A New Method for Preserving Privacy in Data Publishing Against Attribute and Identity Disclosure Risk

R.Mahesh[1] and Dr.T.Meyyappan[2]

[1]Department of Computer Science and Engineering, Alagappa University, Karaikudi
aummahesh@gmail.com
[2] Department of Computer Science and Engineering, Alagappa University, Karaikudi
meylotus@yahoo.com

ABSTRACT

Sharing, transferring, mining and publishing data are fundamental operations in day to day life. Preserving the privacy of individuals is essential one. Sensitive personal information must be protected when data are published. There are two kinds of risks namely attribute disclosure and identity disclosure that affects privacy of individuals whose data are published. Early Researchers have contributed new methods namely k-anonymity, l-diversity, t-closeness to preserve privacy. K-anonymity method preserves privacy of individuals against identity disclosure attack alone. But Attribute disclosure attack makes compromise this method. Limitation of k-anonymity is fulfilled through l-diversity method. but it does not satisfies the privacy against identity disclosure attack and attribute disclosure attack in some scenarios. The efficiency of t-closeness method is better than k-anonymity and l-diversity. But the complexity of Computation is more than other proposed methods. In this paper, the authors contributes a new method for preserving the privacy of individuals' sensitive information from attribute and identity disclosure attacks. In the proposed method, privacy preservation is full filled through generalization of quasi identifiers by setting range values.

Keywords

Data Privacy, generalization, anonymization, suppression, privacy preservation, data publishing.

1. NTRODUCTION

Private companies and government sectors are sharing micro data to facilitate pure research and statistical analysis. Individuals' privacy should be protected. Micro data contains sensitive values of record owners. Generally,microdata stored in table format (T). Adversaries (attackers) associates more than two dataset and apply their background knowledge for deducing the sensitive information. Certain attributes are associates with external knowledge to identify the individual's records indirectly. Such attributes are called Quasi Identifiers(QI). Quasi identifiers are associated with sensitive attribute(S) which should not be disclosed. Data leakage occurs by association of quasi identifiers and background knowledge. There are two types of disclosure namely attribute disclosure and identity disclosure. Anonymization techniques [4] are used to preserve the privacy of individuals and convert the microdata T to anonymized table T*. Generalizations, suppression and data swapping are common anonymization techniques. In this In this paper, a new anonymization based method is proposed for preserving the privacy of sensitive attribute values against identity and attribute disclosure attacks.

2. EXISTING METHODS

k-anonymity[14] achieve the privacy preservation against identity disclosure attack but not achieve for attribute disclosure attack. ℓ-diversity[9] overcome the disadvantage of *k*-anonymity and satisfy the privacy preservation against attribute disclosure attack. But, its effiecieny is not good in case of identity disclosure attack. *t*-closesness method[8] is preserve the privacy against attribute disclosure attack. But the computation complexity is higher than other two methods and also it fails to preserve the privacy against attribute disclosure attack. The attackers comprise the k-anonymity through background knowledge and homogeneity attacks. The modified micro data table T* satisfies (p, α)-sensitive *k*-anonymity[6] property with pre defined condition that each QI-group has at least p distinct sensitive attribute values with its total weight at least α . p-sensitive *k*-anonymity[10] is not achieve the preservation of privacy against Similarity Attack. In Enhanced P sensitive *k*-anonymity model[6], the modified micro data table T* satisfies the condition (p+, α)-sensitive *k*-anonymity property if it satisfies *k*-anonymity, and each QI-group has at least p distinct categories of the sensitive attribute and its total weight is at least α. This method significantly reduces the possibility risks of Similarity Attack and incurs less distortion ratio compared to p-sensitive *k*-anonymity method.

ℓ-diversity method [9] overcomes the drawbacks of *k*- anonymity. In this method, an equivalence class is said to have ℓ-diversity if there are at least ℓ "well-represented" values for the sensitive attribute. A table is said to have ℓ-diversity if every equivalence class of the table has ℓ-diversity. This method is not give the solution for preserving the privacy against skewness and similarity attacks.In (α,*k*)-Anonymity [5] model, a view of the table is said to be an *(α, k)-anonymization,* if the modification of the microdata table satisfies both *k*-anonymity and *α*-deassociation properties with respect to the quasi-identifier. It does not notice the identity disclosure attack. In *t*-closeness method[8], an equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all equivalence classes have *t*-closeness. It preserves the privacy against homogeneity and background knowledge attacks.Tamir Tassa [1] proposed a model of k-type anonymity. It is reduce the information loss and generate anonymized table in less generalization routine. It preserves the privacy against identity disclosure attack. Versatile publishing method[3] is splits the anonymized table T* by framing the privacy rules. Privacy can be compromised by applying the conditional probability in anonymized table.Qian Wang[2] contribute the model to make up the shortage of k-anonymity in protection of attribute disclosure attack. It can prevent attribute disclosure by controlling average leakage probability and probability difference of sensitive attribute value.

