Research Article

Multidomain Fusion Data Privacy Security Framework

Jing Yang, Lianwei Qu ©, and Yong Wang ©

College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China

Correspondence should be addressed to Yong Wang; wangyongcs@hrbeu.edu.cn

Received 30 July 2021; Accepted 17 November 2021; Published 20 December 2021

Academic Editor: James Ying

Copyright © 2021 Jing Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the collaborative collection of the Internet of Things (IoT) in multidomain, the collected data contains richer background knowledge. However, this poses new requirements for the security of data publishing. Furthermore, traditional statistical methods ignore the attributes sensitivity and the relationship between attributes, which makes multimodal statistics among attributes in multidomain fusion data set based on sensitivity difficult. To solve the above problems, this paper proposes a multidomain fusion data privacy security framework. First, based on attributes recognition, classification, and grading model, determine the attributes sensitivity and relationship between attributes to realize the multimode data statistics. Second, combine them with the different modal histograms to build multimodal histograms. Finally, we propose a privacy protection model to ensure the security of data publishing. The experimental analysis shows that the framework can not only build multimodal histograms of different microdomain attribute sets but also effectively reduce frequency query error.

1. Introduction

The Internet of Things (IoT) is widely used in data collection in various fields [1–3] and integrates them to apply data analysis in different fields. Therefore, in the development of 5G, 6G wireless networks, and IoT [4, 5], a lightweight, reliable, and intelligent security privacy protection framework is extremely important [6, 7]. The multidomain fusion data set collected by IoT contains strong background knowledge, and it leads to the higher risk of privacy leakage when data publishing [8, 9]. The issue of personal privacy leakage causes security risk to many aspects such as personal life, property, and family. Early research focuses on the improvement of the privacy protection model based on K-anonymity [10–14] and I-diversity [15, 16], but these improved models cannot entirely resist the strong knowledge attacks and new attack methods. Differential privacy proposed is effectively solving privacy leakage caused by background knowledge and is widely used in the domain of data publishing. Although the paper [16–20] fully improved the histogram data publishing algorithm based on differential privacy, the data availability after privacy protection is not better when publishing multimodal histograms. The paper [21] proposes a personalized privacy protection model that supports the privacy protection needs of publishing histograms, but it is still challenging to balance data privacy and data availability.

Histogram data publishing based on data records is widely used in the domain of data publishing [22]. For example, in the medical data set, we count the number of people suffering from heart disease in different age groups. When facing the multidomain fusion data set, although the histogram data publishing based on data records can provide good data, it is difficult to realize multimode data statistics due to ignoring the attributes sensitivity and relationship between attributes. Therefore, data publishing based on data records cannot publish the relationship between attributes and attribute sensitivity in the multidomain fusion data set. At present, the research on data attributes recognition, classification, grading, and sensitivity focuses on data mining analysis. The domain of data publishing concerns the research of data privacy protection but ignoring the study of data attributes sensitivity and relationship between attributes.

To solve the two problems proposed above, this paper presents a multidomain fusion statistical data publishing privacy protection framework. For the multimode data statistics, we offer two definitions that are microdomain and microdomain attribute set. It applies to the microdomain
2.1. Data Privacy Security Framework. With the increasingly serious problem of privacy leakage, a friendly data privacy security framework is a prerequisite to ensure data sharing, publishing, and mining. The core of the data privacy security framework is the targeted privacy protection methods.

Xiong et al. [6] proposed a data privacy security framework. First, the P-CNN model is implemented based on two ciphertexts, and then, P-CNN is deployed with two edge servers to collaboratively solve the problem of original data sharing among connected and autonomous vehicle information leakage. Xiong et al. [7] proposed a personalized privacy protection framework based on game theory and data encryption. This framework solves the problem of QoS protection imbalance in MCS caused by the unified privacy policy of sensor data. Wu et al. [24] proposed a PETA framework that not only considers the privacy protection of MCS but also uses powerful edge servers deployed between users and platforms to cluster and manage users based on user attributes. Xiong et al. [12] proposed an ATG framework based on the MCS privacy protection framework, which uses AI technology and game theory to solve the problem of public data privacy leakage in MECS.

At present, the data security frameworks are all based on encryption algorithms. This paper exploits differential privacy algorithms to achieve data publishing privacy protection. The purpose is to improve data availability while ensuring that privacy is not leaked.

2.2. Attributes Recognition, Classification, and Grading. Attributes recognition, classification, and grading are important support for data mining, analysis, sharing, and other applications based on data publishing. In contrast to the domain of data publishing, there are relatively few studies on attributes recognition, classification, and grading [25]. The attribute privacy measurement is the key to classification, grading, and privacy protection. The information entropy is one of the effective methods to calculate the amount of information. Therefore, Diaz et al. [26] and Serjantov and Danezis [27] proposed earlier fusion information entropy and other relevant information theory into attribute measurement.

Peng et al. [28] proposed some information entropy privacy protection models based on the Shannon information theory communication framework and solved the problem of privacy measurement from attribute characteristics, background knowledge, and multiple data sources. However, these models ignore the sensitivity between attributes. Yu et al. [29] used the Shannon information theory to measure privacy data and combined it with the BP neural networks to implement privacy data grading. Still, this computational cost of this model is larger, and the grading result of BP neural network depends on samples. Krishnamurthy and Wills [30] calculate the privacy leakage amount of the social network attributes to determine the scope of attribute privacy leakage and propose a privacy protection method, but this paper does not consider attributes recognition, classification, and grading. He and Pen [25] put forward a sensitive attribute classification and grading algorithm for structured data set. First, calculate the privacy attribute sensitivity by clustering information entropy and association rule. Second, calculate the attribute sensitivity average to implement attributes classification and grading. However, the clustering result based on $K$-means depends on the choice of $K$ value. If the selected $K$ value is not appropriate, it will form a local optimal solution leading to inaccurate classification.

The idea of this paper is inspired by the paper [25]. We implement the microdomain recognition of the data set by the proposed definitions for microdomain and microdomain attribute set. Then, we adopt the information gain and attribute sensitivity to represent attributes classification and grading.

2.3. Data Publishing Privacy Protection. Data publishing privacy protection is aimed at protecting privacy information...
security in the process of data publishing [31]. Though the traditional access control and encryption technology have a better privacy protection effect, the availability of the protected data is insufficient, which defeats the purpose of data publishing. Differential privacy is widely used because it can solve the problem of privacy protection models such as K-anonymity [10–14], L-diversity [15, 16], and T-tight [32], which cannot resist the privacy leakage caused by strong background knowledge.

Dwork [33, 34] proposed a differential privacy method based on Laplace and exponential mechanism in 2006 and 2008. Though this method effectively resists the privacy leakage caused by background knowledge, the error between protected data and original data is large. Dwork et al. [35] proposed an equal-width histogram privacy protection method based on differential privacy that is called LP algorithm, and the performance is better in small noise and small range query. However, when the noise is larger or long-range query, too much noise accumulation leads to poor data availability. To improve the accuracy of long-range query, Xu et al. [36, 37] proposed a difference histogram protection method based on Noise First and Structure First, as well as used the idea of V-optimization histogram to optimize the noise histogram and obtain the high accuracy query results. Whatever, this method cannot balance the noise error and reconstruction error, meanwhile the high computation cost of the postoptimization process. Xiao et al. [38] transformed the original histogram into a binary tree of wavelet coefficient to support long-range query. However, this algorithm is not conducive to the practical application due to the high sensitivity of query. Hay et al. [39] proposed a personalized privacy protection method based on constraint inference for unattributed histogram and universal histogram. The histogram with noise is optimized by constraint inference, but the data accuracy does not have an advantage in low noise and small-range query. Piao et al. [18] proposed the MDHP algorithm to implement privacy protection of governmental data publishing. The MDHP algorithm combined the LP algorithm with the grouping reconstruction algorithm based on the maximum difference scaling to achieve privacy protection which satisfies ε-differential privacy. The MDHP algorithm effectively improved the published data availability for small-range query and low noise. However, the performance of data availability was poor in the large noise or long-range query.

