Sun S, Ma S, Song JH *et al.* Experiments and analyses of anonymization mechanisms for trajectory data publishing. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 37(5): 1026–1048 Sept. 2022. DOI 10.1007/s11390-022-2409-x

Experiments and Analyses of Anonymization Mechanisms for Trajectory Data Publishing

She Sun[1](孙 设), Shuai Ma[1,*](马 帅), *Senior Member*, *CCF*, *IEEE*, *Member*, *ACM*, Jing-He Song[1](宋景和) Wen-Hai Yue[1](岳文海), Xue-Lian Lin[1,*](林学练), *Member*, *CCF*, and Tiejun Ma[2](马铁军)

[1] *State Key Laboratory of Software Development Environment, School of Computer Science and Engineering Beihang University, Beijing 100191, China*

[2] *Department of Decision Analytics and Risk, Southampton Business School, University of Southampton Southampton SO17 1BJ, U.K.*

E-mail: {sunshe, mashuai, songjh, yuewh, linxl}@buaa.edu.cn; tiejun.ma@ed.ac.uk

Received April 13, 2022; accepted September 21, 2022.

Abstract With the advancing of location-detection technologies and the increasing popularity of mobile phones and other location-aware devices, trajectory data is continuously growing. While large-scale trajectories provide opportunities for various applications, the locations in trajectories pose a threat to individual privacy. Recently, there has been an interesting debate on the reidentifiability of individuals in the Science magazine. The main finding of Sánchez *et al.* is exactly opposite to that of De Montjoye *et al.*, which raises the first question: "what is the true situation of the privacy preservation for trajectories in terms of reidentification?" Furthermore, it is known that anonymization typically causes a decline of data utility, and anonymization mechanisms need to consider the trade-off between privacy and utility. This raises the second question: "what is the true situation of the utility of anonymized trajectories?" To answer these two questions, we conduct a systematic experimental study, using three real-life trajectory datasets, five existing anonymization mechanisms (i.e., identifier anonymization, grid-based anonymization, dummy trajectories, k-anonymity and ε-differential privacy), and two practical applications (i.e., travel time estimation and window range queries). Our findings reveal the true situation of the privacy preservation for trajectories in terms of reidentification and the true situation of the utility of anonymized trajectories, and essentially close the debate between De Montjoye *et al.* and Sánchez *et al.* To the best of our knowledge, this study is among the first systematic evaluation and analysis of anonymized trajectories on the individual privacy in terms of unicity and on the utility in terms of practical applications.

Keywords anonymization, privacy, reidentification, trajectory, utility

1 Introduction

With the advancing of location-detection technologies (e.g., global positioning systems, cellular networks, Wi-Fi and radio frequency identification) and the increasing popularity of mobile phones and other location-aware devices (e.g., smart-phones, on-board diagnostics, personal navigation devices and wearable smart devices), the trajectory data left by moving objects and daily collected is continuously growing [1–7]. Large-scale trajectories have provided opportunities to fundamentally transform the ways of disease fighting (e.g., COVID-19 pandemic tracking) and urban computing (e.g., traffic analysis and route planning) [1, 2, 7].

The location information in trajectories poses a threat to the privacy of individuals, and it belongs to the privacy and rights stated by General Data Protection Regulation (GDPR)①, which was considered as the most important law in data privacy regulation in

Regular Paper

Special Section on Scalable Data Science

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61925203 and 62172024, and Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

*Corresponding Author (Shuai Ma contributed the key ideas and Xue-Lian Lin was in charge of the experiments.)

①https://gdpr-info.eu/, Aug. 2022.

the past 20 years [8]. For example, a doctor's location can be derived through the correlation of the fine-grain location data with publicly available information [9], and the location data can reveal the habits and sexual preference of individuals that can be abused for unauthorized advertisements [10]. Even worse, personal information is available publicly and globally, as pointed out in GDPR, and the need for data publishing and privacy protection co-exists in various situations [11]. Efforts have also been made to develop anonymization mechanisms for protecting the privacy of trajectories over the past few decades [11, 12], such as simple identifier anonymization [13], grid-based generalization [12, 14], dummy trajectories [12, 15], k-anonymity [10, 16, 17] and differential privacy [18–20].

Recently, there has been an interesting debate on the reidentifiability of individuals in the Science magazine [2, 3, 21]. De Montjoye *et al.* [2] studied the credit card records of three months for 1.1 million users in 10 000 shops, and claimed that "four spatio-temporal points are enough to uniquely reidentify 90% of individuals", where credit card records are essentially treated as trajectories, reidentification is measured by unicity [1], and simple identifier anonymization and grid-based generalization are used as the anonymization mechanisms. This confirms the finding of their earlier work [1], "four spatio-temporal points are enough to uniquely identify 95% of the individuals", which studies the human mobility data of 15 months for 0.5 million individuals, where the location of an individual is specified hourly, and has a spatial resolution equal to that given by the carrier's antennas. Xiao *et al.* [22] studied two large sets of taxi trajectories in Shenzhen and Shanghai in China, and further found that "four spatio-temporal points are sufficient to uniquely identify vehicles, achieving an accuracy of 95.35%", by using grid-based generalization to simulate the basic privacy protection methods.

Sánchez *et al.* [21] claimed that "anonymization can be performed by techniques well established in the literature". They commented and pointed out that there are several limitations in the study of [2], e.g., the implemented anonymization strategies to coarsen the data are unreferenced and fall short of sufficiently protecting privacy, and the grid-based generalization uses fixed range values. To address these concerns, they chose a synthetically generated version of a publicly available patient discharge dataset② with spatio-temporal features, which includes nearly four million patients admitted to California hospitals in 2009, used more sophisticated k-anonymization to group records with similar census and spatio-temporal features, and found zero reidentifications of individuals. The finding successfully justifies their new claim.

The main finding of Sánchez *et al.* [21] is exactly opposite to that of De Montjoye *et al.* [3]. Hence, De Montjoye *et al.* [3] further gave a response and argued that "Sánchez's textbook k-anonymization example does not prove, or even suggest, that location and other big-data datasets can be anonymized and of general use", due to "a fundamental misunderstanding of the size and dimensionality of modern big-data datasets and how they are being used". By presenting more analyses and evidence, they claimed that "deidentification should not be considered a useful basis for policy", which corresponds to the finding of their earlier study [2].

Question 1. "What is the true situation of the privacy preservation for trajectories in terms of reidentification?" To the best of our knowledge, a systematic evaluation and analysis for trajectories on the individual privacy in terms of unicity is still on its way, though there exist a number of surveys on the anonymization mechanisms of trajectories [11, 12, 23–26].

However, anonymization typically causes a decline of data utility [25, 27], and anonymization mechanisms need to consider the trade-off between privacy protection and data utility. Indeed, Sánchez *et al.* [21] already pointed out that anonymized data should also retain its utility for data publishing, and adopt information loss to evaluate the utility of anonymized trajectory data.

Question 2. "What is the true situation of the utility of anonymized trajectories?" This is as important as the privacy in order to achieve a trade-off between privacy protection and data utility. To the best of our knowledge, most existing studies evaluate the utility of anonymization mechanisms independently except [23] that evaluates the utility with the quality loss, and there exist no systematic utility evaluations of anonymized trajectories in terms of practical applications. Moreover, various criteria to evaluate the utility, e.g., information loss based [21, 23, 28, 29] and application oriented [10, 27, 30, 31], application oriented criteria, are more direct or convincing to reflect the utility of anonymized data, such as window range queries [10] and software classification and defect prediction [27, 30].

Contributions. To this end, we provide a systematic evaluation and analysis on the privacy and utility of existing anonymization mechanisms for trajec-

②http://crises-deim.urv.cat/opendata/SPD_Science.zip, Aug. 2022.

tory data publishing. Note that although a number of studies [32–34] have evaluated anonymization mechanisms on various and heterogeneous data types, most of them focus on the performance of anonymization mechanisms, which is different from our purpose, and do not discuss trajectory data.

We conduct a systematic evaluation, using three real-life trajectory datasets [5, 6], five anonymization mechanisms (i.e., identifier anonymization [13], grid based anonymization [14], dummy trajectories [35], k-anonymity [10] and ε-differential privacy [31]), and two practical applications (i.e., travel time estimation [36, 37] and window range queries [38]). We find that reidentification privacy in terms of unicity is not well protected by the classic anonymization methods such as identifier anonymization [13], grid-based anonymization [14] and dummy trajectories [35], but is well preserved by k-anonymity [10] and ε-differential privacy [31]. This somehow confirms Sánchez *et al.*'s finding [21], and the anonymization mechanisms used by De Montjoye *et al.* in [2, 3] indeed have limitations and their finding on reidentification is indeed overestimated. This answers question 1 on the true situation of the trajectory privacy in terms of reidentification, and we also hope that this study closes the debate between De Montjoye *et al.* [2, 3] and Sánchez *et al.* [21].

We also find that the utility is determined by both anonymization mechanisms and concrete application algorithms for trajectory data. Furthermore, no anonymization mechanisms, maybe except identifier anonymization, for trajectory data, successfully satisfy all the needs of practical applications. These give an answer to question 2 on the true situation of the utility of anonymized trajectories.

Organizations. The rest of the paper is organized as follows. Section 2 reviews existing anonymization mechanisms for trajectory data, Section 3 introduces the measurements for privacy and utility, and Section 4 reports and analyzes the experimental findings, followed by conclusions in Section 5.

2 Anonymization Mechanisms

Many applications explicitly or implicitly make use of the trajectories of moving objects, which has raised the problem of individual privacy protection. Accordingly, various anonymization methods for trajectory publishing have been proposed to protect personally identifiable information such that a trajectory is regarded as a record of an individual moving object [10, 12, 31, 35, 39, 40]. In this section, we first introduce basic concepts on trajectories, and then give a brief introduction of these anonymization mechanisms.

Data Points. A data point is defined as a triple $P(x, y, t)$, which represents that a moving object is located at longitude x and latitude y at time t. Note that data points can be viewed as points in the x-y-t 3D Euclidean space.

Trajectories. A trajectory $\mathcal{T}(P_0, \ldots, P_n)$ is a sequence of points in a monotonically increasing order of their associated time values (i.e., $P_i.t < P_j.t$ for any $0 \leqslant i < j \leqslant n$). Intuitively, a trajectory is the path (or track) that a moving object follows through the space as a function of time [5, 6].

Trajectory Data. A trajectory dataset D typically consists of a set of trajectories such that each trajectory is associated with a unique identifier representing an individual moving object, and its sensitive information is the locations.

2.1 Identifier Anonymization

The simplest anonymization could be identifier anonymization, which means a dataset that is lack of names, home addresses, phone numbers, or other obvious identifiers (such as required, for instance, under the U.S. personally identifiable information (PII) "specific-types" approach) [13]. Many datasets are published with this mechanism due to its ease of usage, though it is known unsafe, such as the GAIA data opening program released by DiDi③, Kaggle for an online community of data scientists and machine learners④, and Geolife GPS trajectory data collected by Microsoft Research Asia⑤. Hence, we choose identifier anonymization as a representative method in our analysis.

2.2 Grid-Based Generalization

Generalization essentially means replacing one or multiple specific values with a more general one, such as coarsening the data by dividing the map as the Voronoi diagram [1], and clustering-based generalization [41]. The basic idea of grid-based generalization is to partition the data space into grids such that all points

③https://outreach.didichuxing.com/app-vue/, Aug. 2022.

④https://www.kaggle.com/crailtap/taxi-trajectory/, Aug. 2022.

⑤https://www.microsoft.com/en-us/download/details.aspx?id=52367, Aug. 2022.

falling into the same grid are uniformly represented by the grid. This mechanism is common and designed to anonymize user trajectories for privacy-preserving data mining [12,14], and we hence choose grid-based generalization as a representative method for anonymizing the spatial-temporal resolutions of trajectory data in our analysis.

Example 1. Fig.1 is an example of grid-based generation [14], in which a trajectory (P_1, P_2, ..., P_8) with eight points is fit in a 2D space that is partitioned into six grids denoted as G_1, G_2, ..., G_6. Then the trajectory is transformed into a new format (G_4, G_5, G_2) w.r.t. time intervals $P_1.t$–$P_3.t$, $P_4.t$–$P_6.t$ and $P_7.t$–$P_8.t$.

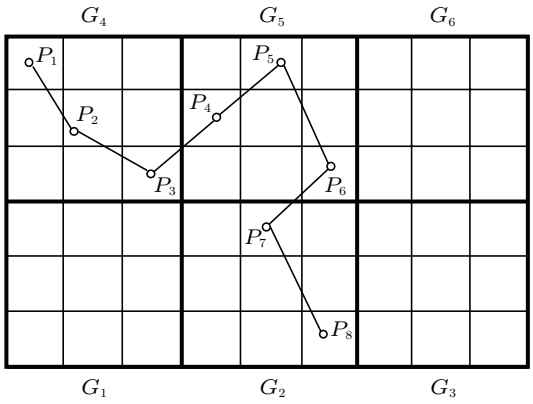

Fig.1. Example of grid-based generation [14].

