Categorical Grading Based Personalized Privacy Preservation Against Attacks

E. Poovammal and Dr. M. Ponnavaikko, Senior IEEE Member

***Abstract*— It is often necessary to publish personal information for research purposes. Re-identification is a major privacy threat to data sets which contain personal sensitive information such as income, a numerical data type and disease, a categorical type. Algorithms such as K-anonymity, L-diversity leave all the sensitive attributes and apply generalization and suppression to the quasi identifiers. So the available truthful data provide good utility for data mining tasks but, here perfect privacy can not be claimed. Privacy can be achieved only by separating Quasi Identifier attributes from sensitive attributes. Then utility of the table gets reduced. Our aim is to design an algorithm that improves both privacy and utility. The numerical sensitive attribute values are protected from proximity attack and the categorical sensitive attribute values are protected from divergence attack by our algorithm. Our experiments which are conducted on Adult data set, proved the improved utility and privacy than the previous methods.**

***Index Terms*— Anonymity, Personalization, Privacy Preservation, Taxonomy.**

I. INTRODUCTION

A vast amount of information and operational data have been stored by different vendors and organizations. Most of the stored data is useful only if the researcher is allowed to analyze them. For example, a hospital may release Patients' diagnosis records so that researchers can study the characteristics of diseases. The raw data or micro data is stored in a trustable server. The server releases the data in such a way that personal privacy is protected while allowing effective mining. Even though the objective of statistical database is to protect confidentiality of any individual, it provides the researchers only with aggregate statistics [1]. Randomization method is a technique of adding noise to the original data. Only the perturbed values and the distribution function used for perturbation are given for analysis. It is impossible to reconstruct the exact distribution of original data, and the accuracy level in estimating the data distribution is sensitive to reconstruction algorithm [2]. But the micro data records contain the actual unperturbed data associated with the individuals to support the development of new data mining algorithms while satisfying legal requirements. Micro data records may contain identifying attributes, sensitive attributes, quasi identifier attributes and neutral attributes. The identifying attributes like Name, Social Security Number are not released to protect privacy. Some of the attributes like Disease, Income are the sensitive attributes whose actual values should not be published without individual's concern. There may be other attributes like age, sex, zip-code which in combination (called quasi identifiers) can be used to recover the personal identities (linking attack problem). Personal identity leads to sensitive attribute disclosure. Micro data may contain neutral attribute like Length of_stay_hospital which is neither sensitive nor quasi identifying attribute. Micro data can only allowed to be released when the individuals are unidentifiable.

II. GENERALIZATION METHODS

Many privacy preserving algorithms rely on generalization and suppression of quasi identifier attributes. K-anonymity [11] [12] is the first effective approach applied to produce K identical tuples within the quasi identifier attributes and form equivalence classes in the table. Since K-anonymity places no constraint on the sensitive attributes in each equivalence class, it may result in homogeneity attack, which allows an adversary to derive actual sensitive information with 100% confidence. Also, anonymization process is inefficient in terms of computational cost, and optimal k-anonymization is NP hard [2]. Homogeneity attack is the motivation of L-diversity [8]. L-diversity principle demands at least 'L' well represented sensitive attribute values in each equivalence class. The (alpha – k) anonymity principle [6] is developed by combining L-Diversity and K-anonymity principles. It requires only alpha% of tuples can carry identical sensitive attribute values in every equivalence class of minimum size k. But, m-invariance principle requires that at least 'm' different sensitive attribute values in every equivalence class of size 'm' [14]. Privacy Level can be increased by increasing the value of 'm', which leads to high information loss.

The privacy preserving transformation of micro data is referred to as recoding. Multidimensional and local recoding methods reduce the amount of generalization [4] [10] on numerical and ordinal quasi identifier attributes. Categorical QI attributes are handled using suppression model by [3]. It is argued in [6] that a generalization nearer to the root of the hierarchy distorts a value more than a generalization further away from the root. Their recoding method based on distance metric, produced higher quality k-anonymity table but with high inconsistency measure compared to global recoding method.

