Hindawi
Journal of Advanced Transportation
Volume 2024, Article ID 5374764, 14 pages
https://doi.org/10.1155/2024/5374764

Research Article

An Efficient and Differential Privacy-Based Scheme for Aggregating Mobility Datasets

Qing Yang, Fujun Ji, and Fei Liu

School of Management and Engineering, Capital University of Economics and Business, Beijing, China

Correspondence should be addressed to Fujun Ji; jfj@cueb.edu.cn

Received 9 February 2023; Revised 30 January 2024; Accepted 22 February 2024; Published 8 March 2024

Academic Editor: Rui Jiang

Mobile smart devices, such as mobile phones, wearable devices, and in-vehicle navigation systems, bring us convenience and have become necessities in modern daily life. The built-in global positioning system (GPS) of these mobile devices collects the users' mobility data to support path planning, navigation and other location-related applications, which also inevitably causes privacy issues. Previous research has shown that employing count-min sketch (CMS) to aggregate mobility datasets is a valid privacy-preserving method for resisting the reconstruction attack on population distributions. However, as the utility/accessibility of the protected datasets is excessively correlated with the size of CMS, decreasing the data transmission cost has become an unsolved issue of that approach. In this paper, we propose an efficient scheme with differential privacy to protect mobility datasets, which releases the privacy-preserving population distributions and achieves better utility as well as a much smaller data transmission cost compared to the CMS-based method. Our proposed scheme is comprised of two collaborative components, global sketch and temporal sketch. The global sketch is responsible for aggregating the raw mobility data and decreasing the data transmission cost, while the temporal sketch is in charge of guaranteeing the utility of the population distributions aggregated by the global sketch. Besides, to enhance the privacy preservation, we employ the Laplace mechanism to make the transmitted data satisfy ϵ-differential privacy. Through our analysis and empirical experiments, compared to the other three state-of-the-art privacy-preserving methods on mobility datasets, our scheme could preserve the privacy of the mobility datasets with much less data transmission cost under the same utility loss.

1. Introduction

Along with the explosive growth of mobile smart devices, such as mobile phones, smart glass, and in-vehicle navigation devices, the massive amount of users' mobility data collected by them enables the fast development of various location-based applications. Recent market studies show that the Apple App Store has more than 2.2 million apps and Google Play has over 2.8 million apps [1]. In the perspective of supporting citizens' real lives, these mobility data enable intelligent transportation systems [2], spatial resource optimization, and even fire emergence response [3]. In the academic research perspective, mining mobility data enables us to better understand human behavior [4] and provide solutions to some global issues, such as controlling a full-blown COVID-19 epidemic [5]. Literally, location-based applications have become a necessity in modern life.

Although these mobile smart devices and applications have brought us considerable benefits and improved citizens' quality of life, the potential privacy risk also has gradually become a hot point of attention since the large amount and fine-grained collection of individual users' mobility data would reveal some of their sensitive information, such as lifestyle, physical condition, home location, and even identity [6]. In [7], the authors infer the "top N" locations for each user from the call records (which contain the location information) and correlate this information with the publicly available census data to find the user's home location. Golle and Partridge [8] found that even though the location records were anonymized, they still could infer the users' identities with some background

knowledge. Besides, studies in [9] have shown that by combining location data, gender, zip code, and birthdate, the majority of the US population can be uniquely identified.

Thus, protecting the privacy of mobile smart devices' users has become a burning problem to be settled. Therefore, large-scale corresponding research has been carried out. The mobility datasets usually contain the trajectory data of the users; thus, a simple and naive idea to solve the privacy problem is releasing the aggregated mobility datasets, such as the number of users in a target block at a specific timestamp, instead of the raw trajectory data. No individual users' information seems to be exposed through this method, and the released data still could support a large scale of applications, such as epidemic controlling [10], transportation scheduling [11], and business intelligence [12]. However, in 2017, Xu et al. [13] developed an attack system based on the uniqueness and regularity characteristics of human mobility to recover individual trajectories from aggregated mobility datasets, and they could achieve an accuracy of 73% ~ 91%. To resist such kind of an attack [14], P4Mobi has been proposed to aggregate mobility datasets, which have achieved privacy preservation at the statistic level with the probabilistic structure count-min sketch (CMS). P4Mobi achieves outstanding performance in resisting the attack, but its capacity for privacy preservation is determined by the size of CMS, which also controls the utility of the released data. To reduce the strong relevance between P4Mobi's privacy-preserving capacity and the utility of its released data, DP-Mobi [15] is carried out. DP-Mobi is an enhanced version of P4Mobi, in which, except CMS, differential privacy is also employed to control the privacy level, which also enhances the privacy of the released data compared to P4Mobi. The workflow's overview for P4Mobi and DP-Mobi is shown in Figure 1. Clearly, the raw mobility data are first processed by the trusted third party, and then, the privacy-preserving population distribution over a period of time is transmitted to the location-based service providers for further processing purposes. Experiments' results in [14, 15] demonstrate that for both P4Mobi and DP-Mobi, the utility of the aggregation mobility datasets is determined by the size of the sketches, and the better the utility is, the larger the sketches would be, and thus, the data transmission between the trusted third party and the service providers is less efficient.

The goal of our work is to improve the data transmission efficiency between the trusted third party and service providers with a desirable utility of the aggregated population distribution datasets. To this end, we propose an efficient and differential privacy-based scheme, which could protect the privacy of the mobility datasets along with improving data transmission efficiency. The workflow of our scheme is similar to that of P4Mobi and DP-Mobi as shown in Figure 1, in which the raw mobility dataset containing the trajectories of the users is collected by the trusted third party, and then, in the trusted third party, the global sketch of our scheme aggregates the raw mobility data and stores the result into the G-array. In the last step of our scheme, the G-array is transmitted to the location-based service providers for further purposes. To reduce the potential privacy risks

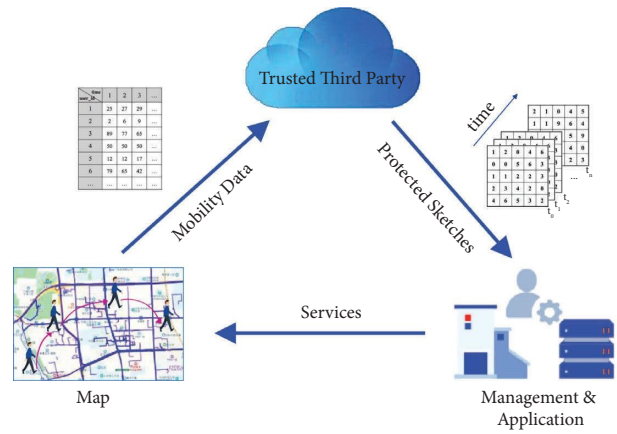


Figure 1: The overview of P4Mobi and DP-Mobi's workflow.

during data transmission, instead of executing the enquiry procedure of the trusted third party, our scheme delivers G-array to the location-based service providers and authorizes them to obtain the privacy-friendly population distribution datasets through enquiring G-array.

The highlight of our scheme is that we develop two collaborative components, global sketch and temporal sketch, to aggregate mobility datasets with better utility preservation. For P4Mobi and DP-Mobi, only one sketch is employed in the aggregation procedure, so that the utility of the released population distributions is directly affected by the size of the sketch. Since in the aggregation procedure, when the size of the sketch is small, the probability of collisions (i.e., two or more different records are stored in the same cell of the sketch) is high, and less utility is preserved. While in our scheme, the size of the temporal sketch is a set larger than that of the global sketch, so that the utility preservation of the temporal sketch is better. Then, in the aggregation stage, a mobility record l_{ij} is first stored in the temporal sketch, and only if the number of l_{ij} stored in the global sketch is smaller than that in the temporal sketch, the mobility record l_{ij} will be stored in the global sketch. Therefore, in our scheme, even if the size of the global sketch is small, the utility of the final released population distributions still could be preserved. Besides, as privacy issues may occur in the data transmission procedure between the trusted third party and the location-based service providers, we employ the Laplace mechanism to make the transmitted data satisfy ϵ-differential privacy.

The contributions of our paper are summarized as follows:

(i) We propose an efficient and differential privacy-based scheme for protecting mobility datasets, which aggregates the raw mobility data with the temporal sketch and global sketch. To reduce the privacy risk existing in the data transmission stage, our scheme sends G-array, which stores the population distributions in the global sketch, to the location-based service providers, and authorizes them to reconstruct the privacy-preserved population distributions from G-array.

(ii) To balance the tradeoff between the utility of the privacy-preserving population distribution and the size of the sketch that aggregates the mobility data, our scheme employs a temporal sketch with a larger size to aggregate the mobility data and be responsible for guaranteeing the utility of the population distributions aggregated by the global sketch. Compared to other CMS-based methods, our scheme could increase the utility of the aggregated population distributions on the premise that the sketches are of the same size.

