*Computers, Materials & Continua* CMC, vol.64, no.2, pp.1025-1038, 2020

# Privacy Protection Algorithm for the Internet of Vehicles Based on Local Differential Privacy and Game Model

**Wenxi Han[1, 2], Mingzhi Cheng[3, *], Min Lei[1, 2], Hanwen Xu[2], Yu Yang[1, 2] and Lei Qian[4]**

**Abstract:** In recent years, with the continuous advancement of the intelligent process of the Internet of Vehicles (IoV), the problem of privacy leakage in IoV has become increasingly prominent. The research on the privacy protection of the IoV has become the focus of the society. This paper analyzes the advantages and disadvantages of the existing location privacy protection system structure and algorithms, proposes a privacy protection system structure based on untrusted data collection server, and designs a vehicle location acquisition algorithm based on a local differential privacy and game model. The algorithm first meshes the road network space. Then, the dynamic game model is introduced into the game user location privacy protection model and the attacker location semantic inference model, thereby minimizing the possibility of exposing the regional semantic privacy of the *k*-location set while maximizing the availability of the service. On this basis, a statistical method is designed, which satisfies the local differential privacy of *k*-location sets and obtains unbiased estimation of traffic density in different regions. Finally, this paper verifies the algorithm based on the data set of mobile vehicles in Shanghai. The experimental results show that the algorithm can guarantee the user's location privacy and location semantic privacy while satisfying the service quality requirements, and provide better privacy protection and service for the users of the IoV.

**Keywords:** The Internet of Vehicles, privacy protection, local differential privacy, location semantic inference attack, game theory.

## 1 Introduction

The Internet of Vehicles is the application of mobile ad hoc networks and the Internet of Things (IoT) in the transportation industry. The specific components of the IoV can be divided into: vehicle nodes, vehicle units, roadside communication units, data collection servers and service providers. The data collection server collects the location service data of the mobile vehicle, including the location, speed, direction and time of the vehicle. The

[1] Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, China.

[2] School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[3] College of New Media, Beijing Institute of Graphic Communication, Beijing, 102600, China.

[4] School of Computer Science, The University of Auckland, Auckland, New Zealand.

[*] Corresponding Author: Mingzhi Cheng. Email: chengmz@bigc.edu.cn.

Received: 20 January 2020; Accepted: 09 March 2020.

service provider of the (IoV) analyzes and mines the data, and draws some conclusions to support the location-based service (LBS), to facilitate the optimization of urban road planning, business decisions. As people pay attention to personal privacy, more and more users are not willing to share their precise location data and most data collection servers are untrustworthy. Users, even more, expects that the car network data has been subjected to privacy protection processing before leaving the vehicle node. This is true even if the data collection server cannot obtain the user's accurate data.

Localized differential privacy (LDP) [Ye, Meng, Zhu et al. (2018)], which has emerged in recent years, is the best way to solve the above problems. The existing localized differential privacy protection technologies are based on a stand-alone system structure, which is a C/S structure composed only of a client (i.e., a mobile device) and a server. The user independently performs a perturbation mechanism according to requirements, directly perturbs the protected data, send to the service provider and obtain the corresponding query result. The system structure is simple to implement as there is no limitation of third-party security bottleneck, but the client only performs the disturbance processing for itself, completely ignoring the real environment information for the release location and the actual location If the deviation is too large, not only the location is easily filtered by the attacker but also reduces the quality of service obtained by the user; if the publishing location is close to the actual location, it is easy to expose the user's location semantics. Therefore, the local differential privacy protection scheme of the IoV needs to be combined with the real environmental information and balance the effectiveness and usability of privacy protection for further research.

This paper studies the application of local differential privacy and game model in the process of data acquisition of the IoV. The main work is summarized as follows:

i. Based on the untrusted environment of data collection, it proposes a location data acquisition method that satisfies the local differential privacy protection algorithm based on optimal *k*-location set with a dynamic game model. By defining the relevant linear programming, in the case of guaranteeing the user's service quality, the location semantic speculation attack in the real environment is most resisted, and the user's protection level of location privacy is optimized.

ii. Proposes a regional traffic density statistical algorithm based on *k*-location set localized differential privacy protection mechanism. Regional traffic density statistical query is performed on randomly disturbed location data, and statistical results can support intelligent transportation system decision.

iii. The method proposed in this paper was verified by experiments. It proves that it has advantages in data availability, algorithm efficiency and scalability.

## 2 Related work

### *2.1 Research status of location privacy protection*

In the current location, privacy protection technologies are mainly divided into four categories, including policy-based methods, encryption-based technologies [Kim, Hong and Chang (2016)], anonymous-based technologies [Li, Lv and Li (2018)], and differential privacy-based technologies [Dwork and Lei (2009)]. The privacy protection policy is the

only method that acts on the service provider to constrain it by developing privacy protection rules, standards, and detailed specifications that rely on the strict enforcement of the service provider. The encryption-based privacy protection method has a high level of privacy protection, but its operation overhead is huge, and a special database needs to be built. The anonymous-based privacy protection method has better data security and usability, and has been currently widely used. However, this method has an obvious drawback, that is, when measuring the level of privacy protection, it is necessary to assume the background knowledge of the attacker. The introduction of differential privacy can well solve the problem based on differential privacy protection technology. It does not limit the attacker's background knowledge. Even if the attacker has mastered all the information except one record, it can still be against the attacker. The records that are mastered are effectively protected and are highly secure. However, the traditional differential privacy location data protection method is mostly based on a trusted third-party data collection server, requiring each user to send their own real data records to the data collection server. The data collection server responds to the query request of the data analyst by using the privacy algorithm that meets the demand, thus causing the problem of server data leakage.

