Computers, Materials & Continua CMC, vol.65, no.3, pp.2665-2685, 2020

A Differential Privacy Based (k-Ψ)-Anonymity Method for Trajectory Data Publishing

Hongyu Chen[1], Shuyu Li[1,*] and Zhaosheng Zhang[1]

Abstract: In recent years, mobile Internet technology and location based services have wide application. Application providers and users have accumulated huge amount of trajectory data. While publishing and analyzing user trajectory data have brought great convenience for people, the disclosure risks of user privacy caused by the trajectory data publishing are also becoming more and more prominent. Traditional k-anonymous trajectory data publishing technologies cannot effectively protect user privacy against attackers with strong background knowledge. For privacy preserving trajectory data publishing, we propose a differential privacy based (k-Ψ)-anonymity method to defend against re-identification and probabilistic inference attack. The proposed method is divided into two phases: in the first phase, a dummy-based (k-Ψ)-anonymous trajectory data publishing algorithm is given, which improves (k-δ)-anonymity by considering changes of threshold δ on different road segments and constructing an adaptive threshold set Ψ that takes into account road network information. In the second phase, Laplace noise regarding distance of anonymous locations under differential privacy is used for trajectory perturbation of the anonymous trajectory dataset outputted by the first phase. Experiments on real road network dataset are performed and the results show that the proposed method improves the trajectory indistinguishability and achieves good data utility in condition of preserving user privacy.

Keywords: Trajectory data publishing, privacy preservation, road network, (k-Ψ)-anonymity, differential privacy.

1 Introduction

In recent years, location-based services are widely applied to a large variety of mobile applications such as route navigation, social games, and Mobile Crowdsensing [Liu, Liu, Zheng et al. (2018); Xiao, Chen, Xie et al. (2018)] and so on. During the application of above location-based services, massive trajectory data of users is generated and collected. Trajectory data analysis has significant value for no matter governments, commercial organizations or individuals. For example, trajectory data analysis can help improve traffic safety and reduce traffic congestion [Xia, Hu and Luo (2017)]. For individuals, it can bring

[1] School of Computer Science, Shaanxi Normal University, Xi'an, 710119, China.

[*] Corresponding Author: Shuyu Li. Email: lishuyu@snnu.edu.cn.

Received: 10 April 2020; Accepted: 14 May 2020.

conveniences to daily life, like optimal route selection when a traffic jam occurs.

However, trajectory data is a sampling sequence of the moving object with position and time information. It usually contains rich explicit spatio-temporal information, and implicit features such as personal behavior patterns, frequent meet and future mobility [Feng and Zhu (2016)]. Directly publishing original trajectory data or publishing trajectory data without carefully taking privacy preservation into consideration may cause problems of privacy disclosure. Nowadays, users pay more attention to personal privacy concerns and the risks of privacy disclosure will severely reduce the enthusiasm of users to contribute his/her trajectory data. Thus privacy preserving data publishing become a hot topic in the data mining field, and its goal is to keep the utility of published data under the constraint of privacy protection. In general, the published data should meet the following two objectives: firstly, to ensure the attackers cannot infer the sensitive information of target individual with high probability. Secondly, to keep the published data still has good utility for third-party users to perform data analysis in condition of privacy preservation.

Since trajectory data is spatio-temporal correlated, high-dimension and context-aware, traditional privacy preserving techniques such as *k*-anonymity [Gruteser and Grunwald (2003)], are not well suited for privacy preserving trajectory data publishing. However, *k*-anonymity model aims at the pre-set attack, and may cannot resist other attacks. Compared with *k*-anonymity model, differential privacy [Dwork and Roth (2014)] is an unconditional privacy protection mechanism that can resist arbitrary attacks and it has been applied to protect trajectory data with two privacy definitions: event-level and user-level, the former protects any single event (e.g., a single location point), whereas the latter protects all the events of any user (e.g., the trajectory of any individual) [Wang, Zheng, Rehmani et al. (2018)].

In this paper, we combine both *k*-anonymity and differential privacy, propose a differential privacy based (*k*-Ψ)-anonymity method to address the privacy threat in trajectory data publishing and defend against malicious attackers with strong background knowledge. In summary, we make the following contributions in this paper:

i. We propose a two-stage trajectory data publishing method that enables the published data against re-identification and probabilistic inference attack. At the first stage, a dummy-based (*k*-Ψ)-anonymity algorithm is presented for the indistinguishability of the trajectory data. And at the second stage, Laplace noise is used for trajectory perturbation under differential privacy.

ii. We propose a dummy-based (*k*-Ψ)-anonymity algorithm. Different from (*k*-δ)-anonymity, the radius δ at each moment is adaptively determined by the query context and the motion model. On this basis, an adaptive threshold set Ψ that considers road network information is constructed for dummy trajectory generation.

iii. For trajectory perturbation, Laplace noise regarding distance of anonymous locations is used to ensure better privacy preservation. The experimental results on the real dataset show that the proposed method improves the trajectory indistinguishability and achieves good data utility in condition of preserving user privacy.

The rest of the paper is organized as follows. Section 2 presents the related works. Section 3 introduces system architecture of privacy preserving trajectory data publishing. Section 4 and Section 5 describe the (*k*-Ψ)-anonymity trajectory algorithm and differentially private trajectory perturbation algorithm respectively. The security analysis of the differential privacy based (*k*-Ψ)-anonymity algorithm is given in Section 6. Section 7 shows the experimental results and analysis of comparative experiments. Section 8 concludes this paper.

2 Related works

A recent survey divide the location privacy protection mechanisms into two scenarios (online or offline scenario) and six categories of techniques including mix-zones, generalization-based, dummy-based, perturbation-based, protocol-based and rule-based techniques [Primault, Boutet, Mokhtar et al. (2018)]. The online scenario mostly focuses on snapshot queries and its extension, that is, users query and want an immediate answer, e.g., Peng et al. [Peng, Liu, Meng et al. (2017)] proposed a collaborative trajectory privacy preserving (CTPP) scheme in continuous LBSs. The scheme can confuse the LBS adversary by issuing fake queries to obfuscate the actual trajectory of users. While the offline scenario refers to an LBS has collected users' trajectory data and want to publish it, whether it is for commercial or non-profit purposes, e.g., Dong et al. [Dong and Pi (2018)] proposed a privacy preserving trajectory data publishing algorithm based on frequent path to strike a balance between data utility and privacy. The workflow of the proposed algorithm is to remove infrequent roads in each trajectory firstly, and divide trajectories into candidate groups, then find the most frequent path and lastly select the representative trajectory to represent all trajectories within a group. Most of above mentioned techniques adopt *k*-anonymity, differential privacy or other privacy preservation model to satisfy requirements of privacy protection.

