An Enhanced Approach for Multiple Sensitive Attributes in Data Publishing

Haiyan Kang(✉), Yaping Feng, and Xiameng Si

Department of Information Security, Beijing Information Science and Technology University,
Beijing 100192, China
kanghaiyan@126.com

Abstract. With the development of the e-commerce and the logistics industry, more and more personal information has been collected by the third-party logistics. The personalized privacy protection problem with multiple sensitive attributes is seldom considered in data publishing. To solve this problem, a method of Multi-sensitive attributes Weights Clustering and Dividing (MWCD) is proposed. Firstly, set the corresponding weight for each sensitive attribute value considering the different requirements of users and then cluster the data based on the weights. Secondly, divide the records by level rule to select record for l-diversity. Finally, publish data based on the idea of Multi-Sensitive Bucketization. The experimental results indicate that the release ratio of the important data though the proposed algorithm is above 95%, and the execution time is shorter.

Keywords: Data publishing · Multi-sensitive attributes · Privacy protection · Clustering · Dividing

1 Introduction

With the development of the Internet, e-commerce has achieved new growth and has driven the development of the logistics industry. Especially since 2011, China's express delivery business has increased at an average rate of more than 50% each year. In the end of 2017, the State Post Bureau announced that China's express delivery had reached 41 billion pieces, ranking the first in the world's express business for four consecutive years. The rapid development of the express delivery industry has improved people's lives to a certain extent and provided great convenience for people's lives. However, information sharing also leads to the leakage of personal private information, which is becoming serious. For example, there are more and more personal accidents in today's society, such as telephone or text message harassment, kidnapping, fraud. Therefore, studying the problem of privacy protection in data publishing can effectively reduce the risk of personal information disclosure.

In real applications, the published data has involved multiple sensitive attributes. Some attributes do not directly contain private information, but when combining with other attributes, they can further infer more private information. Based on the previous

Y.-B. Lin and D.-J. Deng (Eds.): SGIoT 2020, LNICST 354, pp. 79–94, 2021.
https://doi.org/10.1007/978-3-030-69514-9_8

work [1–4], we have presented an algorithm to decrease the leakage of private information. In this paper, we further study the multi-sensitive attribute privacy data release in the field of logistics and propose Multi-sensitive attributes Weights Clustering and Dividing method (MWCD).

Our main contributions are as follows:

(1) We pointed out the problem of privacy protection data publishing with multiple sensitive attributes.
(2) We propose Multi-sensitive attributes Weights Clustering and Dividing algorithm(MWCD), which can keep as much available information as possible, while maintaining data security.
(3) We compare our algorithm with the WMBF algorithm[5] on the real data set. The experimental results show that the release ratio of important data in our algorithm is higher (above 95%), and the execution time is shorter. Therefore, the published data through our algorithm had higher availability and achieved the effect of personalized privacy protection.

2 Related Work

In the study on privacy-preserving data publishing, based on grouping privacy protection have developed amounts of classic privacy protection models. Sweeney et al. firstly proposed *k*-anonymity model [6]. This model only cut off the connection between identifier and sensitive attributes, but it did not make corresponding requirement for sensitive attributes, and it was easy to produce homogeneity attack. Then, domestic and foreign scholars continue to improve the model, and *l*-diversity [7], *p*-sensitive *k*-anonymity model [8, 9], (*a*, *k*)-anonymous model [10], *t*-closeness model [11] and other models were proposed. The above privacy protection models are mainly for single sensitive attribute data, but in the actual situation, logistics data publishing involves multiple sensitive attributes. So data owners should pay more attention to privacy leakage for each sensitive attribute when publishing data. Yang Xiao-chun et al. proposed a Multi-Sensitive Bucketization (MSB) [12] based on lossy join for publishing data with multiple sensitive attributes, which is a breakthrough in this field. Tao Y et al. [13] proposed the (*g*, *l*)-grouping method, whose principle is: based on the MSB, in order to protect the personal privacy information in the data publishing with multiple sensitive attributes, the sensitivity concept of sensitive attributes was introduced, then the *l*-diversity and *g*-difference principles were used to constrain the value of each sensitive attribute in records. In 2010, Liu et al. [14] proposed the *l*-MNSA algorithm, which used anonymity to release data. In 2013, Han et al. [15] proposed a SLOMS algorithm to publish data for microdata with multiple sensitive attributes. Zhu et al. [16] proposed a new (w, y, k)-anonymity model and implemented it in a top-down approach. In 2014, Liu et al. [17] implemented the anonymous data publishing of multiple numerical sensitive attributes by adding noise. Xiao et al. [18] proposed a new privacy protection framework—differential privacy, to prevent any background knowledge attacking. In order to better prevent similarity attacks between multiple numerical sensitive attributes, in 2015, Sowmyarani and

Srinivasan [19] combined t-closeness technology with p-sensitive, k-anonymity technology to form a new privacy protection model for multiple sensitive attributes. Liu et al. [20] proposed the MNSACM method by using clustering and MSB for privacy preserving data publishing.

