P4Mobi: A Probabilistic Privacy-Preserving Framework for Publishing Mobility Datasets

Qing Yang, Yiran Shen, *Senior Member, IEEE*, Dinusha Vatsalan, Jianpei Zhang, Mohamed Ali Kaafar, *Senior Member, IEEE*, and Wen Hu, *Senior Member, IEEE*

***Abstract*—The large-scale collection of individuals' mobility data poses serious privacy concerns. Instead of perturbing data by adding noise to the raw location data to preserve privacy of individuals, we propose an approach that achieves privacy-preservation at the statistics level of aggregating mobility datasets with the probabilistic data structure *Count-Min Sketch* (CMS) [1], which has been widely used to provide efficient statistic functions with a tunable error bound. We use CMS to estimate the population density distributions in the mobility datasets, where the error bound determines utility guarantees. We develop *P4Mobi*, a novel Probabilistic Privacy-Preserving Publishing framework for Mobility datasets that protects individuals' privacy while complying to a specific utility requirement. We empirically validate the performance of *P4Mobi* in terms of utility and privacy-preservation by demonstrating its resilience against a recently proposed reconstruction attack model using two real-world datasets. We compare *P4Mobi* to two state-of-the-art methods and show that with the same level of privacy achieved against our attack model, *P4Mobi* significantly improves the utility of the published mobility datasets by up to 20%. We also provide a theoretical estimate of the utility achieved by *P4Mobi*. We found a very consistent match between the estimated and empirical utility of *P4Mobi* as evaluated on two datasets.**

***Index Terms*—Mobility datasets, count-min sketch, privacy, utility, aggregation, data publishing.**

I. Introduction

With the increasing popularity of mobile devices geared with Internet access and GPS functionalities, a wide range of location-based services (LBSs) has emerged in the last decade [2]–[4]. Examples of applications range from cost effective transportation and traffic-aware recommendations (e.g., Google Map [5], TomTom [6]), to location-based social networks [7]–[9]. These LBSs are powered by mining or analysing the large-scale collection of individuals' mobility data.

Individuals' mobility data reflect the users' trajectories, i.e., the places the users have visited during a period of time. However, they are of significant value and represent a sensitive piece of information as they can be easily linked to a variety of additional personal information, such as occupation, life style, health issues as well as political and religious beliefs [10], [11]. For instance, Krumm *et al.* [10] showed that based on two-weeks GPS traces from 172 individuals, the home addresses (with median error below 60 meters) and some identities of these individuals (with success rate above 5%) can be successfully inferred by joining GPS traces with a reverse geocoder [12] and a Web-based whitepage directory. In [13], Gambs *et al.* built a Mobility Markov Chain (MMC) from the observed mobility traces in the training phase and used the MMC to infer the identity of a particular individual behind a set of mobility traces with a success rate of up to 45%.

In order to overcome the growing privacy concerns and issues that preclude the use of mobility data for LBSs, one approach is to use aggregation of the mobility data of all users prior to the data publishing [14], i.e., only the population density distributions are disclosed. However, the accurate population density distributions obtained from simple aggregation methods are vulnerable to a recently proposed trajectory reconstruction attack [15], which exploits the uniqueness and regularity characteristics of human mobility and recovers individuals' trajectories by associating the same users' mobility records in the neighbouring time slots. The sensitive information of individuals, e.g., their home address or identities, can then be revealed by linking some background knowledge about the individuals with the reconstructed trajectories.

Alternatively, methods relying on random noise perturbation of the location data have been proposed. [16] for instance aims to provide differential privacy (DP) guarantees to the aggregated mobility dataset. However, as shown later in Section V-D, this DP-based method incurs significant utility loss in the resulting population density distributions. In this paper, we propose a **P**robabilistic **P**rivacy-**P**reserving framework for **P**ublishing **Mobi**lity datasets (referred as ***P4Mobi***) that not only protects published aggregated mobility datasets against reconstruction attacks but comes with an improved utility of the published

Manuscript received May 24, 2019; revised November 19, 2019 and April 13, 2020; accepted April 21, 2020. Date of publication May 11, 2020; date of current version July 16, 2020. This work was supported in part by National Natural Science Foundation of China under Grant 61702133, in part by Natural Science Foundation of Heilongjiang province under Grant QC2017069, and in part by the Optus Macquarie University Cyber Security Hub. The work of W. Hu was supported by Cyber Security CRC. The review of this article was coordinated by Prof. Z. Ma. *(Corresponding author: Yiran Shen.)*

Qing Yang, Yiran Shen, and Jianpei Zhang are with the College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China (e-mail: yangqing@hrbeu.edu.cn; shenyiran@hrbeu.edu.cn; zhangjianpei@hrbeu.edu.cn).

Dinusha Vatsalan is with Data61, CSIRO, Eveleigh, NSW 2015, Australia (e-mail: dinusha.vatsalan@data61.csiro.au).

Mohamed Ali Kaafar is with the Macquarie University, Optus Macquarie University Cyber Security Hub, Macquarie Park, NSW 2109, Australia, and also with Data61, CSIRO, Eveleigh, NSW 2015, Australia (e-mail: dali.kaafar@mq.edu.au).

Wen Hu is with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia, and also with Data61, CSIRO, Eveleigh, NSW 2015, Australia (e-mail: wen.hu@unsw.edu.au).

Digital Object Identifier 10.1109/TVT.2020.2994157

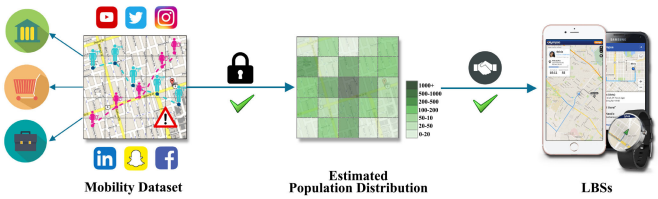


Fig. 1. To protect the sensitive personal information contained in the individuals' trajectories, *P4Mobi* encodes raw mobility dataset and publishes the estimated population density distributions with high practical utility to support LBSs.

datasets than existing DP-based perturbation methods. In particular, our objective is to design a framework that strikes a balance between privacy protection and hence resistance to a recently proposed reconstruction attack [15], and utility compared to a widely accepted DP-based method illustrated in [16]. The novelty behind *P4Mobi*'s approach is to achieve privacy-preservation at the statistics level of aggregation with the probabilistic data structure *Count-Min Sketch* (CMS) [1]. To protect the aggregation of mobility data, the traditional methods often firstly operate on the raw data, such as adding noise to the location/position, and then in the statistic step, they simply count the number of users at different locations. However, *P4Mobi* breaks the traditional aggregation rules, which count the number of users directly, it employs CMS and protects the mobility data in the statistics step.CMS has been widely used to provide efficient statistic functions with a tunable error bound. We then use the CMS to estimate the population density distributions in the mobility datasets, where the error bound provides utility guarantees to the estimated population distributions.

An overview of *P4Mobi* is shown in Fig. 1. The aim is to achieve *privacy-preservation and tunable utility* of mobility data. CMS allows the "encoded" counts of users in different locations (density) to be estimated with guaranteed utility loss (controlled by the parameters of CMS). The intuition behind using CMS is that (i) the utility loss it incurs (measured as the mismatch between the actual and estimated population density distribution) would provide improved privacy preservation to the individual users' locations in the aggregated datasets against the reconstruction attack proposed in [15], while (ii) the errors (i.e. the estimated counts would differ from the actual counts) in CMS based population density can be probabilistically controlled by tuning the parameters of CMS, which enables *P4Mobi* to meet the utility requirements.

In particular, a dataset custodian aiming to share or release a mobility dataset to support a third party LBS may first determine the minimal utility requirement needed, and then adjusts the parameters of *P4Mobi* accordingly and encode the raw mobility data into the CMS data structure.