Various *k*-anonymity models can protect the trajectory data from trail re-identification attack by hiding the connection between user and the trajectory data, and the attackers cannot identify the identity of specific user or the user that the target trajectory belongs to. Based on the *k*-anonymity, Abul et al. [Abul, Bonchi and Nanni (2010)] presented (*k*-δ)-anonymity. They considered the perturbed trajectory of a moving object as a cylinder with radius δ, and the objects in the same cylinder are indistinguishable. A series of methods based on (*k*-δ)-anonymity are proposed for privacy preserving trajectory data publishing. *k*-anonymity cannot protect spatio-temporal trajectory data from the probabilistic attacks, so Gramaglia et al. [Gramaglia, Fiore, Tarable et al. (2017)] introduced $k^{\tau,\varepsilon}$ – anonymity to solve the probabilistic and record linkage attacks for mobile subscriber trajectory data. They discuss the optimal spatio-temporal generalization of *k* trajectories and propose the *k*-merge algorithm that generalizes trajectories with minimal loss of data granularity to guarantee $k^{\tau,\varepsilon}$ – anonymity. Tu et al. [Tu, Zhao, Xu et al. (2017)] introduced a novel attack in publishing trajectory datasets, namely semantic attack. Then for the objective of not only preventing individuals from being re-identified but more importantly protecting semantic information of the trajectory data, they proposed an algorithm to continuously merge trajectories from the original dataset, and obtain a new generalized dataset consisting of all merged trajectories. The

experimental results show that the algorithm achieves *k*-anonymity, *l*-diversity, and *t*-closeness of trajectories.

Differential privacy is an unconditional privacy protection mechanism which achieves privacy preservation by adopting perturbation-based techniques. Wang et al. [Wang and Xu (2017)] proposed a correlated time-series data publication solution based on differential privacy by enforcing Series-Indistinguishability and designing a correlated Laplace mechanism. Series-Indistinguishability guarantees the perturbed series and the original series is indistinguishable for an adversary, and the mechanism uses four Gauss white noise series to produce a correlated Laplace noise series. Based on the utility violation of the released trajectory data, Li et al. [Li, Zhu, Zhang et al. (2017)] proposed a differentially private trajectory data publishing scheme including a bounded Laplace noise generation algorithm and a trajectory merging algorithm. The experimental results show that the scheme can reduce trajectory merging time and have similar data utility with the works of Hua et al. [Hua, Gao and Zhong (2015)]. Gursoy et al. [Gursoy, Liu, Truex et al. (2018)] proposed DP-Star method for publishing trajectory data with differential privacy guarantee and high utility. DP-Star uses the MDL (Minimum Description Length) metric to summarize raw trajectories, and construct a density-aware grid to ensure spatial densities. DP-Star preserves the correlations between trajectories' end points through a private trip distribution, and intermediate points through a private Markov mobility model. At last, DP-Star estimates users' trip lengths using a median length estimation method, and generates synthetic trajectories that preserve both differential privacy and high utility. To solve the extreme sparseness problem of spatio-temporal data, Al-Hussaeni et al. [Al-Hussaeni, Fung, Iqbal et al. (2018)] proposed a solution combined with SafePath to model trajectories as a noisy prefix tree, with the goal of publishing differentially private trajectories while minimizing the impact on data utility. Ou et al. [Ou, Qin, Liao et al. (2018)] consider the mutual correlation between trajectories of two users may leak sensitive social relations, and propose a n-body Laplace framework to prevent social relations inference attack. Under the n-body Laplace framework, two Lagrange Multiplier-based Differentially Private (LMDP) approaches are proposed to optimize the privacy budgets, and the experimental results show that the proposed approaches achieve good privacy and data utility. All these DP-based schemes directly work on trajectory data, but in actual scenarios, the users' location data is stored discretely in a row format, so they are difficult to protect location data in data mining. To solve this problem, Gu et al. [Gu, Yang and Yin (2018)] propose a DP-based scheme to protect location data records. The scheme can query and publish location data on database by using a multi-level query tree structure, and it can provide a balance between data utility and privacy preservation.

3 System model

3.1 Attack model

In dummy-based trajectory data publishing, attackers try to infer the dummy trajectory to weaken privacy protection, or even identify the user's real trajectory. Thereby attackers can illegally get user's sensitive information.

Usually, the malicious attacker can be divided into two types: passive attackers and active attackers. A passive attacker is someone who collect user's information through interception without tampering, replaying or re-injecting data. In practice, some mature encryption methods [Shi, Wang, Zhu et al. (2019); Liu, Chen, Zhu et al. (2017)] such as public key encryption can be used to defend against such attack. An active attacker often takes cooperative attack or inference attack [Wang, Zheng, Rehmani et al. (2018)] as the common attack method, and both of above two type attacks can cause severe privacy disclosure. For example, Huo et al. [Huo, Meng and Zhang (2013)] propose a hidden location inference attack, and the adversaries can infer users' location based on users' historical check-in location data. Cooperative attack is often used as an up-front auxiliary part of inference attack, to get some fundamental data and use this acquired data as the input of inference attack.

In this paper, we will take both cooperative attack and inference attack into consideration, but mainly focus on inference attack.

The inference attack commonly uses probabilistic reasoning, which handles uncertain and probabilistic information and results a range of possible area. For *k*-anonymity method, the inference probability can be expressed as $P(V \to v \mid \Omega)$, where V denotes the anonymity dataset, v is the actual data of the target user. And Ω is the attribute set of the target user gained by the attacker, namely the background knowledge of the attacker, for example, time, location, interests and other attributes. The inference probability of *k*-anonymity is actually the probability of matching degree between the explicit identifier and the sensitive information. The relationship of inference probability is expressed as follows:

$$\underbrace{A_{EI} \xleftarrow{P(A_{QI} \to A_{EI})} A_{QI}}_{T-original\ data} \overbrace{\xleftarrow{P(A_{\overline{QI}} \to A_{QI})}}^{k-anonymity} \underbrace{A_{\overline{QI}} \xrightarrow{P(A_{\overline{QI}} \to A_{SI})} A_{SI}}_{T'-anonymity\ dataset} \tag{1}$$

According to above relationship, the inference probability of *k*-anonymity is:

$$P(A_{EI} \leftrightarrow A_{SI} \mid T \bigcup T') = P(A_{QI} \to A_{EI} \mid T) \times P(A_{\overline{QI}} \to A_{QI} \mid T') \times P(A_{\overline{QI}} \to A_{SI} \mid T') \tag{2}$$

where T is the original attribute dataset, T' is the anonymity attribute dataset, and A_{EI}, A_{QI} are the explicit identifier set and the quasi-identifier set respectively, $A_{\overline{QI}}$ is the anonymity quasi-identifier attribute set, A_{SI} is the anonymity sensitive attribute set, and $\Omega = T \bigcup T'$. $P(A_{EI} \leftrightarrow A_{SI} \mid T \bigcup T')$ is the probability that the explicit identifiers and the sensitive attributes can be derived from each other under the condition of attribute sets Ω. $P(A_{QI} \to A_{EI} \mid T)$ is the probability that the explicit identifiers can be derived according to the quasi-identifiers in T. $P(A_{\overline{QI}} \to A_{QI} \mid T')$ is the probability that the original quasi-identifiers can be derived according to the anonymity quasi-identifiers in T'. $P(A_{\overline{QI}} \to A_{SI} \mid T')$ is the probability that the anonymity sensitive attributes can be derived according to the anonymity quasi-identifiers in T'. Probability threshold is set in advance for each reasoning process, and inference is successful if inference probability exceeds the

threshold. Therefore in addition to guarantee the dataset with *k*-anonymity, considering features of probabilistic reasoning and improving the probabilistic reasoning process, may reduce the success rate of inference attack and enhance the capability of privacy preservation.

