Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 991-1008

http://www.ripublication.com

# Bucketize: Protecting Privacy on Multiple Numerical Sensitive Attribute

**Dharavathu Radha[1] and Prof. Valli Kumari Vatsavayi[2]**

*[2]Department of CS & SE, Andhra University, India,*
*Visakhapatnam-530003, Andhra Pradesh, India.*

**Abstract**

Data Privacy Preservation has depicted significant services in data. k-anonymization method is a solution to data privacy and has been the focus of research in the last several years. Presently anonymized privacy preserving data publication has been receiving considerable amount of attention. We used multi dimension bucketization to anonymize multiple sensitive attributes. This paper proposes bucketizeation method and finds *l*-deversity for sensitive attribute, presents the techniques to generalize quasi-identifiers, which prevents attacks. Here, we propose Protecting Privacy on Multiple Numerical sensitive Attribute through bucketization by taking into account sensitive information like income and loan. We experimentally show that the proposed method generates low suppression ratio and less information loss.

**Keywords:** Clustering, Multi-Sensitive Bucketization (MSB), numerical sensitive attribute, privacy preserving, *k*-anonymity.

## 1. INTRODUCTION

Nowadays society is allowing specifically quality of data well as person-certain information computer information as computer knowledge, network connectivity and disk storage space. The privacy preserving is a publication of personal data. Analyzing personal records requires data to be published while at the same time the individual privacy is protected which has become an issue of rising importance for the time being an anonymization concept is a familiar generalization that swaps quasi-

identifier attributes with those which are vague but logical throughout. Micro-data plays a significant part in data analysis. The propagation and partitioning of micro-data will jeopardize everyone's isolation. Hence, a few anonymity models were suggested to safeguard individual's privacy for recently published micro-data. An uncomplicated and helpful method is [1] *k*-anonymity, this method is used to secure privacy in micro-data and compels for each tuple have at most k identical tuples with regard to quasi-identifier (QID) in the published data. It cannot prevent similarity attacks and background knowledge attack, so some other enhanced anonymity models have been proposed, like *l*-deversity [2] and *t*-closeness [3]. The anonymity model implements by this little method.

Anonymity model is considered the generalization [5] process and it is the main role of anonymity model. This generalization process is to replace value of quasi-identifier. Anatomy [4] all quasi-identifier (QID) and sensitive values are released in two separate tables, here releasing the QID attributes surely may go through from larger breach possibility than generalization. Li Jiuyong et al [9] presented achieving *k*-anonymity by clustering in attribute hierarchical structures, this method is explanation substantially small distortions than an optional global recording *k*-anonymity approach. S.Liu and J.Li and Y.Tao [7] introduced a modern anonymization formula (ϵ,m)-anonymity that ignore the proximity breach in producing numeric sensitive attribute, the indicated paper compresses on micro-data that consists of only a single sensitive attribute. Yang et al [8] suggested a MSB (multiple sensitive bucketization) concept, but it is only suitable for micro-data with less sensitive attribute, for example 2,3 sensitive attributes.

In this paper we implemented a process, called bucketization. The idea of bucketization is to vertically partition the multiple numerical sensitive attribute tables and bucketization the sensitive attributes to apply *l*-deversity. Tuples are partitioned into equivalence-classes, each equivalence-classes are generalized under *k*-anonymity method used for quasi-identifier attributes, at the same way cluster the sensitive attributes, this is called protecting privacy on multiple numerical sensitive attribute. We apply multi-sensitive attribute bucktization and clustering on this paper. We obtain (gender, post-code, age) as QID (quasi-identifier) and (income, loan) as sensitive attributes as illustrated in Table 1. We obtain a 3-devrsity table as shown in Table 2.

### 1.1 Our Contribution

In this paper we proposed Multiple Sensitive Bucktization, which has three types of multiple sensitive bucketizations. Those are MBF, MSDCF, MDCF, later explained in detail. we apply multi dimension bucketization to anonymize multiple sensitive

attributes, to generate anonymity table with low suppression ratio and low information loss (InfoLoss). The remaining of this paper formulated as follows. Section 2 contains preliminary definitions, section 3 contains background and notation, section 4 contains measuring the quality of anonymity, section 5 contains formalization and comparison, section 6 contains bucketization algorithm, section 7 contains experiments of proposed work and section 8 contains conclusion.

**Table 1.** Micro-data

| Tuple | gender | Age | postcode | income | Loan |
|---|---|---|---|---|---|
| t1 | M | 23 | 31200 | 1000 | 600 |
| t2 | F | 27 | 32100 | 2975 | 1010 |
| t3 | M | 24 | 31204 | 1040 | 750 |
| t4 | M | 31 | 42000 | 1010 | 3050 |
| t5 | F | 29 | 32100 | 3050 | 1500 |
| t6 | M | 36 | 42005 | 5000 | 2035 |
| t7 | M | 31 | 42004 | 5100 | 2950 |
| t8 | F | 35 | 32004 | 7950 | 4100 |
| t9 | F | 36 | 31205 | 1050 | 790 |

