# TCPP: Achieving Privacy-Preserving Trajectory Correlation With Differential Privacy

Lei Wu, Chengyi Qin, Zihui Xu, Yunguo Guan, and Rongxing Lu, *Fellow, IEEE*

***Abstract*— The prevalence of mobile Internet, smart terminal devices, and GPS positioning technology has generated a vast number of trajectory data that location-based applications can utilize. However, delivering LBSs based on trajectories without extra protection may expose the personal information of users and even their social ties. Despite the fact that many works have been offered to achieve differential privacy for trajectory correlation, the vast majority of them only consider the trajectory correlation of a single user, and privacy protection for trajectory correlation amongst multiple users is not considered. Directly applying these works to protect correlation amongst multiple users may lead to the low availability of published trajectory data. To address the above challenges, we propose a trajectory correlation privacy-preserving mechanism (TCPP) that fulfills differential privacy. Specifically, we first apply the Euclidean distance to filter out a set of trajectories whose correlation needs to be protected. Then, we employ the Kalman filter to generate a dataset with high availability from the set of trajectories. Finally, we present a mechanism for publishing trajectories that preserves the trajectory correlation based on a customized privacy budget allocation strategy. Rigid security analysis shows that our proposed mechanism can well preserve the correlation privacy of trajectories. Experimental results on real-world datasets further demonstrate the privacy, availability and time efficiency advantages of our mechanism.**

***Index Terms*— Multi-trajectory correlation, differential privacy, privacy budget, privacy-preserving, data availability.**

## I. Introduction

With the popularization and rapid development of mobile Internet, smart terminal devices and GPS positioning technology, location-based services (LBSs) [1] such as point of interest query, traffic path planning and navigation, social network location sharing have become an indispensable part of daily life, bringing us a lot of conveniences. While providing accurate LBSs to users, the location-based service provider (LBSP) collects an enourmous amount of users' locations and trajectories.

However, LBSPs are not always honest and trustworthy. In order to further obtain more valuable personal and commercial information about users for profit, they will mine correlations amongst trajectories generated by single user or multiple users. These correlations can reflect or disclose personal information of a single user or certain social relationship information between two users [2]. As a conclusion, not only the basic location information of users should be preserved in location-based services, but also the leakage of correlation of users' location trajectories must be prevented or limited to preserve the privacy of social relationships amongst users, to achieve perfect LBS privacy preservation.

Research scholars have proposed some methods for trajectory privacy preservation: Generalization [3], Mix zones [4], Inhibition, and Disturbance. Unfortunately, these privacy-preserving mechanisms are challenging to apply to the correlation issue, and are only used to preserve the problem of privacy leakage without correlation within a single trajectory. Nevertheless, in the face of adversaries' background knowledge attacks, combination attacks [5], etc., adversaries can still obtain users' personal privacy information. To overcome the aforementioned attacks, differential privacy [6] was developed, which makes use of strict mathematical theory to protect the relevance of users' trajectories in order to achieve the goal that attackers cannot obtain relevant personal privacy information by the analysis of background knowledge. As detailed in Section VII, research on correlations within a single trajectory has gradually developed and matured [7], [8], using Fourier coefficients [9], Gaussian white noise [10], and Markov models [11] to preserve correlations within a single trajectory under a differential privacy-based preservation model. The relevance literature [12], [13] has also used privacy-preserving techniques such as pufferfish [14] to preserve the correlation of the trajectory within a single individual in recent years. Nevertheless, none of the above literature considered the privacy issues exposed by the trajectory correlation between two individuals.

To the best of our knowledge, there is less literature on the correlation between two trajectories [15], [16], as it is

Manuscript received 14 June 2022; revised 6 February 2023 and 30 May 2023; accepted 24 June 2023. Date of publication 28 June 2023; date of current version 6 July 2023. This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2020MF056 and Grant ZR2020KF011, in part by the Henan Key Laboratory of Network Cryptography Technology under Grant LNCT2021-A12, in part by the National Natural Science Foundation of China under Grant 62071280, and in part by the Major Scientific and Technological Innovation Project of Shandong Province under Grant 2020CXGC010115. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chia-Mu Yu. *(Corresponding author: Lei Wu.)*

Lei Wu is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China, also with the Henan Key Laboratory of Network Cryptography Technology, Zhengzhou 450001, China, and also with the Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250358, China (e-mail: wulei@sdnu.edu.cn).

Chengyi Qin and Zihui Xu are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (e-mail: qcy521111@163.com; xuzihui994@163.com).

Yunguo Guan and Rongxing Lu are with the Faculty of Computer Science (FCS), University of New Brunswick (UNB), Fredericton, NB E3B 5A3, Canada (e-mail: yguan4@unb.ca; rlu1@unb.ca).

Digital Object Identifier 10.1109/TIFS.2023.3290486

challenging to measure the correlation between trajectories and to optimize the privacy budget in differential privacy to improve data availability. If only the correlation of trajectories between two individuals is considered, after adding some noise to them, the attacker can still obtain correlation information from another user's trajectory, resulting in privacy information leakage. To address the above challenges, a trajectory privacy preservation mechanism (TCPP) based on differential privacy is proposed in this paper. This scheme is featured with a strict definition and quantification of trajectory correlation, a personalized and reasonable allocation of privacy budget to protect the correlation between trajectories from adversaries.

The main contribution points of this paper are as follows:

- We investigate the trajectory correlation privacy issues amongst different users, as well as propose a trajectory correlation privacy preservation mechanism (TCPP) based on differential privacy, which designs a Kalman filter-based trajectory prediction algorithm, a differential privacy to preserve trajectory correlation and publish users' trajectories efficiently and securely, achieving the goal of enjoying high-quality location-based services. In addition, we rigorously prove that the TCPP mechanism satisfies $\varepsilon$-differential privacy, and perform a formal analysis of the security, usability, as well as background knowledge of the adversary of the scheme.
- The prediction algorithm based on TCPP uses Kalman filtering for the dynamic prediction of user trajectories to construct highly available correlated datasets of trajectories. Compared with existing filtering mechanisms, this prediction algorithm predicts each location point and corrects them by the prediction error, which is still stable for time-honored real-time trajectories, which improves the availability of trajectory data.
- We propose a personalized privacy budget allocation strategy. Based on the user's time and distance, we design a location privacy level algorithm (LPL), define the privacy weight to achieve the goal of personalized and reasonable allocation of a given privacy budget, which can better balance the noise error and prediction error to improve the utility of published trajectory data.
- We evaluate the proposed TCPP mechanism and compare our approach with other related mechanisms on three real datasets from different aspects, respectively. The analysis of the experimental results shows the superiority of the mechanism.

The remainder of the paper consists of the following sections: In Section II, we introduce the technical approach used in the preparation of the paper and define some relevant knowledge. We detail the system model and the mechanism architecture of our scheme and elaborate on the functionality of each entity in Section III. We describe the system flow and elaborate on the process of each module in Section IV. In Section V, we describe and prove the security of the scheme. In Section VI, we evaluate the performance of the scheme. In Section VII, we detail the literature on trajectory correlation privacy preservation, and finally, in Section VIII, we describe in detail the conclusion and future prospects.

TABLE I
TRAJECTORY DATABASE

| Sequence number | Trajectory $T$ |
|---|---|
| 1 | $T_1 = \{(loc_1, t_1), (loc_2, t_2), (loc_3, t_3)\}$ |
| 2 | $T_2 = \{(loc_2, t_1), (loc_6, t_2), (loc_3, t_3)\}$ |
| 3 | $T_3 = \{(loc_3, t_1), (loc_4, t_2), (loc_1, t_3)\}$ |
| 4 | $T_4 = \{(loc_1, t_1), (loc_4, t_2), (loc_5, t_3)\}$ |

## II. PRELIMINARIES

### A. Data Model

*Definition 1 (Trajectory):* A trajectory is composed of a finite number of location-time nodes:

$$T = (loc_1, t_1) \rightarrow (loc_2, t_2) \rightarrow \ldots \rightarrow (loc_w, t_w), \quad (1)$$

here, it is assumed that all trajectories are of equal length and acquired at the same frequency, where $T(t_i) = loc_i (\forall i, 1 \leq i \leq |T|)$, $|T|$ denotes the length of the trajectory $T$ and $loc_i$ denotes the spatial location point consisting of longitude and latitude at interval $t_i$.

*Definition 2 (Trajectory Database):* The trajectory database $D$ is composed of a sequence number, a trajectory $T$ represented by a time-location sequence. In which, the trajectory $T$ represents a data record of the trajectory database $D$. The range of location nodes in the trajectory database $D$ is $Loc = \{loc_1, loc_2, \ldots, loc_n\}$, which is represented as shown in Table I.

*Definition 3 (Trajectory Distance):* The trajectory distance is the set of distances between different trajectories at the same interval between position points. Suppose the trajectory distance between $User_a$ and $User_b$ is denoted as $Dis(a, b)$, which is formalized and defined as shown below:

$$Dis(a, b) = \{(d_1, t_1), (d_2, t_2), \ldots, (d_w, t_w) | i \in [1, w]\}, \quad (2)$$

where $d_i$ denotes the Euclidean distance between $User_a$ and $User_b$ at the $t_i$ interval, which is denoted as $d_i = (loc_{ai} - loc_{bi})$.

### B. Differential Privacy

Differential privacy can be achieved by adding randomized noise to aggregated query results to preserve individual entries without significantly altering the query results, ensuring that attackers have access to almost as much individual data as they would have from a dataset without such individual record.

*Definition 4 (Differential Privacy [17]):* Suppose $M$ is a randomized algorithm, $D_1$ and $D_2$ are a set of adjacent datasets that differ by only one record, for any output $O$ of algorithm $M$ on adjacent datasets satisfies the inequality:

$$\Pr[M(D_1) = O] \leq e^{\varepsilon} \times \Pr[M(D_2) = O], \quad (3)$$

then the randomized algorithm $M$ satisfies $\varepsilon$-differential privacy.

*Definition 5 (Sensitivity [17]):* Suppose there is a function $F : D \longrightarrow R_m$ with an input dataset $D$ and an output $m$-dimensional real vector with sensitivity for adjacent datasets $D_1$ and $D_2$ as:

$$GF_F = max \| F(D_1) - F(D_2) \|_1, \quad (4)$$

where $\| \|_1$ is called the $L_1$-norm.