3.2 Architecture

The generalization-based trajectory *k*-anonymity technology can protect the user's privacy while guarantee the availability of published trajectory data. But when faced with, attackers with strong background knowledge, it still has the risk of privacy disclosure. Recently, differential privacy is widely used as a privacy protection technology in statistical queries, machine learning, etc. And it has been proven to be a good defense against attackers with strong background knowledge.

We adopt the centralized architecture [Primault, Boutet, Mokhtar et al. (2018)] and use differential privacy based (*k*-Ψ)-anonymity technology to publish trajectory data. The proposed architecture consists of three parts: the trajectory data provider (individual user), the trusted third party, and the data consumer, as shown in Fig. 1. Among them, the tasks completed by the third party include trajectory data collection, trajectory data preprocessing and trajectory data processing to satisfy privacy requirements.

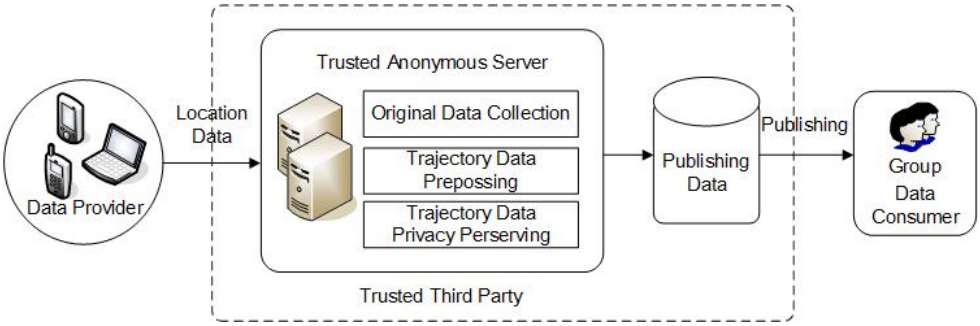


Figure 1: System architecture of privacy preserving trajectory data publishing

3.3 Related concepts

Definition 1 (Trajectory): The trajectory is a spatio-temporal sequence, which consists of the locations of the user over a period of time, it can be expressed as: $trail = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \cdots, (x_n, y_n, t_n)\}$, where (x_i, y_i, t_i) is the spatial location at time t_i, $1 \le i \le n$. As shown in Fig. 2, t1->t2->t3->t4->t5->t6->t7 is a trajectory of certain user in the road network.

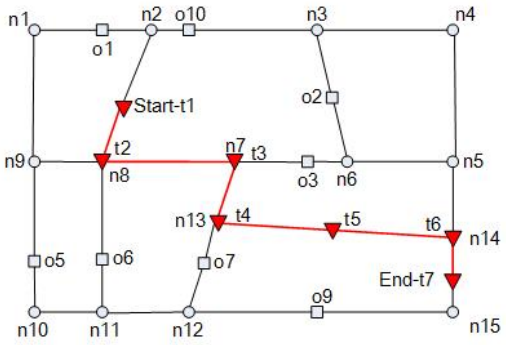


Figure 2: Trajectory of road network

Definition 2 (Trajectory Set): When the trajectory data is publishing, the $k-1$ dummy trajectory generated by the anonymity server and the real trajectory $trail_{real}$ constitute a trajectory set $Trails=\{trail_1, trail_2, \cdots, trail_{k-1}, trail_{real}\}$, $|Trails|$ is the number of elements in the set $Trails$.

Definition 3 (Trajectory (k-Ψ)-Anonymity): For any trajectory in the (k-Ψ)-anonymity sets S which contains the real one $trail_{real}$, there are at least k-1 other trajectories indistinguishable in time and location, where Ψ is a set of uncertain threshold $\delta_i (1 \le i \le k)$.

4 (k-Ψ)-Anonymity trajectory data publishing

4.1 (k-Ψ)-Anonymity trajectory model

The real trajectory $trail_{real}$ in three-dimensional space is denoted as $\{(x_1, y_1, t_1), (x_2, y_2, t_2), \cdots, (x_n, y_n, t_n)\}$, and other $k-1$ dummy trajectories need to be generated based on δ rule to achieve (k-δ)-anonymity. Usually, the dummy trajectories method firstly generate $k-1$ dummy locations at each time t_i during the time period $\{t_1, t_2, \cdots, t_n\}$. Then selecting a dummy location at time t_i respectively to synthetize $k-1$ dummy trajectories. At time t_i, we use $\{(x_i^1, y_i^1, t_i), (x_i^2, y_i^2, t_i), \cdots, (x_i^{k-1}, y_i^{k-1}, t_i)\}$ to denote the $k-1$ dummy locations, and the distance between each of these dummy locations and the true location should be less than δ. However, deciding the uncertain circle that centers on the real location and has a radius δ to generate dummy locations is not enough. If the road network and the moving object are not taken into consideration, there may exists invalid dummy locations in the $k-1$ dummy location sets, and these invalid ones can be easily identified by the attacker. Therefore, we process the trajectory based on the moving object model in the road network to achieve (k-δ)-anonymity.

The threshold δ of all trajectories is same in default. For better workability of anonymity and adapting to road changes of different segments in the road network, different thresholds are adopted and δ_i is the threshold at time t_i. We use $\psi=\{\delta_1, \delta_2, ..., \delta_n\}$ to denote the set of these uncertain thresholds.

Definition 4 (Approximate Trajectory): In the discrete time range $[t_1, t_n]$, if and only if at any time $t_i \in [t_1, t_n]$, any spatio-temporal point (x_{1i}, y_{1i}, t_i) on the trajectory τ_1 and any spatio-temporal point (x_{2i}, y_{2i}, t_i) on the trajectory τ_2 satisfy $Dist((x_{1i}, y_{1i}),(x_{2i}, y_{2i})) \le 2\delta_i$, the two trajectories τ_1 and τ_2 are approximate trajectories with each other, where $1 \le i \le n$ and

$$Dist((x_{1i}, y_{1i}),(x_{2i}, y_{2i})) = \sqrt{(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2} \tag{3}$$

The approximate trajectories τ_1 and τ_2 are represented as $Similar_{\delta_i}(\tau_1, \tau_2)$ with given threshold δ_i, they can be mutually possible moving curves. According to Eq. (3), any two approximate trajectories in the uncertain trajectory dataset are indistinguishable. That is, if τ_1 and τ_2 are approximate trajectories, when the attacker infers τ_1 based on the background knowledge, τ_2 will also be the trajectory recognized. Since τ_2 is also in the cylinder $Vol(\tau_1, \psi)$ composed of the uncertain trajectory dataset, and it can be the possible moving curve of τ_1. Consequently, the attacker is unable to distinguish the approximate trajectories τ_1 and τ_2.

