International Journal of Engineering & Technology, 7 (2.20) (2018) 197-206

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Novel Approach for Personalized Privacy Preserving Data Publishing with Multiple Sensitive Attributes

S. Ram Prasad Reddy[1]*, KVSVN Raju[2], V. Valli Kumari[3]

[1]Department of Computer Science & Engineering, Vignan's of Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India.
[2]GVP College for Degree & PG Courses School of Engineering, Visakhapatnam, Andhra Pradesh, India.
[3] Department of Computer Science & Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.
**Email: reddysadi@gmail.com:*

Abstract

The Personalized Privacy has drawn a lot of attention from diverse magnitudes of the public and various functional units like bureau of statistics, and hospitals. A large number of data publishing models and methods have been proposed and most of them focused on single sensitive attribute. A few research papers marked the need for preserving privacy of data consisting of multiple sensitive attributes. Applying the existing methods such as *k*-anonymity, *l*-diversity directly for publishing multiple sensitive attributes would minimize the utility of the data. Moreover, personalization has not been studied in this dimension. In this paper, we present a publishing model that manages personalization for publishing data with multiple sensitive attributes. The model uses slicing technique supported by deterministic anonymization for quasi identifiers; generalization for categorical sensitive attributes; and fuzzy approach for numerical sensitive attributes based on diversity. We cap the belief of an adversary inferring a sensitive value in a published data set to as high as that of an inference based on public knowledge. The experiments were carried out on census dataset and synthetic datasets. The results ensure that the privacy is being safeguarded without any compromise on the utility of the data.

Keywords: *Anonymity; Categorical Sensitive attributes; Data Publishing; Diversity; Numerical Sensitive Attributes ; Personalized Privacy.*

1. Introduction

The continuous stream of digital information by innumerable segments like public sector units, corporate units, and individuals has facilitated knowledge discovery and information-based decision making. Publishing data for analysis, while maintaining individual privacy, has become a challenging task in today's day to day data. The prime objective is to limit the privacy disclosure risk to an acceptable level while maximizing the benefit due to publication of the data. The personalization perspective that takes into consideration the user's consent for publishing the data in also vital. The traditional approach of anonymization is to remove credential fields such as social security number and name. The universal anonymization approach is generalization, which is semantically consistent. As a result, more records will have the same set of quasi-identifier values by maintaining privacy to some extent as it provides misperception to recognize the value as it is anonymized.

In 2002, Sweeney[1] proposed the k-anonymity model for privacy protection where the corresponding attributes that leak information are suppressed or generalized so that, for every record in the modified table, there are at least k - 1 other records that have exactly the same values for the quasi identifiers. There are many successful applications [2, 3, 4] based on k-anonymity. However, while k-anonymity protects data against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation of k-anonymity, Machanavajjhala et al. [5] introduced a new notion of privacy, called *l*-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least *l* "well represented" values. Li et al.[6] proposed a novel privacy notion called *t* -closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions is no more than a threshold t). This effectively limits the amount of individual specific information that an observer can learn. In addition, several principles were introduced, such as (c; k)-safety [7] and δ_presence [8]. In 2006, Xiao and Tao [9] proposed Anatomy, which is a data anonymization approach that divides one table into two for release. One table includes the original quasi identifier and a group id, and the other includes the association between the group id and the sensitive attribute values.

Many methods have been proposed to ensure privacy. But, most of the methods focused on protecting privacy in the context of a single sensitive attribute. Few authors have also focused on privacy protection models for protecting data with multiple sensitive attributes. In real scenarios data comprises of more than one sensitive attribute which could be numerical or categorical or both. So it is necessary to study privacy preserving data publishing in the context of multiple sensitive attributes. Applying exiting methods such as *k*-anonymity, *l*-diversity in its true form would not ensure privacy and there is every possibility for breach of information.

In this paper, we address the problem of handling privacy for the static datasets consisting of multiple sensitive attributes. Besides this we also consider personalization, i.e., where the users' consent is taken into consideration, while publishing the data. We implement a novel privacy-preserving data publishing method for multivariate data sensitive attributes which uses both horizontal and vertical slicing along with sensitivity threshold. The sensitivity

threshold ensures that each categorical value appears only once in the group. The method also uses the '*k*' and '*l*' parameters.

2. Related work

The recent research work also concentrated on handling of privacy when the datasets consist of multiple sensitive attributes. Gal, Tamas S. et al [16] proposed a model for privacy preserving that protects identity of patients for data with multiple sensitive attributes. The authors assumed that when a distinct sensitive attribute value is deleted from a group, all rows containing that value will be deleted. A variant of this model is also proposed, which allows the user to specify a lower degree of diversity for attributes with very few distinct values. Experiments show that the proposed approach introduces distortion orders of lower magnitude than the distortions introduced by the existing approach in the literature, and introduces small relative error for random SQL queries. To preserve privacy in datasets with multiple sensitive attributes (MSAs) Ye et al [10] applied decomposition which selected one of its MSAs as primary sensitive attribute (PSA) subject to an $(l_1; l_2; \ldots; l_d)$-diversity privacy model which was enforced through noise addition. Das and Bhattacharyyu [11] observed that decomposition is not a dynamic publishing scenario, degrades data utility through noise addition, and enforces diversity on primary sensitive attributes. To address this drawback, Das and Bhattacharyyu [11] used decomposition+, which was dynamic with less data utility degradation. This technique not suitable for high-dimensional datasets is known to suffer from curse of dimensionality. In this scenario, tuples can be added even after anonymization. There is flexibility to add, remove or update tuples in multiple releases of the same the data.

Liu et al [12] used the new *k*-anonymity based on *l*-diversity where *k*-anonymized QID record was linked with *k*-number of sensitive attributes. The sensitive attributes are split into highly and lowly sensitive ones. The tuples are sorted according to according to amount of highly sensitive values first and then distributed to best equivalence classes one by one. The association among sensitive attributes values is destroyed to avoid attack. Han et al. [13] applied the SLicing On Multiple Sensitive (SLOM) and MSB-KACA algorithm based on *l*-diversity for privacy preservation of multiple sensitive attribute (MSA) of a dataset. The quasi-identifier values were generalized based on the k-anonymity principle, and the sensitive values were sliced and bucketized to satisfy the *l*-diversity requirement. This approach may lead to a large suppression ratio and information loss due to tuple suppression of sensitive attributes to enforce *l*-diversity on the one hand and the generalization of quasi-identifier attributes on the other. High data degradation may be the resultant trade-off for privacy preservation.

Liu et al [14] used the MNSACM method, which was based on clustering and multi-sensitive bucketization for anonymizing numerical multi-sensitive attributes of a dataset. The numerical sensitive attributes were placed in multiple groups such that every sensitive attribute corresponds to a single dimension of the multi-dimension bucket. This approach has not been implemented on a real dataset and an algorithm for it has not been proposed. Yi,T. and Shi,M [18] presented that an attack method uses the association rules to get the users' privacy and accordingly presented a protection model. Through theoretical and experimental analysis, the authors proved that the new protection model can provide better protection for privacy, and it was able to preserve useful relationships in released tables. Besides, in order to improve the efficiency of algorithm, the authors presented an improved SID creation method, and proved it is more effective with experiment.