TABLE II
SYMBOL DESCRIPTION

| Symbols | Description $T$ |
|---|---|
| $T_a, T_b$ | $User_a$' trajectory, $User_b$' trajectory |
| $loc_i$ | The location of the user at time point i |
| $D$ | Trajectory Database |
| $Dis(a,b)$ | The distance between $User_a$ and $User_b$ |
| $d_i$ | Euclidean distance between the location of $User_a$ and $User_b$ at interval $i$ |
| $U$ | Trajectory availability metrics |
| $Sim_{ab}$ | Correlation coefficient between $User_a$ and $User_b$ |
| SL | Set of sensitive locations: SL=$\{sl_1, sl_2, ...sl_n\}$ |
| PL | Set of privacy level: PL=$\{pl_1, pl_2, ...pl_n\}$ |
| $weight_{t_i}$ | Weight value of user location at interval $t_i$ |
| $k$ | Privacy Level Weights |
| $noise$ | A noisy set of location points in a trajectory, $noise = \{noise_1, noise_2, ..., noise_w\}$ |
| $b_i$ | The Laplace scale for the location at interval $i$ |

### C. Kalman Filter

The Kalman filter [18] cleverly blends the observation data and the estimation data to limit the error to a certain range, and the error can still remain stable over time. The Kalman filter optimally estimates the system state by the system input and output observation data, and the state equation and observation equation of its dynamic trajectory prediction system are shown as follows:

$$\begin{aligned} X(k+1) &= A(k)X(k) + T(k)W(k) \sim N(0, Q) \\ Z(k) &= H(k)X(k) + V(k) \sim N(0, R), \end{aligned} \tag{5}$$

where $X(k)$ denotes the system state vector; $A(k)$ denotes the state transfer matrix; $T(k)$ denotes the disturbance transfer matrix; $W(k)$ denotes the system state noise of the motion model; $Z(k)$ denotes the observation vector; $H(k)$ denotes the observation matrix; and $V(k)$ denotes the observation noise generated during the motion trajectory. Suppose that $W(k)$ and $V(k)$ are mutually independent Gaussian white noise with covariances $Q$ and $R$, respectively.

### D. Data Availability

*Definition 6 (Data Availability):* Dividing each day into $w$ intervals with time variable $t \in [1, w]$, this scheme uses the distance between the published location $ploc_i$ in the perturbed trajectory and the location $rloc_i$ in the real trajectory as an error evaluation criterion which measures the location of perturbed trajectory availability. Thus, the measure of the availability of a trajectory of length $|w|$ is shown below:

$$U \equiv \frac{1}{w}\sum_{i=1}^{w}\left[\sqrt{|ploc_i - rloc_i|^2}\right], \tag{6}$$

where the larger the $U$, the worse the data availability, and vice versa.

## III. SYSTEM OVERVIEW

### A. System Model

Third party-based architecture consists of mobile users, a Trusted Third Party (TTP), and the LBSP. A TTP server is placed between the users and the LBS server to protect the querying user against untrusted LBS. Its main role is to collect and process users' original queries to protect users' sensitive location information by using some privacy-preserving techniques. Through this method, the LBSP cannot distinguish the user's exact location and the user's query information.

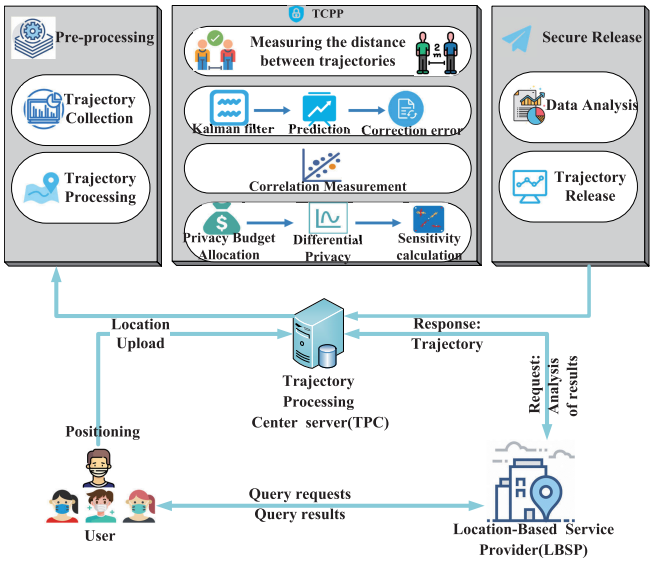



Fig. 1. System model.

Based on the TTP-based model, we propose a mechanism called, TCPP to deal with the trajectory publishing among multiple users. As shown in Fig. 1, the system model of the scheme consists of three entities: Trajectory Processing Center server (TPC), User and Location-Based Service Provider (LBSP). The TCPP of the scheme is described in detail with the example of users enjoying location-based services. TPC provides a trajectory correlation privacy preservation mechanism, which ensures the high availability of published trajectory data as much as possible to provide high-quality location-based services for users. Users collect their trajectories through GPS positioning technology and upload the trajectories to the TPC to request services from LBSP, which obtains the trajectories processed by TPC, and then responds to the user's query to provide high-quality location-based services.

- TPC is a trajectory processing center, which provides three main functions: a) collecting and storing users' trajectory data for pre-processing; b) correlation privacy preservation of pre-processed trajectories; c) providing accurate and efficient trajectory data for LBSP. To be specific, TPC collects location trajectory data from each user for pre-processing, divides the area based on user's predefined sensitive locations by the Voronoi diagram for trajectory processing; Secondly, TPC performs a pre-correlation measure for each trajectory, and if the value is within the threshold then privacy preservation of the correlation of different trajectories using differential privacy is required; Finally, the processed trajectories are published in response to requests from the LBSP.
- As a location service provider, LBSP's function consists of two parts: a) making query request to TPC;

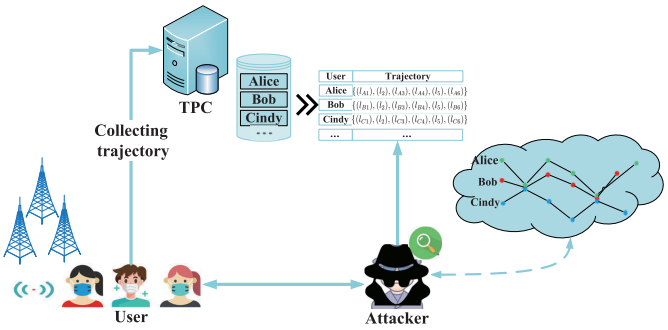



Fig. 2. Threat model.

b) responding to user's request to provide high-quality location services for users. Specifically, after receiving the query request from the user, the LBSP analyzes it and makes a request for query to the TPC, after receiving the feedback result, it stores the trajectory data into its database, actively responds to the user's request to provide high-quality location services.

- User as a consumer of location-based services, whose functions are two parts: a) providing location trajectory data to the TPC; b) making query requests to the LBSP to enjoy location services. Specifically, users collect location trajectory data by GPS positioning technology and upload it to TPC, make query requests to LBSP to obtain its feedback results to meet the demand of location services.

### B. Threat Model

The goal of TCPP is to preserve the trajectory correlation amongst different users and the privacy of users' location trajectory data during its operation, where TPC and User are honest, LBSP is considered to be honest and curious. Briefly, the honest and curious indicates that the participant strictly adheres to the execution of the protocol, but at the same time it may also be curious about the user's sensitive information on which it seeks to analyze and mine the user's private data for its benefit.

The threat model of the scheme is shown in Fig. 2. To conclude, LBSP is considered the most potentially dangerous adversary because it owns the user's location service request as a means to analyze the sensitive information exposed in the user's location trajectory.

### C. Mechanism Architecture

The trajectory correlation privacy preservation mechanism TCPP proposed in this scheme is located in TPC, which consists of six modules: measurement judgment module, prediction module, adjacent datasets construction module, global sensitivity calculation module, privacy budget allocation module, and correlation privacy preservation module, with the detailed flow shown in Fig. 3.

The measurement judgment module uses Euclidean distance to measure the trajectory correlation between users to accurately identify close contacts. Kalman filter skillfully integrates the observation data and estimation data, limits the

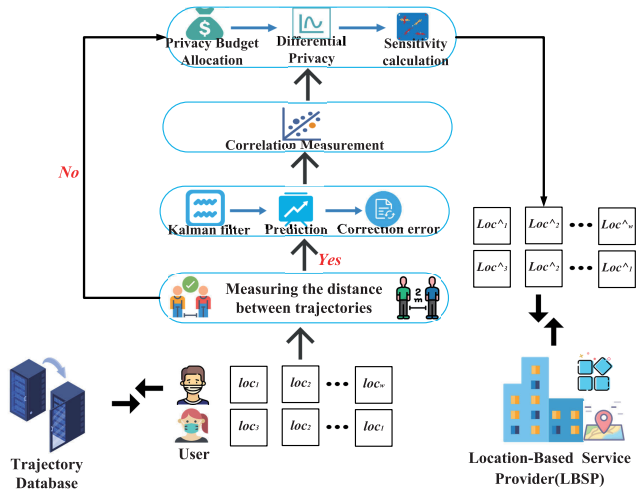



Fig. 3. TCPP Mechanism Architecture.

error to a certain range, and still maintains a stable error for a long time. Therefore the prediction module uses Kalman filter to make dynamic prediction of individuals trajectories and corrects the location points by prediction error when predicting each location point to improve the availability of trajectory data. In order to preserve the correlation amongst trajectories, the adjacency dataset construction module defines a correlation coefficient to filter the set of predicted trajectories with less correlation, and thus constructs adjacency trajectory datasets. The global sensitivity calculation module calculates the global sensitivity of this scheme by analyzing how the traditional differential privacy global sensitivity is calculated. The privacy budget allocation module calculates the privacy level of other locations based on the privacy level of sensitive locations and time factors, etc., and dynamically allocates privacy budgets. The correlation privacy preservation module is actually the process of adding noise to the location, which securely publishes user's trajectory by establishing a trajectory correlation privacy preservation mechanism to provide users with high-quality location services.