Definition 5 (Trajectory Anonymity Set): For a given threshold set Ψ, the trajectories set *Trails* is called an anonymity set if and only if any two trajectories in *Trails* are approximate trajectories. That is:

$$\forall \tau_i, \tau_j \in Trails,\ Similar_{\psi}(\tau_i, \tau_j) \tag{4}$$

Definition 6 ((*k*-Ψ)-Anonymity of Trajectories): For the given trajectories set *Trails*, degree of anonymity *k*, and the threshold set Ψ, *Trails* achieve (*k*-Ψ)-anonymity if and only if there exists a *k*-anonymity set $S \subseteq Trails$ for each trajectory $\tau \in Trails$, which satisfies $\tau \in S$ and $|S| \ge k$.

Fig. 3 is a three-dimensional space map of an anonymity set of trajectories. In the figure, the degree of anonymity *k* is 3. There are three trajectories including a real one and two dummy trajectories, and every two among them are approximate trajectories. In the irregular cylinder $Vol(\tau_1, \psi)$ composed of the uncertain trajectory dataset, a circle with a radius of δ_i at each time t_i can be drawn, and the location of the real trajectory and the locations of the dummy trajectories are all within this uncertain circle. Taking the road network model into consideration is to eliminate invalid dummy locations, and let the generated dummy locations are all in the vicinity of the road network segment, which makes the dummy locations and the real location more difficult to distinguish.

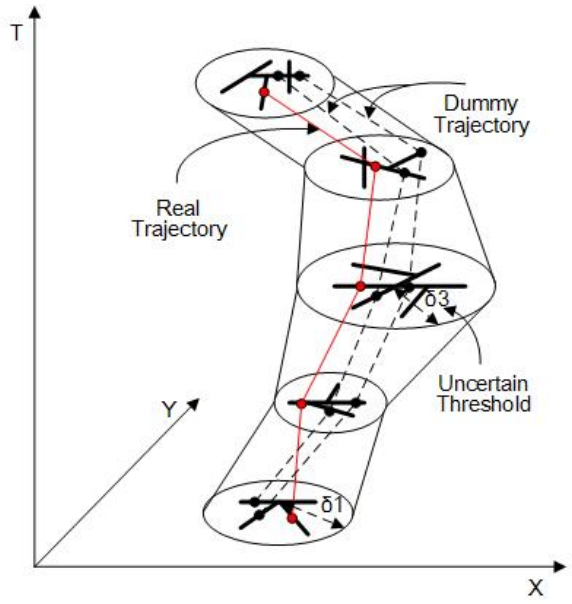


Figure 3: (k-Ψ)-anonymity set of trajectories

4.2 (k-Ψ)-Anonymity trajectory algorithm

(k-Ψ)-anonymity trajectory algorithm generates $k-1$ dummy trajectories based on the real trajectory of the user and a threshold set Ψ. The real trajectory of the user is $trail_{real}=(l_1^{real},l_2^{real},\cdots,l_n^{real})$, and $k-1$ dummy locations are generated based on the real location in the trajectory at each time t_i, we define $L_i=(l_i^1,l_i^2,\cdots,l_i^{k-1},l_i^{real})$ as the set of k locations at time t_i. In order to guarantee the validity of the dummy locations, we used a dummies generation method called TreeGenerate [Li and Li (2018)]. This method firstly generates a valid dummy edges set $Q_i=\{e_1,e_2,\cdots,e_{k-1}\}$ based on the query context and motion model, and then converts the $k-1$ edges into $k-1$ locations in the circle of radius with δ_i, where the real location is also included in. The (k-Ψ)-anonymity trajectory algorithm is given as follows.

Algorithm 1: (k-Ψ)-anonymity trajectory algorithm

Input: the real trajectory $trail_{real}$, the degree of anonymity k

Output: the k-anonymity trajectories set $Trails=\{trail_1,trail_2,\cdots,trail_{k-1},trail_{real}\}$

1: **FOR** each $l_i^{real}\in trail_{real}$

2: $ConvertToe(l_i^{real})\rightarrow e_{real}$ // convert the real location at t_i to an edge in // the road network

3: **WHILE** $|Q_i|<k-1$ // $|Q_i|$ is the number of valid edges generated based

// on e_{real}

4: $TreeGenerate(G, e_{real}) \rightarrow Q_i$

5: **END WHILE**

6: **FOR** each $e_j \in Q_i$ //convert the *k*-1 dummy edges to the locations in the road //network and put them into a dummy locations set L_i

7: $ConvertTol(e_j) \rightarrow l_i^j$

8: $\{l_i^j\} \cup L_i \rightarrow L_i$

9: **END FOR**

10: $\{l_i^{real}\} \cup L_i \rightarrow L_i$

11: **IF** $L_i \in CircularArea(\delta_i)$ //whether all the dummy locations generated at time //t_i are in the circular of radius with δ_i

12: $\{L_i\} \cup LL \rightarrow LL$ // $LL = \{L_1, L_2, \cdots, L_n\}$

13: **ELSE**

14: GOTO 3 // jump to step 3 if the conditions are not met

15: **END IF**

16: **END FOR**

17: **FOR** each $L_i \in LL$ //generate k trajectories based on the locations at each time

18: **FOR** each $l_i^j \in L_i$

19: $\{l_i^j\} \cup trail_j \rightarrow trail_j$

20: **END FOR**

21: **END FOR**

22: **RETURN** *Trails*

5 Trajectory perturbation under differential privacy

Traditional location privacy protection strategies are usually closely related to the attacker's background knowledge and can only resist certain pre-set attacks. For example, the above (*k*-Ψ)-anonymity trajectory algorithm can resist the re-identification attack, but cannot prevent sub-trajectory attack. Differential privacy assumes the attacker has a maximized background knowledge and perturbs the input/output data or intermediate result by adding noise to achieve privacy protection. This method can provide a strong privacy guarantee while keeping a relatively small amount of computation overhead, which can be well applied to solve the privacy disclosure problem in the trajectory data publishing.

5.1 Differential privacy

Differential privacy is a data distortion-based privacy protection method that add noise to the original data. Supposing that D_1 and D_2 are a pair of adjacent datasets that are only differ in one record, the concept of ε-differential privacy is given as follows:

Definition 7 (ε-Differential Privacy [Dwork and Roth (2014)]): A random algorithm q is ε-differentially private if for all $S \subseteq Range(q)$:

$$\frac{\Pr[q(D_1) \in S]}{\Pr[q(D_2) \in S]} \leq e^{\varepsilon} \tag{5}$$

where $\Pr[A]$ is the probability that the algorithm output A and S is the subset of all the output of the algorithm q.