2.3 Dummy Trajectories

Dummy trajectories generate fake trajectories, called dummies, to preserve the location of moving objects [12,15]. Different ways, i.e., classic random and rotation pattern schemes [35] and up-to-date deep generative models [42], can generate fake trajectories. Observing that moving behaviors of users usually follow certain patterns, which adversaries may exploit to distinguish true trajectories from dummies, an approach is proposed to generate intersecting dummy trajectories following certain moving patterns and to decrease the disclosure of individual user trajectories [35]. Two schemes, namely, random and rotation pattern schemes, are designed to generate dummies that exhibit long-term user movement patterns [35]. Random pattern scheme demonstrates that even after a long term observation, it is difficult for adversaries to identify true user trajectories since dummies also exhibit long-term, consistent movement patterns. The rotation pattern scheme generated dummy trajectories that have the same motion pattern with the original trajectory. Indeed [35] alleviates the privacy threat in a long run, and hence, we choose the rotation one as a representative method in our analysis.

Example 2. Fig.2 is an example of the rotation pattern scheme [35] for the original trajectory T. Fig.2(a) shows the result of T rotating 120° around the third point of T and Fig.2(b) shows the result of T rotating 80° around the second point of T.

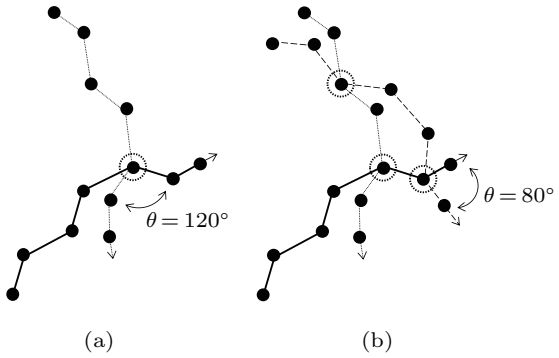

Fig.2. Example of the rotation pattern scheme [35].

2.4 K-Anonymity

K-anonymity is proposed to protect individual privacy such that each record is indistinguishable with at least other $k-1$ records w.r.t. the quasi-identifier, i.e., each equivalence class contains at least k records [16]. However, a k-anonymized equivalence class suffers from a homogeneity attack if all records in the class have less than k values for the sensitive attribute (e.g., disease and salary). To address this issue, l-diversity [43] and t-closeness [44] are proposed to ensure that 1) an equivalence class has at least l values for the sensitive attribute and 2) the distance between the distribution of a sensitive attribute and the distributions of all attributes is no more than a threshold t, respectively. (k, δ)-anonymity is introduced to anonymize trajectory data by extending k-anonymity with the spatial uncertainty $\delta \geqslant 0$ [10,45]. It contains two steps: it first groups k closest trajectories into clusters, and then moves the original trajectories to cylinders with a radius of δ using the space translation. However, it is proved that [10,45] can offer trajectory k-anonymity only when $\delta = 0$ [46]. As k-anonymity offers the better utility compared with its variants l-diversity and t-closeness, we choose the k-anonymity method NWA [10] by setting $\delta = 0$ as a representative method in our analysis.

Example 3. Fig.3 is an instance of k-anonymity [10]. Two trajectories are to be combined and the radius of the volume of trajectory is δ. The blue area represents

the volume of T_1, the red area represents the volume of T_2, τ_c is the track created by T_1 and T_2, and the gray area represents the anonymity set whose radius is $\delta/2$. Finally, we get the black anonymity trajectory.

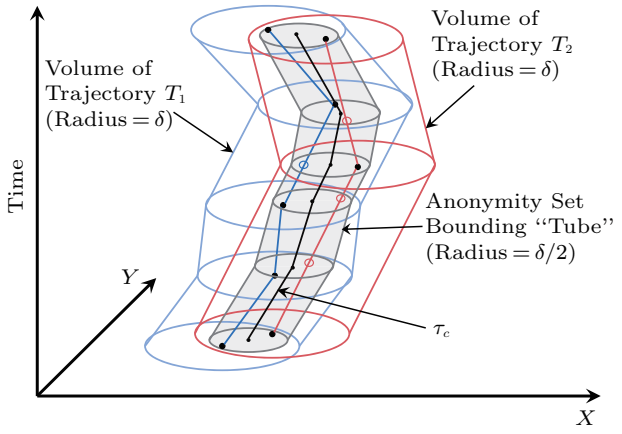

Fig.3. Example of k-anonymity [10].

2.5 Differential Privacy

Differential privacy is proposed in [47], which has become a popular privacy model and has been used in real applications such as Google⑥ and US Census Bureau⑦. Differential privacy requires that any computation on an underlying database is insensitive to the removal and addition of an individual record. That is, it provides a strong individual privacy guarantee.

ε-Differential Privacy. A random algorithm $\mathcal{A}$ satisfies ε-differential privacy if for any two neighboring datasets D_1 and D_2 differing at most one record, and any output O,

$$P(\mathcal{A}(D_1) = O) \leqslant \exp(\varepsilon) \times P(\mathcal{A}(D_2) = O),$$

where P denotes the probability and ε is the positive privacy budget that is believed that the smaller it is, the stronger the privacy guarantee is. The Laplace mechanism [48] and the exponential mechanism [49] are commonly adopted to achieve differential privacy, built on the l_1-norm sensitivity, defined as follows.

l_1-Norm Sensitivity. For any function $f(D) \to R^d$, its l_1-norm sensitivity is the maximum l_1-norm of $f(D_1) - f(D_2)$, where D_1 and D_2 are any two neighboring datasets differing at most one record, and is defined as follows:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|.$$

Sequential and parallel compositions are two important composition properties for differential privacy [31, 50].

Sequential Composition. If differential privacy is provided in each isolation, then it is also ensured on their sequential connection, where the final ε is the sum of all involved privacy budgets.

Parallel Composition. If differential privacy is provided on each disjoint set, then it is also ensured on their union, where the finally ε is the worst of all involved privacy budgets.

Differential privacy is firstly introduced for trajectory data in [51], which relies on the assumption that trajectories contain a lot of identical prefixes. This does not really hold in many applications as pointed out in [31]. The method in [31] removes the assumption, and is composed of two key components: 1) differentially private location generalization, which uses an exponential mechanism to probabilistically partition the location universe into groups at each time point and replaces all the locations belonging to the same group with their centroid, and 2) differentially private release for generalized trajectories, which generates new trajectories over the generalized location domains and publishes their noisy counts based on the Laplace mechanism that share the same Laplace mechanism as [20,52]. As this method shows better performance, we choose it as a representative method in our analysis.

Example 4. Fig. 4 is an example of ε-differential privacy [31]. The original locations in every time of (t_1, t_2, t_3) are transformed into locations P_{11}, P_{12}, P_{21}, P_{22}, P_{31} and P_{32} and then we get the generalized trajectories, i.e., (P_{11}, P_{21}, P_{31}) w.r.t. T_2. Next, a Laplace mechanism is applied to add noises to the counts of generalized trajectories (Table 1). Finally, we release the number of noisy trajectories.

3 Measuring Privacy and Utility

In this section, we introduce the measurements for privacy and utility. Following [1–3], we adopt unicity as the metric for measuring the privacy, and we evaluate the utility in terms of two classic applications: travel time estimation [53, 54] and window range queries [6, 55, 56].

⑥https://github.com/tensorflow/privacy, Aug. 2022.
⑦https://onthemap.ces.census.gov, Aug. 2022.

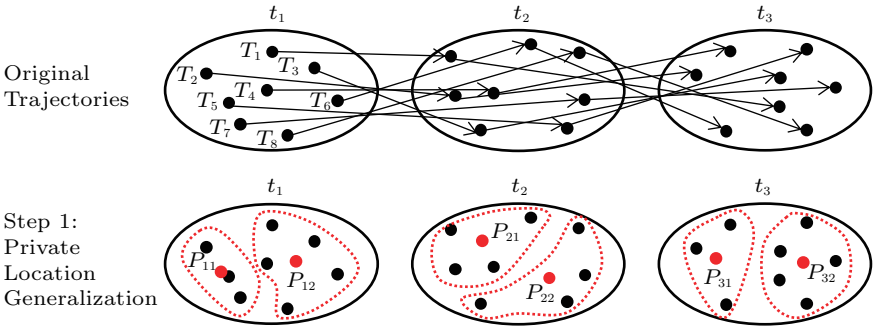

Fig.4. Example of differential privacy [31].

Table 1. Private Release of Generated Trajectories [31]

Generalized Trajectory	Original Trajectory	Number of Real Trajectories	Number of Noisy Trajectories
P_{11}, P_{21}, P_{31}	T_2	1	0
P_{11}, P_{21}, P_{32}	Null	0	1
P_{11}, P_{22}, P_{31}	Null	0	0
P_{11}, P_{22}, P_{32}	T_5, T_7	2	3
P_{12}, P_{21}, P_{31}	T_4, T_6	2	4
P_{11}, P_{21}, P_{32}	T_1, T_3	2	1
P_{12}, P_{22}, P_{31}	Null	0	1
P_{11}, P_{22}, P_{32}	T_8	1	2

3.1 Measuring Privacy

We first introduce unicity, which is used to evaluate the privacy risk in terms of the reidentification of anonymized trajectory data.

Unicity [1]. Unicity, denoted as μ, is the ratio of the number of reidentified (unique) trajectories N_U and the number of total trajectories N_T:

$$unicity\ (\mu) = \frac{N_U}{N_T}.$$

Given a positive integer q, we randomly choose q values $\{k_1, \ldots, k_q\}$ from $[1, n]$, and we say that a trajectory $\mathcal{T}$ in D is unique if there exist no other trajectories $\mathcal{T}'$ in D such that $\mathcal{T}.P_{k_1} = \mathcal{T}'.P_{k_1}$, ..., $\mathcal{T}.P_{k_p} = \mathcal{T}'.P_{k_p}$. That is, a unique trajectory does not have the same locations at the p chosen time points as all the other trajectories. Unicity essentially reveals the possibility of identifying the entire trajectory of a moving object, and p quantifies the amount of information one would need, on average, to reidentify a specific moving object. A larger unicity means a higher possibility to reidentify an object. Considering unicity $\mu = 0.9$ with $p = 2$, if we only know two points in trajectory $\mathcal{T}$, then we have a probability of 90% to identify or recover all the points of trajectory $\mathcal{T}$ by searching trajectories that contain the same two points. When there exists exactly one such trajectory in D, we know exactly the entire trace of the moving object. In this situation, we also say that the moving object (or trajectory) is reidentified.

Remark. For the purpose of unicity evaluation, the trajectory dataset D is also preprocessed by data interpolations such that 1) all trajectories in D have the same number of points, and 2) for any two trajectories $\mathcal{T}$ and $\mathcal{T}'$ in D, $\mathcal{T}.P_i.t = \mathcal{T}'.P_i.t$ for all $i \in [1, n]$.

3.2 Measuring Utility

We then introduce travel time estimation and window range queries, which are used to evaluate the utility of anonymized trajectory data.

3.2.1 *Travel Time Estimation*

Travel time is the total time for a vehicle to travel from one point to another over a specified route [57, 58]. Travel time estimation is to estimate the travel time w.r.t. an origin, a destination and departure time from the historical trips [37]. Travel time estimation has valuable commercial usages for urban computing such as route planning for drivers and passengers [53] and ride-sharing service [59].

SER, MRE, and MAE. Following [36, 58, 60], we adopt successful estimated ratio (SER), mean relative error (MRE) and mean absolute error (MAE) to evaluate the quality of travel time estimation, which are defined as follows:

$$SER = \frac{|i_{\text{tr}} \cap i_{\text{te}}|}{|i_{\text{te}}|},$$

$$MRE = \frac{\sum_{i=1}^{i=n} \frac{|et_i - rt_i|}{et_i}}{n}, \text{ and}$$

$$MAE = \frac{\sum_{i=1}^{i=n} |et_i - rt_i|}{n},$$

where i_{te} and i_{tr} are the trips of the testing data and the training data, respectively, n is the number of esti-

mated trips, et_i and rt_i are the estimated time and the real time of trip i, respectively, and a trip is a segment of a trajectory, i.e., a section from the start state to the stop state of a moving object. Here, the larger the SER is, the better the accuracy of the travel time estimation is, and the smaller the MAE and MRE are, the more accurate the travel time estimation is.

We adopt the trajectory-based simple concatenation (TSC) approach [36] and the temporal speed reference by region (TEMP+R) approach [37] for travel time estimation, which are referenced and used as baselines in many studies [57, 61]. Note that 1) any travel time estimation approach, such as deep learning based approaches [57, 61, 62], can essentially be used here, and 2) we use both raw and map-matched trajectories, i.e., TSC has a map-matching process while TEMP+R uses the raw trajectories. We next briefly introduce these two approaches.