### D. Mechanism Algorithm

The algorithm of TCPP is shown in Algorithm 1. The first step is to judge the similarity of trajectories amongst different users, and its similarity value is required for trajectory privacy preservation if it is less than $\theta$. For any trajectory $T$, the algorithm needs to predict and correct for each location to obtain the set of predicted trajectories ($PT$) needed for the scheme. In addition, the trajectories are filtered based on $Sim < \lambda$ to obtain the set of predicted trajectory protection ($PTP$), which is used to construct the adjacent datasets. Afterwards, the global sensitivity of the query function is calculated, personalize the privacy budget by assigning privacy levels based on locations from the trajectory. Eventually, Laplace noise is added to the trajectory to be published based on differential privacy in order to publish the trajectory securely.

## IV. Modules Design

In general, the area in which the user is located can be defined by an irregular polygon that is made up of location

**Algorithm 1** TCPP

**Input:** $T_a, D = \{T_a, T_b, \ldots, T_n\}, \varepsilon$
**Output:** $T_a{}'$

Calculate the Euclidean distance of every $T_j$ trajectory in $D$ from $T_a$;
Get the trajectory set of $D_a$with $dis(T_a, T_j) < \theta$
**if** $D_a \neq NULL$ **then**
  **for** each predicted trajectory $i \in [1, n]$ **do**
    **for** each timestamp $t \in [1, w]$ **do**
      From prediction algorithm to prediction estimate
      Error correction for predicted trajectory values
      Obtain the predicted trajectory $T$
    **end for**
  **end for**
  Get the set of predicted trajectories $PT$
  Construct the set of predicted trajectory preservation $PTP$
  Get adjacent datasets
  Calculate sensitivity $GF$
  Allocate privacy budget $\varepsilon$ dynamically
  **for** each timestamp $t \in [1, w]$ **do**
    $T_a{}' = T_a + Lap\left(b_i = \frac{GF_F}{\varepsilon_i}\right)$
  **end for**
**else**
  Calculate sensitivity $GF$
  Allocate privacy budget $\varepsilon$ dynamically
  **for** each timestamp $t \in [1, w]$ **do**
    $T_a{}' = T_a + Lap\left(b_i = \frac{GF_F}{\varepsilon_i}\right)$
  **end for**
**end if**
**return** $T_a'$

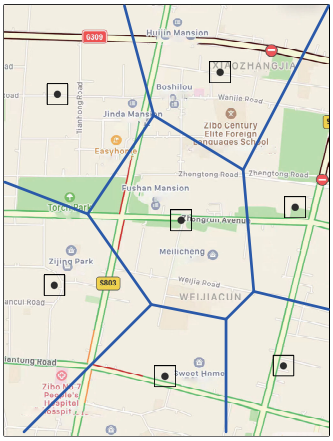



Fig. 4. Map of areas based on self-defined sensitive locations.

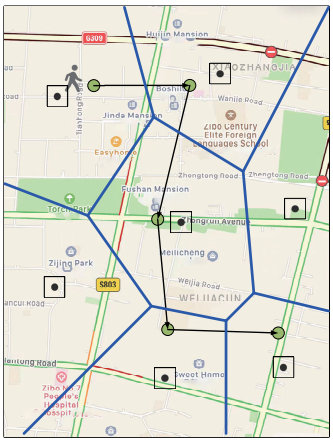



Fig. 5. User trajectory map.

points defined by different sensitivities. Suppose given a map of a city, the Voronoi diagram is constructed according to the sensitive locations provided by the user. Users first predefine themselves sensitive locations and predefine privacy level for each sensitive location, and build a Voronoi diagram of the current area with the sensitive location as the center of the diagram. Since the basic property of Voronoi diagrams is that any location in each cell is at a lower distance from the center of the diagram of the current cell than from the center of the diagram of other cells, which can be used to calculate the privacy level of all locations in the area. In Fig. 4, a city map is overlaid by a Voronoi diagram. Fig. 5 shows the trajectory map in the sensitive area where the user is located.

### *A. Calculate the Euclidean Distance of Both Real Trajectories*

*Definition 7 (Euclidean Distance [19]):* The Euclidean distance requires that the two trajectories have the same length and each location point corresponds to each other. It is defined formally as follows:

$$dis(T_a, T_b) = \frac{1}{w}\sum_{i=1}^{w}\sqrt{Lat_i{}^2 - Lng_i{}^2}, \tag{7}$$

where $Lat_i$ and $Lng_i$ denote the latitude distance and longitude distance between $User_a$ and $User_b$ at interval $i$, respectively, which are calculated as follows:

$$Lat_i = (loc_{ai}.lat - loc_{bi}.lat),$$
$$Lng_i = (loc_{ai}.lng - loc_{bi}.lng),$$

where $loc_{ai}.lat$ and $loc_{ai}.lng$ denote the latitude and longitude of $User_a$'s location, respectively.

If there is a $dis(T_a, T_b) < \theta$ of Euclidean distance between $User_a$ and $User_b$ at different location points, where $\theta$ denotes the distance threshold, it indicates a correlation between their trajectories. The existing trajectory dataset $D = \{T_a, T_b, \ldots, T_n\}$, calculates the Euclidean distance between $T_a$ and each trajectory $T_j$ in $D$, and places trajectories $T_j$ with $dis(T_a, T_j) < \theta$ into the set $D_a$. Therefore, the correlation between trajectories should be preserved to securely publish users' trajectory.

### *B. Predicted Trajectory*

To begin with, the user's trajectory is predicted by using Kalman filter, because of its clever integration of observed data and estimated data, closes the loop management of the error and limits the error to a certain range, it introduces the role that the observed data will correct the estimated data to

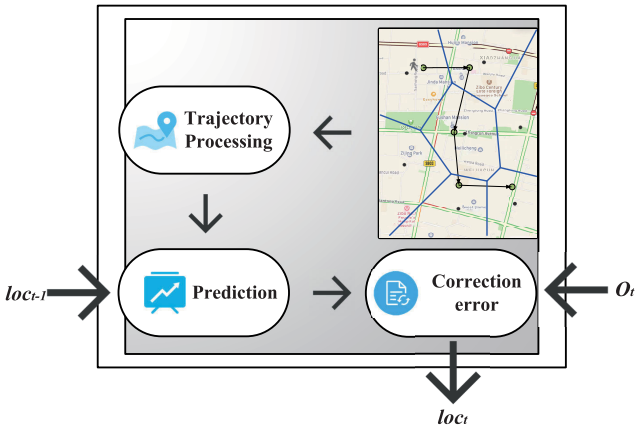



Fig. 6. Prediction Flow.

prevent the error of the estimated data from being so large that it is outrageous, with its use of prediction and correction operation to make the privacy-preserving data closer to the original data. The workflow is shown in Fig. 6.

**Algorithm 2** Prediction Algorithm

**Input:** $T = (loc_1, loc_2, \ldots, loc_w)$
**Output:** $T' = \{loc'_1, loc'_2, \ldots, loc'_n\}$, $Error$
$Data = trajpreprocess(T)$
2: $init()$
$O = getObservationState(Data)$
4: **for** each timestamp $t \in [1, w]$ **do**
$loc_i = Predict(Data)$
6: $error\,[i] = getError(loc_i, loc_i')$
**end for**
8: $Error = \dfrac{\left(\sum_{i=1}^{w} error[i]\right)}{w}$
**return** $T' = \{loc'_1, loc'_2, \ldots, loc'_n\}$

The input of Algorithm 2 input is the user's trajectory data, which is analyzed and corrected, filtered and updated to complete the pre-processing operation. The parameters of the motion model, such as $A$, $Q$ and $R$, are initialized and operated according to the state equation and the observation equation of the system. Predicting the predicted value of the next interval $k$, the covariance array of the estimation error based on the optimal state estimate at interval $k-1$ and the covariance array of the estimation error. In turn, the prediction process is completed based on the observed value at interval $k$ to the optimal state estimate at that interval and the covariance array of the optimal estimation error, and so on. Therefore, it is summarized as predicting the location of the trajectory at that interval based on the optimal state estimate of the preceding interval and the observation of the present interval. Predicted locations for $loc'$ are compared with the real location $loc$, the prediction error is calculated, the operation is repeated a total of $w$ times to complete the prediction of the trajectory points and get the set of predicted trajectory ($PT$), which contains the real trajectory of the user. Finally, the mean value of the error is calculated and output. The set of predicted trajectory is defined as follows:

*Definition 8 (Predicted Trajectory Set):* $User_a$ is in a Kalman filtered environment for motion, with all its possible predictions in $w$ intervals are expressed as:

$$PT = \{PT_1, PT_2, \ldots, PT_n\},$$
$$PT_i = \{loc_{i1}, loc_{i2}, \ldots, loc_{iw}\},$$

where $loc_{iw}$ represents a location in which $User_a$ is at interval $w$ and $PT_i$ represents a possible location trajectory of $User_a$.

### C. Adjacent Trajectory Datasets

In this part, correlation is measured for the predicted trajectories. It is defined as follows:

*Definition 9 (Correlation Coefficient between Trajectories):* Suppose there exists a mean Euclidean distance between the trajectories of $User_a$ and $User_b$ less than $\theta$. The set of predicted trajectories under Kalman filter-based motion are $PT_a$ as well as $PT_b$, respectively. The trajectory correlation amongst them is calculated as follows:

$$Sim_{ab} \equiv \frac{1}{n}\sum_{1}^{n} \frac{\sum_{1}^{w} \frac{Cov(loc_{ai}, loc_{bi})}{\sqrt{Var(loc_{ai})Var(loc_{bi})}}}{w}, \tag{8}$$

where $Cov(loc_{ai}, loc_{bi})$ is the location covariance of the two users at time $i$, $Var(loc_{ai})$ and $Var(loc_{bi})$ denote the location variance of $User_a$ and $User_b$ at time $i$, respectively. $Cov(loc_{ai}, loc_{bi})$ can be calculated as follows:

$$Cov(loc_{ai}, loc_{bi}) = \frac{\sum_{k=1}^{n} \left(loc_{ai}^{k} - loc_{ai}^{R}\right)\left(loc_{bi}^{k} - loc_{bi}^{R}\right)}{n-1},$$

where $loc_{ai}^{k}$ denotes the location of the $k$-th trajectory in $User_a$'s predicted trajectory set $PT_a$ at interval $i$, and $loc_{ai}^{R}$ denotes the real location of $User_a$ at interval $i$.