Some of the above data publishing methods are directed to the data with a single sensitive attribute, while others are directed to the data with multiple sensitive attributes. But they have some shortcomings in terms of algorithm efficiency, background knowledge attacks and others. Therefore, we propose Multi-sensitive attributes Weights Clustering and Dividing method.

3 Preliminary

3.1 Problem Description

Based on different application purposes (population statistics, income statistics, etc.), the third-party will publish data containing sensitive personal information. Once these data are used by attackers, it will lead to the disclosure of individual information and cause immeasurable consequences. Because the fact that different users have different application purposes, we need to protect the published data while meeting the needs of different users. But personalization issues are rarely considered in current data release studies.

The multi-dimensional bucket grouping model is still used in data publishing with multiple sensitive attributes. But, the multi-dimensional bucket of previous methods has the following shortcomings: (1) The grouping efficiency is lower. (2) The order of selecting buckets is improper, causing data suppressing problems. (3) They cannot meet the users' personalized needs. We propose relevant methods for resolving the above problems in the paper.

3.2 Problem Definition

It is assumed that the data table T $\{A_1, A_2,\ldots, A_x, S_1, S_2,\ldots, S_y\}$ is published by the data owner, where $\{A_1, A_2,\ldots, A_x\}$ represents the quasi-identifier and $\{S_1, S_2,\ldots, S_y\}$ represents the sensitive attribute. It is assumed that the number of data records is n in the data table T, then $|T| =$ n, and each data recorded as t_i $(1 \leq i \leq n)$.

Definition 1 (Lossy join). Data publisher divide the data set into two data tables. One includes the Group ID of data record and the quasi-identifier, and the other includes Group ID of data record and sensitive attributes.

Definition 2 (Multidimensional sensitive attributes). In the data table T, all the sensitive attributes form multidimensional sensitive attributes, which can be denoted by S. S_i $(1 \leq i \leq y)$ represents i-th sensitive attribute.

Definition 3 (Grouping). A group is a subset of the data records in the data table T. Each data record belongs to only one group in the data table T. The group of the data table T is denoted as $GT\{G_1, G_2,\ldots, G_m\}$, and $(QI_i \cap QI_j =)$ $(1 \leq i \neq j \leq m)$, where QI is a quasi-identifier attribute.

Definition 4 (Multidimensional sensitive attributes' *l*-diversity) [16]**.** In the group *G*, if each dimension sensitive attribute value of all data records satisfy *l*-diversity respectively, then the group *G* satisfies *l*-diversity of multi-dimensional sensitive attributes. In other words, the group *G* satisfies multi-dimensional sensitive attributes' *l*-diversity.

Definition 5 (d-dimensional bucket). [12]**.** If there are d-dimensional sensitive attributes, the d-dimensional bucket is denoted as Bucket $(S_1, S_2,\ldots, S_d)(2 \le d \le n)$, and each dimension of the multi-sensitive attributes corresponds to a one-dimensional bucket. According to the values of each dimension, the records are mapped to the corresponding bucket.

Definition 6 (Weights clustering). Suppose that there are n records in the data table *T*, each data record has d-dimensional sensitive attributes. For the sake of simplicity, we ignore the identifier and the quasi-identifier when forming multiple clusters which is denoted as $T_c\{C_1, C_2,\ldots, C_q\}(|C_q|\geq 1, 1 \le q \le n)$. The weights of the sensitive attribute values of each dimension in the same cluster are equal or similar, and the weights of different clusters are quite different.

Definition 7 (Weighted average value). It represents the average of the weights all sensitive attribute in each record.

Definition 8 (Weighted standard value). It represents the degree of deviation between the weight of different sensitive attribute in each record. The greater the value is, the greater the degree of difference. On the contrary, the smaller the value is, the smaller the degree of difference.

Definition 9 (Suppression technology). Some records that cannot be released or some records that do not satisfy privacy protection are hidden.

4 Multi-sensitive Attributes Weights Clustering and Dividing Model (MWCD)

4.1 The Overall Framework of Data Publishing

In this paper we improve the MSB grouping technology to achieve safer and more available data release. The main framework of this paper is shown in Fig. 1. It consists of three modules: data collection layer, method layer and data publishing layer. Specific explanations are as follows:

Data Collection Layer

Data collection is obtained through enterprise survey and web crawlers. The enterprise survey aims to understand the storage and release of data information in the express company. Web crawlers use crawler technology to collect information from website pages. The collected information is built into the original database. In order to make the algorithm more convenient, the data is preprocessed and shown in Table 1.

