Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2023. She is currently an Associate Professor with Shandong Computer Science Center (National Supercomputer Center in Jinan), China. Her main research interests include network security, industrial internet security, and computer forensics.

Libo Ban received a B.S. degree in Information and Computing Science from the School of Sciences at Shandong Jiaotong University in 2021. He is currently pursuing an M.S. degree in Computer Technology at Shandong Normal University. His research interests include privacy preservation and mobile crowdsensing.

Hao Wang received his Ph.D degree in Computer Science from Shandong University in 2012. He is currently a professor at Shandong Normal University. He has co-authored over 60 research papers in prestigious journals and conferences. His current research interests include applied cryptography, secure multi-party computation, and Blockchain.

Ye Su received her bachelor degree in School of Mathematical Sciences from the University of Jinan, China, in 2014. She received her Ph.D. degree in school of mathematics in Shandong University, Jinan, China, in 2022. She is currently a lecturer with School of Information Science and Engineering, Shandong Normal University. Her research interests include cryptography and cloud security.

Weizhi Meng (Senior Member, IEEE) received the Ph.D. degree in Computer Science from the City University of Hong Kong. He is currently an Associate Professor with the Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Denmark. He is currently directing the SPTAGE Lab. His primary research interests are blockchain, cyber security and artificial intelligence in security including intrusion detection, smartphone security, biometric authentication, trust management, and IoT security. He was a recipient of the Hong Kong Institution of Engineers (HKIE) Outstanding Paper Award for Young Engineers/Researchers in both 2014 and 2017, and received the IEEE ComSoc Best Young Researcher Award for Europe, Middle East, & Africa Region (EMEA) in 2020. He is an Associate Editor for IEEE TIFS and IEEE TDSC. He is a senior member of IEEE.