**Table 2.** (1000, 3)-anonymity

| Group id | tuple | Age | post code | gender | income | loan |
|---|---|---|---|---|---|---|
| 1 | t1 | [23-26] | 321** | m | 1000 | 600 |
| 1 | t3 | [23-26] | 4200* | f | 1040 | 750 |
| 1 | t9 | [23-26] | 3120* | m | 1050 | 790 |
| 2 | t2 | [27-30] | 312** | * | 2975 | 1010 |
| 2 | t5 | [27-30] | 420** | * | 3050 | 1500 |
| 2 | t6 | [27-30] | 312** | * | 7950 | 4100 |
| 3 | t7 | [31-35] | 32*** | f | 5120 | 2950 |
| 3 | t4 | [31-35] | 32*** | m | 10100 | 3050 |
| 3 | t8 | [31-35] | 42*** | m | 5000 | 2035 |

## 2. PRELIMINARY DEFINATION

### *2.1 Clustering method*

Clustering is a useful method that partition records into groups. However, a collection of items are divided into groups known as cluster, so that items in the equal categories are additional identical defined resemblance standard. Subconsciously, one

ideal resolution of the *k*-anonymization issue is actually a collection of identified classes in those entries. It needs minimum generalization. *k*-anonymization as a clustering problem is normal clustering issue, such as k-means *k*-mens[10], this technology was used in a few papers to accomplish *k* -anonymization and require a different number of clusters to be created in results. Though, the k-anonymity problem does not have a limitation on the number of clusters rather, it needs that every cluster to consists of at most k records.

***Definition 1***: (*k*-means clustering problem). [13] Let us take S={$sa_1$, $sa_2$,…,$sa_m$} is a given a set of numeric attribute and C={$c_1$, $c_2$….,$c_k$} is a set of clusters such that each cluster consists at least *k* ($k \leq m$) records.

$$P(X, C) = \sum_{l=1}^{k} \sum_{i=1}^{m} x_i \, d\,(S_i, D_l) \qquad (1)$$

Where X is m × k partition model, and ($S_i$, $C_l$) is the squared distance between two attributes.

*2.2 MSB (Multi Sensitive Bucket) method*

Actually privacy preserving mechanics target on micro-data with single sensitive attribute. However, we cannot use for micro-data directly with multiple sensitive attributes. An oligo works determined on micro-data with multiple sensitive attribute (MSA). Yang et al [8] presented a structure, to accept privacy preservation in a Multiple Sensitive Bucketization (MSB) approach which contains multiple attributes. They proposed a multi structural bucket grouping method based on the thought of flossy involves, called Multiple Sensitive Bucketization(MSB). They present three types of MSB, which are, MBF (the maximal bucket first), MSDCF (the maximal single dimension capacity first), and MMDCF (the maximal multiple dimension capacity first). Although the multiple sensitive bucketization technique is only applicable to anonymize micro-data with less number of sensitive attributes, (example: 3-4 sensitive attributes. MSB would upshot in large suppression ratios. As instance, table 1 is an initial dataset. We deliberate that {gender, post-code, age} are quasi-identifier (QID), {income and loan} are sensitive attributes. We can obtain Multiple Sensitive Bucketization for 3-diversity table.

MBF: The selection signification of the maximal bucket is described as.

Selection (bukt<$sa^1$, $sa^2$, …. $sa^d$>)=size (bukt<$sa^1$, $sa^2$, …..$sa^d$>)

Where size (bukt<$sa^1$, $sa^2$, ….. $sa^d$>) is the number of tuples in bucket bukt<$sa^1$, $sa^2$, ….$sa^d$>.

MSDCF: The selection signification of maximal single dimension capacity is described as.

Selection (bukt<$sa^1$, $sa^2$, …., $sa^d$>)=max $_{1\le j\le d}$ capa($sa_j$)+ size (bukt<$sa^1$, $sa^2$,….. $sa^d$>)

Where size (bukt<$sa^1$, $sa^2$,….., $sa^d$>) is the size of the bucket (bukt<$sa^1$, $sa^2$,…., $sa^d$>) , max $_{1\le j\le d}$ capa($sa^j$) is the maximal number of tuples in each dimension bucket.

MMDCF: The selection signification of maximal multi dimension capacity is described as..

Selection (bukt<$sa^1$, $sa^2$,…., $sa^d$>)=$\sum_{1\le j\le d}$ capa($sa^j$)+ size (bukt<$sa^1$, $sa^2$,….., $sa^d$>)

Where $\sum_{1\le j\le d}$ capa($sa^j$) is the maximal number of tuples in each dimension bucket, size (bukt<$sa^1$,$sa^2$,…..$sa^d$>) is the size of the bucket (bukt<$sa^1$,$sa^2$,….$sa^d$>).

2.3 (ϵ,m)-anonymity method

(ϵ,m)-anonymity [6] method is new model that removes proximity breach in issuing numerical sensitive attributes. QID group QG, for each sensitive value 's' in QG, at least 1/ m of the tuples in QG can have sensitive values “alike” to ‘s. However, they proposed micro-data with unique sensitive attribute. They straightly employs the (ϵ, m)-anonymity concept.