Manuscript received February 28, 2009.

E. Poovammal is with the Department of Computer Science and Engineering, SRM University, Chennai, India (e-mail: epsrm@yahoo.com).

Dr. M. Ponnavaikko is the Vice Chancellor, Bharathidasan University, Trichy, India (e-mail: ponnav@gmail.com).

III. MOTIVATION

Consider Patients' database i.e., table I, which is released by the trusted server, maintained in the hospital. To preserve personal identity, attribute 'Name' is not released. And table II i.e. voters' database which has no sensitive values, is also released. By linking the values of attributes Age, Sex and the Zip code of these two tables, an adversary may get the sensitive information that Barbie is affected with stomach cancer (Linking attack). In K-anonymity method, the values of quasi-identifier (QI) attributes of each tuple in a table are identical to those of at least (k-1) other tuples (Equivalence class). The larger the value of K, the greater is the implied privacy since no individual can be identified with probability exceeding 1/K through linking attack. But k-anonymity only prevents association between individuals and tuples instead of association between individuals and sensitive values. Consider table III which is derived from table I by 3-anonymization. An adversary, who has QI values of Girija, gets her sensitive information i.e. 'Bronchitis' with 100% confidence. Homogeneity attack is solved by k-anonymous L-diversity principle [8]. Table IV is 3-anonymous, 2-diverse table which is derived from table I. Even from the table IV, an adversary gets the actual disease of Girija with the probability (p) of 50%, since it has two diverse values in the equivalence class. But in real world, she will be linked with other disease, stomach cancer with the probability of 50 %(1-p). This condition is worse than releasing her actual disease. We name this problem as divergence breach which occurs in categorical sensitive attributes. From Table IV, an adversary concludes that Girija's Income is in the interval of [10000 – 10030] with 75% confidence, even though he/she gets the actual value with 25% confidence. This problem which occurs in numerical sensitive attribute is proximity attack. Our task is to protect the sensitive values from both divergence and proximity attack.

IV. CONTRIBUTION

This paper presents a novel approach for privacy preservation in a micro data table, containing both numerical and categorical sensitive attributes. The core of our solution is the concept of personalized transformation of categorical sensitive attributes i.e. an individual can specify the degree of protection for their categorical sensitive values and categorical grading transformation of numerical sensitive attribute. We argue that if the sensitive attribute values are disclosed with the individual's concern, there is no need to do anonymization on QI attributes. We transformed the numerical sensitive values in such a way that the order and rank is maintained, while actual information is not leaked to the adversary.

A. Notations and Definitions

Let T be a relation storing private information about a set of individuals. T= $\{t_1, t_2, t_3, \ldots t_n\}$. Each t_i is a tuple of attribute values representing some individual records. Let A= $\{a_1,a_2,\ldots a_m\}$ be a set of attribute in T. $t[a_i]$ represents the value of attribute a_i for tuple t. A can be classified into four categories: Identifying Attributes A^i, Sensitive Attributes A^s, Quasi Identifying attributes A^q and Neutral Attributes A^n.

Table I - Patients' Database

Age	Sex	Zip Code	Income	Disease
33	M	600018	22000	Flu
29	F	600008	15000	Stomach Cancer
21	M	600006	10000	Bronchitis
31	M	600009	20000	Gastritis
22	M	600006	10020	Bronchitis
60	M	600019	23000	Flu
25	F	600006	10030	Bronchitis

Table II - Voters' Database

Disease	Age	Sex	Zip Code
Anand	33	M	600018
Barbie	29	F	600008
Charles	21	M	600006
Dinesh	31	M	600009
Esra	22	M	600006
Febi	60	M	600019
Girija	25	F	600006