The correlation value is obtained by Definition 9, with the strength of the correlation degree is determined by the absolute value of the correlation coefficient, whose absolute value is $[0.7, 1.0]$ for strong correlation; $[0.4, 0.7]$ for moderate correlation; and $[0.0, 0.4]$ for weak correlation. This scheme defines the absolute value of the correlation coefficient $\geq \lambda$. The threshold can be freely defined by the actual situation. Assuming $\lambda = 0.4$, when $Sim \geq \lambda$, it indicates that there is a correlation between two predicted trajectories. Therefore, the scheme needs to preserve the trajectory correlation for $Sim \geq \lambda$.

In $PT$, the set of trajectories with absolute values of correlations $Sim < \lambda$ between trajectories is calculated to constitute the set of prediction trajectory preservation ($PTP$), which is formally defined as shown below:

*Definition 10 (Prediction Trajectory Preservation Set):* Suppose the set of predicted trajectories of $User_a$ and $User_b$ are denoted as $PT_a$ and $PT_b$, respectively, the correlation between the trajectories of $User_a$ and $User_b$ is calculated. If $Sim_{ab} < \lambda$, the set of predicted trajectories of both users is retained; otherwise, the trajectories with $Sim_{ab} < \lambda$ need to be filtered from the set of predicted trajectories. Thus, if the

absolute value of correlation $\geq \lambda$, $User_a$ gets the following results:

$$PTP_a = (PTP_{a1}, PTP_{a2}, \ldots, PTP_{aw}) = PT_a - D_{Sim_{ab\geq\lambda}},$$

where $D_{Sim_{ab\geq\lambda}}$ denotes the set of trajectories formed by filtering out the set of predicted trajectory $PT_a$ of $User_a$ from the set of predicted trajectory $PT_b$ of $User_b$ with trajectory correlation $Sim_{ab} \geq \lambda$.

Since the Euclidean distance $dis(T_a, T_j)$ between $User_a$ and each trajectory $T_j$ in the trajectory set $D_a$ is less than $\theta$, the set $PT_a$ needs to filter out the part of the correlation coefficient $Sim \geq \lambda$ with each trajectory $T_j$ to get the following results:

$$PTP_a = (PTP_{a1}, PTP_{a2}, \ldots, PTP_{aw}) = PT_a - D_{\sum_j Sim_{aj\geq\lambda}},$$

where $D_{\sum_j Sim_{aj\geq\lambda}}$ denotes the set of trajectories formed by filtering out the set of predicted trajectory $PT_a$ of $User_a$ from the set of predicted each trajectory $PT_i$ of $D_a$ with trajectory correlation $Sim_{aj} \geq \lambda$.

*Definition 11 (Privacy Prediction Trajectory Preservation Set):* Suppose the real trajectory of $User_a$ is denoted as $RT_a$ and the set of prediction trajectory preservation of $User_a$ in time $w$ is denoted as $PTP_a$, then the set of privacy prediction trajectory preservation is denoted as:

$$PPTP_a = PTP_a - RT_a.$$

The construction of adjacency databases will be performed:

*Definition 12 (Adjacent Trajectory Datasets):* Suppose the set of privacy prediction trajectory preservation of $User_a$ at time $w$ is $PPTP_a$, the set of prediction trajectory preservation is $PTP_a$, which only one record is different between these two trajectory datasets, so $PPTP_a$ and $PTP_a$ are adjacent trajectory datasets.

### D. Sensitivity Calculation

In order to calculate the sensitivity, this scheme analyzes the sensitivity of traditional differential privacy by an example, as defined in Definition 5.

Suppose there is a location trajectory data table Tab, which contains the location trajectory data of multiple users, as well as we make the following query:

$F_1$: Query the total number of "Park" in the Tab;

$F_2$: Query the total number of "Sing" in the Tab.

Suppose $D_1$ and $D_2$ are adjacent databases, where $D_2$ has one less or more users' location trajectory data than $D_1$. If $F(D_2) = [2, 11]^T$, then $F(D_1)$ may be the following result:

$$F(D_1) = \begin{bmatrix} 1 & 2 & 2 & 3 & 2 & 1 & 3 \\ 11 & 10 & 11 & 11 & 12 & 10 & 12 \end{bmatrix},$$

in addition, the difference of the above query function F is calculated as follow:

$$F = F(D_1) - F(D_2) = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & 1 & -1 \end{bmatrix},$$

Therefore, its sensitivity is $GF_F = \max|F| = 2$

By analyzing the above example, it can be concluded that the $\tau$-sensitivity of the query function is related to the number of queries. The sensitivity of the query function of this scheme lies in the fact that the location trajectory data of an individual can change the maximum change range of the query function, so the scheme is based on the set of privacy prediction trajectory preservation $PPTP$ for location trajectory correlation preservation. The analysis shows that the sensitivity of the scheme is related to the number of query functions $F$, which is $m$. Therefore, based on Definition 9, the $\tau$-sensitivity is calculated as follows.

*Definition 13 ($\tau$-Sensitivity):* Suppose there is a function $F$ with a query count of $m$. Then the sensitivity of $F$ is:

$$GF_F = m. \tag{9}$$

### E. Personalized Privacy Budget Allocation

Assume that the set of sensitive locations and their corresponding privacy levels are given by SL=$\{sl_1, sl_2, \ldots, sl_n\}$ and PL= $\{pl_1, pl_2, \ldots, pl_n\}$, where the range of the PL set is [0, 1], the larger the value, the higher the privacy level of the location; and vice versa. The time factor is taken into account when calculating the privacy level of other locations based on the user's sensitive locations. At $t_i$, where $i \in [1, w]$, the user pre-specifies the weight values for sensitive locations, denotes the set of weights for each time period by $Weight = \{weight_{t_1}, weight_{t_2}, \ldots, weight_{t_n}\}$, where the range of $Weight$ set is [0, 1], $weight_{t_i} = \{0.5, 0.6, \ldots, 0.8\}$ denote the time weights of each sensitive location at $t_i$.

The scheme takes the relation between nodes into account, assumes that all locations of the user in a certain area are connected to that map center (sensitive locations), i.e., locations on the map have road connectivity with sensitive locations. Privacy levels are assigned to location points based on the distance to the sensitive locations and the connection relationship. In Fig. 5, suppose the user predefined sensitive location is represented by $p$, its privacy level is $pl_p$. Each location in the sensitive area is represented by the set $rangeSet$, its size is the number of locations in the area. Suppose the starting location of a user's trajectory is represented by $v$, whose privacy level is calculated as shown below:

$$pl_v = pl_p \times weight_{t_i(p)} \times \frac{\frac{1}{dis_{v,p}}}{\sum_{v' \in rangeSet} \frac{1}{dis_{v',p}}}, \tag{10}$$

where $pl_v$ denotes the privacy level assigned to node $v$, $dis_{v,p}$ denotes the distance between node $v$ and node $p$, and $weight_{t_i(p)}$ denotes the weight value of node $p$ at $t_i$. The closer the location $p$ is to the sensitive location, the higher privacy levels are assigned. Based on this idea, the Location Privacy Level (LPL) algorithm is proposed, as shown in Algorithm 3.

In LPL algorithm, the function of threshold parameter $\sigma$ is to have a certain moderating effect, i.e., when the privacy level is greater than or equal to $\sigma$, the area needs to perform privacy preservation. Specifically, when $\sigma = 0$, all locations in the area are to perform privacy preservation; when $\sigma = 1$, only the most sensitive locations perform privacy preservation.

From the definition of differential privacy, the $\varepsilon$ is inversely proportional to the degree of data privacy preservation. The $\varepsilon$ is taken in conjunction with specific requirements to achieve

**Algorithm 3** LPL

**Input:** Initial set of sensitive locations: SL, set of privacy levels corresponding to sensitive locations: PL, privacy level thresholds: $\sigma$

**Output:** User per location privacy level: PL

```
Select the head element in the set SL: p
while Sensitive location p ≠ NULL do
    Get the rangeSet set for sensitive location p
    for each point v ∈ rangeSet do
        Assign privacy level: pl_v = newpl
        if newpl < σ then
            continue
        end if
        pl_v = newpl
    end for
    Select the next element in the set SL: v
end while
return PL
```

a balance between security and availability of the output. Since users have different levels of preservation for different locations, we needs to develop a privacy budget allocation scheme to allocate $\varepsilon$ based on privacy levels.

*Definition 14 (Privacy Level Weight):* Assume that each location in trajectory $T = \{loc_1, loc_2, \ldots, loc_w\}$ is represented by $v_i$, where $i \in [1, w]$, and thus the privacy level weighting of each location point at different intervals is shown below:

$$k = \frac{pl_{v_i}}{\sum_{i=1}^{w} pl_{v_i}}, \tag{11}$$

where $v_i.pl$ is calculated by equation (10), therefore, $k \in [0, 1]$.

All locations in the user's trajectory are assigned different privacy levels according to time. It is clear from the above analysis that locations with higher privacy level require higher privacy preservation level, which in turn need to be assigned smaller $\varepsilon$. Therefore, the privacy budget and privacy level are inversely correlated. We combine privacy level weight $k$ and privacy budget $\varepsilon$ to construct $\delta-$privacy budget allocation.

*Definition 15 ($\delta-$Privacy Budget Allocation):* When a location is to be published, it must satisfy that the privacy level weight $k$ of that location is inversely proportional to $\varepsilon$ of the differential privacy assigned to that location:

$$\varepsilon \times k = \delta. \tag{12}$$

It is clear from the above definition 15 that the greater the weight of the privacy level of the user's location at that interval when a certain time, the smaller the allocated privacy budget, indicating a greater degree of privacy preservation, and vice versa. Calculate the privacy budget for each location on the user's trajectory based on the privacy level weighting. In particular, because the privacy level of user defined sensitive locations changes over time, the assigned privacy budget of the user is not fixed, but adaptively changes with the user's time factor.

### F. Trajectory Publishing Mechanism

Adding noise to the trajectory data based on the calculation of the above results to preserve the correlation amongst trajectories and thus securely publish the trajectory data.

