Hindawi
Security and Communication Networks
Volume 2022, Article ID 9075470, 15 pages
https://doi.org/10.1155/2022/9075470

Research Article

Support Personalized Weighted Local Differential Privacy Skyline Query

Guopeng Zhang,[1,2,3] **Xuebin Chen**,[1,2,3] **Yuanli Jia**,[1] **and Ran Zhai**[1,2,3]

[1]*North China University of Science and Technology, 063210 Tangshan, Hebei, China*
[2]*Hebei Key Laboratory of Data Science and Application, 063210 Tangshan, Hebei, China*
[3]*Tangshan Key Laboratory of Data Science, 063210 Tangshan, Hebei, China*

Correspondence should be addressed to Xuebin Chen; chxb@qq.com

Received 6 December 2021; Revised 11 May 2022; Accepted 6 June 2022; Published 14 September 2022

Academic Editor: Mahmood Niazi

The potential privacy risks in certain situations are of concern because of the frequent sharing of data during skyline queries, leading to leakage of users' private information. The most common privacy-preserving technique is to anonymize data by removing or changing certain information, for which an attack with specific background knowledge would render the privacy protection ineffective. To overcome these difficulties, this study proposes a personalized weighted local differential privacy method (PWLDP) to protect data privacy during skyline querying. Compared with existing studies of skyline queries under privacy protection, the degree of privacy protection can be quantitatively analyzed, and the processing of data privacy lies with the user, who quantitatively perturbs the processing according to the sensitivity of the weights of different attributes to avoid substantial information loss. The performance of the proposed PWLDP is verified by comparing PWLDP and LDP on different datasets, the average privacy leakage reduction of 62.22% and 51.67% is obtained for experiments conducted on different datasets relative to the iDP-SC algorithm, and the experimental results demonstrate the efficiency and advantages of the proposed method.

1. Introduction

In the era of IoT, the amount of data generated is growing geometrically, and data have become an essential strategic resource. However, with the massive use of data, the privacy of data owners has become the most significant issue, and with the increasing number of data privacy protection regulations, for example, with the introduction of the EU's "general data protection regulation (GDPR) [1]" in 2016, companies are no longer able to exchange or share data containing user privacy as freely as before, even if they subjectively wish to do so; China's "Data Security Law" clearly states that data processing, including the collection, storage, use, processing, and transmission of data, the departments dealing with data need to take the necessary measures to ensure that the data are in a state of adequate protection and legitimate use; and privacy protection has become an urgent issue at present.

How to dig out the data of interest to users from the huge amount of data and make decisions that satisfy their preferences has led many researchers to pay attention to skyline queries [2]. Skyline queries refer to the discovery of the set of all tuples from a dataset that is not dominated by any other tuples. Liu et al. [3] proposed a new structure, namely, the skyline diagram, to implement precomputation for skyline queries. The skyline diagram consists of skyline regions, called skyline polyominoes, each corresponding to the same set of skyline results, and the final result shows that the proposed algorithm is effective and extensible for both the exact skyline diagram and the approximate skyline diagram. Saad et al. [4] proposed the SkyQUD algorithm to answer skyline queries for data with uncertain dimensions, but many organizational databases may contain various sensitive data, such as personal case data or financial data, the disclosure of which can seriously violate individual privacy and may lead to significant reputational damage and PR

crises for these organizations, and the increasing frequency of data sharing and querying will result in skyline queries that can reveal private user information. Differential privacy [5] (DP), as the mainstream privacy protection technique today, requires particular attack assumptions and background knowledge for K-anonymity, L-diversity, and T-closeness anonymization [6] approaches. In contrast, differential privacy can resist various forms of attacks under the condition that the attacker has maximum background knowledge, and thus, it is robust and reliable. Considering the privacy treatment of local differential privacy [7] (LDP), it has been used and performed well in practice, e.g., Apple [8], Microsoft [9], and other companies have accomplished specific user behavior information statistics with the help of this technique, so skyline queries are performed under local differential privacy protection, but this technique is limited to providing the same level of privacy protection for all individuals. However, not all attributes of users need the same level of privacy, and personalized privacy protection needs to be implemented to avoid providing too much privacy protection for those attributes that do not need too high a privacy level. In this study, a skyline query with personalized weights under local differential privacy is proposed, where each organization assigns subjective and objective weights to the dataset attributes, and both weights are used to calculate new weights, where each data owner has their privacy requirements for each dimension of the data, and the newly calculated weights represent the privacy budgets of the data owner for different dimensions; then, the data owner's data in different dimensions will be subject to different privacy budgets. The data owner does not have to report its privacy allocation to the server, i.e., the server retains only the data that the data owner perturbs to the process. The specific contributions are summarized as follows:

(1) We propose a new privacy method called personalized weighted local differential privacy (PWLDP), which primarily protects users' privacy and allows data owners to add noise to attributes according to different privacy needs, realizing personalized privacy protection.

(2) We propose a skyline query based on personalized weighted local differential privacy, under which each organization performs skyline queries locally and then perturbs them and sends them to the data publisher for integration and final skyline query statistics. The degree of privacy protection can be quantitatively analyzed to make up for the shortcomings of existing studies.

(3) Experiments on different datasets validate that the PWLDP algorithm with a personalized privacy policy has higher privacy and accuracy of its skyline query results compared to the LDP algorithm, and the PWLDP algorithm yields more minor errors in the results compared to the LDP algorithm.

2. Related Work

Skyline query research is divided into two areas: optimization of skyline query algorithm and application of skyline query algorithm to related research areas [10]. Skyline query algorithm mainly includes block nested-loop algorithm, nearest-neighbor algorithm, branch-and-bound algorithm, etc. Yang et al. [11] extended for skyline single-point queries and proposed a top-k group skyline query method based on skyline layers to optimize and speed up by pruning the points on high skyline layers. It is verified that when the k value is small, the skyline group can be quickly found, and the query efficiency is improved. The effectiveness of naive skyline query decreases overall due to the expansion of data volume or the rise in dimensionality, which will lead to the increase in cost for comparison between data, so Choi et al. [12] proposed HI-Sky. This method can perform fast skyline computation by using hash indexes, which exploits the fact that grid location address (GLAD) has column priority ordering and data spatial location information, and HI-Sky can efficiently manage data by hash indexing and can be good at cleaning up unnecessary comparisons during the comparison process.

At this stage, the research work on data privacy handling in the skyline query process is mainly focused on homomorphic encryption techniques [13, 14], but the encrypted keys are brute force broke, which still leaks privacy during skyline queries, and the level of privacy protection cannot be quantitatively analyzed [15]. Zaman et al. [16] proposed a new method for computing skyline in a multiparty computing environment in the MapReduceHadoop framework without revealing an object's value to another party, demonstrating the validity and extensibility of the proposed secure skyline computation. Liu et al. [17] proposed a new framework, PUSC, which introduces a user-defined vector-dominated secure protocol that compares the vector-dominated relationship between two cryptographic vectors based on the user's preferences, which is not efficient enough due to the complexity of the different protocols and the complexity of the computational process, which takes a lot of time to execute. Hua et al. [18] proposed an efficient and privacy-preserving online medical primary diagnosis (CINEMA) framework. Within the CINEMA framework, maintaining the privacy of users' dynamic skyline queries is considered, and users can accurately access online medical primary diagnosis services without revealing medical data. Qaosar et al. [19] proposed to compute skyline in a secure multiparty computing environment using the Paillier cryptosystem [20] to transform object attribute values without changing the order of objects on each attribute; each participant collaborates with the other participants to securely prepare the encrypted order of objects on each feature. The skyline is then computed based on the order of the object attribute values on each dimension, without obtaining the original attribute values of the objects. Qaosar et al. [21] proposed a new privacy-preserving multiparty skyline query framework that utilizes additive homomorphic encryption

and data anonymization, perturbation, and randomization techniques that do not reveal data to others during multiparty skyline queries.