Table III - Patients' Database- 3-Anonymous

Age	Sex	Zip Code	Income	Disease
21-25	Person	600006	10000	Bronchitis
21-25	Person	600006	10020	Bronchitis
21-25	Person	600006	10030	Bronchitis
29-60	Person	600008 – 600019	22000	Flu
29-60	Person	600008 – 600019	15000	Stomach Cancer
29-60	Person	600008 – 600019	20000	Gastritis
29-60	Person	600008 – 600019	23000	Flu

Table IV - Patients' Database- 3-Anonymous, 2-Diverse

Age	Sex	Zip Code	Income	Disease
21-29	Person	600006 – 600008	10000	Bronchitis
21-29	Person	600006 – 600008	10020	Bronchitis
21-29	Person	600006 – 600008	10030	Bronchitis
21-29	Person	600006 – 600008	15000	Stomach Cancer
31-60	Person	600009 – 600019	15000	Flu
31-60	Person	600009 – 600019	20000	Gastritis
31-60	Person	600009 – 600019	23000	Flu

Proximity Breach: It is a privacy threat specific to numerical sensitive attribute. It occurs when an adversary concludes with high confidence that the sensitive value of individual must fall in a short interval even though the adversary may have less confidence about his/her actual value [7].

Divergence Breach: It is a privacy threat specific to categorical sensitive attribute. It occurs when an adversary concludes with high confidence that the sensitive value of individual is totally irrelevant even though the adversary may have less confidence about his/her actual value.

Our objective is to publish a table T' derived from T containing all the attributes except A^i and all the tuples in T in such a way that T' possesses as much privacy and information as possible and deprived from both proximity and divergence attack.

B. Personalized Privacy

Personalized privacy is a method by which the individual gives his/her preference or decides the categorical sensitive information to be published. The individual is given three choices. Choice 1 is to be chosen by the individual who is willing to release actual value for research purposes. In this case we get the best data for analysis. Choice 2 is to be chosen by the individual who is not willing to publish actual but doesn't mind giving generalized value. In this case we get better data for analysis. Choice 3 is for the people who are not willing to publish even generalized value. In this case, the actual values are to be recoded with alias names so that data is useful for analysis but actual values can not be guessed.

1) Taxonomy tree with alias Names

For any categorical sensitive attribute Acs, the taxonomy tree is published by the domain experts along with micro data. All the leaf nodes are sensitive values in the table T. In our approach, all the leaf nodes are given two different alias names alias1 and alias2 as shown in fig. 1.These alias names are not to be published but to be preserved in trusted server with the table T. If the choice is 1, then the actual values are to be transferred to the table T'. If the choice is 2, then the actual values are to be replaced with alias1 otherwise with alias2.

C. Graded Grouping

Although various generalization principles such as, L-Diversity [8], T-closeness [10] deals with numerical sensitive values to preserve privacy, they failed to protect actual values from proximity attack. For generalization of numerical attribute, grouping methods were followed in some previous works. Our approach to sensitive numerical attribute is graded grouping as shown in fig. 2.To convert the actual values into a new form the following steps are followed. First step is to fix the number of categories (k) for the given range. Second step is for each category C1 …Ck, the max and min value is to be fixed in such a way that non overlapping continuous range results. Third step is to fix the category (Ci) for each actual value and find the membership value m(x) using

$m(x) = 0.0$ if $x = \min(Ci)$

$= (x - \min(Ci))/(\max(Ci) - \min(Ci))$ if $\min(Ci) > x < \max(Ci)$

$= 0.999$ if $x = \max(Ci)$

The fourth step is to replace the actual value with a new value obtained by adding category number and the membership value.