The LDP that has emerged in recent years is a powerful means for data privacy protection on the client side, has been introduced into the field of location privacy protection, and has made certain progress. The literature Gao et al. [Gao, Cui, Du et al. (2019)] uses the pre-arranged information collection points to generate a random response candidate location set, without considering the location reachability of the candidate location set, and improving the location privacy protection level at the expense of service quality, resulting in low location service availability. The literature Chen et al. [Chen, Li, Qin et al. (2016)] proposed a personalized count estimation protocol (PCEP) to establish a random response candidate location set based on user preference constraints, without considering the constraints in the real road network environment. In the PCEP algorithm, the computational cost of the S-Hist perturbation algorithm used is positively correlated with the number of users. When the user is busy for a long time, the computational cost is huge, and the sampling process also brings a certain precision loss, and the availability of the algorithm needs to be improved. The literature Xiao et al. [Xiao and Xiong (2015)] considers the timing impact of the released disturbance location on the upcoming release location, describes the time correlation with Markov chain, and constructs the random response candidate location set according to the location transition prior probability. The computational complexity of the algorithm is relatively high. Moreover, when the user's response location set appears at a higher probability possible location, it is easy to expose the user's interest point, failing to consider the information leakage problem caused by the attacker's semantic speculation attack. The literature Zhen et al. [Zhen, Ping and Yan (2019)] uses the Voronoi Diagram division method to make a Voronoi grid contain at least one road node, and there is no case where an unreachable area is divided into a safe area, such as a river, a lake, etc. However, other real locations of users within the Voronoi boundary are directly selected as the candidate location set, without considering the location accessibility of Voronoi grid under the constraints of real speed and driving direction.

### *2.2 Game theory*

An attacker can use the collected data to infer private information. Shokri [Shokri (2015)] proposed a protection strategy based on Stackelberg game, which assumes that the attacker

has acquired prior knowledge, allowing users and attackers to play in turn. The user maximizes the level of privacy protection while ensuring that the quality of service loss is less than a given threshold, and the attacker seeks to minimize the level of privacy protection based on prior knowledge and offset location. Through the game, the strategy can ultimately ensure that the quality of service loss is less than a given threshold while optimizing the level of privacy protection.

## 3 Privacy protection algorithm for Internet of Vehicles based on local differential privacy and game model

### *3.1 Description of the problem*

**Data availability of LDP.** Based on the LDP location protection algorithm, usually preset $N$ information collection point $\mathrm{C} = \{c_1, c_2, \ldots, c_N\}$, $n$ collection points cover the entire urban area R. The user location ($1 \le i \le N$) is marked by the strongest communication signal i between the vehicle and the collection point, and the generalized vehicle coordinates are the labels of the information collection points. The n-bit array A encodes the current location of the vehicle as shown in Eq. (1):

$$\mathrm{A}_k = \begin{cases} 1, & k = i; \\ 0, & otherwise \end{cases} \tag{1}$$

$\mathrm{A}_k$ represents the $k$th bit of the location array A. The strongest signal corresponds to the location code 1, and the other locations are 0.

Although LDP has strict and provable privacy protection features, its disadvantage is that it is less usable. Usually, the number of information collection points $n$ is as much as possible, coverage R is as wide as possible. If the vehicle's disturbance location is far from the real location, the vehicle will use the disturbance location instead of the real location to send to the service provider, which will greatly reduce the quality of the location service. At the same time, a large number of information collection points increase the length of the location code, resulting in a high transmission cost between the vehicle and the information collection point. For the vehicle user, it is of little significance to shift the real location to a farther location, and the location disturbance satisfying the service availability is more in line with the actual requirements.

**Location Semantic Inference Attack.** In the process of location privacy protection based on local differential privacy mechanism, the road network semantic information of the real environment, and information for the location privacy protection mechanism $\mathrm{P}(\mathrm{x}_t'|\mathrm{x}_t)$, $k$-location set $\mathrm{K}_{set}$ and the release location $x'$ is public. Therefore, service providers and malicious attackers can formulate corresponding countermeasures based on the privacy protection methods adopted by vehicle users, and combine precise background knowledge for location semantic inference attacks.

Although the nearest $k$-location set is the location with a high probability of users' occurrence, for attackers who master the semantic location information of the road network, the semantic information contained in the nearest $k$-location set still exposed the behavior and activity privacy of vehicle users through location semantic inference attacks [Ma, Du, Li et al. (2016)]. As shown in Fig. 1, the square areas constitute the $k$-location set, which contains four points of interest and three types of locations. Although the

semantic diversity is satisfied, if the vehicle requests LBS at 12 o'clock in the evening, considering that the school and bank are closed at this time, it's easy to infer that the user is currently in the hospital. In the case where the topic selection rule is unchanged, the influence of time on the region implies semantics. That is, the implicit semantics of the same $k$-location set will change with time. The attacker can dynamically generate the geographically implicit semantic information based on the temporal topic model and guess the real location semantic information of the vehicle user.

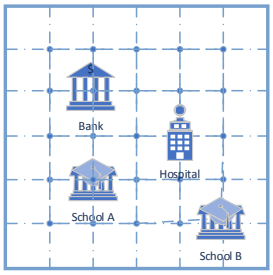



**Figure 1:** Semantic inference attack

In summary, although nearest $k$-location set maximizes the data availability of the perturbed location. If the attacker and the user have the same semantic data background knowledge, the LDP also exposes the location semantic information of the vehicle. Therefore, achieving a reasonable balance between privacy and availability under the LDP condition is a big challenge.