Differential privacy (DP) has been widely used in data publishing as a privacy-preserving method with solid privacy guarantees [22]. Differential privacy can be achieved by using privacy metrics and utility metrics to achieve a trade-off between privacy and utility. However, third-party servers are assumed to be trusted in the DP model. The local differential privacy model sets a stricter notion of privacy, and the application of LDP to data is of increasing concern. Sun et al. [23] applied random response techniques to the frequent itemset mining for personalized privacy requirements. Ouyang et al. [24] introduced a set-valued data collection approach (SetLDP) based on a category hierarchy under a local differential privacy model, whose central concept is to first to provide a random response to the presence of a category, and the results show that it can well protect the privacy information in the set-valued data. Lan et al. [25] proposed personalized differential privacy (iDP-SC) based on a spectral clustering algorithm to reduce the local sensitivity by the introduction of the spectral clustering algorithm, and the noise reduction generated by spectral clustering compensates for the information distortion error introduced by itself. Xiong et al. [26] proposed a new (ε, δ)-LDP concept for capturing users' privacy needs by accounting for the temporal relevance of spatiotemporal data at the same time as guaranteeing sensible utility, demonstrating its superiority in achieving a better trade-off between privacy and utility for real-time spatiotemporal data integration and rigorous privacy protection.

3. Basic Theory

3.1. Local Differential Privacy. If the output of any input after random encoding M as the same result is similar, then the observer cannot infer the original data from the production, so the privacy of the data is protected, and based on such an idea, the definition of local differential privacy is described as follows.

Definition 1 (ε-LDP [7]). For a given privacy budget $\varepsilon \in R^+$, if the random perturbation mechanism M satisfying ε-LDP, when and only when any inputs $x, x\prime$, and output $y \in \text{Range}(M)$ satisfy the following conditions:

$$\Pr(M(x) = y) \leq \exp^{\varepsilon} * \Pr(M(x') = y). \tag{1}$$

ε in equation (1) is a parameter that controls the strength of privacy protection, and the closer the value is to 0, the higher the degree of privacy protection of the algorithm M.

Differential privacy has essential properties such as sequential composition and parallel composition, and the compositional nature of differential privacy can help designers partition the privacy budget ε.

Property 1. Sequence combinatoriality [27]. Suppose that given a dataset U with n random response algorithms M such that $M_i = \{M_1, M_2, \ldots, M_n\}$, algorithm M_i satisfies local differential privacy, and then, the sequence consisting of $nM_i(U)$ algorithms also satisfies $\sum \varepsilon_i$ -local differential privacy.

Property 2. Parallel combinatoriality [27]. Suppose that given a dataset $U = \{U_1 \cup U_2 \cup \cdots \cup U_n\}$, where the datasets $U_i (1 \leq i \leq n)$ and $U_j (1 \leq j \leq n, i \neq j)$ are disjoint subsets, algorithm M_i satisfies ε_i -local differential privacy, and then, the sequence consisting of $n M_i(U_i)$ algorithms also satisfies $\max\{\varepsilon_i\}$ -local differential privacy.

3.2. Skyline Query Calculation. With the booming development of internet technology, a large amount of data is generated all the time. Skyline query can effectively analyze and find accurate results from this high-value, colossal amount of data, which can make appropriate decisions according to user needs. It plays an important role in multitarget decision-making, data mining, and other needed fields.

Definition 2 (Skyline calculation [2]). For a d-dimensional dataset U, i.e., $U = \{X_1, X_2, \ldots, X_n\}$, the attribute of the dataset U is denoted by $x_{ij} (1 \leq i \leq n, 1 \leq j \leq d)$, and suppose any two data records X_a and X_b are said to be dominated by X_a if they satisfy the following condition, then $X_a \prec X_b$ is denoted.

(1) $\forall X_{ij} \in X_i, X_{aj} \leq X_{bj}$;

(2) $\exists X_{ij} \in X_i, X_{aj} < X_{bj}$;

For any set of data records X_a, X_b, if $X_a \prec X_b$, and $X_b \prec X_a$, the data records X_a and X_b are said not to have any dominance relationship.

Definition 3 (Skyline points). For a d-dimensional dataset U, if there is a data record X_i and $X_i \in U$, for any one data record, $X_i' \in U$ other than X_i, there is no $X_i' \prec X_i$, then the data record X_i is the Skyline point of the dataset U.

Definition 4 (Skyline query), it refers to the process of filtering a set of data records that are relatively better in all attribute dimensions from a known dataset, and the result of the above query process is the skyline result set. In addition, any data record in the result set is a skyline point, and the result set is denoted as SKY, which is formally represented as follows.

$$\text{SKY} = \{X_a \in U | \forall X_i \in U, X_i \nprec X_a\}, \tag{2}$$

Definition 5 (Skyline query additivity [28]). We assume n dataset $U = \{U_1, U_2, \ldots, U_n\}$, if the dataset satisfies $U = \{U_1 \cup U_2 \cup \cdots \cup U_n\}$, then

$$\begin{aligned} \text{SKY}(U) &= \text{SKY}(U_1 \cup U_2 \cup \cdots \cup U_n) \\ &= \text{SKY}(\text{SKY}(U_1) \cup \cdots \cup \text{SKY}(U_n)). \end{aligned} \tag{3}$$

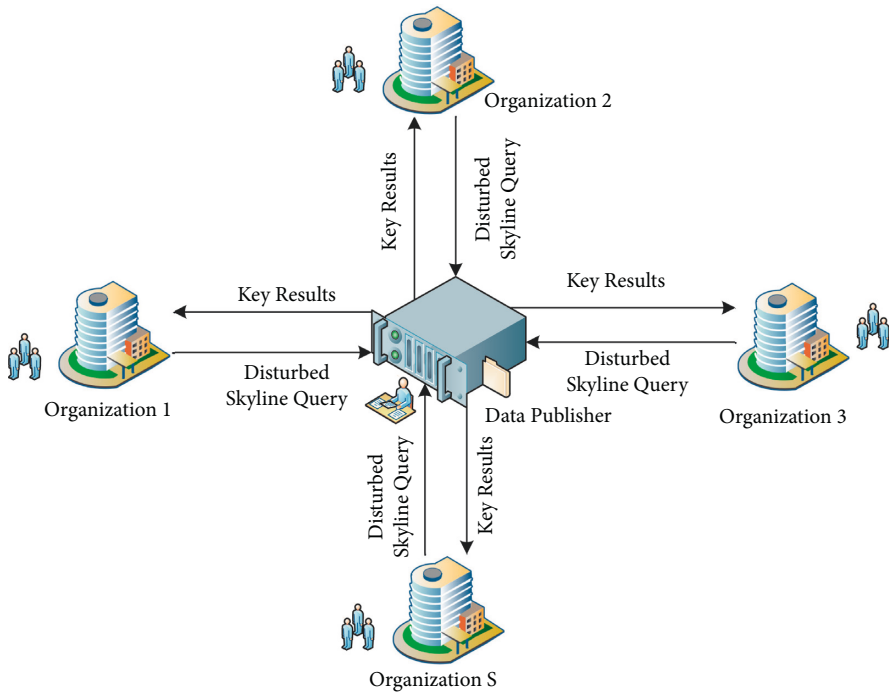

Figure 1: Skyline query scenario between different organizations and publishers.

In skyline query, local skyline query (LSQ) represents the set of nonexisting dominance relation objects in U_i, i.e., SKY(U_i). The global skyline query (GSQ) represents the set of objects in U that do not exist in the dominance relation, i.e., SKY(U).

4. Privacy-Protected Skyline Queries

4.1. Problem Description. The scenario is shown in Figure 1, with S organizations, a central server, i.e., a data publisher. The organization's data usually affect the publisher's skyline query of the data. Given that the data publisher shares data with multiple organizations, the publisher needs to comprehensively evaluate the organization's data situation from multiple perspectives so that the organization can better assess the situation of its data and develop corresponding measures, and the data publisher assumes the role of the central server in the scenario. Let $U \in R^{n\times d}$ be the dataset of all institutions, we divide the dataset U horizontally into datasets $U_1, U_2, \ldots, U_S$, where the $i-$th institution has the dataset $U_i = \{X_{i1}, X_{i2}, \ldots, X_{in_i}\} \in R^{n_i\times d}$, where d is the number of data dimensions. The number of data dimensions is the same for each organization, n_i is the amount of data owned by the $i-$th organization, and $n = \sum_{i=1}^{s} n_i$ is the total number of data. Organizations do not want to send their local data to the central server, which would cause the server to leak the data. To solve the problem, each organization can perform a local skyline query on its data by preference relation to get the LSQ. Then, it will only send the perturbed LSQ to the data publisher. This process avoids unnecessary data wastage and data leakage and also avoids excessive leakage of raw data to the server.