(iii) We enhance the privacy of our scheme by employing the Laplace mechanism, and the transmitted data G-array satisfies ϵ-differential privacy. For the users of our scheme, they could tune the privacy parameter λ of our scheme to meet different privacy requirements.

(iv) We conduct an empirically experimental evaluation of our scheme by comparing it with other three state-of-the-art privacy-preserving methods (DP-simple, P4Mobi, and DPCMS) for mobility datasets. The results of the experiments demonstrate that the privacy-preserving capacity of our scheme is affected by the size of the temporal sketch, global sketch, and privacy parameter λ. Compared to the other three methods, the volume of the transmitted data in our scheme is much smaller on the premise that the utility of their released population distributions is at the same level.

We organize the rest of the paper as follows. In Section 2, we present the preliminaries related to our scheme, including (a) count-min sketch, (b) differential privacy, and (c) mobility datasets. A detailed introduction of our scheme is presented in Section 3. Section 4 evaluates our scheme by conducting experiments to compare it with the other three privacy-preserving methods. Finally, we conclude our work and corresponding further work in Section 5.

2. Preliminary

We present in this section a set of definitions (Sections 2.1 and 2.2) related to our scheme and other three kinds of mobility datasets with privacy-preserving mechanisms (Section 2.3), which will be the comparison objects to our scheme in Section 4.

2.1. Count-Min Sketch. Count-min sketch (CMS) is proposed by Cormode and Muthukrishnan [16], which is a probabilistic data structure to store the frequencies of items in an array and returns an estimate of the frequency of any given item when enquired. Due to the relatively small memory footprint and high accuracy, CMS has been a popular choice for a broad spectrum of applications, such as processing distributed datasets [17, 18], aggregating statistics in sensor networks [19], and detecting attacks in routers [20].

A CMS consists of an array of d rows and w columns, and we present the cell in the i^{th} row and j^{th} column of the array with $A_i[j]$, where $0 \le i \le d$ and $0 \le j \le w$. Each row A_i of the array is associated with an independent hash function $h_i(\cdot)$, which has a uniformly distributed output. To estimate the frequency of a given item by CMS, the values of all cells in the array are first initialized to 0, and then, two operations are executed, i.e., *insert* and *enquiry*.

2.1.1. Insert. To insert an item e into a CMS, i.e., to increment the stored frequency of item e by 1, the CMS first computes d hash functions $h_1(e), h_2(e), \ldots, h_d(e)$ and then increases the values of cells $A_1[h_1(e)], A_2[h_2(e)], \ldots, A_d[h_d(e)]$ by 1, respectively. The abovementioned process can be formulated as $\forall 1 \le k \le d$: $A_k[h_k(e)] \longleftarrow A_k[h_k(e)] + 1$.

2.1.2. Enquiry. We get the estimation of the frequency of item e through the *enquiry* operation, whose first step is similar to that of the *insert* operation, i.e., computing d hash functions $h_1(e), h_2(e), \ldots, h_d(e)$. Then, the smallest value among $A_1[h_1(e)], A_2[h_2(e)], \ldots, A_d[h_d(e)]$ is selected as the estimated frequency of item e to return.

By observing the description of CMS, it is not difficult to find that the size of the array (i.e., the value of w and d) in CMS is a key factor in determining the accuracy of the estimated frequency of the target items. Since in the *insert* operation, different items would be inserted into the same cell of the array (referred to as collisions) if the size of the array is smaller than the items' distribution, so that the frequency of these items would be overestimated, and the accuracy would be decreased. Inspired by such characteristics of CMS, we propose a scheme based on CMS, which employs a temporal sketch to guarantee the accuracy of the estimated population distributions even if the size of the array is small. The detail of our scheme is introduced in Section 3.

2.2. Differential Privacy. In this section, we give some preliminaries on differential privacy, which is an important step in our scheme for protecting the privacy of mobility datasets. Differential privacy was introduced by Dwork et al. [21], which is a framework to quantify the privacy level of a dataset while releasing useful aggregate information about the dataset.

Considering two neighbouring datasets $D_1, D_2 \in \mathscr{D}^n$, where $\mathscr{D}^n$ is the set of all possible datasets, and a real-valued query function $q: \mathscr{D}^n \longrightarrow \mathbb{R}$, a randomized query-answering mechanism $\mathscr{K}$ for the query function q will randomly output a number with probability distribution depending on query output $q(D)$, where D is the dataset. (ϵ, δ) – differential privacy is defined as follows.

Definition 1. A randomized mechanism $\mathscr{K}$ gives (ϵ, δ)-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subset \text{Range}(\mathscr{K})$.

$$\Pr[\mathscr{K}(D_1) \in S] \le e^{\epsilon} \Pr[\mathscr{K}(D_2) \in S] + \delta, \quad (1)$$

where ϵ and δ are the privacy budget and confidence of mechanism $\mathscr{K}$, respectively.

Obviously, based on Definition 1, the smaller the value of ϵ is, the closer the two probabilities ($\Pr[\mathscr{K}(D_1) \in S]$ and $\Pr[\mathscr{K}(D_2) \in S]$) will be, and the mechanism will achieve a better performance in preserving privacy. The smaller the value of δ is, the better the mechanism will comply with the definition of differential privacy more strictly.

The sensitivity [22] of a real-valued query function is defined as follows.

Definition 2. For a real-valued query function q: $\mathscr{D}^n \longrightarrow \mathbb{R}$, the sensitivity of q is defined as

$$\Delta \coloneqq \max_{D_1, D_2 \in \mathscr{D}^n} \left| q(D_1) - q(D_2) \right|, \tag{2}$$

for all D_1 and D_2 differing in at most one element.

In [22], basic techniques, such as randomized response, Laplace mechanism, and exponential mechanism, which could protect datasets while satisfying differential privacy are introduced. Among them, the Laplace mechanism is a classic technique satisfying $(\epsilon, 0)$-differential privacy, in which the privacy budget is determined by the value of ϵ, and the confidence value is 0, that is, the Laplace mechanism could comply with the definition of differential privacy strictly. Before introducing the Laplace mechanism, we first give the definition of Laplace distribution.

Definition 3. The Laplace distribution (centered at 0) with scale b is the distribution with probability density function represented as

$$\mathrm{Lap}(x \mid b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b} \right). \tag{3}$$

Sometimes, we write $\mathrm{Lap}(b)$ to denote the Laplace distribution with scale b.

As the name of the Laplace mechanism suggests, it first computes the query function q and then perturbs each coordinate with noise drawn from the Laplace distribution. The scale of the noise will be calibrated to the sensitivity of q (divided by ϵ).

Definition 4. Given any query function q: $\mathbb{N}^{|\mathscr{X}|} \longrightarrow \mathbb{R}^k$, the Laplace mechanism is defined as

$$\mathscr{M}_L(x, q(\cdot), \epsilon) = q(x) + (Y_1, \ldots, Y_k), \tag{4}$$

where Y_i are random variables drawn from $\mathrm{Lap}(\Delta/\epsilon)$ and Δ is the sensitivity of function q.

In our scheme, the privacy of the mobility datasets is still at high risk in the data transmission step between the trusted third party and the location-based service providers. Although, in this step, our scheme transmits the sketches instead of the estimated population distributions to enhance privacy, it is still possible for the evildoers to recover some information by enough observation of sketches, such as the specific hash functions related to each row of sketches or population distribution. Thus, our scheme will conduct the Laplace mechanism on the sketches and transmit these differentially private sketches. The specific parameters' setting of the Laplace mechanism is introduced in Section 3.

2.3. Privacy-Preserving Mechanisms on Mobility Datasets. To resist the threats affecting location privacy, various privacy-preserving mechanisms for mobility datasets have been carried out, and in this part, we put emphasis on three of them, which are relevant to our scheme.

2.3.1. DP-Simple. A simple and naive method to protect the privacy of mobility datasets is simply aggregating the trajectories and releasing the population distributions to cover the sensitive personal information existing in the trajectories. To enhance the privacy of this simple method, the concept of differential privacy has been adopted in a mechanism called geo-indistinguishability (which we refer to as DP-simple) [23]. In DP-simple, the Laplace mechanism is first applied to the raw trajectories to ensure differential privacy, i.e., noise drawn from the Laplace distribution is added to each location point in the trajectories. Then, DP-simple aggregates the encrypted trajectories to obtain the population distribution for release. As in such a mechanism, it is more difficult to recover sensitive individual information compared to simple aggregation, and the privacy-preserving level of DP-simple is determined by the parameters of the Laplace mechanism.

In the DP-simple method, the Laplace mechanism is applied to the individual user's trajectory data in the client site before transmitting it to the third party, and then, the third party generates the population distribution based on the protected trajectory data. In such a scenario, there are fewer constraints on the third party. The utility of the released population distribution is only affected by the parameter of the Laplace mechanism.