Method Layer

We firstly divide different records into multiple categories by clustering. Then we build weighted multi-dimensional buckets for multi-sensitive attributes, and map the records in

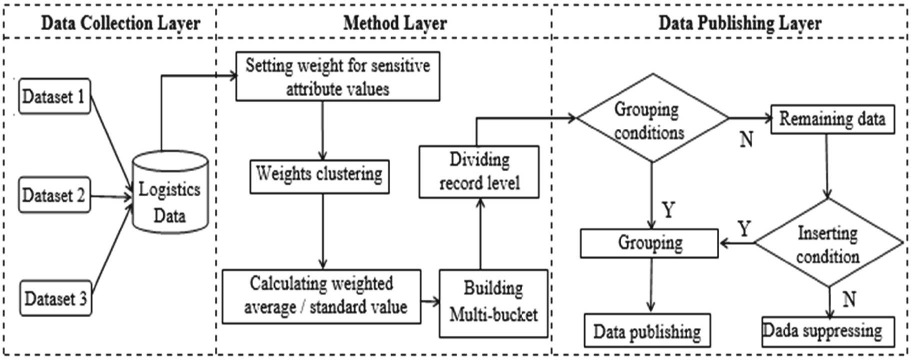


Fig. 1. The flow chart of data publishing.

Table 1. Part of the customer data of a logistics company.

Identifier	Quasi-identifier		Sensitive attribute			
ID	Zip	Age	Home address	Phone number	Name of goods	Credibility
t_1	0052	29	HeBei **(0.8)	132****7752(0.9)	Food	B
t_2	0029	31	TianJin**(0.6)	135****6851(0.2)	Clothing	A
t_3	0031	35	ShanDong**(0.5)	134****5612(0.7)	Phone	B
t_4	0058	28	HeBei **(0.1)	136****5762(1.0)	Toy	B
t_5	0062	32	TianJin**(0.5)	136****8623(0.1)	Clothing	A
t_6	0046	36	BeiJing(0.4)	137****6752(0.4)	Toy	A
t_7	0039	40	HeBei **(0.2)	139****4231(0.9)	Food	B
t_8	0075	30	ShanDong**(0.8)	187****1234(0.8)	Phone	C
t_9	0048	46	BeiJing**(0.6)	152****2564(0.2)	Book	A
t_{10}	0089	38	TianJin**(0.2)	136****8962(0.8)	Book	B

the data table to corresponding multi-dimensional buckets. Finally, we select important records by the hierarchical division method to satisfy the users' different needs.

Data Publishing Layer

By considering the max weight first selection algorithm, we select the data record in the multi-dimensional bucket to build a group satisfying l-diversity. Then we judge the data that do not satisfy the group for the first time. Finally, the quasi-identifiers of the groups that satisfy the condition are generalized. We suppress the remaining data and publish anonymous tables.

Our paper mainly improves the method layer and solves the following problems. (1) According to the weights that users set, we can meet the user's personalized privacy requirements. (2) Our core idea is clustering and dividing. We group the records that satisfy *l*-diversity by adopting record level strategy to protect data privacy. (3) We further process the clustered data and the remaining data to reduce data suppressing. Finally, we realize the data publishing.

4.2 Multi-sensitive Attributes Weights Clustering and Dividing Method

Setting Weight

According to different needs of users, we set the weight of sensitive attribute value for each record. As shown in Table 1, there are parts of the customer data of Logistics Company. Assuming that Table 1 is to be published, the data are divided into three parts: identifier, quasi-identifier and sensitive attribute. The data contains five sensitive attributes, such as home address, phone number, name of goods and credibility. In this paper, the first Two sensitive attributes are analyzed in Table 1, where the home address and phone number match corresponding weights for the user's requirements. For example, in Table 1, for the record t_1, the weight of the home address and phone number is 0.8 and is 0.9, respectively. For the record t_2, the weight of the home address and phone number is 0.6 and 0.2 respectively and so on.

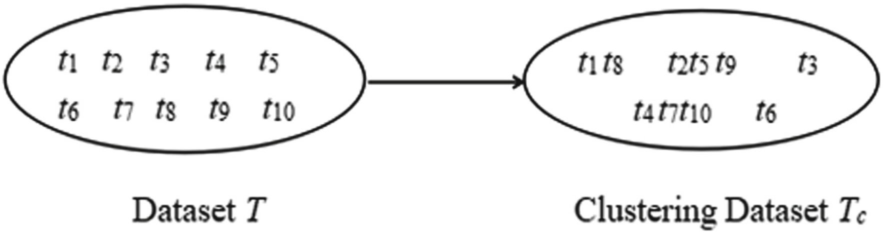

t_1 t_2 t_3 t_4 t_5
t_6 t_7 t_8 t_9 t_{10}
$t_1 t_8$ $t_2 t_5 t_9$ t_3
$t_4 t_7 t_{10}$ t_6
Dataset T
Clustering Dataset T_c

Fig. 2. The process of clustering.