In this scenario, it is set up that there is no collusion among organizations and between organizations and publishers, i.e., each organization and publisher are honest and curious, and the calculations are performed strictly according to the regulations.

4.2. Personalized Weighted Local Differential Privacy

4.2.1. Analytic Hierarchy Process. The analytic hierarchy process (AHP) is a widely used and effective method to determine the weights. It is a method that simulates the way of thinking of people's decision-making process and mathematizes the thinking process of decision-making by using less quantitative information based on in-depth research on the nature of complex decision-making problems, influencing factors, and their internal relationships Table 2 .

(1) Constructing judgment matrix

Each attribute of user X_i is selected as the index, and the judgment matrix Q is established by pairwise comparison between attribute indexes. The comparison influence degree assignment between attributes needs the help of "scale," as shown in Table 1.

$$Q = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & \cdots & \cdots & a_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ a_{d1} & a_{d2} & \cdots & a_{dd} \end{bmatrix}. \tag{4}$$

(2) Consistency check

First, the eigenvector $\omega = (\omega_1, \omega_2, \ldots, \omega_d)$ of the maximum eigenvalue $\lambda_{\max}$ of the judgment matrix Q is calculated, and the vector ω is normalized to obtain ω'. The sum of the elements in the vector is 1. We

Table 1: Scale.

Scale	Implication
1	This indicates that two attributes have the same importance compared to each other
3	This indicates that one attribute is slightly more important than the other when compared to the other attribute
5	This indicates that one attribute is significantly more important than the other when compared to the two attributes
7	This indicates that one attribute is extremely more important than the other when compared to the other attribute
9	This indicates that one attribute is strongly more important than the other when compared to the two attributes
2, 4, 6, 8	The median of the above two adjacent judgments
Reciprocal	Judgment $a_{ij} = 1/a_{ij}$ or comparing attribute j with i
a_{ij}	Judgment $a_{ij}, a_{ij} > 0\,(1 \le i, j \le d)$ for comparison of attributes i and j

calculate the consistency index CI and consistency ratio CR.

$$\mathrm{CI} = \frac{\lambda_{\max} - d}{d - 1}, \tag{5}$$

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}}. \tag{6}$$

The d in equation (5) denotes the order of the judgment matrix Q. The RI value in equation (6) is known from Table 2.

When $\mathrm{CR} < 0.1$, it means that the consistency test is passed and the judgment matrix is reasonably constructed, and the larger the CI is, the more serious the degree of inconsistency of the judgment matrix is. The judgment matrix is modified and adjusted until $\mathrm{CR} < 0.1$, so that it has a good consistency.

(3) Determining weights

The judgment matrix Q passes the consistency test. The elements of vector ω' generated after the normalization process are the weights of each attribute.

Table 2: Random consistency index RI values.

n	RI values
1	0
2	0
3	0.58
4	0.90
5	1.12
6	1.24
7	1.32
8	1.41
9	1.45
10	1.49

4.2.2. Entropy Weight Method. The entropy weight method is an objective method of attribute weight assignment. By calculating the information entropy of each attribute, the weight of the attribute is determined according to the impact of the degree of variation of the attribute on the dataset as a whole, and the attribute with a higher degree of variation receives a more considerable weight so that the attribute receives a more objective attribute weight.

(1) Data processing

Suppose the dataset has d attributes, $U = \{X_1, X_2, \ldots, X_n\}$, where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}\}$, the value x_{ij}' after processing the attributes of the dataset.

$$x_{ij}' = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}. \tag{7}$$

where x_{ij} is the $j-$th attribute of the $i-$th user, $x_{\max}$ is the maximum value of the $j-$th attribute, and $x_{\min}$ is the minimum value of the $j-$th attribute.

(2) Calculating the information entropy of an attribute

$$e_j = -\frac{1}{\ln n} \sum_{i}^{n} \rho_{ij} * \ln \rho_{ij}. \tag{8}$$

In equation (8), $\rho_{ij} = x_{ij}' / \sum_i^n x_{ij}'$, $0 \le \rho_{ij} \le 1$.

(3) Determining weights

$$\omega'' = \frac{1 - e_j}{\sum_{j=1}^{d} 1 - e_j}. \tag{9}$$

4.2.3. Personalized Privacy Budget Allocation. The analytic hierarchy process has an advantage over the entropy weight method in determining the weights according to the decision-maker's wishes but is less objective and more subjective. The entropy weight method has objective advantages, but it cannot reflect the importance of decision-makers to different attributes, and there will be a certain weight and degree opposite to the actual attribute. Given the strengths and weaknesses of the two alternative weighting methods, we hope to control the subjective randomness within a certain range and achieve a neutral weighting between subjective and objective. Therefore, when assigning weights to each attribute, the inherent statistical laws and authoritative value among the feature data should be considered. To make up for the shortcomings of a single method, a reasonable attribute assignment method is proposed, i.e., a combined assignment approach combining the hierarchical analytic hierarchy process and entropy weight method. The weight $\widetilde{\omega}$ of the attribute of each data record is as follows.

$$\widetilde{\omega} = \eta * \omega' + (1 - \eta)\omega''. \tag{10}$$

The η in equation (10) is the weighting factor, which is used to balance the parameters of the two weights, and the value of η is temporarily taken as 0.5 in this study.

Therefore, we propose a new LDP concept called personalized weighted local differential privacy (PWLDP), in which the user sets the privacy budget for different features based on attribute weights as described in Algorithm 1.

4.2.4. Comparison of PWLDP and LDP. Most of the traditional LDP approaches are to assign an even privacy budget to all attributes of the data owner; however, not all attributes of the user need the same privacy level, for example, when comparing the user's age and identity ID, the ID is more sensitive than the age, so the two should not have the same degree of noise addition, so PWLDP can avoid providing too much privacy protection to those attributes that do not need too high a privacy level. In a multidimensional data scenario, PWLDP and LDP privacy budget allocation are related as follows: $\varepsilon_{\text{All}} = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_d = d * \varepsilon_{\text{Avg}}$. Thus, when PWLDP provides an averaging privacy budget for the attributes of user-owners, then it becomes LDP, so PWLDP is a generalization of LDP, and the relation between the two is stated by the following theorem.

Theorem 1. *For any d-dimensional user record X_a, X_b, if a perturbation mechanism M satisfies $\varepsilon - LDP$, then it also satisfies $\varepsilon_{All} - PWLDP$, where $\varepsilon = \varepsilon_{All}$.*

4.3. Skyline Query Based on PWLDP

4.3.1. The Perturbation Mechanism of PWLDP

Definition 6. (ε-PWLDP), given a privacy budget $\varepsilon \in R^+$, the data owner assigns different privacy budgets ε_j to different features according to the feature weights, i.e., the privacy budget ε_j for each attribute is $\widetilde{\omega}_{mm} * \varepsilon (1 \le m \le d)$, and a random perturbation mechanism M satisfies ε -PWLDP when and only when any input $x, x\prime$ and output $y \in \text{Range}(M)$ satisfy the following.

$$\frac{\Pr(M(x) \in y)}{\Pr(M(x\prime) \in y)} \le \exp^{\widetilde{\omega}_{11} * \varepsilon + \widetilde{\omega}_{22} * \varepsilon * \ldots * \widetilde{\omega}_{ij} * \varepsilon} = \exp^{\varepsilon}. \quad (11)$$

The RR mechanism [29] is the mainstream perturbation mechanism of LDP, the main idea is to give stochastic answers to the private data, to overcome the problem that this method is for binary variables, and later, researchers proposed a more generalized definition in the form of K-RR [30, 31]. For any input $x \in R$, the output $x\prime \in R$ of its response is given in the following way as shown in equation.

$$pr[kRR(x) = x\prime] = \begin{cases} \dfrac{e^{\varepsilon}}{e^{\varepsilon} + k - 1} & \text{if}, x' = x, \\ \dfrac{1}{e^{\varepsilon} + k - 1} & \text{if}, x' \ne x. \end{cases} \quad (12)$$

In equation (12) $k(k > 2)$ is the number of attributes containing different candidate values, Pr is the probability in different perturbation cases, the probability response of $\exp^{\varepsilon}/\exp^{\varepsilon} + k - 1$ is the true value, and the probability response of $1/\exp^{\varepsilon} + k - 1$ is any one of k-1 outcomes except the true value, and algorithm 2 is as follows.