***Definition 2 [6] :***

Let we take ‘ t’ be a tuple in table T, its closest district D(t) with it field of sensitive attribute S, its value is t.S of t. For example Table 3 shows the micro-data and Table 4 shows the (1000, 3)-anonymity of income. This income can't loss of privacy, but it will shortfall of privacy in order to get the linked bonus. However, let us suppose John in the first QID group, several tuple in the first QID group can belong to John. However, without extra information, an attacker supposes that every tuple in this group has an equivalent probability as long as had by John. Hence, the attacker infers that John’s bonus may be 1010, 5000, 1000. John’s income is around 1000 with a 70% probability.

**Table 3:** Micro-data

| Id | Age | Post Code | income | bouns |
|---|---|---|---|---|
| 1 | [25,29] | 13000 | 1000 | 1010 |
| 1 | [25,29] | 16000 | 4200 | 5000 |
| 1 | [25,29] | 27000 | 4100 | 6000 |
| 2 | [32,38] | 21000 | 5200 | 2100 |
| 2 | [32,38] | 18000 | 3020 | 1000 |
| 2 | [32,38] | 23000 | 3100 | 4000 |

**Table 4:** ☹1000,3)-Anonymity

| id | Age | Postcode | income | Bouns |
|---|---|---|---|---|
| t1 | 25 | 13000 | 1000 | 1010 |
| t2 | 26 | 16000 | 4200 | 5000 |
| t3 | 38 | 27000 | 4100 | 6000 |
| t4 | 32 | 21000 | 5200 | 2100 |
| t5 | 29 | 18000 | 3020 | 1000 |
| t6 | 37 | 23000 | 3100 | 4000 |

## 3. BACKGROUND AND NOTATION

***Definition 3***: QID (Quasi-identifier attribute) [9]l : A table has QID (quasi-identifier attribute set), this table hypothetically disclose private data, probably by linking along with another tables. For example, attribute set {gender, age, post-code} in Table1 is a quasi-identifier(QID).

***Definition 4***: [9] (Partition/QID-group): A partition contains of a few subsection of table T, such that every tuple in table T belongs to explicitly one subset. We mention these subsets as QID -groups, and denote them as $QID_1$, $QID_2$,…, $QID_m$. Namely, $\cup_{j=1}^{m} QID_j = T$ and, for any $1 \leq j_1 \neq j_2 \leq m$, $QID_D \cap QID_{j2} = \phi$.

***Definition 5*:(**Additional information loss) : Let $G_i$ exist a group within a table of *l*-diversity, n be the number of groups in the table of *l*-diversity, therefore the Additional information loss(AddInfoLoss) of single sensitive attribute taking part in group $G_i$ can be determined during.

$$\sum_{i=1}^{b} \frac{|G_i|-1}{b\times l} \quad (2)$$

The Additional information loss(AddInfoLoss) of the integral table T is determined as .

$$AddiInfoLoss = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{b_j} \frac{|G_i-1|}{b\times l} \quad (3)$$

Whereas m is the number of sensitive attribute. If at all possible, each tuple in real table ought to given way to one group in anonymity table, for all that, the limitation of *l*-diversity, few tuples cannot belong to a few group. Such tuples should be suppressed.

***Definition 6***: (Suppression ratio): let we take the number of suppressed(Suppratio) tuples are $m_s$ and table T, then the suppression ratio is described as.

$$SuppRatio = \frac{m_S}{T} \quad (4)$$

Automatically, the Suppression ratio is 0; the quality of anonymized data is superlative. Whereas the lower Suppression ratio the higher the quality of anonymized data. We use the Buketizeation algorithm to analyze effectiveness of the suppression ratio.

## 4. MEASURING THE QUALITY OF K-ANONYMITY

This segment we consider some metrics for calculating the quality of generalization. Here we have two aspects of information loss, that is the one generalization and anther one is above mentioned as suppression ration and information loss.

***Definition 7***: (Weighted Hierarchical Distance(WHD)).

We consider 'ht' is the orbit hierarchy of height, and levels 1, 2, ht-1, 'ht' are the area level. Whereas a block is generalized from level *k* to level *l*, whereas $k \geq 1$, the WHD (weighted hierarchical distance) of this generalization is well-defined as.

$$WHD(k,l) = \frac{\sum_{j=l+1}^{k} w_{j,j-1}}{\sum_{j=2}^{h} w_{j,j-1}} \quad where\ w_{i,i-1}\ \frac{1}{(j-1)}\ \ (2 \leq j \leq h) \quad (5)$$

Where $w_{j,j-1}=1/(j-1)(2\leq j\leq h)$

Take post-Code as an example [9]. Let post-Code hierarchy be {53500, 5350*, 535**, 53***, 5****, ****}, WHD from 5350* to 53*** is WHD(6, 4) = (1/5+1/4)/(1/6+1/4+1/4+1/3+1/2 +1) = 0.183.

Shows the Fig1 the right side the numbering methods of hierarchical levels and the left side weights between hierarchical levels. Level 1 is always the most general level of a hierarchy and contains one value. We can define weight $w_{j,j\text{-}1}$ to enforce a priority in generalization.