Radha,D and Valli Kumari, V [19] suggested a bucketization approach to anonymize multiple sensitive attributes on micro-data. The authors used the idea of clustering with MSB to develop the model. The authors showed that the bucketization has low additional information loss and suppression ratio. They later concluded that the process is a demanding issue by cause of an attacker may exploit the complex association between varieties of published accounts to raise the opportunity of breaching the privacy of a distinct. S. A. Onashoga et al [20] introduced a new approach to anonymizing multiple sensitive attributes (MSAs) through the combination of the LKC-Privacy model, slicing technique and cell suppression; enhancing MSAs anonymization through dynamic and web-based features; and anonymizing MSAs with improved utility gain and reduced data degradation.

3. System Architecture

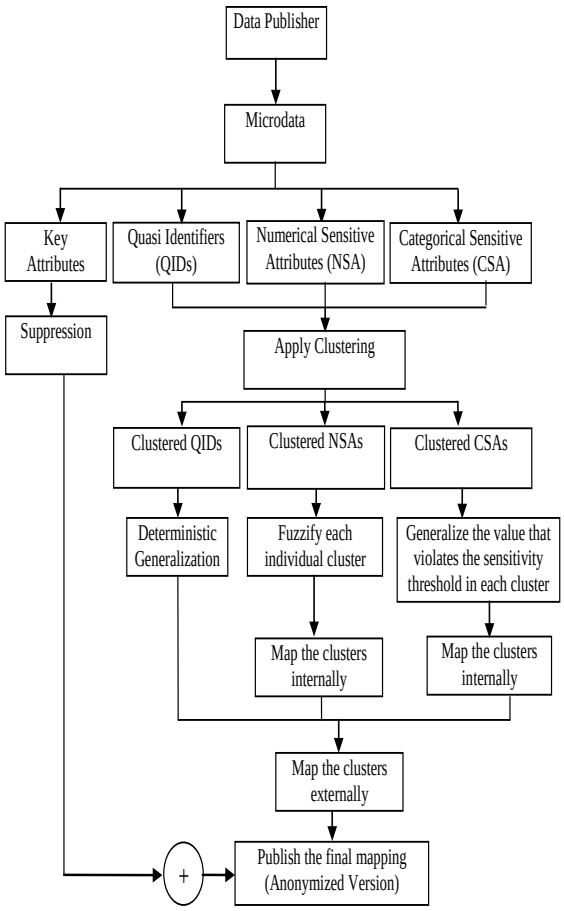

Fig. 1: System architecture

The proposed privacy preserving model primarily has two objectives: preserving privacy while revealing useful information for sensitive i) multiple numerical attributes, and ii) multiple categorical (non-numerical) attributes and to find a generalized table T', such that it includes all the attributes of T and an individual tuple from T is non-identifiable in T'. It also considers users' consent into account to address personalization. Let T be the micro data holding information about a set of individuals each associated with a tuple as shown in table1. Table 1 is composition of key identifiers, quasi-identifiers, numerical and categorical sensitive attributes. The basic intuition is to publish such kind of data without much loss of data and at the same privacy is not to be compromised. The overall system architecture is depicted in Fig. 1.

Table 1: Micro data with multiple sensitive attributes

PID	PNAME	AGE	GENDER	ZIPCODE	SALARY	BONUS	LOAN	EDUCATION	JOB	DISEASE

P01	James	40	Male	12411	90000	9000	108000	Bachelors	Tech support	Flu
P02	Alice	30	Female	22311	120000	12000	144000	Preschool	Sales	HIV
P03	Bill	25	Male	22629	60000	6000	72000	HS-grad	Exec-managerial	Malaria
P04	Bob	50	Male	42411	200000	20000	240000	Prof-school	Adm-clerical	Typhoid
P05	Williams	47	Male	42344	50000	5000	60000	Masters	Adm-clerical	HIV
P06	Henry	38	Female	12523	250000	25000	300000	Doctorate	Exec-managerial	HIV
P07	Angel	43	Female	42413	125000	12500	175000	Bachelors	Adm-clerical	HIV
P08	Small	60	Male	32266	70000	7000	84000	Preschool	Exec-managerial	Flu
P09	Mary	55	Female	32243	300000	30000	360000	HS-grad	Tech support	Diabetes
P10	Adam	49	Male	42512	58000	5800	69600	Masters	Exec-managerial	Cancer
P11	Mercy	39	Female	42123	76000	7600	91200	Bachelors	Sales	Diabetes
P12	Anil	31	Male	41234	45000	4500	54000	HS-Grad	Tech support	Cancer

Personalization is inducted into the publishable micro data by introducing a new attribute, privacy disclosure (PD) into the original micro data. The privacy disclosure is a user defined value. The value specifies whether to publish the tuple or not. The value PD considered is a Boolean value. The Boolean "TRUE (T)" value specified by the user signals that the user has given his consent to disclose the information after adhering to the privacy guidelines. The Boolean value "FALSE (F)" stipulates that the user is unwilling to disclose his/her data and is therefore suppressed. The table of data after introducing the privacy disclosure is shown in table 2. The last two tuples of table 2 consists of "FALSE" for privacy disclosure(PD). So these two tuples will be supressed while publishing the table of information.

Table 2: Micro data after introducing privacy disclosure

PD	PID	PNAME	AGE	GENDER	ZIPCODE	SALARY	BONUS	LOAN	EDUCATION	JOB	DISEASE
T	P01	James	40	Male	12411	90000	9000	108000	Bachelors	Tech support	Flu
T	P02	Alice	30	Female	22311	120000	12000	144000	Preschool	Sales	HIV
T	P03	Bill	25	Male	22629	60000	6000	72000	HS-grad	Exec-managerial	Malaria
T	P04	Bob	50	Male	42411	200000	20000	240000	Prof-school	Adm-clerical	Typhoid
T	P05	Williams	47	Male	42344	50000	5000	60000	Masters	Adm-clerical	HIV
T	P06	Henry	38	Female	12523	250000	25000	300000	Doctorate	Exec-managerial	HIV
T	P07	Angel	43	Female	42413	125000	12500	175000	Bachelors	Adm-clerical	HIV
T	P08	Small	60	Male	32266	70000	7000	84000	Preschool	Exec-managerial	Flu
T	P09	Mary	55	Female	32243	300000	30000	360000	HS-grad	Tech support	Diabetes
T	P10	Adam	49	Male	42512	58000	5800	69600	Masters	Exec-managerial	Cancer
F	P11	Mercy	39	Female	42123	76000	7600	91200	Bachelors	Sales	Diabetes
F	P12	Anil	31	Male	41234	45000	4500	54000	HS-Grad	Tech support	Cancer

The micro data (T) that the publisher wishes to publish comprises of key attributes (PID, PNAME), quasi-identifiers (QIDs – AGE, GENDER, ZIPCODE), multiple numerical sensitive attributes (NSAs – SALARY, BONUS, LOAN) and multiple categorical sensitive attributes (CSAs – EDUCATION, JOB, DISEASE). The data to be published is initially vertically partitioned into four independent tables namely (i) table of key attributes (T^I), (ii) table of quasi-identifiers (T^Q), (iii) table of numerical sensitive attributes (T^N) and (iv) table of multiple categorical sensitive attributes (T^C) as given in tables 3,4,5 and 6.