Weights Clustering

Based on the weight of sensitive attribute value in each record, all records are clustered to form multiple clusters, recorded as clustered dataset Tc. In each cluster, the corresponding weights of all records are similar or equal. As shown in Fig. 2, it is a simple process of clustering for Table 1.

In Fig. 1, ten records are clustered to form 5 clusters, C_1 $\{t_1, t_8\}$, C_2 $\{t_2, t_5, t_9\}$, C_3 $\{t_3\}$, C_4 $\{t_4, t_7, t_{10}\}$, C_5 $\{t_6\}$. It can bring benefits for the next step. The process not only reduces time to calculate the weighted average value and the weighted standard value, but also contributes to comparison and selection results between the records when grouping, thus improve the efficiency of the algorithm and reduce the running time of CPU.

Calculating the Weighted Average and Weighted Standard Value for Each Record

For Table 1, the records have four sensitive attributes, recorded as S_1, S_2, S_3, and S_4. In t_1, the weight of each sensitive attribute is recorded as ${S_1}^1$, ${S_1}^2$, ${S_1}^3$, ${S_1}^4$. In t_2, the weight of each sensitive attribute is recorded as ${S_2}^1, {S_2}^2, {S_2}^3, {S_2}^4$, and so on. The Weight Average Value (WAve$^{\mathrm{n}}$) is denoted as

$$WAve^n = \frac{1}{d}\sum\nolimits_{i=1}^{d} S_i^n \tag{1}$$

The Weight Standard Value (WSve$^{\mathrm{n}}$) is denoted as

$$WSve^n = \frac{1}{d}\sum\nolimits_{i=1}^{d} \left(S_i^n - WAve^n\right)^2 \tag{2}$$

After weight clustering, calculate the weighted average and weighted standard value of each record based on the Eq. (1) and (2), and then write the result of each record in the corresponding place.

Building Multi-sensitive Bucketization

The previous method for building a multi-sensitive bucketization is that every dimension of the multi-sensitive bucketization corresponds to the dimension of the multiple sensitive attributes, and the value of each dimension corresponds to different sensitive attribute values. However, in practical applications, some sensitive attribute values are mostly different or even completely different. If there are a big amount of data records, the previous method cannot work. In this case, we first generalize the sensitive attribute values in order to reduce the classification of sensitive attribute values. Then build a multi-sensitive bucketization. We built a multidimensional bucket for Table 1, showed in Table 2. In the first two sensitive attributes, the attribute values of the phone number are different. We generalize each sensitive attribute value, according to respective dimension of the multi-sensitive bucketization.

Table 2. 2-*d* bucket structure.

	Class 1	Class 2	Class 3	Class 4	Class 5
HeBei**	$\{t_1, t_7\}$				$\{t_4\}$
TianJin**		$\{t_2, t_5\}$		$\{t_{10}\}$	
ShanDong**			$\{t_3, t_8\}$		
BeiJing**				$\{t_9\}$	$\{t_6\}$

Dividing Record Level

We divide the records based on the weighted average value, then selects important records and builds a group that satisfies *l*-diversity. The larger the weighted average value, the more important the record is. As shown in Table 2, the record level represents the importance of the record, and all records are divided into three levels. ‘A’ represents important, ‘B’ represents mediocre and ‘C’ is not important. The purpose of the record level table is to judge the importance of the records. The important records are firstly released.

In order to release important data as much as possible, while satisfying the privacy requirements of users, we use the maximum weight first selection strategy (where the ‘weight’ is not the weight of each sensitive attribute but considering the weighted average and the weighted standard value synthetically) to select the record for group which satisfy multi-sensitive attributes *l*-diversity, which ensure that important data are released.

Four conditions between the weighted average and the weighted standard value of each record in the same cluster are listed as follows:

(1) Both the weighted average value and the weighted standard value are different.
(2) The weighted average value is different, but the weighted standard value is equal.
(3) The weighted average value is equal, but the weighted standard value is different.
(4) Both the weighted average value and the weighted standard value are equal.

Therefore, the steps for defining the maximum weight first selection strategy are as follows:

Step1: If there exists the cluster of $WAve^n \in [0.6, 1]$ in data set, these data are very important. Alternately select the record with larger weighted average and smaller weighted standard value to form group until all the records are traversed in the cluster. Otherwise, goto Step2.

Step2: If there exists the cluster of $WAve^n \in [0.4, 0.6)$ in data set, these data are important generally. Alternately select the record with larger weighted average and smaller weighted standard value to form group until all the records are traversed in the cluster. Otherwise, goto Step3.

Step3: If it exists the cluster of $WAve^n \in [0, 0.4)$ in data set, these data are not very important. In turn, select the record with larger weighted average and larger weighted standard value form group until all the records are traversed in the cluster.

Table 3. Record level table.