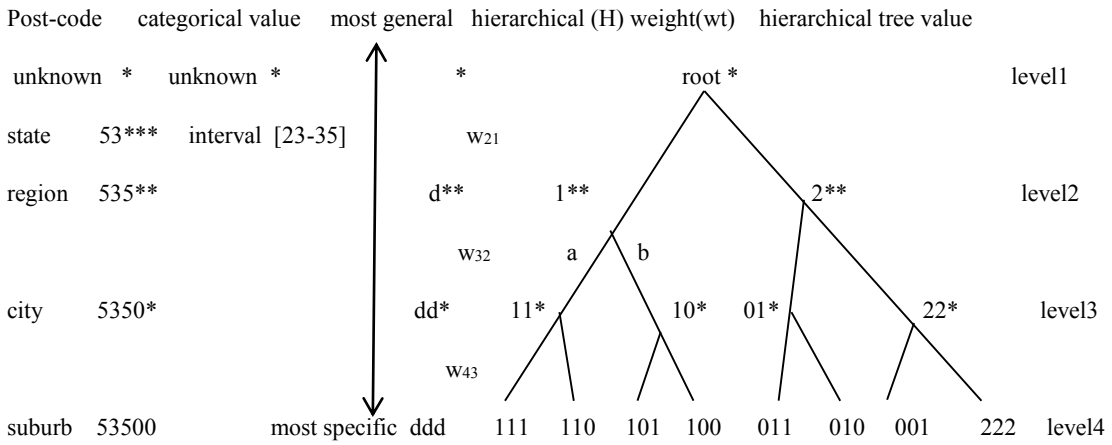



**Fig 1.** Heirarchical tree

Shows the Fig1 the right hand side the numeric hierarchical levels and the left hand side weights between hierarchical levels. Level 1 is every time the most general level of a hierarchy and consists of one value. We can describe weight $w_{j,j\text{-}1}$ to impose a priority in generalization.

***Definition 8***: (Tuples generalization)

Let us take G={x1, $x_2$,…, $x_m$} is a tuple with 'm' QID values and G' = {x'1, x'2,…., x'm} is the generalized tuple of t. Allow the level($x_j$) is the domain level of $x_j$ in an attribute hierarchy. The deformation of that generalization is described as.

$$Distrotion(G, G') = WHD(level(v_j), level(v'_j) \quad (6)$$

Take the example of [9], let the weight of WHD, hierarchy of attribute gender {m/ f ,*} and hierarchy of post-code {53500, 5350*,535**,53***,5****,****}. Let tuple $t_1$ in Table 1 and tuple $t'_1$ in Table 2, gender WHD=1, age WHD=2, post-code WHD=1/5=0.2. So that the Distortion of ($t_1$, $t'_1$)=3.2.

***Definition 9***: (Distortion of tables generalization)

Let us take quasi-identifier(QID) table $T_a$ and the generalized table of $T_a$' and $t_a^i$ is the generalization tuple of $t_a$ the distortion of $T_a$' is described as.

$$Distrotion(T_a, T_a') = \sum_{i=1}^{n} Distrotion(G_a, G_a') \quad (7)$$

Where | T | is the number of tuples in T.

From table 1 and table 2 WHD($t_1$,t'$_1$)= 3.2 (i.e For gender WHD=1, age WHD=2 and post-code WHD=1/5=0.2),….., WHD($t_9$, t'$_9$ )=3.2). The two tables of distortion is Distortion ($T_a$, T '$_a$ ) = 29.25.

***Definition 10***: (Closet common generalization)

From a hierarchical value tree accept all values of attribute. Every cost is defined when a node in the tree, and a node have a number of child nodes equivalent to its particular cost $t_{12}$ is the closest common generalization of $t_1$ and $t_2$ and its cost is defined as.

$$qv_{12}^i = \begin{cases} qv_1^i & if\ qv_1^i = qv_2^i \\ \text{the value of common ancestor} & if\ qv_1^i \neq qv_2^i \end{cases} \quad (8)$$

Where, qv$^i_1$ , qv$^i_2$ and qv $^i_{12}$ obtain the values of the i$^{th}$ QID in tuples are $t_2$, $t_5$ and $t_{25}$.

For example, Fig1b (RHS: weight between region levels and specified hierarchical value tree.) show a hierarchical value tree as well as region levels and $2^{(l-1)}$ values as every region level *l*. Let us take nodes 0** is the very closest common dynasty of nodes 110 and 101 in the hierarchical value tree. Allow for other case, $t_2$= {f, 27, 53503}, $t_5$= {f, 29, 53505}, $t_{25}$ = {*, [27-30], 53***}.

***Definition 11***: (Tuples distances)

Let us calculate the distance between $a_1$ and $a_2$, here, the closest common generalization is $a_{12}$. So its distance described as.

$$dist(a_1, a_{12}) = Distrotion(a_1, a_{12}) + Distrotion((a_2, a_{12}) \quad (9)$$

For example take attributes tuples in table 1, $t_2$ and $t_5$ (gender, age, post-code), $t_{25}$= {*, [27-30], 53***}. Dist($t_2$, $t_{25}$) = Distortion ($t_2$, $t_{25}$)+ Distortion ($t_5$, $t_{25}$) = 3.2+3.2=6.4