Table 3: T^I-Key attributes

PID	PNAME
P01	James
P02	Alice
P03	Bill
P04	Bob
P05	Williams
P06	Henry

P07	Angel
P08	Small
P09	Mary
P10	Adam

Table 4: T^Q-Quasi-identifiers

AGE	GENDER	ZIPCODE
40	Male	12411
30	Female	22311
25	Male	22629
50	Male	42411
47	Male	42344
38	Female	12523
43	Female	42413
60	Male	32266
55	Female	32243
49	Male	42512

Table 5: T^N-Numerical sensitive attributes

SAL-ARY	BONUS	LOAN
90000	9000	108000
120000	12000	144000
60000	6000	72000
200000	20000	240000
50000	5000	60000
250000	25000	300000
125000	12500	175000
70000	7000	84000
300000	30000	360000
58000	5800	69600

Table 6: T^C-Categorical sensitive attributes

EDUCATION	JOB	DISEASE
Bachelors	Tech support	Flu
Preschool	Sales	HIV
HS-grad	Exec-managerial	Malaria
Prof-school	Adm-clerical	Typhoid
Masters	Adm-clerical	HIV
Doctorate	Exec-managerial	HIV
Bachelors	Adm-clerical	HIV
Preschool	Exec-managerial	Flu
HS-grad	Tech support	Diabetes
Masters	Exec-managerial	Cancer

The identifiable attributes are to be suppressed. So, the table T^I is suppressed totally as it comprises of identifiable attributes. The remaining three tables T^Q, T^N, T^C are handled independently. The table T^Q is clustered horizontally based on similarity measure such that each group contains at least 'K' tuples. If any group consists of less than 'K' tuples it is merged with the closest similar cluster. For each group, deterministic generalization is applied.

Definition: (Deterministic generalization): This process of anonymization depends on multi-set based generalization. The values in each group are not generalized using taxonomies instead they are represented by sets. Each set comprises of elements of the cluster given by the frequency of each item as shown in table 8.

Table 7: T^Q-After clustering with k=2

Group ID	AGE	GENDER	ZIPCODE
GID1	30	Female	22311
GID1	25	Male	22629
GID2	40	Male	12411
GID2	38	Female	12523
GID3	50	Male	42411
GID3	47	Male	42344
GID3	43	Female	42413
GID3	49	Male	42512
GID4	60	Male	32266
GID4	55	Female	32243

Table 8: T^Q - After multi-set based generalization

Group ID	AGE	GEN-DER	ZIPCODE
GID1	25:1,30:1	Male:1, Female:1	22311:1,22629:1
GID1	25:1,30:1	Male:1, Female:1	22311:1,22629:1
GID2	38:1,40:1	Male:1, Female:1	12411:1,12523:1
GID2	38:1,40:1	Male:1, Female:1	12411:1,12523:1
GID3	43:1,47:1,49:150:1	Male:3, Female:1	42411:1,42344:1, 42413:1,42512:1
GID3	43:1,47:1,49:150:1	Male:3, Female:1	42411:1,42344:1, 42413:1,42512:1
GID3	43:1,47:1,49:150:1	Male:3, Female:1	42411:1,42344:1, 42413:1,42512:1
GID3	43:1,47:1,49:150:1	Male:3, Female:1	42411:1,42344:1, 42413:1,42512:1
GID4	55:1,60:1	Male:1, Female:1	32266:1,32243:1
GID4	55:1,60:1	Male:1, Female:1	32266:1,32243:1

The numerical sensitive attributes are generalized using the Fuzzification process [17] to bring the numerical values to linguistic terms. The cluster of values would be transformed to the linguistic term using Fuzzification process. To generalize the numerical sensitive attributes, the attributes are initially clustered into groups such that each group consists of distinct values. The process is carried out by initially finding the frequency of each element in the first attribute of the NSA set. This frequency helps us to fix the initial number of buckets required. The final number of buckets depends on the diversity (l) value. Now, the values are distributed into the respective clusters such that each cluster consists of distinct values. After distribution of the values, we shall now check if each cluster consists of a minimum of 'k' values. If all the clusters satisfy then it is checked for 'l' requirement. The clusters that neither satisfies 'k' nor 'l' requirement are merged with that cluster which still satisfies distinctness. The clusters of values se values are now fuzzified. The similar procedure is repeated for the remaining numerical sensitive attributes. After clustering, each group is fuzzified independently. Suppose that the numerical sensitive attribute, income, of table 5 is to be fuzzified. Then, the following procedure is employed to transform the cluster into a publishable form. We apply the following rule for numerical sensitive attributes for transforming its values. L is the linguistic term set with $\{l_1, l_2, l_3 \ldots\}$ as the linguistic values; NSA_i is the i^{th} sensitive variable in the numerical sensitive attribute (NSA) and 'n' is the number of linguistic values. We transform all tuples in T_N to T'_N.

Rule: Given L={l_1,l2,l3,...,l_n}, then

$$\forall t \in T'_N\ F(t.A_i^s) \rightarrow l, such\ that\ l \epsilon L$$

Suppose the linguistic term set for the variable income L(NSA=income) is: *{High, Medium, Low}* with membership functions defined as below. The minimum and maximum values of income according to the business organization rules are *min* and *max* respectively and $a_1, a_2, \ldots, a_k$ are the midpoints of each fuzzy set and k is the number of fuzzy sets. The k fuzzy sets will have ranges of: $\{min\text{-}a_2\}$, $\{a_1\text{-}a_3\}$, $\{a_{(i-1)}\text{-}a_{(i+1)}\}, \ldots, \{a_{(k-1)}\text{-}max\}$.