TSC Approach [36]. Given a set D of trajectories, TSC partitions D into the training data and the testing data, where only those trajectories with more than 1 000 points are chosen. It first calculates the travel time of the training data, and then uses the travel time of the training data to estimate the travel time of the testing data. More specifically, TSC selects the fifth day as the testing data and the first four days as the training data, where the five days fall in the middle of the entire time range. It segments the long trajectories into smaller trips. Here the moving object is considered in a stop state when its speed is smaller than 1 m/s lasting for more than 120 s following [63]. Second, for fitting the input requirement of k-anonymity [10], it uses interpolations to unify the training data such that the start and the end time points of different trajectories are the same. The time interval between any two neighbouring points of a trajectory must be the same (6 s for Ucar, 6 s for Taxi, and 10 s for Truck for our experiments in Section 4). Third, it anonymizes the training data, where the map-matching method [64] is also implemented for trips. The result of map-matching is a table with attributes (road ID, time slot, travel time), and a day is divided into 48 time slots with 30 minutes each. Fourth, it calculates the real travel time of trips of the testing data using the map-matching results. Fifth, it estimates the travel time in the testing data by aligning the two tables generated by the training and the testing data, and by comparing the records with the same road ID and time slot. Finally, MAE, MRE and MR are calculated.

TEMP+R Approach [37]. Given a set D of trajectories, TEMP+R also partitions D into the training data and the testing data, and segments trajectories into trips along the same lines as TSC. Trip t_i is a neighbor of trip t in the testing data if the origin and the destination of t_i are spatially close to the origin and the destination of t, respectively, measured in terms of the Euclidean distance. In order to quickly retrieve the neighboring trips, it employs a grid partition of a city (e.g., 50 m × 50 m grids). Let $N(t)$ be the neighbors of trip t whose Euclidean distances are in the nearest τ grids. It estimates the travel time of trip t using its neighboring trips as follows:

$$TravelTime(t) = \frac{1}{|N(t)|} \sum_{t_i \in N(t)} t_i \times \frac{AVG(t_i)}{AVG(t)},$$

where $AVG(t_i)$ denotes the average speed of all trips whose start time point falls into the same time slot as trip t_i (for example, using a daily pattern with one hour as the basic unit, we have 24 time slots per day), and $AVG(t)$ denotes the average speed of trip t.

Note that the calculations of SER using TSC and TEMP+R are different as TSC has a map-matching procedure. 1) For TSC, i_{te} is the set of trips in the testing data, i_{tr} is the set of trips in the training data that are successfully map-matched, and $i_{\text{tr}} \cap i_{\text{te}}$ is the set of successfully estimated trips when aligning the two tables generated by the training data and the testing data. 2) For TEMP+R, i_{te} is the set of trips in the testing data, i_{tr} is the set of trips in the training data, and $i_{\text{tr}} \cap i_{\text{te}}$ is the set of successfully estimated trips in the neighbouring trips of the testing data from the training data.

3.2.2 Window Range Queries

Spatio-temporal queries are fundamental operations for trajectory data [6, 38, 55, 56, 65], among which we choose window range queries to evaluate the utility of anonymized trajectory data. Window range queries are in particular useful for vehicle flow monitoring that explores the traffic flow information to help make better travel decisions, alleviate traffic congestion, and improve the urban planning [66], and have been commonly used for the utility evaluation [10, 31].

Given a cube $(x_1, x_2, y_1, y_2, t_1, t_2)$, a window range query (W-RQ) finds all the trajectories with at least one point $p = (x, y, t)$ such that $x_1 \leqslant x \leqslant x_2$, $y_1 \leqslant y \leqslant y_2$ and $t_1 \leqslant t \leqslant t_2$ [31, 55, 65]. Such a W-RQ answers the question about how many trajectories pass through the region (x_1, x_2, y_1, y_2) during the time period $[t_1, t_2]$.

*F*1 *Measure.* Following [38], we adopt the $F1$ measure to evaluate the quality of window range queries such that the higher it is, the better the quality is. Given a W-RQ Q, its $F1$ is defined as follows:

$$F1\,(Q) = \frac{2 \times precision\,(Q) \times recall\,(Q)}{precision\,(Q) + recall\,(Q)},$$

such that the precision and the recall are computed as follows:

$$precision\,(Q) = |R_o \cap R_a|/|R_o|,$$
$$recall\,(Q) = |R_o \cap R_a|/|R_a|,$$

where R_o and R_a denote the two sets of trajectories returned from the original trajectory data and the anonymized data, respectively.

4 Experiments and Analyses

In this section, we evaluate the unicity and utility with five anonymization methods, and present our findings and analyses. As data cleaning has few impacts on the unicity and utility, we use the original datasets for our experiments.

4.1 Experimental Settings

4.1.1 Datasets

We use three real-life datasets reflecting the trajectories in cities [5, 6], shown in Table 2. 1) Ucar is a dataset containing trajectories from a Chinese car rental company located in Beijing, China. 2) Taxi is a dataset containing taxi trajectories in Beijing, China. 3) Truck is a dataset containing truck trajectories mainly in Nanning, China. These datasets have various sampling rates, ranging from one point per six seconds to one point per 60 seconds. The trajectories of Ucar and Taxi are mostly located in city centers, and those of Truck sometimes move in a group and have a wide spatial distribution. The latitude and the longitude accuracies of all datasets are kept with six decimal numbers.

4.1.2 Anonymization Mechanism Implementation

We implement five anonymization mechanisms.

1) *Identifier Anonymization.* We simply randomly generate a pseudo identifier for each trajectory.

2) *Grid-Based Anonymization.* We set different spatial-temporal resolutions for the trajectory datasets. The entire spatial area is divided into different squares along the latitude and the longitude, and the spatial resolution means the side length of squares, varying from 20 m to 10 000 m. A temporal resolution divides the whole time period into time bins with a time interval (temporal resolutions), where the start time of the first time bin is 0:0 of the first day.

3) *Dummy Trajectories.* Following [35], dummy trajectories are generated as follows. For each trajectory, we first find the point locating in the middle time of the trajectory, and then we rotate the trajectory around the point anticlockwise with varying angles that fall in {5°, 10°, 15°, 20°, 25°, 30°, 45°, 90°, 135°, 180°}.

4) *K-Anonymity.* We adopt the (k, δ)-anonymity approach [10] that consists of three steps. i) All training datasets are pre-processed using interpolations such that the time intervals between any two neighboring points are exactly 6 s for Ucar, 60 s for Taxi, and 10 s for Truck, and all the trajectories in a training dataset have the same start and end time. ii) The processed training datasets are then clustered. First, a list of pivot trajectories that are acted as the cluster centers are selected, where the first one is the farthest trajectory from the average trajectory of the entire dataset, and the remaining pivot trajectories are the farthest trajectories from previous chosen pivot ones. Second, a pivot trajectory together with its $k-1$ nearest neighboring trajectories forms a cluster ($k \in \{2, 4, 6, 8, 10\}$). A constraint is enforced for each cluster whose radius is not larger than a threshold max_radius, which is initially set to 0.5% of the semi-diagonal of the spatial minimal bounding box of the datasets. If a cluster cannot be created around a new pivot, it is not used as a pivot, but as a member of some other clusters. If a trajectory cannot be added to any cluster, it is simply

Table 2. Three Real-Life Trajectory Datasets

Dataset [5, 6]	Number of Trajectories	Sampling Rate (points/s)	Number of Points	Latitude		Longitude	
				Min.	Max.	Min.	Max.
Ucar	2 023	6	48 936 975	39.680	40.252	116.007	116.738
Taxi	8 715	60	76 543 949	39.680	40.252	116.007	116.738
Truck	675	10, 30, 60	12 886 273	22.552	23.259	107.636	108.939

trashed. This process may lead to many trash trajectories such that the clustering process is restarted by multiplying *max_radius* with 1.5. This process repeats until the number of the trashed trajectories is smaller than *max_trash*. iii) Finally, a k-anonymized aggregate trajectory is formed for each cluster by setting the locations of all points at the same time to their arithmetic means. Note that the (k, δ)-anonymity approach can offer trajectory k-anonymity only when δ=0 [46], and hence we always set $\delta = 0$ in our tests.

5) *Differential Privacy.* We adopt the ε-differential privacy approach [31], which consists of three steps. i) All training datasets are firstly pre-processed using interpolations such that the time intervals between two neighboring are exactly 3 600 s for Ucar, Taxi and Truck, and all the trajectories have the same start and end time. The interval here is much larger than k-anonymity, as the computation of ε-differential privacy is much more expensive. ii) This second step is called differentially private location generalization, which uses an exponential mechanism to probabilistically partition the location universe Γ into groups at each time point and replaces all the locations belonging to the same group with their centroid. To improve the efficiency, the total number of partition candidates is reduced from $m^{|\Gamma|}$ to $\varphi + 1 + |\Gamma|$, where m is the expected number of partitions and Γ is the original location domain ($\varphi = \left\lceil \frac{|\Gamma|}{10} \right\rceil$). This makes the exponential scheme become feasible in practice. First, it uses the k-means algorithms to partition the original trajectories into m groups based on their pairwise Euclidean distances ($k \in \{20, 40, 60, 80, 100\}$). Second, it produces another set of partition candidates τ, which consists of φ partitions producing the next φ greatest utilities, of which each adds a distinct trajectory nearest to the centroid, and a total of $|\Gamma|$ k-means partitions based on the datasets, of which each removes a distinct trajectory from D. When conducting the exponential mechanism, the probability of choosing each partition $p \in \tau$ is $\frac{\exp(\frac{\varepsilon_1}{2\Delta U} U(D,p))}{\sum_{p\in\tau} \exp(\frac{\varepsilon_1}{2\Delta U} U(D,p))}$, where $\varepsilon_1 = 0.001$ is the privacy budget of the exponential mechanism, U is the utility function such that $U = \frac{\min_{p\in\tau}(AvgDist(p))}{AvgDist(p)}$, where $AvgDist$ is the average of the mean distances between all pair of points at every time point in a cluster and ΔU is the sensitivity of the utility function (equal to 1). iii) This step is called differentially private release for generalized trajectories, which generates new trajectories over the generalized location domains, and publishes their noisy counts based on the Laplace mechanism. We add Laplace noise $\mathrm{Lap}(1/\varepsilon_2)$, where $\varepsilon_2 \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$, to the real count of trajectories.

As shown in [31], this approach holds the following.

Theorem 1. *The approach in* [31] *satisfies ε-differential privacy, where $\varepsilon = |\Gamma| \times \varepsilon_1 + \varepsilon_2$, and $|\Gamma|$ is the number of trajectories.*

4.2 Experimental Results

In this subsection, we report the findings of individual privacy in terms of unicity, and utility in terms of two classic applications. Each test is repeated over five times, and the average is reported.

4.2.1 Privacy Tests with Unicity

We find that 1) for identifier anonymization and dummy trajectories with the rotation pattern scheme, the unicity is always kept to 1, 2) for (k, δ)-anonymity, the unicity is always 0 (which can be easily inferred from the definition of k-anonymity), and the unicity remains close to 0 even if we keep the trashed trajectories as 2 for $k \in \{2, 4, 6, 8, 10, 20, 40, 60, 80, 100\}$, and 3) for differential privacy, the unicity is always close to 0. These imply that the individual privacy is not well preserved by identifier anonymization and dummy trajectories, but is well preserved for k-anonymity and differential privacy. Next, we only report the unicity for grid-based generalization.

Exp-1: *Impact of the Number of Points.* In this test, to evaluate the impact of the number of points m, we fix the spatial resolution $x = 1$ km and the temporal resolution $y = 1$ h, and vary the number of points m from 1 to 10. The unicity results are reported in Fig.5(a).

The results show that the unicity μ is high, i.e., 1.0 for Ucar, 0.92 for Taxi and 0.68 for Truck when $m = 4$, respectively. This means that knowing four random spatio-temporal points is enough to uniquely reidentify most of the individual objects and to uncover their entire records. Further, the unicity increases with the increment of m. This is obvious as it becomes more difficult to find trajectories with more knowing points. When $m = 1$, the unicity of Truck is higher than that of Ucar and Taxi. This is because truck trajectories are sparser. We also find that the unicity of Truck is smaller when $m > 3$ because trucks often move together in groups. While four points are enough to uniquely reidentify all considered trajectories for Ucar ($\mu > 0.99$), eight points are needed for Taxi ($\mu > 0.99$) and Truck ($\mu > 0.80$). Hence, the individual privacy is not well preserved.