### *3.2 Local differential privacy protection algorithm based on optimal k-location set with dynamic game model*

In this paper, a vehicle location protection algorithm optimal *k*-LDPM (optimal *k*-location Set based Local Differential Privacy Mechanism) is proposed. A dynamic game model is established to optimize the selection of $k$-location set. In this game, the vehicle user gives the $k$-location sets $\mathrm{K}_{set}$ first, which called $fuzzy(K_{set}^{t}|l_t)$, while the service provider optimizes the location semantic attack strategy according to the real environment information and the $k$-location sets $\mathrm{K}_{set}$, which called $guess(\widehat{s_t}|K_{set}^{t})$. Therefore, this paper introduces a dynamic game model, which takes the vehicle user as the leader ($\mathrm{N}_1$), the service provider as the follower ($\mathrm{N}_2$). Through the game between the protection model of user location privacy and the location conjecture model of attacker, the privacy protection process of the IoV is optimized, and the balance between the quality of service and the risk of privacy disclosure is made to further improve the level of location privacy protection. Based on the optimal *k*-location set, randomly respond to a release location $\mathrm{o}_t$ that meets LDP and expose it to the service provider to protect the location privacy of the vehicle. The specific process of optimal *k*-LDPM is shown below.

Dividing road network space, vehicles move in a space area Ω, define the grid size for grid partition granularity omega, meshing Ω space area, and generating a set of road network space locations $\mathrm{P} = \{p_1, p_2, \ldots, p_{|\mathrm{P}|}\}$. Each grid intersection corresponds to a location, and the total number of divisions is |P|. After the road network space division is completed, the server transmits the division results to the vehicle terminals.

Generate offset location $x_t$. After the vehicle receive the road network space division result, the vehicle calculates nearest offset location $x_t$ , which replaces vehicle real location $l_t$ to $k$-location set. The distance between the two spatial locations is measured by Haversine distance, as shown in Eq. (2).

$$dis(p_i, p_j) = \mathrm{R} \cdot \arccos\left[\cos\beta_i \cos\beta_j \cos(\alpha_i - \alpha_j) + \sin\beta_i \sin\beta_j\right] \quad (2)$$

where r is the radius of the earth, β is the longitude angle, α is the latitude angle.

Construct a set of optimal $k$-location. The game between the vehicle user and the service provider is a strict zero-sum game. The Nash equilibrium is to calculate its own maximum revenue on the basis of considering the best choice of the other party. The probability of Exposing Semantic (PES, Probability of Exposing Semantic) is used to measure the revenue of participants in the game, PES's calculation is shown in Eq. (3),

$$\mathrm{PES} = \frac{\widehat{ms}_{valid}^{K_{set}^t}}{ms_{valid}^t}, (0 \le PES \le 1) \quad (3)$$

Among them, $ms_{valid}^t t$ represents the effective semantic number of t time, and $\widehat{ms}_{valid}^{K_{set}^t}$ represents the effective semantic number of t time-$k$-location set $K_{set}^t$ region inferred by the attacker. The smaller PES is, the less the amount of the location semantics is filtered by the attacker, the more the vehicle user's revenue, and the better the privacy protection level of the privacy protection algorithm.

First of all, considering that the $k$-location sets of given $\mathrm{N}_1$ at the first stage of the selection time $t$ is $K_{set}^t$, the game needs to hide the true location semantics of the user on the premise of ensuring the quality of service of the user. In order to ensure the quality of service, the maximum loss of the service quality acceptable to the user is $\mathrm{Q}_{loss}^{max}$, as shown in Eq. (4):

$$dis(p_t, l_t) \le \mathrm{Q}_{loss}^{max}, \ \forall x_t' \in K_{set}^t \quad (4)$$

where $p_t$ is any single location point of $k$-location sets under the moment $t$; $l_t$ is the true location of vehicle user under the moment $t$.

Currently, the goal of $\mathrm{N}_2$ is to minimize the location semantics protection ability. According to the background knowledge, a semantic speculation model [Sarda, Eickhoff and Hofmann (2016)] of regional location based on the time theme was built by $\mathrm{N}_2$.

*Modeling.* The key premise of semantic attack lies in that the attacker owns the same map data with that of the vehicle user. Therefore, divide the road network space as mentioned above and generate the spatial location sets $\mathrm{P} = \{p_1, p_2, \ldots, p_{|\mathrm{P}|}\}$, extract the urban road network semantics and generate the road network interest point sets $\mathrm{M} = \{m_1, m_2, \ldots, m_{|\mathrm{M}|}\}$ and interest semantic sets $\mathrm{MS} = \{ms_1, ms_2, \ldots, ms_{|\mathrm{MS}|}\} ms_i = (class\_ms_i, num\_ms_i)$.

*Semantic annotation of location point.* Describe the actual semantics of location point and use $\mathrm{ps}_{\mathrm{t}_i} = (\mathrm{ps}_i, \mathrm{t})$ represents the actual semantics of the single point location $p_i$, of which, among them $\mathrm{ps}_i = \left\{ps_i^1, ps_i^2, \ldots, ps_i^{|\mathrm{MS}|}\right\}$ is coding with "01", to represent the semantic possibility of the single point location $p_i$; $\mathrm{t}$ represents the time that the semantic possibility of the single point location $p_i$ is $\mathrm{ps}_i$. Such as, M={A hospital, B primary school, C Rehabilitation, D Bank}, MS={(healthcare, 2), (finance, 1), (education, 1)}, $\mathrm{s\_t_i}$=((1, 0, 1),

6) represents the location point that the vehicle user appears at $p_i$, from 6:00 a.m. to 7:00 a.m. and the user activity may mean attending school or going to hospital. The process of semantic annotation of location point is as follows.