For fuzzy set with midpoints $a_1, a_2, a_3, \ldots a_{k-1}$, the membership function is given by f_1, f_2 & f_3 for *Low*, *Medium and High* respectively. For fuzzy set with midpoint a1, the membership function is given by

$f_1(x)$ = 1.0 if x=min

= (x- a_2) / (min-a_2) if x< a_2

$= 0$ if $x >= a_2$

For the fuzzy set with midpoint a_i, $2<=i<=k-1$, the membership function is given by

$f_i(x) = 0$ if $x<=a_{(i-1)}$

$= (x-a_{(i-1)})/(a_i-a_{(i-1)})$ if $a_{(i-1)}<x<a_i$

$= 1.0$ if $x= a_i$

$= (a_{(i+1)}-x)/(a_{(i+1)}- a_i)$ if $a_i<x< a_{(i+1)}$

$= 0$ if $x>=a_{(i+1)}$

For fuzzy set with midpoint a_k, the membership function is given by

$f_k(x) = 0$ if $x<=a_{(k-1)}$

$= (x- a_{(k-1)}) / (max- a_{(k-1)})$ if $x>a_{(k-1)}$

$=1.0$ if $x=max$

In the similar manner, the remaining numerical sensitive attributes are transformed using fuzzification process. The income attribute values of table 5 after applying the above transformations along with the values of weight (f1, f2, f3) are as given in table 9. This helps the end user of the data to make out the distinction between two attribute values, even though they are mapped to the same linguistic term. For instance, in table 9, both 50000 and 70000 are mapped to low. The relativeness (informativeness) is still maintained by the weight. The weight associated tells that low associated with 50000 is still lower than the low associated with 70000. The data in publishable form will have weight associated with every transformed value as in table 9. However, the Income attribute values in its original form are not published. This is how we claim informativeness in data while preserving privacy. The data is then given by internal association among the clusters as shown in Fig. 2.

Table 9: T^N – Transformed numerical sensitive attribute -income

INCOME	Weight	Transformed Value
50000	1	Low
58000	0.92	Low
60000	0.9	Low
70000	0.8	Low
90000	0.6	Low
120000	0.3	Low
125000	0.25	Low
200000	0.5	Medium
250000	0	Medium
300000	1	High

Table 10: T^N – Numerical sensitive attributes – after clustering

SALARY		BONUS		LOAN	
SB1	SB2	BB1	BB2	LB1	LB2
120000	60000	12000	6000	144000	72000
50000	200000	5000	20000	60000	240000
70000	250000	7000	12500	84000	300000
58000	125000	58000	30000	69600	175000
90000	300000	9000	25000	108000	360000

Table 11: Transformed numerical sensitive attributes

SALARY		BONUS		LOAN	
SB1	SB2	BB1	BB2	LB1	LB2
Low (0.3)	Low (0.9)	Low (0.3)	Low (0.9)	Low (0.3)	Low (0.9)
Low (1.0)	Medium (0.5)	Low (1.0)	Medium (0.5)	Low (1.0)	Medium (0.5)
Low (0.8)	Medium (0)	Low (0.8)	Medium (0)	Low (0.8)	Medium (0)
Low (0.92)	Low (0.25)	Low (0.92)	Low (0.25)	Low (0.92)	Low (0.25)
Low (0.6)	High (1)	Low (0.6)	High (1)	Low (0.6)	High (1)

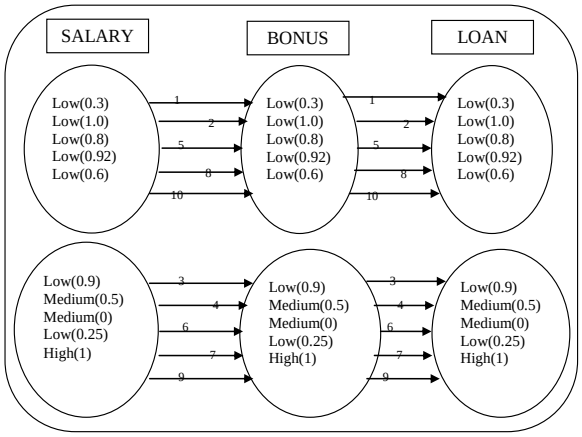

Fig. 2: Tuple associations among clusters of numerical sensitive attributes

The final set of attributes to be handled is categorical sensitive attributes. Each categorical sensitive attribute is independently handled. The initial number of buckets considered depends on the '*k*' value. The initial number of buckets required is defined by the frequency of each value. The maximum frequency of the item is considered to be the initial requirement. The final number of buckets is given by relies on '*l*' value.

Each attribute is handled independently. The first categorical sensitive attribute values are placed in each bucket such that the similar values are distributed into different buckets. We ensure that each bucket at the maximum contains distinct values. This is to ensure diversity within each group. The same is applied for the remaining categorical sensitive attributes. If all the clusters satisfy the '*k*' property `then the clusters are verified for '*l*' property. If any of the clusters does not satisfy the '*k*' requirement then that cluster is merged with one of the clusters. If any cluster violates the property then that cluster is merged with the cluster with cluster that also violates the property. If there is only cluster that is violating the property then the members of that clusters are distributed into the cluster that show maximum distinctness even after adding the new members. If all the clusters satisfy the 'k' property and one of the clusters is not satisfying the '*l*' property then the repeated value is generalized to next higher level using the taxonomy.

In this paper, we considered the sensitivity threshold as '1'. This implies that at any cluster at any point of time should contain only one occurrence of each value. Suppose that a cluster contains 6 values satisfying the requirements '*k*=5' and '*l*=5'. If the cluster contains duplicate values then we say that the values of the cluster are violating the sensitivity threshold. One of the occurrences of the duplicate value is transformed to next higher level taxonomy value. This is how the complete distinctness is obtained in each cluster. If all the members of the cluster are distinct then that cluster need not be generalized further. If the cluster contains common members then one occurrence of that value is left unchanged in the cluster and the remaining values are generalized to the next higher levels in the taxonomy so that all the values are distinct.

The similar procedure is employed for the remaining categorical sensitive attributes. After generalization, the clusters of each categorical sensitive attribute are mapped to the other clusters accordingly so that the mapping reveals the combination of the tuple.

After applying the process to each of the quasi-identifiers, numerical sensitive attributes and categorical sensitive attributes the clusters are mapped from QIDs -> NSAs -> CSAs. This would be the final publishable data.

