ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

# Lightweight adaptive differential privacy protection mechanism for mobile group perception data auction

*Hui Zhao*[1,*] *and Kai Ma*[1]

[1] Department of Electronics and Information Engineering, Bozhou University, Bozhou 236800, Anhui, China

**SUMMARY:** *In response to the problem of balancing privacy protection and data utility in mobile group perception data auction scenarios, this study proposes Lightweight adaptive differential privacy protection mechanism based on sparse shared neighbors for density peak clustering LDCR(SNDPC-LDCR). Firstly, the GRU network is used to predict time-series data, calculate data change rate, and combine PID error with remaining privacy budget to achieve adaptive sampling; Secondly, the recycling factor is introduced to optimize the allocation of sliding window privacy budget, and SNDPC clustering is used to complete data grouping. Noise is applied according to the minimum budget within the group and smoothed by Kalman filtering; Finally, non-sampled data will be published using the previous results, balancing privacy strength and computational efficiency. The experimental results show that the proposed algorithm has lower MAE than the compared methods under different sliding window lengths and overall privacy budgets, with smaller data distortion and better usability; The allocation of privacy budget is more reasonable, with stronger robustness to changes in window and budget parameters, and significant lightweight characteristics. Research has shown that the SNDPC-LDCR mechanism can achieve a balance between privacy protection and data utility in mobile group aware data auctions, providing efficient technical support for secure data transactions and open sharing.*

***KEYWORDS:*** *mobile group perception; Data auction; Differential privacy protection; Lightweight; Density peak clustering; Privacy budget allocation*

## 1 Introduction

With the rapid development of modern electronic information technology, the global data volume is growing exponentially, characterized by large scale, fast updates, multiple types, and strong authenticity. This has promoted the iterative innovation of data collection, storage, processing, and analysis technology systems, enabling deep exploration and efficient utilization of the potential value of data, and providing scientific basis for social production decisions. Currently, data has been included alongside land, capital, labor, and technology as core production factors [1]. For example, in the field of public health, medical big data can support epidemic monitoring, research hypothesis construction, and causal relationship inference; In the field of traffic management, road data analysis provides theoretical and practical guidance for traffic accident prevention and control, traffic flow prediction, and travel path planning; In the field of power grid dispatching, the large-scale collection and application of smart meter data effectively supports the implementation of business such as load forecasting, user grouping,

*huizhao_66@163.com
https://doi.org/10.65102/is20261255 Published: 10 June, 2026

and abnormal state detection. With the rapid evolution of emerging digital industries such as the Internet of Things, cloud computing, artificial intelligence, and blockchain, data has become a fundamental strategic resource and a core driving force for empowering enterprise transformation and high-quality innovation and development in society [2].

The fusion of mobile group perception data has dual dimensional attributes of time and space, originating from the behavioral activities of intelligent terminals or individuals in road networks, social scenes, and other environments. It can accurately mark the geographic spatial information of mobile targets at specific times or periods. The deep integration of mobile Internet and high-precision positioning technology has enriched the types of mobile Internet access terminals. Mobile phones, tablets, intelligent wearable devices, vehicle terminals, etc. can generate and return mobile group awareness data in real time, greatly reducing the collection and generation costs of mobile group awareness data. Various mobile applications rely on real-time positioning to push personalized services, such as ride hailing, bike sharing, map navigation and other location service platforms that operate based on user mobile group perception data; The real-time location sharing and hot spot check-in functions of social platforms also rely on mobile group perception data support [3]. In addition, mobile group perception data has been deeply integrated into government decision-making scenarios such as epidemic prevention and control, urban planning, and disaster prevention and reduction, highlighting its social application value. Mobile users are both the main creators of mobile group perception data and the beneficiaries of data services, facing severe risks of personal privacy breaches. If mobile group perception data is freely disclosed or abused in violation of regulations, it will directly threaten user privacy and security, and easily lead to the leakage of sensitive information such as personal identity characteristics, behavioral preferences, health status, and asset levels [4]: (1) The privacy protection standards of the data collection party lack unified control, and there is a behavior of privately reselling user data, which exposes user location privacy for a long time. For example, the Life360 platform sells user mobile group perception data, and the data black market can easily obtain the complete location trajectory information of tens of thousands of mobile phone users throughout the year. (2) Some service providers, in order to maximize the scale of data collection, implement out of scope collection and unauthorized access without the knowledge of users. For example, in 2023, the Southern Personal Information Protection Center found that nearly half of mobile finance apps had a frequency of over 55 location permission calls per minute. (3) The iteration of information technology and the expansion of information acquisition channels have made the privacy attack methods for mobile group perception data increasingly diversified. With just five days of historical trajectory data, over 80% of individual trajectory matching accuracy can be achieved. Even with the national level large-scale mobile group perception dataset, there is still a high risk of individual weight recognition, and the risk of privacy leakage cannot be avoided. The existing attack models can restore the complete travel trajectory of users from aggregated mobile data without additional background information support. The original aggregated data can accurately infer the user's location information and core mobile features, and the individual recognition accuracy of small-scale groups and discrete time period aggregated data is more prominent [5].

In the context of big data, frequent privacy breaches can reduce the public's willingness to share data, trigger resistance, and restrict the development of the mobile group perception data industry, as well as hinder the process of social digital transformation. Relying solely on legal constraints is passive and cannot eliminate the hidden dangers of mobile group perception data privacy breaches from the source. To improve the privacy and security protection system for mobile group perception data, it is urgent to rely on technological means to complete sensitive information desensitization and security reinforcement before data release and application,

enhance the pre emptive and proactive nature of privacy protection, and reduce the risk of leakage. This study proposes a privacy protection scheme for mobile group perception data publishing scenarios that balances privacy, security, and data availability. A lightweight adaptive differential privacy protection mechanism based on sparse shared neighbors and density peak clustering (LDCR) is proposed for mobile group perception data auction, which has important theoretical and engineering application value for promoting the open sharing and industrialization of mobile group perception data resources, safeguarding the legitimate rights and interests of data providers, and improving the social data security governance system.