4.3.2. Skyline Query Flow Based on PWLDP. The skyline query returns the object that cannot be dominated by any other object given the dominance relationship in the dataset. Suppose a two-dimensional dataset with a total of 16 points and the dominance relationship as the value of each dimension is the smallest, as shown in Figure 2 the skyline point $\{a, b, e, h, k, m\}$ is returned.

In a scenario where organizations and data publishers share data with each other, the data publisher acts as the central server, and each organization is responsible for providing the data. The publisher is not trusted by the organization because the publisher is likely to disclose the privacy of users in each organization. Therefore, each organization first performs a subset of skyline queries internally based on the preference relationships published by the data publisher, and each organization is responsible for uploading the scrambled skyline query results, avoiding unnecessary data leakage. As can be seen from Definition 5, the GSQ of the data publisher is based on the original data query of all organizations, which is the same as the skyline result that is computed based on the LSQ of each organization. Therefore, each organization can use the original data to obtain the LSQ locally against the skyline query and later perform the perturbation process, and it is only responsible for sending the perturbed LSQ to the data publisher. This process does not affect the accuracy of skyline queries and prevents organizations from leaking large amounts of original data to publishers. The data publisher shares the key results with each organization after the global skyline query, and each organization only knows its own key important data and the overall percentage of data to make the corresponding measure plan, and cannot get the details of other organizations from the data, and the sharing of data in this process does not disclose user privacy.

From algorithm 3 and Figure 3, it can be seen that PWLDP first provides local privacy processing for users to perform their own privacy perturbation processing of data, so that privacy is guaranteed, and then provides the allocation of a personalized privacy budget, considering that data owners have personal privacy requirements for each dimension of the data, providing different levels of noise addition to achieve personalized privacy protection.

4.3.3. Privacy Analysis. For the privacy of PWLDP, we prove that it satisfies ε -LDP. Then, by extending this proof to ε-PWLDP, we can prove that the probability that the query result is indistinguishable on any two data records is not greater than $\exp^{\varepsilon}$.

Input: $U = \{X_1, X_2, \ldots, X_n\}$: each data record $X_i (1 \le i \le n)$ has d attributes; ε: denotes the overall privacy budget.
Output: $\varepsilon' = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d\}$: the privacy budget allocated to each attribute $x_j (1 \le j \le d)$.

(1) Construct the judgment matrix Q of the data set according to equation (4)
(2) $\lambda_{\max} \leftarrow Q$
(3) According to equation (5), we obtain the consistency index
(4) According to equation (6), we calculate CR
(5) **if** CR ≥ 0.1 **then**
(6) Denotes a failure of the consistency checking
(7) **else**
(8) By passing the consistency check, the feature vector ω' is calculated based on $\lambda_{\max}$
(9) **end if**
(10) $\omega' = \{\omega_1, \omega_2, \ldots, \omega_d\}$: normalization
(11) Attributes of the standardized data record $X_i, x_{ij}' = x_{ij} - x_{\min}/x_{\max} - x_{\min}$
(12) $\rho_{ij} = x_{ij}'/\sum_i^n x_{ij}', 0 \le \rho_{ij} \le 1$
(13) $e_j = -(1/\ln n)\sum_i^n \rho_{ij} * \ln \rho_{ij}$
(14) $\omega\prime\prime = 1 - e_j/\sum_{j=1}^d 1 - e_j, \omega\prime\prime = \{\omega_1, \omega_2, \ldots, \omega_d\}$
(15) According to equation (10), we can get $\widetilde{\omega}$
(16) $\varepsilon\prime = \varepsilon * \widetilde{\omega}_j$
(17) **return** $\varepsilon' = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d\}$;

Algorithm 1: Personalized privacy budget allocation.

Input: $x_{ij} (1 \le i \le n, 1 \le j \le d)$: attribute value; ε: privacy budget; $k_j (1 \le j \le d)$: the maximum value in the value range of each attribute.
Output: x_{ij}': property values after perturbation processing

(1) **if** $pr_1 = 1/(np.e^{\varepsilon} + k_j - 1)$ **then**
(2) $x_{ij}' = \left\{x_{i\prime j\prime} | i\prime \ne i, j\prime \ne j\right\}$
(3) **else** $pr_2 = np \cdot e^{\varepsilon}/(np \cdot e^{\varepsilon} + k_j - 1)$
(4) $x_{ij}' = \left\{x_{ij} | i = i, j = j\right\}$
(5) **end if**
(6) **return** x_{ij}';

Algorithm 2: K-RR numerical perturbation mechanism.

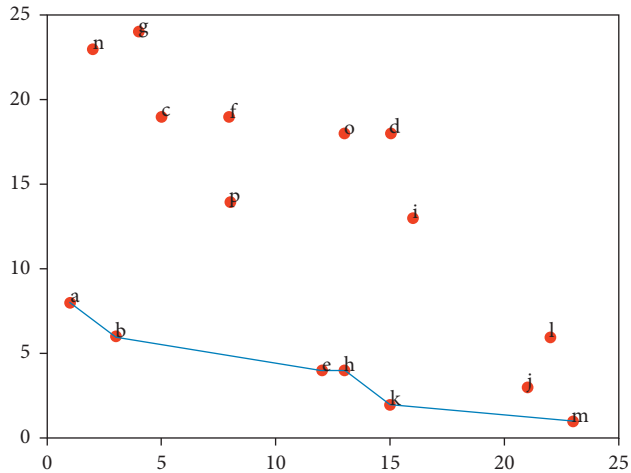

Figure 2: Skyline query results.

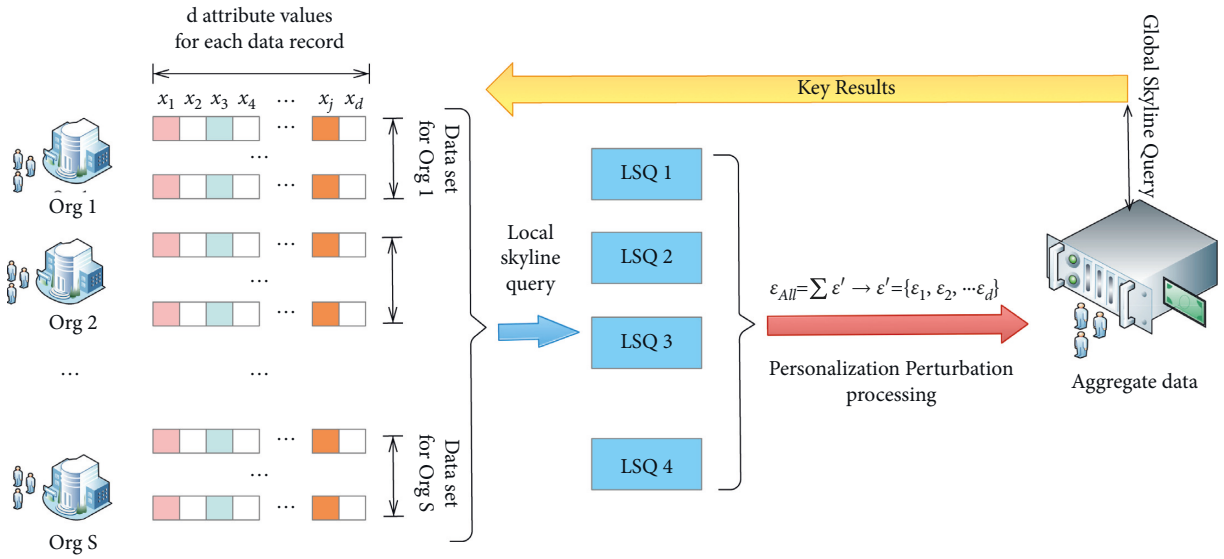

Figure 3: Skyline query process based on PWLDP.