The main contributions of this paper are:

- We propose ***P4Mobi***, a **P**robabilistic **P**rivacy-**P**reserving **P**ublishing framework that encodes **Mobi**lity datasets and publishes their estimated population density distributions using CMS. To the best of our knowledge, this is the first piece of work addressing the privacy issues of the reconstruction attacks on aggregated mobility datasets while preserving a reasonable degree of practical utility as a constraint while releasing the dataset.
- We formulate the utility loss of *P4Mobi* in terms of Probability of Collisions (PoC) using the key CMS parameters, and then deduce the relationship between PoC and the utility of the published datasets, so that it allows conveniently to tune the parameter settings of CMS according to different requirements of utility.
- We extensively evaluate the performance of *P4Mobi* first against the reconstruction attack (measured as the reconstruction errors) and then as a tradeoff between privacy and utility using two real-world mobility datasets collected from different application scenarios. We compare *P4Mobi* to two different approaches. First one is a simple mobility dataset aggregation technique proposed in [14] referred as (S-MDA) and the second one is the DP-based geo-indistinguishable Mobility Dataset Aggregation approach (referred as DP-MDA) [16]. The S-MDA method serves as the benchmark for our utility evaluation as it is assumed to be 100% accurate in terms of published population density. The DP-MDA method provides improved privacy against the reconstruction attack [15] at the cost of some utility loss. We show that *P4Mobi* is able to improve the utility by up to 20% when maintaining the same privacy preservation performance against reconstruction attacks.
- Finally, we validate the practical usage of the estimated utility guarantee on real-world datasets under different parameter settings. The results show that the empirical utility can be well approximated by our estimate and the correlation between the empirical and theoretical utility is consistently over 98%, which indicates that the key CMS parameters can be tuned according to different utility requirements without repeatedly processing the whole dataset to find the parameter setting satisfying the utility requirement.

The rest of the paper is organised as follows. Section II provides some background material including a brief introduction to CMS and an overview of two state-of-the-art methods for privacy-preserving publishing of mobility datasets, which we compare *P4Mobi* to. Section III describes the threat model and introduces the trajectory resconstruction attack we focus on in this paper. Section IV presents the framework of *P4Mobi* and analyses the privacy and utility-preserving properties of our proposed approach. *P4Mobi* is empirically evaluated in Section V, and Section VII concludes the paper.

II. Background

A. The Count-Min Sketch Data Structure

Count-Min Sketch [1] (CMS) is a probabilistic data structure typically used to efficiently store (encode) the frequencies of events (referred also as items) in a database. The two major components of CMS are: (i) encoding or updating of items in the sketch and (ii) estimating the counts of encoded items in the sketch.

1) Encoding and Updating: Given a pair of parameters (θ, δ), the sketch parameters can be set as $w = \lceil e/\theta \rceil$ and

$d = \lceil ln(1/\delta) \rceil$, where e is Euler's number, θ and δ mean that the error in answering a query is within a factor of θ with probability $1 - \delta$, w and d indicate the width and depth of the sketch. All the entries of the sketch are initialised to zeros. Before the encoding and updating procedure, d hash functions $h_1, h_2, .., h_j, .., h_d : \{0,1\}^* \rightarrow \{0,1\}^w$, are chosen uniformly from a pairwise-independent hash family [1].

To update the item i with the count c_i in the sketch, the corresponding entry in each row is incremented by c_i. The position of the entry to be updated in each row is determined by the output of corresponding hash function, e.g., the position of the entry in the j^{th} row is the output of the j^{th} hash function, i.e., $h_j(i)$. To be more specific, to update the count c_i of item i in the sketch, for the j^{th} row,

$$CMS[j, h_j(i)] \leftarrow CMS[j, h_j(i)] + c_i, \tag{1}$$

where CMS is the sketch to store the statistics of the items. The sketch is updated iteratively until all the items are considered. It is worth noting that, different items can be updated to the same entry in one row after applying hash functions, which causes the so-called *collisions*. *Collisions* lead to false positive probability, which will be discussed in detail in Section IV-B.

2) Estimating Counts: After updating all items in the CMS, the final count of each item can be estimated. In the estimation procedure, the minimal value of all the entries corresponding to an item from different rows across the whole sketch is regarded as the final count of the very item (corresponds to the term *count-min*). Formally, for item i, its estimated count $\hat{c_i}$ is the minimum of all the entries related to item i from all rows of CMS, i.e.,

$$\hat{c_i} = \min_{j=1}^{d} CMS[j, h_j(i)]. \tag{2}$$

As different items can be mapped to the same entry of the sketch, the estimated counts of the items are always larger than or equal to their actual counts. In other words, CMS allows false positives, but not false negatives, which means the estimated count is larger than or equal (if no collision) to the actual count. This is important in certain applications, for example, in disease outbreak detection systems that aim to issue alerts when the consumption of a certain drug exceeds a threshold at all or some of the hospitals. In such applications, false negatives have more cost than false positives. The same applies to other example applications of transport planning and traffic management.

B. The Mobility Data Aggregation Method (S-MDA)

In order to protect individuals' privacy, several aggregation methods, which generally summarise or anonymise individuals' location traces prior to releasing them have been developed.

One simple approach is aggregating mobility data [14] by simply counting the number of users within some area during specific time period, and then only the aggregated results are published. The aggregated results are the population density distributions which can be easily utilised to support many LBS applications.

S-MDA provides useful summary information (without any utility loss) that can support a wide range of statistical functions, such as predicting events, studying the effect of "shocks" in transport and detecting traffic anomalies. After aggregating all the users' mobility data together, it is expected to be highly unlikely to seek out a specific user's trajectory. Therefore, it provides some sort of guarantee on privacy-preserving of individual's mobility data until the emergence of the reconstruction attack algorithm proposed in [15].

C. The Geo-Indistinguishability Method

In [16] Andrés *et al.* propose a differentially private [17] method for protecting location-based data. The method gurantees Geo-indistinguishability by perturbing the data with a differentially private mechanism K that remaps each location point x to the closest point (by adding noise) in the discrete domain. Authors use Laplacian noise [18] so that the mechanism K satisfies $d_p(K(x), K(x')) \leq \epsilon d_e(x, x')$ for all x, x', where $d_p(K(x), K(x'))$ is defined as the *multiplicative distance* between two distributions $K(x)$ and $K(x')$, and $d_e(x, x')$ denotes the *Euclidean distance* between two different points x and x'.

III. Threat Model

A. Overview

In [15], authors present that the aggregating mobility dataset does not preserve users' privacy, since a user's mobility pattern is regular while different from others'. Based on the characteristics of human mobility, they transform the population density distribution to a *location-time* format and propose the trajectories reconstruction attack that iteratively associates the same users' mobility records in the neighbouring time slots. They exploit the regularity of mobility data to estimate the next location of the user and choose the location in the aggregated data with the largest similarity to the estimated next location as the reconstructed next location according to the uniqueness pattern of human mobility data.

To recover trajectories from the aggregated dataset, the first step is transforming the population density distribution $C^t = [c_1^t, c_2^t, \ldots, c_i^t, \ldots, c_q^t]$ into a *location-time* record $P^t = [p_1^t, p_2^t, \ldots, p_j^t, \ldots, p_m^t]$, where c_i^t represents the number of mobile users at location i during time slot t, p_j^t represents the location of the j^{th} user at time slot t, and q represents the total number of possible locations, while m is the total number of users . To link the *location-time* records that represent the same user across different time slots, the reconstruction attack is modeled as a *Linear Sum Assignment Problem* [19], which has been extensively studied and can be solved in polynomial time based on *Hungarian algorithm* [20].

B. Reconstructing Individuals' Trajectories

Specifically, we assume a set of recovered trajectories until time slot t as $S^t = [s_1^t, s_2^t, \ldots, s_k^t, \ldots, s_m^t]$, where $s_k^t = [l_k^1, l_k^2, \ldots, l_k^t]$ is the k^{th} recovered trajectory and l_k^t is the recovered location at time slot t. To recover the next position l_k^{t+1} from the *location-time* records $P^{t+1} = [p_1^{t+1}, p_2^{t+1}, \ldots, p_m^{t+1}]$, an estimated location $\hat{p}_k^{t+1}$ is first generated based on the continuity

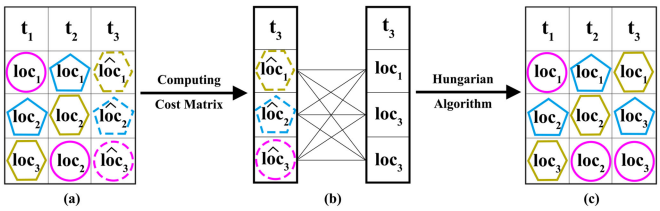


Fig. 2. Procedure of reconstruction attack on transformed mobility dataset.

feature of human mobility, and then the location in the *location-time* record P^{t+1} with the largest likelihood to the estimated location $\hat{p}_k^{t+1}$ will be chosen as the recovered next location, i.e., l_k^{t+1}. In the daytime, users move frequently, and their locations are continuous, which makes it possible to estimate the next location with the current location and the velocity. Formally, for the $k^{th}(1 \leq k \leq m)$ recovered trajectory, the estimated location is

$$\hat{p}_k^{t+1} = l_k^t + (l_k^t - l_k^{t-1}). \tag{3}$$

To quantify the likelihood between the estimated location and those in the *location-time* records, Fengli *et al.* [15] formulate the cost matrix $G^t = \{g_{i,j}^t\}_{m \times m}$, where $g_{i,j}^t$ is the distance between the estimated next location $\hat{p}_i^{t+1}$ and the actual location p_j^{t+1}.