1) Algorithm for graded grouping

Function Group_grade(Ans)
Input: n records of numerical data type (actual values)
Output: n records of numerical data type (transformed values)

```
1. Get the value of k \\ Number of categories
2. For i= 1 to k
   Get min(Ci) and max(Ci) \\ fix range for each category
3. For j= 1 to n \\ number of records = n
      Let i=1
      Do while i< k
          If min(Ci) ≤ X[j] ≤ max(Ci)
                CX[j] = i
             If X[j] = min(Ci)
              MX[j]=0.0
                 Else
                  If X[j] = max(Ci)
                       MX[j]=0.999
                          Else
            MX[j] = X[j]-min(Ci)/ max(Ci)-min(Ci)
             NX[j] = MX[j]+CX[j]
             i=k
          Endif
       Else
         i++
     Endif
Endfor
```

D. Quasi Identifier Attributes

Quasi identifier Attributes A^q is a set of attributes A^{q1} to A^{qd} whose actual values are able to fetch a unique record in the table T. This property of Quasi Identifier attributes leads to the problem of linking attack which discloses the sensitive information. But in our approach the sensitive categorical attributes are having the values as per the user choice. Also the sensitive numerical attributes are graded to maintain the rank, but the actual values are not published. So, the linking attack can not disclose any sensitive information.

E. Privacy Attack

In any K-anonymized table the privacy level increases with increase in K- value. But, increase in k-value increases the information loss. The improved K-anonymity method like, L-diversity, T-closeness even though try maintaining some diversity in the sensitive values, they suffer from proximity and divergence breach as defined in section IV. Our new definition of privacy breach is not only leaking the actual sensitive information of the individuals but also leaking the close information and also linking the individuals with totally irrelevant information. Both proximity and divergence breach can not be solved if there exists an equivalence class. In our approach, since we are not anonymizing, no equivalence class exist and hence proximity and divergence breach is totally eliminated. Since there is no generalization, there is no information loss. Whatever may be the mining techniques and tools, they can be applied to the transformed table T' without any modification.

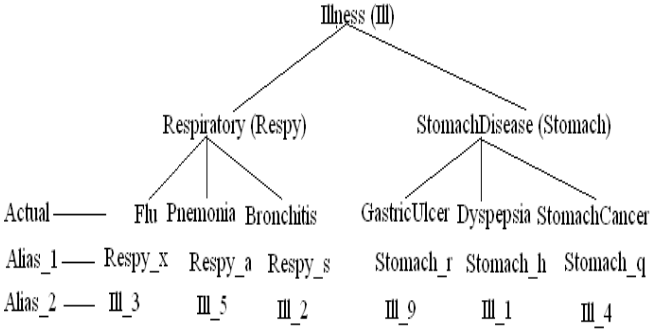

Fig. 1 Taxonomy Tree with Alias Names

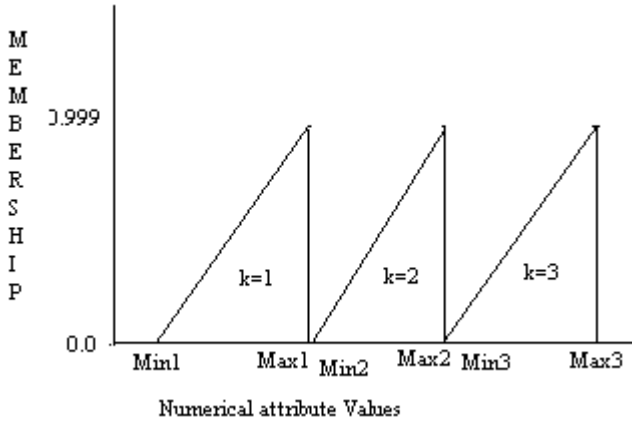

Fig. 2 Graded grouping

The table T' is generated from the original table T, by the algorithm explained in this section and maintains the structure of T. The only limitation in our approach is that both the adversary and the researcher can not get the actual values directly from the data table. For example, if the association rules or classification rules framed from the table T' has the transformed values with high confidence and support, and then the researcher can get the actual values from the trusted server. The server should check for the authentication of the researcher and interpret the actual results from the transformed values.