The range of the weighted average	Record level
[0.6, 1]	A
[0.4, 0.6)	B
[0, 0.4)	C

When $WAve^n \in [0, 0.4)$, firstly choose the record with larger weighted standard value. The reason is that in some records, the individual sensitive attributes have higher weights and other sensitive attributes have lower weights. In order to guarantee the release of sensitive attributes with higher weights, we, only choose the record with larger weighted standard value to join the group. For example, if the weighted average of two records is 0.3 in the cluster, but the sensitive attribute weights are (0.8, 0.1, 0.1, 0.2), (0.3, 0.2, 0.4, 0.3), respectively. It can be seen that the former contains a sensitive attribute with higher weight, and the weighted standard values are 8.5 and 0.5, respectively by calculation, so that the former is first selected to ensure that important data is firstly released, which further satisfies the users’ personalized privacy requirements.

Publishing Data

It generalizes the quasi identifier of the data that satisfies l-diversity. The released data are divided into two data tables. One contains the records with the quasi identifier attributes and group number, called the quasi identifier attribute table QIT. The other contains the records with the sensitive attribute and group number, called the sensitive attribute table ST.

4.3 Multi-sensitive Attributes Weights Clustering and Dividing Algorithm

In this paper, we propose MWCD methods to release data, and we use Algorithm 1 to denote them. In this algorithm, we bring in clustering, division and generalization methods, and then select important records to release.

Take Table 1 as an example to illustrate the process of the proposed algorithm. We can see the weights of the first two-dimensional sensitive attributes and Table 3 shows the corresponding d-dimensional bucket ($d = 2$). By clustering, the dataset $T_c$contains 5 clusters: $C_1\{t_1, t_8\}$, $C_2\{t_2, t_5, t_9\}$, $C_3\{C_3\}$, $C_4\{t_4, t_7, t_{10}\}$, $C_5\{t_6\}$. According to the maximum weight first selection strategy, we first select record t_1 of cluster C_1 into group G_1, $G_1\{t_1\}$, while deleting t_1 in bucket; then t_8 is selected, and t_1 and t_8 satisfy multiple sensitive attributes l-diversity, so we insert t_8 into the group G_1, as $G_1\{t_1, t_8\}$, while deleting t_8 in bucket; the next step is to select record t_3 in cluster $C_{3.}$ Due to multiple sensitive attributes l-diversity rule, t_3 cannot be inserted into the group G_1, so a new group G_2 is built, and denoted as $G_2\{t_3\}$, and deleting t_3 in bucket; then record t_4 in the cluster C_4 is selected, and record t_3 and t_4 satisfy multiple sensitive attributes l-diversity, so we insert t_4 into the group G_2, as $G_2\{t_3, t_4\}$, and deleting t_4 in bucket; Then we can get the groups $G_1\{t_1, t_8, t_{10}\}$, $G_2\{t_2, t_3, t_4\}$, $G_3\{t_5, t_7, t_9\}$, and the remaining record is t_6. Once again, determine whether the remaining data can be inserted into the group. At this time, the remaining record t_6 can be inserted into the group G_1, as $G_1\{t_1, t_6, t_8, t_{10}\}$. Therefore, the final groups are $G_1\{t_1, t_6, t_8, t_{10}\}$, $G_2\{t_2, t_3, t_4\}$, $G_3\{t_5, t_7, t_9\}$. The final released data is shown in Table 4 and Table 5.

Algorithm 1 Multi-sensitive Weights Clustering and Dividing

```
Input: Data table $T$, Diversity parameter $l$.
Output: Quasi-identifier attribute table QIT, Sensitive attribute table ST.
Setting weight;
Weights clustering;
Calculating the weighted average value and the weighted standard value for each record and save them in the record;
Building multi-sensitive bucket;
Dividing records level;
while
    Selecting the record $t_i$ by the maximum weight first selection strategy;
    if (the record $t_i$ satisfies $l$-diversity)
        Inserting $t_i$ into the group and deleting $t_i$ from the bucket;
    else
        creating a new $l$-diversity group and deleting $t_i$ from the bucket;
end while (no record is selected)
for each the records which do not satisfy $l$-diversity group
    Selecting the record $t_i$ by the maximum weight first selection strategy;
    if ( a group $G_i$ still satisfy $l$-diversity after adding the record $t_i$)
        Inserting $t_i$ into the group $G_i$;
    else
        Inserting $t_i$ into the remaining dataset;
end for
Suppression all the remaining records;
Generalizing the quasi-identifier attributes of data in all groups;
Output QIT and ST;
```

5 Experimental and Analysis

5.1 Experimental Data and Environment

Experimental Data

The customer information of the logistics company is used as experimental data and we collect 7000 records. The description of partial data is shown in Table 1.