# 2 Related research

## 2.1 Methods based on data interference technology

Data interference technology mainly achieves privacy desensitization and security protection by directly tampering or blurring the original trajectory data content. Data interference techniques mainly include three types: (1) fake data methods. The fake data method interferes and confuses real travel trajectories by implanting false positioning points or fictional trajectory segments into the original trajectory sequence. For example, Orabe et al. [6] introduced a false trajectory obfuscation strategy to generate disguised paths and constructed an enhanced privacy protection framework, which can effectively resist the leakage of user location information. Kazan & Reiter [7] proposed a segmented false trajectory privacy protection scheme through a three-layer process of false position generation, segmented trajectory construction, and global trajectory fusion. The fake data method has the advantages of simple principle and easy engineering implementation, but it will occupy a large amount of database storage resources and have low data availability; If the forged trajectory does not conform to the spatiotemporal travel rules, it is easy for attackers to identify and distinguish it. In addition, when attackers have access to some real prior data, they can use correlation analysis and trajectory matching techniques to restore the user's original real travel trajectory. (2) Inhibition method. The core idea is to selectively delete or conceal high-frequency sensitive spatiotemporal nodes in the trajectory to reduce privacy exposure risks. For example, Imola et al. [8] achieved trajectory data privacy protection by drawing spatiotemporal position information entropy flow maps, constructing optimization cost functions, and implementing local spatiotemporal node removal. Rahman et al. [9] implemented global suppression based on trajectory frequency statistics and performed local node suppression on highly sensitive segments within the trajectory to further avoid the risk of raw data leakage. The suppression method has the advantages of easy operation and significant protective effect, but residual trajectory fragments are prone to attackers relying on contextual inference to reverse mine sensitive privacy information; If the proportion of sensitive nodes removed is too high, it can easily cause serious information loss and significantly reduce the subsequent analysis and application value of trajectory data. (3) Generalization method. By replacing precise location information with fuzzy spatial ranges, trajectory privacy protection is achieved by reducing data granularity. For example, Sebastian [10] proposed a fake trajectory generation framework that integrates position similarity measurement and equivalent probability interval partitioning, relying on dual similarity constraints to achieve trajectory K anonymous privacy protection. Poojari [11] generates K-1 virtual trajectories based on real movement trajectory sequences, completing the anonymization and privacy protection of user trajectory sets. K-anonymity is the mainstream implementation technique of generalization methods, but if attackers have sufficient prior knowledge of the background, the effectiveness of privacy protection in traditional K-anonymity mechanisms will significantly decrease.

## 2.2 Methods based on cryptographic technology

Cryptography technology is a privacy protection method that relies on the fundamental theory of cryptography to encrypt and transform raw data and computational processes. For example, Nagarajan [12] combines homomorphic encryption algorithm with secure comparison protocol, and introduces ciphertext compression algorithm to effectively solve the privacy leakage problem of trajectory data in similarity calculation process. Williamson & Prybutok [13] constructed a three factor authentication protocol based on elliptic curve cryptography, which integrates user passwords, biometric features, and smart device hardware identifiers to complete triple identity verification. The privacy protection method based on cryptography utilizes one-way irreversible encryption functions to anonymize user location information, which can ensure the integrity of location service functions while avoiding the risk of user identity and geographic location sensitive information leakage. However, the encryption and decryption operations of cryptographic technology will increase the system's time and resource consumption, and the overall computational complexity is relatively high.

## 2.3 Methods based on differential privacy technology

Differential privacy utilizes noise perturbation mechanisms to resist attackers' background knowledge attacks, achieving quantitative characterization and controllable constraints of privacy leakage risks, and effectively suppressing the privacy leakage risks of positional sequence data. Existing research combines differential privacy and generalization techniques, first performing granular generalization on trajectory sensitive location nodes, and then adding noise disturbances to further enhance privacy protection strength. For example, Hotz et al. [14] proposed an anonymization model that combines differential privacy and generalization strategies to achieve privacy protection for sensitive vehicle trajectories. Singh & Gupta [15] introduced K-means vector space clustering to achieve trajectory dimensionality reduction optimization, and embedded differential privacy noise mechanism to achieve privacy desensitization of trajectory data. Al-Hawawreh & Hossain [16] constructed a noise prefix tree and combined it with Markov chains to model and reconstruct the spatiotemporal correlation of trajectories, generating privacy enhanced equivalent trajectory sequences. However, generalization operations can easily cause irreversible loss of trajectory information, thereby weakening the usability of subsequent mining and analysis of the dataset. Another study used graph and tree structures to model and express trajectory features and associated attributes. For example, Chukwunweike et al. [17] extended the trajectory similarity index structure based on the traditional R-tree and constructed a privacy enhanced trajectory tree DPTS tree, which can resist multi-dimensional inference attacks. However, the construction and iterative processing of graph and tree structures require a significant amount of computational resources and time overhead. Biswas et al. [18] used graph models to implement differential privacy constraints for a single trajectory and proposed a trajectory publishing framework based on differential privacy called PTDP. They used density clustering to identify trajectory hotspots and outliers, and achieved position blur processing through generalization. Some studies have constructed a differential privacy protection paradigm for semantic perception based on trajectory semantic features. For example, Feretzakis et al. [19] proposed a semantic preserving differential privacy trajectory synthesis scheme, which combines Markov chain theory to generate synthesized trajectories that combine semantic consistency and privacy security. However, the complexity of trajectory semantic parsing and high-precision semantic model construction is high, and the difficulty of algorithm design and implementation is significantly increased. Pasham [20] designed the semantic aware personalized trajectory differential privacy framework OPTDP, which adaptively adjusts the noise injection intensity based on the semantic sensitivity of

different regions, achieving semantic driven differential trajectory privacy protection. In addition, many studies have introduced localized differential privacy architectures. For example, Akter et al. [21] proposed a localized differential privacy trajectory synthesis framework LDPTrace, which designs trajectory synthesis algorithms through feature probability modeling to generate high simulation privacy trajectories. Rodríguez et al. [22] proposed the differential privacy location protection algorithm LPADP, which embeds Laplacian noise to achieve location information desensitization. Kobayashi et al. [23] designed a personalized differential privacy trajectory publishing mechanism based on Hilbert curves, and implemented adaptive generalization for differentiated privacy preferences to meet users' personalized privacy protection needs.