Fig. 2 presents an intuitive demonstration to explain the procedure of recovering the trajectories from transformed *location-time* mobility dataset. In this example, an aggregated dataset with three possible locations and three time slots is shown. We assume that trajectories until time slot t_2 have been recovered, and then the estimations of the next locations are generated based on the continuity feature of mobility data as shown in Fig. 2(a). The distance between the estimated locations and those in *location-time* records (Fig. 2(b)) is formulated as the cost matrix. In the last step, *Hungarian algorithm* is applied to minimise the cost matrix and find each trajectory's associated location in the *location-time* record. Fig. 2(c) demonstrates the final recovered trajectories and those annotated with the same colour and shape belong to the same trajectory.

In our prototype implementation of the reconstruction attack for the experimental evaluation, we consider that the adversary, trying to recover users' trajectories from published aggregated mobility datasets, has some background knowledge of the target users. Generally, the adversary could have different kinds of background knowledge based on various sources such as social networks. However, in our specific attack model, to ease the presentation, we assume that the adversary has the target users' location information in the first two time slots as the background knowledge.

IV. Probabilistic Privacy-Preserving Publishing Framework

P4Mobi is composed of four major blocks: raw data preprocessing, sketch initialisation, mobility data encoding and population density distribution estimation. In this section, we first present the main modules and then theoretically quantify and analyse the privacy and utility of the published mobility dataset using the parameters of CMS, i.e., the number of hash functions d and the output range w of the hash functions.

A. *Framework of* P4Mobi

Fig. 3 presents the overall framework of *P4Mobi*.

1) Raw Data Preprocessing: Trajectories corresponding to raw mobility data can be discontinued or mislabeled, duplicated or altogether missing. This often leads to inconsistencies and poor utilities. Here we present the steps in preprocessing the mobility data in *P4Mobi*.

We assume that the whole covering area is divided into a grid of q blocks (in the example presented in Fig. 3, the value of q is 35) and the total duration of the dataset is divided into n time slots, i.e., $\{t_1, t_2, .., t_n\}$. During preprocessing, we need to determine the locations of the users at each time slot, the specific steps are: (1) each GPS record is assigned with a space block and a time slot according to its GPS reading and timestamp; (2) for the users having multiple different locations (space blocks) at the same time slot due to duplication or erroneous readings, we choose the space block having the highest occurrence frequency to represent the real location of the user at this time slot; (3) if a time slot of a user is vacant due to missing points, we use linear interpolation to determine its location. Fig. 3 (b) demonstrates the preprocessed mobility data records in a matrix-formation where $\{t_1, t_2, .., t_n\}$ are the time slots, $\{u_1, u_2, .., u_m\}$ are different users and each entry of the matrix is the location of a user at some specific time slot.

2) Sketch Initialisation: At the time slot t_i, we first create a $d \times w$ matrix, i.e., a sketch, and then initialise all the entries as zeros for future processing. The parameters d and w of the sketch refer to the number of hash functions and their output range, respectively. The choice of d and w determines the trade-off between privacy and utility of the final aggregated mobility dataset. In the initialisation step, we set the parameters d and w given the targeted utility level which is theoretically formulated using d and w in Section IV-B.

3) Mobility Data Encoding: After preprocessing, mobility data records are grouped into multiple sets of location points: $L_1, L_2, \ldots, L_i, \ldots, L_n$, where $L_i = [l_{i,1}, l_{i,2}, \ldots, l_{i,j}, \ldots, l_{i,m}]$ is a collection of all users' locations at time slot t_i. In mobility data encoding step, we use the set of locations L_i to update the sketch corresponding to time slot t_i which is initialised in the previous step. The mobility data is encoded into the sketch using d different hash functions which are chosen uniformly at random from a pairwise-independent family [1]. In details, for each element $l_{i,j} \in L_i$, we apply d hash functions on $l_{i,j}$ iteratively to find the position of the entry to be updated in the sketch. For example, the k^{th} hash function corresponds to the k^{th} row of the sketch and the output $h_k(l_{i,j})$ determines the column position, i.e., the entry at $(k, h_k(l_{i,j}))$ needs to be updated: the value of corresponding entry is incremented by 1. The above process can be formulated as,

$$\forall 1 \leq k \leq d : CMS[k, h_k(l_{i,j})] \leftarrow CMS[k, h_k(l_{i,j})] + 1, \tag{4}$$

where $CMS[k, h_k(l_{i,j})]$ is the value of entry at position $(k, h_k(l_{i,j}))$ of the sketch. As shown in Fig. 3 (d), a location point

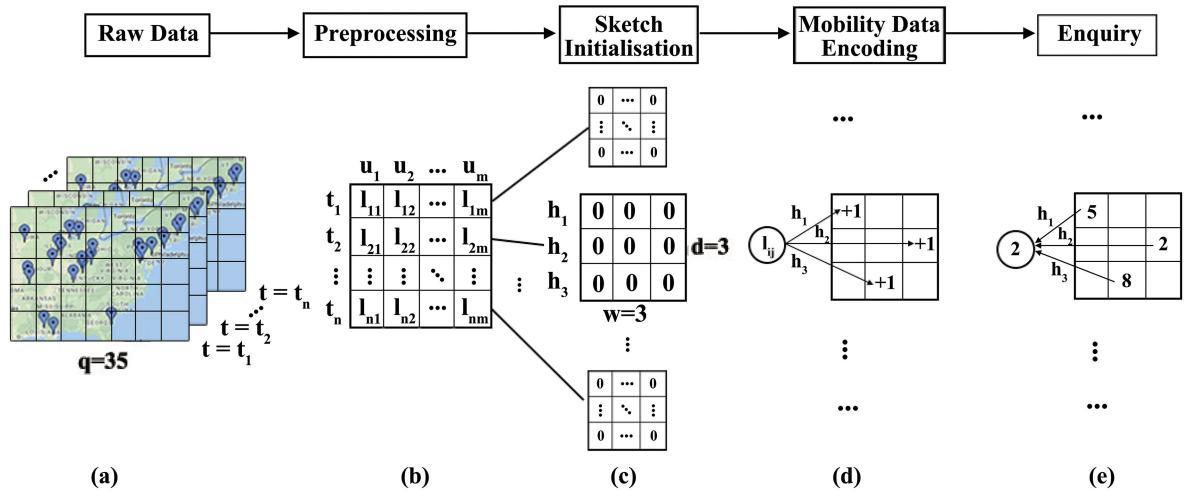


Fig. 3. Framework of *P4Mobi*.

$l_{i,j}$ is used to update multiple entries (one entry in each row) using multiple hash functions. After mobility data encoding, we obtain multiple sketches from the mobility dataset and each sketch corresponds to one specific time slot.

The procedure of encoding mobility data into sketches is essential for preserving privacy in the published mobility dataset. The targeted level of privacy and utility can be achieved through tuning the parameters d and w, or more intuitively, controlling the number of hash functions and their output range, respectively. We will discuss it in detail in Section IV-B.

4) Population Distribution Estimation Through Enquiry: At the final step, the sketches are used to estimate the population distribution at each time slot, through a mechanism termed as *enquiry*. We propose the enquiry function $enquiry(B_{j,i})$ to estimate the number of users at the block $B_j (1 \leq j \leq q)$ during time slot t_i. As shown in Fig. 3 (e), function $enquiry(B_{j,i})$ first finds all the entries related to the location of the block $B_{j,i}$ in the sketch based on the outputs of d hash functions used in the data encoding step. Then it compares the values of these entries to find the minimum to represent the estimated number of users at block $B_{j,i}$. Formally,

$$enquiry(B_{j,i}) = \min_{k=1}^{d} CMS[k, h_k(B_{j,i})]. \tag{5}$$

This enquiry step finds the minimum among the values of the entries corresponding to the same location (block), therefore, with larger d (or more hash functions), more candidate values are obtained when searching for the minimum. As the minimum is accepted, the final estimated result will be different from the ground truth if all of the d candidate values are larger than the ground truth value due to collisions.