Input: $U = \{LSQ_1, LSQ_2, \ldots, LSQ_S\}$: local skyline query results for each organization; ε: privacy budget; $\tilde{\omega}_j (1 \le j \le d)$: the proportion of weights under each attribute.
Output: U = GSQ: global skyline query results.

(1) **for** local skyline query data set LSQ_S from 1 to S **do**
(2) **for** data record X_i from 1 to n **do**
(3) **for** each attribute x_{ij} from 1 to d **do**
(4) $\varepsilon_j \leftarrow \tilde{\omega}_j * \varepsilon$
(5) $k_j \leftarrow$ Max_*att*
(6) $pr_1 = 1/(np \cdot e^{\varepsilon_j} + k_j - 1)$
(7) $pr_2 = np \cdot e^{\varepsilon_j}/(np \cdot e^{\varepsilon_j} + k_j - 1)$
(8) $pr' =$ np.full(shape = k_j, fill_value = pr_1)
(9) $pr'[x_{ij} - 1] = pr_2$
(10) $x'_{ij} =$ np.random.choice(a = range(1, $k_j + 1$), $p = pr'$
(11) **end for**
(12) **end for**
(13) **end for**
(14) **for** each X_i in List **do**
(15) flag = True
(16) **for** each X_j in List **do**
(17) **if** $X_i \neq X_j$ **then**
(18) **if** Compare (X_i, X_j) **then**
(19) return True or Flase
(20) flag = Flase
(21) **end if**
(22) **end if**
(23) **end for**
(24) **return** X_i. Index
(25) **end for**
(26) GSQ$\leftarrow U' = \{LSQ'_1, LSQ'_2, \ldots, LSQ'_S\}$
(27) **return** GSQ to each organization

Algorithm 3: Skyline query based on PWLDP algorithm.

Theorem 2. *PWLDP algorithm satisfies ε -local differential privacy.*

Proof. Any two different data records $x, x\prime$ and x_i, x'_i represent the attributes of different data records, y denotes the possible output value after random perturbance, and y_i denotes the possible output value after perturbance of different attributes.

According to Definition 1, the PWLDP algorithm satisfies ε-local differential privacy.

Table 3: Different dataset weight parameters.

Different dataset weights	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
Cancer	0.124114385	0.385456285	0.090549575	0.178479455	0.221400305
Mass	0.056889285	0.10520268	0.284250805	0.492551555	0.061105175

$$\begin{aligned}\frac{pr[kRR(x)=y]}{pr[kRR(x')=y]} &= \prod_{i=1}^{d}\frac{pr[kRR(x_i)=y_i]}{pr[kRR(x_i')=y_i]}\\ &\le \prod_{i=1}^{d}\frac{e^{\varepsilon_i}/(k-1+e^{\varepsilon_i})}{1/(k-1+e^{\varepsilon_i})}\\ &= \prod_{i=1}^{d}e^{\varepsilon_i}\\ &= e^{\varepsilon}.\end{aligned} \tag{13}$$

According to Definition 1, the PWLDP algorithm satisfies ε-local differential privacy. □

5. Evaluation

To verify the performance of the PWLDP algorithm and its effectiveness, and we design multiple sets of experiments to test the PWLDP algorithm. We verify the effectiveness of the PWLDP algorithm from two aspects. First, the difference between the original data skyline query results and the perturbed skyline results is measured on different institutional datasets. The mean square error (MSE) is the average of the sum of squares of the difference between the perturbed skyline results and the corresponding real skyline results. The square absolute error (MAE) is the average of the sum of absolute values of the difference between the perturbed skyline results and the corresponding real skyline results. These two measures have been broadly applied to evaluate the utility of noisy results relative to true query results. Finally, after the publisher performed the final skyline query, the PWLDP algorithm was compared with the LDP algorithm results by using the precision rate and F-measure, and the degree of privacy protection of this study's algorithm relative to the iDP-SC [25] algorithm is measured by the metric privacy leakage. To better avoid randomness, the algorithm query performance final metrics were based on the average of 1000 tests run in the same environment.

5.1. Experimental Setup

5.1.1. Experimental Environment

(1) Hardware environment: Intel(R) Core(TM) i5-7200U CPU @ 2.50 GHz Windows 10 PC with 8 GB RAM.

(2) Programming environment: Python3, PyCharm platform, and Jupyter platform.

5.1.2. Dataset Settings. The mammographic mass dataset and the breast cancer dataset were selected on the publicly available UCI dataset to evaluate both algorithms. To better simulate the scenarios in this study, we suppose that each dataset has three organizations. For the breast cancer dataset, it has 32 features, we selected 5 representative attributes by dimensionality reduction, and the dominant rule of skyline query is that if the attribute values are equally smaller in all dimensions, the record takes precedence over another record. The mammographic mass dataset selects five attributes, and again if the attribute values are equally small in all dimensions, the record takes precedence over another record.

5.1.3. Evaluation Metrics

(1) MAE (L1 loss)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \check{Y}_i\right|. \tag{14}$$

(2) MSE (L2 loss)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \check{Y}_i\right)^2. \tag{15}$$

(3) F-measure

$$F = \frac{(\alpha+1)*P*R}{\alpha(P+R)}, \tag{16}$$

where $\alpha = 1$, P is precision, and R is recall.

(4) Amount of privacy leakage (APL) [32].

$$\mathrm{APl} = \mathrm{TPR} - \mathrm{FPR}, \tag{17}$$

where TPR indicates true-positive rate, and FPR indicates false-positive rate.

5.1.4. Parameter Setting

(1) Privacy parameters: the privacy budget ε is set to 0.1, 0.3, 0.5, 0.7, 0.9, and 1.

(2) Weight parameters.

The different attribute weights for different datasets are shown in Table 3.