## 2.4 Urgent Problems to be Solved

There are many key issues that urgently need to be addressed for differential privacy protection of trajectory data [24]: (1) accurate screening of privacy sensitive location points. Trajectory data contains a large amount of user stay location information, and there are significant differences in the degree of threat to privacy and security at each location point. Prioritizing the identification and protection of highly sensitive location points is the core prerequisite for improving privacy protection efficiency. How to design efficient and accurate filtering algorithms to quickly locate the key sensitive points that are most likely to leak user identity characteristics and behavior patterns from massive trajectory data, and implement targeted graded and precise protection, is a hot issue that urgently needs to be addressed. (2) Scientific allocation and rational distribution of privacy budget. The privacy budget (ε value) is used to characterize the level of noise injection in the dataset and is a core parameter for adjusting the strength of system privacy protection. At present, there is a lack of unified standardization criteria for setting the ε value, which needs to be dynamically adjusted according to specific application scenarios and is highly subjective. If excessive noise is applied to a few points in the trajectory, it can easily lead to conflicts between the perturbed position and the spatial continuity of the original trajectory. Therefore, the setting and allocation of privacy budget need to comprehensively consider multidimensional influencing factors, construct an efficient privacy budget optimization model, and achieve a balance between privacy protection intensity and resource consumption. (3) Balancing privacy protection and data utility optimization. The parameter configuration of privacy budget shows a significant negative correlation with the intensity of data disturbance. When the ε value is small, the system will inject more noise to enhance privacy protection capabilities, but this will lead to an increase in data distortion and a decrease in availability; When the value of ε increases, the amount of noise injection decreases, the strength of privacy protection weakens, and the availability of data significantly improves. If the noise injection is excessive, it can easily mask the key spatiotemporal features of the trajectory, which seriously affects the application value of trajectory data in practical scenarios such as range queries, frequent pattern mining, and path planning. Therefore, how to minimize data distortion and achieve a balanced optimization of privacy protection and data utility while strictly ensuring user privacy and security is also a key challenge that urgently needs to be overcome.

# 3 SNDPC-LDCR algorithm

## 3.1 Overall Framework

In the mobile crowdsensing scenario, the trusted perception platform aggregates mobile crowdsensing data within various geographic regions based on business needs, and applies differential privacy protection to the aggregated data sequence using the SNDPC-LDCR algorithm to generate a sequence with privacy and security guarantees, supporting subsequent data auction transactions and business analysis processes. The overall architecture of SNDPC-LDCR algorithm is shown in Figure 1.

The algorithm execution process shown in Figure 1 is as follows: (1) Time series prediction is performed on aggregated data from various geographical regions, and the data time series change rate is synchronously calculated. Through an adaptive sampling mechanism, the aggregated data sample set that requires noise disturbance and publishing processing at the current time is selected. Then, the initial allocation of privacy budget is completed by combining the regional data change rate and budget recovery factor; (2) Taking into account the data change rate and the remaining privacy budget of the sliding window at the next moment, the SNDPC clustering algorithm is introduced to cluster the sampled dataset, and differential privacy noise disturbance is applied to the total amount of data in each group with the minimum privacy budget within the group as the constraint; (3) According to the weight ratio of a single predicted data to the total predicted data within the group, the fine allocation of global perturbation values is completed, and the independent noise perturbation values corresponding to each sampled data are calculated. (4) The noise disturbance values of each sampled data are introduced into the Kalman filter to achieve smooth optimization. For non sampled data that is not included in the sampling set, the published result of the previous sampling time is directly used as the approximate value of the current time to complete the overall data publishing.

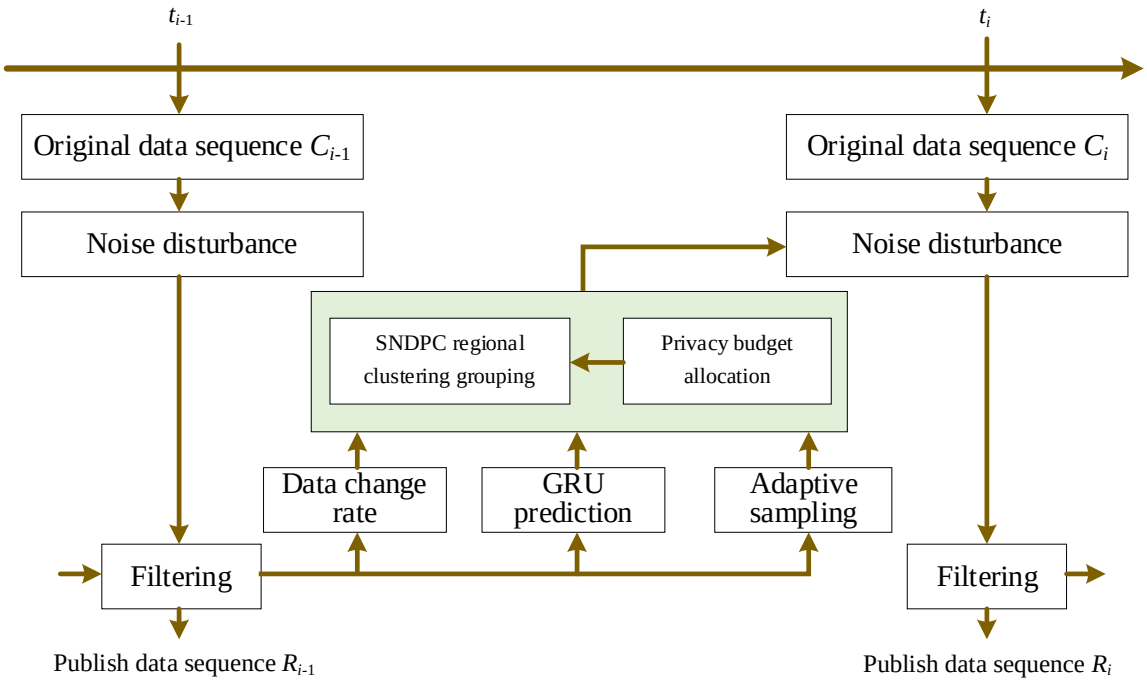



*Figure 1: Model of SNDPC-LDCR algorithm*

## 3.2 Calculation of Data Change Rate

The continuous generation of mobile group perception data over time will evolve and change user behavior patterns. During peak hours in the morning and evening, the travel behavior of mobile entities tends to cluster and have strong regularity. However, during peak transition periods, the spatial distribution rate of mobile groups accelerates, and the aggregated data quickly transitions between sparse and peak states [25]. In the early hours of the morning, individual small travel behaviors can cause significant fluctuations in aggregated data, and the probability of inferring individual movement trajectories based on aggregated data is significantly increased, which will exacerbate the risk of privacy leakage of sensitive user information. In response to the above issues, this article introduces the data change rate index to calculate the slope and inclination changes of aggregated data at adjacent times, and uses it as the core basis for subsequent adaptive privacy policy regulation. It dynamically adapts the privacy protection strength based on the dynamic evolution law of data, which can effectively maintain and improve the overall availability of trajectory publishing data [26].