1) Privacy Preserving Algorithm

Input: Table T with 'n' tuples
Output: Table T' with 'n' tuples in which individuals are unidentifiable.
1. All the attributes A^1 to A^m are categorized into four groups A^i, A^q, A^s and A^n (Refer section IV A)
2. T= T- A^i \\Suppressing identifying attribute
3. A^s is categorized into A^{ns} and A^{cs} \\ Numerical and Categorical sensitive attribute
4. For i= 1 to n \\ number of records
ti[A^{cs}]= mapping table value for { ti[A^{cs}] && ti[choice]}
5. Call the function Group_grade(A^{ns})

V. Experimental Methodology

The main goals of our experiments are to investigate the performance implication of our approach in terms of data mining utility, information content and the privacy level achieved. To enable direct comparison with previous micro data works [5] [6] [15] we have used the same adult database from UCI machine learning repository [9] with 45,222 records and the attributes age, sex, race, work class, Education, marital status, occupation are considered. Here, Education and Age are treated as sensitive categorical, numerical attribute respectively. One more attribute namely 'choice' is added whose domain values are {1,2,3} and all the values for that attribute were filled manually, considering the general tendency of the people. For example, people with diseases like flu, headache etc may choose choice 1 but people with diseases like cancer, tuberculosis may choose 3. But, few people may be exceptions. These factors were considered while filling the attribute choice. The only reason to publish generalized quasi identifiers and the sensitive attributes together is to support data mining tasks that consider both types of attributes in the data base, for e.g. construction of classifier.

A. Classification accuracy

We have used Weka with the default settings for C.4.5 classifier learning on the original Adult dataset. Then privacy preserving algorithm was applied to the table T, and T' was generated. The algorithm was implemented in Java standard Edition 5.0 and made to run on Intel® Core2 Duo, 1.8 GHz, 1GB RAM system which took only 28sec for generating privacy preserving Adult data set T'. Then C.4.5 is applied on T' and decision tree learning accuracy of different attributes of T and T' are listed as shown in table V. From the table V we conclude that the data mining utility (classifier accuracy) does not get affected with our approach, which in turn speaks about the truthfulness of the data available in T'. The adult dataset with the original values T and transformed table T' are able to produce the same rules and distribution as shown in fig. 3.

B. Proximity Breach

In K- anonymization (or improved methods), let t be the tuple in T, and G the QI-group in T' that t is generalized to. The risk of Proximity Breach of t, denoted as Pb(t), equals x / |G| where x is the number of tuples in G whose sensitive values fall in very short interval and |G| the size of G. In our approach, we are not generating any equivalence class. Hence the size of G to be equal to the size of table 'n' and there is least probability of getting the x value high. For example, the category is fixed as one and the maximum and the minimum value of the category is 40 and 10, then the actual values of age 21, 22 are replaced with 1.366 and 1.4 respectively. Even though the transformed values fall in very close range (proximity attack), the actual values can not be guessed, because the transformation depends on number of categories fixed and minimum and maximum values of each category. Since denominator value increases in the privacy breach calculation, Pb(t)is very low. Hence, high privacy there exists, which protects from Proximity attack. Previous works proved their algorithms by testing the classification accuracy by increasing the privacy level. The privacy level is increased by increasing the size of equivalence class or increasing the value of K. But the optimal K-anonymization even with k=2 is NP hard. Otherwise, when k increases the generalization also increases which in turn increases the information loss. We tested classification accuracy of original and transformed Census data set, generated by our privacy preserving algorithm and the result is shown in fig. 4.

Table V Comparison of learning accuracy

Attribute	Learning Accuracy in percentage	
	T	T'
Work class	74.87	74.87
Education	41.67	39.64
Marital Status	69.36	69.36
Occupation	32.23	32.23