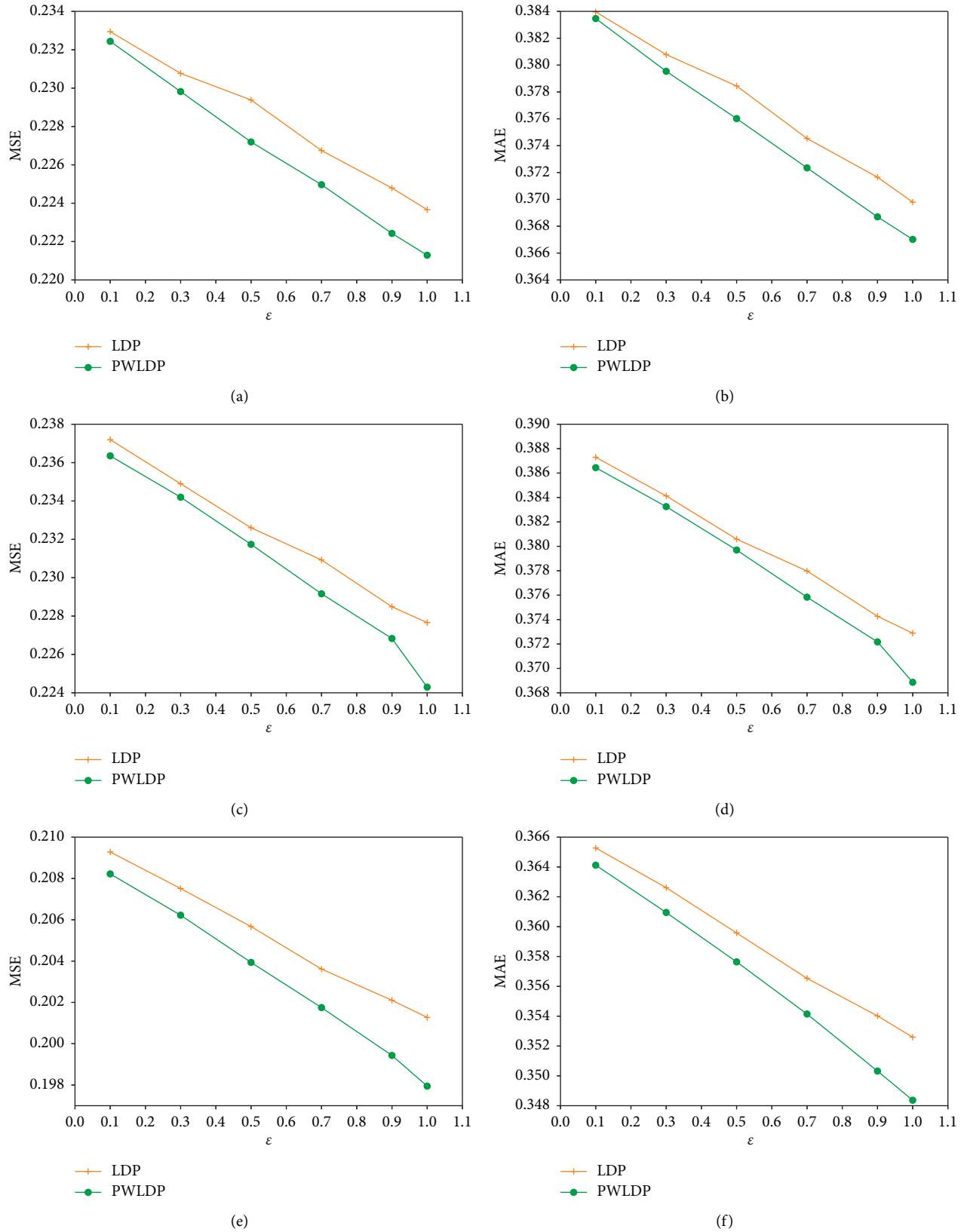

Figure 4: Comparison of MAE and MSE on mammographic mass dataset under different ε indicators. (a) Organization 1. (b) Organization 1. (c) Organization 2. (d) Organization 2. (e) Organization 3. (f) Organization 3.

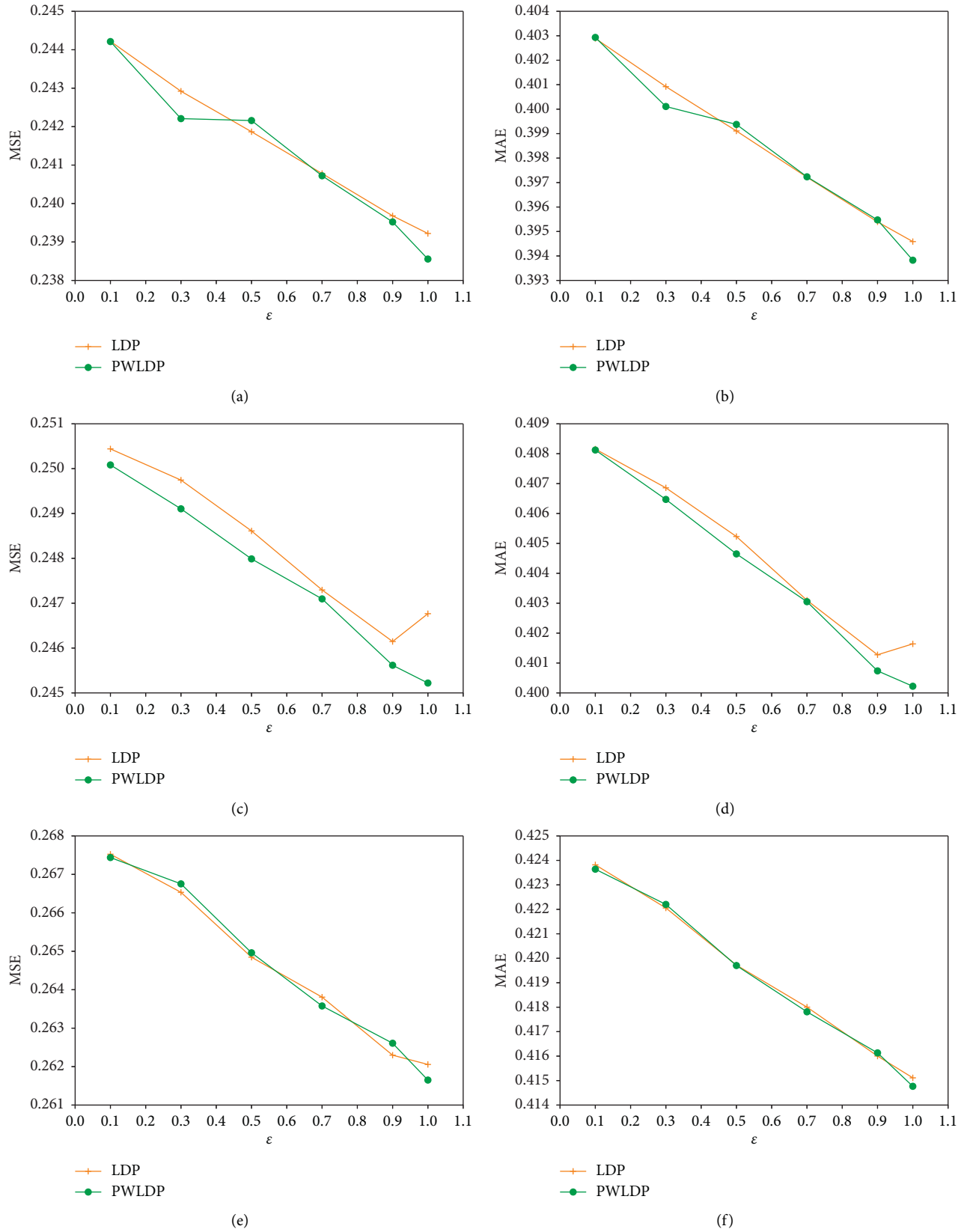

Figure 5: Comparison of MSE and MAE on breast cancer dataset under different ε indicators. (a) Organization 1. (b) Organization 1. (c) Organization 2. (d) Organization 2. (e) Organization 3. (f) Organization 3.

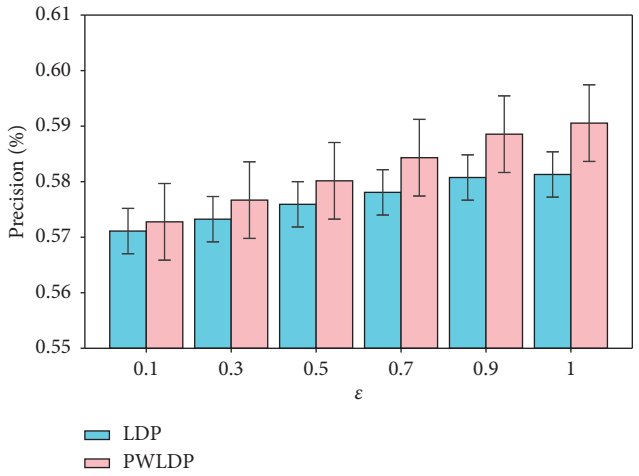

Figure 6: Comparison of precision rates on the mammographic mass dataset at different ε indicators.

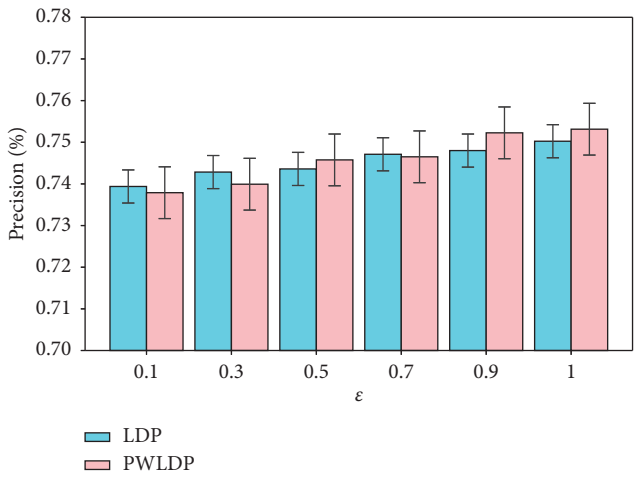

Figure 7: Comparison of precision rates on the breast cancer dataset at different ε indicators.

(3) Comparison algorithms: the PWLDP algorithm, LDP algorithm, and iDP-SC algorithm.