By perturbing the output, differential privacy guarantees a randomized algorithm behaves similarly on adjacent datasets. Moreover, for numeric data, we can use the Laplace mechanism which adding noise subject to Laplace distribution to the output.

Definition 8 (Laplace Distribution): If the probability density function of a random variable is:

$$\begin{aligned} f(x \mid \mu, b) &= \frac{1}{2b}\exp\left(-\frac{x-\mu}{b}\right) \\ &= \frac{1}{2b}\begin{cases} \exp\left(-\frac{\mu - x}{\mathrm{b}}\right) & if\ (x < \mu) \\ \exp\left(-\frac{x-\mu}{b}\right) & if\ (x \geq \mu) \end{cases} \end{aligned} \tag{6}$$

then the random variable subject to the Laplace distribution, where μ is the location parameter, and $b > 0$ is the scale parameter.

Definition 9 (Sensitivity): For the random algorithm q, the sensitivity is defined as the difference between the adjacent datasets D_1 and D_2:

$$\Delta q = \max_{D_1, D_2} \| q(D_1) - q(D_2) \|_1 \tag{7}$$

where $\| q(D_1) - q(D_2) \|_1$ is the l_1 norm distance between $q(D_1)$ and $q(D_2)$.

Definition 10 (Laplace Mechanism [Dwork and Roth (2014)]): For the random algorithm q, the Laplace mechanism is defined as:

$$M(x, q, \varepsilon) = q(x) + (Y_1, Y_2, \cdots, Y_k) \tag{8}$$

where x is a k-dimensional input and Y_i is the Laplace noise. The Laplace mechanism perturbs the output with Laplace noise, and preserves ε-differential privacy. For each noise, they are random variables drawn from the Laplace distribution centered at 0 and the scale is calibrated to the sensitivity of q (divided by ε), that is, $Y_i \sim Lap(\frac{\Delta q}{\varepsilon})$.

5.2 Perturbation by adding laplace noise

The algorithm 1 output a (k-Ψ)-anonymity trajectory set *Trails* , and to achieve differentially private perturbation, we first extract the locations set $\{(x_{1i}, y_{1i}, t_i), (x_{2i}, y_{2i}, t_i), \cdots, (x_{ki}, y_{ki}, t_i)\}$ drawn from the all k trajectories at time t_i, where $1 \le i \le n$. For convenience, we use a simple version $L = \{l_1, l_2, \cdots, l_k\}$ to denote the locations set at time t_i. For two locations l_i and l_j in L, they may get a same perturbed result $l_p = (x_p, y_p)$ by adding random noise drawn from the Laplace distribution. The whole trajectory set is ε-differentially private if for any two locations l_i and l_j in every L, the probability of the algorithm output $l_p = (x_p, y_p)$ satisfy the following formula:

$$P(x_p \mid x_i) \le e^{\varepsilon} P(x_p \mid x_j) \tag{9}$$

$$P(y_p \mid y_i) \le e^{\varepsilon} P(y_p \mid y_j) \tag{10}$$

where $\varepsilon \ge 0$, $i, j \in \{1, 2, \cdots k\}$, $P(x_p \mid x_i)$ and $P(y_p \mid y_i)$ are the probability of outputting the perturbed results along x and y axis. It has been proved that adding noise to each coordinate point of the location is better than adding to the location directly to achieve privacy protection. Therefore, the Laplace noise with the scale b is added to each coordinate point respectively for achieving the perturbed location $l_p = (x_p, y_p)$

$$P(x_p \mid x_i) = \frac{1}{2b} e^{-\frac{|x_i - x_p|}{b}} \tag{11}$$

$$P(y_p \mid y_i) = \frac{1}{2b} e^{-\frac{|y_i - y_p|}{b}} \tag{12}$$

The random noise in each axis is $-b sign(r_{nd}) \ln(1\text{-}2|r_{nd}|)$, where r_{nd} is a uniform random value within range [-1/2,1/2] . For scale b , it is set to $(\max_n x_n - \min_n x_n)/\varepsilon$ when computing x_p , and set to $(\max_n y_n - \min_n y_n)/\varepsilon$ when computing y_p , where $\max_n x_n$ and $\min_n x_n$ are the maximum and minimum value within set $\{x_1, x_2, ..., x_n\}$ respectively. In general, for the locations l_i , l_j and l_p , we can get the following triangle inequality:

$$|l_j - l_p| \le |l_j - l_i| + |l_i - l_p| \tag{13}$$

we get a new inequality by transforming with dividing by b , taking the power exponent with e as the base, and multiplying by $1/2b$ in both sides:

$$\frac{1}{2b} e^{-\frac{|l_i - l_p|}{b}} \le \frac{1}{2b} e^{-\frac{|l_j - l_p|}{b}} e^{\frac{|l_i - l_j|}{b}} \tag{14}$$

According to Formulas (11) and (12), (14) can be turned into:

$$P(l_p \mid l_i) \le P(l_p \mid l_j) e^{\frac{|l_j - l_i|}{b}} \tag{15}$$

Considering the independence of x and y coordinates, we can get:

$$P(x_p \mid x_i) \le e^{\frac{|x_j - x_i|}{b}} P(x_p \mid x_j) \tag{16}$$

$$P(y_p \mid y_i) \le e^{\frac{|y_j - y_i|}{b}} P(y_p \mid y_j) \tag{17}$$

The maximum value of $|x_j - x_i|$ and $|y_j - y_i|$ are $|\max_n x_n - \min_n x_n|$ and $|\max_n y_n - \min_n y_n|$ respectively. Therefore, we further transform (16) and (17) to:

$$P(x_p \mid x_i) \le e^{\frac{|\max_n x_n - \min_n x_n|}{b}} P(x_p \mid x_j) \tag{18}$$

$$P(y_p \mid y_i) \le e^{\frac{|\max_n y_n - \min_n y_n|}{b}} P(y_p \mid y_j) \tag{19}$$

Therefore, when the scale b is calibrated to $|\max_n x_n - \min_n x_n|$ and $|\max_n y_n - \min_n y_n|$ (divided by ε) respectively, we can limit the probability of the random algorithm output the same value to the constant factor e^{ε}.

5.3 Differentially private trajectory perturbation algorithm

In the proposed differentially private trajectory perturbation algorithm, trajectory perturbation is achieved by the location perturbation. For location set $L_i = (l_i^1, l_i^2, \cdots, l_i^{k-1}, l_i^{real})$ at time t_i, we add Laplace noise $-bsign(r_{nd})\ln(1-2|r_{nd}|)$ to each location in L_i. Based on the perturbation, we can limit the probability ratio that the inference attack algorithm output the same location to the constant factor e^{ε}, and prevent the real location information from being inferred by the attacker with strong background knowledge. The differentially private trajectory perturbation algorithm is given as follows.