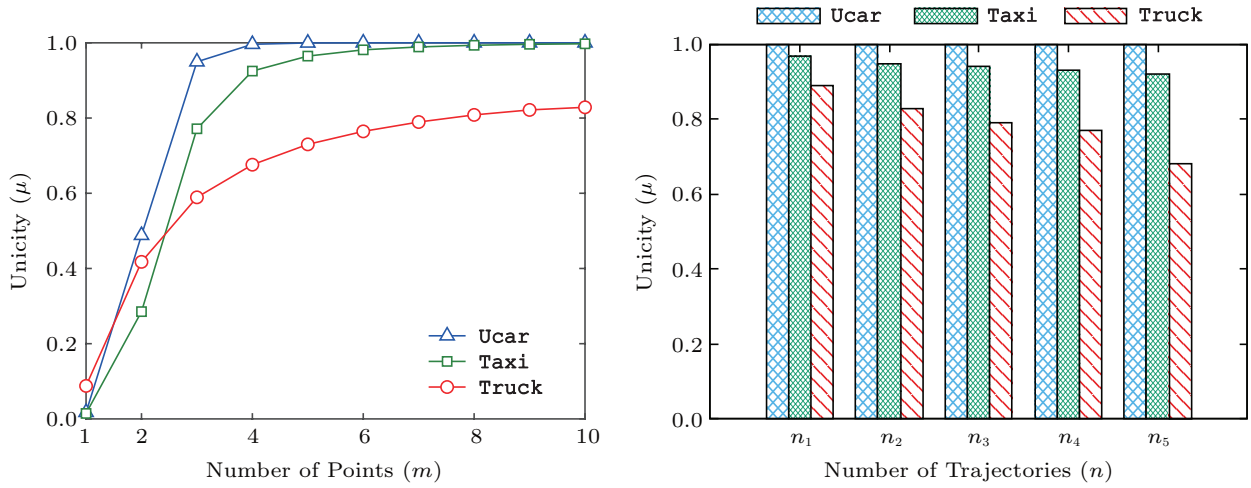

Fig.5. Unicity μ w.r.t. grid-based generalization. (a) w.r.t. number of points. (b) w.r.t. number of trajectories.

Exp-2: *Impact of the Number of Trajectories.* In this test, we use the same setting as Exp-1, fix the number of points $m = 4$, and set the number of trajectories n to $\{500, 1\,000, 1\,500, 2\,000, 2\,023\}$ for Ucar, $\{2\,000, 4\,000, 6\,000, 8\,000, 8\,715\}$ for Taxi, and $\{150, 300, 450, 600, 675\}$ for Truck, respectively. The results are reported in Fig.5(b).

These results show that the unicity decreases with the increment of the number of trajectories on all datasets. This is because when the number of trajectories increases, there are more possibilities to find the trajectories sharing p points, which obviously leads to the decrease of μ. However, the unicity μ remains high, and the individual privacy is still not well preserved.

Exp-3: *Impact of Spatial and Temporal Resolutions.* In this test, we use the same setting as Exp-2 on the entire datasets, and vary the spatial resolution (the length of square sides) from 20 m to 10 000 m and the temporal resolution from 20 s to 100 000 s. We also test the unicity with very low spatial (4 000 m) and temporal (12 800 s) resolutions. The results are reported in Figs.6 and 7.

The results show that the unicity decreases with the decrement of both spatial and temporal resolutions on all datasets. To explain this phenomenon, we calculate the entropies of the anonymized datasets with different resolutions shown in Fig.8, which has been used to indicate the difficulty of predicting the user locations [67]. Let N_i be the number of points that user u located in area L during a time period, and M be the total number of his/her appearances, and then user u appears at location L with a probability $pr = N_i/M$. The entropy E is defined as follows: $E = \sum_{i=1}^{i=n} p \log_2 p$, where n is the total number of points. The results show that the entropy decreases with the decrement of spatial and temporal resolutions. A large entropy indicates a low degree of location concentration, which leads to a larger unicity as it becomes more difficult for trajectories to share

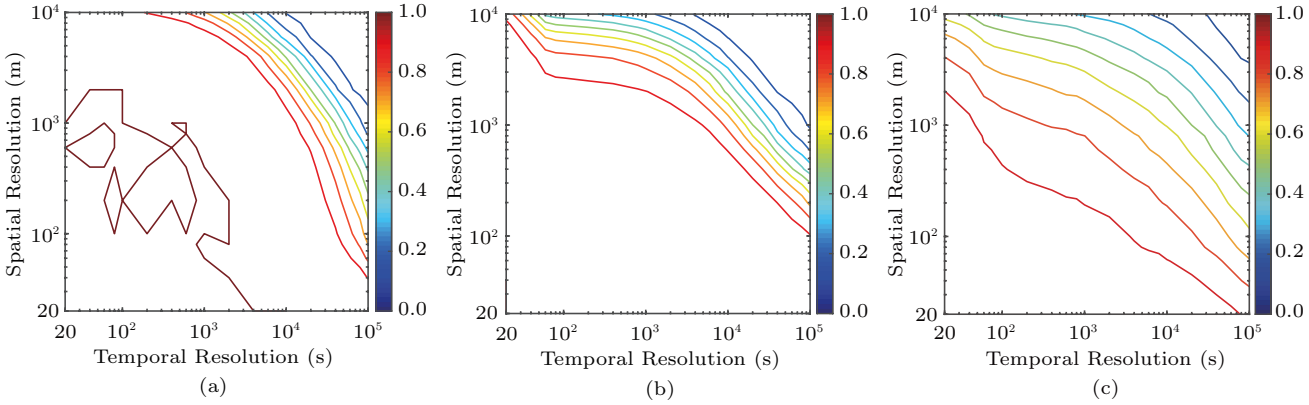

Fig.6. Contour map of unicity μ w.r.t. spatial and temporal resolutions. (a) Ucar. (b) Taxi. (c) Truck.

common points. This also explains why the unicity of Truck, Taxi and Ucar in general obeys an increasing order, as their entropies obey a decreasing order. Fig.7 further shows that data generalization is not enough to protect the privacy of individuals even with very low spatial and temporal resolutions. Although the unicity decreases with the decrement of the resolutions, it decreases slowly along the spatial and temporal axes. Furthermore, this decrease can be easily eliminated by collecting a few more points.

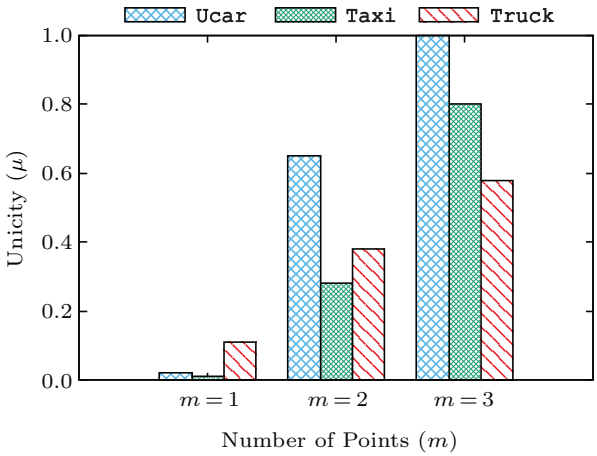

Fig.7. Unicity μ with very low spatial and temporal resolutions (4 000 m × 12 800 s).

4.2.2 *Utility Tests with Travel Time Estimation*

As explained in Subsection 3.2, map-matching [64] is a key step for the TSC approach [36] for travel time estimation. However, the successful map-matching ratios for grid-based generalization and ε-differential privacy are almost 0, as road searches are restricted within 200 m in map-matching, and larger distances are normally unnecessary, which otherwise makes the computation costs unacceptable [64]. For grid-based generalization, each grid is treated as a point, and hence TEMP+R fails for travel time estimation [37]; and ε-differential privacy involves grouping on each time point, and hence the time interval between two neighboring points needs to be large to make its computation practical. However, large time intervals typically make TEMP+R [37] fail for travel time estimation. Hence, we only report travel time estimation for identifier anonymization, dummy trajectories with rotation pattern scheme and k-anonymity, where we choose to omit the trashed trajectories as they have few impacts on the evaluation. One issue is how to distinguish the proper set of trips for estimation. It is reported that passengers wait around five minutes for pickup [68]. Hence, we choose longer trips whose total travel time is more than 10 minutes. The average travel time of the resulting training data on (Ucar, Taxi, Truck) is (48.4, 24.8, 20) minutes, and that of the resulting testing data on (Ucar, Taxi, Truck) is (36.9, 34.2, 25.4) minutes, respectively.

Exp-4: *Successful Estimated Ratio Test.* In this test, we evaluate the successful estimated ratios of the travel time estimation for identifier anonymization, dummy trajectories and k-anonymity using methods TSC and TEMP+R respectively with τ set to 5 (i.e., the five nearest grids). More specifically, we evaluate the impact of the rotation angle θ for dummy trajectories by varying its values to $\{0°, 5°, 10°, 15°, 20°, 25°, 30°, 45°, 90°, 135°, 180°\}$, respectively, and the impact of k for k-anonymity by varying its values to $\{1, 2, 4, 6, 8, 10\}$, respectively. Note that for the case of dummy trajectories with $\theta = 0°$ or k-anonymity with

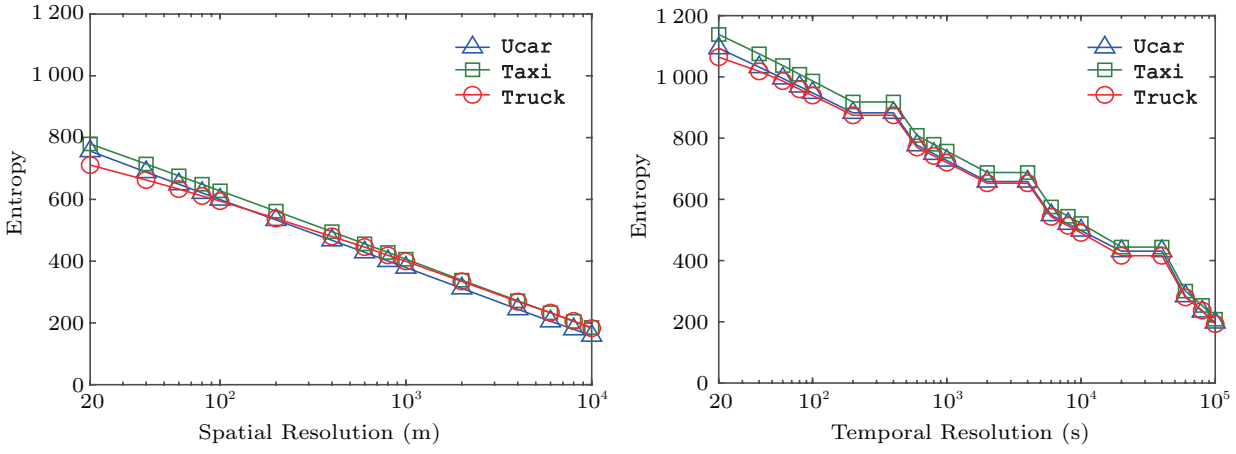

Fig.8. Entropy w.r.t. (a) spatial and (b) temporal resolutions.

$k = 1$, it is essentially identifier anonymization. The results are reported in Figs.9(a), 10(a) and 11(a).

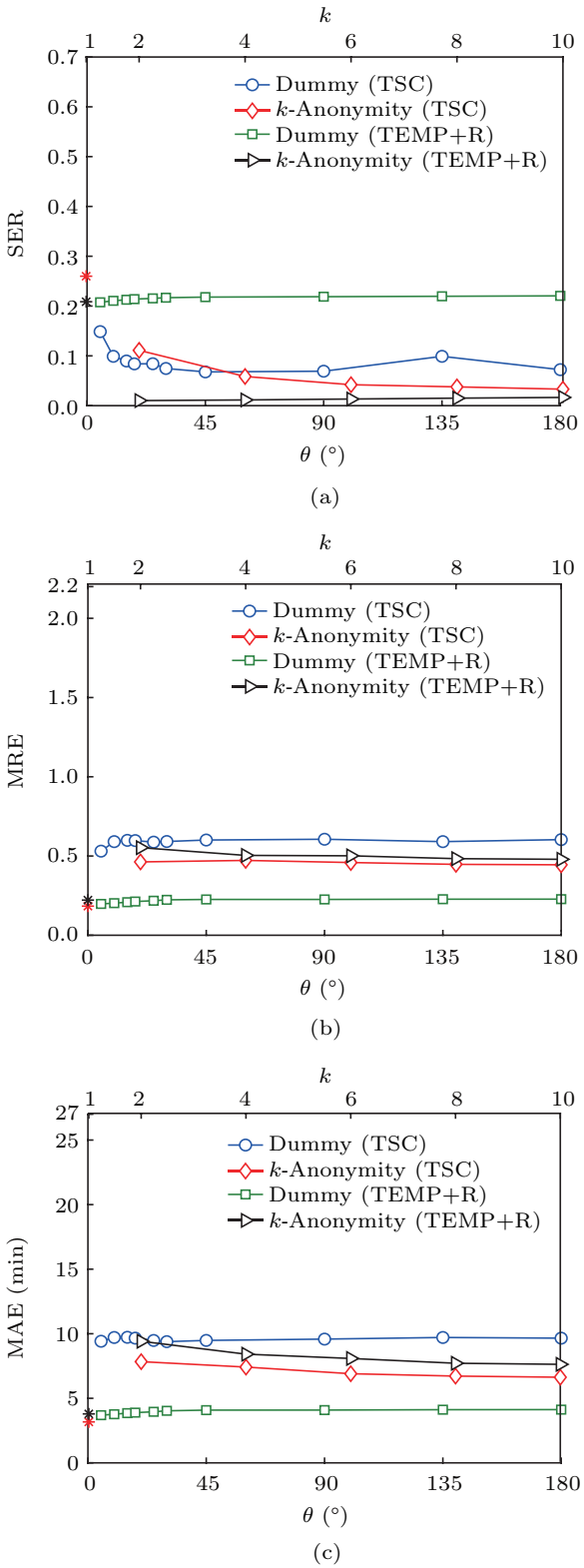

Fig.9. Travel time estimation on Ucar. (a) SRE. (b) MRE. (c) MAE.