*Distance measurement.* Calculate the Haversine distance between each location point in the spatial location sets P of road network and $|M|$ interest point in the road network interest point set $M$.

*Valid POI screening.* Set the max valid semantic radius to be r, if $dis(p,s) >$ r, then the interest point is unreachable, then exclude it. And the number of valid POI screened is $m_{valid}(m_{valid} \leq |M|)$ at last. Valid interest semantics screening: Based on the time theme, use 24 |MS| -bit "01" code to represent whether each interest semantics of each hour every day can be accessed $V = \{v_0, v_1, \ldots, v_{23}\}$, $v_t = \left\{v_t^{ms_1}, v_t^{ms_2}, \ldots, v_t^{ms_{|MS|}}\right\}$ if $v_t^{ms_i}$ is 0, then the access possibility of the interest semantics $ms_i$ is 0 in t time period and the interest semantics can be excluded. And the number of valid interest semantics screened is $ms_{valid}^t\left(ms_{valid}^t \leq |MS|\right)$ at last. Where the access possibility of the semantics based on the time theme is supposed to be known.

*Calculation of semantic possibility.* The number of the valid interest point implying the interest semantics $ms_j$ at the location point $p_i$ is ${ms_i}_{valid}^t$ and the way of implying the interest semantics $ms_j$ at the location point $p_i$ for $|MS| ps_i^j$ bit array coding vehicle user is as shown in Eq. (5):

$$ps_i^j \begin{cases} 1, & {ms_i}_{valid}^t > 0 \\ 0, & {ms_i}_{valid}^t = 0 \end{cases} \tag{5}$$

By parity of reasoning, the semantic annotations of location point $p_i$ are obtained, as shown in Eq. (6):

$$ps_i = \left\{ps_i^1, ps_i^2, \ldots, ps_i^{|MS|}\right\} \tag{6}$$

*Regional semantic speculation.* The regional semantic distribution of $k$-location sets $K_{set}^t$ is speculated according to semantic annotation of location point within the region, as shown in Eq. (7):

$$\hat{s}_{set}^t = guess(\hat{s}_{set}^t | K_{set}^t) = \frac{\sum_{K_{set}^t} ps_i}{k} \tag{7}$$

To sum up, $N_2$ finds out the semantic location sets $\hat{s}_{set}^t$ which minimizes semantic privacy degree. The privacy attack earnings of $N_2$ are shown in Eq. (8):

$$U_2{}^{MAX} = max\left[U_2\left(fuzzy\left(K_{set}^t | l_t\right), guess\left(\hat{s}_{set}^t | K_{set}^t\right)\right)\right] = max\left[\frac{\widehat{ms}_{valid}^{K_{set}^t}}{ms_{valid}^t}\right] \tag{8}$$

Because the optimal attack strategy of $N_2$ is $guess(\hat{s}_{set}^t | K_{set}^t)$ if $N_1$ selected $k$-location set $K_{set}^t$. Then, at the first stage of the game, the goal of $N_1$ is to maximize the ability of privacy protection under the condition of ensuring the availability of private data, thus maximizing his own earnings. The privacy protection earnings of $N_1$ are as shown in Eq. (9):

$$U_1{}^{MAX} = max\left[U_1\left(fuzzy\left(K_{set}^t | l_t\right), guess\left(\hat{s}_{set}^t | K_{set}^t\right)\right)\right] = min\left[\frac{\widehat{ms}_{valid}^{K_{set}^t}}{ms_{valid}^t}\right] \tag{9}$$

To sum up, the $k$ optimal location set $K_{set}^t$ may use the linear programming to solve it. The final definition of linear programming is as shown in Eq. (10):

$$\text{Maximize } argmax\left[\mathrm{U}_1\left(fuzzy^*(K_{set}^t|l_t), guess^*(\hat{s}_{set}^t|K_{set}^t)\right)\right] \tag{10}$$

s.t.

$$\mathrm{U}_2\left(fuzzy^*(K_{set}^t|l_t), guess^*(\hat{s}_{set}^t|K_{set}^t)\right) \geq \mathrm{U}_2\left(fuzzy(K_{set}^t|l_t), guess(\hat{s}_{set}^t|K_{set}^t)\right) \quad \text{(a)}$$

$$dis(x_t{}', l_t) \leq \mathrm{Q}_{loss}^{max}, \forall x_t{}' \in K_{set}^t \quad \text{(b)}$$

$$\exists K_{set}^t : fuzzy(K_{set}^t|l_t), \forall l_t \in \Omega \quad \text{(c)}$$

Solve the objective function under the constraint condition of Eq. (11), in the strategy maximizing the earnings expectation, find the $k$-optimal location set $K_{set}^t$ and maximizes the earnings of $\mathrm{N}_2$ on the premise of meeting the service quality, thus realizing the objective of maximizing privacy protection of location semantic for the vehicle user. Where, condition (a) maximizes the earnings of the attacker; condition (b) restrains the quality loss of service, condition (c) represents for arbitrary $l_t$ in Ω, there exists at least one $K_{set}^t$ which supports $fuzzy(K_{set}^t|l_t)$.