5.2. Experimental Analysis

5.2.1. Impact of Privacy Budget on MSE and MAE. The privacy budget ε is used as a parameter for privacy protection, determining the degree of privacy protection. From Figures 4 and 5, it can be observed that the MSE and MAE values of PWLDP and LDP decrease with increasing ε. From Figures 4 and 5, the MSE and MAE values of the PWLDP algorithm and LDP algorithm are close to each other for $\varepsilon = 0.1$ with the increase of ε, the error values of both are gradually separated, and the most tremendous variability in the results of MSE and MAE values of the two algorithms is observed when $\varepsilon = 1$, it is because when ε is too small, and the probability of the perturbation being other values and the actual value is infinitely close concerning the KRR response mechanism. Under the same privacy budget ε, the MSE and MAE values of PWLDP are smaller than those of the LDP algorithm, which is because the personalized privacy allocation is provided. When the attribute value weight of the PWLDP algorithm is large, the allocated privacy budget is too large, the perturbance intensity decreases, and the privacy protection decreases. When the attribute weight is small, the budget is allocated too small, the perturbance intensity increases, and the protection is enhanced. Because the more extensive the weight is, the more important the value is for users. The PWLDP has higher security and lower privacy budget consumption.

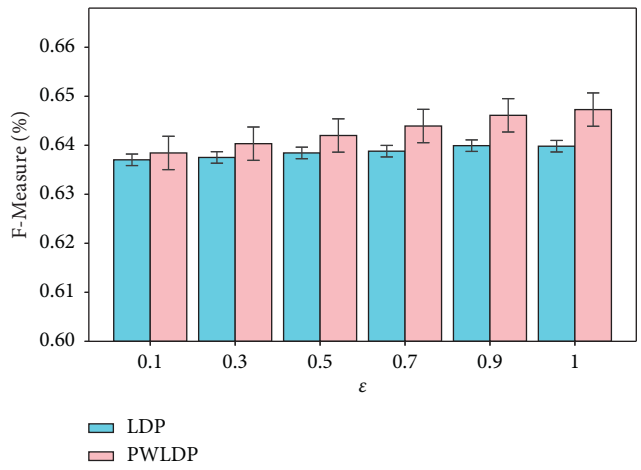

Figure 8: Comparison of F-measure on the mammographic mass dataset at different ε indicators.

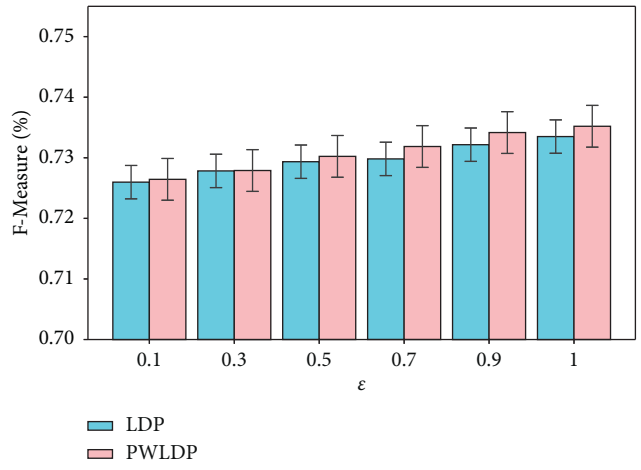

Figure 9: Comparison of F-measure on the breast cancer dataset at different ε indicators.

5.2.2. Impact of Privacy Budget on Precision Rate and F-Measure. The precision rate refers to the percentage of skyline query results that are correct records in the dataset after perturbation, and as observed in Figures 6 and 7, the precision rate gradually increases with the increase in the

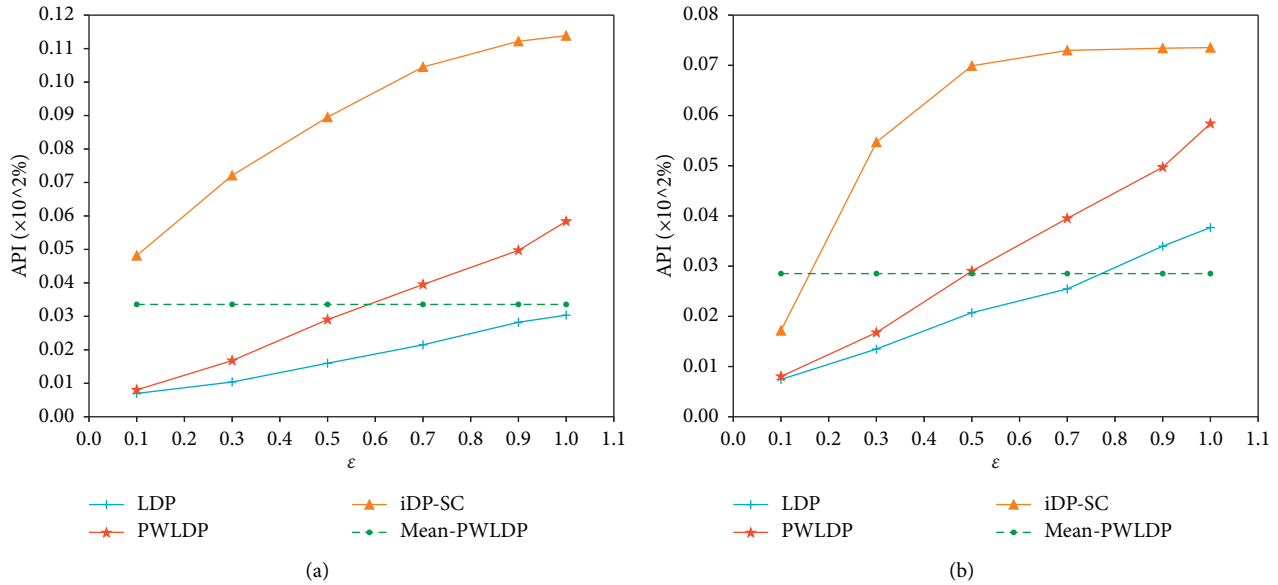

Figure 10: Comparison of privacy leakage (APL) for different datasets with different ε metrics. (a) Mammographic mass dataset. (b) Breast cancer dataset.

privacy budget ε. In Figure 6, it can be seen that for $\varepsilon = 0.1$, the precision rate of PWLDP is 0.573 and that of LDP is 0.571, with an improvement of 0.3%, and for $\varepsilon = 1$, the precision rate of PWLDP is 0.591 and that of LDP is 0.581, with an improvement of 1.7%. The improvement of PWLDP can also be observed in Figure 7. The F-value measures the effectiveness of the experimental method and also indicates the availability of the data, as can be seen from Figures 8 and 9, as ε increases, the PWLDP algorithm F-value also gradually increases and is higher than the LDP algorithm, because the degree of privacy protection decreases while the amount of noise added also decreases, the PWLDP as the privacy budget ε increases, the attribute values with higher weights retain the information of the original data with a higher probability and smaller probability perturbing for other data information, it is possible to make important data require light privacy protection, unimportant data attributes use light privacy budget, which has less impact on data availability and ensures that the availability of data is improved, and therefore, the F-value is also improved. From Figure 9, it can be seen that the F-value gradually approaches 1, and the effectiveness of the algorithm is greatly improved. The F-value in Figure 8 is relatively slow to improve because of the significant difference in the value domains of different attributes in the mammographic mass dataset because the RR method has better performance when the attribute takes a lower value domain.

5.2.3. Impact of Privacy Budget on Privacy Leakage. As can be seen from Figure 10, the smaller the privacy budget ε, the lower the privacy leakage, and the better the privacy protection, and as the privacy budget ε increases, the privacy leakage gradually increases, the iDP-SC algorithm has the most obvious privacy leakage, and the privacy protection becomes worse. There is no significant difference in the privacy leakage amount between the PWLDP and LDP algorithms when the privacy budget ε is set low. In Figures 10(a) and 10(b), the privacy leakage between the two gradually differs when $\varepsilon = 0.3$, the PWLDP algorithm adds noise based on the personalized weights of user attributes, and when the privacy budget ε is small, the difference between this personalized weight allocation budget and the uniform allocation budget is not very obvious, and as the budget ε increases, the budget allocated to certain attributes in PWLDP gradually increases and is larger than the budget uniformly allocated to attributes in LDP, so it causes a small difference between the two when the privacy budget increases. When $\varepsilon = 0.1$, the privacy leakage of the PWLDP in Figure 10(a) is reduced by 83.3% relative to the iDP-SC algorithm, and the privacy leakage of the PWLDP in Figure 10(b) is reduced by 56.6% relative to the iDP-SC algorithm. When $\varepsilon = 1$, the privacy leakage of the PWLDP in Figure 10(a) is reduced by 48.7% relative to the iDP-SC algorithm, and the privacy leakage of the PWLDP in Figure 10(b) is reduced by 36.1% relative to the iDP-SC algorithm. As can be seen from Table 4, the average privacy leakage for PWLDP on the mass dataset is 0.034 and 0.09 for iDP-SC, with an average reduction of 62.22%, and the average privacy leakage for the PWLDP on the cancer dataset is 0.029 and 0.06 for iDP-SC, with an average reduction of

Table 4: Comparison of the average privacy leakage of different algorithms on different datasets.