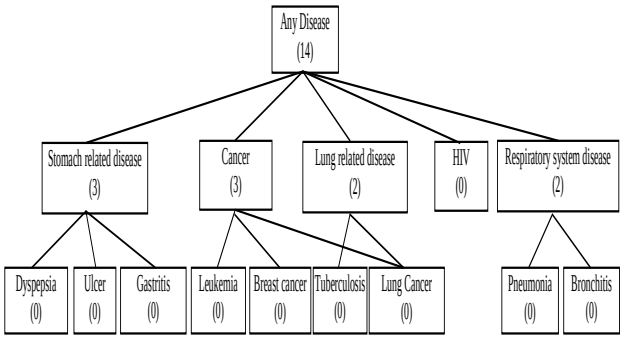

Fig. 3: Taxonomy for Disease Attribute

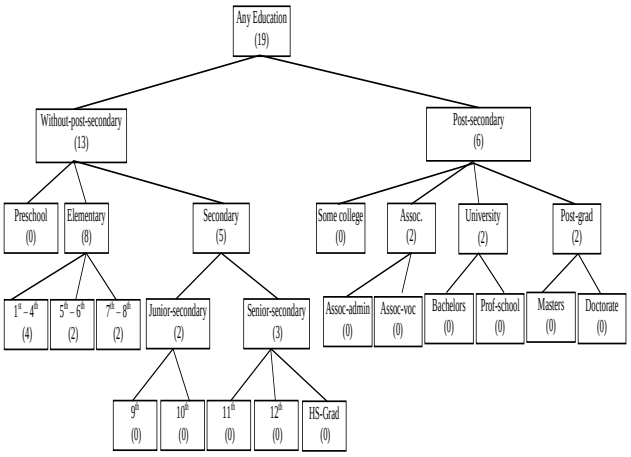

Fig. 4: Taxonomy for education

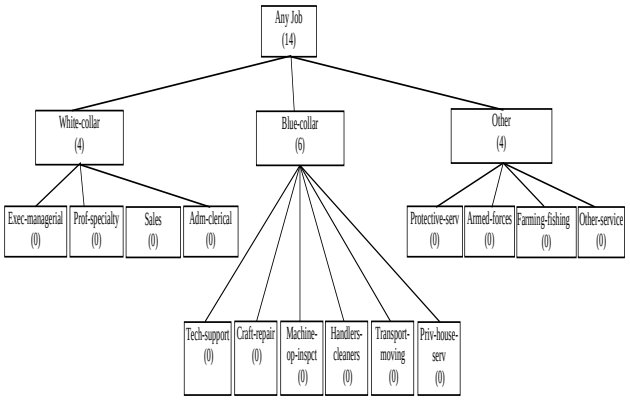

Fig. 5: Taxonomy for job

Table 12: T^C – Categorical sensitive attributes – after bucketization

EDUCATION			JOB			DISEASE		
EB1	EB2	EB3	JB1	JB2	JB3	DB1	DB2	DB3
Bachelors	Bachelors	Pre-school	Tech support	Sales	Exec-managerial	Flu	HIV	Malaria
Preschool	Prof-school	Masters	Adm-clerical	Adm-celrical	Adm-clerical	Typhoid	Flu	HIV
HS-Grad	Hs-Grad	Doctorate	Exec-managerial	Exec-managerial	Tech support	HIV	Cancer	Diabetes
Masters			Exec-managerial			HIV		

Table 13: Transformed categorical sensitive attributes

EDUCATION			JOB			DISEASE		
EB1	EB2	EB3	JB1	JB2	JB3	DB1	DB2	DB3
Bachelors	Bachelors	Pre-school	Tech support	Sales	Exec-managerial	Flu	HIV	Malaria
Preschool	Prof-school	Masters	Adm-clerical	Adm-clerical	Adm-clerical	Typhoid	Flu	HIV
HS-Grad	Hs-Grad	Doctorate	Exec-managerial	Exec-managerial	Tech support	HIV	Cancer	Diabetes
Masters			White-collar			Any Disease		

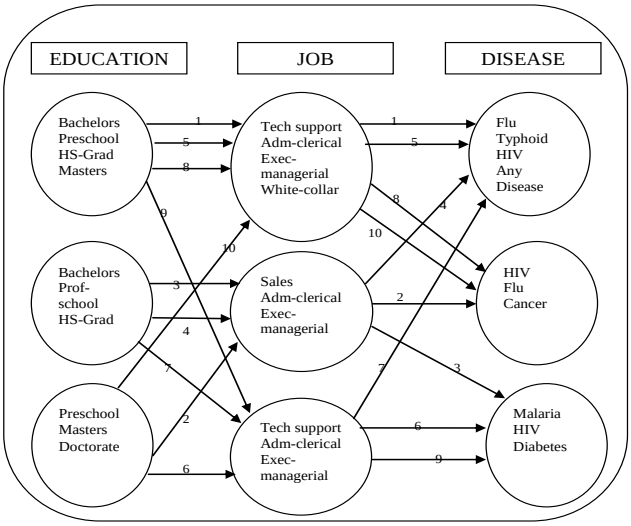

Fig.6: Tuple associations among clusters of categorical sensitive attributes

All the categorical sensitive attributes satisfy the parameter value l >=2. The overall publication of data involves the association of clusters of quasi-identifiers, numerical sensitive attributes and categorical sensitive attributes.

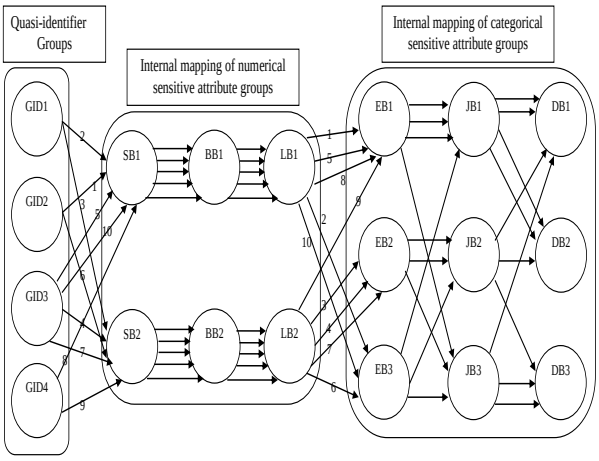

Fig. 7: Tuple associations among clusters of QIDs, NSAs and CSAs

3.1 Algorithm

Input: Dataset D
Anonymity parameter K
Output: Associative mapping of anonymized data

1. Partition the dataset vertically into four tables T^I, T^Q, T^N, T^C

{T^I consists of key attributes; T^Q consists of quasi-identifiers; T^N consists of Numerical sensitive attributes; T^C consists of Categorical sensitive attributes}
2. Suppress the key attributes in T^I
3. (a) Horizontally cluster the table T^Q based on similarity measure such that each cluster consists of a minimum of 'K' values
(b) Clusters not satisfying the anonymity requirement parameter 'K' are merged with another closest cluster
4. Apply deterministic generalization to anonymize each cluster of T^Q formed such that all the members in the group are identical
5. For each numerical sensitive attribute of T^N
(a) Horizontally cluster such that each cluster consists of distinct values and a minimum of 'K' values (The initial number of clusters formed depends on the most frequent item)
(b) (i) If only one cluster is not satisfying the 'K' parameter, the members of that cluster is distributed into other satisfying clusters
(ii) If there is more than one cluster not satisfying the 'K' parameter then the non-satisfying clusters are merged accordingly
(c) Each cluster of values is transformed into linguistic terms using the fuzzification process
6. End for
7. Internally map the clusters of T^N such that the combination defines a tuple
8. For each categorical sensitive attribute of T^N
(a) Horizontally cluster such that each cluster consists of distinct values and a minimum of 'k' values (The initial number of clusters formed depends on the most frequent item)
(b) (i) If only one cluster is not satisfying the 'k' parameter, the members of that cluster is distributed into other satisfying clusters uniformly
(ii) If there is more than one cluster not satisfying the 'k' parameter then the non-satisfying clusters are merged accordingly
(c) (i) Each cluster is now verified for 'l' requirement – The clusters of values that do not satisfy are generalized accordingly
(ii) Now each cluster is checked for distinctness property using the sensitivity threshold - if it is being violated then the repeated values are transformed to higher levels of taxonomy else the values remain unchanged
9. End for
10. Internally map the clusters of T^C such that the combination defines a tuple
11. Apply associative mapping among the clusters of T^Q, internal mapping of T^N clusters and internal mapping of T^C clusters
12. Publish the result (final associative mapping)