*Definition 16 (Trajectory Publishing Mechanism):* For the user's trajectory $T$, add noise to obtain the trajectory $T'$, which is defined formally as follows:

$$T' = T + noise, \tag{13}$$

where $T$ is the original true trajectory of the user and $noise$ obeys the Laplace distribution [17]:

$$\Pr(noise) = \prod_{i=1}^{w} \frac{e^{\frac{-|loc_i|}{b_i}}}{2b_i}, \tag{14}$$

where $noise$ added at each location point in the user's trajectory $T$ during time $w$ is denoted as $noise = (noise_1, noise_2, \ldots, noise_w)$; the set of Laplace noise scales is $b = (b_1, b_2, \ldots, b_w)$, where $b_i = \frac{GF_F}{\varepsilon_i}$, which is determined by the sensitivity $GF$ and privacy budget $\varepsilon$.

## V. Analysis

In this phase, the security and availability of the scheme and the background knowledge of the adversary will be analyzed in detail, in which the process is as follows.

### A. Security Analysis

*Theorem 5.1:* Given a privacy budget $\varepsilon$, the TCPP algorithm is satisfying $\varepsilon$-differential privacy in the privacy level assignment phase in time $t$.

*Proof:* In accordance with a definition of differential privacy, it is clear that only one tuple or one individual differs between adjacent datasets. There is only one record difference between the output of the scrambled trajectory $OT$ and the output of the true trajectory $RT$, the published trajectory is $OT$. The posterior probability of determining the true trajectory based on the published trajectory is $\Pr(RT|PT)$, the formal proof is shown below:

If $OT$ is generated from the dataset $PPTP$, it is obvious from Theorem 5.2 that $PT$ is satisfying differential privacy, so the following equation holds.

$$\frac{\Pr(RT|OT)}{\Pr(RT)} \leq e^{\varepsilon}. \tag{15}$$

If $PT$ is not generated from the dataset $PPTP$, there exists $T_k \in PPTP$ such that the following equation holds.

$$\frac{\Pr(RT|OT)}{\Pr(RT)} = \frac{\sum_k \Pr(RT|T_k)\Pr(T_k|OT)}{\sum_k \Pr(RT|T_k)\Pr(T_k)} \leq e^{\varepsilon}. \tag{16}$$

It is known from differential privacy that TCPP satisfies $\varepsilon$-differential privacy in the privacy level assignment phase. □

*Theorem 5.2:* Given a privacy budget $\varepsilon$, the TCPP algorithm is satisfying $\varepsilon$-differential privacy in the add Laplace noise phase in time $t$. Assume that function $F$ has sensitivity

$GF_F$, $M$ is an algorithm that adds independent noise to the output of function $F$, in which the algorithm $M$ satisfies $\varepsilon$-differential privacy if the noise obeys the Laplace distribution of $\frac{GF_F}{\varepsilon}$.

*Proof:* From the definition of differential privacy, in TCPP algorithm, using the knowledge of conditional probability, suppose the output trajectory is $OT$ based on the adjacent datasets $PTP$ and $PPTP$. For ease of presentation, $PTP$ and $PPTP$ are denoted by $D_1$ and $D_2$, respectively. The process is as follows:

$$\begin{aligned}\frac{\Pr[M(D_1)=OT]}{\Pr[M(D_2)=OT]} &= \prod_{i=1}^{w} \frac{e^{-\frac{\varepsilon(F(D_1)_i - OT_i)}{GF_F}}}{e^{-\frac{\varepsilon(F(D_2)_i - OT_i)}{GF_F}}} \\ &= \prod_{i=1}^{w} e^{\frac{\varepsilon[(F(D_2)_i - OT_i)-(F(D_1)_i - OT_i)]}{GF_F}} \\ &\le \prod_{i=1}^{w} e^{\frac{\varepsilon(F(D_2)_i - F(D_1)_i)}{GF_F}} \\ &= e^{\frac{\varepsilon\|F(D_2)-F(D_1)\|_1}{GF_F}} \le e^{\varepsilon},\end{aligned}$$

it is known from differential privacy that TCPP satisfies $\varepsilon$-differential privacy in the add Laplace noise phase. □

For the correlation preservation amongst trajectories, this scheme needs to be analyzed from two aspects: 1)the personal trajectory privacy; 2)the correlation preservation of trajectories amongst different users.

At first, for the personal trajectory privacy is analyzed, assuming that the real trajectory of $User_a$ is $RT_a$, and the real trajectory of $User_b$ is $RT_b$, $OT$ generated from the dataset $PPTP$. The Kalman filter is employed to predict the user's trajectory while processing the real trajectory. Eventually, the real trajectories of users are not published, so the privacy of personal trajectories is guaranteed. Its formal analysis is as follows:

$$e^{\varepsilon} \le \frac{\Pr(T_a|OT)/\Pr(T_a|OT)}{\Pr(T_a|RT)/\Pr(T_b|RT)} \le e^{\varepsilon}, \tag{17}$$

it follows that the adversary cannot distinguish between $T_a$ as well as $T_b$, therefore, the scheme preserves the privacy of individual trajectories.

Secondly, it can be obtained $RT$ from the above analysis that the real location trajectory of user is hidden, and the trajectory correlation amongst different users, is also hidden, hence the trajectory correlation amongst different users is preserved.

### B. Availability Analysis

For the availability of perturbed trajectory, the analysis is performed according to Definition 6. The data availability is influenced by two factors: 1)the distance between the perturbed location and the real location, the larger the distance, the larger the value of $U$, the lower the trajectory data availability; 2)the length of the trajectory, the longer the trajectory, the more locations need to be considered, if the goal is to achieve higher accuracy, the corresponding interval division is also more, causing the lower data availability.

In the privacy-preserving mechanism proposed in this paper, the probability ratio of the Euclidean distance between the true location and the published location is $e^{\varepsilon}$, from which $\varepsilon$ can be controlled to adjust the magnitude of the probability ratio as a way to solve the Euclidean distance between both the true location and the published location. Therefore, this scheme has some advantages in terms of trajectory data availability.

For the availability of statistical analysis, assume that there exists statistical analysis $S$. The probability of $S$ when the perturbed trajectory is published is $\Pr(S|PT)$, the probability of $S$ when the true trajectory is released is $\Pr(S|RT)$. The difference between the perturbed and true trajectory is only one record, therefore, the ratio between them is:

$$\frac{\Pr(S|PT)}{\Pr(S|RT)} \le e^{\varepsilon}, \tag{18}$$

the above analysis shows that when $\varepsilon$ is small, $\Pr(S|PT)$ and $\Pr(S|RT)$ are almost the same, so the availability of the scheme proposed TCPP mechanism for statistical analysis remains steady state.

### C. Adversary's Background Knowledge

The purpose of this scheme is to make the adversary's a priori knowledge and a posteriori knowledge essentially identical to prevent the adversary from obtaining additional information

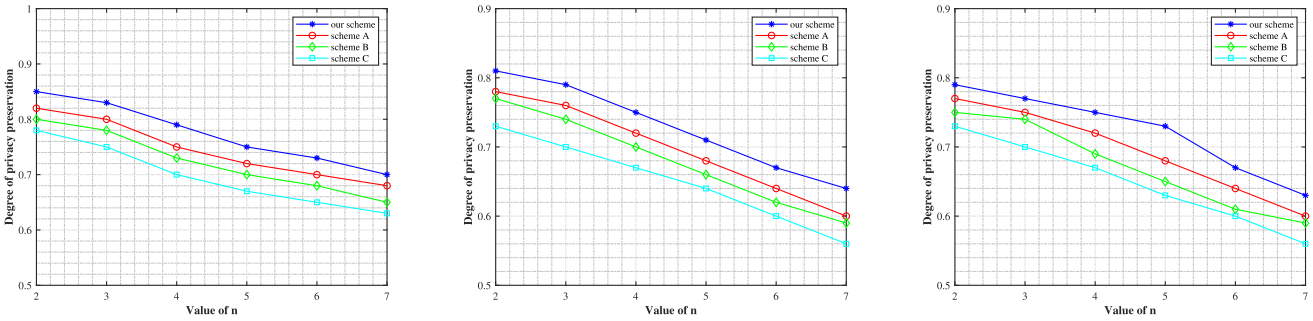
to obtain efficient privacy preservation for users. Therefore, this scheme needs to restrict the impact of the output of the perturbed trajectory publish on the prior knowledge of the adversary to a larger impact for poorer privacy preservation and a smaller impact for better privacy preservation. Suppose the user's trajectory dataset is $PTP$, its true trajectory is $RT \in PTP$, and the published scrambled trajectory is $PT \in PPTP$, so $\Pr(RT)$ is taken as the prior knowledge probability of the adversary, $\Pr(RT|PT)$ as the posterior knowledge probability of the adversary. As shown below:

$$\begin{aligned}&\frac{\Pr(RT_1|PT)/\Pr(RT_2|PT)}{\Pr(RT_1)/\Pr(RT_2)} \\ &\quad\le \frac{\Pr(RT_1|PT)/\Pr(RT_2|PT)}{\Pr(RT_1|PTP)/\Pr(RT_2|PTP)} \\ &\quad\le e^{\varepsilon},\end{aligned} \tag{19}$$

amongst them, $RT_1$, $RT_2$ belong to the trajectory dataset.

## VI. Experimental Evaluation

The simulation was performed using python 3.9 on a laptop with an 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40 GHz 2.42 GHz, 16.0 GB RAM, and Windows 10, and using three datasets, including Geolife [20], ShangHai [21], and Gowalla [22]. The above three different datasets differ in the size of the data and the size of the map area. To better represent the performance of this scheme, we simulate and analyze this scheme with Li et al. [23], Ghane et al. [24] and Gursoy et al. [25], which are named as Scheme A, Scheme B, and Scheme C in order. Where the experimental results are taken as the average of the results of five experiments.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 7. The effect of $n$ with PL=0.6.