3. PROPOSED METHOD

In existing Anonymization techniques, preservation of individual privacy is not ensure against either identity or attribute disclosure attack, but not preserve the privacy against both. The method proposed in this paper attempts to reduce the limitations of existing anonymization techniques. It adopts generalization technique and preserves the privacy in a new way.

A table T of micro data contains quasi identifiers Q_i(i=1,2…n) and sensitive attribute S. Initally, Suppression techniques is applied over selected quasi identifier Qi and then generalized by following a new procedure proposed in this paper.

The proposed new method performs generalization operation and generate the anonymized table T. Suppression technique [12] [13] is applied over selected Quasi identifier Q_i which having more frequency. After the suppression process, the records in the table T are sorted and arranged in n

groups $G_1,G_2,G_3 \ldots G_n$. Each group is ordered by suppressed value of Quasi identifier attribute B_i (i=1, 2, … m). Among the Quasi identifiers Q_i, one with more distinct values is selected. In each group G_i, the next nearest minimum integer value L_i, and next nearest maximum integer value M_i are found. In each group G_i, Attribute values of quasi identifier is rewritten as a range value Li <= M_i. This process is recursively until all the Qi values in each group G_i are suppressed.

An unanonymized database table T can be generalized on quasi identifier to maintain privacy of a sensitive attribute S. A new method given below performs generalization operation that converts the table T to T*. Dataset in a table T with k number of tuples, n number of Quasi identifiers and sensitive attribute S are chosen.

Input: Table T with k tuples containing Quasi identifier Q and Sensitive attribute S.

Step 1:

Arrange the records in the table T into n groups $G_1,G_2,G_3 \ldots G_n$ by value s_i of Quasi Identifier B_i. where i = 1, 2, … k

Step3:

Repeat steps 4 to 6 varying j from 1 to k

Step 4:

Let $L_j = q_j$.

Find the next nearest integer L_j less than q_j in group G_i where i=1,2,3,…n and if found, Let Lj = q_j

Step5:

Let $M_j = q_j$.

Find the next nearest integer M_j greater than q_i in group G_i where i=1,2,3,…n and if found, Let $L_j = q_j$

Step6:

If L_j and M_j are found in the same group G_m, the generalization condition is set as

set $q_j = L_j <= M_j$

Step7:

If L_j and M_j are not found in the same group G_m, the generalization condition is set as

set $q_j = <= Q_j$

Output:

Anonymized table T^*

3.1 Functional Procedure for anonymizing table T*

```
Function Anonymize(T)
        Array q,w;
        Int, nextminimum,nextmaximum,i;
        String sv1,sv2;
        T=functiongroup(T[B])
        T=functionsort(T[Q],G);
        q=T[Q];
        w=T[S];
        if(G.count==1)

        q0='<='+q0
        else
                While (u=0 to G.count)
                  While(i=0 to T.rowcount in u )
                        Nextminimum=if(Findnextminimum(q[i],T[Q]))
                        If(Nextminimum==0)
                                Nextminimum=q[i];
```

```
                    end if
                    Nextmaximum=FindNextmaximum(q[i].T[Q])
                       q[i]=Nextminimum+"<="+Nextmaximum

                End while
            End While
End if
T*=Arrange(T[Q],q)
End function
```

Sub procedure functiongroup() finds number of groups in the given data set. Sub procedure functionsort() sorts the quasi identifier values within each group.

4. RESULTS AND DISCUSSION

Proposed method is applied over the micro data in table T as shown below. Zipcode, Age, Salary are quasi identifiers and one sensitive attribute Disease present in original table I. Quasi identifier attribute age has unique values.

Table: 1 Original Table

Sno	**Zipcode**	**Age**	**Salary**	**Disease**
1	47677	29	3000	Gastric ulcer
2	47602	22	4000	Gastritis
3	47678	27	5000	stomach cancer
4	47905	43	6000	Gastritis
5	47909	52	11000	Flu
6	47906	47	8000	Bronchitis
7	47605	30	7000	Bronchitis
8	47673	36	9000	Pneumonia
9	47607	32	10000	stomach cancer

As a first step, suppression technique is applied over Zipcode attribute to transform the dataset into Table1

Table: 2 Transformed from Table 1

Sno	**Zipcode**	**Age**	**Salary**	**Disease**	
1	4760*	22	4000	Gastritis	- G_1
2	4760*	30	7000	Bronchitis	
3	4760*	32	10000	stomach cancer	
4	4767*	27	5000	Stomach cancer	- G_2
5	4767*	29	3000	Gastric ulcer	
6	4767*	36	9000	Pneumonia	
7	4790*	43	6000	Gastritis	- G_3
8	4790*	47	8000	Bronchitis	
9	4790*	52	11000	Flu	