At any timestamp *i*, three states of data can be obtained from the slope change: (1) the data change accelerates, and the slope $k_2$ will be greater than $k_1$; (2) The rate of change of data at time *i* is slower than in the previous period, and at this time, the slope $k_1$ will be greater than $k_2$; (3) $k_2 = k_1$, The predicted rate of change is the same as the previous adjacent time interval. Taking region j as an example, first, the predicted data $p_i^j$ obtained from the GRU model at time *i* and historical release data $r_{i-2}^j$ and $r_{i-1}^j$ can be used. Then, the data slopes $k_1$ and $k_2$ at adjacent time intervals $[i-2, i-1]$ and $[i-1, i]$ can be calculated using the following equation:

$$k_1 = \frac{r_{i-1}^j - r_{i-2}^j}{(i-1)-(i-2)} \tag{1}$$

$$k_2 = \frac{p_i^j - r_{i-1}^j}{i-(i-1)} \tag{2}$$

The current time interval data has not changed, and the tilt angle is 0. At this time, the data change rate for the next time period is defined as the proportion of the relative maximum tilt angle of 90 °. When data is changing in two time periods, the data change rate is defined as the proportion of the second change angle to the total of the two change angles. When the second change accelerates, the change rate is greater than 0.5, and when the second change slows down, the change rate is less than 0.5. When the two data changes remain unchanged, the change rate is 0.5.

$$v_i^j = \begin{cases} \left|\arctan k_2\right| / 90^\circ, k_1 = 0 \\ \left|\arctan k_2\right| / \left(\left|\arctan k_1\right| + \left|\arctan k_2\right|\right), k_1 \neq 0 \end{cases} \tag{3}$$

## 3.3 Adaptive Data Sampling Mechanism

The SNDPC-LDCR algorithm dynamically adjusts the sampling interval by combining PID error with the current remaining privacy budget: (1) when the PID error increases, the sampling frequency is appropriately increased to improve data timeliness and accuracy; (2) Constrain the increase in sampling frequency through the remaining privacy budget to avoid privacy budget depletion caused by oversampling. Unlike traditional LDCR algorithms, in the SNDPC-LDC model, feedback error is defined as the absolute error between the predicted value *p* at the

current timestamp *i* and the filtered output value *r* at the previous sampling time. The adaptive adjustment process of the sampling interval for the aggregated position range *j* can be defined as follows:

$$E_{k_l} = \left| p_i^j - r_{k_l}^j \right| \tag{4}$$

$$\delta^j = G_P E_{k_l}^j + G_I \sum_{o=l-\Phi-1}^{l} E_{k_o}^j \Big/ \Phi + G_D E_{k_l}^j / k_l - k_{l-1} \tag{5}$$

$$T = \max\left\{1, T_l + \theta\left(1 - \left(\varepsilon_{ri}{}^j \cdot \delta_i^j\right)^2\right)\right\} \tag{6}$$

where, *j* is the sampling geographic range, $\varepsilon_{ri}{}^j$ is the remaining budget at the current time, $\delta^j$ is the PID error at the current sampling time, $E_{k_l}^j$ is the proportional error, and $G_P$ is the proportional gain; $G_I \sum_{o=l-\Phi-1}^{l} E_{k_o}^j / \Phi$ is the integration error, $G_I$ is the integration gain, $\Phi$ represents the cumulative average of errors over the past $\Phi$ time points, $G_D E_{k_l}^j / k_l - k_{l-1}$ is the differential gain, and $G_D$ is the differential gain.

The new sampling interval is determined through the coordinated regulation of data change rate, remaining privacy budget, and PID error. The data change rate reflects the temporal evolution trend of the predicted data in real time, while the remaining privacy budget can increase the sampling interval in a timely manner when the budget is insufficient, effectively reducing data utility loss. The PID error accurately characterizes the actual change characteristics of the published data. The organic integration of the three can ultimately achieve efficient and adaptive sampling control.

## 3.4 Privacy Budget Allocation Mechanism

To improve the rationality and utilization efficiency of privacy budget allocation within the sliding window [27]: (1) With data change rate as the core, allocate more privacy budget to areas with drastic data changes, ensuring effective protection of data temporal trends; (2) Introducing a recycle factor to quantify the difficulty of recycling privacy budget within the sliding window, and using it as an important regulatory factor for privacy budget allocation, to accurately control the current remaining budget expenditures and avoid data utility damage caused by insufficient budget in the future.

The recycling factor can be defined as the time interval between the first sampling point $t_{fs}$ in the current timestamp sliding window and the starting point $t_{ws}$ of the window plus 1. The specific definition is as follows:

$$\text{recycle} = t_{fs} - t_{ws} + 1 \tag{7}$$

The larger the recycling factor value, the longer the privacy budget recycling cycle, and the privacy budget available for subsequent publishing nodes is difficult to increase in a short period of time. Reasonably controlling the current budget expenditure can effectively ensure the availability and accuracy of subsequent timestamp data publishing.

For the sampling geographic range *j*, the privacy budget allocation process is as follows: (1) Calculate the remaining privacy budget within the sliding window at the current time and obtain the current recycling factor recycle value; (2) Calculate the allocation ratio $\tau_{\text{alllocated}}$ of the remaining privacy budget, introduce a recovery factor and sampling interval to constrain this

allocation ratio, and avoid serious damage to the subsequent data utility caused by the long-term inability to recover the privacy budget; (3) Allocate the remaining privacy budget based on the calculated allocation ratio $\tau_{\text{alllocated}}$, and set $\tau_{\min}$ and $\tau_{\max}$ as the lower and upper limits of the allocation ratio, respectively; (4) Compare the calculated remaining budget allocation value $\varepsilon_{\text{allocation}}^{j}$ with the maximum privacy budget allocation value $\tau_{\max}$, and take the smaller of the two as the final privacy budget allocation for this time.