B. Privacy and Utility Analysis of P4Mobi

Next we theoretically analyse the privacy and utility-preserving properties of our proposed *P4Mobi* framework and formulate how the parameters, i.e., the depth and width pair (d, w) (in other words, the number of hash functions and their output range) determine the privacy and utility performance of the published dataset.

We first consider two extreme scenarios with the same original mobility dataset consisting of $q\tilde{}(q > 1)$ location blocks and m users. In the first scenario, both parameters w and d are set to be 1, which indicates only one hash function is used and the output of the hash function is always 1 irrespective of the input location. Therefore the outputs of the enquiry function will also be the same for all location blocks, i.e., m. The published population distribution has no practical use in this scenario. While to the other extreme scenario, we set the parameters w and d the same as the number of locations q, which indicates each original location has its unique position in each row of the sketch (we assume no collision in this scenario). The population distribution generated by *P4Mobi* under this situation is the same as that resulting from simple aggregation method [14], which has high utility, but is significantly susceptible to the reconstruction attack proposed in [15].

From the privacy perspective, the manipulations of *P4Mobi* can be regarded as adding some *noise* into mobility datasets in terms of collisions, i.e., opportunistically mapping/encoding different physical locations into the same position of sketch. Collisions could lead the estimated population distribution deviating from the actual value when enquiring a location block using the encoded sketches. For example, in the first extreme scenario described above, the collision rate is maximum as all the physical locations are mapped to the same position in the sketch. The collision rate is zero in the second extreme scenario as all the input physical locations have their unique positions in the sketch. The probability of collisions decreases when w and d increase.

We now formulate the *Probability of Collisions (PoC)* of *P4Mobi* using the CMS parameters. We assume the number of distinct locations is q, the probability of the $i^{th} (1 \leq i \leq q)$ location loc_i in the mobility dataset being hash-mapped to any one position in one row of sketch is $1/w$, the *PoC* of the i^{th}

location $Pr_c(loc_i)$ is,

$$Pr_c(loc_i) = \left(1 - \left(1 - \frac{1}{w}\right)^{q-1}\right)^d, \forall i \in q, \tag{6}$$

where $(1 - \frac{1}{w})^{q-1}$ is the probability that the other $(q-1)$ locations (except loc_i) are not mapped to the position of loc_i in one row of sketch, which means loc_i is mapped to a position with value of zero (not occupied by other elements) in this row. The probability that loc_i is mapped to a position with non-zero value (i.e., a collision occurs) in this row of sketch is therefore $(1 - (1 - \frac{1}{w})^{q-1})$. In the final step of *P4Mobi*, we use $enquiry(B_{j,i})$ to find the minimum among values of the entries determined by the outputs of the d hash functions as the estimated number of users at location j in time slot i. Therefore, a false positive or collision (the minimum value is larger than the ground truth) only occurs when collision happens in every row of sketch, i.e., all independent d hash functions map loc_i to the positions with non-zero values and hence the *PoC* is $(1 - (1 - \frac{1}{w})^{q-1})^d$. In the scenario of mobility dataset, the number of distinct locations q for a certain time slot is not user defined but can be easily obtained from preprocessed raw data. We therefore consider the parameters w and d as the impact factors of PoC.

From the utility perspective, we aim for an estimation of the population distribution that is as accurate as possible. As in [1], we assume the actual population distribution of q locations is $C = \{c_1, c_2, \ldots, c_i, \ldots c_q\}$ and the corresponding estimation is $\hat{C} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_i, \ldots \hat{c}_q\}$. Given two small positive values θ and δ range in (0,1), we need to set the number of hash functions $d = \lceil ln(1/\delta) \rceil$ and the range of output $w = \lceil e/\theta \rceil$, where e is Euler's number, so that for all $c_i \in C$, there exist an estimation $\hat{c}_i$ satisfying

$$|\hat{c}_i - c_i| \leq \theta ||\hat{C}||_1, \tag{7}$$

with probability at least $1 - \delta$. Here $||\hat{C}||_1$ is ℓ_1 norm of the estimated population distribution, i.e. the sum of the estimated users numbers of all q locations. $|\hat{c}_i - c_i|$ is the difference between the estimated population and the ground truth, which can be regarded as the utility loss. Intuitively, the smaller the values of θ and δ are, the smaller the difference between the estimated population and the actual population (e.g., smaller utility loss) will be. Smaller θ and δ also indicate larger values for w and d of CMS.

In practice, as *P4Mobi* publishes the estimated population density distribution, i.e., the estimated number of users at each location, the PoC of *P4Mobi* can be computed as,

$$PoC = \frac{n_d}{n}, \tag{8}$$

where n_d is the number of locations in the published dataset with more users than ground truth, i.e., number of collisions, and n is the number of distinct locations in the published dataset. To compute the utility of the published dataset, we compare it with the actual population distribution. The utility of the processed dataset is defined as the proportion of locations which have the identical population estimation to the ground truth, i.e., without any collision after being encoded into CMS, so that

$$utility = \frac{n - n_d}{n}. \tag{9}$$

This leads to

$$utility = \frac{n - n_d}{n} = 1 - \frac{n_d}{n} = 1 - PoC, \tag{10}$$

where PoC is controlled by the parameters of CMS. Therefore, we can tune the parameters of CMS according to the utility requirement. In essence, this shows that the smaller the values of w and d, the larger the PoC (and the lesser the utility). Increasing w and d leads to smaller PoC and better utility of the released dataset.

Comparing to the state-of-the-art method based on the widely used notion of differential privacy, our proposed *P4Mobi* seems to be a little deficient in theoretical depth. However, the experimental results shown in Section V demonstrate that *P4Mobi* outperforms the differential privacy based method in terms of privacy against the recently shown attack on mobility data and utility. Besides, we also note that the threat model considered in this paper is that the server is trusted while the querier is an untrusted third party. In this setting, we show that the privacy guarantees provided by probability of collisions are sufficient for aggregated counts. In the future, we plan to extend our approach for the threat model of untrusted server as well by providing privacy guarantees for users' input/updates to the server by using local differential privacy, which is similar to approaches [21] and [22], introduced by Google and Apple, respectively.

V. Evaluation

We now evaluate the performance of *P4Mobi*, first against the threat model described in Section III, then we demonstrate its performance on estimating the population density distribution of mobility datasets (utility). We also compare with other mobility dataset aggregation approaches on the trade-off between privacy and utility (under the threat model we consider in this paper) and finally we empirically validate the accuracy of the theoretical utility formulation or Probability of Collisions (PoC) on estimating the practical utility in real world datasets.

A. Evaluation Metrics

We use the *Reconstruction error* to quantify the robustness of *P4Mobi* under the reconstruction attacks. *Reconstruction error* is defined as the Euclidean distance between the recovered users' trajectories and the ground truth. Larger *reconstruction errors* (distances) indicate less privacy leakage and better privacy protection.

We propose *L-Jaccard Index*, a location-based transformation of *Jaccard Index* [23], to calculate the similarity between the published datasets processed by *P4Mobi* or DP-MDA and the ground truth as the metric of *utility* of the processed datasets. The results demonstrate how well the published dataset supports higher-level LBS applications for analysis and decision making by exhibiting higher utility preserved in the published datasets.

1) Reconstruction Error: The reconstruction attacks on the aggregated mobility datasets aim to infer users' private daily routines by recovering their mobile trajectories. We use the relative value of the average Euclidean distance (*reconstruction error*) between the recovered trajectories and the ground truth corresponding to the length of the whole working area (50km in our datasets) as a measure of the performance on privacy-preservation. To compute the *reconstruction error*, we first uniquely pair the recovered trajectories with the most similar ground truth trajectories in terms of Euclidean distance between vectors of trajectories. We use a greedy-based heuristic algorithm [15] to achieve effective and efficient pairing. We then compute the average value of all the Euclidean distances between the paired trajectories and use its proportion to the scale of the working area ($50\tilde{}km$) as the *reconstruction error*.