3.2 Transformation Ratio

This is defined as the number of tuples transformed during anonymization to publish the data. The overall suppression is given by

$$SR_i = \frac{|Transformed\ values\ of\ the\ i^{th}\ attribute|}{|Dataset|} \quad (1)$$

In this paper, the personalization is being taken into consideration. The tuple suppression takes place only when the individual would not like to disclose the information. This is specified by privacy disclosure value. When the privacy disclosure value is "FALSE (F)", then only the tuple is suppressed. In case if all the individuals would like to disclose the information then the suppression ratio is zero percent. The suppression ration is calculated only when the individual does not want to disclose the information. But here we compute the transformation ration of numerical and categorical sensitive attributes which is given by the equation 1.

3.3 Privacy Gain

Let T be the micro data table to be published. T contains 'd' attributes: A = {A_1,A_2, . . . ,A_d} and their attribute domains are {D[A_1],D[A_2], . . . ,D[A_d]}. A tuple t ∈T can be represented as t = (t[A_1], t[A_2], ... , t[A_d]) where t[A_i] ($1 \le i \le d$) is the A_i value of t. Table 1 is the sample micro data considered to illustrate the notion. The tuples are distributed into diverse clusters. A tuple partition consists of several subsets of T, such that each tuple belongs to exactly one subset. Specifically, let there be 'k' clusters C_1,C_2, ... ,C_k then $\cup_{i=1}^{k} C_i = T$ and for any $1 \le i_1 \ne i_2 \le k$, $C_{i_1} \cap C_{i_2} = \emptyset$.
The adversary is initially interested in finding the matching cluster. Once the matching cluster is identified, the adversary computes the conditional probability $P_i(SA|QI)$ that measures the strength of the link between QID and SA.

$$P_i(SA|QID) = \frac{\sum_{\langle q \in QID, a \in SA \rangle \in T'} \langle q, a \rangle}{\sum_{\langle a \in SA \rangle \in T'} \langle a \rangle} \quad (2)$$

The privacy gain metric depends on the diversity parameter and the generalization taxonomy for the categorical sensitive attribute. It depends on the fuzzified value for numerical sensitive attributes. The strength of privacy is given by the distribution of sensitive values for each equivalence class and the anonymization of quasi-identifier attribute values. During data publishing, the values of the QI are modified using multi-set based generalization.
Property 1. Given $t \in B$, $PG_{(t)} = A_{(t[QI])} + PG_{(t[SA])}$; where $A_{(t[QI])}$ is the anonymization of QI in a bucket, $PG_{(t[SA])}$ is the privacy gain of a value in SA and $PG_{(t)}$ is the privacy gain of the total tuple. The privacy gain for categorical sensitive attribute, CSA is given by

$$PG_{(t[CSA])} = \frac{1}{|CSA|} \sum_{S_j \in CSA} \frac{N_{S_j}}{N_T - 1} \quad (3)$$

where N_{s_j} is the no of descendants the node has in the sub-tree which is given in parenthesis for each node in the taxonomy in Fig. 3 and N_T is the total number of nodes in the taxonomy which is given in parenthesis in the root of the taxonomy in Fig. 3. '0' specifies that no privacy gain is being provided and '1' implies complete privacy gain.
Property 2. $\forall_{q_i \in QI, S_j \in SA} PG_{(q_i)} \in [0,1] and\ PG_{(S_j)} \in [0,1]$.
The privacy gain for numerical sensitive attribute, NSA is defined by the fuzziness property. This is given by the weight of the value along with linguistic term. The linguistic term specifies uncertainty. So, if a value belongs to all the sets with different weights, the lower weight specifies a higher privacy gain. The privacy gain cannot be measured exactly and it would be only an approximation.

3.4 Utility Gain

The utility gain is the percentage of information that is being provided to the beneficial user. The information related to quasi-identifiers is being published using multi-set based generalization. The data is not being modified but it is being replicated in the set format across the equivalence class or the cluster. This clearly addresses that information is being published in its original format.
The numerical sensitive attributes are generalized using fuzzification process. The clusters of numerical sensitive attributes are replaced with linguistic terms. The linguistic term along with the weight specifies the belongingness of the data to the fuzzy set. The linguistic term along with the weight '1' specifies that it completely belongs to that set and the utility if the data is 100%. The linguistic term along with the weight '0.1' specifies that the value belongs to the set only with a 10% assurance and the utility of the data falls below expectations. So the weight plays a vital role in measuring the utility of the data while the data is being published. The weight only specifies an approximation of the utility gain.
The categorical sensitive attributes are generalized using the predefined taxonomy trees. The categorical sensitive attributes are

generalized only if the sensitivity threshold is violated for the respective cluster. The clusters are initially checked for diversity. If the diversity is satisfied then the cluster is checked for sensitivity threshold violation. The sensitivity threshold considered is '1' in this scenario which implies that only occurrence of the value should be present in the cluster. If the value appears more than once in the cluster then one of the occurrences is replaced by the more generalized term by moving up the taxonomy. This is given by the equation

$$UG_{t[CSA]} = \frac{1}{|CSA|}\sum_{s_i \in CSA} \frac{H_{s_i}}{H_T - 1} \quad (4)$$

where H_{s_j} is the height of the generalized node that is, the sub-tree; H_T is the height of the tree and s_j is the j^{th} value to be generalized.

4. Results & Analysis

The experiments were performed on an Intel i5 processor machine with 8 GB of RAM. The operating system on the machine was Microsoft Windows 10. The implementation of the method was built and run in python and the graphs were drawn in RStudio. The dataset used in our experiments was the adult census dataset from the Irvine machine learning repository [15], since this dataset was the closest to a common k-anonymization benchmark that we are aware of. The actual dataset consists of 14 attributes with 48442 tuples. It has missing values also. This dataset used for result analysis consists of 8 attributes and 30,162 records. These are age, work class, education, occupation, relationship, capital-gain, capital-loss, and gender. Records with missing values are discarded because of limitations in our prototype system. The table structure is defined in Table 14. As our results are to be verified for datasets comprising of multiple sensitive attributes, synthetic dataset is created accordingly.

To analyze our model in terms of computational effort, we have implemented the model with multiple sensitive attributes. These programs have been tested by using the dataset that was taken from UCI Machine Learning Repository and the synthetic dataset. To analyze the computational effort, we have considered datasets of different sizes.