## 3.5 SNDPC clustering algorithm

Given dataset $X_{N\times D}=\{x_1,x_2,\cdots,x_N\}$, $N$ represents the number of data points, and $D$ is the data dimension. Assuming $G=(V,E)$ is the complete graph of dataset $X_{N\times D}$, $T_X=(V,E_X)$ is the minimum spanning tree constructed on $G$, where $V$ is the set of vertices consisting of $N$ data points, and $E_X$ represents the set of all edges on the minimum spanning tree. The weight of the connecting edge between data points $x_i$ and $x_j$ on $T_X$ is represented as $d(x_i,x_j)$, which is the Euclidean distance between data points $x_i$ and $x_j$.

Assuming that the paths of $x_i$ and $x_j$ on the minimum spanning tree are $\{p_1,p_2,\cdots,p_n\}$, with edges $(p_m,p_{m+1})\in E_X$ and $1\le m\le n-1$. The geodesic distance between $x_i$ and $x_j$ is defined as:

$$g(x_i,x_j)=\sum_{m=1}^{n-1}d(p_m,p_{m+1}) \tag{8}$$

where, $d(p_m,p_{m+1})$ represents the Euclidean distance between $p_m$ and $p_{m+1}$.

Next, the density of data point $x_i$ is:

$$\rho(x_i)=\sum_{j=1,j\neq i}^{N}\exp\left(-\frac{g(x_i,x_j)^2}{2\sigma^2}\right) \tag{9}$$

where, $\sigma$ is the variance. The calculation method for $\sigma$ is as follows: first, sort the geodesic distances between all point pairs in ascending order, i.e. $\{g_1,g_2,\cdots,g_{N(N-1)/2}\}$; Secondly, let $t_{\text{h}}=[d_{\text{c}}n(n-1)/2]$, where $d_{\text{c}}$ is the truncation distance and $[\cdot]$ is the rounding operation. Finally, let $\sigma=g_{t_{\text{h}}}$.

When $d_{\text{c}}$ increases, $t_{\text{h}}$ also increases, and the variance $\sigma$ also increases accordingly. According to the characteristics of Gaussian distribution, the density $\rho(x_i)$ also increases accordingly. On the contrary, when $d_{\text{c}}$ is reduced, the density $\rho(x_i)$ also decreases accordingly.

In addition, relative distance can be defined as:

$$\delta(x_i)=\begin{cases}\min_{j:\rho(x_i)<\rho(x_j)}\left(g(x_i,x_j)\right), \rho(x_i)<\rho(x_j)\\ \max_j\left(g(x_i,x_j)\right), \text{otherwise}\end{cases} \tag{10}$$

Finally, let $\gamma(x_i)=\rho(x_i)\times\delta(x_i)$, take the $K$ data points with larger $\gamma$ as the cluster centers, denoted as $\{C_1,C_2,\cdots,C_K\}$.

Assigning $d_{\text{c}}$ values of 0.02 and 0.2 respectively to calculate the local density of data

points and determining the optimal number of clusters through clustering evaluation indicators can improve clustering performance, but it will increase the computational time cost of the algorithm. In this regard, this article dynamically adjusts the $d_c$ of data points through the sparsity factor of data points.

**Definition 1** (Sparse Factor): The sparsity factor $\xi(x_i)$ of data point $x_i$ is the average distance from data point $x_i$ to its *k*-nearest neighbors, which can be expressed as:

$$\xi(x_i) = \frac{1}{k} \sum_{x_j \in kNN(x_i)} d(x_i, x_j) \tag{11}$$

where, $kNN(x_i)$ is a set of *k* nearest neighbors of $x_i$.

The sparsity factor of data points located in dense areas is relatively small, while the sparsity factor of points located in sparse areas is relatively large. Therefore, this article assigns $d_c$ to data point $x_i$ based on $\xi(x_i)$:

$$d_c(x_i) = m_1 \xi(x_i) + m_2 \tag{12}$$

where, $m_1 = (d_{\max} - d_{\min})/(0.02 - 0.2)$, $m_2 = 0.02 - m_1 \times d_{\max}$, $d_{\max}$ and $d_{\min}$ are the maximum and minimum values of the sparsity factor for all data points, respectively.

**Definition 2** (Inconsistency Factor): The inconsistency factor between data points in dataset *X* and $(x_i, x_j)$ is defined as:

$$\zeta(x_i, x_j) = \frac{\sum_{u \in \overline{SNN}(x_i|x_j), v \in \overline{SNN}(x_i|x_j)} d(u,v) / N_C}{\sum_{m,n \in \overline{SNN}(x_i|x_j)} d(m,n) / N_{C_1} + \sum_{p,q \in \overline{SNN}(x_i|x_j)} d(p,q) / N_{C_2}} \tag{13}$$

where, $N_C = \left|\overline{SNN}(x_i|x_j)\right| \left|\overline{SNN}(x_j|x_i)\right|$, $N_{C_1} = \left|\overline{SNN}(x_i|x_j)\right|$, $NC_2 = \left|\overline{SNN}(x_j|x_i)\right|$.

The specific steps of the SNDPC clustering algorithm are as follows:

Input: Dataset $X_{N \times D} = \{x_1, x_2, \cdots, x_N\}$, number of clusters *K*, number of nearest neighbors *k*.

Output: Clustering results.

Step 1: Calculate the Euclidean distance $d(x_i, x_j)$ between all point pairs in dataset *X*, generate the minimum spanning tree $T_X$, and determine the *k*-nearest neighbors of each data point.

Step 2: Calculate the geodesic distance $g(x_i, x_j)$ between all point pairs according to equation (8).

Step 3: Calculate the $d_c$ of each point based on the k-nearest neighbors of the data points and equations (12) and (13).

Step 4: Calculate the density $\rho(x_i)$ of all data points according to equation (10).

Step 5: Sort all data points in descending order of their $\rho(x_i)$ values, and calculate the relative distance $\delta(x_i)$ of all data points according to equation (11).

Step 6: Calculate the $\gamma(x_i)$ of all data points and use the *K* data points with $\gamma$ as the clustering center $C = \{C_1, C_2, \cdots, C_K\}$.

Step 7: Construct the minimum spanning tree $T_C$ with *K* cluster centers $C = \{C_1, C_2, \cdots, C_K\}$.

Step 8: Calculate the inconsistency factor $\zeta(x_i, x_j)$ between adjacent points on *K*-1 paths according to equation (13), take the edge connected to the point pair with the highest $\zeta$ on *K*-1 paths as the inconsistent edge, delete all inconsistent edges of $T_X$, and obtain the clustering result.