This paper is inspired by the paper [18, 39]. We improve the MDHP algorithm by the order inference and linear estimation. The algorithm in this paper not only satisfies the privacy requirement of the published multimodal histogram but also effectively balances the privacy and availability of the published data.

3. Preliminaries

3.1. Information Entropy

**Definition 1 Information entropy** [25]. The information entropy is the self-information expected value for each discrete event and is denoted by $H(X)$.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i).$$

(1)

$p(x_i)$ is probability when $X$ is equal to $x_i$.

**Definition 2 Maximum entropy** [25]. The maximum entropy is when the probability of each discrete event $x_i$ is equal. The maximum entropy formula is as follows. $n$ is the number of discrete events.

$$H_{\text{max}}(X) = \log n.$$  

(2)

**Definition 3 Conditional entropy** [25]. When we set the condition $X$ and the uncertainty of the random variable $Y$, $p(x, y)$ is joint distribution $n$ probability, and $p(y | x)$ is conditional probability.

$$H_n(Y | X = x) = -\sum_{x \in x, y \in Y} p(x, y) \log p(y | x).$$

(3)

**Definition 4 Information gain.** The information gain is when we give the condition $X$, calculating the reduction in the uncertainty of $Y$. The information gain formula is as follows. $H(Y)$ is information entropy, and $H(Y | X)$ is conditional entropy.

$$IG_i = H(Y) - H(Y | X).$$

(4)

3.2. Differential Privacy

**Definition 5 ε-Differential privacy** [30]. A randomized mechanism $S$ is differentially private for any pair of neighboring data set $I$ and $I'$ and for any set of possible sanitized output $R$.

$$\Pr [S(I) \in R] \leq \exp (\varepsilon) \times \Pr \left[S(I') \in R \right].$$

(5)

The privacy budget is denoted by $\varepsilon$ which used to represent the privacy level. The smaller $\varepsilon$ represents the higher privacy protection level. The privacy level is inversely proportional to privacy budget $\varepsilon (\varepsilon \in [0, 1])$.

**Definition 6 Global sensitivity** [30]. Given a random query function $f : D \rightarrow R^d$, $D_1$ and $D_2$ are neighbor data sets at most one different data record. The global sensitivity formula is as follows:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1.$$  

(6)

**Definition 7 Laplace mechanism** [38]. Suppose a random query sequence $f$ with a query of length $n$. Given a function $f : D \rightarrow R^n$, the global sensitivity is $\Delta f$ and the privacy budget is $\varepsilon$. The formula for $\varepsilon$-differential privacy is as
Figure 1: Multidomain fusion data privacy security framework.

Figure 2: Recognition results of the multidomain data set for campus.

Table 1: Attribute sensitivity level table.

| Grade  | Value range          |
|--------|----------------------|
| Higher | 0 < SV(Att_i) < 0.6  |
| Moderate | 0.6 ≤ SV(Att_i) < 0.8 |
| Low   | SV(Att_i) ≥ 0.8      |

Table 2: Grading condition table for SPA sets.

| Grade   | Condition                                                                 |
|---------|---------------------------------------------------------------------------|
| Higher  | 1/2 < Num[IG(X_a) > γ × H(X_i)]/Num(DMA) ≤ 1                            |
| Moderate| 1/3 ≤ Num[IG(X_a) > γ × H(X_i)]/Num(DMA) ≤ 1/2                          |
| Low     | 0 < Num[IG(X_a) > γ × H(X_i)]/Num(DMA) < 1/3                            |
follows:

\[ f(\hat{D}) = f(D) + \text{Lap}\left(\frac{\Delta f}{\varepsilon}\right)^n. \]  

(7)

4. Multidomain Fusion Data Privacy Security Framework

To solve the personal privacy security problem caused by multidomain personal privacy fusion data set containing strong background knowledge, we proposed a multidomain fusion data privacy security framework shown in Figure 1 and the framework includes four models: (1) the input model includes multidomain fusion data sets and expert knowledge; (2) Multidomain Fusion Data Recognition, Classification, and Grading model (MRCG); (3) Multimodal Histogram Building model (MHB) according to the result of recognition, classification, and grading builds multimodal histograms; and (4) Multimodal Histogram Publishing Privacy Protection model (MHPP), through constraint inference algorithm and grouping reconstruction algorithm to achieve the multimodal publishing histogram privacy protection.

4.1. Multidomain Fusion Data Recognition Classification Grading Model. In the MRCG model, we present the definition of microdomain and microdomain attribute set and combine them with expert knowledge, information entropy, condition entropy, information gain to realize recognition classification and grading for the multidomain fusion data.

4.1.1. Microdomain Attribute Recognition Module. If data set \( \hat{D} \) contains communication, location tracking, personal information, health and other personal privacy information, we call data set \( \hat{D} \) multidomain personal privacy fusion data set. The different domains’ personal privacy information in data set \( \hat{D} \) is called microdomain, i.e., \( \hat{D} = \{MF_1, \ldots, MF_n\} \), which is a collection of different subfields. In this case, any subdomain \( F_{\text{micro}} \) in \( \hat{D} \) is called the microdomain, i.e., \( \hat{D} = \{MF_1, \ldots, MF_n\} \).

Definition 8 Microdomain. Define a multidomain personal privacy fusion data set \( \hat{D} = \{\text{Att}_1, \ldots, \text{Att}_n\} \) and a domain expert knowledge set \( ES = \{F_1, \ldots, F_n\} \). According to the domain expert knowledge \( ES \), the data set \( \hat{D} \) is transformed into \( \hat{D} = \{F_1, \ldots, F_n\} \), which is a collection of different subfields. In this case, any subdomain \( F_{\text{micro}} \) in \( \hat{D} \) is called the microdomain, i.e., \( \hat{D} = \{MF_1, \ldots, MF_n\} \).

Definition 9 Microdomain attribute set. Define a multidomain personal privacy fusion data set \( \hat{D} = \{\text{Att}_1, \ldots, \text{Att}_n\} \) and the set of attributes that make up microdomain \( MF_{k\in D} \) is called microdomain attribute set, i.e., \( MF_{\text{AS}} = \{\text{Att}_m, \ldots, \text{Att}_{n-1}\} \).
In this section, take the multidomain personal privacy fusion data set for the campus as an example. The data set contains seven attributes: grade, absences, phone, E-mail, address, health, and personal basic information. Figure 2 shows the recognition result of data attributes based on Definitions 9 and 10 and expert knowledge.

4.1.2. Microdomain Attribute Set Classification Module. When we select any microdomain attribute set $MFAS_k$ to publishing, the personal privacy in $MFAS_k$ can directly represent the privacy characteristics of the microdomain $MF_k$. Therefore, the attributes of the published microdomain attribute set $MFAS_k$ are called direct privacy attribute. The definition is as follows.

**Definition 10 Direct privacy attribute (DPA).** Set the data attribute recognition result is $\hat{D} = \{MF_1 \cdots MF_{n-1}, MF_n\}$, $MFAS_k = \{Att_m \cdots Att_{m-1}, Att_m\}$. When we select the attribute set $MFAS_k$ of any microdomain $MF_k$ in $\hat{D}$ to publish, the attribute in $MFAS_k$ is called the direct privacy attribute of microdomain $MF_k$.