The proposed anonymization based method is applied over Table 2 to get the data set transformed to T*

Table: T* Anonymized table

Sno	**Zipcode**	**Age**	**Salary**	Disease	
1	4760*	22<=30	4000	Gastritis	**- G_1**
2	4760*	22<=32	7000	Bronchitis	
3	4760*	30<=32	10000	Stomach cancer	
4	4767*	27<=29	5000	Stomach Cancer	- G_2
5	4767*	27<=36	3000	Gastric Ulcer	
6	4767*	29<=36	9000	Pneumonia	
7	4790*	43<=47	6000	Gastritis	**- G_3**
8	4790*	43<=52	8000	Bronchitis	
9	4790*	47<=52	11000	Flu	

The performance of the proposed algorithm is evaluated in terms of two data metrics namely information loss and privacy gain. The proposed method and three existing methods namely k-anonymity (k=3), ℓ-diversity(l=3) and t-closeness are experimented with the same data set and their performance were compared in terms of information loss and privacy gain. The following formulae are used to measure information loss ILoss[9] and privacy gain *PG*[11,12].

$$ILOSS(v_g) = \frac{|v_g| - 1}{|D_A|}$$

where;

$|v_g|$ is the number of domain values that are descendants of v_g

$|D_A|$ is the number of domain values in the attribute A of v_g.

$ILOSS(v_g)=0$ if v_g is an original data value in the table.

In words, $ILOSS(v_g)$ measures the fraction of domain values generalized by v_g.

The loss of a generalized record r is given by

$$\text{ILoss(r)} = \sum_{Vg \in r} (w_i \times ILoss(v_g))$$

Where w_i is a positive constant specifying the penalty weight of attribute A_i. The overall loss of a generalized table T is given by

$$\text{ILoss(T)} = \sum_{r \in T} Iloss(r).$$

$PG = \text{avg}\{A(QID_j) - A_s(QID_j)\}$.

Where;

$A(QID_j)$ and $A_s(QID_j)$ denote the anonymity of QID_j before and after specialization.

The Principle of information/privacy trade-off can also be used to select a generalization g, in the which case it will minimize.

$$ILPG = \frac{IL(g)}{PG(g)}$$

Where *IL*(g) denotes the information loss and *PG*(g) denotes the privacy gain by performing g.

Table R

Methods	**Information Loss**	**Privacy Gain**	**ILPG**
k anonymity	1.488	12	.124
l- diversity	1.488	10.5	.1417
t-closeness	.990	10.5	.0942
Proposed Method	**.832**	**11.7**	**.0711**

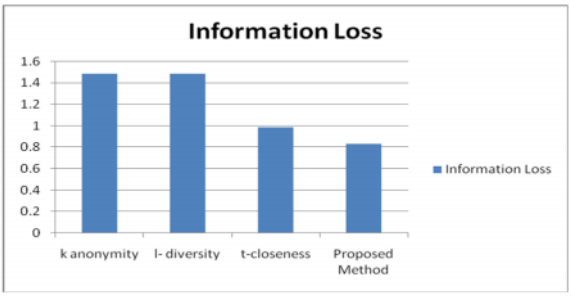

Figure.1. Measure the information loss with sample dataset

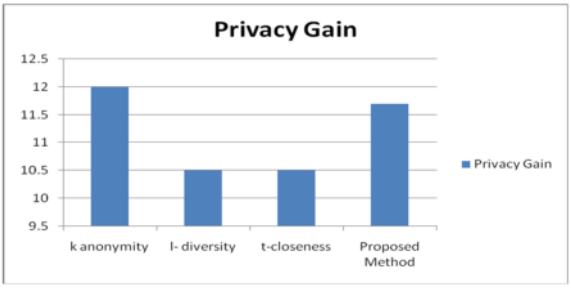

Figure.2. Measure the Privacy gain with sample dataset

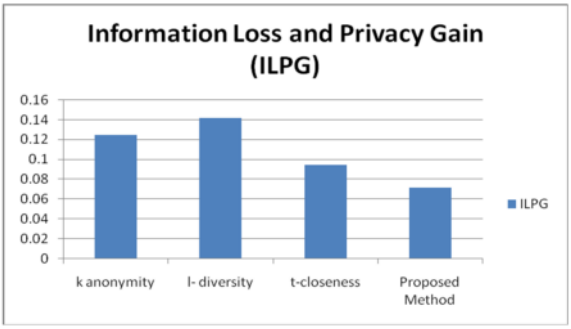

Figure.3. Measure the overall performance ILPG with sample dataset

It is observed that the proposed method reduces the information loss compared to existing methods, as shown in table R. It is also observed that the proposed method performs well in terms of privacy gain and ratio of information loss to privacy gain(ILPG). The overall performance of

the methods is shown in the last column as ILPG. The overall performance of the proposed method is encouraging.