Algorithm 2: differentially private trajectory perturbation algorithm

Function 1: *GetSourceData*$(Trails, T)$

Input: (k-Ψ)-anonymity trajectory dataset $Trails = \{trail_1, trail_2, \cdots, trail_k\}$,

time set $T = \{t_1, t_2, \cdots, t_n\}$

Output: original locations set $D = \{L_1, L_2, \cdots, L_n\}$

1: **FOR** each $t_i \in T$

2: **FOR** each $trail_j \in Trails$

3: $\{loc_j^i\} \cup L_i \to L_i$ // loc_j^i is the location of j-th trajectory at time t_i

4: **END FOR**

5: $\{L_i\} \cup D \rightarrow D$

6: **END FOR**

7: **RETURN** D

Function 2: $LapDifferential(D, \varepsilon)$

Input: original locations set $D = \{L_1, L_2, \cdots, L_n\}$, privacy budget ε

Output: perturbed location set D'

1: **FOR** each $L_i \in D$

2: **FOR** each $l_j \in L_i$

3: $l_j^{'}(x) = l_j(x) + [-bsign(r_{nd}) \ln(1 - 2|r_{nd}|)]$

4: $l_j^{'}(y) = l_j(y) + [-bsign(r_{nd}) \ln(1 - 2|r_{nd}|)]$

5: $\{l_j^{'}\} \cup L_i^{'} \rightarrow L_i^{'}$

6: **END FOR**

7: $\{L_i^{'}\} \cup D' \rightarrow D'$

8: **END FOR**

9: **RETURN** D'

6 Security analysis

For trajectory data publishing, (k-δ)-anonymity reduces the probability of the attacker identifying the real trajectory to $1/k$, but it is still easy for the attacker with strong background knowledge to infer the real trajectory. The differential privacy based (k-Ψ)-anonymity algorithm proposed in this paper, can effectively resist cooperative attack and inference attack.

(1) Cooperative attack usually requires interactions between groups of users. However, the trajectory dataset we used is composed of multiple independent sub datasets, namely each sub dataset is only related to one user, so the cooperation attack has no effect on other users. Therefore, the algorithm we proposed can resist the cooperative attack. Ideally, the active attacker can capture the location service providers and get information about all relevant users, then the attacker can perform the inference attack.

(2) Inference attack against (k-Ψ)-anonymity trajectory dataset focuses on inferring the location at each time on the trajectory. In other words, inference attack is successful if the attacker can infer the real trajectory by inferring the real location at each time. Ideally, the attacker cannot find any connection between the user and the perturbed location. However, the attacker may has a maximized background knowledge, such as the entire road network information, approximate locations related to the user, and even the general distribution of the adding noise. Based on these auxiliary information, the attacker can

access to user's real location by inference attack. The trajectory data published by the proposed algorithm is $L_i = \{l_i^1, l_i^2, \cdots, l_i^{k-1}, l_i^{real}\}$, where $1 \le i \le n$, and we assume that the attacker can access all k trajectory data and has the knowledge of differential privacy perturbation and the distribution of noise $P(l_i)$. As a result, the attacker can exclude some locations where the perturbation is highly unlikely to be generated according to these knowledge, and may increase the probability of successful inference attack. But according to the proposed algorithm, the size of Laplace noise added to the location coordinates (x, y) is determined by the maximum distance of any two points in each coordinate component and the privacy budget ε. When the scale parameter b is set to the maximum distance, any two locations at the same time point in the anonymity dataset are reciprocal and satisfy differential privacy. For a perturbed location $l_p = (x_p, y_p)$, any location at the same time point in the anonymity dataset may generate this perturbed location by noise injection with nearly same probability (confined by e^{ε}). Thus the attacker knowing these knowledges still cannot increase the probability of successful inference attack.

7 Experimental evaluation

In this section, we evaluate the effectiveness of the differential privacy-based (k-Ψ)-anonymity algorithm for trajectory data publishing. The hardware environments of simulation experiment are Intel i5-6600 3.30 GHz CPU with 16 GB memory and 1T hard disk, and the software environments are Windows 10 (64bit), Microsoft Visual Studio 2013 and MATLAB 2016. C++ and MATLAB are adopted as the programming languages.

The experiment is conducted into two groups, the first group is to examine the indistinguishability of the trajectory data published by the (k-Ψ)-anonymity algorithm, and the second group is to examine the effectiveness of differentially private trajectory perturbation algorithm, the dataset used in the experiment is the road network of Oldenburg (OL) city, Germany. This OL dataset is a real-life dataset with 6,105 nodes and 7,035 edges, and the road network diagram simulated by MATLAB is presented in Fig. 4.

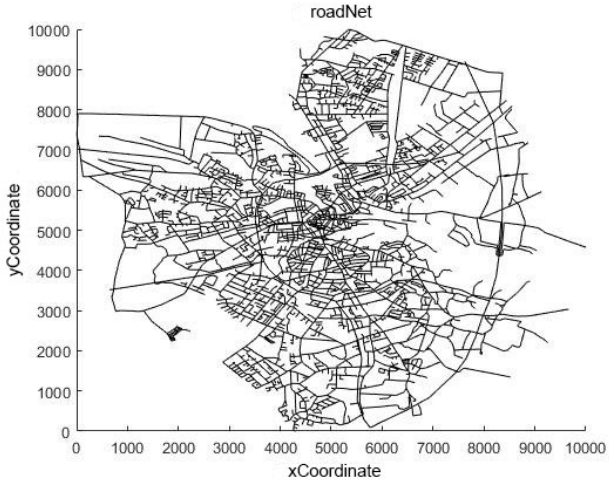


Figure 4: The road network of oldenburg city

7.1 Analysis on trajectory data indistinguishability

In order to examine (k-Ψ)-anonymity algorithm, we firstly generate a real trajectory with n time-location points in the Oldenburg Road Network, and set $n = 13$. As shown in Fig. 5, $Trail8$ is the real trajectory of the user. Then, we generate seven dummy trajectories $\{trail1, trail2, \cdots, trail7\}$ based on the (k-Ψ)-anonymity algorithm. Any two trajectories in this anonymity set are approximate trajectories with each other.

The DM (discernibility metric) is an important indicator commonly used to test the quality of anonymity trajectory dataset, which measures the indistinguishability between the trajectories in the anonymity trajectory dataset. Given a real trajectory $trail_{real}$, we can get a (k-Ψ)-anonymity trajectory dataset $Trails = \{trail_1, \cdots, trail_i, \cdots, trail_{k-1}, trail_{real}\}$, where $trail_i = \{loc_i^1, \cdots, loc_i^j, \cdots, loc_i^n\}$, the DM can be defined as:

$$DM(Trails) = \frac{1}{k \times (k-1)} \times \sum_{m=1}^{k} \sum_{\substack{i=1 \\ i \neq m}}^{k} \frac{1}{Max(Dist(loc_i^j, loc_m^j))} \tag{20}$$

where $Max(Dist(loc_i^j, loc_m^j))$ is the maximum distance between the any two trajectory i and trajectory $m\,(i \neq m)$ at time j, $1 \leq j \leq n$. The DM value reflects the indistinguishable degree of the trajectory in the anonymity trajectory dataset, the larger of the DM value means the better indistinguishability.