Formally, we denote the ground truth trajectory of the i^{th} $(1 \leq i \leq m)$ user as $s_i = [l_i^1, l_i^2, \ldots, l_i^n]$, where each element represents the location of the user at the specific time slot and the corresponding recovered trajectory as $\hat{s}_i = [\hat{l}_i^1, \hat{l}_i^2, \ldots, \hat{l}_i^n]$, where n is the total number of time slots, m is the total number of users in the mobility dataset and the scale of the working area is $50\tilde{}km$. The *reconstruction error* E_r can be computed as,

$$E_r = \frac{\sum_{i=1}^{m} ||(\hat{s}_i - s_i)||_2^2}{50 \times m}, \tag{11}$$

where $||(\hat{s}_i - s_i)||_2^2$ is the Euclidean distance between the vectors of recovered and ground truth trajectories. Larger *reconstruction error* E_r indicates better privacy protection performance against the threat model.

2) Utility: As mentioned earlier, S-MDA simply counts the statistics of the population density distribution without injecting any additional noise. We treat it as having no *utility* loss. *P4Mobi* and DP-MDA inject collisions and random noise, respectively, during dataset aggregation, however, resulting in a loss of *utility*.

We propose *L-Jaccard Index*, which is a location-based transformation of *Jaccard Index* [23], to quantify the *utility* of the estimated datasets. The original *Jaccard Index* computes the proportion of the intersections between different datasets for comparing their similarity. In our paper, however, the application scenario is mobile population density, where the actual location information and the corresponding population density are both important features in supporting LBSs applications. The location information is ignored when we employ the original *Jaccard Index* to compare the similarity between different datasets, so we propose *L-Jaccard Index*, which considers both the location information and population density to quantify the similarity between population distributions in two mobility datasets by extending the original *Jaccard Index*.

We calculate the *L-Jaccard Index* between the estimated and real population density distributions by comparing the density (count) values appended with the corresponding location information. Formally, We denote the actual population density distribution of q locations from S-MDA at time slot t_i $(1 \leq i \leq n)$ as $C^i = [c_1^i, c_2^i, \ldots, c_q^i]$ and the estimated ones from *P4Mobi* or DP-MDA as $\hat{C}^i = [\hat{c}_1^i, \hat{c}_2^i, \ldots, \hat{c}_q^i]$, where each element is the actual/ estimated number of users at some specific location. The *L-Jaccard Index* between C^i and $\hat{C}^i$ can be computed as,

$$L\_JI(C^i, \hat{C}^i) = \frac{iden(C^i, \hat{C}^i)}{q}, \tag{12}$$

where $iden(C^i, \hat{C}^i)$ returns the number of elements in C^i and $\hat{C}^i$ having identical location information and population density. The *L-Jaccard Index* measures the similarity between the outputs of *P4Mobi* or DP-MDA and S-MDA. Larger value of *L-Jaccard Index* indicates higher *utility*. We compute the *utility* of the published datasets as,

$$Utility = \frac{1}{n} \sum_{i=1}^{n} L\_JI(C^i, \hat{C}^i), \tag{13}$$

where n is the number of time slots in the dataset. Intuitively, higher *utility* value corresponds to more accurate population density distributions comparing with S-MDA (or ground truth) and implies better practical usability.

B. Mobility Datasets

We use two real-world mobility datasets in our empirical evaluation to demonstrate the performance of *P4Mobi* on privacy and utility-preserving properties for publishing aggregated mobility datasets.

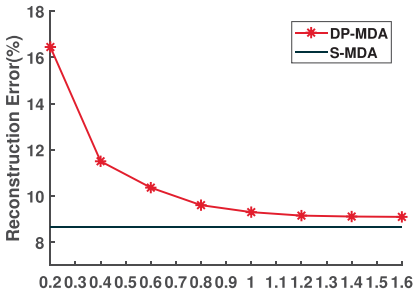
1) MoMo Mobile App Dataset (MoMo): MoMo dataset [24] has been collected from the GPS of mobile devices using a very popular social network application, MoMo, from 21 May, 2012 to 26 June, 2012, in Beijing, China. It contains approximately 3.89 million users' check-in trajectories. Each record consists of four attributes: user ID, timestamp, latitude and longitude.

2) San Francisco Cabs Dataset (SFC): SFC dataset [25] contains mobility traces of taxi cabs in San Francisco, USA. It includes GPS coordinates of approximately 500 taxis collected over 30 days in the San Francisco bay area. Each record has four attributes: cab ID, timestamp, latitude and longitude.

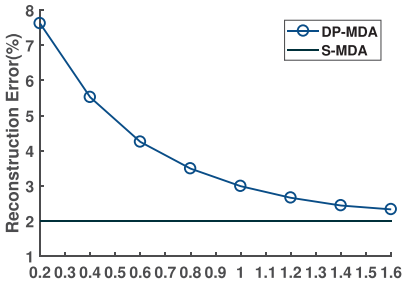
3) Dataset Preprocessing for Metadata: Our two real-world datasets contain very dense spatio-temporal records of each individual's mobility trajectories. To obtain the metadata for evaluation purpose, we trim both datasets. As during nighttime, most individuals tend to not move too much, the mobility data does not provide significant information to recover. Therefore, we set the working period as from 9:00 to 17:00, and the size of the whole working area is 50km $\times$ 50km. Discussion in [15] indicates that the reconstruction algorithm is robust under different spatial and temporal resolutions. We choose the size of the space blocks to a moderate value of 2km in our evaluation. According to different application scenarios (e.g., travelling speed) we set the duration of the time slots for the two datasets as 30 minutes for MoMo and 2 minutes for SFC, respectively, because on average cabs travel significantly more dynamic and faster than human users. Then the corresponding number of time slots are 16 and 240, respectively. An overview of the two datasets is shown in Table I.

TABLE I
METADATA OF THE TWO DATASETS

Dataset	Users	Spacial Resolution	Locations	Temporal Resolution
MoMo	198	2km	625	30 minutes
SFC	127	2km	625	2 minutes



(a)

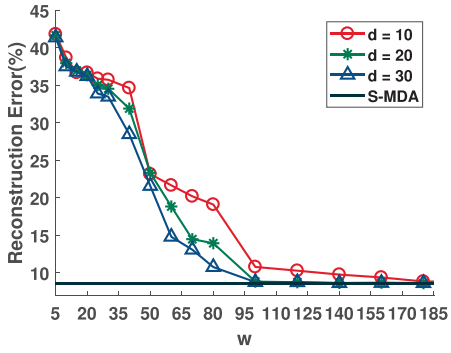


(b)

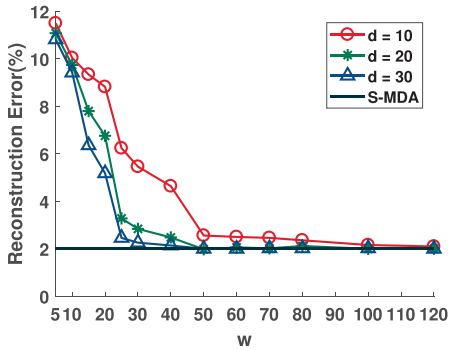
Fig. 4. The reconstruction error of DP-MDA under different noise factor ϵ. (a) MoMo dataset. (b) SFC dataset.

C. Resilience Against the Reconstruction Attack

We first evaluate the *reconstruction errors* obtained for S-MDA as a baseline. We apply the reconstruction algorithm on the aggregated results of both MoMo and SFC datasets processed by S-MDA and then compute the average Euclidean distance between the reconstructed and ground truth trajectories. The average *reconstruction errors* for MoMo and SFC datasets are 8.64% and 2%, respectively. The *reconstruction errors* from S-MDA are represented as horizontal lines in Fig. 4 and Fig. 5 to benchmark the performance of the other two approaches. We then evaluate the privacy-preserving performance of DP-MDA against reconstruction attacks. Results are shown in Fig. 4. The x-axis stands for the value of factor ϵ controlling the added Laplacian noise while the y-axis stands for the *reconstruction error* in percentage. The higher *reconstruction error* indicates better performance on privacy-preservation. We vary ϵ from 0.2 to 1.6 and compute the corresponding *reconstruction errors*. The results show that, in general, compared to the S-MDA, DP-MDA achieves significantly better performance on both datasets for all parameter (ϵ) values. Specifically, when $\epsilon < 0.6$, DP-MDA outperforms the benchmark by at least 2%, which is interpreted as at least 1000 meters in real-world. This is not suprising as a lower ϵ leads to a higher noise ratio, which makes the aggregated results more difficult to be reconstructed.



(a)



(b)

Fig. 5. The reconstruction error of *P4Mobi* on different size of sketch. (a) MoMo dataset. (b) SFC dataset.