# 4 Experimental analysis

## 4.1 Experimental setup

The experiment uses the T-driver dataset, which contains GPS trajectory information of 10267 taxis in Beijing from April 3 to April 9, 2025 [28]. The average sampling interval of the data is about 153 seconds, and the total number of location points is about 14.9 million. This article aggregates mobile group perception data from the T-driver dataset with longitude ranges of [116.37E 116.48E] and latitude ranges of [39.67 ° N, 40.13N], and publishes taxi count values at 15-minute intervals. By dividing the selected latitude and longitude range with different accuracies, two map regions with varying degrees of data sparsity were obtained, as shown in Table 1.

*Table 1: Regional Division Information Table*

| T-driver | Single area range (km²) |
|---|---|
| Region 1 | 1.07×0.83 |
| Region 2 | 2.65×2.05 |

This experiment measures the utility of privacy protected published data by calculating the Mean Absolute Error (MAE) between the privacy protected published data and the original aggregated data [29]. The smaller the MAE, the lower the degree of data distortion and the higher the usability. This article establishes a GRU network for each region in two maps, with the previous 15 historical published data as inputs and 1 predicted data output for each network. The number of neurons in each GRU network layer in region 1 is 15, 30, and 1, respectively. The hidden layer uses the softsign function as the activation function; The number of neurons in the GRU network hierarchy in region 2 is 15, 45, and 1, respectively. The relu function is used as the hidden layer activation function, and the specific settings are shown in Table 2. Set the parameters of the PID controller to: $G_P = 0.9$, $G_I = 0.1$, $G_D = 0$, $\Phi = 3$, $\theta = 10$. The acceptable minimum rate of change for the regional grouping mechanism is set to 0.1.

*Table 2: Hyperparameter Settings for GRU Model*

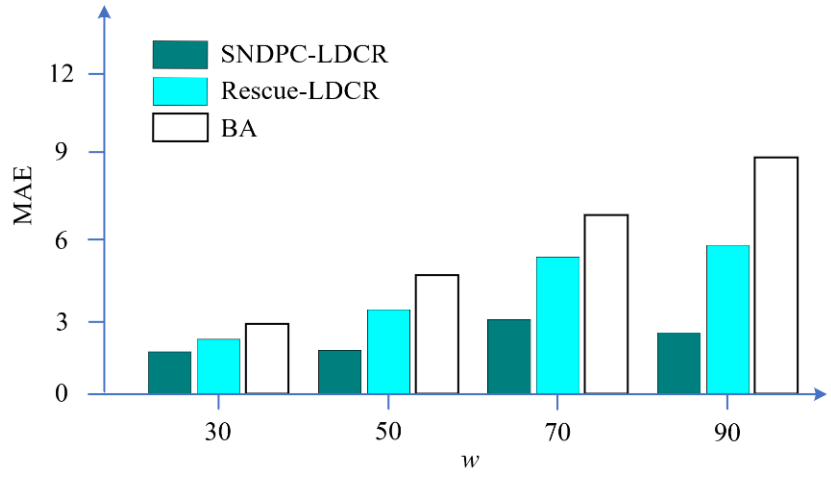
| T-driver | network architecture | Hidden layer activation function | loss function | Epochs |
|---|---|---|---|---|
| Region 1 | (15,30,1) | softsign | MSE | 900 |
| Region 2 | (15,45,1) | relu | MSE | 700 |

To increase the credibility of the experimental results, the SNDPC-LDCR algorithm model was compared with two existing differential privacy data protection schemes, BA and Rescue-LDCR [30], to verify the effectiveness of the SNDPC-LDCR algorithm model.

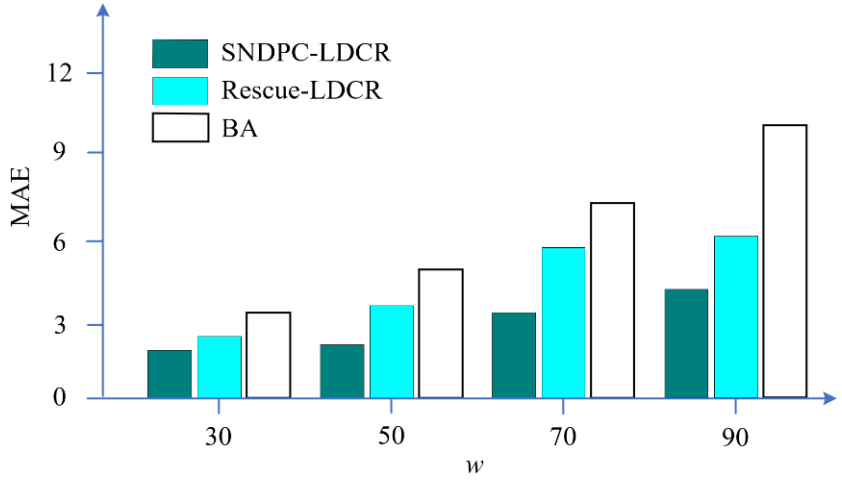
## 4.2 Result Analysis

Differential privacy continuously recovers and allocates privacy budget through the continuous movement of sliding windows over time series, and the length of the window to some extent

affects the difficulty of recovering privacy budget. Figure 2 evaluates the data utility of differential privacy protection results of existing schemes under different sliding window lengths w.