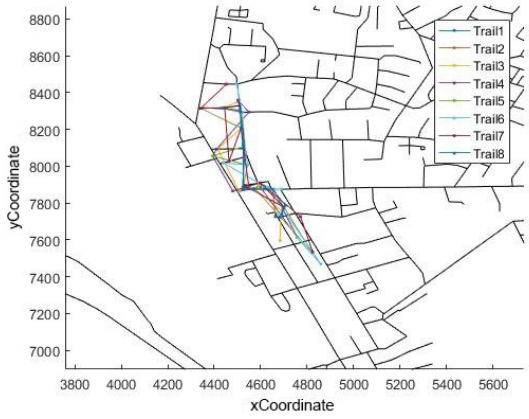


Figure 5: (k-Ψ)-anonymity trajectories in the road network

To verify the validity of the $k-1$ dummy locations of the proposed method, we compare our dummy locations regarding road network with those dummy locations generated ignoring road network information, to analyze the discernibility metric in two cases. If the location coordinates of any of loc_i^j and loc_m^j are not in the vicinity of road segment, the diameter value $2\delta^j$ is taken as $Max(Dist(loc_i^j, loc_m^j))$. As shown in Fig. 6, DM values of the two anonymity trajectory datasets (whether regarding road network or not) increase with the increasing of k value, this means that the risk of privacy leakage decreases as k increases. Another found is that the risk of privacy leakage of the dataset regarding

road network is smaller than the dataset ignoring road network. As k increases, computation cost also increases and the availability of the anonymity dataset decrease. Therefore, we need to balance between the quality of the dataset and the risk of privacy leakage.

Figure 6: DM values of two anonymity trajectory datasets

7.2 Analysis of differentially private trajectory perturbation algorithm

To resist attackers with powerful background knowledge effectively, the differentially private trajectory perturbation algorithm is integrated with the (k-Ψ)-anonymity algorithm to generating anonymity trajectory dataset. In preparatory step, we firstly set $k=8$, $n=13$, and generate trajectories in the road network. Fig. 7 is the 3-dimentional map of the anonymity trajectory dataset $Trails$, where $trail_8$ is the real trajectory, and $trail_1,\cdots,trail_7$ are dummy approximate trajectories, the blue square indicates the plane in which the location of the k trajectories at time $t=1$.

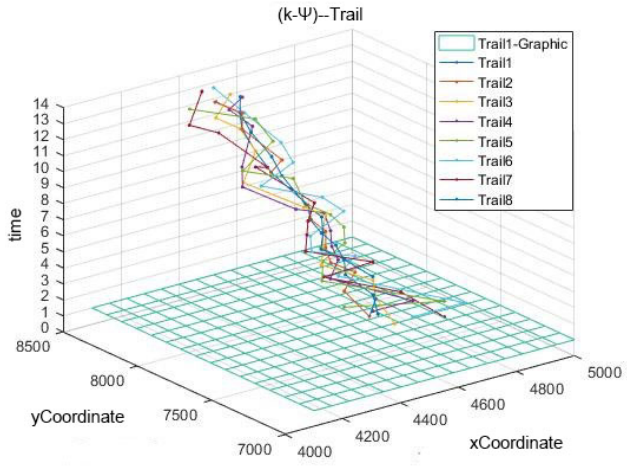
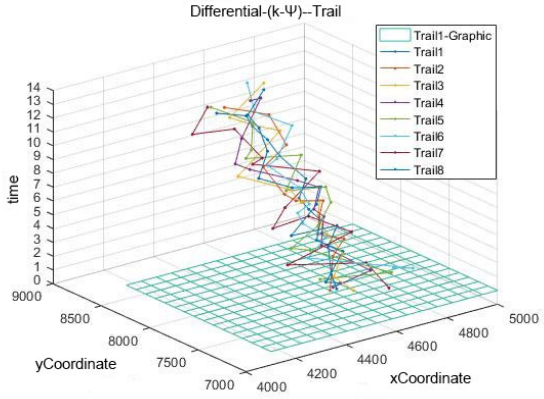
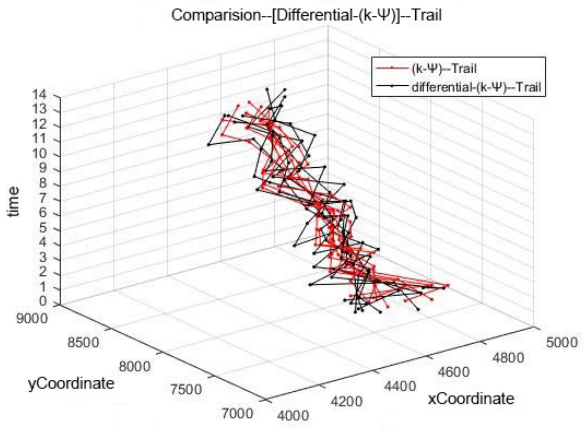


Figure 7: 3-Dimentional map of the anonymity trajectory dataset

In next step, we get a location dataset $L=\{L_1,\cdots,L_j,\cdots,L_{13}\}$ from *Trails* , where $L_j=\{l_{1j},l_{2j},\cdots,l_{8j}\}$, and add Laplace noise to each coordinate of the location based on the proposed algorithm. In the differential perturbation, random noise $-bsign(r_{nd})\ln(1-2|r_{nd}|)$ is independently calculated and added to the X coordinate and the Y coordinate of each location in L_j , and the scale b is calibrated to $(\max_j(x)-\min_j(x))/\varepsilon$, where we set $\varepsilon=1$. As shown in Fig. 8, (a) is the 3-dimentional map of 8 anonymity trajectories with Laplace noise perturbation, (b) is the 3-dimentional map of *k* trajectories with and without perturbation, where red polylines are trajectories without perturbation, black polylines are trajectories with differential perturbation. It can be seen from figure (b) that the perturbed trajectories are similar in overall shape with the undisturbed ones, thereby the proposed algorithm can protect the trajectory privacy of the user without reducing the quality of the published trajectory data a lot.



(a)



(b)

Figure 8: 3-Dimentional map of perturbation (*k*-Ψ)-anonymity trajectory

The indicator closeness degree is used to evaluate the effectiveness of the differential privacy based (*k*-Ψ)-anonymity algorithm, where the closeness degree is the ratio of the perturbed locations that are near to the real ones. We use *dist* to denote the distance between the perturbed location and the real location, and compute the average perturbation percentage of *dist* is less than 100/500/1000 meters when privacy budget ε is taken different value, the result is shown in Fig. 9.