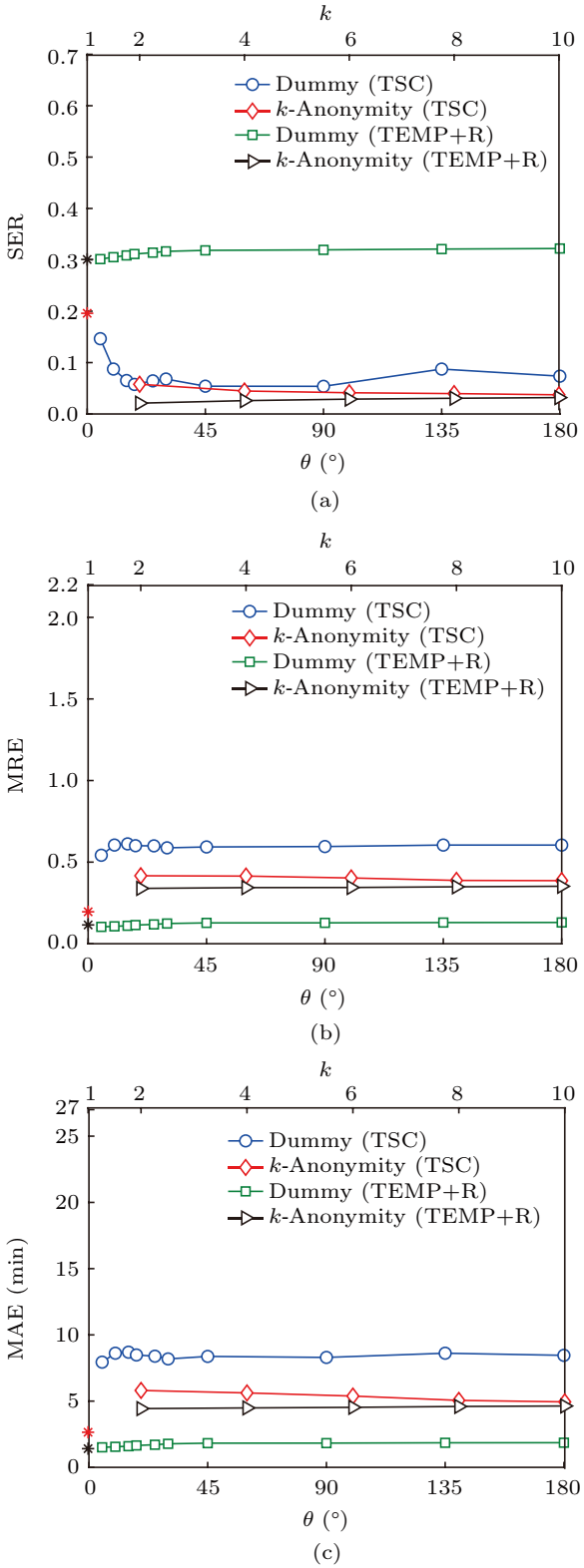

Fig.10. Travel time estimation on Taxi. (a) SRE. (b) MRE. (c) MAE.

In Figs.9(a), 10(a) and 11(a), for identifier anonymization, i.e., the left most points marked with $*$ ($\theta = 0°$

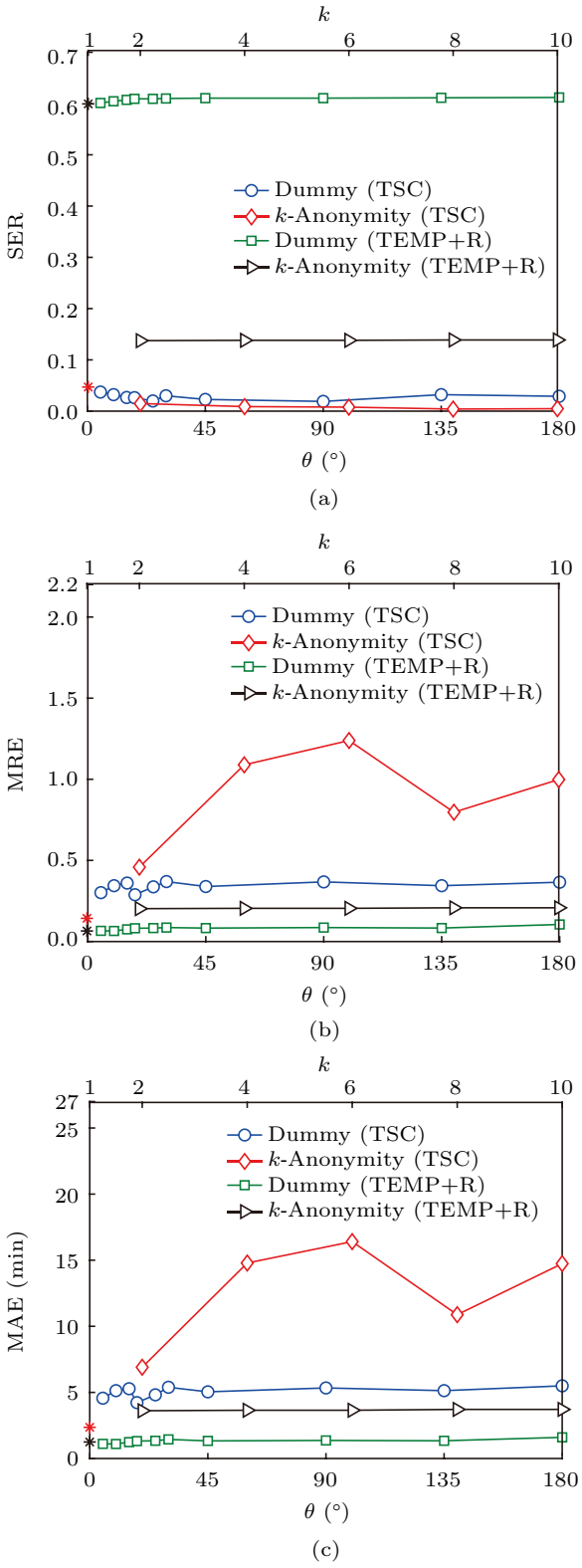

Fig.11. Travel time estimation on Truck. (a) SRE. (b) MRE. (c) MAE.

or $k = 1$), the results show that SER using TSC is (0.26, 0.19, 0.05) on (Ucar, Taxi, Truck) and SER using TEMP+R is (0.20, 0.30, 0.60) on (Ucar, Taxi, Truck), respectively. TEMP+R also shows better performance than TSC. Note that identifier anonymization obviously has no impact on the SER of travel time estimation.

For dummy trajectories, the results show that the average SER of TSC decreases by $(65\%, 60\%, 41\%)$ on (Ucar, Taxi, Truck), but that of TEMP+R slightly increases by $(6\%, 6\%, 2\%)$ on (Ucar, Taxi, Truck). For TSC, its map-matching procedure becomes more inaccurate for dummy trajectories, which leads to the decrease of its SER. For TEMP+R, there is no map-matching, and its neighboring trips sharing spatially close origins and destinations benefit from the gridding. These results also tell us that dummy trajectories have a significant impact on the SER of TSC, but not that of TEMP+R.

For k-anonymity, in Figs.9(a), 10(a) and 11(a), the results show that the average SER of TSC decreases by $(80\%, 81\%, 83\%)$ on (Ucar, Taxi, Truck), and that of TEMP+R decreases by $(97\%, 93\%, 77\%)$ on (Ucar, Taxi, Truck), respectively. The map-matching procedure of TSC and the neighboring trips sharing spatially close origins and destinations of TEMP+R both become more inaccurate for k-anonymity, which leads to the decrease of their SER. These results also tell us that k-anonymity has a significant impact on the SER of both TSC and TEMP+R.

Exp-5: *Mean Relative Error Test.* In this test, we evaluate the mean relative errors of travel time estimation for identifier anonymization, dummy trajectories and k-anonymity using TSC and TEMP+R, using the same setting as Exp-4. The results are reported in Figs.9(b), 10(b) and 11(b). From these three figures, we set the following findings.

For identifier anonymization (the left most points marked with $*$), the results show that the MRE using TSC is (0.16, 0.18, 0.14) on (Ucar, Taxi, Truck) and the MRE using TEMP+R is (0.20, 0.10, 0.06) on (Ucar, Taxi, Truck), respectively. Note that identifier anonymization obviously has no impact on the MRE of travel time estimation.

For dummy trajectories, the results show that the average MRE of TSC significantly increases by $(274\%, 228\%, 148\%)$ on (Ucar, Taxi, Truck), and that of TEMP+R increases by $(11\%, 19\%, 37\%)$ on (Ucar, Taxi, Truck), respectively. TEMP+R also shows better performance than TSC mainly due to the extra errors introduced by map-matching in TSC. These results also tell us that dummy trajectories have a significant impact on the MRE of TSC.

For k-anonymity, the results show that the average MRE of TSC increases by $(175\%, 115\%, 562\%)$ on (Ucar, Taxi, Truck), and that of TEMP+R increases by $(145\%, 233\%, 236\%)$ on (Ucar, Taxi, Truck), respectively. This is mainly because the trips in the training set become much smaller and more inaccurate after k-anonymity. These results tell us that k-anonymity has significant impacts on the MRE of both TSC and TEMP+R.

Exp-6: *Mean Absolute Error Test.* In this test, we evaluate the mean absolute errors of travel time estimation for identifier anonymization, dummy trajectories and k-anonymity using methods TSC and TEMP+R, using the same setting as Exp-4. The results are reported in Figs.9(c), 10(c) and 11(c).

For identifier anonymization (the left most points marked with $*$), the results show that the MAE using TSC is (3.05, 2.70, 2.12) minutes on (Ucar, Taxi, Truck) and the MAE using TEMP+R is (3.67, 1.47, 1.00) minutes on (Ucar, Taxi, Truck), respectively. Note that identifier anonymization obviously has no impact on the MAE of travel time estimation.

For dummy trajectories, the results show that the average MAE of TSC significantly increases by $(215\%, 212\%, 138\%)$ on (Ucar, Taxi, Truck), and that of TEMP+R increases by $(8\%, 17\%, 32\%)$ on (Ucar, Taxi, Truck), respectively. TEMP+R also shows better performance than TSC mainly due to the extra errors introduced by map-matching in TSC. These results also tell us that dummy trajectories have a significant impact on the MAE of TSC.

For k-anonymity, the results show that the average MAE of TSC increases by $(129\%, 100\%, 495\%)$ on (Ucar, Taxi, Truck), and that of TEMP+R increases by $(212\%, 233\%, 246\%)$ on (Ucar, Taxi, Truck), respectively. These results tell us that k-anonymity has significant impacts on the MAE of both TSC and TEMP+R.

4.2.3 Utility Tests with Window Range Queries

It is easy to know that the $F1$ scores of window range queries are always 1 for identifier anonymization. Here we only report the results for grid-based generalization, dummy trajectory anonymization, k-anonymity and ε-differential privacy.

The cube $(x_1, x_2, y_1, y_2, t_1, t_2)$ for window range queries is set as follows. 1) Following [66], its half time period $halfT = (t_2 - t_1)/2$ is chosen as one hour and three hours, and its half length $halfL$ of latitudes $(x_2 - x_1)/2$ and longitudes $(y_2 - y_1)/2$ is identical and chosen as {2 km, 4 km, 6 km, 8 km, 10 km}, respectively. 2) For each dataset, we also select five places as the centres of the cubes. For Ucar and Taxi, the centres are Beijing Railway Station, Beijing West Railway Station, Tian'anmen Square, Beijing Olympic Forest Park and China Central Television. For Truck, the centres are Nanning Railway Station, Nanning South Railway Station, Nanning Bridge, Guangxi Province Government and Nanning Government. 3) Finally, we set the time at the centres of the cubes as 12 o'clock at noon on the fourth day. The average $F1$ scores of the five places for each half time period and each half length on each dataset are reported.

Exp-7: *Grid-Based Generalization.* In this test, we evaluate the impacts of spatial and temporal resolutions for grid-based generalization, the half time period and the half length of the cubes for window range queries. We fix the temporal resolution to 1 h, and vary the spatial resolution to 2 km, 4 km and 6 km, respectively, to test the impacts of spatial resolutions. Besides, we fix the spatial resolution to 1 km, and vary the temporal resolution to 1 h, 3 h and 6 h, respectively, to test the impacts of temporal resolutions. The impacts of the cubes are also tested by varying the half time period and the half length as introduced above. The results are reported in Tables 3–5.