Experimental Environment

Processor Intel(R)Core(TM)i5-5200U CPU, memory 4GB, operating system Windows 10(×64), MATLAB and Java are used as the main test language.

5.2 Evaluation Criteria of the Algorithm

Release Ratio of Important Data

If the weighted average value of the record belongs to the range of [0.4, 1], it is defined as “important data”, the release ratio of important record is defined as Eq. (3).

$$Relration = \frac{number'(WAve \in [0.4, 1])}{number(WAve \in [0.4, 1])} \tag{3}$$

In the Eq. (3), $number'(WAve \in [0.4, 1])$ represents the number of important records in the published data, and $number(WAve \in [0.4, 1])$ represents the number of important records in the initial data. The larger the Relratio, the more important the record is. Important records are released first, so that personalized demand of users can be satisfied.

Additional Information Loss
In data set T, if there exist multi-sensitive attribute l-diversity group $G\{G_1, G_2,\ldots, G_m\}$, $|G_i| \geq l(1 \leq i \leq m)$, where m represents the number of groups, then the additional information is defined as Eq. (4).

$$AddInfo = \frac{\sum_{1 \leq i \leq m} |Gi - l|}{ml} \tag{4}$$

Suppression Ratio
When publishing data, if the number of suppression records is $|T_s|$ and the number of total data is $|T|$, then the suppression ratio is defined as Eq. (5).

$$Suppratio = \frac{|T_s|}{|T|} \tag{5}$$

Execution Time
The execution time means the time it takes to execute the proposed algorithm in the testing dada.

5.3 Experimental Analysis

The experiment will analyze the performance of the proposed algorithm from four aspects: release ratio of important data, additional information loss, suppression ratio, and execution time. With various data size $|T|$ ($\mathrm{k} = 10^3$), various diversity parameters l, and various the number of sensitive attribute d, we compare algorithm 1 with WMBF algorithm, and the results are shown in the following chart.

Analysis for Release Ratio of Important Data
Figure 3(a–c) gives the release ratio of important data in two algorithms under different parameters. For example, Fig. 3(a) shows the release ratio of important data under different data size $|T|$, when l is 3 and d is 3.

Three phenomena can be acquired from the experiment results:(1)With the increase of different parameters, the release ratio of important data has been changed in two algorithms, but it always keep above 0.80; (2) In Fig. 3(b), there is a big gap for the release ratio of important data in two algorithms. The reason is that between MWBF algorithm, the bigger the diversity parameter is, the more l-diversity groups are, and the number of records in each bucket is less, then the random selection of the records becomes relatively large, which may lead to selecting the unimportant records preferentially. In contrast, important records are preferentially selected in the algorithm 1. Therefore, the release ratio of important data is higher. (3) For the algorithm 1, the release rate of important data is always higher than the algorithm MWBF, which can reach above 95%.

Figure 4(a–c) shows the additional information loss under different parameters in two algorithms.

We can see that (1) In Fig. 4(a), the additional information loss is below 0.12 for different data size when $l = 3$ and $d = 3$ and it takes a significant trend as the increasing number of data size; (2) From Fig. 4(b), it can be seen that with the increase of diversity parameter l, the additional information loss also increases when $|T| = 6$ k and $d = 3$; (3) The additional information loss generated in the two algorithms is small, which is below 0.03, and means that the algorithm is close to optimal.

Analysis for Suppression Ratio

Figure 5(a, b) shows the suppression ratio under different parameters in two algorithms. When the parameters are different, the suppression ratio changes, too. When the suppression ratio is the shortest, it also indicates that the algorithm is optimal under this condition.

Three phenomena can be acquired from the experiment results: (1) In Fig. 5(a), the suppression ratio decreases as the data size $|T|$ increasing, when $l = 3$, $d = 3$, and it will close to 0 when data size $|T| = 7$ k. The reason is that the greater the number of data size, the better the diversification of sensitive attribute values of the record, and then the effect of the grouping becomes better, and the number of suppressed data also gradually decreases. (2) In Fig. 5(b, c), with the increase of l and d, the suppression ratio continues to grow in two algorithm. With the increase of diversity parameter l, we need insert more records into the group. If the number of records in a group does not satisfy diversity parameter l, this group is incomplete and the remaining data will increase; (3) In Fig. 5(a, c), the suppression ratio is much lower than the results of Fig. 4(b).

Analysis for Execution Time

Figure 6(a–c) shows the execution time under different parameters in two algorithms.

Three phenomena can be acquired from the experiment results: (1) From Fig. 6(a, c), it can be seen that with the increase of the number of data $|T|$ (or the number of sensitive attributes d), the execution time also increases linearly, but below 25 s; (2) In Fig. 6(b), the execution time of the two algorithms are between 15 s and 20 s when $|T| = 6$ k and $d = 3$. (3)The efficiency of the algorithm 1 is higher than that of the algorithm WMBF. The algorithm 1 cost less time in the weights clustering process, and thus the total execution time of algorithm is lower.