5. CONCLUSION

In this information age, Data publishing and data sharing are increasing the transaction flow in every year. While collecting the data for statistical analysis and research purpose, privacy of the individuals whose data are published should not be challenged. In contrast to cryptographic methods which transform the plain text to ciphertext, privacy methods protects the privacy of owners whose data are published for public. The methods developed for privacy preservation should not result in information loss. In this paper, the authors proposed a new method which preserves the privacy and reduces information loss. Performance of the proposed method is compared with existing methods in terms of privacy gain and information loss metrics. The proposed new method is implemented using MatLab coding. Results are tabulated and plotted. The results show improvement in privacy gain and reduction in information loss compared to existing methods. The proposed method is achieved the preservation of individuals' sensitive information which is represent as numeric values. Further research is in progress to find out the solution for preserving the individual privacy in pubished data which having non-numeric quasi identifiers.

REFERENCES

[1] Tamir Tassa,ArnonMazza and Aristides Gionis,(2012), "k-Concealment: An Alternative Model of k-Type Anonymity", TRANSACTIONS ON DATA PRIVACY 5, pp189–222

[2] Qiang Wang,Zhiwei Xu and Shengzhi Qu,(2011) "An Enhanced K-Anonymity Model against Homogeneity Attack", Journal of software, Vol. 6, No.10, October 2011;1945-1952

[3] Xin Jin,Mingyang Zhang,Nan Zhang and Gautam Das,(2010) "Versatile Publishing For Privacy Preservation",KDD10,ACM

[4] Benjamin C.M.Fung,KE Wang,Ada Wai-Chee Fu and Philip S. Yu,(2010) "Introduction to Privacy-Preserving Data Publishing Concepts and techniques", ISBN:978-1-4200-9148-9,2010

[5] Raymond Wong, Jiuyong Li,Ada Fu and Ke wang,(2009), "(α,k)-anonymous data publishing", Journal Intelligent Information System, pp209-234.

[6] Xiaoxun Sun, Hua Wang, Jiuyong Li and Traian Marius Truta,(2008) "Enhanced P-Sensitive K-Anonymity Models for privacy Preserving Data Publishing", Transactions On Data Privacy, pp53-66

[7] B.C.M. Fung, Ke Wang and P.S.Yu,(2007) "Anonymizing classification data for privacy preservation", IEEE Transactions on Knowledge and Data Engineering(TKDE), pp711-725

[8] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam,(2007) "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity", International Conference on Data Engineering, pp106-115

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,(2006) "ℓ-diversity: Privacy beyond k-anonymity", In Proc. 22nd Intnl international Conference on data engineering. (ICDE), pp24

[10] T. M. Truta and V. Bindu,,(2006) "Privacy Protection: "p-sensitive k-anonymity property", International Workshop of Privacy Data Management (PDM2006), In Conjunction with 22th International Conference of Data Engineering (ICDE), pp94

[11] X. Xiao and Y. Tao,(2006) "Personalized privacy preservation", In Proceedings of ACM Conference on Management of Data (SIGMOD'06"), pp229-240

[12] B.C.M. Fung, Ke Wang and P.S.Yu ,(2005) ," Top-down specialization for information and privacy preservation", In Proc. of the 21st IEEE international Conference on data engineering(ICDE),pp205-216

[13] L. Sweeney,(2002) "An Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based System,pp571-588

[14] L. Sweeney,(2002) "k-anonymity: a model for protecting privacy International Journal on Uncertainty,Fuzziness and Knowledge-based Systems, pp557-570.

AUTHORS

Mahesh.R M.C.A, Currently Ph.D Research Scholar, Department of Computer Science and Engineering, Alagappa University,Karaikudi, Tamilnadu,India, His research work is in Data mining and publishing, web mining

Dr. T. Meyyappan M.Sc., M.Phil., M.B.A., Ph.D., currently, Professor, Department of Computer Science and Engineering, Alagappa University, Karaikudi, TamilNadu, India. He has obtained his Ph.D. in Computer Science and Engineering in January 2011 and published a number of research papers in National, International journals and conferences. He has been honored with Best Citizens of India Award 2012 by International Publishing House, New Delhi. He has developed Software packages for Examination, Admission Processing and official Website of Alagappa University. His research areas include Operational Research, Digital Image Processing, Fault Tolerant computing, Network security and Data Mining.