TABLE III
MECHANISM'S COMPARISON

| Scheme | Data prediction and correction | $\varepsilon$ automaticlly allocation | Correlation preservation |
|---|---|---|---|
| Scheme A [23] | ∨ | ∨ | × |
| Scheme B [24] | × | ∨ | × |
| Scheme C [25] | × | √ | × |
| Scheme D [15] | ∨ | × | ∨ |
| Scheme E [16] | × | × | ∨ |
| Our scheme | √ | √ | √ |

**Mechanisms' Comparison.** We briefly present the comparison between the comparative scheme and our scheme in table III, where the symbol ∨, the symbol ×, and the symbol √ indicate semi-conformity, nonconformity, and conformity, respectively. Scheme A proposed a differential privacy location trajectory privacy preservation scheme based on a Markov model, but did not correct the error of the predicted data to improve data availability. Scheme B proposed a trajectory generation mechanism (TGM), which encodes trajectories as graphical generations and generates trajectories of any length. However, the correlation between location trajectory is not considered. Scheme C proposed a practical perceptual location trajectory synthesizer AdaTrace, which generates synthesized trajectories with differential privacy protection. But the correlation between location trajectories was not considered. Scheme D proposed a location correlation preservation scheme based on hidden Markov models and differential privacy, Scheme E presents an efficient and optimal location trajectory correlation protection scheme based on differential privacy. LM method is used to constrain optimization to solve the privacy problem of correlation between trajectories. Scheme D and Scheme E only consider the location dependency between two users in terms of their correlation preservation, not the case of multiple users. The above schemes consider the dynamic allocation of the privacy budget and suppresses the publication of noisy data, but does not fully consider the personalized needs of users, thus reducing the availability of data.

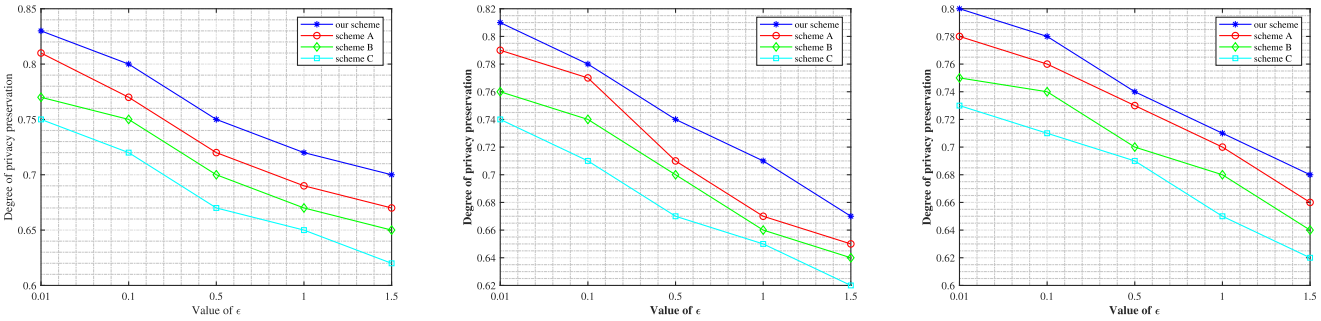
### *A. Degree of Privacy Preservation*

The degree of privacy preservation in the manuscript is regulated and controlled according to the privacy budget $\varepsilon$ of differential privacy, and the distance between the real location and the published location is obtained from the probability ratio as $e^{\varepsilon}$. To facilitate the measurement of the degree of privacy preservation, the specific result is restricted to between 0 and 1, so its specific formula is $PD = e^{-\varepsilon}$. For the degree of privacy preservation, we will analyze the different schemes in terms of three factors: the number of sensitive locations of users $n$, the size of privacy budget $\varepsilon$, and the size of privacy level PL of locations.

At first, the scheme considers the impact of the number of sensitive locations of users on the privacy preservation level of different schemes based on different datasets, assuming PL=0.6, and from Fig. 7a, as the number of sensitive locations of users increases, the required $\varepsilon$ gradually increases and the degree of privacy preservation decreases. The scheme exhibits better performance because its privacy budget's are executed on demand. Fig. 7b shows similar results to Fig. 7a, but the degree of privacy preservation is reduced because the semantic set of Geolife dataset is slightly larger than that of the ShangHai dataset. Similarly, for Fig. 7c, the semantic set of the ShangHai dataset is slightly larger than that of Gollawa dataset.

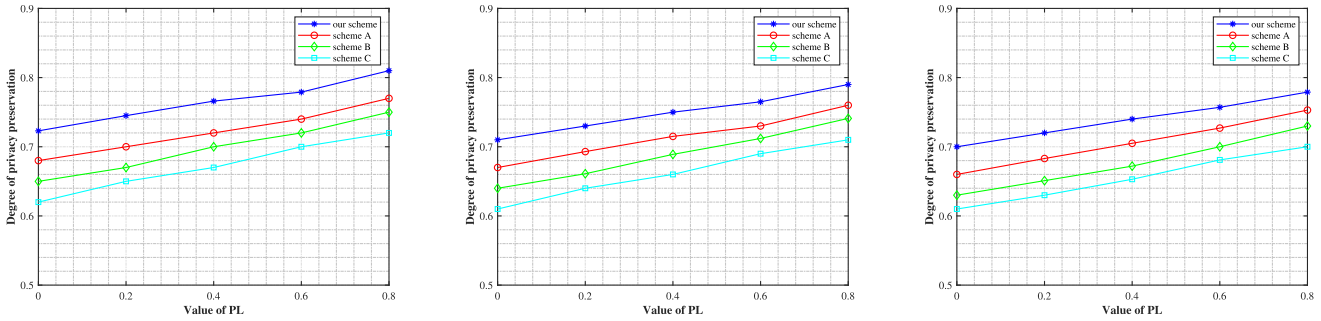
Secondly, in Fig. 8, the scheme considers the impact of the size of the user's $\varepsilon$ on the PL of different schemes on different datasets. For the purpose of analysis, assuming n=6 and PL=0.6. In Fig. 8a, as the $\varepsilon$ increases, it is clear from the definition of differential privacy that the degree of privacy preservation decreases with a larger $\varepsilon$. This scheme shows better performance because it has an on-demand $\varepsilon$ that takes into account the user's time, distance and other factors. For Fig. 8b and Fig. 8c, the results are similar to Fig. 8a. Since the semantic set of Geolife dataset is slightly larger than that of ShangHai dataset, which in turn is slightly larger than that of Gollawa dataset, the degree of privacy preservation is reduced.

In Fig. 9, the scheme considers the impact of the user's location privacy level PL on the privacy preservation of different schemes under different datasets. The number of sensitive locations of the user is assumed to be n=6. In Fig. 9a, the degree of privacy preservation decreases as the user location PL increases, which is because the definition of differential privacy shows that as PL increases, the required $\varepsilon$ increases and the degree of privacy preservation decreases. The same is true for Fig. 9b and Fig. 9c, but the degree of privacy



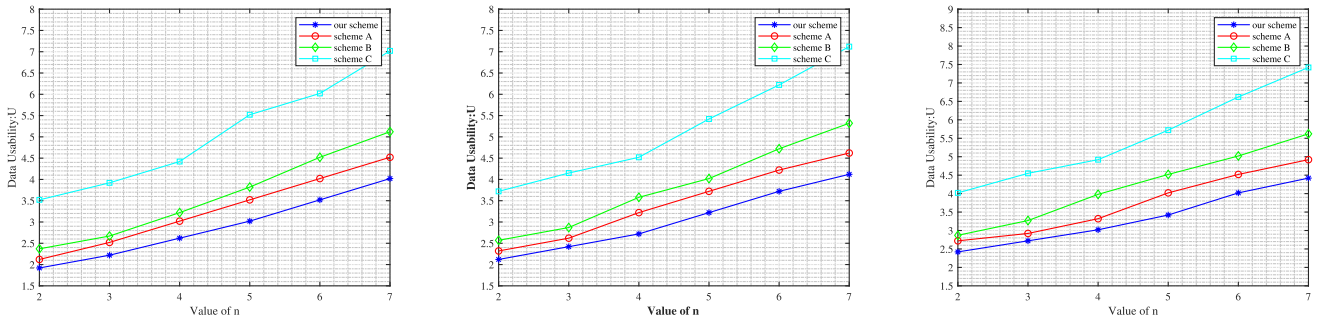

(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 8. The effect of $\varepsilon$ with $n$=6 and PL=0.6.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 9. The effect of PL with $n$=6.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 10. The effect of $n$ with PL=0.6.

preservation is reduced due to the fact that the area of the delimited region in Geolife dataset is more detailed than that in ShangHai dataset, which in turn is more detailed than that of Gollawa dataset.

### B. Data Availability

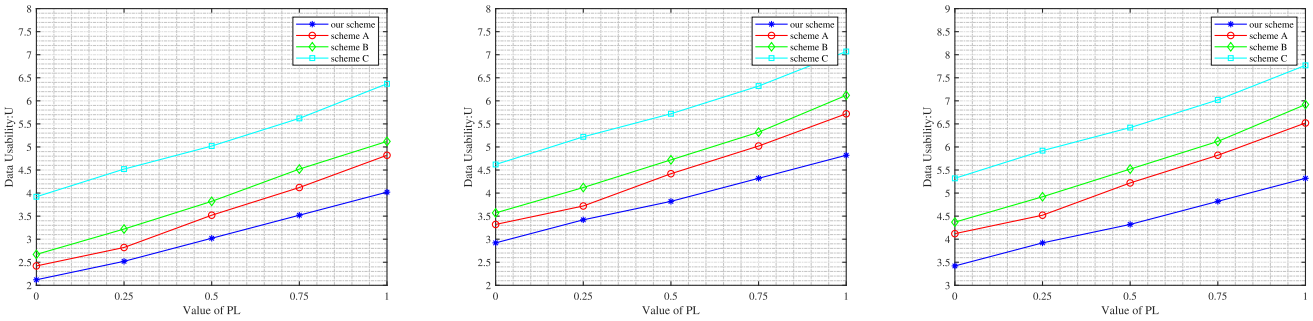
For the data availability, we will analyze the different schemes from two factors: the number of sensitive locations of users $n$, and the size of privacy level PL of locations.

In Fig. 10, for analysis purposes, the user's location privacy level PL = 0.6 is assumed. In Fig. 10a, as the $n$ increases, the required $\varepsilon$ also gradually increases, adding noise to the data and resulting in lower availability of data. According to Definition 6, the metric $U$ for assessing data availability increases as the number of sensitive locations increases. This scheme shows better performance and scheme C shows worse performance. Since the semantic set of Geolife dataset is slightly larger than that of ShangHai dataset, the metric $U$ is increased in Fig. 10b. The results in Fig. 10c are likewise.

Next, in Fig. 11, the scheme considers the impact of the user's location PL on the data availability of different schemes on different datasets, the number of sensitive locations of the user is assumed to be $n = 6$. In Fig. 11a, as the PL increases, the noise to be added increases and the data availability becomes worse. From Definition 6, the metric $U$ increases with the increase of PL. This scheme shows better performance in this aspect. Since the partition area in geolife dataset is more detailed than that in Shanghai dataset, the measure $u$ in Fig. 11b increases. So do the results shown in Fig. 11c.