---

**Algorithm 1:** Constraint inference for unattributed histogram.

```plaintext
Input: $\widetilde{SQ}[i]$: the noise frequency;
Output: $\overline{SQ}$: the constraint inference frequency
1 define listB, array[i], flag=false, start =0, end=0, sum=0, avg=0;
2 for i ∈ [0, i] // Line2-13 Judgment whether $\widetilde{SQ}[i]$ is a order query sequence
3 if i=0
4 listB ← $\widetilde{SQ}[i]$; //add the $\widetilde{SQ}[i]$ to listB
5 else
6 listB ← $\widetilde{SQ}[i]$;
7 if $\widetilde{SQ}[i] < listB$ get(i-1)
8 if flag = false;
9 start = i-1;
10 end = i;
11 flag =!flag;
12 else
13 End = i;
14 Handle(); //The recursive method of the order inference
15 end for
16 Handle() // Lines16-26 the order inference
17 for i∈[start,end]
18 sum=sum+listB.get(i);
19 end for
20 avg = sum/(end – start +1);
21 for i∈[start,end]
22 listB.set(i,avg);
23 end for
24 if start >=1 && listB.get(start-1) >avg;
25 start = start – 1;
26 Handle();
27 $\overline{SQ}[i] ← ar[i]$;
28 return the constraint inference frequency $\overline{SQ}[i]$;
```

---

**Figure 4:** Query sequence tree based on the full binary tree.
This section introduces information entropy and personali-
vacy attributes for DPA set. When the information gain is larger, it means that attributes in OA are more important to the attributes of the DPA; otherwise, it is less important to attributes in DPA. The attributes in OA are classified as sensitive privacy attributes and implicit privacy attributes for DPA set.

According to Definition 10, the microdomain is divided into DPA set and other attribute sets (OA, DPA ∪ OA = D). This section introduces information entropy and personalization parameter γ to realize the attribute classification in the OA set. The value of parameter γ is any value within the range of effective classification. The larger the information gain value, the stronger the correlation between the attributes in OA and the attributes in DPA. Therefore, the attributes in OA include two categories: sensitive privacy attributes (SPAs) and implicit privacy attributes (IPAs). The definitions are as follows.

**Definition 11 Sensitive privacy attribute (SPA).** If IG(Xi) ≥ γ × H(Xi) (Xi ∈ DPA, Xi ∈ OA), the attributes in the MFAS are called sensitive privacy attribute, where the parameter γ is threshold.

**Definition 12 Implicit privacy attribute (IPA).** If IG(Xi) < γ × H(Xi) (Xi ∈ DPA, Xi ∈ OA), the attributes in the MFAS are called implicit privacy attribute, where the parameter γ is a threshold.

For example, let the data set be $D = \{\text{Att}_1, \text{Att}_2, \text{Att}_3, \text{Att}_4, \text{Att}_5\}$. Firstly, according to Definitions 9 and 10, get $\tilde{D} = \{MF_1, MF_2\}$, $MFAS_1 = \{\text{Att}_1, \text{Att}_2\}$, and $MFAS_2 = \{\text{Att}_2, \text{Att}_3, \text{Att}_4\}$ and choose published microdomain attribute set $MFAS_1$. Secondly, we combine with Definition 11 to get $\text{DPA} = \{\text{Att}_1, \text{Att}_5\}$ and $\text{OA} = \{\text{Att}_2, \text{Att}_3, \text{Att}_4\}$, and then, according to Definition 4, calculate the information gain. Finally, based on Definitions 11 and 12, realize the attribute classification.

### 4.1.3. Microdomain Attribute Set Grading Module

Information entropy can measure the expectation of the amount of data attribute information. The maximum entropy reflects the data attribute maximum information expectation, so we denote the sensitivity of the attributes by the relative rate of attributes information entropy and maximum entropy. The formula for attribute sensitivity is as follows.

$$SV(Att_i) = \frac{H(Att_i)}{H_{\text{max}}(Att_i)}; \quad (SV(Att_i) \in (0, 1]).$$  \(8\)

As the formula is known, the smaller distance between the information entropy and maximum entropy, the stronger attribute sensitivity, otherwise the less sensitive of the attribute.

Sensitivity can effectively describe the importance of attribute privacy. In this section, we use the attributes sensitivity level table to achieve grading of attributes. Based on the personal privacy level grading in Personal Privacy Protection Law, we propose the sensitivity level table for the microdomain attribute set, as shown in Table 1.

Because there is a strong correlation between the attributes in the SPA set and the DPA set, we propose the attribute grading process in the SPA set from the perspective of information gain. And the grading condition for SPA sets as shown in Table 2.

1. Count the number of attributes satisfying $IG(Xi) > \gamma \times H(Xi)$ in SPA set, denoted as $\text{Num}(IG(Xi) > \gamma \times H(Xi))$
2. Count the number of attributes in DPA set, denoted by $\text{Num}(DMA)$;
3. We implement the attribute grading in SPA set according to the following conditions

```plaintext
Algorithm 2: Constraint inference for universal histogram.
```

**Input:** \(\text{NL} [i]\): the noise frequency array, \(h[i]\): the height of each node in the tree array

**Output:** \(\text{NL}\): the constraint inference frequency

1. Define array \(\text{CV}[j]\) storing the estimated value;
2. For each \(i\) in \(\text{NL}\) // The full binary tree from the bottom-up
3. If \((h[i]=0)\) \(i\) is the leaf node
4. \(\text{CV}[j] \leftarrow \text{NL}[i]\);
5. Else
6. Calculating \(\text{CV}[j]\) according to formula (11);
7. End for;
8. For each \(j\) in \(\text{CV}[j]\) // The full binary tree from the top-down
9. If \((h[j]=\log_2 n+1)\) \(j\) is the root node;
10. \(\text{NL}[j] = \text{CV}[j]\);
11. Else
12. Calculating \(\text{NL}[j]\) according to formula (12);
13. End for;
14. Return \(\text{NL}\);
Finally, according to the results of recognition, classification, and grading, combine the unattributed histogram and universal histogram [37] to realize multimodal mathematical statistics based on attribute sensitivity for multidomain fusion data sets.

4.2. Multimodal Histogram Publishing Privacy Protection Model. To solve the problem of privacy leakage caused by strong background knowledge in multidomain personal privacy fusion data set, in this section, we propose a privacy protection model for multimodal histogram publishing. According to the differences between the unattributed histogram and the universal histogram in paper [39], we use the two different constraint inference algorithms and combine them with grouping reconstruction algorithm to achieve multimodal histogram data publishing privacy protection [35, 36].

4.2.1. Model Overview. Figure 3 shows the privacy protection model in this paper, including added noise, constraint inference, and grouping reconstruction.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Microdomain & Attribute \\
\hline
Personal & Sex, age, school, guardian, AddressType \\
Family & Pastatus, Fedu, Medu, Fjob, Mjob, Famrel, Famsup \\
Entertainment & Internet, activities, Freetime, Dalc, Walc, Goout \\
Campus & Higher, Schoolreason, G1, G2, G3, absence \\
After-school & Studytime, Schoolsup, paid, Famsup \\
Health & Health \\
Spatial & Traveltime \\
Emotion & Romantic \\
\hline
\end{tabular}
\caption{Microdomain recognition results.}
\end{table}

(1) \textit{Added Noise}. Firstly, we classify histograms by step ① in Figure 3 and add the Laplace noise to histograms by step ②.

(2) \textit{Constraint Inference for Differential Privacy (CDP)}. We use positive-order inference and linear estimation for unattributed histogram and universal histogram with noise through step ③.
Figure 5: Continued.
(3) **Grouping Reconstruction Based on Constraint Inference (CDPR).** According to the grouping reconstruction algorithm of step ④, obtain the best grouping of histogram publishing.