Table 14: Attributes for adult census dataset

Attribute	Type	# leaves
Age	Continuous	17-90
Work class	Categorical	8
Education	Categorical	16
Occupation	Categorical	14
Relationship	Categorical	6
Capital-gain	Continuous	10-45
Capital-loss	Continuous	10-45
Sex	Categorical	2

As we are dealing with personalized privacy, the user consent is also taken into consideration. Fig. 8 presents the computational effort for different sizes of datasets with 20%; 50%; and 100% privacy disclosure. 100% privacy disclosure implies that every user has given the consent to publish the data whereas 50% privacy disclosure specifies that only 50% of the users have no objection in releasing their data. It is evident from the Fig. 8 that the time complexity with respect to 100% privacy disclosure significantly varies as all the users have given their willingness to publish the data and needs a lot of transformation. Fig. 9 highlights the computational complexity for different values of 'k' with 100% privacy disclosure.

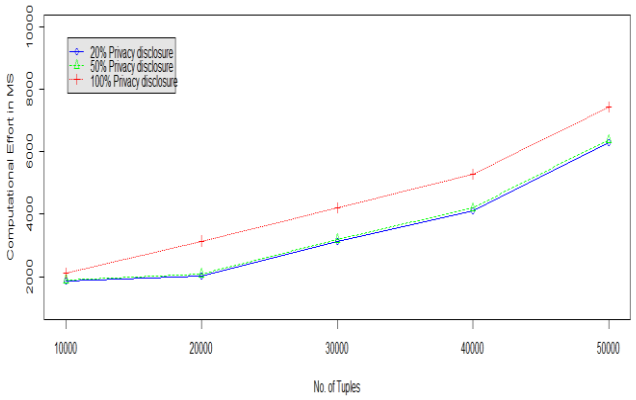

Fig. 8: Computational effort with different privacy disclosure levels

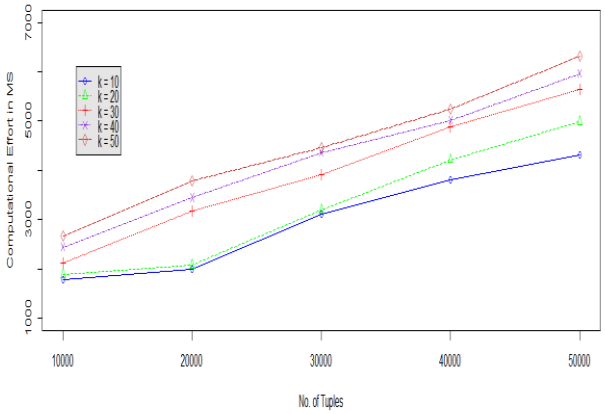

Fig. 9: Computational effort for different sizes of datasets for k=10,20,30,40,50 with 100% privacy disclosure

Fig. 10 and Fig. 11 illustrate the utility gain and privacy gain for the model constructed. It is clearly evident from the graphs that utility gain and privacy gain are more or less balanced eventually. In both the cases the average utility gain privacy gain are above 80%. Fig. 12 and Fig. 13 demonstrate the transformation percentage for numerical and categorical sensitive attributes for each of the attribute independently. The transformation percentage is less than 8% for both numerical and categorical sensitive values which ensures that most of the data remains unchanged.

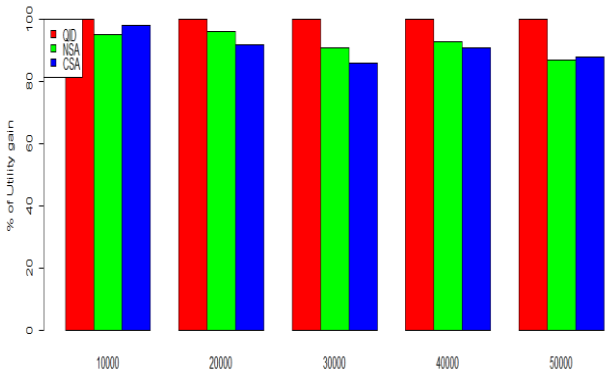

Fig. 10: Utility gain - QIDs, NSAs, CSAs for different sizes of dataset

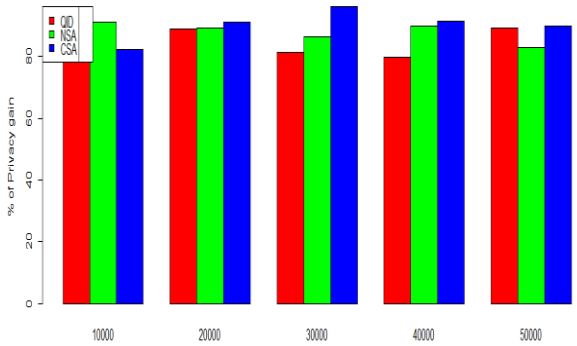

Fig. 11: Privacy Gain - QIDs, NSAs, CSAs for different sizes of dataset

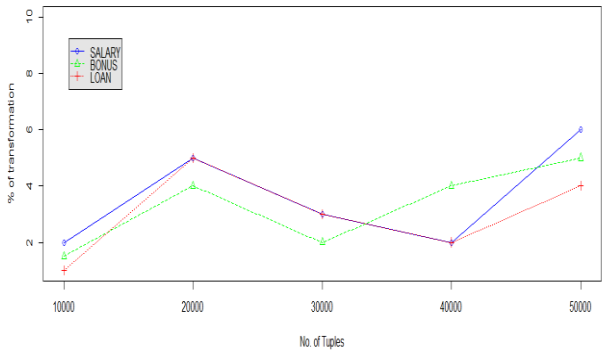

Fig. 12: Percentage of transformation with respect to each individual attribute of NSAs

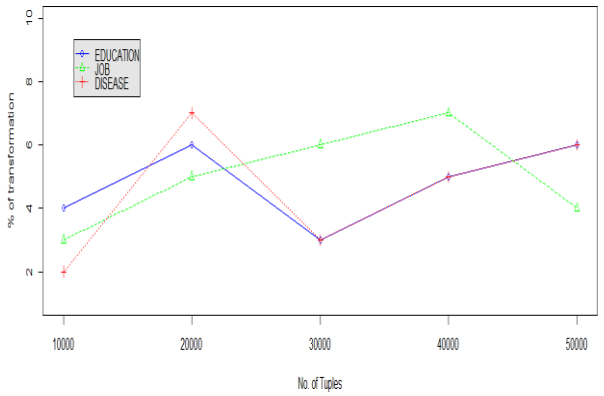

Fig. 13: Percentage of transformation with respect to each individual attribute of CSAs