2.3.2. P4Mobi. Different from the DP-simple mechanism, which offers theoretical guarantees to protect the privacy of the mobility datasets, P4Mobi [14] reaches the objective of preserving the privacy of mobility datasets by employing a probabilistic data structure called count-min sketch (CMS) [16] in the statistic step of generating population distributions from mobility datasets and offers practical guarantees in protecting the privacy of mobility datasets. Descriptions in Section 2.1 have shown that the size of CMS would affect the accuracy of the estimated frequency of items, because when the size of CMS is smaller than the distribution of items, collisions would happen, which decreases the final accuracy. Inspired by this, P4Mobi formalizes the relationship between the utility (accuracy) loss of the final estimated population distributions and the size of CMS as

$$\text{Utility loss} = \left(1 - \left(1 - \frac{1}{w} \right)^{q-1} \right)^d, \tag{5}$$

where d and w denote the number of rows and columns in CMS, respectively, and q refers to the number of locations in the target area. In equation (5), $(1-1/w)^{q-1}$ denotes the probability that the other $(q-1)$ locations (except the current location) are not mapped to the position of the current location record in one row of the sketch, which means that the current location is mapped to a position with a value of zero (not occupied by other elements) in this row. Therefore, the probability that the current location is mapped to a position with a nonzero value (i.e., a collision occurs) in all d rows is $(1-(1-1/w)^{q-1})^d$. Equation (5) formulates the probability of collisions happening in CMS by using the CMS parameters, which also reflects the utility loss of CMS. Straightforwardly, the utility loss also reflects the privacy level of the released data; therefore, for a specific raw mobility dataset, the value of q is constant, and the users could tune the value of w and d in equation (5) to satisfy the different utility loss (privacy) requirements of the released population distribution.

2.3.3. DPCMS. Although P4Mobi achieves better performance in resisting the attack model [13] compared to traditional mechanisms, the tradeoff between the utility loss and the privacy level of the final released population distribution is the main concern of P4Mobi. As a progressive version of P4Mobi, the authors in [15] introduce a differential privacy-based probabilistic mechanism for mobility datasets releasing (referred to as DPCMS). Similar to P4Mobi, CMS is also the main component of DPCMS, which could enhance privacy at the statistic level. The main contribution of DPCMS is that it employs the Laplace mechanism to sketch before the enquiry step of CMS, which not only makes the sketches differentially private but also relieves the strong correlation between utility loss and privacy in P4Mobi.

While the abovementioned mechanisms offer various guarantees for the privacy of mobility datasets, it is difficult to guarantee the privacy of data in the transmission stage. In this paper, we show how our scheme protects the privacy of the transmitted data and guarantees a desirable utility (accuracy) of the final released population distributions.

3. Scheme

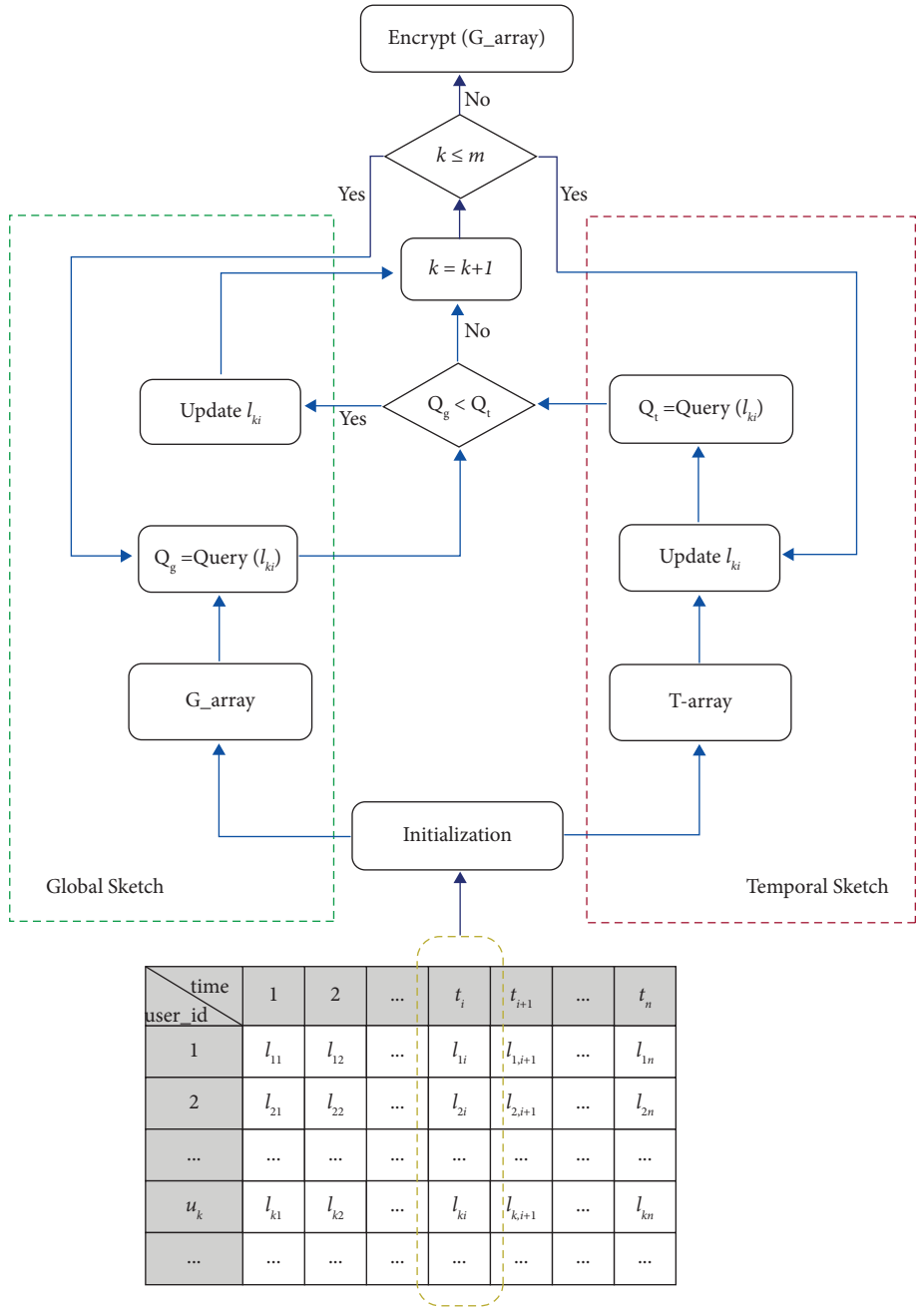
As shown in Figure 1, the application scenario of our scheme consists of three main participants, the mobility datasets' providers (individual users), the trusted third party, and the location-based service providers. To obtain location-based services, the individual users first send their mobility data to the trusted third party, and then, in the third party, our scheme aggregates these mobility data and stores the aggregated population distributions in the form of a protected array (G-array), which will be transmitted to the location-based service providers. Finally, the service providers recover the population distributions through the protected G-array and provide services to the individual users. In this section, we will describe each step of our scheme in detail and analyze the privacy and utility-preserving capacity of our scheme.

3.1. Mobility Dataset. Generally, the mobility dataset contains the trajectory information of individual users, which records the users' whereabouts (locations) associated with a series of timestamps. To describe our scheme intuitively, we assume that the target area is represented as a grid, and each location corresponds to a cell in the grid. Formally, let $L=\{L_1, L_2, \ldots, L_q\}$ be the universe of q locations in the target area and $T=\{t_1, t_2, \ldots, t_n\}$ denote n timestamps. For m users $U=\{u_1, u_2, \ldots, u_m\}$, we present their trajectories around n timestamps as $\mathrm{Tr}=\{\mathrm{tr}_1, \mathrm{tr}_2, \ldots, \mathrm{tr}_m\}$, where $\mathrm{tr}_{i(1\le i\le m)}=\{l_{i1}, l_{i2}, \ldots, l_{\mathrm{in}}\}$ and $l_{\mathrm{ij}(1\le j\le n)}\in L$. As mentioned earlier, our scheme aggregates the mobility datasets and releases the population distributions around n timestamps to support the location-based service providers, which are represented as $D=\{d_1, d_2, \ldots, d_n\}$, where $d_i=\{p_{1,i}, p_{2,i}, \ldots, p_{q,i}\}$ and $p_{k,i}$ denotes the number of users in location L_k at timestamp t_i.

3.2. Framework of Our Scheme. We summarize the framework of our scheme in Figure 2. Broadly, there are two components in our scheme, the global sketch and the temporal sketch, among which three operations, the *initialization*, *update*, and *query* are conducted in different orders.