### 4.2.2. Constraint Inference for Differential Privacy

Constraint inference makes noise data approach actual data by query constraint conditions such as order and nonnegative. Firstly, add Laplace noise to original query sequence \( Q \), which has constraint inference \( C_Q \) to obtain the query sequence \( q_1 = Q_1(I) \). Then, by the constraint inference rules and L2 distance, calculate \( q_2 \), which is closest to \( q_1 \). The minimum L2 solution is defined as follows.

**Definition 13 Minimum L2 solution** [39]. Let \( Q \) be a query sequence with constraints \( C_Q \). Given a noisy query sequence \( q_1 = Q_1(I) \), a minimum L2 solution and denoted \( q_1 \), that is a vector, satisfy the constraints \( C_Q \) and at the same time minimize \( ||q_1 - q||_2 \).

### (1) Unattributed Histogram

**Add Noise.** Define the original unattributed histogram query sequence \( Q \). Since we only care about the frequency distribution of the unattributed histogram. Therefore, any sort of query sequence is equivalent. In this section, we use the positive-order query sequence \( SQ \) to replace the original query sequence \( Q \). For example, if the original query sequence \( Q = \{5,10,2,3,9\} \), then the positive-order query sequence \( SQ = \{2,3,5,9,10\} \).

According to Definitions 6 and 7, the global sensitivity of the query sequence is \( \Delta f = 1 \), and the randomized algorithm

| Microdomain  | Attribute | Sensitivity   |
|-------------|-----------|---------------|
| Personal    | Sex       | 0.997960      |
|             | Age       | 0.962287      |
|             | School    | 0.519086      |
|             | Address   | 0.765220      |
|             | Guardian  | 0.724476      |
|             | Paststatus| 0.480919      |
|             | Fedu      | 0.872324      |
|             | Medu      | 0.860186      |
| Family      | Fjob      | 0.726524      |
|             | Mjob      | 0.928905      |
|             | Famrel    | 0.760564      |
|             | Famsize   | 0.866916      |
|             | Internet  | 0.651001      |
|             | Activities| 0.999773      |
| Entertainment| Freetime | 0.869073      |
|             | Dalc      | 0.569988      |
|             | Walc      | 0.915546      |
|             | Gout      | 0.921601      |
|             | Higher    | 0.289079      |
|             | Schoolreason| 0.933194    |
| Campus      | G1        | 0.735019      |
|             | G2        | 0.735018      |
|             | G3        | 0.739303      |
|             | Absence   | 0.965337      |
|             | Studytime | 0.850267      |
|             | Schoolsup | 0.555003      |
|             | Paid      | 0.994959      |
|             | Famsup    | 0.963063      |
Figure 6: Continued.
Figure 6: Continued.
Theorem 14. Let $S = SQ(1)$, this algorithm is aimed at finding a query sequence $S$ that satisfies constraint condition $S[i] \leq S[i + 1], 1 \leq i < n$ and minimizes $||S - \bar{S}||/2$.

Positive-Order Inference. Given a query sequence $\bar{S} = SQ(l)$, the algorithm is aimed at finding a query sequence $\bar{S}$ that satisfies constraint condition $\bar{S}[i] \leq \bar{S}[i + 1], 1 \leq i < n$ and minimizes $||\bar{S} - \bar{S}||/2$.

The query sequence $SQ$ satisfies $SQ[i] \leq SQ[i + 1], 1 \leq i < n$.

We use the two cases to expand. The first case assumes the noise sequence is $\bar{S} = \{2, 6, 8\}$, it satisfies the positive-order constraint condition, and the final constraint inference sequence result is $\bar{S} = \bar{S}$. In the second case, if the query sequence $\bar{S} = \{2, 8, 6\}$ is unordered, according to Theorem 14, get the $\bar{S} = \{2, 7, 7\}$. The constraint inference is described in detail in Algorithm 1.

Algorithm 1 is a constraint inference for unattributed histograms. Line 1 defines list B, the array arr[$\bar{k}$], the sum variable sum, and the average variable avg in the recursive method. Lines 2-15 are the sequential check and positive-order constraint inference of the current noise sequence. Lines 7 to 13 are used to check whether the noise sequence is order. In line 14, we use the recursive method of the order inference to realize the order of the sequence. Lines 16 to 26 are the implementation of the order inference. Finally, we obtain an inference query sequence $\bar{S}[i]$ that satisfies constraint condition $\bar{S}[i] \leq \bar{S}[i + 1], 1 \leq i < n$ and minimizes $||\bar{S} - \bar{S}||/2$.

(2) Universal Histogram

Add Noise. Define the original universal histogram query sequence $L$. The unit interval of the universal histogram has meaningful. Therefore, overmuch noise accumulation in long-range queries leads to low accuracy and poor availability of the query sequence results. In order to reduce the cumulative error, this section replaces the original query sequence $L$ by creating a query sequence that supports long-range queries.

Universal histograms support any interval frequency queries, and the query of any interval frequency is based on unit interval frequency statistics. The frequency of the unit interval is the same as the leaf nodes of the tree, and any other interval is the same as other nodes of the tree. In this section, we use the full binary tree to create a long-range sequence $NL$ replace the original query sequence $L$.

Building a full binary tree of height $h$, the node set $V$ includes the leaf node set $lv$ and the other node set $v$. In a full binary tree, the parent nodes in each layer calculates from the query interval of the corresponding child nodes and the leaf nodes are the unit interval in the original query sequence $L$.

For example, the original query sequence is replaced by the full binary tree with $k = 2$ and $h = 4$ as shown in Figure 4. The original query sequence is $L = \{0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, 0.9-1.0\}$ and the replace query sequence is $NL$ = $\{0.2-1.0, 0.2-0.6, 0.6-1.0\}$.
According to Definition 6, the global sensitivity of the query sequence NL is the height of the full binary tree, and we get the randomized algorithm $f_{NL}$ which satisfies $\varepsilon$-differential privacy by Definition 7.

$$\tilde{NL} = NL + \text{Lap}\left(\frac{h}{\varepsilon}\right) \times n. \quad (10)$$

**Linear Estimation.** After noise added to query sequence NL, the parent node frequency is not equal to the sum of corresponding child node frequency in the full binary tree, so we use linear estimation to constraint inference. Finally, we find a query sequence $\tilde{NL}$ which satisfies the constraint condition and minimizes $\|NL - \tilde{NL}\|_2$.

Firstly, we calculate the linear estimate of the full binary tree from the bottom-up. If the node is leaf node $e(\nu) = \tilde{NL} (\nu)$ and if it is not leaf nodes use the current node noise value $\tilde{NL} (\nu)$ and its child node linear estimate $e(\nu)$ recursive computation, the linear estimation formula is shown below.

$$e(\nu) = \begin{cases} \tilde{NL} (\nu) & \text{if } \nu \text{ is leaf node}, \\ \frac{V^{h-1}}{V^{h-1}} \tilde{NL} (\nu) + \frac{k^{h-1}-1}{k^{h-1}-1} \sum_{m \in \text{child}(\nu)} e(m) & \nu = V - l\nu. \end{cases} \quad (11)$$

**Figure 7:** Unattributed histograms in campus, entertainment, and family.
Figure 8: Universal histograms in campus, entertainment, and family.
In the formula, the $k$ denotes full binary tree fan-out, $h$ is the height of the current node $v_p$, and the $s(v_p)$ represents the child node set of the current node $v_p$.

Based on the current estimate $e(v)$, we adopt the top-down linear estimation calculation to full binary tree, if the current node is root node $NL(v_{root}) = e(v_{root})$. In the top-down traversal, if the parent node frequency is not equal to the sum of child node frequency, we use the following formula for linear estimation to constraint inference. The details of the algorithm are shown in Algorithm 2.