Table 4. Algorithm 1 QIT.

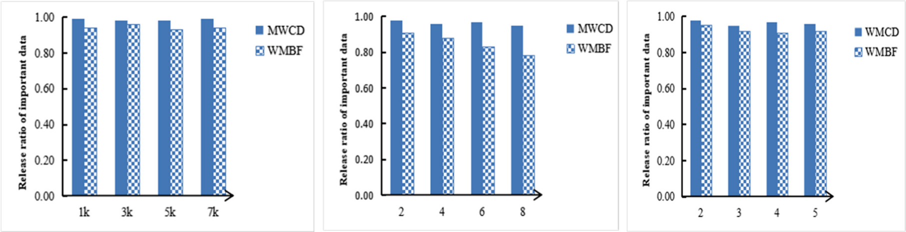
Zip	Age	Group ID
[0050–0090]	[28–38]	1
[0020–0060]	[28–40]	2
[0020–0060]	[28–40]	2
[0020–0060]	[28–40]	2
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1

Table 5. Algorithm 1 ST.

Group ID	Home address	Phone number
1	HeBei**	132****7752
	TianJin**	137****6752
	TianJin**	187****1234
	BeiJing**	136****8962
2	TianJin**	135****6851
	ShanDong**	134****5612
	HeBei**	136****5762
3	TianJin**	136****8623
	HeBei**	139****4231
	BeiJing**	152****2564

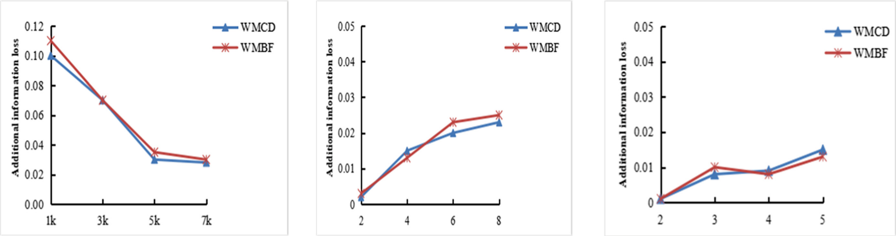
6 Conclusion

In this paper, we introduce data publishing for multi-sensitive attributes in logistics, and analyze the personalized privacy-preserving problem of multi-sensitive attributes values. Based on the idea of multi-sensitive bucketization, we proposed a method of Multi-sensitive attributes Weights Clustering and Dividing (MWCD) to satisfy the requirements of users. We adopted the clustering and dividing method to release data. Then we compared the proposed (MWCD) with WMBF algorithm. The experimental results show that the additional information loss and suppression ratio of two algorithms have a little difference, but the release rate of the important data in the proposed (MWCD) algorithm is above 95%, and the execution time is lower. Therefore, the published data



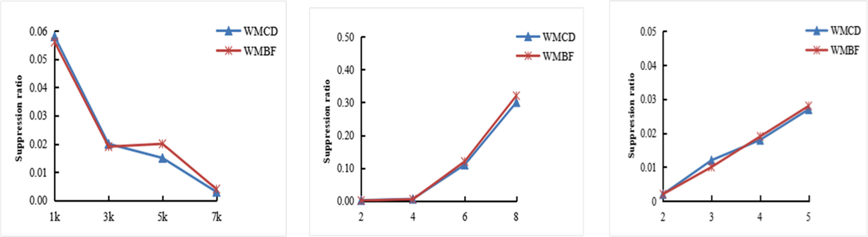
(a)Data size |*T*| (*l*=3, *d*=3) (b)Diversity parameter *l* (|*T*|=6k, *d*=3) (c) Number *d* (|*T*|=6k, *l* =3)

Fig. 3. Release ratio of important data under different parameters



(a)Data size |*T*| (*l*=3, *d*=3) (b)Diversity parameter *l* (|*T*|=6k, *d*=3) (c)Number *d* (|*T*|=6k, *l* =3)

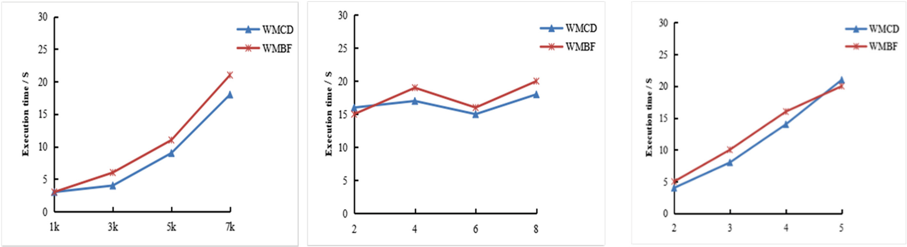
Fig. 4. Additional information loss under different parameters.