As shown in Figure 2, when a new timestamp t_i arrives, our scheme will first call the *initialization* operation, which will initialize the G-array and T-array for the global sketch and temporal sketch, respectively. Then, for each user's location record $l_{k,i}, 1\le k\le m$ at timestamp t_i, the temporal sketch would first update the T-array with it and then query T-array about $l_{k,i}$ and return Q_t, which stands for the number of $l_{k,i}$ in T-array, e.g., the number of users at location $l_{k,i}$ stored in T-array. With respect to the global sketch, before updating the location record $l_{k,i}$, it conducts the *query* operation on G-array with $l_{k,i}$ first and compares the result Q_g with Q_t. On condition that the value of Q_g is smaller than that of Q_t, $l_{k,i}$ will be updated into G-array; otherwise, the next user's location record $l_{k+1,i}$ at timestamp t_i will be the input for both temporal sketch and global sketch. Successively, when $k\le m$ is false, i.e., all users' location records at timestamp t_i have been updated, our scheme will conduct the *encryption* operation on G-array in the global sketch (as shown at the top of Figure 2) and transmit the protected G-array to the location-based service providers for further processing. In the following section, we will describe the four operations *initialization*, *update*, *query*, and *encryption* of our scheme in detail.

3.2.1. Initialization. At each timestamp t_i of a mobility dataset, our scheme aggregates all users' location records in an array and transmits the array to the third party for the supplement of the population distribution of the target area at t_i. Therefore, the *initialization* operation will be conducted when a new timestamp arrives. In this operation, our scheme will preset four parameters, (w_g, d_g) and (w_t, d_t), and then, the *initialization* creates the G-array and T-array for global sketch and temporal sketch, respectively. Both the G-array and T-array are zero-valued arrays, and their sizes are w_g columns, d_g rows, w_t columns, and d_t rows, respectively. In



time / user_id	1	2	...	t_i	t_{i+1}	...	t_n
1	l_{11}	l_{12}	...	l_{1i}	$l_{1,i+1}$	...	l_{1n}
2	l_{21}	l_{22}	...	l_{2i}	$l_{2,i+1}$	...	l_{2n}
...	...	...	...	...	...	...	...
u_k	l_{k1}	l_{k2}	...	l_{ki}	$l_{k,i+1}$	...	l_{kn}
...	...	...	...	...	...	...	...

Figure 2: The framework of our scheme.

the final step of this operation, the *initialization* arranges d_g independent hash functions to associate with each row in G-array, and the same arrangement is applied to T-array. Later, in Section 3.3, we analyze how the values of parameters (w_g, d_g) and (w_t, d_t) could affect the utility of our scheme and how to present them to satisfy different utility requirements.

3.2.2. Update. To aggregate a location record l_{ki} into an array, *update* first computes d_g (or d_t) hush functions $h_i(l_{\mathrm{ki}})$ and then increases the value of the cell in the i^{th} row and $[h_i(l_{\mathrm{ki}})]^{\mathrm{th}}$ column of the array by 1. In Figure 3, we summarize the *update* operation.

3.2.3. Query. The *query* operation of our scheme is responsible for returning the frequency of the location l_{ki} in the array. Similar to the enquiry operation in CMS, the first step of *query* is computing d_g (or d_t) hush functions $h_i(l_{\mathrm{ki}})$, which stand for the column index of the target cells in each row of the array. Then, the smallest value among the target cells is selected as the result to return. Formally, for an array A (G-array or T-array) in our scheme, we have

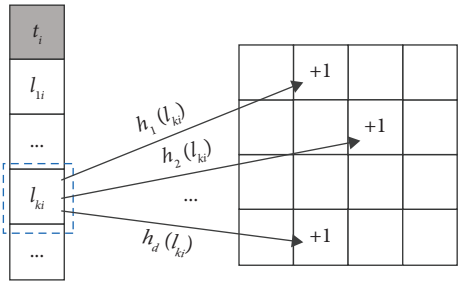


Figure 3: The *update* operation.

$$\text{Query}(l_{\text{ki}}) = \min_{j=1}^{d} A\left[j, h_j(l_{\text{ki}})\right], \tag{6}$$

which returns the frequency of the location record l_{ki} stored in the array A. In the raw mobility dataset, the location record l_{ki} denotes the position of the user u_k at timestamp t_i, which also could be construed as there was 1 user in l_{ki} at t_i. Therefore, the frequency of l_{ki} stored in the array represents the population in l_{ki}, i.e., the *query* operation of our scheme returns the population of users in the target location.

3.2.4. Encryption. When the last record of the timestamp t_i has been updated, our scheme would call encryption to protect the G-array in the global sketch. In this operation, we employ the Laplace mechanism to preserve the privacy of G-array and make it satisfy ϵ-differential privacy. According to the requirement of the privacy level, we would preset the value of parameter $\lambda = 2/\epsilon$, and then, random variables drawn from the Laplace distribution $\text{Lap}(\lambda)$ are added to the G-array. The *encryption* operation is formalized as

$$\text{Encryption}(G - \text{array}) = G - \text{array} + X \sim \text{Lap}(\lambda). \tag{7}$$

In Section 3.3, we will analyze how the *encryption* operation could make the protected G-array satisfy ϵ-differential privacy and how to tune the parameter λ to satisfy different privacy requirements.

3.3. Utility and Privacy Analysis of Our Scheme. In this section, we will analyze how our mechanism can meet the needs of strong utility, ideal privacy preservation, and cost-efficient data transmission.

3.3.1. Utility Preservation. Retrospecting the workflow of our scheme, it first updates the raw mobility dataset into the G-array and then triggers *encryption* to protect the G-array before being transmitted to the location-based service providers. An obvious solution to improve the efficiency of data transmission between the trusted third party and service providers is trimming the amount of the transmitted data, i.e., downsizing G-array. However, the discussion in the last paragraph in Section 2.1 indicates that the size of the G-array determines the accuracy of the final estimated population distributions, and the smaller the size is, the more inaccurate the results will be. In our scheme, to balance the tradeoff between the utility (accuracy) of the final results and the data transmission cost, we introduce the temporal sketch to guarantee the utility (accuracy) of the final population distributions while maintaining G-array in a relatively small size. The temporal sketch works in the trusted third party, where the storage space is abundant, so that the size of T-array, i.e., (w_t, d_t), in the temporal sketch could be set without regard to the constraint of storage space. By trying to avoid collisions happening in the G-array, before updating the current record l_{ki}, our scheme first queries G-array about the number of l_{ki} (Q_g) and then compares it with that aggregated by T-array (Q_t), and when the value of Q_g is smaller than that of Q_t, it could be considered that the probability of collisions happening in the G-array when updating l_{ki} is relatively low, which indicates that the utility(accuracy) of the G-array could be guaranteed.

On the strength of the abovementioned description, we will quantitatively analyze the utility-preserving ability of our scheme by associating the probability of collisions happening in the G-array with the preset parameters (w_t, d_t) and (w_g, d_g). For a location record l_{ki}, the precondition for the collisions happening in G-array is that collisions first happen in T-array, and in T-array, the probability of collisions of l_{ki} is calculated as

$$\Pr_t(l_{\text{ki}}) = \left(1 - \left(1 - \frac{1}{w_t}\right)^{q-1}\right)^{d_t}, \tag{8}$$

where q is the number of locations in the mobility dataset and $(1 - 1/w_t)^{q-1}$ is the probability that the other $(q-1)$ locations (except l_{ki}) in the dataset are not updated to the position of l_{ki} in one row of T-array, i.e., l_{ki} is updated to a position, which is not occupied by the others in this row of T-array. Therefore, the probability that l_{ki} is updated to a position in one row that has been occupied by others is $(1 - (1 - 1/w_t)^{q-1})$. According to the *query* operation of our scheme, only when a collision happens in all the rows of the T-array, the results of *query* will be affected, and thus, the probability of collisions of l_{ki} in T-array is $(1 - (1 - 1/w_t)^{q-1})^{d_t}$. The *update* and *query* operations in the global sketch are the same as that in the temporal sketch, besides a condition that the results of queryt (l_{ki}) in the global sketch should be smaller than that of the temporal sketch; i.e., collisions happen in the global sketch only when they first happens in the temporal sketch, and therefore, the probability of collisions of l_{ki} in G-array is

$$\begin{aligned}\Pr_g(l_{\text{ki}}) &= \Pr_t(l_{\text{ki}})\left(1 - \left(1 - \frac{1}{w_g}\right)^{q-1}\right)^{d_g} \\ &= \left(1 - \left(1 - \frac{1}{w_t}\right)^{q-1}\right)^{d_t}\left(1 - \left(1 - \frac{1}{w_g}\right)^{q-1}\right)^{d_g}.\end{aligned} \tag{9}$$

For a given mobility dataset, the value of q is a constant, and to improve the data transmission efficiency of our scheme, we could preset the values of (w_g, d_g), and then, according to the utility requirement, we set the value of (w_t, d_t). Following equation (9), for a value fixed (w_g, d_g),

when the value of w_t or d_t is set larger, the probability of collisions in G-array becomes lower, that is to say, the utility loss of our scheme is less. Thus, we could briefly sum up that for a given data transmission cost requirement, the utility of our scheme is relevant to the value of (w_t, d_t), and the larger their values are, the more utility our scheme will achieve.