Defining the noise query sequence $NL = NL(I)$, the bottom-up estimation value is $e(v)$; then, the bottom-up
4.2.3. Grouping Reconstruction Algorithm Based on Constraint Inference. After constraint inference, the published histogram error is caused by noise error (NE), and the total error of grouping reconstruction based on constraint inference includes noise error (NE) and reconstruction error (CE). The sum of squares due to error (SSE) can measure the total error between the published histogram with privacy protection and the original histogram. The sum of squares error (SSE) formula is as follows:

\[
SSE = \sum_{i=1}^{n} \sum_{j=1}^{n} (D_i - ND_j)^2.
\]

\(D_i\) is the original data and the \(ND_j\) is the noise data. When the SSE is smaller, the absolute error value becomes smaller and the published histogram data availability becomes better. We find the minimum SSE to implement the best the grouping reconstruction. The core steps of grouping reconstruction algorithm are as follows.

The idea of this algorithm is to find the best grouping strategy by calculating the SSE between the group reconstruction histogram and the original histogram. The input is constrained inference sequence \(\text{NL}[i]\) and the original query sequence \(L[i]\). The output is the best group strategy and minimum SSE. Lines 1 to 3 define some variables to support the algorithm. According to lines 4-6, we calculate the absolute value of the difference values between adjacent buckets for the constraint inference result \(\text{NL}[i]\) and stored in \(DV[j]\). Lines 7 and 8 calculate the SSE when all buckets are combined into a group. The core algorithm of grouping reconstruction was based on SSE from lines 9 to 28. From lines 13-17, find the maximum value of the DV to group the sequence; then, according to lines 18-28, calculate the SSE of the current grouped sequence until satisfied conditions of the lines 10 and 11 to stop grouping. We sort the results of SSE to find the minimum SSE and get the best group strategy in lines 29 and 30.

The CDPR algorithm includes the CDP algorithm and the grouping reconstruction algorithm. In the CDP algorithm, both the attribute-free histogram and the general histogram are added with Lap \((\alpha/\varepsilon)\) noise through mathematical inferences based on conditional constraints to improve data availability. The grouping reconstruction algorithm structurally optimizes the data availability of the published histogram based on the CDP algorithm and still satisfies the \(\varepsilon\)-differential privacy. The grouping reconstruction algorithm realizes data optimization and improves the availability of the published histogram data based on the histogram structural characteristics. Moreover, the grouping reconstruction algorithm also satisfies \(\varepsilon\)-differential privacy.

In group reconstruction based on constraint inference algorithm, the time complexity of the histogram group reconstruction algorithm is \(O(n^2)\). Although the CDPR algorithm has a high time complexity, the time complexity is within an acceptable range, and the CDPR algorithm can improve long-distance query accuracy and data availability.

5. Experimental Results and Analysis

In this section, we use the real data set to analyze the multimodal histogram data publishing framework experiment. Firstly, we analyze the microdomain recognition experiments and the microdomain classification grading experiments. Secondly, we build multimodal histograms based on three microdomain and analyze the risk of histograms. Moreover, we use the multimodal histograms for the comparison experiment of the privacy protection model. Finally, comparing the LP algorithm and MDHP algorithm to prove the MHPP model has the advantage of low data error.

5.1. Experimental Setting

5.1.1. Experimental Data Set. The experimental data comes from a questionnaire filled out by 788 students [40], which contains 37 questions, and the final data set consists of 32 attributes. In the preprocessing of the data set, a valueless attribute was deleted, and the value in the attributes grade 1, grade 2, and grade 3 was converted into five levels: A, B, C, D, and F. Finally, after preprocessing and recognition processing, the data set contains 8 microdomains, 31 attributes, and 395 records.

Table 3 shows the 8 microdomains included in the multidomain fusion data set are as follows: personal, family, entertainment, campus, after-school, health, spatial, and emotion. Among them, campus is the performance of students in school and after-school is the content related to students’ spare time.
There is only one attribute in the three microdomains of health, spatial, and emotion which leads to low experimental significance. Therefore, the following experiments in this paper select personal, family, entertainment, campus, and after-school for experimental analysis.

5.1.2. Experimental Parameter Set

(1) Parameter $\gamma$ Set. In this section, we compare and analyze the recall rate and precision rate of five microdomains (personal, family, after-school, campus, and entertainment) at different parameters to determine the range of experimental parameters. The value range of the experimental parameters in this section is $[0.31, 0.34]$.

Figures 5(a)–5(e) show the recall rate and precision rate under different parameter values in 5 microdomains. When the parameter $\gamma = 0.31$, the recall rate in the five microdomains is 100%. The reason for this problem is that the recall rate is to calculate how many positive samples in the original sample are predicted correctly. Assume that the SPA set is a positive sample, and the HPA set is a negative sample, when the parameter $\gamma = 0.31$, the attributes in the five microdomains are basically classified into the positive sample SPA set. However, the HPA set cannot be effectively classified from the OA set under the parameter $\gamma = 0.31$. When the parameter $\gamma = 0.335$, the precision rate is 0 in the two microdomains of family and entertainment. At the same time, when the value range is $[0.335, 0.34]$, the recall rate is stored as 0, which indicates that the parameter $\gamma = 0.335$ cannot effectively classify the OA set. In summary, when the parameter $\gamma \in (0, 0.31]$ or $\gamma \in (0.335, 1]$, the OA set cannot be effectively classified into the SPA set and the HPA set. Therefore, the experimental parameter $\gamma\in (0.31, 0.335)$ in this paper can effectively classify the OA set into SPA set and HPA set. In the classification process, the recall rate and the precision rate have the same importance. This section realizes the weighted average calculation of different parameter $\gamma$ based on the average recall rate and average precision rate of different parameter $\gamma$ in 5 microdomains. According to the weighted average curve shown in Figure 5(f), it can be seen that the weighted average of the parameter $\gamma = 0.315$ performs better within the effective value range. Therefore, the experimental parameter $\gamma$ in this paper will be 0.315.

(2) Parameter $\varepsilon$ Set. This experiment uses the LP algorithm [25] and the MDHP algorithm [18] to compare and analyze the privacy protection model for multimodal histogram data publishing. In this experiment, we set the privacy budget $\varepsilon$ to 0.01, 0.1, and 1 [30] based on ensuring a reasonable allocation of the privacy budget.

5.2. Multidomain Attribute Set Classification and Grading Results. We calculate the attribute sensitivity $\text{SV(Att}_i\text{)}$ in the microdomain attribute set by the attribute sensitivity formula (X). When the $\text{SV(Att}_i\text{)}$ is larger, the uncertainty of the
information is more substantial and the information value of the attribute is greater. The attribute sensitivity results are shown in Table 4.

Figure 6 shows the classification results of DPA, SPA, and IPA attribute sets for different microdomains. The attribute set of DPA and IPA is graded by the attribute sensitivity of Table 4 and the attribute sensitivity level of Table 1 in Section 4.1.3. In contrast, the SPA attribute set uses the grading condition of Table 2 in Section 4.1.3.

5.3. Multimodal Histogram Building. In the multimodal histogram building experiment, we select campus, entertainment, and family to build multimodal histograms. In addition, we also analyze the risk of privacy leakage in the multimodal histogram.