As discussed in Section IV-B, the privacy level of *P4Mobi* can be controlled by tuning the number of hash functions d and their output range w. We gradually change the output range w from 5 to 180 with a different number of hash functions $d = [10, 20, 30]$ and compute the corresponding *reconstruction errors*. The results on two different datasets are shown in Fig. 5. Compared with S-MDA that produce average *reconstruction errors* of 8.64% for MoMo and 2% for SFC datasets, *P4Mobi* yields higher *reconstruction errors* when $w < 95$ for all settings of d. The performance gain grows rapidly when w decreases below 50. For example, for the MoMo dataset, when we fix $d = 30$ and $w = 50$, the corresponding *reconstruction error* is over 21% which is significantly higher than S-MDA (8.64%). The difference stands for 6.5km in real-world scenario.

We cannot compare the performance of *P4Mobi* and DP-MDA directly at this stage because they adopt a different set of parameters. However, intuitively, *P4Mobi* seems to achieve significantly better performance when preserving privacy than DP-MDA (comparing the value range of the y-axes in Fig. 4 and Fig. 5). To compare *P4Mobi* with other methods more comprehensively, in the next section, we consider the utility-preservation performance of these approaches.

D. Evaluation on the Utility-Preservation

This set of evaluations aim to validate the utility of the DP-MDA and *P4Mobi* approaches when computing or estimating

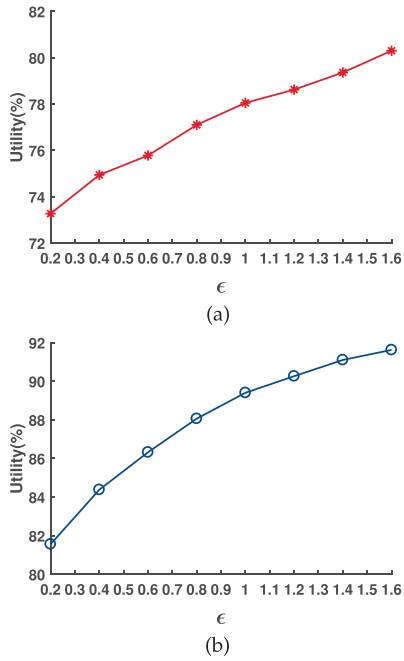


Fig. 6. Utility Performance of DP-MDA. (a) MoMo dataset. (b) SFC dataset

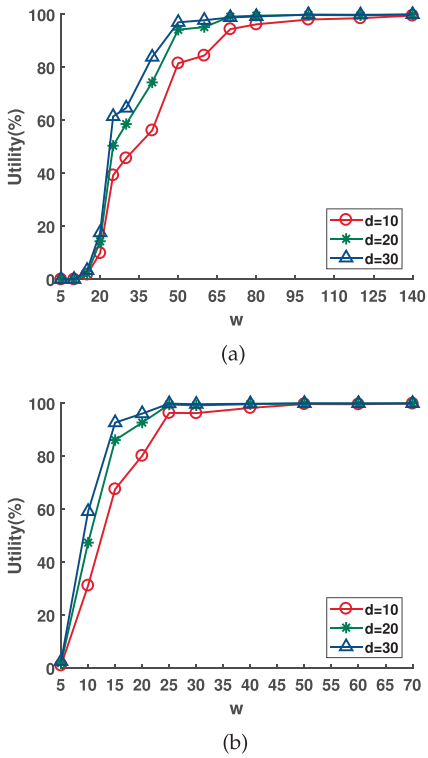


Fig. 7. Utility Performance of *P4Mobi*. (a) MoMo dataset (b) SFC dataset

the population density distribution of the published mobility datasets.

The evaluation of the utility metric is presented in Fig. 6. The x-axis stands for the value of ϵ that controls the rate of Laplacian noise added and a smaller ϵ indicates a higher noise ratio. The y-axis stands for the average *utility* preserved by DP-MDA in percentage. We vary ϵ from 0.2 to 1.6 and compute the corresponding *utility*. We observe that the larger ϵ, the higher the similarity between the two density distribution. For instance, in Fig. 6, when ϵ is 1.6, the *utility* of DP-MDA is close to 80% and 91% for MoMo and SFC, respectively. We then evaluate the utility obtained by *P4Mobi* under different parameter settings. The results are shown in Fig. 7. Again, we gradually change the value of w from 5 to 180 when $d = [10, 20, 30]$ and compute the corresponding *utility*. The results in Fig. 7 show that for both datasets, the *utility* increases steeply with the growing value of w for all three values of d when $w < 50$. For example, in Fig. 7(a), when we set $d = 30$ and increase w from 5 to 50, the *utility* grows sharply from 0% to 95%. On the other side, when w is fixed, a larger d leads to a higher *utility*.

E. Privacy-Utility Trade-Off Comparison

We further investigate and compare the performance of *P4Mobi* and DP-MDA on the trade-off between privacy and utility-preserving which is evaluated in terms of *reconstruction error* and *utility*. In the evaluation, we set $d = 30$ and change the value of w to compute the corresponding *reconstruction error* and *utility* of *P4Mobi*. The results of the evaluations on two real-world datasets are shown in Fig. 8, where the x-axis stands for the *reconstruction error* and the y-axis is the *utility*. As discussed before, higher values for both *reconstruction error*

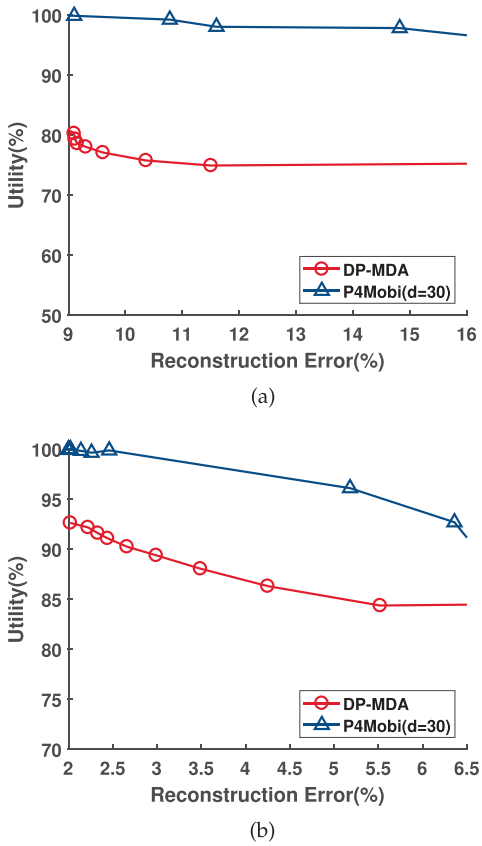


Fig. 8. Comparison of *P4Mobi* and DP-MDA over Privacy-Utility Trade-off. a) MoMo dataset (b) SFC dataset

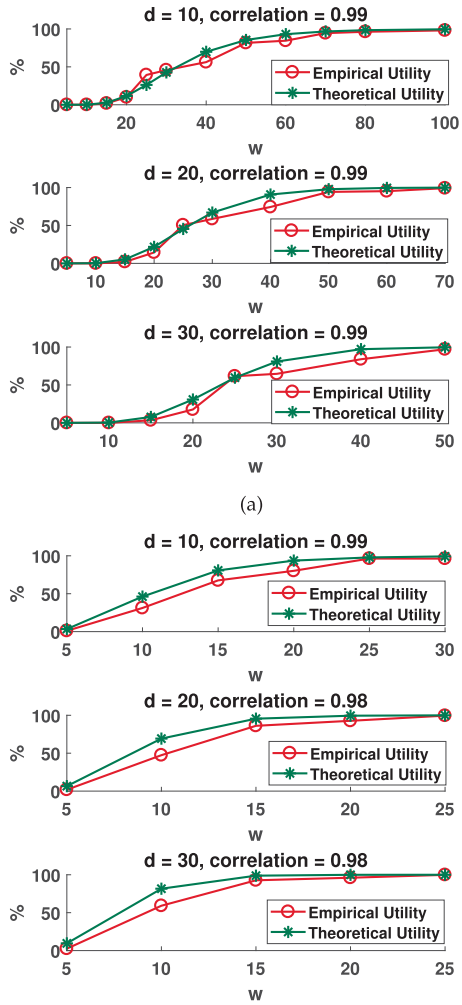
and *utility* (higher privacy and practical utility) are desired. From the results shown in Fig. 8, we can observe that, with the same *reconstruction error*, *P4Mobi* achieves up to 20% and 10% higher $utility$ than that of DP-MDA on MoMo and SFC datasets, respectively. For example, as the results shown in Fig. 8(a), even when the *reconstruction error* (refers to privacy-preserving) is as high as 16% (which is almost twice of the benchmark, S-MDA, i.e., 8.64% for MoMo dataset), *P4Mobi* still preserves more than 95% *utility*, which is only 75% for DP-MDA with the same *reconstruction error*. Similar results are obtained for *P4Mobi* with $d = 10$ and $d = 20$. We do not provide these results in this paper due to space limitation.