To enhance the privacy of the global sketch, the Laplace mechanism is involved in our scheme. Before transporting the global sketch, our scheme generates the Laplace noise and adds them to the sketch. Through this step, the utility of the released population will also be affected. To make the utility analysis intuitive, we demonstrate in Figure 4 the probability of the noise generated by the Laplace mechanism under different values of parameter λ. Intuitively, when the value of λ is 0.5, the probability of generating a noise value of 0 is almost 99%, and in such a situation, the utility effect would be tiny. While when the value of λ increases to 2, the probability of value 0 decreases to about 21%, and the probability of value 4 increases to nearly 10%, that is to say, the value of the noise is larger, and the utility of the results will also decrease a lot. Above all, in our scheme, the value of parameter λ could also affect the utility of the results, and the larger the value of λ is, the more utility loss of the results will be.

3.3.2. Privacy Analysis. The privacy preservation and utility of the final results interact with each other in privacy-preserving mechanisms. An ideal privacy-preserving mechanism is always accompanied by the sacrifice of utility. In our scheme, we employ the temporal sketch and global sketch to aggregate mobility datasets and protect the users' privacy. By observing equation (9), it is apparent that the value of the probability of collisions of G-array is in the range of (0, 1), and that is to say, no matter how large the values of (w_d, d_t) and (w_g, d_g) are set, collisions will always exist in G-array, and the aggregated results will be different from the real population distributions of the mobility dataset. From the perspective of privacy preservation, our scheme protects the privacy of the users by introducing collisions in G-array.

However, it is insufficient to preserve both the privacy and utility of the mobility dataset by tuning the parameters (w_t, d_t) for temporal sketch and (w_g, d_g) for global sketch, as under such circumstances, the correlation between the privacy and utility of the protected mobility dataset is excessively strong. On account of such an issue, our scheme includes the *encryption* operation to enhance the privacy preservation of the mobility datasets.

Assuming that in the data transmission stage, our scheme sends the G-array directly to the service providers, so it would be possible for the evildoers to access the G-array and reconstruct the sensitive individual information of the users. Through the *encryption* operation, variables drawn from the Laplace distribution are added to the G-array, and the protected G-array $(G - \widehat{\text{array}})$ is sent to the service provider. Considering the evildoers' attack on G-array is $f(\cdot)$, then $f(G - \text{array})$ returns the cell in G-array that satisfies the evildoers' requirements. After involving the

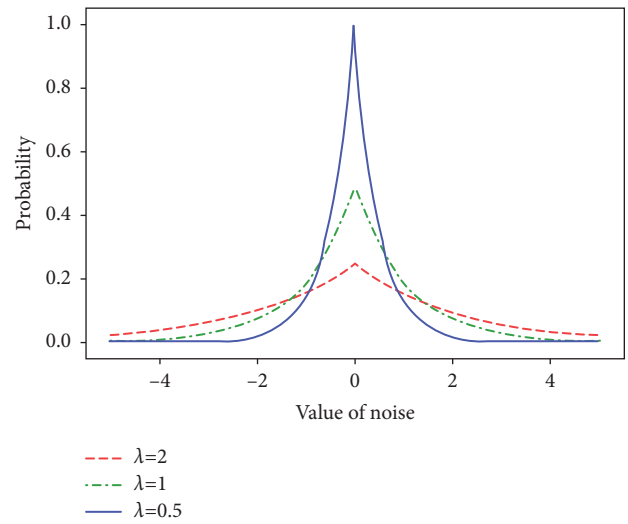


Figure 4: The probability of the noise generated by the Laplace mechanism under different values of λ.

encryption operation, the sensitivity (definition) of the attack function $f(\cdot)$ will be $\Delta f = \max|f(\text{G} - \text{array}) - f(\text{G} - \widehat{\text{array}})|$. After adding the noise, the position of the cell in the array containing the information that the evildoers are interested in may change or not, and therefore, the maximum difference between $f(G - \text{array})$ and $f(G - \widehat{\text{array}})$ is 2, i.e., $\Delta f = 2$. Based on the Laplace mechanism, our scheme chooses variables drawn from the Laplace distribution with scale $\lambda = 2/\epsilon$, i.e., $X \sim \text{Lap}(2/\epsilon)$, to protect the G-array. Here, we compare the probability density function of G-array and $G - \widehat{\text{array}}$ (denoted as P_{G} and $P_{\widehat{\text{G}}}$ in equation (10), respectively) at some arbitrary point $z \in \mathbb{R}^k$ as

$$\frac{P_{\text{G}}(z)}{P_{\widehat{\text{G}}}(z)} = \prod_{i=1}^{k}\left(\frac{\exp\left(-\epsilon\left|f\left(\text{G}_i - z_i\right)\right|/\Delta f\right)}{\exp\left(-\epsilon\left|f\left(\widehat{\text{G}}_i - z_i\right)\right|/\Delta f\right)}\right) \leq \exp(\epsilon). \quad (10)$$

The detailed derivation of equation (10) is provided in [22].

The abovementioned analysis indicates that the *encryption* operation makes the protected G-array satisfy ϵ-differential privacy. To achieve different privacy requirements, the users of our scheme could tune the value of ϵ, and the larger the value of ϵ is, the more the privacy will be preserved in the protected G-array.

4. Evaluation

In this section, we empirically evaluate our scheme on a real-world mobility dataset and a synthetic dataset and compare its performance with the other three privacy-preserving methods for mobility datasets, DP-simple, CMS, and DP-CMS, which have been introduced in Section 3.1. The experiments were performed on a MacBook Pro PC with a 2.70 GHz Intel Core-i5 processor and an 8.00 GB of RAM. The Python programming language was used to implement our proposed scheme.

4.1. Data. In our experiments, we use a GPS trajectory dataset collected from the GeoLife project [24–26] in a period of over four years (from April 2007 to August 2012) with a sampling interval of approximately 5s. In this dataset, a raw trajectory record contains a sequence of time-stamp points, each of which is associated with the information of latitude, longitude, and altitude. Ahead of the implementation of our scheme, we first preprocess the raw mobility dataset, and through observation of the distribution of the raw dataset, we find that the trajectories collected within the area of longitude $116.25^{\circ} \sim 116.50^{\circ}$ and latitude $39.85^{\circ} \sim 40.10^{\circ}$ are more intensive; therefore, we select the trajectory data in this area as the research target. For formalization purposes, we split the target area into a 25×25 grid and set the time granularity to one minute. In the preprocessed dataset, the total number of users is 354, the number of location cells is 625, and the number of time slots is 1440. The population distribution in the GeoLife dataset is sparse, and therefore, we generate a synthetic trajectory dataset as a supplement to the evaluation. In the synthetic dataset, the total number of users is 5000, the number of location cells is 400, and the number of time slots is 100.

4.2. Evaluation Criteria. Inspired by [27], we use the error between the real and published population distributions to evaluate the practical utility of the scheme. In our scheme, we assume that the population distribution transmitted to the location-based service providers at the timestamp t_i is $\widehat{D}_i$, and the real population distribution at this timestamp is D_i; then, the error between them is defined as

$$\mathrm{Err}_i = \frac{1}{q} \times \sum_{j \in [1,q]} \left| d_j - \widehat{d}_j \right|, \tag{11}$$

where q is the number of locations in the area and d_j and $\widehat{d}_j$ are the number of users at the j^{th} location at timestamp t_i in the real and published population distribution datasets, respectively. Then, the error of the scheme is defined as

$$\mathrm{ARE} = \frac{1}{n} \times \sum_{i \in [1,n]} \mathrm{Err}_i, \tag{12}$$

where n is the number of total time slots in the mobility dataset.

4.3. Experimental Results. In this part, we present the performance of our scheme in preserving the utility of the final released population distributions with different parameter settings: the size of G-array, the size of T-array, and the value of ϵ in the *encryption* operation.