5.3.1. Unattributed Histogram. Figure 7 shows the frequency of attribute distribution that satisfies $\text{IG}(\text{Att}_i) > 0.315 \times H(\text{Att}_i \in \text{DPA})$ and $\text{SV}(\text{Att}_i \in \text{SPA}) \geq 0.8$ in the SPA set of the campus, entertainment, and family. In the histogram, the abscissa is the attribute of the DPA set, and the ordinate is the frequency distribution of highly sensitive attributes in the DPA attribute sets and give the following formal conditions.

$$\text{Att}_i \in \text{SPA} \cap \text{SV}(\text{Att}_i) \geq 0.8 \cap \text{IG}(\text{Att}_i) > 0.315 \times H(\text{Att}_i \in \text{DPA})$$

\(14\)
Take Figure 7(a) as an example, if the attacker grasps the total number of highly sensitive attributes in the DPA attribute sets of campus is 14. Moreover, we still know the 13 attributes in the DPA attribute set and the highly sensitive attributes in higher attribute set. Then, the attacker can infer the remaining highly sensitive attributes by combining the attribute sensitivity in Table 4. At this time, the highly sensitive attributes are not only leaked, but also lead to privacy leaks and even malicious recommendations through the attacker’s data mining.

5.3.2. Universal Histogram. We select campus, entertainment, and family to build universal histograms. Figure 8 shows the distribution of attributes with a sensitivity level of higher and over 0.6 in the SPA attribute set. The sensitivity range of the abscissa in the universal histogram is [0.6, 1], and the ordinate is the number of SPA collection attributes satisfying the level of higher in the corresponding sensitivity range.

In Figure 8(a), assume that the attacker knows the frequency of campus in the sensitivity range of [0.9, 1] is 5 and knows the names of the 4 attributes. At this time, the attacker combines the attributes of the higher level in the SPA set with the attribute sensitivity in Table 4 can infer more private information and lead to private leaks. Due to the microdomain attribute set contains strong privacy information, they face a greater risk of privacy leakage.

5.4. Microdomain Privacy Data Publishing of Privacy Protection Result Analysis. In this experiment, the mean absolute error (MAE) is used to calculate the error between

![Graphs showing distribution of attributes with higher sensitivity levels for different privacy data sets.](image-url)
the original frequency and the frequency after privacy protection. The MAE can reflect the frequency availability of multimodal histograms published after privacy protection. The formula of MAE is as follows.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |Q_i(m) - \tilde{Q}_i(m)|^2.$$  \hspace{1cm} (15)

In the formula, \(n\) is the number of buckets of the histogram, \(Q_i(m)\) represents the original histogram frequency, and \(\tilde{Q}_i(m)\) is the histogram frequency value after privacy protection. According to the formula, if the value of MAE is smaller, the distance between the histogram frequency value after privacy protection and the original histogram frequency value is closer, which shows that the availability of frequency data is better.

5.4.1. Unattributed Histogram. The core algorithms of the privacy protection model for multimodal publishing histograms proposed in this paper are the constraint inference for differential privacy algorithm (CDP) and the grouping reconstruction based on constraint inference (CDPR). First, we compare and analyze the query accuracy and the mean absolute error of the CDP algorithm and the LP algorithm in this section experiment. Second, we choose the LP algorithm, the MDHP algorithm, and the CDPR algorithm proposed in this paper to compare and analyze the mean absolute error of the query. Through the analysis of the above comparative experiments, it proved that the privacy protection model for multimodal publishing histograms proposed in this paper can not only effectively guarantee the privacy of publishing histograms but also improve the availability of data.

(1) Analysis of the Query Results. In this experiment, the original unattributed histogram frequency set \(L\) is used as the baseline, and observe the distance from the baseline \(L\). In the differential privacy protection process, since the noise added to the frequency is random with negative numbers and decimals, this experiment uses nonnegative processing and rounding processing for the noise frequency. In the experiment, the privacy budget \(\varepsilon\) is set to 1, 0.1, and 0.01. As the privacy budget \(\varepsilon\) decreases, the requirement for privacy protection is stronger, and it shows that more random noise was added. We define the frequency results of the CDP algorithm and the LP algorithm as CDP-L and LP-L.

Figure 9 shows the frequency query results of unattributed histograms protected by CDP and LP algorithms in different microdomains. We observe that under different privacy budget \(\varepsilon\), and the frequency of unattributed histograms published based on the CDP algorithm is closer to the baseline \(L\). By observing that when the privacy budget \(\varepsilon = 1\), the CDP-L and the LP-L are both close to the baseline \(L\). The reason is that when \(\varepsilon = 1\), there is less random noise added, which makes the frequency disturbance of the original histogram smaller. With the privacy budget \(\varepsilon\) decreases, the distance of CDP-L and LP-L from the baseline \(L\) increases. When the privacy budget is 0.1 and 0.01, the noise added to the original frequency gradually increases, causing the deviation between the query result after privacy protection and the baseline \(L\) to gradually increase. As the privacy budget decreases, the histogram frequency published by the CDP algorithm is closer to the baseline \(L\) than the LP algorithm.

(2) Analysis of the CDP Algorithm. In this experiment, the value of the privacy budget \(\varepsilon\) is still 1, 0.1, and 0.01. Then, take the mean absolute error of 100 random queries under different privacy budgets and calculate the average value after repeating the experiment 50 times. Finally, we obtain the mean absolute error table and histogram shown in Table 5 and Figure 10. Figures 10(a)–10(c), respectively, show the mean absolute error of the CDP algorithm and the LP algorithm under different privacy budgets for campus, entertainment, and family. From the analysis of the mean absolute error results in Table 5, there are two reasons for the closer error when the privacy budget \(\varepsilon = 1\). One is that the added noise is small, resulting in small frequency disturbance of the original histogram, and the other is that the frequency of the buckets in the original histogram is relatively close, and the number of buckets is small. From the mean absolute error and privacy budget, the noise error in the CDP algorithm is significantly smaller than the error in the LP algorithm with the privacy budget decreases. It shows that the unattributed histograms published by the CDP algorithm have more higher accuracy under the same privacy budget.

(3) Analysis of the CDPR Algorithm. In the experiment, we take the mean absolute error of 50 samples under the same query range to calculate the average, and the privacy budget is set to 1, 0.1, and 0.01.

Table 6: Mean absolute error table.

| Privacy budget | Algorithm | Campus   | Entertainment | Family    |
|----------------|-----------|----------|---------------|-----------|
| \(\varepsilon = 1\) | CDP       | 2.1547   | 2.1127        | 2.1518    |
|                | LP        | 3.9398   | 3.9252        | 3.9509    |
| \(\varepsilon = 0.1\) | CDP       | 21.3226  | 21.2638       | 21.1145   |
|                | LP        | 39.5529  | 39.8937       | 39.1742   |
| \(\varepsilon = 0.01\) | CDP       | 214.8347 | 211.2819      | 212.6625  |
|                | LP        | 391.8511 | 396.3250      | 388.6677  |

Figures 11(a)–11(c), respectively, show the mean absolute error trend of the CDPR algorithm, MDHP algorithm, and LP algorithm under different privacy budgets for campus, entertainment, and family. As the query range increases under the same privacy budget, the CDPR algorithm has a lower error than the MDHP algorithm and the LP algorithm, which shows that the CDPR algorithm proposed in this paper satisfies the \(\varepsilon\)-differential privacy and improves the data availability. By observing the average absolute error curves under three different privacy budgets, it is found that
when the privacy budget decreases, the random noise increases, which leads to the increase of the mean absolute error of the CDPR algorithm, MDHP algorithm, and LP algorithm. However, the mean absolute error of the CDPR algorithm is still smaller than the mean absolute error of the MDHP algorithm and the LP algorithm. When the privacy budget decreases or the number of queries increases, the CDPR algorithm not only satisfies the \( \varepsilon \)-differential privacy but also publishes unattributed histograms with low error.

It can be observed from Figure 11 that there are two reasons why the CDPR algorithm proposed in this paper has no obvious advantages compared with the LP algorithm and the MDHP algorithm in a small range of queries. One reason is due to the frequency of the unattributed histogram experimental cases is similar, and the number of buckets is fewer, and when adding small random noise to the unattributed histogram, the frequency values of the unattributed histograms published by the CDP algorithm and the LP algorithm are similar. On the other hand, since the frequency of the unattributed histogram fluctuates greatly after noise is added, there may be cases of no merging in the grouping reconstruction stage, causing the results of the CDPR algorithm to be similar to the LP results. The last reason is that the frequency of adjacent buckets in the unattributed histogram published by the CDP algorithm is the same, which causes the CDPR algorithm to directly merge adjacent buckets with the same frequency.