## 5. FORMALIZATION AND COMPARISON

Allow T to be a dataset by attributes $\{a_1, a_2, \ldots, a_n, sa_1, sa_2, \ldots, sa_m\}$, where $\{a_1, a_2, \ldots, a_n\}$ are QID (quasi-identifier attributes) and $\{sa_1, sa_2, \ldots, sa_m\}$ are SA (sensitive-attributes), *m* is the number of sensitive-attributes. In view of this section illustrate the implementation process of the method for simplicity; the QID table has generalization of all QID attributes and to implement *k*-anonymity. The generalized table has 3-quasi identifier (QID) attributes into equivalence group, at the same time, it clusters the numerical sensitive attributes as $SA_1, SA_2, \ldots, SA_m$. For each $SA_i$ $(1 \leq i \leq m)$ , we put its cost into multiple groups based on the approximate degree, which are marked as $SA_{i1}, SA_{i2}, \ldots, SA_{ij}$ , and the coupling of the groups can enclose entire the cost of $SA_i$ $(1 \leq i \leq m)$. The interchange of any two groups is the unfilled set instantaneously. Such as, suppose $SA_1$ is income, this implies that there are m numerical values in $SA_1$, on which we can use the numerical cluster process to place the m numerical values in that several groups, where the size of each group could be a distinct. Assume z is 20, then we cluster the m into 5 groups, that is ($SA_{11}$ to $SA_{15}$) Fig.2 shows the process of clustering.

Once the multi-dimensional bucket is constructed, we can choose various records to form the identical QID group. Yfei Tao [11], the certain greedy based Multiple Sensitive Bucketization (MSB) algorithms are suggested, it holds these three things MBF, MSDCF and MMDCF. We measure the suppression ratio and additional information loss consequently. Subsequently, to form we obtain many dissimilar records to produce the reciprocal QID group. Since the records in every QID group are chosen from various rows and various columns of the multi-dimensional bucket, those sensitive attributes are held from different $S_{ij}$.

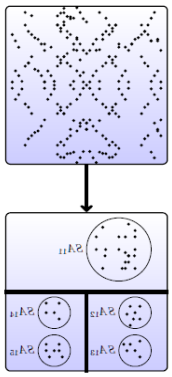


**Fig.2:** The process of clustering

To compare protecting privacy on multiple numerical sensitive attribute with KACA (Achiving *k*-Anonymity by Clustering in Attribute Hierarchical Structures), first we

formalize generalization with *l*-deversity, then after bucketization the sensitive attribute. It will give the amazing results then the KACA (Achiving *k*-Anonymity by Clustering in Attribute Hierarchical Structures). The supperssion ratio and additional information loss is better than the KACA (Achiving *k*-Anonymity by Clustering in Attribute Hierarchical Structures). Next, our proposal algorithm is Bucketization algorithm.

## 6. BUCKETIZATION ALGORITHAM

In this section we introduce our idea and technique as shown in Table 1. In that two numerical sensitive attributes, which are income and loan and QID is an age, gender and post-code. Mainly, we clustering sensitive attributes (such as income and loan). We set the income cluster within five groups $SA_{11}$= {1000, 1040, 1050}, $SA_{12}$= {2975, 3050}, $SA_{13}$= {5000, 5100}, $SA_{14}$= {7950}, $SA_{15}$= {10100}. Likewise, we also set the loan cluster within five groups: $SA_{21}$= {600, 750 and 790}, $SA_{22}$= {1010, 1500}, $SA_{23}$= {2035, 2950}, $SA_{24}$= {4100}, $SA_{25}$= {3050}. We formulate income and loan to be the first element and the second element subsequently are shown in table 5. Here, cluster groups of income and loan approach to rows and columns. Every cell in the table performs a bucket, during each and every records could be designed into the identical bucket lookalike through their individual sensitive attributes. Since the example is with $t_1$, its income belongs to $SA_{11}$ and its loan belongs to $SA_{21}$. Therefore, we put $t_1$ in the upper left bucket. Then as well, we place entire another records as well. Subsequently, we design a two-dimensional bucket, as shown in Table.5.

**TABLE 5.** TWO DIMENSINAL BUCKET

| | $SA_{11}$ | $SA_{12}$ | $SA_{13}$ | $SA_{14}$ | $SA_{15}$ |
|---|---|---|---|---|---|
| $SA_{21}$ | {$t_1$,$t_9$,$t_3$} | | | | |
| $SA_{22}$ | | {$t_2$} | | | |
| $SA_{23}$ | | {$t_5$} | {$t_6$,$t_7$} | | |
| $SA_{24}$ | | | | | {$t_4$} |
| $SA_{25}$ | | | | {$t_8$} | {$t_9$} |

Let us take three approaches, MBF, MSDCF, and MMDCF, we could choose dissimilar records to formulate the identical QID group. We take {$t_1$, $t_5$, $t_6$} when we use MBF. We get {$t_1$, $t_5$, $t_6$} and need to suppress the record {$t_2$, $t_3$, $t_4$, $t_5$, $t_7$, $t_8$, $t_9$} and the additional information loss is 0.036 and the suppression ratio is 0.6. However, MSDCF in we get {$t_1$, $t_2$, $t_5$, $t_6$, $t_7$, $t_9$} and need to suppress the record {$t_4$, $t_8$} and the additional information loss (AIL) is 0.012 and the suppression ratio is 0.2. In the

MMDCF is the highest selection of bucket <$SA_{13}$, $SA_{23}$ > of the value is 7 refuse to formula of MMDCF , which is the highest priority, then we echo the process, we can obtain bucket <$SA_{11}$, $SA_{21}$> of the value is 3 and bucket <$SA_{12}$, $SA_{22}$> of the value is 4. MMDCF, the result is {$t_1$, $t_3$, $t_9$}; {$t_2$, $t_5$, $t_6$}, {$t_4$, $t_7$, $t_8$} and there is no records for suppress. Here the two cases are zero, that is, the suppression ratio and the additional information loss. As seen, we conclude that MMDCF is the best. We get 3 QID groups in the table and that the sensitive attributes of every QID group are chosen from various clustering groups. We block the proximity breach successfully. As shown in Fig.3, the total structure of anonymity process, like QID (Qusi-Identifier), SA (sensitive attribute).