4.3.1. Effect of the G-Array's Size. The size of the G-array is determined by the value of (w_g, d_g), and to measure the impact of the size of the G-array with the utility of our scheme, we conduct two experiments to observe the effect on the utility with the value of w_g and d_g, respectively. In the first experiment, for the GeoLife dataset, we set the depth of G-array as 6, i.e., $d_g = 6$. The size of the T-array is $w_t = 40$ and $d_t = 30$. The value of w_g varies from 3 to 21. For comparison purposes, we also conduct experiments with DP-CMS, CMS, and DP-simple methods. For the DP-CMS and CMS methods, the size of the sketches is set as the same as our scheme. While for DP-CMS, DP-simple, and our scheme, we set the parameter of the Laplace mechanism as 1. The results are shown in Figure 5(a).We can observe that in Figure 5(a), ARE decreases along with the increase in the value of g_w, which indicates that more utility of the population distributions released by our scheme will be preserved, when the width of the G-array is large. For the DP-simple method, its performance is irrelevant to w_g, and therefore, the average relative error of DP-simple in this experiment is about 1. When the value of w_g is smaller than 9, the ARE of our scheme is larger than DP-simple, and the reason is that in our scheme, except for the *encryption* operation, the small size of G-array causes more collisions and decreases the utility of the final result. While when the width of the G-array increases to 9, the utility-preserving ability of our scheme oversteps that of DP-simple, and the reason is that in our scheme, the Laplace mechanism is applied to the G-array instead of each location in DP-simple, so that the utility of the released population distributions is better. The tendency of the CMS method's utility-preserving performance is similar to that of the DP-CMS method, except that when the value of w_g is larger than 12, the average relative error of CMS is smaller than our scheme; i.e., the CMS method achieves better utility preservation performance than our scheme. Thus, when the size of the G-array becomes larger, the probability of collisions happening in both our scheme and CMS method will be smaller, and the utility of the CMS method will be better. However, in our scheme, the *encryption* operation employs the Laplace mechanism to protect the privacy of our scheme, so the utility-preserving ability of our scheme becomes weaker than CMS, which also indicates that our scheme achieves better privacy preservation compared to the CMS with the same size of sketch. For the synthetic dataset, similar results are shown in Figure 5(b).

Besides the value of w_g, we also conduct experiments to verify the effect of the depth of the G-array. In this experiment, for the GeoLife dataset, we set the width of the G-array constant as $w_g = 6$, and d_g varies from 3 to 30. The results of this experiment are shown in Figure 6, and obviously, the utility of our scheme becomes better when the value of d_g is increased. The comparison to the other three methods is similar to that in Figure 5. We could summarize the results shown in Figures 5 and 6 which show that the utility of our scheme is affected by the size of the G-array, and when the size is larger, the utility preservation is better. Compared to the other three methods, our scheme could achieve better utility preservation with a suitable size of G-array.

4.3.2. Effect of the T-Array's Size. In our scheme, T-array is an important component responsible for preserving the utility (accuracy) of the population distribution released by our scheme. In this part, we will show the experiments'

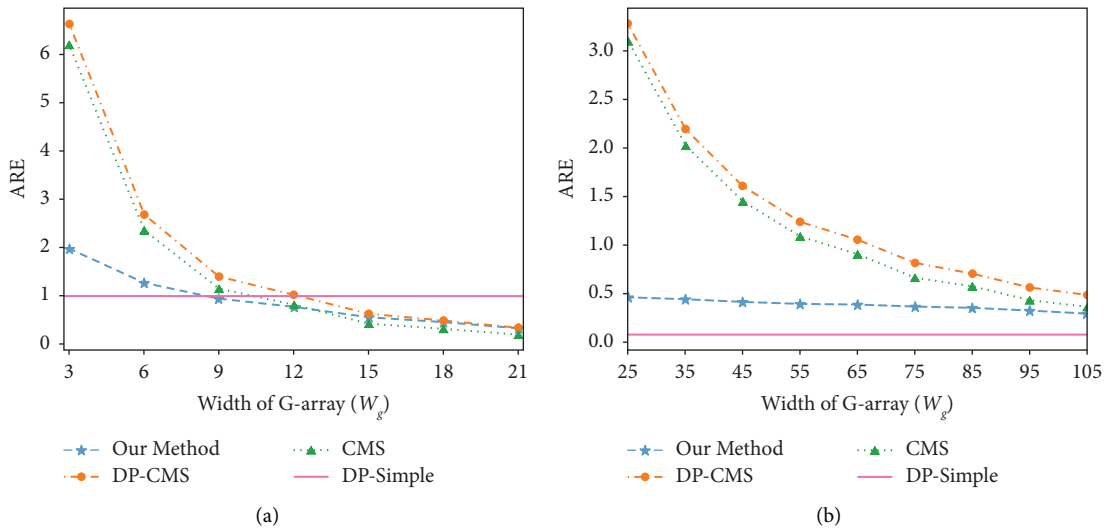


Figure 5: The average relative error of the released population distributions with different values of w_g. (a) GeoLife dataset. (b) Synthetic dataset.

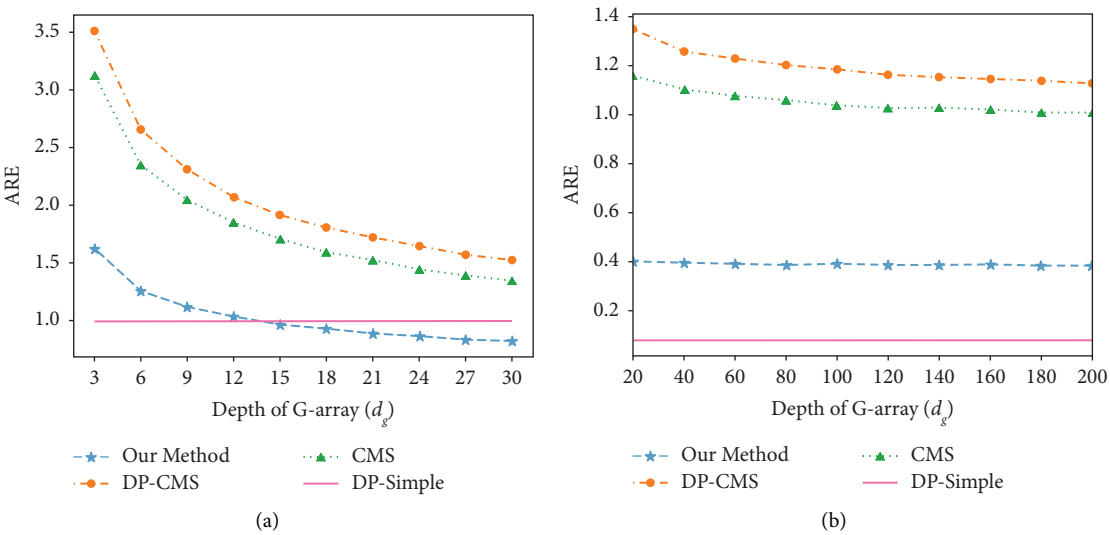


Figure 6: The average relative error of the released population distribution with different values of d_g. (a) GeoLife dataset. (b) Synthetic dataset.

results related to the effect of the T-array's size on the utility of our scheme. Similar to the experiments conducted in Section 4.3.1, we also conducted two experiments to study the effect of the T-array's width and depth, and the results are presented in Figures 7 and 8, respectively.

In Figure 7, for the GeoLife dataset, we set the G-array's size of our scheme as $w_g = 6$ and $d_g = 15$. The size of the CMS and DP-CMS methods is the same as that of the G-array. The parameters of the Laplace mechanism of our scheme, DP-simple, and DP-CMS are all set as 1 in this experiment. The value of w_t is in the range of 5–40. The T-array is a unique component of our scheme compared to the DP-CMS, CMS, and DP-simple methods, and therefore, in this experiment, the average relative error of these three methods is stable. Given the overall tendency shown in Figure 7, the utility of our scheme is affected by the value of w_t, and when the width of the T-array increases, the utility of our scheme will also increase. For the GeoLife dataset, specifically when the value of w_t is smaller than 20, the utility-preserving performance of our scheme outperforms CMS and DP-CMS, while being less reliable than DP-simple. Compared to CMS and DP-CMS, even though the sizes of the sketches that store the population distributions are the same, the existence of T-array in our scheme guarantees the utility of the released population distributions. While for the DP-simple method, when the size of the T-array is smaller, the probability of collisions happening in our scheme is relatively high, so the utility is decreased. When the value of w_t is larger than 20, our scheme could achieve the best utility preservation compared to the DP-CMS, CMS, and DP-simple methods.