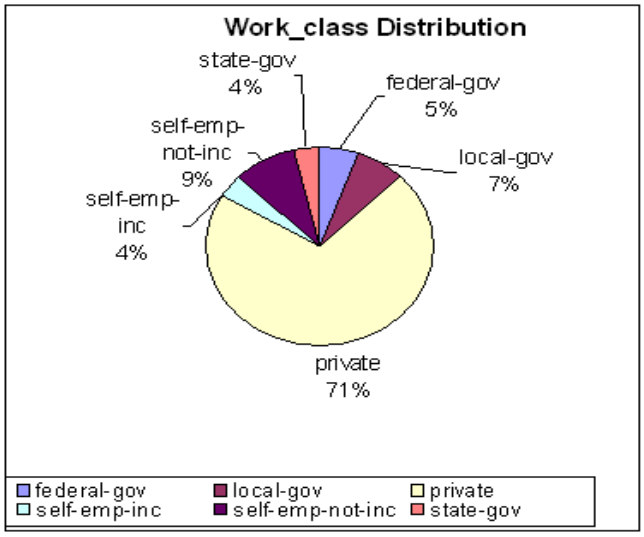

Fig. 3 Work Class Distribution (T & T')

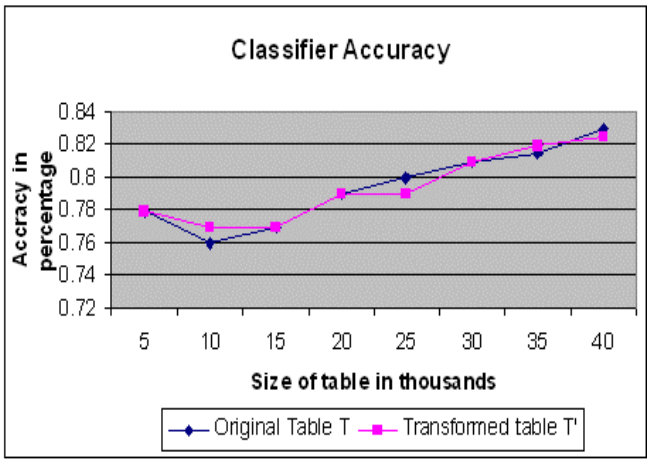

Fig. 4 Comparison of T and T'

C. Divergence Breach

The amount of data distortion occurs by generalization of equivalence class e in K-anonymization is, denoted by IL(e)=|e||G|/|D| where |e| is the number of records in the equivalence class, |D| is the domain size and |G| the amount of generalization. The amount of generalization is zero because of taxonomy tree with alias names. Hence the information loss is zero. Since Education is treated as sensitive attribute, it is defined as class variable and the distribution is found in both T and T' as shown in fig. 5 and 6.