Dataset	Algorithm		
	Mean-PWLDP	Mean-iDP-SC	Mean-LDP
Mass	0.033569449	0.090055232	0.018901954
Cancer	0.028515301	0.060279844	0.023127942

51.67%. This shows that the PWLDP has higher privacy, lower privacy leakage, and personalized privacy protection as the privacy budget ε increases.

6. Conclusions

In this study, we consider the data privacy of skyline query under the data sharing scenario based on local differential privacy. It is worth exploring how to solve the data leakage problem in skyline query, and most of the current privacy protection methods about skyline query focus on encryption and anonymization, and the privacy is still hidden and cannot be quantitatively analyzed under the premise that data availability is improved. A personality weight assignment for local differential privacy (PWLDP) is proposed, first the privacy treatment of data lies in the hands of users, quantitative privacy protection is provided for different attributes according to individual privacy requirements, and experiments based on real datasets verify the effectiveness of the scheme in this study. The following two aspects are investigated in future work: (1) to study the appropriate weighting parameters to achieve a balanced and efficient personalization weighting; (2) to consider the user's privacy needs and data sensitivity from multiple perspectives so that the personalization mechanism can be more widely expanded and applied.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. U20A20179).

References

[1] P. Voigt and A. V. D. Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide*, Springer International Publishing, vol. 10, Springer International Publishing, Cham, 1st edition, Article ID 3152676, 2017.

[2] S. Borzsony, D. Kossmann, and K. Stocker, "The skyline operator," in *Proceedings of the 17th international conference on data engineering*, pp. 421–430, IEEE, Heidelberg, Germany, April 2001.

[3] J. Liu, J. Yang, L. Xiong et al., "Skyline diagram: efficient space partitioning for skyline queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 271–286, 2019.

[4] N. H. M. Saad, H. Ibrahim, F. Sidi, R. Yaakob, and A. A. Alwan, "Efficient skyline computation on uncertain dimensions," *IEEE Access*, vol. 9, Article ID 96994, 2021.

[5] C. Dwork, "Differential privacy: a survey of results," in *Proceedings of the International conference on theory and applications of models of computation*, pp. 1–19, Springer, Hong Kong, China, May 2013.

[6] A. M. Sazdar, S. A. Ghorashi, V. Moghtadaiee, A. Khonsari, and D. Windridge, "A low-complexity trajectory privacy preservation approach for indoor fingerprinting positioning systems," *Journal of Information Security and Applications*, vol. 53, Article ID 102515, 2020.

[7] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, IEEE, Berkeley, CA, USA, October 2013.

[8] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting Telemetry Data Privately," 2017, https://arxiv.org/abs/1712.01524.

[9] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: local differential privacy in practice," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 1655–1658, Houston, TX, USA, June 2018.

[10] Y. Tang and S. Chen, "Supporting continuous skyline queries in dynamically weighted road networks," *Mathematical Problems in Engineering*, vol. 2018, Article ID 6749650, 14 pages, 2018.

[11] Y. Yang, W. Lu, and C. Tang, "A fast top-k group skyline query method based on skyline layer," in *Proceedings of the 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, pp. 146–151, IEEE, Sanya, China, December 2020.

[12] J.-H. Choi, F. Hao, and A. Nasridinov, "Hi-sky: hash index-based skyline query processing," *Applied Sciences*, vol. 10, no. 5, p. 1708, 2020.

[13] X. Liu, R. Lu, J. Ma, L. Chen, and H. Bao, "Efficient and privacy-preserving skyline computation framework across domains," *Future Generation Computer Systems*, vol. 62, pp. 161–174, 2016.

[14] J. Liu, J. Yang, L. Xiong, and J. Pei, "Secure skyline queries on cloud platform," in *Proceedings of the 2017 IEEE 33rd international conference on data engineering (ICDE)*, pp. 633–644, IEEE, San Diego, CA, USA, April 2017.

[15] P. Saravanakumar, T. Sundararajan, R. Kumar Dhanaraj, K. Nisar, F. Hussain Memon, and A. Ibrahim, "Lamport certificateless signcryption deep neural networks for data aggregation security in wsn," *Intelligent Automation & Soft Computing*, vol. 33, no. 3, pp. 1835–1847, 2022.

[16] A. Zaman, M. A. Siddique, Y. Annisa, and Y. Morimoto, "Secure computation of skyline query in mapreduce," in *Proceedings of the International Conference on Advanced Data Mining and Applications*, pp. 345–360, Springer, Foshan, China, November 2016.

[17] X. Liu, K.-K. R. Choo, R. H. Deng, Y. Yang, and Y. Zhang, "Pusc: privacy-preserving user-centric skyline computation over multiple encrypted domains," in *Proceedings of the 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering*, pp. 958–963, IEEE, NY, USA, August 2018.

[18] J. Hua, H. Zhu, F. Wang et al., "Cinema: efficient and privacy-preserving online medical primary diagnosis with skyline query," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1450–1461, 2019.

[19] M. Qaosar, A. Zaman, M. Siddique, Y. Annisa, and Y. Morimoto, "Privacy-preserving secure computation of skyline query in distributed multi-party databases," *Information*, vol. 10, no. 3, p. 119, 2019.

[20] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proceedings of the International conference on the theory and applications of cryptographic techniques*, pp. 223–238, Springer, Trondheim, Norway, June 2022.

[21] M. Qaosar, K. M. R. Alam, A. Zaman et al., "A framework for privacy-preserving multi-party skyline query based on homomorphic encryption," *IEEE Access*, vol. 7, Article ID 167496, 2019.

[22] E. Antwi-Boasiako, S. Zhou, Y. Liao, Q. Liu, Y. Wang, and K. Owusu-Agyemang, "Privacy preservation in distributed deep learning: a survey on distributed deep learning, privacy preservation techniques used and interesting research directions," *Journal of Information Security and Applications*, vol. 61, Article ID 102949, 2021.

[23] C. Sun, Y. Fu, J. Zhou, and H. Gao, "Personalized privacy-preserving frequent itemset mining using randomized response," *The Scientific World Journal*, vol. 2014, Article ID 686151, 10 pages, 2014.

[24] J. Ouyang, Y. Xiao, Y. Xiao, S. Liu, Z. Xiao, and X. Liao, "Set-valued data collection with local differential privacy based on category hierarchy," *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2733–2763, 2021.

[25] Q. Lan, J. Ma, Z. Yan, and G. Li, "Utility-preserving differentially private skyline query," *Expert Systems with Applications*, vol. 187, Article ID 115871, 2022.

[26] X. Xiong, S. Liu, D. Li, Z. Cai, and X. Niu, "Real-time and private spatio-temporal data aggregation with local differential privacy," *Journal of Information Security and Applications*, vol. 55, Article ID 102633, 2020.

[27] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30, New York, NY, June 2009.

[28] K. Hose and A. Vlachou, "A survey of skyline processing in highly distributed environments," *The VLDB Journal*, vol. 21, no. 3, pp. 359–384, 2012.

[29] N. Holohan, D. J. Leith, and O. Mason, "Optimal differentially private mechanisms for randomised response," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2726–2735, 2017.

[30] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2879–2887, 2014.

[31] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proceedings of the 26th {USENIX} Security Symposium ({USENIX} Security 17)*, pp. 729–745, Vancouver, BC, Canada, August 2017.

[32] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *Proceedings of the 28th USENIX Security Symposium*, pp. 1895–1912, Santa Clara, CA, USA, August 2019.