The results show that the $F1$ scores decrease with the decrement of both spatial and temporal resolutions, as the positions in the trajectories of individual objects become more inaccurate for lower resolutions such that the trajectories originally passing through the cubes may not pass through the cubes any more, and the $F1$ scores increase with the increment of both the half time period and the half length, as large cubes are more tolerant to the inaccuracy of the positions in the trajectories of individual objects, giving the same set of anonymized trajectories. Further, the spatial and the temporal resolutions have significant impacts on the $F1$ scores, compared with the cubes. The $F1$ scores even become 0 when the spatial and temporal resolutions are 1 h and 6 km, respectively, as the cubes are small and the positions in the trajectories are very inaccurate.

These also tell us that given a window range query, we need to properly choose the spatial and temporal resolutions in order to reach a high utility (e.g., high $F1$ scores). For instance, when given a query cube with $halfT = 1$ h and $halfL = 2$ km, the spatial and temporal resolutions need to satisfy (1 h, $\leqslant$ 4 km) or ($\leqslant$ 3 h, 1 km) for Ucar and Taxi, and (1 h, $\leqslant$ 2 km) or ($\leqslant$ 6 h, 1 km) for Truck, respectively, to reach an $F1$ score

Table 3. $F1$ Scores of Window Range Queries for Grid-Based Generalization (UCar)

$halfT$ (h)	$halfL$ (km)	Resolution					
		(1 h, 2 km)	(1 h, 4 km)	(1 h, 6 km)	(1 h, 1 km)	(3 h, 1 km)	(6 h, 1 km)
1	2	0.91	0.83	0.00	0.96	0.83	0.66
	4	0.96	0.91	0.81	0.97	0.86	0.73
	6	0.98	0.95	0.95	0.99	0.89	0.79
	8	0.99	0.95	0.92	0.99	0.92	0.83
	10	0.99	0.98	0.96	0.99	0.86	0.93
3	2	0.93	0.86	0.00	0.97	0.90	0.88
	4	0.97	0.94	0.86	0.98	0.93	0.91
	6	0.99	0.97	0.97	0.99	0.95	0.94
	8	0.99	0.97	0.96	1.00	0.97	0.96
	10	0.99	0.99	0.98	1.00	0.97	0.96

Table 4. $F1$ Scores of Window Range Queries for Grid-Based Generalization (Taxi)

$halfT$ (h)	$halfL$ (km)	Resolution					
		(1 h, 2 km)	(1 h, 4 km)	(1 h, 6 km)	(1 h, 1 km)	(3 h, 1 km)	(6 h, 1 km)
1	2	0.92	0.83	0.00	0.96	0.84	0.65
	4	0.96	0.90	0.84	0.97	0.88	0.74
	6	0.98	0.95	0.94	0.98	0.91	0.81
	8	0.98	0.95	0.92	0.99	0.93	0.85
	10	0.98	0.98	0.96	0.99	0.95	0.88
3	2	0.95	0.87	0.00	0.97	0.89	0.89
	4	0.98	0.95	0.89	0.99	0.93	0.93
	6	0.99	0.97	0.97	0.99	0.95	0.95
	8	0.99	0.97	0.96	0.99	0.96	0.96
	10	0.99	0.99	0.98	0.99	0.97	0.97

Table 5. $F1$ Scores of Window Range Queries for Grid-Based Generalization (Truck)

$halfT$ (h)	$halfL$ (km)	Resolution					
		(1 h, 2 km)	(1 h, 4 km)	(1 h, 6 km)	(1 h, 1 km)	(3 h, 1 km)	(6 h, 1 km)
1	2	0.88	0.59	0.00	0.96	0.90	0.90
	4	0.91	0.81	0.77	0.97	0.93	0.93
	6	0.92	0.84	0.80	0.95	0.91	0.91
	8	0.93	0.89	0.81	0.98	0.94	0.94
	10	0.96	0.94	0.90	0.98	0.95	0.95
3	2	0.88	0.59	0.00	0.96	0.90	0.90
	4	0.91	0.81	0.77	0.97	0.91	0.93
	6	0.92	0.84	0.80	0.95	0.91	0.91
	8	0.93	0.89	0.81	0.98	0.94	0.94
	10	0.96	0.94	0.90	0.98	0.95	0.95

greater than 0.80.

Exp-8: *Dummy Trajectories*. In this test, we evaluate the impacts of rotation angles for dummy trajectories, the half time period and the half length of the cubes for window range queries. We vary the rotation angle θ to {5°, 10°, 15°, 20°, 25°, 30°, 45°, 90°, 135°, 180°}, respectively, to evaluate the impacts of rotation angles. The impacts of cubes are also tested by varying the half time period and the half length following Exp-7. The results are reported in Tables 6–8.

The results show that the $F1$ scores decrease significantly with the increment of rotation angles, as the positions in the trajectories of individual objects become more inaccurate for larger rotation angles such that the trajectories originally passing through the cubes may not pass through the cubes any more. The same impacts of the cubes as Exp-7 are also confirmed, i.e., the $F1$ scores increase with the increment of both the half time period and the half length.

These also tell us that given a window range query, we need to properly choose the rotation angles in order to reach a high utility. For instance, when given a

Table 6. $F1$ Scores of Window Range Queries for Dummy Trajectories (UCar)

$halfT$ (h)	$halfL$ (km)	θ (°)									
		5	10	15	20	25	30	45	90	135	180
1	2	0.87	0.70	0.59	0.50	0.44	0.40	0.30	0.19	0.17	0.15
	4	0.91	0.82	0.74	0.68	0.63	0.60	0.49	0.33	0.28	0.26
	6	0.96	0.89	0.83	0.78	0.74	0.71	0.62	0.45	0.38	0.35
	8	0.97	0.93	0.89	0.84	0.80	0.77	0.69	0.54	0.47	0.44
	10	0.98	0.96	0.93	0.90	0.86	0.83	0.76	0.61	0.55	0.52
3	2	0.88	0.74	0.66	0.59	0.54	0.49	0.40	0.28	0.26	0.23
	4	0.93	0.87	0.80	0.75	0.71	0.68	0.60	0.45	0.41	0.39
	6	0.97	0.92	0.88	0.83	0.80	0.77	0.71	0.57	0.52	0.49
	8	0.99	0.96	0.93	0.89	0.86	0.83	0.78	0.66	0.60	0.58
	10	0.99	0.98	0.96	0.93	0.91	0.88	0.83	0.73	0.68	0.65

Table 7. $F1$ Scores of Window Range Queries for Dummy Trajectories (Taxi)

$halfT$ (h)	$halfL$ (km)	θ (°)									
		5	10	15	20	25	30	45	90	135	180
1	2	0.85	0.71	0.63	0.57	0.53	0.48	0.41	0.31	0.28	0.27
	4	0.92	0.84	0.77	0.72	0.69	0.65	0.57	0.45	0.41	0.39
	6	0.96	0.90	0.85	0.81	0.78	0.75	0.68	0.56	0.52	0.50
	8	0.98	0.94	0.90	0.87	0.84	0.82	0.76	0.65	0.60	0.59
	10	0.98	0.97	0.94	0.91	0.89	0.87	0.82	0.71	0.67	0.66
3	2	0.87	0.75	0.68	0.63	0.59	0.55	0.54	0.43	0.40	0.39
	4	0.93	0.86	0.81	0.77	0.74	0.71	0.68	0.57	0.55	0.54
	6	0.97	0.93	0.88	0.85	0.82	0.79	0.76	0.67	0.64	0.63
	8	0.98	0.96	0.93	0.89	0.87	0.85	0.82	0.74	0.72	0.70
	10	0.99	0.98	0.96	0.93	0.91	0.89	0.86	0.80	0.77	0.76

Table 8. $F1$ Scores of Window Range Queries for Dummy Trajectories (Truck)

$halfT$ (h)	$halfL$ (km)	θ (°)									
		5	10	15	20	25	30	45	90	135	180
1	2	0.69	0.68	0.63	0.58	0.52	0.48	0.40	0.32	0.28	0.25
	4	0.74	0.70	0.65	0.61	0.59	0.56	0.57	0.41	0.37	0.33
	6	0.76	0.72	0.69	0.66	0.63	0.61	0.56	0.49	0.45	0.43
	8	0.77	0.74	0.71	0.69	0.66	0.65	0.60	0.53	0.50	0.50
	10	0.80	0.78	0.77	0.75	0.73	0.72	0.68	0.59	0.56	0.54
3	2	0.76	0.69	0.63	0.59	0.53	0.48	0.42	0.30	0.28	0.26
	4	0.80	0.74	0.70	0.66	0.63	0.60	0.54	0.42	0.39	0.36
	6	0.80	0.77	0.73	0.70	0.68	0.66	0.60	0.52	0.48	0.47
	8	0.81	0.78	0.76	0.73	0.71	0.69	0.65	0.58	0.55	0.54
	10	0.83	0.82	0.81	0.79	0.78	0.76	0.73	0.64	0.61	0.60

query cube with $halfT = 1$ h and $halfL = 2$ km, the rotation angles need to be no larger than 5° for Ucar and Taxi, to reach an $F1$ score greater than 0.80, and there may even have no choice to reach such a high $F1$ score for Truck.

Exp-9: *k-Anonymity*. In this test, we evaluate the impacts of parameter k for k-anonymity, the half time period and the half length of the cubes for window range queries. We vary k to $\{2, 4, 6, 8, 10\}$, respectively, to evaluate its impacts. The impacts of the cube are also tested by varying the half time period and the half length following Exp-7. The results are reported in Tables 9–11.

The results show that the $F1$ scores decrease with the increment of k, as the positions in the trajectories of individual objects become more inaccurate for larger k values such that the trajectories originally passing through the cubes may not pass through them any more. The same impacts of the cubes as Exp-7 are also confirmed, i.e., the $F1$ scores increase with the increment of both the half time period and the half length.

Table 9. $F1$ Scores of Window Range Queries for k-Anonymity (UCar)

$halfT$ (h)	$halfL$ (km)	k				
		2	4	6	8	10
1	2	0.36	0.24	0.20	0.20	0.15
	4	0.55	0.42	0.36	0.35	0.33
	6	0.64	0.50	0.50	0.47	0.47
	8	0.68	0.59	0.57	0.55	0.54
	10	0.70	0.64	0.62	0.60	0.60
3	2	0.50	0.39	0.35	0.34	0.31
	4	0.69	0.58	0.56	0.53	0.52
	6	0.79	0.71	0.69	0.67	0.66
	8	0.83	0.78	0.76	0.75	0.74
	10	0.85	0.82	0.80	0.80	0.79

Table 10. $F1$ Scores of Window Range Queries for k-Anonymity (Taxi)

$halfT$ (h)	$halfL$ (km)	k				
		2	4	6	8	10
1	2	0.52	0.41	0.34	0.32	0.27
	4	0.71	0.61	0.56	0.54	0.51
	6	0.81	0.74	0.71	0.69	0.68
	8	0.87	0.82	0.80	0.78	0.77
	10	0.91	0.86	0.84	0.84	0.82
3	2	0.63	0.50	0.44	0.40	0.36
	4	0.78	0.68	0.64	0.61	0.58
	6	0.85	0.79	0.76	0.75	0.72
	8	0.90	0.85	0.83	0.82	0.80
	10	0.92	0.89	0.87	0.86	0.85

Table 11. $F1$ Scores of Window Range Queries for k-Anonymity (Truck)

$halfT$ (h)	$halfL$ (km)	k				
		2	4	6	8	10
1	2	0.71	0.48	0.39	0.39	0.40
	4	0.82	0.69	0.62	0.63	0.60
	6	0.89	0.83	0.77	0.75	0.74
	8	0.92	0.86	0.84	0.85	0.84
	10	0.94	0.93	0.91	0.91	0.91
3	2	0.73	0.52	0.44	0.43	0.42
	4	0.84	0.74	0.64	0.65	0.63
	6	0.90	0.84	0.79	0.78	0.77
	8	0.93	0.88	0.86	0.86	0.85
	10	0.95	0.94	0.92	0.91	0.91

These also tell us that given a window range query, we need to properly choose k in order to reach a high utility. For instance, when given a cube with $halfT = 1$ h and $halfL = 6$ km, k needs to be 2 for Taxi and no more than 4 for Truck, to reach an $F1$ score greater than 0.80, and there may even have no choice to reach such a high $F1$ score for Ucar.

Exp-10: *ε-Differential Privacy.* In this test, we evaluate the impacts of the key parameter k used for partitioning for ε-differential privacy, the half time period and the half length of the cubes for window range queries. We vary k to $\{20, 40, 60, 80, 100\}$, respectively, to evaluate its impacts. The impacts of the cubes are also tested by varying the half time period and the half length following Exp-7. Note that we calculate the mean probability of a trajectory based on the probabilities of the $\varphi + 1 + |\Gamma|$ groups, which is used to compute the weighted $F1$ scores for window range queries.