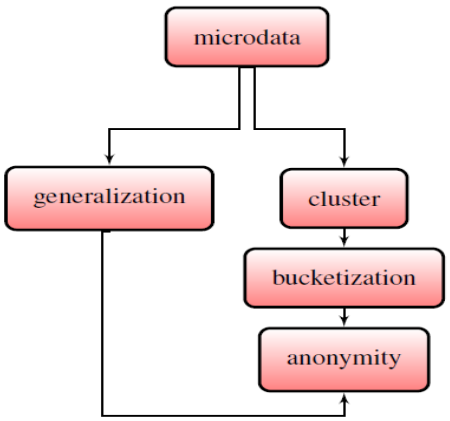



**Fig. 3:** Structure of the Anonymization

---

Algorithm 1 : Algorithm of MNSACB

---

**Original dataset** $T\{a_1, a_2, \ldots\ldots, a_n, s_1, s_2 \ldots\ldots, s_m\}$, parameter and $k$.

**Output**: Anonymous table T'

Procedure:

1. Begin
2. Select ECL from the real data set
3. While ∃ an ECL of size < $k$ do
4. Arbitrarily select an ECL< $k$
5. Calculate distance between ECL and all another ECL
6. find the E*CL'* with the smallest distance to E*CL*
7. the E*CL* and E*CL'* generalize
8. End while

9. Get numerical sensitive attributes $SA_1$, $SA_2$, … $SA_m$.
10. for each $Sa_i$,($1 \leq i \leq m$);
11. cluster values into approximate groups, i.e. $Sa_{ij}$,($1 \leq i \leq m$; $1 \leq j \leq n$);
12. end for;
13. each $Sa_i$.($1 \leq i \leq m$)correspond to one dimension into a bucket, establish m dimension bucket Bucket.$Sa_1$;$Sa_2$; …; $Sa_m$);
14. calculate the capacity of approximate groups for each $Sa_i$($1 \leq i \leq m$);
15. while (can extract records constitute a group)
16. set unshielded marker for all buckets, Grouping $G_i \neq \phi$, $i \leftarrow 1$;
17. Calculation selection of non empty bucket;
18. for (.j =1;j ≤ 1;j ++)
19. if (there is non empty and unshielded bucket)
20. select a record t from the maximum selection bucket buk and add it into group $G_i$;
21. delete t from buk, and size(buk)=size(buk)-1;
22. recalculate the capacity of buk for each dimension;
23. Shielding the bucket which has the same approximate group with t
24. }
25. Else
26. if (there is no record can choose)
27. the end of the group process;
28. }
29. End for;
30. i+ +;
31. End while;
32. if(Bucket($SA_1$; $SA_2$; …; $SA_m \neq \phi$ );
33. suppress all the remaining records in multi-dimensional bucket;
34. generalize quasi identifier in each $G_i$;
35. End if
36. Return on anonymity table T';

---

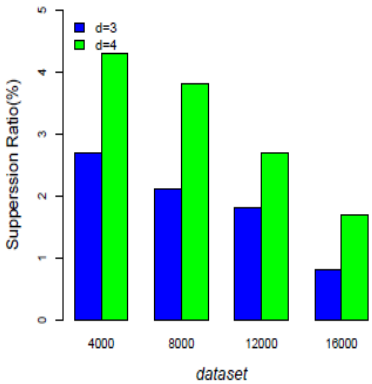
## 7. EXPERIMENT

The main aim of the exploration is to consider the presentation of our resolution in conditions of suppression ratio data size, information loss and runtime of CPU. To precisely estimate our resolution, we distinguished our implementation with another other algorithms, namely MSB-KACA. The Bucketization is better than these algorithms in suppression ratio and information loss. In this section, we governed different experiments to show the ability of our algorithm. All the experiments were

conducted on real data set. We considered (sex, age, race, marital-status, work class, education, education number, occupation, income). However, age, sex, race, marital-status treated as categorical attributes and work class, education, education number, occupation and income were treated as a numerical attribute. The depiction of adult data set is as shown in Table.7 , we chosen 16000 records of real dataset and experiments were conducted on an Intel (R) Core(TM) i3-4005U CPU @1.70GHz and 4GB memory running the Microsoft Windows 8.1 operating system. All the algorithms were implemented in Java with JDK version 1.8 on Windows 8.1. The real dataset used contains 16000 tuples and SA is 3 and 5 (i.e sensitive attributes number) tuple. We select three attributes to be sensitive attributes, i.e. Work class, Education, occupation ,education number and income, and corresponding on different *l* when n is 16000 and SA is 3 and 5.