5.4.2. Universal Histograms

(1) Analysis of the Query Results. In this experiment, the universal histogram original frequency set \( S \) used as the baseline. The histogram frequency sets published by the CDP algorithm and the LP algorithm are defined as CDP-S and LP-S. The accuracy of histogram data published based on the CDP algorithm was verified by comparing the distance between CDP-S, LP-S, and the baseline \( S \) when the privacy budget \( \varepsilon \) is 1, 0.1, and 0.01.

Figure 12 shows the results of frequency query under different privacy budgets. When the privacy budget \( \varepsilon = 1 \), the noise added to the original frequency is small, making the frequency inferred by the CDP algorithm similar to the frequency published by the LP algorithm, so the distance advantage between CDP-S and baseline \( S \) is not apparent. With the privacy budget is set to 0.1 and 0.01, the noise content of the original frequency increases. The frequency set CDP-S is closer to the baseline \( S \) than the frequency set LP-S, which shows that the CDP algorithm has lower error than the LP algorithm.

(2) Analysis of the CDP Algorithm. According to the mean absolute error results in Table 6, we draw the mean absolute error distribution as shown in Figure 13. In this experiment, the privacy budget \( \varepsilon \) is still selected as 1, 0.1, and 0.01, and take the average of the mean absolute error after 100 random queries for the count results of any interval. The whole
experiment was repeated 50 times, and the average value was taken.

According to the Table 6 and Figure 13, in the same microdomain, no matter the privacy budget is 1, 0.1, or 0.01, the average absolute error produced by the CDP algorithm is smaller than the average absolute error produced by the LP algorithm. However, since the random noise added by $\varepsilon=1$ is small, and the original frequency value in the experimental case is small, resulting in a small frequency fluctuation range after adding noise, the mean absolute error of the CDP algorithm is close to that of the LP algorithm. When the privacy budget is 0.1 or 0.01, more noise will be added to the original frequency. The mean absolute error of the CDP algorithm is significantly smaller than that of the LP algorithm. With the privacy budget decreases, the histogram published by the CDP algorithm has a lower error than the histogram published based on the LP algorithm.

(3) Analysis of the CDPR Algorithm. The long-range query frequency results of the general histogram are to calculate the mean absolute error of the CDPR algorithm, MDHP algorithm, and LP algorithm under different privacy budgets and compare and analyze the three algorithms. During the experiment, the interval size taken as $2^h$, where $h$ is the tree’s height. For the same interval size, we take the average of the

![Figure 14: Mean absolute error of universal histograms in campus, entertainment, and family.](image-url)
same number of times and repeat the entire experiment 50 times to get the average.

Observing Figures 14(a)–14(c), it is found that under the same privacy budget, as the query range increases, the mean absolute error curve of the CDPR algorithm is lower than the mean absolute error curve of the LP algorithm and the MDHP algorithm. When the privacy budget decreases, the mean absolute error curve of the CDPR algorithm is still lower than the mean absolute error curve of the LP algorithm and the MDHP algorithm. It shows that the universal histogram published based on the CDPR algorithm not only supports long-range queries, but also with the privacy budget decreases, the CDPR algorithm has lower error than the LP algorithm and the MDHP algorithm.

From the trend of the mean absolute error curve in Figure 14, it was found that the mean absolute error curve of the LP algorithm increases sharply with the increase of the query number. The reason is that due to the increase in the number of queries, too much noise is accumulated in the frequency of the unit interval, so more noise is accumulated when calculating the frequency of any interval. With the query range increases, the mean absolute error curve of the MDHP algorithm compared to the LP algorithm does not have a sharp increase significantly. Although the MDHP algorithm optimizes the noise error accumulated in the unit interval by merging adjacent buckets, the other arbitrary interval frequency is still calculated based on the unit interval frequency, so more errors are accumulated in the face of long-range queries or small privacy budgets. However, this paper replaces the original query sequence by building a full binary tree before adding noise, to avoid the noise accumulated in the unit interval from affecting the frequency of other arbitrary intervals.

(4) CDPR Algorithm Time Complexity. The CDPR algorithm is composed of constraint inference for differential privacy algorithm and group reconstruction based on constraint inference algorithm. In constraint inference for differential privacy algorithm, the time complexity of the differential privacy algorithm is $O(n)$. Then, execute the constraint inference algorithm, where the time complexity of the unattributed histogram positive sequence inference algorithm is $O(n^2)$, and the time complexity of the linear inference algorithm of the universal histogram is $O(\log 2n)$. In group reconstruction based on constraint inference algorithm, the time complexity of the histogram group reconstruction algorithm is $O(n^2)$.

In summary, the algorithm proposed in this paper has high time complexity, but the time complexity is within an acceptable range. At the same time, the method proposed in this paper can support long-range queries and improve the availability of data. The method proposed in this paper is suitable for small-scale, small-span, and small-change statistical data, such as government statistics, traffic statistics, and other related data. It is reasonable to improve the availability of data by sacrificing acceptable time complexity.

6. Conclusions

Due to the multidomain fusion data based on the collaborative collection of the IoT has rich background knowledge and attribute characteristics, it causes problems such as privacy leakage based on background knowledge and difficulty in multimodal data publishing. To address these problems, we propose a multidomain fusion data privacy security framework that includes three models such as the MRCG model, MHB model, and MHPP model. In the MRCG, first, we perform microdomain recognition through the proposed microdomain and microdomain attribute set definitions. Second, we use information entropy, conditional entropy, and information gain to realize the microdomain attribute set classification and grading. Finally, based on the results of MRCG, determine the attributes sensitivity and relationship between attributes to solve the problem of multimodal data statistics difficulties in multidomain fusion data sets. In the MHB, according to the results of multimodal data statistics, combine unattributed histogram and universal histogram to build multimodal histograms for multidomain fusion data set. In the MHPP, we improve the MDHP algorithm by positive-order inference and linear estimation to implement privacy protection for multimodal histogram data publishing.

Based on the real data set, the experimental results show that the multidomain fusion data privacy security framework not only realizes the recognition, classification, and grading of microdomain attribute set but also builds multimodal histograms by combining multimode data statistics with the unattributed histogram and the universal histogram. Furthermore, the CDPR algorithm in the MHPP model ensures the privacy security of the published multimodal histogram data. Compared with the LP algorithm and the MDHP algorithm, the CDPR algorithm has a lower frequency query error and improves the data availability of the published histogram.

In the future, we will research the recognition, classification, and grading model for streamed data and the privacy protection model for dynamic data publishing.

Data Availability

The data (multidomain fusion structured data) used to support the findings of this study are included within the article (“Using data mining to predict secondary school student performance”).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by (1) the National Natural Science Foundation of China under Grant nos. 61672179 and 61370083, (2) the Project funded by China Postdoctoral Science Foundation under Grant no. 2019M651262, and (3) the Youth Fund Project of Humanities and Social Sciences.
Research of the Ministry of Education of China under Grant no. 20YJCZH172.

References

[1] Y. Liu and T. Zhang, “Personal privacy protection in the era of big data,” Journal of Computer Research and Development, vol. 52, no. 1, pp. 229–248, 2015.

[2] T. Ya, Y. Lin, J. Wang, and J.-U. Kim, “Semi-supervised learning with generative adversarial networks on digital signal modulation classification,” CMC-Computers Materials & Continua, vol. 55, no. 2, pp. 243–254, 2019.