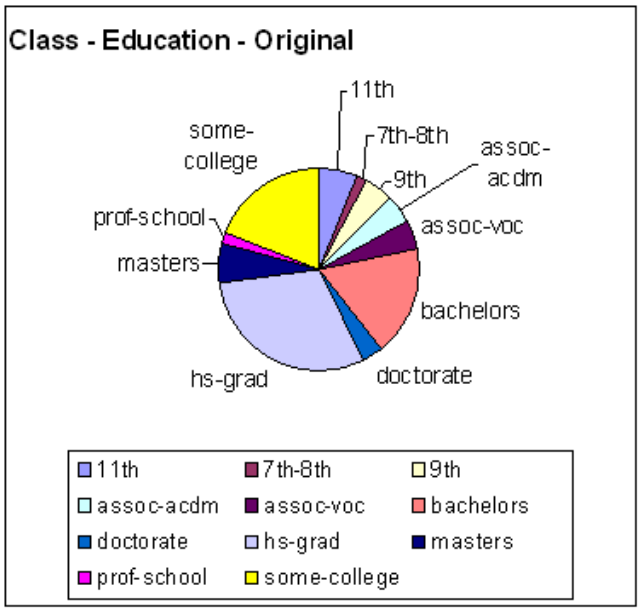

Fig. 5 Distribution of Education in T

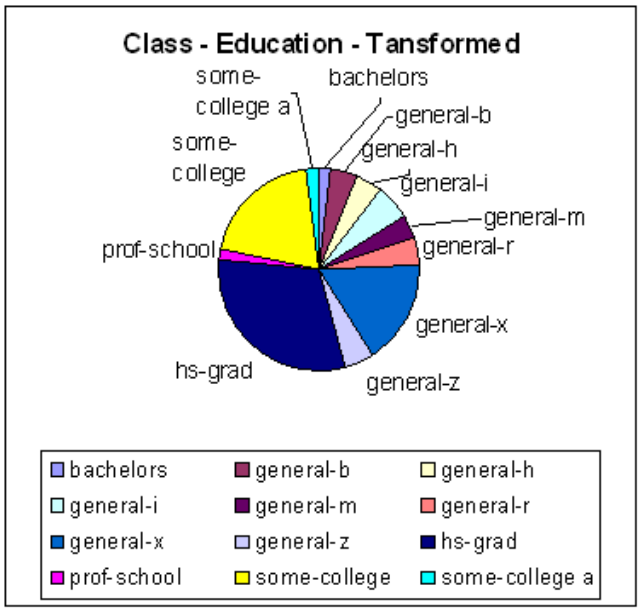

Fig. 6 Distribution of Education in T'

VI. CONCLUSION

K-anonymity and the related techniques, increase the privacy by increasing the K-value but lead to high information loss, if not optimized. Irrespective of size of the equivalence class there is always a hope for proximity and divergence attack. The personalization method combined with transformation of sensitive values avoids the need for creating an equivalence class. Even if all the other attributes (except sensitive) act as quasi identifiers and fetch unique record an adversary can not guess the actual value. The researcher can perform any task on the published data as if he/she is using original table. Only if the result contains transformed values, then he/she has to interpret the result. For this interpretation he/she has to prove authentication in the trusted server.

REFERENCES

[1] Adam N., Wortmann J. C., "Security-Control Methods for Statistical Databases: A Comparison Study", *ACM Computing Surveys*, 21(4), 1989

[2] Agrawal D, Aggarwal C. C., "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms", *ACM PODS Conference*, 2002

[3] Aggarwal G., Feder T.,Kenthapadi K, ,MotwaniR., PanigrahyR., Thomas D., Zhu A. " Anonymizing Tables," ICDT Conference, 2005.

[4] Y. Du, T. Xia, Y. Tao, D.Zhang, F. Zhu, "On multidimensional K-anonymity with local recoding generalization", I*CDE*, pp. 1422-1424,2007

[5] Justin Brickell, Vitaly Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing", *KDD*'08

[6] J. Li, Raymond chi wing wong, Ada Fu, J. pei, "Anonymization by local recoding in data with attribute hierarchical taxonomies", *IEEE transaction on Knowledge and data Engg*, Vol 20, No. 9, pp. 1181-1194,sep 2008

[7] J. Li, Y. Tao, X. Xiao, "Preservation of proximity privacy in publishing numerical data", *ACM SIGMOD '08*

[8] Machanavajjhala A., Gehrke J., Kifer D.,and Venkitasubramaniam M, "l-Diversity: Privacy Beyond k-Anonymity", *ICDE*, pp.24-35, 2006

[9] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "*UCI Repository of Machine Learning Database*s", Available at www .ics. uci. edu/~ mlearn/MLRepository.html, University of California, Irvine, 1998

[10] Ninghui Li, Tiancheng Li and Suresh.V, "t-Closeness: Privacy beyond k-anonymity and l-diversity", *ICDE*, 2007

[11] P. Samarati, "Protecting respondents identities in micro data release*", TKDE, 13(6),* 1010-1027, 2001

[12] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), pp. 571-588, 2002

[13] R. C. W. Wong, J. Li, A. W. C. Fu , K. Wang, " (Alpha – K) anonymity: an enhanced K-anonymity model for privacy preserving data publishing,", *ACM SIGKDD*, pp.754-759,2006

[14] X. Xiao and Y. Tao, "m-variance: towards privacy preserving re-publication of dynamic dataset", *ACM SIGMOD*, pp 689-700, 2007

[15] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, AWC Fu, "Utility based anonymization using local recoding", *ACM SIGKDD,* pp.785-790, 2006