**Table 6:** Depiction of adult data set

| S.No | Name of attribute | Type of Attribute | height | Values |
|---|---|---|---|---|
| 1 | Sex | Categorical | 4 | 85 |
| 2 | Age | categorical | 2 | 2 |
| 3 | Race | categorical | 3 | 5 |
| 4 | Marital-status | categorical | 3 | 7 |
| 5 | Work class | sensitive attribute | | 8 |
| 6 | Education | sensitive attribute | | 17 |
| 7 | Occupation | sensitive attribute | | 60 |
| 8 | income | sensitive attribute | | 25 |
| 9 | Education number | sensitive attribute | | 10 |



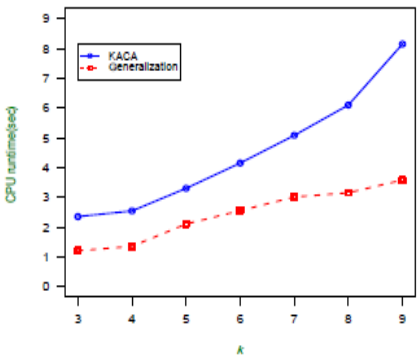




**Fig. 4:** reduce data size of suppression ratio **Fig. 5:** Comparison of Equivalence group(dataset=16000,QID=4)

As shown in Fig. 4 is the suppression ratio changes with data size while *l*=3. The green line shows the suppression ratio when the data has four numerical sensitive attributes. The blue line show the suppression ratio when the data has three numerical sensitive attributes (occupation and income, Education). In Fig.4, *l* value and data size are equal, and the green line is greater than the blue line. It shows that the suppression ratio of data is greater when the sensitive attribute dimension is larger. This is because there is a greater restriction in selecting records, when the sensitive attribute dimension is larger. At last, the remaining records increase leading to an increase in the suppression ratio. From a global context, where data size increase, the height of line step reduces, i.e, the suppression ratio has a decreasing trend. This is because as the data size increases, we need to build more record groups and there are more records that can be selected. Subsequently, when the original remaining records are added to some record groups, the proportion of remaining records is reduced and the suppression ratio decreases.

In Fig. 5 the line chart shows that the comparison of CPU runs time with equivalence group *k*. The cost of measuring publishable table by KACA (Achieving *k*-Anonymity by Clustering in Attribute Hierarchical Structures) and generalization on different *k*-anonymity when the real data set is 16000 and QID is 4. It is obvious that the cost of *k* value is increases, the CPU run time increases. The blue line shows the CPU run time when data set is 16000 and QID is 4, it's execution time more than the red line, that is generalization. However, our algorithm can completed in very less time. We conclude that the generalization is better than the KACA. So which is agreed for data-Anonymization?

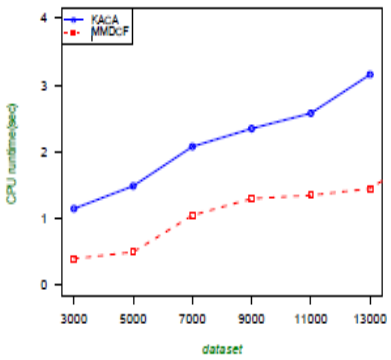



**Fig. 6:** Comparison of CPU Run Time

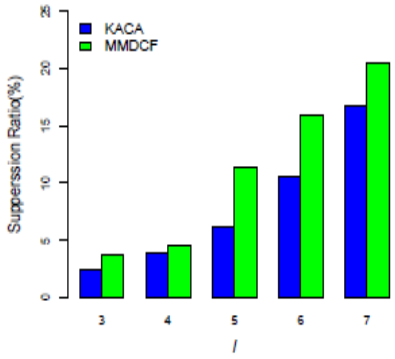



**Fig. 8:** Supperssion Ratio where sensitive attribute=3 (dataset=13000, QID=4, l=4)

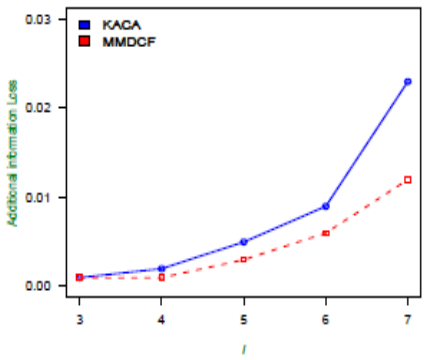



**Fig. 9:** Additional information loss

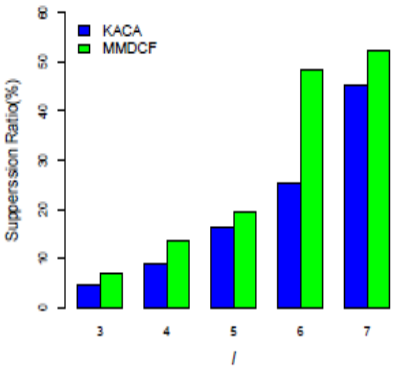



**Fig. 8:** Supperssion Ratio where sensitive attribute=5

As shown in Fig. 6. is the CPU run time changes with data set using KACA and MMDCF on various real data set although QID is 4, *l* is 4 and n is 13000. In Fig.6 the data set is increases, the time complexity is increases. However, the blue line is greater than the red line, it shows so that the CPU performance time increases when the real data set is increases.