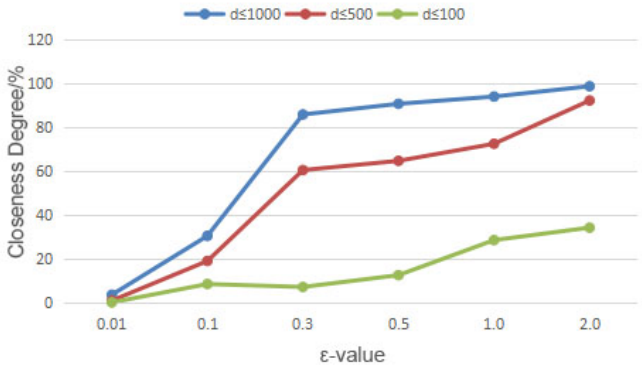


Figure 9: The closeness degrees of differential perturbation

It can be seen that with increase in privacy budget ε, the added Laplace noise decreases, so the *dist* between the perturbed locations and the real ones decreases. In general, the closeness degree increases as ε increases.

As shown in Fig. 9, and in the case of the same ε, the smaller the specified range is, the smaller the closeness degree is. For example, when ε is 0.5, more than 90% of the perturbed locations are within 1000 meters of the real ones, and more than 60% are within 500 meters. Although we hope the proposed algorithm can output a similar result on the dummy locations and the real ones to achieve privacy protection, we want to retain the data utility of the published dataset. Privacy budget $\varepsilon \geq 0.5$ is a better choice.

8 Conclusions

To achieve privacy preserving trajectory data publishing for real road network, we discuss the drawbacks of single invariant δ value and ignoring road network information in traditional (*k*-δ)-anonymity technologies. To get a better balance between data utility and privacy, we propose a differential privacy based (*k*-Ψ)-anonymity method for trajectory data publishing in the paper. The proposed method firstly generate $k-1$ dummy trajectories based on the real trajectory and road network using an adaptive threshold set Ψ, then the outputted anonymous trajectory dataset is perturbed by Laplace noise regarding distance of anonymous locations. The results of experiment with real road network dataset show that the proposed method improves the trajectory indistinguishability and achieves good data utility while satisfying the requirements of user privacy. In future research, how to decrease the amount of Laplace noise and improve the computation efficiency of the proposed method will be our main research direction.

Funding Statement: This work is supported by the Fundamental Research Funds for the Central Universities (No. GK201906009), CERNET Innovation Project (No. NGII20190704), Science and Technology Program of Xi'an City (No. 2019216914GXRC005CG006-GXYD5.2).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Abul, O.; Bonchi, F.; Nanni, M. (2010): Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, vol. 35, no. 8, pp. 884-910.

Al-Hussaeni, K.; Fung, B. C.; Iqbal, F.; Dagher, G. G.; Park, E. G. (2018): SafePath: Differentially-private publishing of passenger trajectories in transportation systems. *Computer Networks*, vol. 143, pp. 126-139.

Dong, Y. L.; Pi, D. C. (2018): Novel privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowledge-Based Systems*, vol. 148, pp. 55-65.

Dwork, C.; Roth, A. (2014): The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211-407.

Feng, Z. N.; Zhu, Y. M. (2016): A survey on trajectory data mining: techniques and applications. *IEEE Access*, vol. 4, pp. 2056-2067.

Gramaglia, M.; Fiore, M.; Tarable, A.; Banchs, A. (2017): $k^{\tau,\epsilon}$-anonymity: Towards privacy-preserving publishing of spatiotemporal trajectory data. arXiv preprint arXiv:1701.02243.

Gruteser, M.; Grunwald, D. (2003): Anonymous usage of location-based services through spatial and temporal cloaking. *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, pp. 31-42.

Gu, K.; Yang, L. H.; Yin, B. (2018): Location data record privacy protection based on differential privacy mechanism. *Information Technology and Control*, vol. 47, no. 4, pp. 639-654.

Gursoy, M. E.; Liu, L.; Truex, S.; Yu, L. (2018): Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing*, vol. 18, no. 10, pp. 2315-2329.

Hua, J. Y.; Gao, Y.; Zhong, S. (2015): Differentially private publication of general time-serial trajectory data. *IEEE Conference on Computer Communications*, pp. 549-557.

Huo, Z.; Meng, X. F.; Zhang, R. (2013): Feel free to check-in: Privacy alert against hidden location inference attacks in geoSNs. *International Conference on Database Systems for Advanced Applications*, vol. 7825, pp. 377-391.

Li, M.; Zhu, L. H.; Zhang, Z. J.; Xu, R. X. (2017): Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences*, vol. 400, pp. 1-13.

Li, Y. L.; Li, S. Y. (2018): A real-time location privacy protection method based on space transformation. *14th International Conference on Computational Intelligence and Security*, pp. 291-295.

Liu, B. Z.; Chen, L.; Zhu, X. Q.; Zhang, Y.; Zhang, C. Q. et al. (2017): Protecting location privacy in spatial crowdsourcing using encrypted data. *Advances in Database Technology-EDBT*, pp. 478-481.

Liu, L.; Liu, W.; Zheng, Y.; Ma, H. D.; Zhang, C. (2018): Third-eye: A mobile phone-enabled crowdsensing system for air quality monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1-26.

Ou, L.; Qin, Z.; Liao, S. S.; Hong, Y.; Jia, X. H. (2018): Releasing correlated trajectories: towards high utility and optimal differential privacy. *IEEE Transactions on Dependable and Secure Computing*, pp. 1-13.

Peng, T.; Liu, Q.; Meng, D. C.; Wang, G. J. (2017): Collaborative trajectory privacy preserving scheme in location-based services. *Information Sciences*, vol. 387, pp. 165-179.

Primault, V.; Boutet, A.; Mokhtar, S. B.; Brunie, L. (2018): The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2772-2793.

Shi, W. B.; Wang, J. Q.; Zhu, J. X.; Wang, Y. P.; Choi, D. (2019): A novel privacy-preserving multi-attribute reverse auction scheme with bidder anonymity using multi-server homomorphic computation. *Intelligent Automation and Soft Computing*, vol. 25, no. 1, pp. 171-181.

Tu, Z.; Zhao, K.; Xu, F. L.; Li, Y.; Su, L. et al. (2017): Beyond k-anonymity: protect your trajectory from semantic attack. *14th Annual IEEE International Conference on Sensing, Communication, and Networking*, pp. 1-9.

Wang, H., Xu, Z. Q. (2017): CTS-DP: Publishing correlated time-series data via differential privacy. *Knowledge-Based Systems*, vol. 122, pp. 167-179.

Wang, T.; Zheng, Z. G.; Rehmani, M. H.; Yao, S. H.; Huo, Z. (2018): Privacy preservation in big data from the communication perspective-a survey. *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753-778.

Xia, Z. Q.; Hu, Z. Z.; Luo, J. P. (2017): UPTP vehicle trajectory prediction based on user preference under complexity environment. *Wireless Personal Communications*, vol. 97, no. 3, pp. 4651-4665.

Xiao, L.; Chen T. H.; Xie, C. X.; Dai, H. Y.; Poor, H. V. (2018): Mobile crowdsensing games in vehicular networks. *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1535-1545.