Generate release location $x_t{}'$. Combined with the *k*-RR [Warner (1965)] random response mechanism, the vehicle offset location $\mathrm{x}_t$ is perturbed to generate the release location $\mathrm{o}_t$ that satisfies the local differential privacy, making it meets the indistinguishability of *k*-locations.

Given local differential privacy parameters based on random response mechanisms $\epsilon$. Set the privacy of mobile vehicles within the same space area Ω budget are the same. At time t, the way each vehicle releases the location is as shown in Eq. (11):

$$\mathrm{P}(\mathrm{o}_t|x_t) = \frac{1}{k-1+e^{\epsilon}} \begin{cases} e^{\epsilon}, & if\ \mathrm{o}_t = x_t \\ 1, & if\ \mathrm{o}_t \neq x_t \end{cases} \tag{11}$$

That is, using a probability of $\frac{e^{\epsilon}}{k-1+e^{\epsilon}}$ sending its offset location $x_t$, using a probability of $\frac{1}{k-1+e^{\epsilon}}$ responding to any of the rest of the $k-1$ locations and making them meet the $\epsilon$-local differential privacy. $\epsilon$ used to balance degree of privacy and data availability. The smaller $\epsilon$, the higher the degree of privacy and the lower the statistical data availability.

### *3.3 The regional traffic density statistical algorithm based on meeting the localized difference privacy of* k *location set*

The vehicle user responses randomly to the vehicle location within the $k$-location set through the random response mechanism to disturb the vehicle release location within the location set coverage area $K$ in order to protect the location privacy of the user. Therefore, different vehicles $k$ has different location set coverage area. The disturbance statistics and correction of the regional traffic density need to be analyzed according to the specific scenario.

Suppose that the total number of vehicle involving in the random location response is $n$ and the vehicle user sends its deviation location $x_t$ with the probability of $p_1$ and responses to any location of the rest $k-1$ locations with the probability of $p_2$. The regional traffic density statistical algorithm involved in this paper is realized in the following four scenarios:

a) The density statistics area R completely covers $k$-location set composition area $\mathrm{K}_n$ of n vehicles, i.e., $\mathrm{K}_n \cap \mathrm{R} = \mathrm{K}_n$. There are n moving vehicles in the $k$-set composition

area $\mathrm{K}_n$ of n vehicles, then the regional traffic density is $|TD| = \mathrm{n}$.

b) The density statistics area intersects with the location set area of each vehicle of/vehicles, if and only if: $\mathrm{Rn}k\mathrm{K}_n^i\mathrm{K}_n^i \cap \mathrm{R} \neq \emptyset\mathrm{K}_n^i \cup \mathrm{R} \neq \mathrm{RK}_n^i \cup \mathrm{R} \neq \mathrm{K}_n^i$.

Suppose that in the statistical results, the number of vehicles of which response location is in the density statistics area R is $n'$, then the number not in that area is $n - n'$. The disturbance statistics of the proportion of vehicle user of which response location is/isn't in the density statistics area is as shown in Eq. (12):

$$\begin{cases} \Pr(X_i = \text{"in"}) = \frac{n'}{n} = \pi\big(p_1 + p_2(\tilde{r} - 1)\big) + (1 - \pi)p_2 \cdot \tilde{r} \\ \Pr(X_i = \text{"}not\ in\text{"}) = \frac{n - n'}{n} = \pi p_2 \cdot (k - \tilde{r}) + (1 - \pi)\big(p_1 + p_2(k - \tilde{r} - 1)\big) \end{cases} \tag{12}$$

where, $\pi$ is the proportion of vehicle user of which true location is in the density statistics area R and $\tilde{\mathrm{r}}$ is the average of the n number of locations in which the vehicle intersects with the density statistics area, as shown in Eq. (13):

$$\tilde{\mathrm{r}} = \frac{\sum_{i=0}^{n} K_n^i \cap R}{n} \tag{13}$$

According to the maximum likelihood estimation, build the likelihood function as shown in Eq. (14):

$$L(\pi) = \big[\pi\big(p_1 + p_2(\tilde{r} - 1)\big) + (1 - \pi)p_2 \cdot \tilde{r}\big]^{n'}\big[\pi p_2 \cdot (k - \tilde{r}) + (1 - \pi)\big(p_1 + p_2(k - \tilde{r} - 1)\big)\big]^{n - n'} \tag{14}$$

Then, the corrected statistical value of regional traffic density $R$ is computed with maximum likelihood estimation of $\pi$, as shown in Eq. (15):

$$|TD| = \hat{\pi} \times n = \frac{-n(p_1 - p_2)}{np_2\tilde{r} - n'[p_2(k-1) + p_1]} \times n \tag{15}$$

c) The density statistics area only intersects with the $k$- location set area $\mathrm{K}_{n_1}^i$ R of each vehicle of $n_1$ vehicles, that is $\mathrm{K}_{n_1}^i \cup \mathrm{R} \neq \mathrm{R}$ and $\mathrm{K}_{n_1}^i \cup \mathrm{R} \neq \mathrm{K}_{n_1}^i$, $n_1 < n$.

Suppose that the total number of vehicles is $n_1$ in this scenario, then the calculation method of $R$ traffic density is the same to that of (b).

d) The density statistics area R completely covers the $k$-location set colocation area $\mathrm{K}_{n^1}$ of $n_1$ vehicles and intersects with the $k$-location set area $\mathrm{K}_{n_2}^i$ of each vehicle of $n_2$ vehicles, that is $\mathrm{K}_{n^1} \cap \mathrm{R} = \mathrm{K}_{n^1}$ and $\mathrm{K}_{n_2}^i \cap \mathrm{R} \neq \emptyset$ and $\mathrm{K}_{n_2}^i \cup \mathrm{R} \neq \mathrm{R}$ and $\mathrm{K}_{n_2}^i \cup \mathrm{R} \neq \mathrm{K}_{n_2}^i$, $n_1 + n_2 \leq n$.