In our tests, the mean probability is always close to 1, and the $F1$ scores for all cases are close to 1. Compared with k-anonymity that clusters trajectories, the ε-differential privacy approach [31] clusters locations at each time, which significantly improves the chances of trajectories originally passing through the cubes to pass through the cubes again. This also brings ε-differential privacy the capability to accurately support window range queries.

4.3 Summarization and Analyses

In this subsection, we summarize the findings on the individual privacy of anonymization mechanisms in terms of unicity, and the utility in terms of travel time estimation and window range queries, where Table 12 summarizes the main findings of experimental results.

4.3.1 Privacy with Unicity

For the identifier anonymization [13] and dummy trajectories [35] with the rotation pattern scheme, unicity μ is always kept to 1 on all tested datasets, as these two mechanisms essentially have no impact on the unicity.

For grid-based generalization [14], unicity μ varies from 0.3 to 1.0, and is affected by the number of points, the number of trajectories, and the spatial and temporal resolutions. Moreover, the unicity decrease can be easily eliminated by collecting a few more points, i.e., larger p. Essentially, four points are enough to uniquely reidentify all considered trajectories for Ucar ($\mu > 0.99$), and eight points are needed for Taxi ($\mu > 0.99$) and Truck ($\mu > 0.80$) respectively.

For k-anonymity [10], the unicity μ is always 0, which can be easily inferred from the definition of k-anonymity, and the unicity remains close to 0 even if

Table 12. Summary of Experimental Results

Method	Dataset [5, 6]	Privacy (Unicity) [1]	Utility		
			Travel Time Estimation		Window Range Queries [38]
			TSC [36]	TEMP+R [37]	
Identifier [13]	Ucar	×	✓	✓	✓
	Taxi	×	✓	✓	✓
	Truck	×	✓	✓	✓
Grid-based [14]	Ucar	×	×	×	○
	Taxi	×	×	×	○
	Truck	×	×	×	○
Dummy [35]	Ucar	×	×	✓	○
	Taxi	×	×	✓	○
	Truck	×	×	✓	○
K-anonymity [10]	Ucar	✓	×	×	○
	Taxi	✓	×	×	○
	Truck	✓	×	×	○
Differential [31]	Ucar	✓	⊗	⊗	✓
	Taxi	✓	⊗	⊗	✓
	Truck	✓	⊗	⊗	✓

Note: Here ✓and × represent "success" and "failure", respectively, ○ means "partially success" that proper choices are needed to reach a good utility and may not have opportunities to reach, and ⊗ means the method fails partially due to the limitations of tested methods.

we keep the trashed trajectories as 2, a parameter that may have impacts on the unicity.

For ε-differential privacy [31], the unicity is always close to 0, mainly because it has a grouping procedure of locations such that each location group uses its centroid to represent all its locations.

These imply that the reidentification privacy in terms of unicity is not well protected by identifier anonymization, dummy trajectories and grid-based generalization, but is well preserved by k-anonymity and differential privacy. This is in particular consistent with the findings of both De Montjoye *et al.* [2, 3] based on identifier anonymization and grid-based generalization and Sánchez *et al.* [21] based on k-anonymity. Our findings also confirm that well-established existing anonymization mechanisms (e.g., k-anonymity and differential privacy) can effectively protect individual privacy in terms of reidentification attacks.

This is somehow similar to Sánchez *et al.*'s finding [21], and the anonymization mechanisms used in De Montjoye *et al.* [2, 3] indeed have limitations and their finding on reidentification is indeed overestimated. This answers question 1 on the true situation of the privacy preservation for trajectories in terms of reidentification, and we also hopefully close the debate between De Montjoye *et al.* [2, 3] and Sánchez *et al.* [21] through our systematic evaluation.

However, we should remember that there are more than reidentification attacks and multi-source data may be used for attackers [11, 12, 23–26], and hence there is a long way to go for the privacy preservation of trajectories in the general sense.

4.3.2 *Utility with Travel Time Estimation*

For identifier anonymization [13], it obviously has no impact on travel time estimation.

For grid-based generalization [14], both TSC [36] and TEMP+R [37] fail for travel time estimation.

For dummy trajectories [35], 1) the successful estimated ratios of TSC decrease by (65%, 60%, 41%) on (Ucar, Taxi, Truck) on average, and its errors (mean relative errors and mean absolute errors) of TSC increase by $(245\%, 220\%, 143\%)$ on (Ucar, Taxi, Truck) on average, and 2) the successful estimated ratios of TEMP+R increase by $(6\%, 6\%, 2\%)$ on (Ucar, Taxi, Truck), and its errors (mean relative errors and mean absolute errors) of TSC increase by $(9.5\%, 18\%, 34.5\%)$ on (Ucar, Taxi, Truck) on average. That is, TSC fails for estimating the travel time, but TEMP+R is successful for estimating the travel time.

For k-anonymity [10], 1) the successful estimated ratios of TSC and TEMP+R decrease by $(80\%, 81\%, 83\%)$ and $(97\%, 93\%, 77\%)$ on (Ucar, Taxi, Truck) on average, respectively, and 2) their errors (mean relative errors and mean absolute errors) increase by $(157\%, 107.5\%, 528.5\%)$ and $(178.5\%, 233\%, 241\%)$ on

(Ucar, Taxi, Truck) on average, respectively. That is, TSC and TEMP+R fail for estimating the travel time.

For ε-differential privacy [31], both TSC and TEMP+R fail for estimating the travel time majorly because during the trajectory grouping involved in the ε-differential privacy, we coarsen input trajectories to make computation practical.

4.3.3 *Utility with Window Range Queries*

For identifier anonymization [13], it obviously has no impact on window range queries.

For grid-based generalization [14], the sizes of query cubes and spatial and temporal resolutions all have impacts on the accuracy of window range queries. Given a window range query, we need to properly choose the spatial and temporal resolutions in order to reach a high utility (e.g., high $F1$ scores). For instance, when given a query cube with $halfT = 1$ h and $halfL = 2$ km, the spatial and temporal resolutions need to satisfy (1 h, $\leqslant$ 4 km) or ($\leqslant$ 3 h, 1 km) for Ucar and Taxi, and (1 h, $\leqslant$ 2 km) or ($\leqslant$ 6 h, 1 km) for Truck, respectively, to reach an $F1$ score greater than 0.80.

For dummy trajectories [35], the sizes of query cubes and rotation angles all have impacts on the accuracy of window range queries. Given a window range query, we need to properly choose the rotation angles in order to reach a high utility. For instance, when given a query cube with $halfT = 1$ h and $halfL = 2$ km, the rotation angles need to be no larger than 5° for Ucar and Taxi, to reach a $F1$ score greater than 0.80, and there may even have no choice to reach such a high $F1$ score for Truck.

For k-anonymity [10], the sizes of query cubes and k all have impacts on the accuracy of window range queries. Given a window range query, we need to properly choose k values in order to reach a high utility. For instance, when given a query cube with $halfT = 1$ h and $halfL = 6$ km, k needs to be 2 for Taxi and no more than 4 for Truck, to reach an $F1$ score greater than 0.80, and there may even have no choice to reach such a high $F1$ score for Ucar.

For ε-differential privacy [31], the $F1$ scores of window range queries for all cases are close to 1. This also brings ε-differential privacy the capability to accurately support window range queries.

From the privacy evaluation with unicity and utility using practical applications (i.e., travel time estimation and window range queries), we find the followings. 1) There are no anonymization mechanisms, maybe except identifier anonymization, for the trajectory data that successfully satisfy all the needs of practical applications. 2) For grid-based anonymization, dummy trajectories and k-anonymity, the utility of window range queries appears to be partially successful by carefully adjusting the corresponding settings. However, the sizes of query cubes may not be fixed in practice, which significantly limits their de facto utility. 3) Except ε-differential privacy for window range queries, no anonymization mechanisms for trajectory data can successfully achieve a good trade-off between privacy and utility.

While the privacy is typically only determined by anonymization mechanisms, the utility is determined by both anonymization mechanisms and concrete application algorithms for trajectory data that are typically developed independently. Further, different applications may have different privacy and utility requirements. These make it hard to reach a trade-off between privacy and utility. We argue that efforts remain needed for designing better application algorithms that are more tolerable to anonymization mechanisms and better anonymization mechanisms that have less impacts on utility. These give an answer to question 2 on the true situation of the utility of anonymized trajectories.

5 Conclusions

We systematically evaluated the individual privacy in terms of unicity and the utility in terms of practical applications of the anonymized trajectory data. We confirmed Sánchez *et al.*'s finding [21], as k-anonymity and ε-differential privacy preserve well the reidentification privacy of trajectory data. This revealed the true situation of the privacy preservation for trajectories in terms of reidentification, and essentially closed the debate between De Montjoye *et al.* [2, 3] and Sánchez *et al.* [21]. We also found that the utility of existing anonymization mechanisms is not optimistic (especially when both privacy and utility are considered), and there is a long way to go for the privacy preservation for trajectories in the general sense. This revealed the true situation of the utility of anonymized trajectories.

Finally, we argued that efforts remain needed for designing better application algorithms that are more tolerable to anonymization mechanisms and better anonymization mechanisms (such as deep generative models to generate fake trajectories) that have less impacts on utility in the future.

References

[1] De Montjoye Y A, Hidalgo C A, Verleysen M, Blondel V D. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 2013, 3(6): Article No. 1376. DOI: 10.1038/srep01376.

[2] De Montjoye Y A D, Radaelli L, Singh V K, Pentland A S. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 2015, 347(6221): 536-539. DOI: 10.1126/science.12562.

[3] De Montjoye Y A D, Pentland A S. Response to comment on "unique in the shopping mall: On the reidentifiability of credit card metadata". *Science*, 2016, 351(6279): 1274. DOI: 10.1126/science.aaf15.

[4] Rocher L, Hendrickx J M, De Montjoye Y A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 2019, 10(1): Article No. 3069. DOI: 10.1038/s41467-019-10933-3.

[5] Lin X, Ma S, Zhang H, Wo T, Huai J. One-pass error bounded trajectory simplification. *Proceedings of the VLDB Endowment*, 2017, 10(7): 841-852. DOI: 10.14778/3067421.3067432.

[6] Lin X, Jiang J, Ma S, Zuo Y, Hu C. One-pass trajectory simplification using the synchronous Euclidean distance. *The VLDB Journal*, 2019, 28(6): 897-921. DOI: 10.1007/s00778-019-00575-8.

[7] Lin X, Ma S, Jiang J, Hou Y, Wo T. Error bounded line simplification algorithms for trajectory compression: An experimental evaluation. *ACM Trans. Database Syst.*, 2021, 46(3): Article No. 11. DOI: 10.1145/3474373.

[8] Zaeem R N, Barber K S. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Trans. Manag. Inf. Syst.*, 2021, 12(1): Article No. 2. DOI: 10.1145/3389685.

[9] Wicker S B. The loss of location privacy in the cellular age. *Communications of the ACM*, 2012, 55(8): 60-68. DOI: 10.1145/2240236.2240255.

[10] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proc. the 24th IEEE International Conference on Data Engineering*, April 2008, pp.376-385. DOI: 10.1109/ICDE.2008.4497446.

[11] Fung B C M, Wang K, Chen R, Yu P S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 2010, 42(4): Article No. 14. DOI: 10.1145/1749603.1749605.

[12] Chow C, Mokbel M F. Privacy of spatial trajectories. In *Computing with Spatial Trajectories*, Zheng Y, Zhou X (eds.), Springer, 2011, pp.109-141. DOI: 10.1007/978-1-4614-1629-6_4.

[13] Schwartz P M, Solove D J. Reconciling personal information in the United States and European Union. *California Law Review*, 2014, 102(4): 877-916. DOI: 10.2139/ssrn.2271442.

[14] Gidófalvi G, Huang X, Pedersen T B. Privacy-preserving data mining on moving object trajectories. In *Proc. the 2007 International Conference on Mobile Data Management*, May 2007, pp.60-68. DOI: 10.1109/MDM.2007.18.

[15] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services. In *Proc. the 2005 International Conference on Pervasive Services*, July 2005, pp.88-97. DOI: 10.1109/PERSER.2005.1506394.

[16] Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, 10(5): 557-570. DOI: 10.1142/S0218488502001648.

[17] Zhao K, Tu Z, Xu F, Li Y, Member S, Zhang P, Pei D, Su L, Jin D. Walking without friends: Publishing anonymized trajectory dataset without leaking social relationships. *IEEE Transactions on Network and Service Management*, 2019, 16(3): 1212-1225. DOI: 10.1109/TNSM.2019.2907542.

[18] Gursoy M E, Liu L, Truex S, Yu L. Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing*, 2019, 18(10): 2315-2329. DOI: 10.1109/TMC.2018.2874008.

[19] He X, Cormode G, Machanavajjhala A, Procopiuc C M, Srivastava D. DPT: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 2015, 8(11): 1154-1165. DOI: 10.14778/2809974.2809978.

[20] Andrés M E, Bordenabe N E, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: Differential privacy for location-based systems. In *Proc. the 2013 ACM SIGSAC Conference on Computer and Communications Security*, Nov. 2013, pp.901-914. DOI: 10.1145/2508859.2516735.