At last, we conclude the comparison of different privacy-preserving mobility dataset publishing approaches by presenting some examples of the published population density maps from the three approaches with two different mobility datasets in Fig. 10. We use colours with different gradations to present the different numbers of users on the map, and the lighter the colour is, the larger the number of users in that location is. The examples shown on the left column are from S-MDA, which are regarded as accurate benchmark of the population density distribution, i.e., the *utility* is 100%. Examples from DP-MDA are in the middle while those from *P4Mobi* are on the right. By comparing the maps across columns, an intuitive observation is that the density maps from *P4Mobi* are significantly more similar to the benchmark (S-MDA) compared with DP-MDA (98% v.s. 75% for MoMo and 90% v.s. 84.5% for SFC) meanwhile they achieve better privacy-preserving against the attack model (*reconstruction error:* 21% v.s. 11.5% for MoMo and 6.8% v.s. 5.4% for SFC).

F. Evaluation on Utility Estimation

We propose the theoretical formulation of the utility derived from *PoC* in Section IV-B to estimate the practical utility of population distributions to be published, so that the followers are able to determine the suitable CMS parameter settings conveniently according to their utility requirements. In this section, we evaluate the consistency of the theoretical and empirical utility of *P4Mobi* under different parameter settings.

The evaluation results are presented in Fig. 9. Across different parameter settings, we observe that the tendency of theoretical utility (calculated using Equation 6 and Equation 10) is always consistent with the empirical results, and as expected, with the same d, utility increases with the increase of w. To provide more concrete evidence, we also compute the correlation between the curves of theoretical and empirical utility with different values of d. The correlation results are over 98% which indicate a close matching between empirical and theoretical utility.



(a)

(b)

Fig. 9. Theoretical and empirical utility of *P4Mobi*. (a) MoMo dataset (b) SFC dataset

VI. Related Works

Location-based services (LBSs) support human daily life by studying mobility patterns from trillions of trails and footprints [26]. Urban planning [27], traffic forecasting [28], market campaign [29], prediction of epidemics [30] and designing of mobile network protocols [31] are all powered by citizen's trajectories. Such services not only bring convenience to people's life, but also followed by privacy issues towards individual users.

To address the privacy issues in releasing mobility datasets for empowering LBSs. Some researches focus on the raw mobility data, i.e. encrypting or encoding the trajectory records before releasing. Laplacian noise is used in [16] to publish geo-indistinguishable mobility datasets that achieve ϵ-differential privacy. In this work, noise is added drawn from Laplace distribution to satisfy differential privacy and the location of each user is re-mapped. Another method is proposed in [32] to combine differential privacy and k-anonymisation by adding a random sampling step before performing "secure" k-anonymisation for publishing mobility data. Zhang *et al.* [33]proposed a dual-K mechanism, which inserts multiple anonymizers between the user and LBSs to protect the user's trajectory privacy. Markov model is utilized in [34] to predict the next query location according to the user mobility and form spatial K-anonymity to enhance user location privacy. Direct operations on the raw mobility data impacts the utility of the published datasets, and the trade-off between utility and privacy is a common challenge for these techniques.

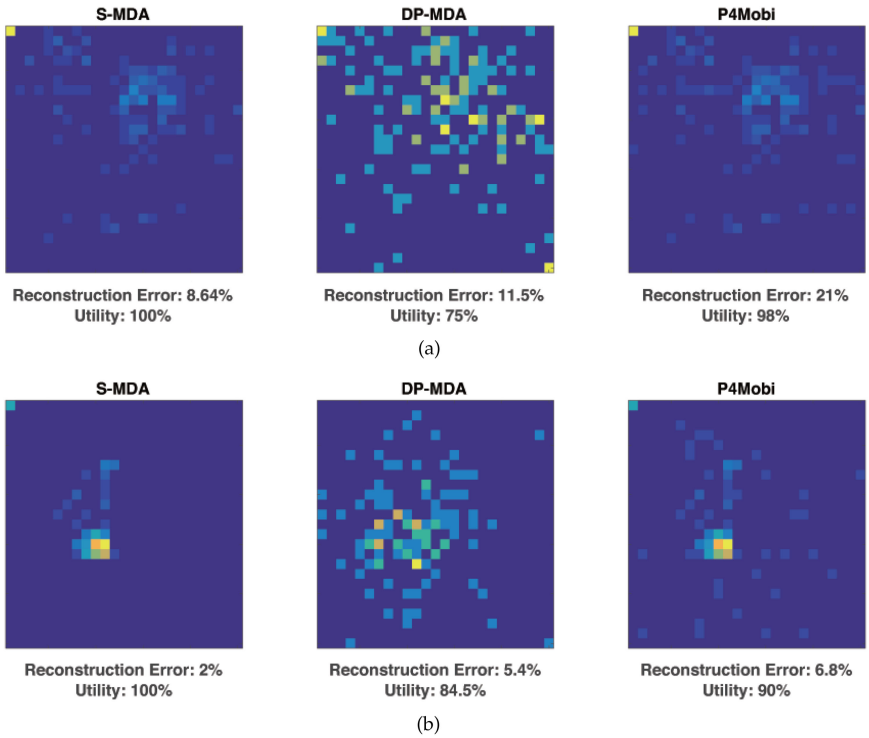


Fig. 10. Examples of population density distributions obtained from S-MDA (left column), DP-MDA (middle column) and *P4Mobi* (right column). (a) MoMo dataset. (b) SFC dataset

Some researches focus on the statistics of mobility datasets. They protect the users' trajectoy records by releasing the statistical properties. For instance, [14] introduces the French XData project that only reports the population density of each region. Mobile operators in China share the real-time counts of mobile users at specific locations with some real estate companies [35]. Ma *et al.* [36] proposed a mechanism called RPTR, which protects a vehicle's real-time trajectory data release. It samples the vehicles density distribution data, and predicts the next release based on the previous sampled data. Fan *et al.* [37] proposed a sampling processing method based on Kalman filter to obtain a trade-off between the privacy degree and system performance, which allows differentially private aggregate sharing and time-series analysis. Yang *et al.* [38] proposed $l-$trajectory privacy to achieve user-level privacy in each l length trajectory statistics, they made the privacy budget for each user assigned to their l-length trajectory. k^m-anonymity and p-confidentiality are introduced in [39] to protect the privacy of population density. All of these approaches protect the users' privacy based on statistic properties of the raw trajectory data, which could incur privacy issues if the statistic properties are revealed.

VII. Conclusion

In this paper, we propose ***P4Mobi***, a **P**robabilistic **P**rivacy-**P**reserving **P**ublishing framework for **Mobi**lity datasets. *P4Mobi* is designed by facilitating the efficient data structure Count-Min Sketch (CMS). It publishes the estimated population distribution derived from the mobility datasets to support LBSs with tunable utility. As our extensive evaluations on two different typical mobility datasets have shown, *P4Mobi* achieves up to 20% and 10% higher utility, respectively, with the same reconstruction error premise against reconstruction attacks on aggregated datasets, comparing with the state-of-the-art mobility dataset aggregation and privacy-preserving approaches. At last, we use the two real-world datasets to validate that the theoretical formulation of utility is able to predict the practical utility of the dataset processed by *P4Mobi* under different parameter settings accurately which enables the followers choosing parameters according to their required utility without processing the datasets repeatedly to find the suitable settings.

As future work, we aim to investigate other probabilistic data structures, such as counting Bloom filters [40] and Cuckoo filters [41], for privacy-preserving publishing of mobility datasets. Another direction of our future work is to study other advanced reconstruction attacks and evaluate the performance of our approach in resilient against those threat models. Last but not least, considering the privacy concerns during the communication between individual devices and mobility data collectors (such as the mobile operators), we will extend the application of our approach to the individual level, which will be implemented on the individual's devices (such as smartphones, smart wearables and other internet-enabled devices) to protect individuals' mobility data before being collected.