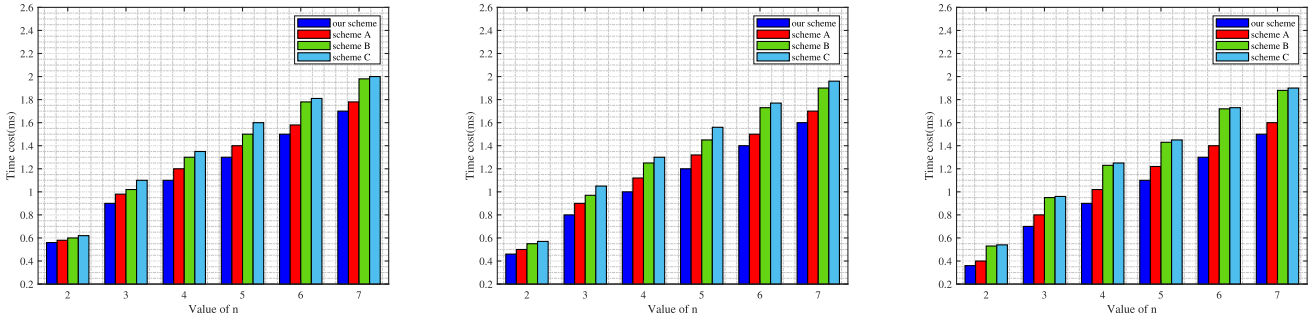
### C. Running Time

For the runtime, the different schemes are analyzed from two factors: the number of sensitive locations of users $n$, and the size of privacy level PL of locations.



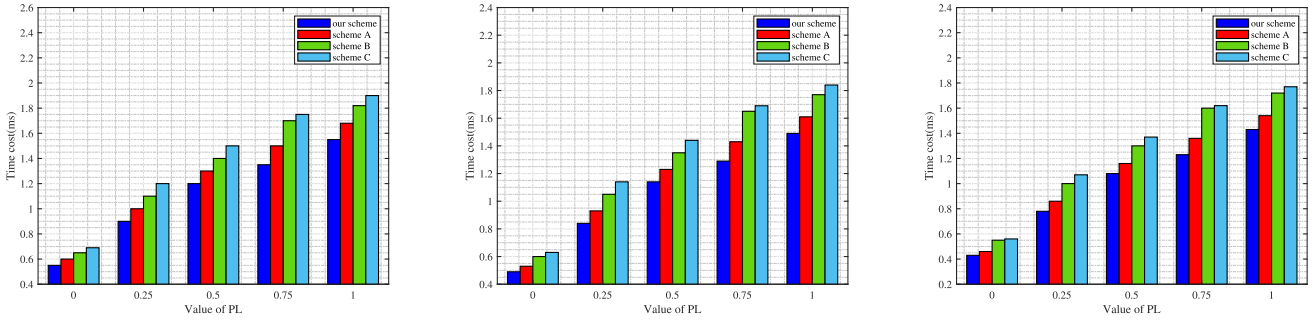

(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 11. The effect of PL with $n$=6.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 12. The effect of $n$ with PL=0.6.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 13. The effect of PL with n=6.

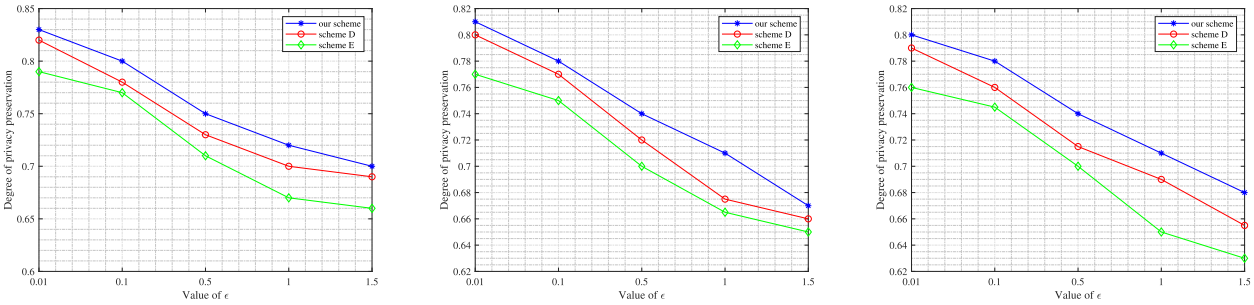
To facilitate the analysis of the impact of the number of sensitive locations n on the running time of the scheme, it is assumed that the PL = 0.6. In Fig. 12a, as the $n$ increases, the running time of the scheme increases. Since the number of geographic locations traversed by the scheme increases as $n$ increases, the time cost to be consumed increases as well. As the scheme requires only prediction and noise addition, shows better performance. The comparison results in Fig. 12b are similar to Fig. 12, but the running time gradually decreases, as does Fig. 12c. As the semantic set of Geolife dataset is slightly larger than that of ShangHai dataset, which is slightly larger than that of Gollawa dataset.

In Fig. 13, to facilitate the analysis of the impact of PL on the running time of the scheme, the number of sensitive locations of user is assumed to be $n = 6$. As shown in Fig. 13a, as the PL increases, more noise needs to be added and more time needs to be consumed. The scheme shows superior performance. The results of comparing Fig. 13b with Fig. 13c are similar to Fig. 13a, but the running time is relatively reduced as the delimited area of Geolife dataset is more detailed than that of ShangHai dataset, which is more detailed than that of Gollawa dataset.

### *D. Simulation of Correlation Between Two Trajectories*

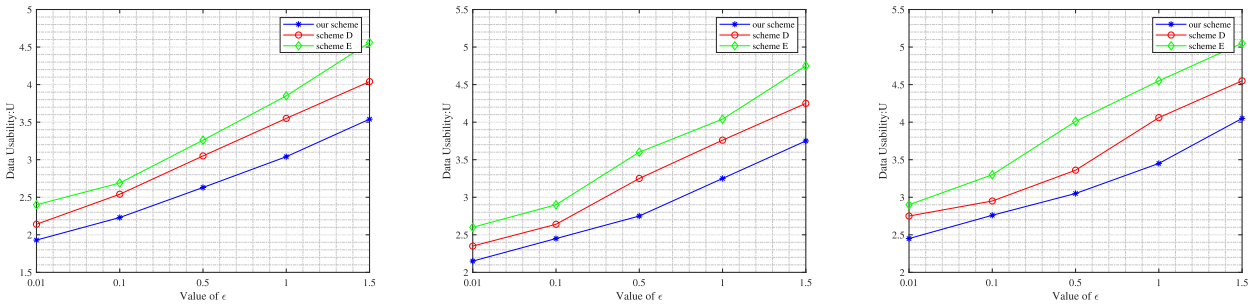
To better represent the performance of this scheme, the algorithm of this scheme will be simulated and analyzed with the algorithms of the literature [15], [16] in terms of privacy preservation and data availability, starting from the aspect of privacy preservation of trajectory correlation between two different users. For the sake of description, the above two schemes are named as Scheme D, Scheme E in order.

*1) Degree of Privacy Preservation:* For the degree of privacy preservation of trajectory correlation between two different users, this scheme will be analyzed and compared




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 14. The effect of $\varepsilon$.




(a) Geolife. (b) ShangHai. (c) Gowalla.

Fig. 15. The effect of $\varepsilon$.

with schemes D and E in terms of the size of the privacy budget $\varepsilon$. This scheme considers the effect of the size of the user's privacy budget $\varepsilon$ on the degree of privacy preservation of different schemes under different datasets, and the results are shown in Fig. 14. As the privacy budget $\varepsilon$ increases, it is clear from the definition of differential privacy that the larger the privacy budget, the lower the degree of privacy preservation $PD$. The scheme shows better performance because of its on-demand privacy budget, which takes into account the user's time, distance and other factors. The comparison results in Fig. 14a, Fig. 14b and Fig. 14c are similar, the reason for the slight difference is that the semantic sets of Geolife, ShangHai and Gollawa datasets gradually decrease.

*2) Data Availability:* The data availability for the trajectory correlation between two different users will be analyzed in terms of the size of the privacy budget $\varepsilon$. This scheme will be compared with schemes D and E. This scheme considers the impact of the size of the user's privacy budget $\varepsilon$ on the data availability of different schemes with different datasets, and the results are shown in Fig. 15. The ratio of the data availability parameter $U$ increases with the coefficient $\varepsilon$. Therefore, data availability decreases with the coefficient $\varepsilon$. This scheme makes the distance between the real location and the published location obtainable by the probability ratio as $e^{\varepsilon}$. The privacy budget is reduced by controlling the privacy budget, and the privacy budget is implemented for personalized allocation, and by inference, the scheme is optimal in terms of data availability.

## VII. Related Work

A detailed description of the research related to privacy preservation of trajectory correlation is as follows.

### A. No Correlation Within a Single Trajectory

Currently, the main approaches to trajectory privacy preservation include cryptography, $k$-anonymity, and differential privacy [6], some valuable works of research were achieved. In 2019, a privacy-preserving system based on game theory was proposed by Xiong et al. [26], a privacy-based personalization metric algorithm is used to establish a reasonable uploading strategy using game theory, and a privacy-based data aggregation technique is given. In 2020, Zhang et al. [27] proposed a double-$k$ mechanism (DKM) to preserve the user's continuous location privacy, which sends the individual query locations to different anonymizes, as trajectories contain sensitive information, adversaries are able to analyze them based on background knowledge for profit, none of the above works can provide strict privacy preservation for them.

Hence, differential privacy [28] was established. In 2018, Wu et al. [29] proposed a new differential privacy-based approach for trajectory preservation. This paper proposes a method to evaluate the privacy level of each location using geospatial correlation to find the exact privacy budget of each location. In 2020, Zhao et al. [30] introduced a method to preserve personal trajectory using prefix trees. The paper proposes a parameter minimum description length (PDML) algorithm based on the idea of minimum description length (DML), which applies differential privacy to preserve trajectories, constrains the noise by Markov chains to improve the data availability. However, the above related works don't consider the privacy leakage issues with correlation within a single trajectory.