Fig. 7 shows that the suppression ratio changes with *l*, where the tuples is 16000 and SA=3(sensitive attribute number=3. In order to have a comparison, we set *l* to be 3, 4, 5 and 7 respectively. As shown in Fig. 7. MMDCF shows the green line is higher than the blue line. Especially, MMDCF dimension is larger when the suppression ratio of data is higher. From a global aspect, with an increase in *l*, the height of the suppression ratio increases gradually, showing that the suppression ratio has an increasing trend. This is because the larger the *l*, the higher the privacy requirements. As a result, ensuring the group *l*-diversity in each dimension is more difficult. Thus the overall effect of variation grouping becomes of inferior quality (the suppression ratio of data increase. Gradually, KACA produces significantly lower suppression ratio then the MMDCF.

Fig. 8 shows that the suppression ratio changes with *l*, where the tuples is 16000 and SA=5(sensitive attribute number=3. In order to have a comparison, we set *l* to be 3, 4, 5 and 7 respectively. As shown in Fig.8 MMDCF shows the green line is higher than the blue line. Especially, MMDCF dimension is larger when the suppression ratio of data is higher. From a global aspect, with an increase in *l*, the height of the suppression ratio increases gradually, showing that the suppression ratio has an increasing trend. This is because the larger the *l*, the higher the privacy requirements. As a result, ensuring the group *l*-diversity in each dimension is more difficult. Thus the overall effect of variation grouping becomes of inferior quality (the suppression ratio of data increases. Gradually, KACA produces significantly lower suppression ratio then the MMDCF.

Fig. 9 exhibits the dissimilarity of additional information loss with the help of KACA and MMDCF on various *l* values at the time n is 16000, however, we estimate additional information loss using the formula(3). As shown in Fig. 9 shows the red line is low additional information loss then the blue line. We conclude that MMDCF is better than the KACA. Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures

## 8. CONCLUSION

A Bucketization approach is suggested in this paper to anonymize multiple sensitive attributes on micro-data. We use the idea of clustering with MSB to develop our model and through our experiments we can see the results of the model itself. An example is demonstrated that shows this technique could keep security with multiple numerical sensitive attribute acceptably. It was shown by experimental results that the bucketization has low additional information loss and suppression ratio. We conclude that the process is a demanding issue by cause of an attacker may exploit the complex association between varieties of published accounts to raise our opportunity of breaching the privacy of a distinct. So in this paper we conclude that the bucketization has very less suppression ratio and additional information loss. It is better than the Achiving *k*-Anonymity by Clustering in Attribute Hierarchical Structures.

## REFERENCES

[1] L. Sweeney K-anonymity A model for protecting privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5), 2002, pp. 557-570.

[2] Machanavajjhala A, Gehrke J, Kifer D l-Diversity: Privacy beyond K-anonymity In: Proc. of the 22nd International Conference on Data Engineering. Atlanta: IEEE Computer Society, 2006, pp. 24-35.

[3] Lining-hui,LiTi an-cheng and Venkatasubramanian .S. t-Closeness: privacy beyond k-anonymity and l-diversity In: Proc. of the 23rd ICDE, 2007, pp. 106-115.

[4] Xiao .X, Tao Y. Anatomy Simple and effective privacy preservation In: Proc. Of the 32nd International Conference on Very Large Data Bases. Seoul: VLDB Endowment, 2006, pp. 139150.

[5] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In PODS, page 188, 1998.

[6] J. Li,Y. Tao and X. Xiao, Preservation of proximity privacy in publishing numerical sensitive data, in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, Vancouver, Canada, 2008, pp. 473486.

[7] S. Liu, H. Jin, and S. Ju, Privacy preserving technology for multiple sensitive attributes in data publishing, (inChinese), Application Research of Computers, vol. 28, no.6, pp. 5965, 2011.

[8] X. Yang, Y. Wang, B. Wang, and G. Yu, Privacy preserving approaches for multiple sensitive attributes in data publishing, Chinese Journal of Computers, vol. 31,no. 4, pp. 574587, 2009.

[9] Li Jiuyong, Wong Raymond Chi-Wing, Fu Ada Wai-Chee,et al Achieving k-anonymity by clustering in attribute hierarchical structure [C]. DaWak.LNCS4081, Springer verlag, Berlin, Heidelberg, 2006, pp. 405-416.

[10] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In the fifth Berkeley Symposium on Mathematical Statistics and Probability, pages 281297, 1967.

[11] Yufei Tao, Hekang Chen, Xiaokui Xiao, ShuigengZhou, Member, IEEE Computer Society, and DonghuiZhang. ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication IEEE Transaction on Knowledge and Data Engineering .2009, vol. 21.No.7. pp.1073-1087.

[12] T. Liu, W. Ni, Z. Chong, and Y. Zhang, Privacy preserving data publishing methods for multiple numerical sensitive attributes, (in Chinese), Journal of Southeast University (Natural Science Edition), vol. 40, no. 4, pp. 669703,2010.

[13] ZHEXUE HUANG huang@mip com.au ACSys CRC ,Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values Data Mining and Knowledge Discovery 2, 283304 (1998) c 1998 Kluwer Academic Publishers. Manufactured in the Netherlands.