References

[1] G. Cormode and S. Muthukrishnan, "An improved data stream summary: The count-min sketch and its applications," *J. Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.
[2] A. Küpper, "Location-based services," *Fundamental and Operation*. Hoboken, NJ, USA: Wiley, 2005.

[3] I. Bisio, F. Lavagetto, M. Marchese, and A. Sciarrone, "GPS/HPS-and Wi-Fi fingerprint-based location recognition for check-in applications over smartphones in cloud-based LBSs," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 858–869, Jun. 2013.
[4] R. Ferraro and M. Aktihanoglu, *Location-Aware Applications*. Manning Publications Co., 4849:377–383, 2011.
[5] F.-M. Hsu, Y.-T. Lin, and T.-K. Ho, "Design and implementation of an intelligent recommendation system for tourist attractions: The integration of EBM model, Bayesian network and Google Maps," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3257–3264, 2012.
[6] M. Wojnarski, P. Gora, M. Szczuka, H. S. Nguyen, J. Swietlicka, and D. Zeinalipour, "IEEE ICDM 2010 contest: Tomtom traffic prediction for intelligent GPS navigation," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2010, pp. 1372–1376.
[7] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proc. 20th Int. Conf. Adv. Geographic Inf. Syst.*, 2012, pp. 199–208.
[8] A. Kofod-Petersen, P. A. Gransaether, and J. Krogstie, "An empirical investigation of attitude towards location-aware social network service," *Int. J. Mobile Commun.*, vol. 8, no. 1, pp. 53–70, 2009.
[9] H. Wang, M. Terrovitis, and N. Mamoulis, "Location recommendation in location-based social networks using user check-in data," in *Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2013, pp. 374–383.
[10] J. Krumm, "Inference attacks on location tracks," in *Proc. Int. Conf. Pervasive Comput.*, 2007, pp. 127–143.
[11] V. Pandurangan, "On taxis and rainbows lessons from NYCs improperly anonymized taxi logs," Jun. 2014. [Online]. Available: https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1
[12] M. Zarem, E. Vuillermet, and J. DeAguiar, "Intelligent reverse geocoding," May 20 2014, uS Patent 8,731,585.
[13] S. Gambs, M.-O. Killijian, and M. N. del P. Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, 2014.
[14] G. Acs and C. Castelluccia, "A case study: Privacy preserving release of spatio-temporal density in paris," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1679–1688.
[15] 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 *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1241–1250.
[16] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2013, pp. 901–914.
[17] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.
[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptography Conf.*, Springer, 2006, vol. 3876, pp. 265–284.
[19] G. Dantzig, *Linear Programming and Extensions*. Princeton, NJ, USA: Princeton Univ. Press, 2016.
[20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics*, vol. 2, no. 1-2, pp. 83–97, 1955.
[21] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun Secur*, 2014, pp. 1054–1067.
[22] D. Team *et al.*, "Learning with privacy at scale," 2017. [Online]. Available: https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html
[23] L. Hamer *et al.*, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula," *Inf. Process. Manage.*, vol. 25, no. 3, pp. 315–318, 1989.
[24] T. Chen, M. A. Kaafar, and R. Boreli, "The where and when of finding new friends: Analysis of a location-based social discovery network," in *Proc. 7th Int Conf Weblogs Social Media*, 2013, pp. 61–70.
[25] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset epfl/mobility (v. 2009-02-24)," Feb. 2009. [Online]. Available: https://crawdad.org/epfl/mobility/20090224
[26] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu, "The scaling of human mobility by taxis is exponential," *Physica A: Statist. Mech. Appl.*, vol. 391, no. 5, pp. 2135–2144, 2012.
[27] H. D. Rozenfeld, D. Rybski, J. S. Andrade, M. Batty, H. E. Stanley, and H. A. Makse, "Laws of population growth," *Proc. Nat. Acad. Sci.*, vol. 105, no. 48, pp. 18 702–18 707, 2008.
[28] B. Jiang, J. Yin, and S. Zhao, "Characterizing the human mobility pattern in a large street network," *Phys. Rev. E*, vol. 80, no. 2, 2009, Art. no. 021136.
[29] E. Agliari, R. Burioni, D. Cassi, and F. Maria Neri, "Word-of-mouth and dynamical inhomogeneous markets: An efficiency measure and optimal sampling policies for the pre-launch stage," *IMA J. Manage. Math.*, vol. 21, no. 1, pp. 67–83, 2009.
[30] L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 42, pp. 15 124–15 129, 2004.
[31] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *Proc. INFOCOM*, 2009, pp. 855–863.
[32] N. Li, W. H. Qardaji, and D. Su, "Provably private data anonymization: Or, k-anonymity meets differential privacy," *CoRR*, vol. abs/1101.2604, 2011. [Online]. Available: http://arxiv.org/abs/1101.2604
[33] 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. 427, pp. 406–419, 2019.
[34] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, "A caching and spatial k-anonymity driven privacy enhancement scheme in continuous location-based services," *Future Gener. Comput. Syst.*, vol. 94, pp. 40–50, 2019.
[35] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-Preserving Data Mining*. Berlin, Germany: Springer, 2008, pp. 11–52.
[36] 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.
[37] L. Fan, L. Xiong, and V. Sunderam, "Fast: Differentially private real-time aggregate monitor with filtering and adaptive sampling," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 1065–1068.
[38] Y. Cao and M. Yoshikawa, "Differentially private real-time data release over infinite trajectory streams," in *Proc. 16th IEEE Int. Conf. Mobile Data Manage.*, 2015, vol. 2, pp. 68–73.
[39] G. Acs, G. Biczók, and C. Castelluccia, "Privacy-preserving release of spatio-temporal density," in *Handbook of Mobile Data Privacy*. Berlin, Germany: Springer, 2018, pp. 307–335.
[40] F. Bonomi, M. Mitzenmacher, R. Panigrahy, S. Singh, and G. Varghese, "An improved construction for counting bloom filters," in *Proc. Eur. Symp. Algorithms*, 2006, pp. 684–695.
[41] B. Fan, D. G. Andersen, M. Kaminsky, and M. D. Mitzenmacher, "Cuckoo filter: Practically better than bloom," in *Proc. 10th ACM Int. Conf. Emerg. Netw. Experiments Technol.*, 2014, pp. 75–88.

Qing Yang received the bachelor's degree in software engineering from HEU. She is currently working toward the Ph.D. degree at the College of Computer Science and Technology, Harbin Engineering University(HEU). She was a Visiting Student in the Commonwealth Scientific and Industrial Organization(CSIRO) in Australia. Her main research interests are wearable/ mobile computing and privacy preserving in mobile sensor networks.

Yiran Shen (Senior Member) received the Ph.D. degree in computer science and engineering from the University of New South Wales. He is an Associate Professor in the College of Computer Science and Technology, Harbin Engineering University(HEU). He publishes regularly at top-tier conferences and journals. His current research interests are wearable/ mobile computing, wireless sensor networks and applications of compressive sensing.

Dinusha Vatsalan is a Research Scientist at Data61-CSIRO, Australia, and an Honorary Lecturer in the Research School of Computer Science at the Australian National University. Her research interests are mainly in privacy preserving techniques, privacy in data matching and mining, privacy in social media, privacy risk evaluation and prediction, health informatics, and population informatics. She is currently working on privacy in web search and social media as well as privacy preserving data matching.

Jianpei Zhang received the Ph.D. degree from Harbin Engineering University. He is the Professor with College of Computer Science and Technology, Harbin Engineering University(HEU). He is also the Leader of Software and Social Computing Research Group. His research interests include data mining, database systems and software engineering.

Mohamed Ali (Dali) Kaafar (Senior Member) is the Executive Director of the Optus Macquarie University Cyber Security Hub and Professor at the Faculty of Sciences and Engineering in Macquarie University. He is also the recipient of the CAS President's International Fellowship award and is a Visiting Professor at the Computer Network Information Center (CNIC). He was a Research Leader with National ICT Australia (NICTA), Sydney, Australia, a Research Scientist with INRIA Rhone-Alpes, Grenoble, France and the Group Leader of the Information Security and Privacy Group with Data61, CSIRO Australia. His research interests include cyber security, privacy preserving technologies, and networks measurement.

Wen Hu (Senior Member) is a Senior Lecturer at School of Computer Science and Engineering, the University of New South Wales(UNSW). Much of his research career has focused on the novel applications, low-power communications, security and compressive sensing in sensor network systems and Internet of Things(IoT). He published regularly in the top rated sensor network and mobile computing venues. He is a Senior Member of ACM.