[21] Sánchez D, Martínez S, Domingo-Ferrer J. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science*, 2016, 351(6279): 1274. DOI: 10.1126/science.aad9295.

[22] Xiao Z, Wang C, Han W, Jiang C. Unique on the road: Re-identification of vehicular location-based metadata. In *Proc. the 12th International Conference on Security and Privacy in Communication Networks*, Oct. 2016, pp.496-513. DOI: 10.1007/978-3-319-59608-2_28.

[23] Chatzikokolakis K, ElSalamouny E, Palamidessi C, Pazii A. Methods for location privacy: A comparative overview. *Found. Trends Priv. Secur.*, 2017, 1(4): 199-257. DOI: 10.1561/3300000017.

[24] Henriksen-Bulmer J, Jeary S. Re-identification attacks—A systematic literature review. *Int. J. Inf. Manag.*, 2016, 36(6): 1184-1192. DOI: 10.1016/j.ijinfomgt.2016.08.002.

[25] Wagner I, Eckhoff D. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.*, 2018, 51(3): Article No. 57. DOI: 10.1145/3168389.

[26] Primault V, Boutet A, Mokhtar S B, Brunie L. The long road to computational location privacy: A survey. *IEEE Commun. Surv. Tutorials*, 2019, 21(3): 2772-2793. DOI: 10.1109/COMST.2018.2873950.

[27] Peters F, Menzies T, Gong L, Zhang H. Balancing privacy and utility in cross-company defect prediction. *IEEE Trans. Software Eng.*, 2013, 39(8): 1054-1068. DOI: 10.1109/TSE.2013.6.

[28] Xu J, Wang W, Pei J, Wang X, Shi B, Fu A W. Utility-based anonymization using local recoding. In *Proc. the*

12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2006, pp.785-790. DOI: 10.1145/1150402.1150504.

[29] Jr R J B, Agrawal R. Data privacy through optimal k-anonymization. In *Proc. the 21st International Conference on Data Engineering*, April 2005, pp.217-228. DOI: 10.1109/ICDE.2005.42.

[30] Peters F, Menzies T. Privacy and utility for defect prediction: Experiments with MORPH. In *Proc. the 34th International Conference on Software Engineering*, June 2012, pp.189-199. DOI: 10.1109/ICSE.2012.6227194.

[31] Hua J, Gao Y, Zhong S. Differentially private publication of general time-serial trajectory data. In *Proc. the 2015 IEEE Conference on Computer Communications*, April 26-May 1, 2015, pp.549-557. DOI: 10.1109/INFOCOM.2015.7218422.

[32] Cunha M, Mendes R, Vilela J P. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, 2021, 41: Article No. 100403. DOI: 10.1016/j.cosrev.2021.100403.

[33] Casas-Roma J. DUEF-GA: Data utility and privacy evaluation framework for graph anonymization. *International Journal of Information Security*, 2020, 19(4): 465-478. DOI: 10.1007/s10207-019-00469-4.

[34] Ni C, Cang L S, Gope P, Min G. Data anonymization evaluation for big data and IoT environment. *Information Sciences*, 2022, 605: 381-392. DOI: 10.1016/j.ins.2022.05.040.

[35] You T, Peng W, Lee W. Protecting moving trajectories with dummies. In *Proc. the 2007 International Conference on Mobile Data Management*, May 2007, pp.278-282. DOI: 10.1109/MDM.2007.58.

[36] Wang Y, Zheng Y, Xue Y. Travel time estimation of a path using sparse trajectories. In *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2014, pp.25-34. DOI: 10.1145/2623330.2623656.

[37] Wang H, Tang X, Kuo Y, Kifer D, Li Z. A simple baseline for travel time estimation using large-scale trip data. *ACM Trans. Intell. Syst. Technol.*, 2019, 10(2): Article No. 19. DOI: 10.1145/3293317.

[38] Eldawy A, Alarabi L, Mokbel M F. Spatial partitioning techniques in spatial Hadoop. *Proceedings of the VLDB Endowment*, 2015, 8(12): 1602-1605. DOI: 10.14778/2824032.2824057.

[39] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In *Proc. the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2012, pp.186-194. DOI: 10.1145/2339530.2339561.

[40] Jiang K, Shao D, Bressan S, Kister T, Tan K. Publishing trajectories with differential privacy guarantees. In *Proc. the 25th International Conference on Scientific and Statistical Database Management*, July 2013, Article No. 12. DOI: 10.1145/2484838.2484846.

[41] Nergiz M E, Atzori M, Saygin Y, Güç B. Towards trajectory anonymization: A generalization-based approach. *Trans. Data Privacy*, 2009, 2(1): 47-75.

[42] Zhang C, Han J, Shou L, Lu J, Porta T L. Splitter: Mining fine-grained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 2014, 7(9): 769-780. DOI: 10.14778/2732939.2732949.

[43] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. the 23rd IEEE International Conference on Data Engineering*, April 2007. DOI: 10.1109/ICDE.2007.367856.

[44] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 2007, 1(1): Article No. 3. DOI: 10.1145/1217299.1217302.

[45] Abul O, Bonchi F, Nanni M. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 2010, 35(8): 884-910. DOI: 10.1016/j.is.2010.05.003.

[46] Trujillo-Rasua R, Domingo-Ferrer J. On the privacy offered by (k, δ)-anonymity. *Information Systems*, 2013, 38(4): 491-494. DOI: 10.1016/j.is.2012.12.003.

[47] Dwork C, McSherry F, Nissim K, Smith A D. Calibrating noise to sensitivity in private data analysis. In *Proc. the 3rd Theory of Cryptography Conference*, March 2006, pp.265-284. DOI: 10.1007/11681878_14.

[48] Dwork C, Roth A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014, 9(3/4): 211-407. DOI: 10.1561/0400000042.

[49] McSherry F, Talwar K. Mechanism design via differential privacy. In *Proc. the 48th Annual IEEE Symposium on Foundations of Computer Science*, Oct. 2007, pp.94-103. DOI: 10.1109/FOCS.2007.66.

[50] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proc. the ACM SIGMOD International Conference on Management of Data*, June 29-July 2, 2009, pp.19-30. DOI: 10.1145/1559845.1559850.

[51] Chen R, Fung B C M, Desai B C. Differentially private trajectory data publication. arXiv:1112.2020, 2011. https://arxiv.org/abs/1112.2020, July 2022.

[52] Yao L, Chen Z, Hu H, Wu G, Wu B. Privacy preservation for trajectory publication based on differential privacy. *ACM Trans. Intell. Syst. Technol.*, 2022, 13(3): Article No. 42. DOI: 10.1145/3474839.

[53] Yuan N J, Zheng Y, Zhang L, Xie X. T-finder: A recommender system for finding passengers and vacant taxis. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(10): 2390-2403. DOI: 10.1109/TKDE.2012.153.

[54] Yuan J, Zheng Y, Xie X, Sun G. T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(1): 220-232. DOI: 10.1109/TKDE.2011.200.

[55] Zhang D, Ding M, Yang D, Liu Y, Fan J, Shen H T. Trajectory simplification: An experimental study and quality analysis. *Proceedings of the VLDB Endowment*, 2018, 11(9): 934-946. DOI: 10.14778/3213880.3213885.

[56] Ali M E, Eusuf S S, Abdullah K, Choudhury F M, Culpepper J S, Sellis T. The maximum trajectory coverage query in spatial databases. *Proceedings of the VLDB Endowment*, 2018, 12(3): 197-209. DOI: 10.14778/3291264.3291266.

[57] Yuan H, Li G, Bao Z, Feng L. Effective travel time estimation: When historical trajectories over road networks matter. In *Proc. the 2020 ACM SIGMOD International Conference on Management of Data*, June 2020, pp.2135-2149. DOI: 10.1145/3318464.3389771.

[58] Shah D, Kumaran A, Sen R, Kumaraguru P. Travel time estimation accuracy in developing regions: An empirical case study with Uber data in Delhi-NCR*. In *Proc. Companion of the 2019 World Wide Web Conference*, May 2019, pp.130-136. DOI: 10.1145/3308560.3317057.

[59] Ma S, Yu Z, Wolfson O. T-share: A large-scale dynamic taxi ridesharing service. In *Proc. the 29th IEEE International Conference on Data Engineering*, April 2013, pp.410-421. DOI: 10.1109/ICDE.2013.6544843.

[60] Wang Y, Lin X, Wei H, Wo T, Huang Z, Zhang Y, Xu J. A unified framework with multi-source data for predicting passenger demands of ride services. *ACM Transactions on Knowledge Discovery from Data*, 2019, 13(6): Article No. 56. DOI: 10.1145/3355563.

[61] Li Y, Fu K, Wang Z, Shahabi C, Ye J, Liu Y. Multi-task representation learning for travel time estimation. In *Proc. the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2018, pp.1695-1704. DOI: 10.1145/3219819.3220033.

[62] Fang X, Huang J, Wang F, Zeng L, Liang H, Wang H. ConSTGAT: Contextual spatial-temporal graph attention network for travel time estimation at Baidu maps. In *Proc. the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, pp.2697-2705. DOI: 10.1145/3394486.3403320.

[63] Wang L, Ma W, Fan Y, Zuo Z. Trip chain extraction using smartphone-collected trajectory data. *Transportmetrica B: Transport Dynamics*, 2019, 7(1): 255-274. DOI: 10.1080/21680566.2017.1386599.

[64] Newson P, Krumm J. Hidden Markov map matching through noise and sparseness. In *Proc. the 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, Nov. 2009, pp.336-343. DOI: 10.1145/1653771.1653818.

[65] Cao H, Wolfson O, Trajcevski G. Spatiotemporal data reduction with deterministic error bounds. *The VLDB Journal,* 2006, 15(3): 211-228. DOI: 10.1007/s00778-005-0163-7.

[66] Gao Z, Zhai R, Wang P, Yan X, Qin H, Tang Y, Ramesh B. Synergizing appearance and motion with low rank representation for vehicle counting and traffic flow analysis. *IEEE Transactions on Intelligent Transportation Systems*, 2018, 19(8): 2675-2685. DOI: 10.1109/TITS.2017.2757040.

[67] Zang H, Bolot J. Anonymization of location data does not work: A large-scale measurement study. In *Proc. the 17th Annual International Conference on Mobile Computing and Networking*, Sept. 2011, pp.145-156. DOI: 10.1145/2030613.2030630.

[68] Shokoohyar S, Sobhani A, Sobhani A. Impacts of trip characteristics and weather condition on ride-sourcing network: Evidence from Uber and Lyft. *Research in Transportation Economics*, 2020, 80: Article No. 100820. DOI: 10.1016/j.retrec.2020.100820.

She Sun received his B.S. degree in environment science from Northwestern Polytechnical University, Xi'an, in 2013, and his M.S. degree in applied mathematics from Northwestern Polytechnical University, Xi'an, in 2016. He is working toward his Ph.D. degree in the School of Computer Science and Engineering, Beihang University, Beijing. His current research include trajectory privacy and differential privacy.

Shuai Ma received his Ph.D. degrees in computer science from the University of Edinburgh, Edinburgh, in 2010, and from Peking University, Beijing, in 2004, respectively. He is currently a professor with the School of Computer Science and Engineering, Beihang University, Beijing. He was a postdoctoral research fellow with the Database Group, University of Edinburgh, a summer intern with Bell Labs, Murray Hill, and a visiting researcher of MSRA, Beijing. He is a recipient of the Best Paper Award for VLDB 2010, the Best Challenge Paper Award for WISE 2013 and ICDM 2019 Best Paper Candidate. He is/was an associate editor of the VLDB Journal, and IEEE Transactions on Big Data and Knowledge and Information Systems. His current research interests include database theory and systems, and big data. He is a senior member of IEEE.

Jing-He Song received his B.S. degree in quality and reliability system engineering from Beihang University, Beijing. He is working toward his Ph.D. degree in the School of Computer Science and Engineering, Beihang University, Beijing. His current research interests include database systems and data mining.

Wen-Hai Yue received his B.S. degree in software engineering from Southeast University, Nanjing. He is working toward his M.S. degree in software engineering, Beihang University, Beijing. His current research interests include database systems, data anonymization and differential privacy.

Xue-Lian Lin received his M.S. and Ph.D. degrees in computer science and technologies from Beihang University, Beijing, in 2002 and 2013 respectively. He is an associate professor with the School of Computer Science and Engineering, Beihang University, Beijing. His current research interests include large scale data management systems, data intensive computing, mobile computing, and time series analysis.

Tiejun Ma received his Ph.D. degree in computer science from the University of Edinburgh, Edinburgh, in 2007. He is an associate professor in the Centre for Risk Research, Department of Decision Analytics and Risk, Southampton Business School at the University of Southampton, Southampton. Before he joined the University of Southampton, he worked in the Department of Computing, Imperial College London, and the Oxford e-Research Centre, University of Oxford, Oxford. His research focuses on risk analysis and decision-making using quantitative modelling and real-time big data analysis techniques applies to FinTech, Cyber-Risk, and Resilience of distributed systems.