5. Conclusions and Future Work

This model brings out a practical problem of maintaining anonymity against datasets with multiple sensitive attributes and proposes an effective solution. Maintaining anonymity against datasets with multiple sensitive attributes is an important and practical problem as we cannot always go with an assumption that datasets contain only one sensitive attribute. Although good progress on some scenarios have been made in [10, 11, 12, 13, 14, 16, 18, 19, 20] and this paper, the problem still at large remains open and challenging. All these paper have addressed the problem of multiple sensitive attributes but not in the context of personalization. We have made an attempt to provide a simple solution using slicing technique for maintaining privacy in datasets with multiple sensitive attributes. We have applied deterministic anonymization for quasi-identifiers. For both numerical and categorical sensitive attributes, the clustering is applied based on diversity. Fuzzification is considered for providing privacy to numerical sensitive attributes. Taxonomy based generalization is applied for categorical sensitive attributes by considering the sensitivity threshold. It is assumed that each categorical group should contain only occurrence of the value. A duplicate value is represented by the higher level taxonomy value. The numerical and categorical sensitive buckets are internally mapped. The final publishing set is externally mapped by considering the quasi-identifiers, numerical mapping and categorical mapping. The proposed model maintained privacy as well as utility.

This work motivates several directions for future research. First, in this paper, we consider a static dataset with multiple sensitive attributes. An extension is the notion of considering the dataset at fly. We may also consider web based data publishing of multiple sensitive attributes with secured access and authorization. Privacy-preserving data mining of datasets with multiple sensitive attributes can also be considered in this dimension. There is still a need to standardize the privacy and utility metrics.

Acknowledgement

I sincerely thank my supervisors for their continuous support. I also thank the anonymous referees for their careful reading of the paper and their valuable comments that significantly improved its quality.

References

[1] L. Sweeney, “k-anonymity: A model for protecting privacy”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, (2002), pp. 557–570, http://dx.doi.org/10.1142/S0218488502001648.

[2] K. Stokes and V. Torra, “n-confusion: A generalization of k-anonymity”, *Proceedings of the 2012 Joint EDBT/ICDT Workshops, ACM*, (2012), pp. 211–215, https://dl.acm.org/citation.cfm?id=2320824.

[3] J. Liu and K.Wang, “Enforcing vocabulary k-anonymity by semantic similarity based clustering”, *Proceedings of the 2010 IEEE 10th International Conference on Data Mining*, (2010), pp. 899–904, http://dx.doi.org/10.1109/ICDM.2010.59.

[4] C. Wang, L. Liu, and L. Gao, “Research on k-anonymity algorithm in privacy protection”, *Advanced Materials Research*, Vols. 756-759, (2013), pp. 3471–3475, https://doi.org/10.4028/www.scientific.net/AMR.756-759.3471.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “*l*-diversity: Privacy beyond k-anonymity”, *ACM Transactions on Knowledge Discovery from Data, Vol.1, No. 1,* (2007), pp. 1–47, http://dx.doi.org/10.1145/1217299.1217302.

[6] N. Li, T. Li, and S. Venkatasubramanian, “t -closeness: Privacy beyond k-anonymity and l-diversity”, *Proceedings of the IEEE 23rd International Conference on Data Engineering* (2007), pp. 106–115, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4221659&isnumber=4221635.

[7] D. J Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y Halpern, “Worst-case background knowledge for privacy-preserving data publishing”, *Proceedings of the IEEE 23rd International Conference on Data Engineering*, (2007), pp. 126–135, http://doi.ieeecomputersociety.org/10.1109/ICDE.2007.367858.

[8] M. E. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases”, *Proceedings of the 2007 ACM International Conference on Management of Data*, (2007), pp. 665–676, https://doi.org/10.1145/1247480.1247554.

[9] X. Xiao and Y. Tao, “Anatomy: Simple and effective privacy preservation”, *Proceedings of the 32nd International Conference on Very Large Data Bases*, (2006), pp. 139–150, https://dl.acm.org/citation.cfm?id=1164141.

[10] Ye, Y., Liu, Y., Lv, D., & Feng, J., “Decomposition: Privacy preservation for multiple sensitive attributes”, *Database Systems for Advanced Applications, Lecture Notes in Computer Science, Springer*, Vol. 5463, (2009), pp. 1-15, https://doi.org/10.1007/978-3-642-00887-0_42.

[11] Das, D., & Bhattacharyyu, D. K., “Decomposition+: Improving *l*-diversity for Multiple Sensitive Attributes”, *Advances in Computer Science and Information Technology, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer*, Vol. 85, (2012), pp. 1-10, https://doi.org/10.1007/978-3-642-27308-7_44.

[12] Liu, F., Jia, Y., & Han, W., “A new K-anonymity algorithm towards multiple-sensitive attributes”, *Proceedings of the IEEE 12th International Conference on Computer and Information Technology* (2012), pp. 768-772, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6391995&isnumber=6391864.

[13] Han, J., Luo, F., Lu, J., & Peng, H., “SLOMS: A privacy preserving data publishing methods for multiple sensitive attributes micro data”,

Journal of Software, Vol. 8, No. 12, (2013), pp. 3096-3104, https://doi.org/10.4304/jsw.8.12.3096-3104.

[14] Liu, Q., Shen, H., & Sang, Y., "Privacy-preserving data publishing for multiple numerical sensitive attributes", *Tsinghua Science and Technology*, Vol. 20, No. 3, (2015), pp. 246–254, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7128936&isnumber=7128931.

[15] Dua, D. and Karra Taniskidou, E., "UCI Machine Learning Repository", *Irvine, CA: University of California, School of Information and Computer Science*, (2017) , http://archive.ics.uci.edu/ml.

[16] Gal, Tamas S., Zhiyuan Chen, Aryya Gangopadhyay, "A privacy protection model for patient data with multiple sensitive attributes", *International Journal of Information Security and Privacy*, *Vol. 2, No. 3*, (2008), pp. 28-44, http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jisp.2008070103.

[17] V. Valli Kumari, S. Ram Prasad Reddy, M. Aruna Sowjanya, B. Jhansi Vazram, KVSVN Raju, "A novel approach for privacy preserving publication of data", *Proceedings of the 2008 International Conference on Data Mining*, (2008), pp. 506-512, https://dblp.org/rec/bib/conf/dmin/VallikumariRSVR08.

[18] Yi,T. Shi,M., "Privacy protection method for multiple sensitive attributes based on strong rule", *Mathematical Problems in Engineering* (2015), Vol. 2015, pp. 1-14, http://dx.doi.org/10.1155/2015/464731.

[19] Radha,D Valli Kumari, V., "Bucketize: protecting privacy on multiple numerical sensitive attributes", *Advances in Computational Sciences and Technology*, Vol. 10, No. 5, (2017), pp. 991-1008, https://www.ripublication.com/acst17/acstv10n5_32.pdf.

[20] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale & J. A. Oguntuase , "KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes", *Information Security Journal: A Global Perspective*, *Vol 26, No.3*, (2017), pp. 121-135, https://doi.org/10.1080/19393555.2017.1319522.