The calculation method of regional traffic density in this scenario can be divided into two parts, the first is the traffic density of statistical area R $\mathrm{K}_{n^1}$ and the calculation method is the same to that of (a); the second is that the calculation method of traffic density of statistical area $\mathrm{K}_{n_2}^i \cap \mathrm{R}$ is the same to that of (c).

## 4 Experimental analysis

This paper uses true data set to release the vehicle disturbance location and conduct the simulation experiment of regional traffic density statistics. The vehicle track data set comes from Smart City Research Group [Liu, Liu, Lionel et al. (2010)], including the driving records for 24 h of 4,000 taxis on Feb. 20, 2007 in Shanghai. The sample interval of vehicle driving

data is 1 min. and vehicle track data format is: vehicle ID-time-longitude and latitude-speed. The map POI data set comes from Open Street Map [Haklay and Weber (2008)], including large amount of latitude and longitude information of interest points in Shanghai. The interest point format is: Interest point ID-name of the interest point-longitude and latitude.

### *4.1 Data pre-processing*

**Division of the urban network.** Conduct the meshing of the map of Shanghai and take the granularity of meshing $\omega = 1000$ m as the example, the number of mesh intersection is 7527.

**Semantic annotation of the urban road network POI.** 7 semantic categories of road network interest semantic set attributes MS $= \{ms_1, ms_2, \ldots, ms_7\}$ are as shown in Tab. 1.

**Table 1:** Semantic annotation of Shanghai Road Network POI

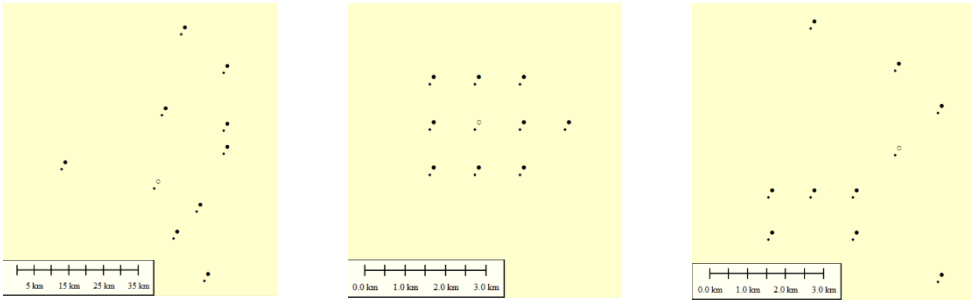
| No. | Semantic | Category Number of POI | Total Number of POI |
|---|---|---|---|
| **1** | Education | 170 | |
| **2** | Finance | 624 | |
| **3** | Healthcare | 204 | |
| **4** | Food & Beverage | 2590 | 4704 |
| **5** | Market | 225 | |
| **6** | Residence | 480 | |
| **7** | Leisure Spots | 411 | |

Define that probability vectors for accessing the interest semantics M $= \{m_1, m_2, \ldots, m_{4704}\}$ of Shanghai road network at 12:00 p.m. and 24:00 p.m. are $v_{12} = \{1,1,1,1,1,1,1\}$ and $v_{12} = \{0,1,1,1,0,1,0\}$, respectively and define that the interest point within 1 km away from Haversine is the valid point of distance M $= \{m_1, m_2, \ldots, m_{4704}\}$ as well as complete the semantic annotation MS $= \{ms_1, ms_2, \ldots, ms_6\}$ of POI of spatial location P $= \{p_1, p_2, \ldots, p_{|P|}\}$ of Shanghai road network.

**Vehicle Location Deviation.** Calculate the distance between the vehicle location and proximity space location point of 12:05 a.m. and 24:05 p.m., select the most proximity space location point to generate the vehicle deviation location to replace the actual location of vehicles. At 12:05, the total number of vehicle deviation is 1470. At 24:05, the total number of vehicle deviation is 1004.

### *4.2 Utility analysis of privacy protection mechanism*

To verify the effectiveness of algorithm, this paper would compare LDPM, proximity $k-$ LDPM and optimization $k-$ LDPM. Take $\omega = 1000$ m, $k = 10$ as an example, for the vehicle with ID of 86349 at 12:05, the location distribution of its $k$-location set is shown in Fig. 2.



(a) Random response location set (LDPM)

(b) Location set $k$ ( nearest $k-$ LDPM)

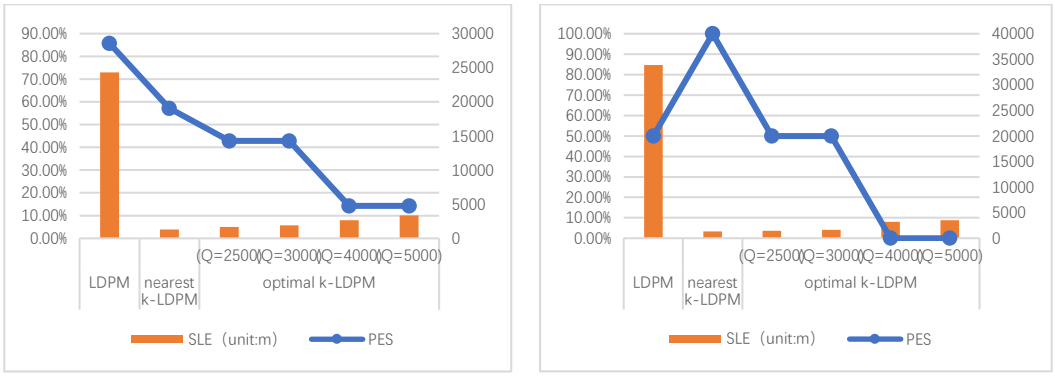
(c) Location set $k$ (optimal $k-$ LDPM)