(a) Region 1



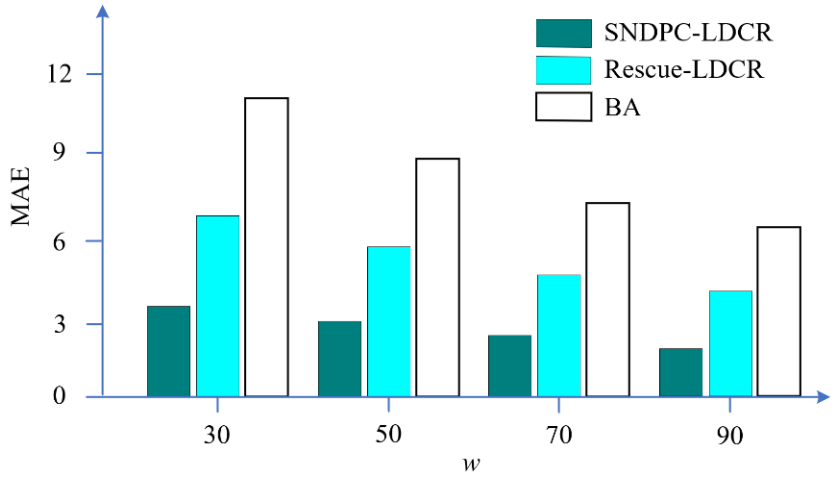

(b) Region 2

*Figure 2: MAE test results (w influence)*

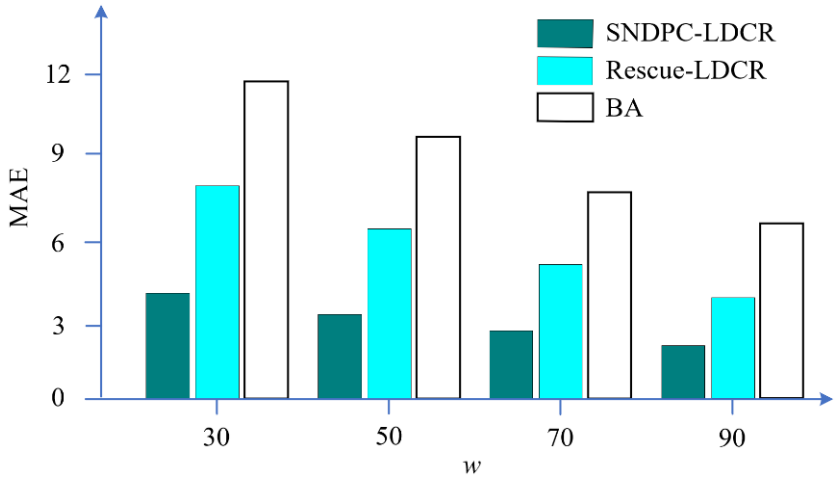
According to the results in Figure 2, it can be seen that as the sliding window length w increases, the average absolute error MAE shows an upward trend, with the BA error showing the most obvious trend as w increases. The error increase of the Rescue-LDCR and SNDPC-LDCR algorithm models proposed in this paper is relatively slow. It can be seen that the change in w has a significant impact on the BA algorithm. The reason is that BA uses window length as one of the parameters in the privacy budget allocation process. The larger the window, the less privacy budget is allocated, and the degree of noise disturbance is intensified, resulting in a significant deterioration of data utility. However, Rescue-LDCR and LDCR do not involve window length in the privacy budget allocation process, so they can have a smoother and more stable error than BA when w changes. In addition, the SNDPC-LDCR algorithm model performs better than BA and Rescue-LDCR on both datasets. The main reason is that the SNDPC-LDCR algorithm model determines the adjustment of sampling intervals during the sampling process through three indicators: data change rate, PID control error, and remaining

budget of the region. It is more cautious about adding noise, and at the same time, special consideration is given to data change rate in the privacy protection process, so that the noise disturbance results can better preserve the original data features.

The overall privacy budget ε determines the upper limit of the disposable budget within the sliding window, and also determines the strength of privacy protection that the algorithm can provide. The smaller the ε, the more noise is added to the disturbance process of the sampling area data, and the stronger the privacy protection, but the greater the damage to the data utility. Figure 3 shows the comparative evaluation results of the average error changes of each scheme under the overall privacy budget variation.




(a) Region 1




(b) Region 2

*Figure 3: MAE test results (ε influence)*

According to the results in Figure 3, it can be seen that the error values of BA, Rescue-LDCR, and SNDPC-LDCR all show a decreasing trend with increasing. The BA algorithm is most sensitive to changes in ε, because in the privacy budget allocation process of the BA algorithm, only half of the overall privacy budget is used to perturb the original data with noise, resulting in a smaller privacy budget that can be controlled by the perturbation mechanism, a larger amount of noise introduced at each sampling moment, and severe loss of data utility.

In addition, the comparison shows that SNDPC-LDCR is still superior to the other two

privacy protection schemes in protecting the availability of differential privacy perturbation data. The main reason is that the privacy budget allocation of SNDPC-LDCR depends on the changes in regional aggregated data, and can adjust the amount of privacy budget used according to the data change status, greatly preserving the original data change trend. At the same time, the regional grouping mechanism has divided regions with similar data change rates into groups, so that even using the minimum privacy budget within the group for noise disturbance will not excessively damage the data utility, but can protect the data utility of regions with larger changes as much as possible. Finally, with the cooperation of adaptive sampling mechanism, the use of privacy budget is more cautious, improving the rationality of budget application and effectively preventing the waste of privacy budget.

# 6. Concluding remarks

This article proposes a lightweight adaptive differential privacy protection mechanism based on sparse shared neighbor density peak clustering (SNDPC-LDCR), which takes mobile group perception spatiotemporal data as the processing object, and deeply integrates GRU temporal prediction, data change rate analysis, adaptive sampling, dynamic allocation of privacy budget, and SNDPC clustering to construct a lightweight and adaptive privacy protection framework. The comparative experimental results on the T-driver taxi trajectory dataset show that the SNDPC-LDCR mechanism proposed in this paper has significantly lower mean absolute error (MAE) than traditional schemes such as BA and Rescue-LDCR under different sliding window lengths and overall privacy budget conditions. The data distortion is smaller, the availability is better, and it has stronger robustness to parameter changes. Its lightweight characteristics are outstanding, and it can be efficiently deployed in resource limited mobile group sensing terminals and platform terminals.

Future research direction: (1) Building a lightweight differential privacy model with spatiotemporal joint constraints for continuous dynamic trajectories and multi-user collaborative perception scenarios, to enhance privacy protection capabilities under complex mobile behaviors. (2) Combining reinforcement learning to achieve online autonomous optimization of budgets, dynamically adjusting noise intensity based on real-time attack risk and data value, further enhancing adaptive capabilities. (3) Integrating multi-scale clustering and adaptive smoothing algorithms to enhance the ability to differentiate sparse, hotspot, and outlier data, maintaining stable utility under extreme data distributions. (4) Research on a lightweight privacy collaboration mechanism that supports multi-party secure computing, achieving unified cross domain privacy control under the premise of trusted data transactions. (5) Integrate the proposed mechanism with the actual mobile group perception system, conduct performance testing and engineering implementation in real scenarios, and promote the practical and industrial transformation of theoretical achievements.

# Funding

This work was supported by the Key Research Project of Natural Science in universities of Anhui Province (No.2023AH052269), and the High-Level Talent Research Initial Fund Project of Bozhou University (No. BYKQ202413).