(a)Data size |*T*| (*l*=3, *d*=3) (b)Diversity parameter *l* (|*T*|=6k, *d*=3) (c)Number *d* (|*T*|=6k, *l* =3)

Fig. 5. Suppression ratio under different parameters.

in the proposed (MWCD) algorithm have high availability and achieved the effect of personalized privacy protection.



(a)Data size $|T|$ (l=3, d=3) (b)Diversity parameter l ($|T|$=6k, d=3) (c)Number d ($|T|$=6k, l =3)

Fig. 6. Execution time under different parameters.

Acknowledgment. This work is partially supported by Humanities and social sciences research project of the Ministry of Education (No. 20YJAZH046), fund of Bistu promoting the connotation development of universities and improving the scientific research level (No. 2019KYNH219), and Natural Science Foundation of China (No. 61370139).

References

1. Lu, Q.W., Wang, C.M., Xiong, Y., et al.: Personalized privacy-preserving trajectory data publishing. Chin. J. Electron. **26**(2), 285–291 (2017)
2. Li, J., Bai, Z.H., Yu, R.Y., et al.: Mobile location privacy protection algorithm based on PSO optimization. Acta Comput. Sinica **41**(05), 1037–1051 (2018)
3. Wang, H.Y., Lu, J.X.: Personalized privacy protection method for group recommendation. J. Commun. **40**(09), 106–115 (2019)
4. Zhou, C.L., Chen, Y.H., Tian, H., et al.: Network k nearest neighbor query method for protecting location privacy and query content privacy. Acta Softw. Sinica **31**(02), 229–250 (2020)
5. Lv, G.J.: Research on Privacy Protection Method of Multi Sensitive Attribute Data in Data Publishing. Chongqing University, Chongqing (2018)
6. Latanya, S.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(05), 557–570 (2002)
7. Ashwin, M., et al.: L-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Disc. Data **2006**(1), 24–36 (2007)
8. Truta, T.M., Vinay, B..: Privacy protection: p-sensitive k-anonymity property. In: Proceeding of the 22th International Conference on Data Engineering 2006, pp. 94–103. ICDE (2006)
9. Sun, X.X., Wang, H., Li, J.Y., Ross, D.: Achieving P-sensitive K-anonymity via anatomy. In: Proceedings of the 2009 IEEE International Conference on e-Business Engineering 2009, pp. 199–205. ICEBE (2009)
10. Wong, R.C.W., Li, J., et al.: (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining 2006, pp. 754–759. ACM (2006)
11. Li, N.H., Li, T.C., Venkatasubramanian, S.: T-Closeness: privacy beyond k-anonymity and l-diversity. In: International Conference on Data Engineering 2007, pp. 106–115. IEEE (2007)
12. Yang, X.C., Wang, Y.Z., Wang, B., et al.: Privacy preserving approaches for multiple sensitive attributes in data publishing. Chin. J. Comput. **31**(04), 574–587 (2008)
13. Tao, Y., Chen, H., Xiao, X., et al.: Angel: enhancing the utility of generalization for privacy preserving publication. Trans. Knowl. Data Eng. **21**(07), 1073–1087 (2009)

14. Liu, T., Ni, W., Chong, Z., et al.: Privacy-preserving data publishing methods for multiple numerical sensitive attributes. J. Southeast Univ. (Nat. Sci. Ed.) **40**(04), 699–703 (2010)
15. Han, J., Luo, F., Lu, J., et al.: SLOMS: a privacy preserving data publishing method for multiple sensitive attributes microdata. J. Softw. **8**(12), 3096–3104 (2013)
16. Zhu, H., Tian, S., Xie, M., et al.: Preserving privacy for sensitive values of individuals in data publishing based on a new additive noise approach. In: Proceeding of the 3rd International Conference on Computer Communication and Networks 2014, pp. 1–6. IEEE (2014)
17. Liu, Q., Shen, H., Sang, Y.: A privacy-preserving data publishing method for multiple numerical sensitive attributes via clustering and multi-sensitive bucketization. In: Proceeding of the sixth International Symposium on Parallel Architectures, Algorithms and Programming 2014, pp. 220–223. IEEE (2014)
18. Guo, X.L., Zhang, J., Qu, Z.Y., et al.: MADARS: a method of multi-attributes generalized randomization privacy preserving. Int. J. Multimedia Ubiq. Eng. **10**(10), 119–126 (2015)
19. Sowmyarani, C.N, Srinivasan, G.N.: A robust privacy-preserving model for data publishing. In: 2015 International Conference on Computer Communication and Informatics 2015, pp. 1–6. IEEE (2015)
20. Liu, T.T., Ni, W.W., Chong, Z.H., et al.: Privacy-preserving data publishing methods for multiple numerical sensitive attributes. J. Southeast Univ. (Nat. Sci. Ed.) **40**(4), 699–703 (2010)