**Figure 2:** Random response location set generated under different algorithms

In Fig. 2, ○ labels the actual location of vehicle users, ● labels the false location of $k$-location concentration. Fig. 2 (a) is the distribution situation of the LDPM selected random response location set, Fig. 2 (b) is the location set distribution $k-$ LDPM selected by the proximity and Fig. 2 (c) is the optimized $k$-location set distribution and definition $Q_{loss}^{max} = 3000$.

Probability of Exposing Semantic (PES) and Service Loss Expectation (SLE) are adopted to measure the privacy protection degree and service quality loss degree of location semantic of algorithm. SLE is calculated as the Eq. (16):

$$\text{SLE} = \frac{\sum_{K_{set}^t} dis(p_t, l_t)}{k} \tag{16}$$




(a) 12:00 a.m.

(b) 24:00 p.m.

**Figure 3:** Comparison of location semantic privacy protection and location service availability of vehicles under different algorithms

Analyzed from Fig. 3, the random response location set of LDPM is the most scattered, and some locations appear in the unreachable location, which is not conducive to privacy protection and service availability; $k-$ LDPM $k$ -location sets selected are more centralized, which reduce the service quality loss to the maximum extent, but it is most easily to expose the location semantic information of vehicle users; $k-$ LDPM $k$-location set selected by optimization is more scattered. Under the condition of ensuring the

maximum service availability, it enriches the location semantics of vehicles and reduces the possibility of behavior activity exposure of vehicle users. The greater the exposure $Q_{loss}^{max}$, the better the privacy protection effect of location semantics.

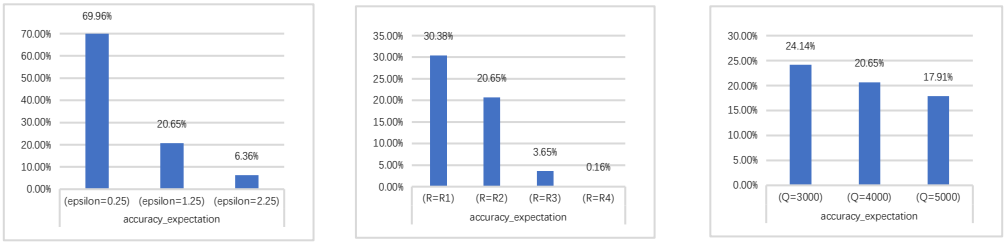
Random response location set of LDPM is the set of acquisition points of all locations of road network space. Under the RAPPOR perturbation mechanism and *k*-RR perturbation mechanism, the transmission cost of its private data is the vector and single value of |P| in length. *k*-location LDPM uses *k*-RR perturbation mechanism, and its privacy data transmission cost is k+1 value, which is positively correlated with the number of *k*-location. Based on location of $k$-LDPM and the comparison of LDPM RAPPOR perturbation mechanism, it greatly reduces the data transmission cost. Although it cannot use the data performance of *k*-RR perturbation mechanism, it balances the applicability of vehicle users to obtain the service quality.

### *4.3 Statistical analysis of regional traffic density*

The section conducts the comparative analysis of experimental features based on the local differential privacy technology of $k$-location set. It mainly includes three aspects: The accuracy impact of privacy budget $\epsilon$ on the traffic density statistical results, the accuracy impact of area coverage of density statistics on the traffic density statistical results, and the accuracy impact of service quality loss limit $Q_{loss}^{max}$ on the traffic density statistical results.

The experiment sets up four area coverage of density statistics: $R_1$—$R_4$. The sizes of spatial scale are $3\omega \times 3\omega$, $5\omega \times 5\omega$, $20\omega \times 20\omega$, $50\omega \times 50\omega$ respectively. Three groups of different privacy budgets are: $\epsilon_{low} = 0.25, \epsilon_{mid} = 1.25$ and $\epsilon_{high} = 2.25$. Three groups of loss limit of service quality set are: $Q_{loss\ low}^{max} = 3000$, $Q_{loss\ low}^{max} = 4000$, $Q_{loss\ low}^{max} = 5000$. The accuracy expectation measurement function of the regional traffic density statistical results is constructed by using the relative error [18], as shown in Eq. (17).

$$Accuracy = \frac{\sum_n \frac{\| |TD|_{real} - |TD|_{estimate} \|}{\| |TD|_{real} \|}}{n} \tag{17}$$




(a) under Different $\epsilon$ (b) under Different R (c) under the Different $Q_{loss}^{max}$

**Figure 4:** Regional traffic density statistical results

The accuracy expectation comparison of regional traffic density statistical results is shown in Fig. 4(a): Under the high privacy budget, the traffic density statistical result is closest to the real count; Under the low privacy budget, the accuracy of traffic density statistical results is low due to the large introduced noise.

The accuracy comparison of regional traffic density statistical results is shown in Fig. 4(b): In the larger area coverage, the traffic density statistical result is closest to the real count; In the smaller area coverage, the accuracy of traffic density statistical results is low.