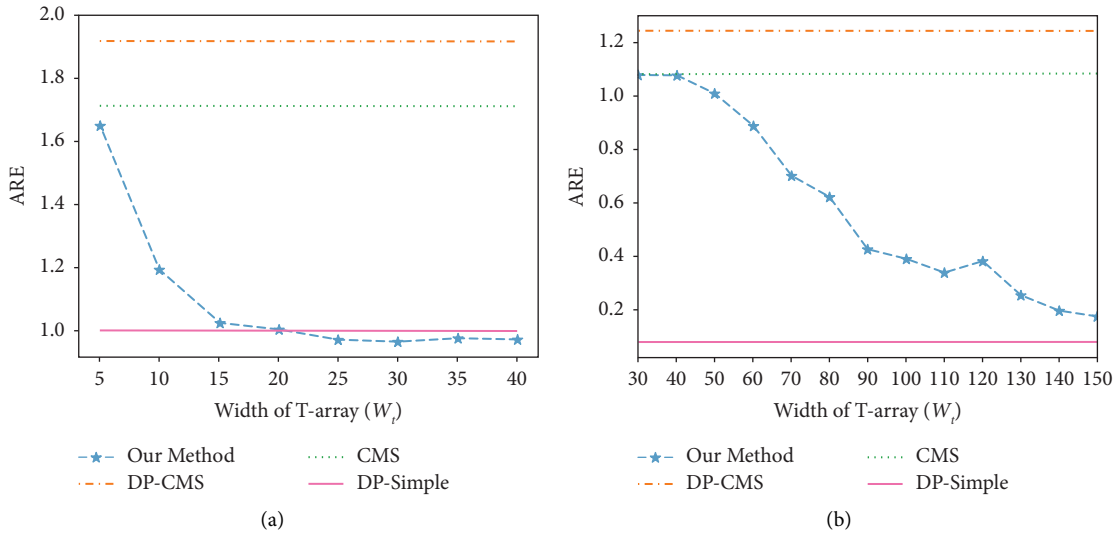


Figure 7: The average relative error of the released population distributions with different values of w_t. (a) GeoLife dataset. (b) Synthetic dataset.

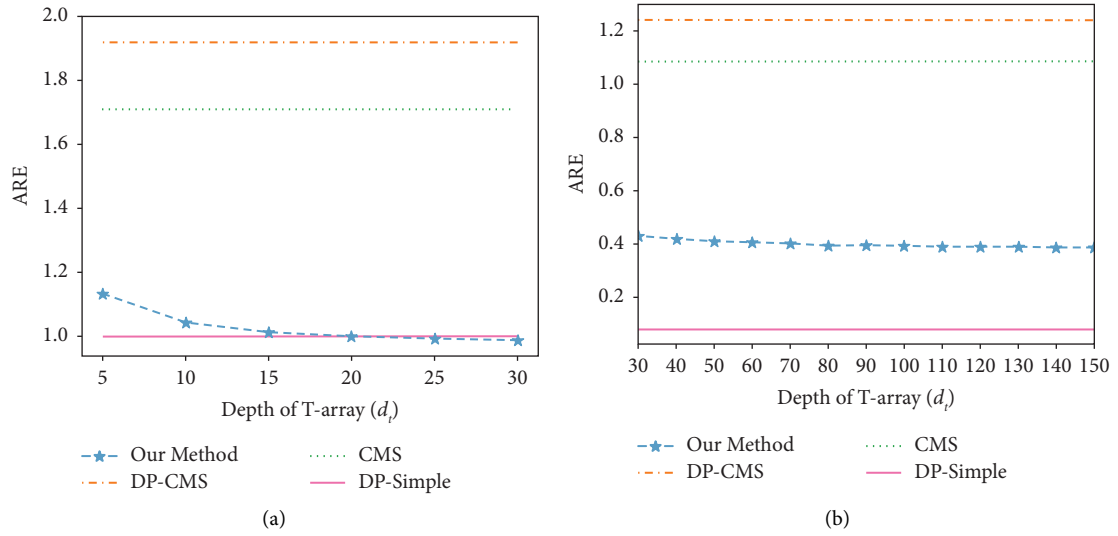


Figure 8: The average relative error of the released population distributions with different values of d_t. (a) GeoLife dataset. (b) Synthetic dataset.

Figure 8 shows the average relative error of the released population distributions with different depths of T-array in our scheme. For the GeoLife dataset, we set the width of the T-array as $w_t = 20$, the depth of the T-array varies from 5 to 30, and the size of the G-array, CMS, and DP-CMS is set as $w_g = 6, d_g = 15$. Besides, for DP-Ssimple, DP-CMS, and our scheme, the parameter of the Laplace mechanism is set as 1. The results shown in Figure 8 are similar to that presented in Figure 7, which indicates that the utility of our scheme will become better when the depth of the T-array increases, and when the value of w_t is larger than 20, the average relative error of our scheme is the smallest among these four methods, and thus, the utility preservation of our scheme is the best in this situation.

4.3.3. Effect of ϵ in the Encryption Operation. In Sections 4.3.1 and 4.3.2, we have demonstrated that the size of the G-array and T-array in our scheme could affect the utility of the released population distributions, and with appropriate settings of (w_g, d_g) and (w_t, d_t), our scheme could achieve a better utility preservation compared to the CMS, DP-CMS, and DP-simple methods. In this section, we will study the effect of the ϵ in our scheme on the final released population distributions.

As differential privacy is not involved in the CMS method, in this experiment, only comparisons between our scheme, DP-CMS method, and DP-simple method will be conducted. For our scheme, the size of the T-array is set as $w_t = 20$ and $d_t = 12$ and the size of the G-array is $w_g = 6$ and $d_g = 15$, which is the same as the DP-CMS method. For the

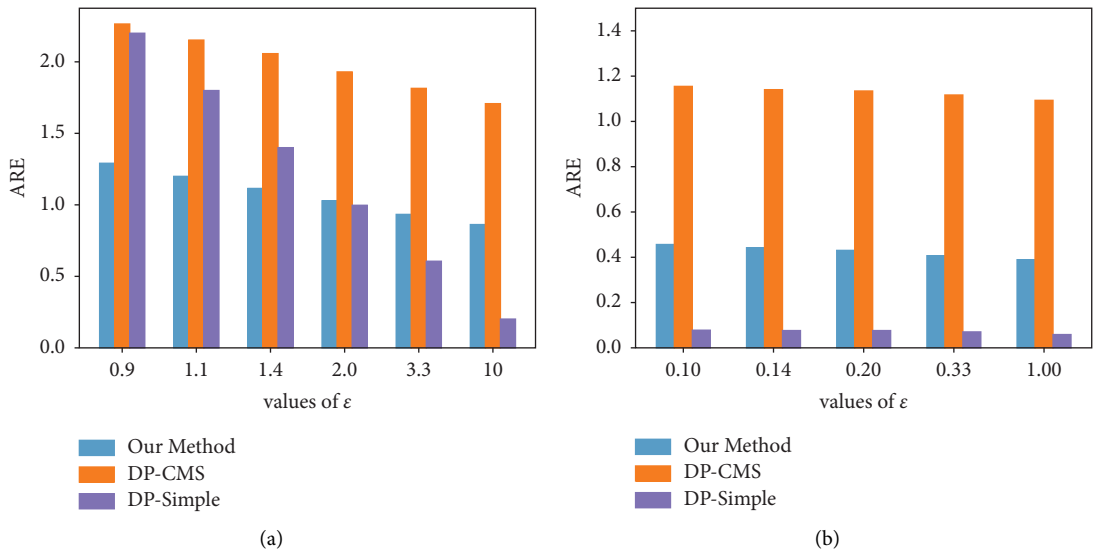


Figure 9: The average relative error of the released population distributions with different values of ϵ in the *encryption* operation. (a) GeoLife dataset. (b) Synthetic dataset.

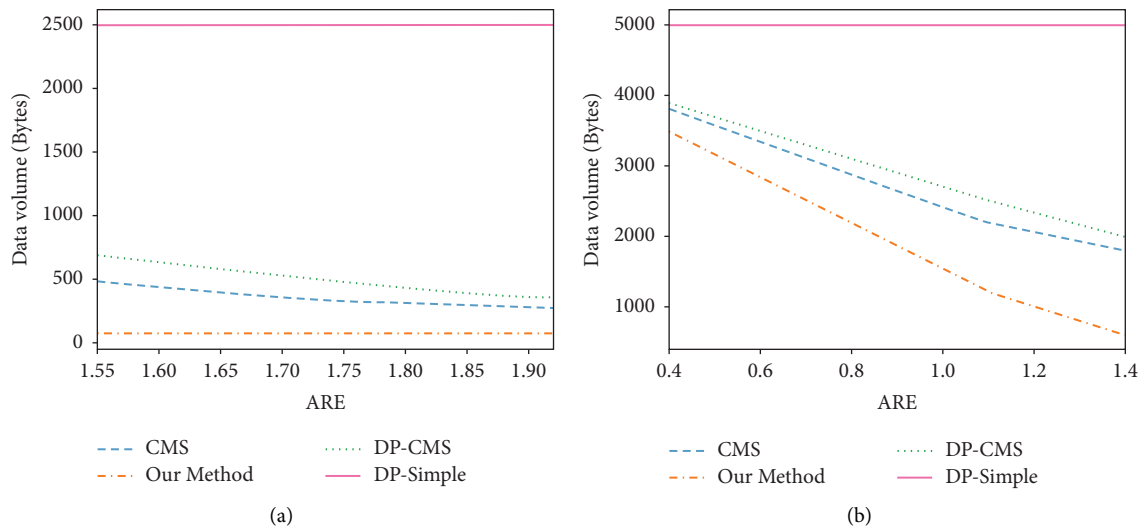


Figure 10: The volume of the transmitted data under different average relative errors. (a) GeoLife dataset. (b) Synthetic dataset.