# About The Author

Hui Zhao received the PhD degree from the University of Science and Technology of China, in

2021. She is currently a lecturer in Bozhou University. Her main research direction is data trading in MCS, auction theory, and data security and privacy.

Kai Ma received the MS degree from the University of Science and Technology of China, in 2019. He is currently a teaching assistant in Bozhou University. His main research direction is spatial crowdsourcing, vehicular ad hoc networks, auction theory, and privacy-preserving mechanism.

# References

[1] Blanco-Justicia A, Sánchez D, Domingo-Ferrer J, et al. A critical review on the use (and misuse) of differential privacy in machine learning[J]. ACM Computing Surveys, 2022, 55(8): 1-16.

[2] El Ouadrhiri A, Abdelhadi A. Differential privacy for deep and federated learning: A survey[J]. IEEE access, 2022, 10: 22359-22380.

[3] Nanayakkara P, Smart M A, Cummings R, et al. What are the chances? explaining the epsilon parameter in differential privacy[C]//32nd USENIX Security Symposium (USENIX Security 23). 2023: 1613-1630.

[4] Aziz R, Banerjee S, Bouzefrane S, et al. Exploring homomorphic encryption and differential privacy techniques towards secure federated learning paradigm[J]. Future internet, 2023, 15(9): 310.

[5] Batool H, Anjum A, Khan A, et al. A secure and privacy preserved infrastructure for VANETs based on federated learning with local differential privacy[J]. Information Sciences, 2024, 652: 119717.

[6] Orabe Z, Vasankari A, Pahikkala T, et al. Securing deep learning models with differential privacy for cardiovascular disease prediction[J]. Biomedical Signal Processing and Control, 2026, 112: 108502.

[7] Kazan Z, Reiter J P. Prior-itizing privacy: A Bayesian approach to setting the privacy budget in differential privacy[J]. Advances in Neural Information Processing Systems, 2024, 37: 90384-90430.

[8] Imola J, Murakami T, Chaudhuri K. {Communication-Efficient} triangle counting under local differential privacy[C]//31st USENIX security symposium (USENIX Security 22). 2022: 537-554.

[9] Rahman M H, Mowla M M, Shanto S. Differential privacy enabled deep neural networks for wireless resource management[J]. Mobile Networks and Applications, 2022, 27(5): 2153-2162.

[10] Sebastian G. Privacy and data protection in ChatGPT and other AI chatbots: strategies for securing user information[J]. International Journal of Security and Privacy in Pervasive Computing (IJSPPC), 2023, 15(1): 1-14.

[11] Poojari R. Privacy-Preserving Generative AI in Healthcare Systems Using Federated Learning Approaches[J]. International Journal of Data Science and IoT Management

System, 2026, 5(1): 78-88.

[12] Nagarajan G. AI-Integrated Cloud Security and Privacy Framework for Protecting Healthcare Network Information and Cross-Team Collaborative Processes[J]. International Journal of Engineering & Extended Technologies Research (IJEETR), 2023, 5(2): 6292-6297.

[13] Williamson S M, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare[J]. Applied Sciences, 2024, 14(2): 675.

[14] Hotz V J, Bollinger C R, Komarova T, et al. Balancing data privacy and usability in the federal statistical system[J]. Proceedings of the National Academy of Sciences, 2022, 119(31): e2104906119.

[15] Singh A K, Gupta R. A privacy-preserving model based on differential approach for sensitive data in cloud environment[J]. Multimedia Tools and Applications, 2022, 81(23): 33127-33150.

[16] Al-Hawawreh M, Hossain M S. A privacy-aware framework for detecting cyber attacks on internet of medical things systems using data fusion and quantum deep learning[J]. Information Fusion, 2023, 99: 101889.

[17] Chukwunweike J N, Yussuf M, Okusi O, et al. The role of deep learning in ensuring privacy integrity and security: Applications in AI-driven cybersecurity solutions[J]. World Journal of Advanced Research and Reviews, 2024, 23(2): 2550.

[18] Biswas S, Jung K, Palamidessi C. Tight differential privacy guarantees for the shuffle model with k-randomized response[C]//International Symposium on Foundations and Practice of Security. Cham: Springer Nature Switzerland, 2023: 440-458.

[19] Feretzakis G, Papaspyridis K, Gkoulalas-Divanis A, et al. Privacy-preserving techniques in generative AI and large language models: A narrative review[J]. Information, 2024, 15(11): 697.

[20] Pasham S D. Privacy-preserving data sharing in big data analytics: A distributed computing approach[J]. The Metascience, 2023, 1(1): 149-184.

[21] Akter M, Moustafa N, Lynar T, et al. Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(12): 5805-5816.

[22] Rodríguez E, Otero B, Canal R. A survey of machine and deep learning methods for privacy protection in the internet of things[J]. Sensors, 2023, 23(3): 1252.

[23] Kobayashi M, Fujioka A, Chida K, et al. km-anonymization Meets Differential Privacy under Sampling[J]. Journal of Information Processing, 2025, 33: 646-656.

[24] Yanamala A K Y, Suryadevara S, Kalli V D R. Balancing innovation and privacy: The intersection of data protection and artificial intelligence[J]. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 2024, 15(1): 1-

43.

[25] Ikwuanusi U F, Adepoju P A, Odionu C S. Advancing ethical AI practices to solve data privacy issues in library systems[J]. International journal of multidisciplinary research updates, 2023, 6(1): 033-044.

[26] Adnan M, Kalra S, Cresswell J C, et al. Federated learning and differential privacy for medical image analysis[J]. Scientific reports, 2022, 12(1): 1953.

[27] Yadav N, Pandey S, Gupta A, et al. Data privacy in healthcare: in the era of artificial intelligence[J]. Indian dermatology online journal, 2023, 14(6): 788-792.

[28] Abaoud M, Almuqrin M A, Khan M F. Advancing federated learning through novel mechanism for privacy preservation in healthcare applications[J]. Ieee Access, 2023, 11: 83562-83579.

[29] Jadon A, Kumar S. Leveraging generative AI models for synthetic data generation in healthcare: Balancing research and privacy[C]//2023 International Conference on Smart Applications, Communications and Networking (SmartNets). IEEE, 2023: 1-4.

[30] Gstrein O J, Beaulieu A. How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches[J]. Philosophy & Technology, 2022, 35(1): 3.