The accuracy expectation comparison of regional traffic density statistical results is shown in Fig. 4(c): The greater the loss limit of service quality is, the closer the traffic density statistical result is to the real count; The less the limitation of service quality loss is, the lower the accuracy of the traffic density statistical result is.

## 5 Conclusions

Based on the data acquisition environment of the unbelievable Internet of Vehicles, this paper puts forward a location data acquisition method that satisfies the local differential privacy of $k$-location set, introduces the dynamic game model to optimize the $k$-location set, resists the location semantic inference attack in the real environment to the greatest extent and optimizes the level of location privacy protection of users under the condition of ensuring the service quality of users. On the basis of it, this paper further puts forward the regional traffic density statistical algorithm to satisfy the local differential privacy protection mechanism of the $k$-location set. The experimental analysis result shows that the method put forward by this paper has the advantages on the privacy protection effectiveness, data availability and high efficiency of data transmission.

**Funding Statement:** This work is supported by Major Scientific and Technological Special Project of Guizhou Province (20183001), Research on the education mode for complicate skill students in new media with cross specialty integration (22150117092), Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ014), Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ019) and Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ022).

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

## Reference

**Chen, R.; Li, H.; Qin, A. K.; Shiva, P. K.; Jin, H. X.** (2016): Private spatial data aggregation in the local setting. *Proceedings of the IEEE, International Conference on Data Engineering*, pp. 289-300.

**Dwork, C.; Lei, J.** (2009): Differential privacy and robust statistics. *41st Annual ACM Symposium on Theory of Computing*, pp. 371-380.

**Gao, Z. Q.; Cui, X. L.; Du, B.; Zhou, S.; Yuan, C. et. al.** (2019): Location data acquisition method to satisfy the local differential privacy. *Journal of Tsinghua University (Natural Science Version)*, vol. 59, no. 1, pp. 25-29.

**Haklay, M.; Weber, P.** (2008): Openstreetmap: user-generated street maps. *IEEE Pervasive Computing*. vol. 7, no. 4, pp. 12-18.

**Kairouz, P.; Oh, S.; Viswanath, P.** (2014): Extremal mechanisms for local differential privacy. *Advances in Neural Information Processing Systems*, pp. 2879-2887.

**Kim, H.; Hong, S.; Chang, J.** (2016): Hilbert curve-based cryptographic transformation scheme for spatial query processing on outsourced private data. *Data & Knowledge Engineering*, pp. 77-82.

**Liu, S. Y.; Liu, Y. H.; Lionel, M. N.; Fan, J. P.; Li, M. L.** (2010): Towards mobility-based clustering. *Proceedings of ACM KDD*, pp. 919-928.

**Li, C. L.; Lv, X.; Li, X.** (2018): Prediction of attack dynamics based on historical trajectory-anonymous algorithm. *Computer Engineering and Applications*, vol. 54, no. 2, pp. 119-124.

**Ma, M. J.; Du, Y. J.; Li, F. H.; Li, J. W.** (2016): Overview of location service privacy protection based on semantics. *Chinese Journal of Network and Information Security*, vol. 2, no. 12, pp. 1-11.

**Qu, Z. G.; Wu, S. Y.; Wang, M. M.; Sun, L.; Wang, X. J.** (2017): Effect of quantum noise on deterministic remote state preparation of an arbitrary two-particle state via various quantum entangled channels. *Quantum Information Processing*, vol. 16, no. 306, pp. 1-25.

**Qu, Z. G.; Cheng, Z. W.; Liu, W. J.; Wang, X. J.** (2019): A novel quantum image steganography algorithm based on exploiting modification direction. *Multimedia Tools and Applications*, pp. 7981-8001.

**Qu, Z. G.; Li, Z. Y.; Xu, G.; Wu, S. Y.; Wang, X. J.** (2019): Quantum image steganography protocol based on quantum image expansion and Grover search algorithm. *IEEE Access*, pp. 50849-50857.

**Shokri, R.** (2015): Privacy games: optimal user-centric data obfuscation. *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 2, pp. 299-315.

**Sarda, S.; Eickhoff, C.; Hofmann, T.** (2016): Semantic place descriptors for classification and map discovery. https://arxiv.org/pdf/1601.05952.pdf.

**Warner, S. L.** (1965): Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63-69.

**Wang, P. S.; Wang, Z. C.; Chen, T.; Ma, Q. J.** (2019): Personalized privacy protecting model in mobile social network. *Computers, Materials & Continua*, vol. 59, no. 2, pp. 533-546.

**Xiao, Y.; Xiong, L.** (2015): Protecting locations with differential privacy under temporal correlations. *22nd ACM Conference on Computer and Communications Security*.

**Xue, M.; Kalnis, P.; Pung, H.** (2009): Location diversity: enhanced privacy protection in location-based services. *The 4th International Symposium on Location and Context Awareness*, pp. 70-87. https://arxiv.org/pdf/1410.5919.pdf.

**Ye, Q. Q.; Meng, X. F.; Zhu, M. J.; Huo, Z.** (2018): Research overview of local differential privacy. *Journal of Software*, vol. 29, no. 7, pp. 159-183.

**Zhou, A. Y.; Yang, B.; Jin, C. Q.** (2011): Location based services: architecture and progress. *Chinese Journal of Computers*, vol. 34, no. 7, pp. 1155-1171.

**Zhen, H.; Ping, H.; Yan, W. W.** (2019). Crowdsourcing location data acquisition to satisfy the local differential privacy. *Journal of Computer Applications*, vol. 39, no. 3, pp. 763-768.