GeoLife dataset, we range the value of ϵ from 0.9 to 10, and the results are shown in Figure 9(a). For the synthetic dataset, as the number of users is larger than the GeoLife dataset, we set the value of ϵ from 0.1 to 1, which would increase the added noise, and the results are shown in Figure 9(b)

Apparently, all these three methods' utility-preserving ability is correlated with the value of ϵ, and when the value of ϵ increases, the utility for all three methods will increase. Particularly, the average relative error of DP-CMS is the largest all along, i.e., the utility-preserving ability of DP-CMS is weaker than DP-simple and our scheme in this experiment. The reason for this result is that in DP-CMS, both collisions happened in the update procedure and the added noise drawn from the Laplace mechanism reduced the utility of the final results. While for our scheme and the DP-simple method, when the value of ϵ is larger than 1.0, the average relative error of DP-simple is smaller than ours, but when ϵ is decreased to smaller than 2.0, the average relative error of our scheme becomes smaller. In other words, with the value of ϵ becoming larger, the privacy level of the results also becomes stronger, and the utility-preserving ability of our scheme will be better than that of the DP-simple and DP-CMS methods.

4.3.4. The Comparison of Data Transmission Cost. Except for the ideal performance in preserving the utility of the released population distributions, another highlight of our scheme is the high efficiency in the data transmission stage. For CMS, DP-CMS, and our scheme, the array that stores the frequency of users is transmitted to the service providers, while

for the DP-simple method, the volume of the data that are transmitted to the service providers is $4 \times q$, where q is the number of locations in the raw mobility dataset, and each value occupies 4 bytes in the transmission stage. In this experiment for the dataset and the synthetic dataset, the transmitted data volume of the DP-simple method is 2500 bytes and 5000 bytes, which is shown as the solid line in Figure 10(a) and Figure 10(b).

For CMS, DP-CMS, and our scheme, we combine the results demonstrated in Figures 5–9 and show the correlation between the average relative error and the transmitted data volume in Figure 10. Horizontally, when the average relative error becomes larger, the transmitted data volume will decrease, which could be understood as when the volume of the transmitted data is reduced, the utility of the final released population distributions will also decrease. While vertically observing Figure 10, we see that when the utility preservation is consistent, the transmitted data volume of our scheme is much smaller than the CMS, DP-CMS, and DP-simple methods.

5. Conclusion

In this paper, we propose a scheme that protects the privacy of the mobility datasets by employing a temporal sketch, a global sketch, and a Laplace mechanism, which updates the users' trajectories' data into a G-array and sends it to the location-based service providers, who will get the population distributions through enquiring about G-array. Compared to other CMS-based methods, our scheme could balance the tradeoff between the utility and privacy preservation of the released population distributions. Beyond this, the joining of T-array in our scheme enables the data transmission cost between the trusted third party and location-based service providers to decrease sharply, and the evaluation results show that compared with the DP-CMS, CMS, and DP-simple methods, our scheme could save approximately 80%, 75%, and 96% data transmission cost when the average relative error (utility loss) is about 1.8.

Nevertheless, in the research area of preserving the privacy of mobility datasets, another interesting research direction is on improving the efficiency of differential privacy, as location data collected by GPS could contain sensing noise [28], and for such location data, it is pointless to add more noise. Therefore, research on distinguishing location data with sensing noise would be of great significance to improve the efficiency of differential privacy. There are still other factors affecting the privacy or utility-preserving capacity. In our work, optimizing the spatial [29] and temporal granularity [30] in the data collection stage to protect the privacy of mobility datasets would also be another direction of future research.

Data Availability

The data that support the findings of this study are available at https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qing Yang proposed the main idea, designed the model of the methodology, and wrote the initial draft. Fujun Ji is the corresponding author of this paper. He managed the raw GPS data and prepared the figures of the paper. Fei Liu was responsible for the privacy analysis part of the paper and proved that our method satisfied differential privacy. Qing Yang, Fujun Ji, and Fei Liu contributed equally to this work.

Acknowledgments

The authors are very grateful for the strong support of the funders and to the authors of the references, on which this study was carried out. This work was supported in part by the Fundamental Research Fund for the Capital University of Economics and Business under Grant no. XRZ2022031 and the National Natural Science Foundation of China under Grant no. 61806197.

References

[1] J. Clement, *Number of Apps Available in Leading App Stores 2019*, Web page, Hamburg, Germany, 2020.

[2] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Vehicular Technology Magazine*, vol. 5, no. 1, pp. 77–84, 2010.

[3] M. Ben Lazreg, J. Radianti, O.-C. Granmo, M. B. L. Palen, T. Comes, and A. Hughes, *Smartrescue: Architecture for Fire Crisis Assessment and Prediction*, ISCRAM, Omaha, NE, USA, 2015.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[5] L. Ferretti, C. Wymant, M. Kendall et al., "Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, Article ID eabb6936, 2020.

[6] J. Krumm, "Inference attacks on location tracks," in *Proceedings of the International Conference on Pervasive Computing*, pp. 127–143, Springer, Berlin Heidelberg, May 2007.

[7] H. Zang and B. Jean, "Anonymization of location data does not work: a large-scale measurement study," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, pp. 145–156, Las Vegas, ND, USA, September 2011.

[8] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Proceedings of the International Conference on Pervasive Computing*, pp. 390–397, Springer, Berlin, Heidelberg, May 2009.

[9] L. Sweeney, *Uniqueness of Simple Demographics in the Us Population*, LIDAP-WP4, Pittsburgh, PA, USA, 2000.

[10] C. O. Buckee, S. Balsari, J. Chan et al., "Aggregated mobility data could help fight covid-19," *Science*, vol. 368, no. 6487, pp. 145-146, 2020.

[11] K. McKinzie and J. W. Barnes, "A review of strategic mobility models supporting the defense transportation system," *Mathematical and Computer Modelling*, vol. 39, no. 6-8, pp. 839–868, 2004.

[12] T. Andrade, B. Cancela, and J. Gama, "From mobility data to habits and common pathways," *Expert Systems*, vol. 37, no. 6, 2020.

[13] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: user privacy is not preserved in aggregated mobility data," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1241–1250, International World Wide Web Conferences Steering Committee, Geneva, Switzerland, April 2017.

[14] Q. Yang, Y. Shen, D. Vatsalan, J. Zhang, M. A. Kaafar, and W. Hu, "P4mobi: a probabilistic privacy-preserving framework for publishing mobility datasets," *IEEE Transactions on Vehicular Technology*, vol. 69, 2020.

[15] J. Zhang, Q. Yang, Y. Shen, Y. Wang, X. Yang, and B. Wei, "A differential privacy based probabilistic mechanism for mobility datasets releasing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 201–212, 2020.

[16] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," *Journal of Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[17] Y. Tong, A. X. Liu, M. Shahzad et al., "A shifting bloom filter framework for set queries," 2015, https://arxiv.org/abs/1510.03019.

[18] T. Yang, A. X. Liu, M. Shahzad et al., "A shifting framework for set queries," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 3116–3131, 2017.

[19] Y. Zhou, Y. Tong, J. Jiang et al., "Cold filter: a meta-framework for faster and more accurate stream processing," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 741–756, Houston, TX, USA, November 2018.

[20] D. Barman, P. Satapathy, and G. Ciardo, "Detecting attacks in routers using sketches," in *Proceedings of the 2007 Workshop on High Performance Switching and Routing*, pp. 1–6, IEEE, Poznan, Poland, June 2007.

[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer, Zagreb, Croatia, May 2006.

[22] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[23] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security*, pp. 901–914, ACM, Berlin, Germany, November 2013.

[24] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, pp. 791–800, Madrid, Spain, April 2009.

[25] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, pp. 312–321, Seoul, Korea, January 2008.

[26] 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, 2010.

[27] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, "Differentially private event sequences over infinite streams," *Proceedings of the VLDB Endowment*, vol. 7, 2014.

[28] Y. Sei and A. Ohsuga, "Private true data mining: differential privacy featuring errors to manage internet-of-things data," *IEEE Access*, vol. 10, pp. 8738–8757, 2022.

[29] X. Sun, Q. Ye, H. Hu et al., "Synthesizing realistic trajectory data with differential privacy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5502–5515, 2023.

[30] S. Shamshad, K. Mahmood, U. Shamshad, I. Hussain, S. Hussain, and A. K. Das, "A provably secure and lightweight access control protocol for ei-based vehicle to grid environment," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 16650–16657, 2023.