### B. Correlation Within a Single Trajectory

The single-trajectory intra-correlation method takes into account that such spatio-temporal correlations are becoming

increasingly common in researches [10], [31], [32]. In early 2012, Chen et al. [33] solved the problem of applying differential privacy to trajectory data by using a model with variable arbitrary length n-grams, which were optimized using search trees and Markov assumptions in order to optimize the noise added. To publish more accurate trajectory data, Wang and Sinnott [32] constructed a privacy reference system for the original discrete trajectories based on clustered anchors on the X-order Markov assumption to publish noisy calibrated trajectories by using differential privacy prefix trees. To effectively preserve the trajectory release against various types of inference attacks, Gursoy et al. [25] in 2018 proposed a different privacy trajectory synthesis method that provides strong privacy preservation for trajectory release. Cao et al. [34] proposed an enhanced version of the differential privacy trajectory privacy preservation method to prevent temporal privacy leakage of trajectories by analyzing the shortcomings of traditional differential privacy under temporal correlation. Wang et al. [35] researched the problem of securely publishing trajectories in a crowdsourcing scenario and designed a RescueDP mechanism based on differential privacy, which accurately predicts the location at each interval, adaptively allocates the privacy budget required, preserves the temporal correlation within the trajectory, and securely publishes the trajectory in real-time. However, these methods only consider spatio-temporal correlation within a single trajectory, ignore correlation between different trajectories that may lead to serious privacy leakage.

### C. Correlation Between Two Trajectories

In order to achieve the preservation of trajectory correlation between different users, Ou et al. [15] presented a differential privacy-based trajectory publishing mechanism to preserve location correlation amongst multiple users, which uses the Hidden Markov to construct candidate sets and uses similarity to measure the correlation amongst trajectories. In 2018, an n-body Laplace framework based on differential privacy was proposed in the literature [16] to preserve trajectory correlation. And a way to measure trajectory correlation is designed as a way to preserve social relationships amongst different users. To improve the availability of trajectory data, the system defines two important tools based on Lagrange multipliers as a way to obtain the optimal Laplace noise. Unfortunately, the above researches are limited to correlation privacy preservation between two users only, and do not apply to privacy preservation amongst multiple users. Currently, research on the privacy preservation of correlations amongst multiple trajectories is not yet well researched. Since the leakage of social relationships can have a huge impact on users, this is an urgent problem.

## VIII. Conclusion

In this paper, a trajectory correlation privacy preservation mechanism (TCPP) satisfying differential privacy was proposed to solve the trajectory correlation privacy leakage problem. First, the similarity of different trajectories was premeasured with Euclidean distance, which is to determine whether correlation preservation is needed. Second, a prediction trajectory algorithm was designed based on Kalman filtering to construct high accuracy and low error datasets to improve the availability of data. Then, a personalized privacy budget allocation strategy was presented. Finally, Laplace noise was added to the trajectory data to preserve the correlation amongst trajectories and ensured that the trajectory data are securely released to trajectories. Evaluations on real-world datasets demonstrated the superiority of the method in terms of privacy, availability as well as time efficiency.

Due to the difficulty of automatic calculation of query function sensitivity caused by the variety of query functions. The direction of future research will concentrate on establishing an intelligent differential privacy preservation framework.

## References


[1] V. K. Yadav, N. Andola, S. Verma, and S. Venkatesan, "P2LBS: Privacy provisioning in location-based services," *IEEE Trans. Services Comput.*, vol. 16, no. 1, pp. 466–477, Jan. 2023.

[2] P. Zhao et al., "Synthesizing privacy preserving traces: Enhancing plausibility with social networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 6, pp. 2391–2404, Dec. 2019.

[3] M. Gramaglia, M. Fiore, A. Furno, and R. Stanica, "GLOVE: Towards privacy-preserving publishing of record-level-truthful mobile phone trajectories," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 3, pp. 1–36, Aug. 2021.

[4] L. Hou, N. Yao, Z. Lu, F. Zhan, and Z. Liu, "Tracking based mix-zone location privacy evaluation in VANET," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10957–10969, Oct. 2021.

[5] J. Hua, Y. Gao, and S. Zhong, "Differentially private publication of general time-serial trajectory data," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 549–557.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2006, pp. 265–284.

[7] L. Fan, L. Xiong, and V. Sunderam, "Differentially private multi-dimensional time series release for traffic monitoring," in *Data and Applications Security and Privacy* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2013, pp. 33–48.

[8] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private data release via Bayesian networks," *ACM Trans. Database Syst.*, vol. 42, no. 4, pp. 1–41, Oct. 2017.

[9] L. Ou, Z. Qin, S. Liao, H. Yin, and X. Jia, "An optimal pufferfish privacy mechanism for temporally correlated trajectories," *IEEE Access*, vol. 6, pp. 37150–37165, 2018.

[10] H. Wang and Z. Xu, "CTS-DP: Publishing correlated time-series data via differential privacy," *Knowl.-Based Syst.*, vol. 122, pp. 167–179, Apr. 2017.

[11] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. ACM Int. Conf. Manage. Data*, May 2017, pp. 1291–1306.

[12] L. Ou, Z. Qin, S. Liao, J. Weng, and X. Jia, "An optimal noise mechanism for cross-correlated IoT data releasing," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 4, pp. 1528–1540, Jul. 2021.

[13] C. Niu, Z. Zheng, S. Tang, X. Gao, and F. Wu, "Making big money from small sensors: Trading time-series data under pufferfish privacy," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 568–576.

[14] D. Kifer and A. Machanavajjhala, "PufferFish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 1–36, Jan. 2014.

[15] L. Ou, Z. Qin, Y. Liu, H. Yin, Y. Hu, and H. Chen, "Multi-user location correlation protection with differential privacy," in *Proc. IEEE 22nd Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2016, pp. 422–429.

[16] L. Ou, Z. Qin, S. Liao, Y. Hong, and X. Jia, "Releasing correlated trajectories: Towards high utility and optimal differential privacy," *IEEE Trans. Dependable Secure Comput.*, vol. 17, no. 5, pp. 1109–1123, Sep. 2020.

[17] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1619–1638, Aug. 2017.
[18] Z. Ma, T. Zhang, X. Liu, X. Li, and K. Ren, "Real-time privacy-preserving data release over vehicle trajectory," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8091–8102, Aug. 2019.
[19] S. P. Patel and S. H. Upadhyay, "Euclidean distance based feature ranking and subset selection for bearing fault diagnosis," *Expert Syst. Appl.*, vol. 154, Sep. 2020, Art. no. 113400.
[20] Y. Zheng, X. Xie, and W.-Y. Ma, "GeoLife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, Jun. 2010.
[21] Hong Kong Univ. Sci. Technol., Smart City Res. Group, Shanghai, China, Feb. 2007.
[22] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 1082–1090.
[23] H. Li, Y. Wang, F. Guo, J. Wang, B. Wang, and C. Wu, "Differential privacy location protection method based on the Markov model," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–12, Jun. 2021.
[24] S. Ghane, L. Kulik, and K. Ramamohanarao, "TGM: A generative mechanism for publishing trajectories with differential privacy," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2611–2621, Apr. 2020.
[25] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, "Utility-aware synthesis of differentially private and attack-resilient location traces," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 196–211.
[26] J. Xiong et al., "A personalized privacy protection framework for mobile crowdsensing in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4231–4241, Jun. 2020.
[27] S. Zhang, X. Mao, K.-K.-R. Choo, T. Peng, and G. Wang, "A trajectory privacy-preserving scheme based on a dual-K mechanism for continuous location-based services," *Inf. Sci.*, vol. 527, pp. 406–419, Jul. 2020.
[28] S. Cai, X. Lyu, X. Li, D. Ban, and T. Zeng, "A trajectory released scheme for the based on differential privacy," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16534–16547, Sep. 2022.
[29] Y. Wu et al., "Differentially private trajectory protection based on spatial and temporal correlation," *Chin. J. Comput.*, vol. 41, no. 2, pp. 309–322, 2018.
[30] X. Zhao, D. Pi, and J. Chen, "Novel trajectory privacy-preserving method based on prefix tree using differential privacy," *Knowl.-Based Syst.*, vol. 198, Jun. 2020, Art. no. 105940.
[31] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "RescueDP: Real-time spatio-temporal crowd-sourced data publishing with differential privacy," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.
[32] S. Wang and R. O. Sinnott, "Protecting personal trajectories of social media users through differential privacy," *Comput. Secur.*, vol. 67, pp. 142–163, Jun. 2017.
[33] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *Proc. ACM Conf. Comput. Commun. Secur.*, Oct. 2012, pp. 638–649.
[34] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy in continuous data release under temporal correlations," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1281–1295, Jul. 2019.
[35] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 591–606, Jul. 2018.

**Lei Wu** received the Ph.D. degree in applied mathematics from the 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.

**Chengyi Qin** is currently pursuing the degree in information science and engineering with Shandong Normal University. Her research interests include privacy preservation and differential privacy.

**Zihui Xu** is currently pursuing the degree in information science and engineering with Shandong Normal University. His research interests include privacy preservation and fog computing.

**Yunguo Guan** is currently pursuing the Ph.D. degree with the Faculty of Computer Science, University of New Brunswick, Canada. His research interests include applied cryptography and game theory.

**Rongxing Lu** (Fellow, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Canada, in 2012. He is currently a Mastercard IoT Research Chair, a University Research Scholar, and an Associate Professor with the Faculty of Computer Science (FCS), University of New Brunswick (UNB), Canada. Before that, he was an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, from April 2013 to August 2016. He was a Post-Doctoral Fellow with the University of Waterloo from May 2012 to April 2013. His research interests include applied cryptography, privacy enhancing technologies, and the IoT-big data security and privacy. He has published extensively in his areas of expertise. He was a recipient of nine best (student) paper awards from some reputable journals and conferences. He received the most prestigious Governor Generas Gold Medal for the Ph.D. degree. He received the 8th IEEE Communications Society (ComSoc) Asia Pacific (AP) Outstanding Young Researcher Award in 2013. He is the Winner of the 2016–2017 Excellence in Teaching Award, FCS, UNB. He serves as the Chair for the IEEE ComSoc Communications and Information Security Technical Committee (CIS-TC). He serves as the Founding Co-Chair for the IEEE TEMS Blockchain and Distributed Ledgers Technologies Technical Committee (BDLT-TC).