[3] J. Xiong, J. Ren, L. Chen et al., “Enhancing privacy and availability for data clustering in intelligent electrical service of IoT,” IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1530–1540, 2019.

[4] Q. Li, B. Xia, H. Huang, Y. Zhang, and T. Zhang, “TRAC: traceable and revocable access control scheme for mHealth in 5G-enabled IoT,” IEEE Transactions on Industrial Informatics, p. 1, 2021.

[5] M. Wang, Y. Lin, Q. Tian, and G. Si, “Transfer learning promotes 6G wireless communications: recent advances and future challenges,” IEEE Transactions on Reliability, vol. 70, no. 2, pp. 790–807, 2021.

[6] J. Xiong, R. Bi, M. Zhao, J. Guo, and Q. Yang, “Edge-assisted privacy-preserving raw data sharing framework for connected autonomous vehicles,” IEEE Wireless Communications, vol. 27, no. 3, pp. 24–30, 2020.

[7] J. Xiong, R. Ma, L. Chen et al., “A personalized privacy protection framework for mobile crowdsensing in IoT,” IEEE Transactions on Informational Informatics, vol. 16, no. 6, pp. 4231–4241, 2020.

[8] Y. C. Yang, L. F. Wu, G. Yin, L. Li, and H. Zhao, “A survey on security and privacy issues in Internet-of-Things,” IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1250–1258, 2017.

[9] M. Ammar, G. Russello, and B. Crispo, “Internet of things: a survey on the security of IoT frameworks,” Journal of Information Security and Applications, vol. 38, pp. 8–27, 2018.

[10] L. Sweeney, “K-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[11] W. Q. Zhang, G. S. Yin, Y. Sha, and J. Yang, “Protecting the moving user’s locations by combining differential privacy and k-anonymity under temporal correlations in wireless networks,” Wireless Communications and Mobile Computing, vol. 2021, 12 pages, 2021.

[12] J. B. Xiong, M. F. Zhao, M. Z. A. Bhuiyan, L. Chen, and Y. Tian, “An AI-enabled three-party game framework for guaranteed data privacy in mobile edge crowdsensing of IoT,” IEEE Transactions on Industrial Informatics, vol. 17, no. 2, pp. 922–933, 2021.

[13] B. Gedik and L. Liu, “Protecting location privacy with personalized k-anonymity: architecture and algorithms,” IEEE Transactions on Mobile Computing, vol. 7, no. 1, pp. 1–18, 2008.

[14] K. El Emam and F. K. Dankar, “Protecting privacy using k-anonymity,” Journal of the American Medical Informatics Association, vol. 15, no. 5, pp. 627–637, 2008.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: privacy beyond k-anonymity,” in Proceedings of the 22nd International Conference on Data Engineering (ICDE), pp. 24–35, Atlanta, Georgia, USA, 2006.

[16] B. Zhou and J. Pei, “The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks,” Knowledge and Information Systems, vol. 28, no. 1, pp. 47–77, 2011.

[17] X. Zhang, “An accurate method for mining top-k frequent pattern under differential privacy,” Journal of Computer Research and Development, vol. 51, no. 1, pp. 104–114, 2014.

[18] C. Piao, Y. Shi, J. Yan, C. Zhang, and L. Liu, “Privacy-preserving governmental data publishing: a fog-computing-based differential privacy approach,” Future Generation Computer Systems, vol. 90, pp. 158–174, 2019.

[19] X. Yang and L. Gao, “Balanced correlation differential privacy protection method for histogram publishing,” Chinese Journal of Computers, vol. 43, no. 8, pp. 1415–1432, 2020.

[20] K. Li and C. Wang, “High-precision histogram publishing method based on differential privacy,” Journal of Computer Applications, vol. 40, pp. 3242–3248, 2020.

[21] Z. Chong and W. Ni, “A privacy-preserving data publishing algorithm for clustering application,” Journal of Computer Research and Development, vol. 47, no. 12, pp. 2083–2089, 2010.

[22] H. Li, J. T. Cui, X. Meng, and J. Ma, “IHP: improving the utility in differential private histogram publication,” Distributed and Parallel Databases, vol. 37, no. 4, pp. 721–750, 2019.

[23] S. Zhang, L. Liu, Z. Chen, and H. Zhong, “Probabilistic matrix factorization with personalized differential privacy,” Knowledge-Based Systems, vol. 183, p. 104864, 2019.

[24] D. Wu, Z. Yang, B. Yang, R. Wang, and P. Zhang, “From centralized management to edge collaboration: a privacy-preserving task assignment framework for mobile crowdsensing,” IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4579–4589, 2020.

[25] W. He and C. Pen, “Sensitive attribute recognition and classification algorithm for structure dataset,” Application Research of Computers, vol. 37, no. 10, pp. 3077–3082, 2020.

[26] C. Diaz, S. Seys, J. Claessens, and B. Preneel, “Towards measuring anonymity,” in Proc. of the 2nd Int’l Conf. on Privacy Enhancing Technologies, R. Dingledine and P. Syverson, Eds., pp. 54–68, Springer-Verlag, Berlin, Heidelberg, 2002.

[27] A. Serjantov and G. Danezis, “Towards an information theoretic metric for anonymity,” in Proc. of the 2nd Int’l Conf. on Privacy Enhancing Technologies, R. Dingledine and P. Syverson, Eds., pp. 41–53, Springer-Verlag, Berlin, Heidelberg, 2002.

[28] C. Pen, “Information entropy models and privacy metrics methods for privacy protection,” Journal of Software, vol. 27, no. 8, pp. 1891–1903, 2016.

[29] Y. Yihan, “Metric and classification model for privacy data based on Shannon information entropy and BP neural network,” Journal on Communications, vol. 39, no. 12, pp. 11–17, 2018.

[30] B. Krishnamurthy and C. E. Wills, “Characterizing privacy in online social networks,” in Proceedings of the First Workshop on Online Social Networks, pp. 37–42, Seattle, WA, USA, 2008.

[31] X. Yong and X. Qin, “A QI weight-aware approach to privacy preserving publishing data set,” Journal of Computer Research and Development, vol. 49, no. 5, pp. 913–924, 2012.

[32] N. Li and T. Li, “T-closeness: privacy beyond k-anonymity and l-diversity,” in Proceedings of the 23rd International
Conference on Data Engineering (ICDE), pp. 106–115, Istanbul, Turkey, 2007.

[33] C. Dwork, “Differential privacy,” in Proceedings of the 33rd International Colloquium on Automata Languages and Programming (ICALP), pp. 1–12, Venice, Italy, 2006.

[34] C. Dwork, “Differential privacy: a survey of results,” Theory and Applications of Models of Computation Proceedings, vol. 4978, pp. 1–19, 2008.

[35] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Proceedings of the 3th Theory of Cryptography Conference (TCC), pp. 363–385, New York, USA, 2006.

[36] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, “Differentially private histogram publication,” in Proceedings of IEEE 28th International Conference on Data Engineering (ICDE), pp. 32–43, Washington, DC, USA, 2012.

[37] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, “Differentially private histogram publication,” International Journal of Very Large Database (VLDBJ), vol. 22, no. 6, pp. 797–822, 2013.

[38] X. Xiao, G. Wang, and J. Gehrke, “Differential privacy via wavelet transforms,” IEEE Transactions on Knowledge Data Engineering (TKDE), vol. 23, no. 8, pp. 1200–1214, 2011.

[39] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, “Boosting the accuracy of differentially private histograms through consistency,” in Proceedings of the 36th Conference of Very Large Databases (VLDB), pp. 1021–1032, Istanbul, Turkey, 2010.

[40] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” in Conference Committee of the ECEC-FUBUTEC’2008 Conference, Porto, 2008.