I J C T A, 9(40) 2016, pp. 227-235

Performance Evaluation of *k-A*nonymization Algorithms for Generalized Information Loss

Deepak Narula[1], Pardeep Kumar[2] and Shuchita Upadhyaya[3]

ABSTRACT

Advancement of technology has increased the size of data sets which may cause the risk of re-identification of individual's information. Many techniques have been suggested for anonymizing the data sets. Such techniques ensure the individual's identity to remain anonymous. As a result of that, privacy preservation in data publishing has become an active area for research. In this paper an evaluation of various k-anonymity algorithms has been carried out with the objective of identifying the value of general information loss that occurs due to anonymization. An experiment has been performed to determine information loss based on the type of attribute(s) on three publically available data sets that carries different dimensions.

Keywords: Metrics, Equivalence Class, Privacy Preserving Data Publishing (PPDP), Quasi identifier (QI), Discernibility Metric(DM)

I. INTRODUCTION

Data publishing deals with large amount of data collection and the huge collection of data always contain personal sensitive information that need not be disclosed. But this huge personal information about individuals always attract the attention of those who want to steal information. There is always a challenge to protect the privacy of individual. Thus, to publish data without disclosing the personal identity has become an important area of interest for researchers. The aim of PPDP is to make changes in the data by making it less specific via generalization by appropriate conversions and protect the individual's privacy while keeping the utility of anonymized data.

Due to availability of various anonymizing techniques, selection of the most appropriate one is always a challenge for the professionals. Moreover, it is not only the selection of appropriate technique but also the appropriate parameters that should be taken care of, as they are the major basis for data utility components. *k*-anonymity is one of the widely used approach for data anonymity and in this approach anonymization is achieved using generalization and suppression. Various algorithms for *k*-anonymity have been found in literature like Datafly[1], Mondrian[2], Incognito[3] etc.

In this paper an evaluation of Datafly, Mondrian and incognito anonymity algorithms have been done. Original data is anonymized using publically available software and further data utility metric is applied to calculate general information loss on different data sets to determine which algorithm is most suitable. Analysis on various data sets have been done to determine how these algorithms performs when characteristic of quasi attributes is taken into consideration and to check whether general information loss depends upon number of quasi attributes.

[1] Research Scholar, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India, *E-mail: dnarula123@yahoo.com*

[2] Associate Professor, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India, *E-mail: mittalkuk@gmail.com*

[3] Professor, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India, *E-mail: shuchita_bhasin@yahoo.com*

II. BACKGROUND AND RELATED WORK

The main objective of PPDP is to secure information while publishing the data, whereas the prime aim of adversary is to obtain the information about sensitive attribute that can be determined by linking various attributes of relation with each other. In literature various methods have been given for data classification purpose. Different approaches are given to maintain privacy of individual, with the consideration of keeping the data useful [4, 5]. In a relation variety of attributes exists which are classified as key attributes, quasi-attributes, sensitive attributes and insensitive attributes.

Key attributes are those which are concerned with unique identification of records and these attributes are generally removed while publishing the information. Examples of key attributes are Registration no., Name etc.

Quasi-attributes are those which are used for linkage of anonymized data set with the aim of reaching the sensitive information e.g. Birth Date, Age, Gender, Zip Code etc.

Sensitive attribute are those which need not to be disclosed and aim of an attacker is to determine these e.g. Disease of person, Salary etc.

There are various data models used for anonymization such as *k*-anonymity, *l*-diversity, *t*-closeness etc. This paper focuses on *k*-anonymity model as it has been widely discussed in the literature. Moreover, this also has been identified that *k*-anonymity model is vulnerable to certain attacks and also in contrast to some robust models, might hamper the utility of anonymized data to maintain privacy[6].

***k*-anonymity** This was the first model for anonymizing the data and base for the others to which further extensions have been made. The formal definition of *k*-anonymity for relation is as[1,7]. "A table T is *k*-anonymous with respect to Quasi-Identifiers $Q_i(Q_1,\ldots\ldots, Q_d)$ if every unique tuple $(q_1,\ldots q_d)$ in the projection of T on $Q_1,\ldots Q_d$ occurs at least *k* times". For example Table1 represents the original table containing data about school employees where as Table 2 represents the anonymized data with *k*=3.

Table 1
Represents records for School Employees

Sno	*ID*	*QID*			*Sensitive Attribute*
	Name	*Designation*	*Age*	*Pin Code*	*Salary*
1	Ana	TGT	49	132042	42000
2	Ali	PGT	40	132021	58000
3	Joe	PPRT	44	132024	35000
4	Karim	TGT	48	132046	43000
5	Durga	PPRT	45	132045	34000
6	Raghav	PGT	43	132027	55000

Table 2
Represents an anonymized table (*k*=3) for School Employees

Sno	*EQ*	*QID*			*Sensitive Attribute*
		Designation	*Age*	*Pin Code*	*Salary*
1	A	Teaching	[45-50)	13204$	42000
4		Teaching	[45-50)	13204$	43000
5		Teaching	[45-50)	13204$	34000
2	B	Teaching	[40-45)	13202$	58000
3		Teaching	[40-45)	13202$	35000
6		Teaching	[40-45)	13202$	55000

2.1. Operations to achieve *k*-anonymization

2.1.1 Generalization is a process of substituting the substantive value against less specific but semantically consistent value. It is achieved using the purpose of hierarchical tree and associated with attribute of category quasi-identifiers. In Fig. 1 the nodes PGT, TGT or PRT are more specific as compared with node teaching, whereas it can be seen the node School Employee is at the top of hierarchal level with highest level of generalization.

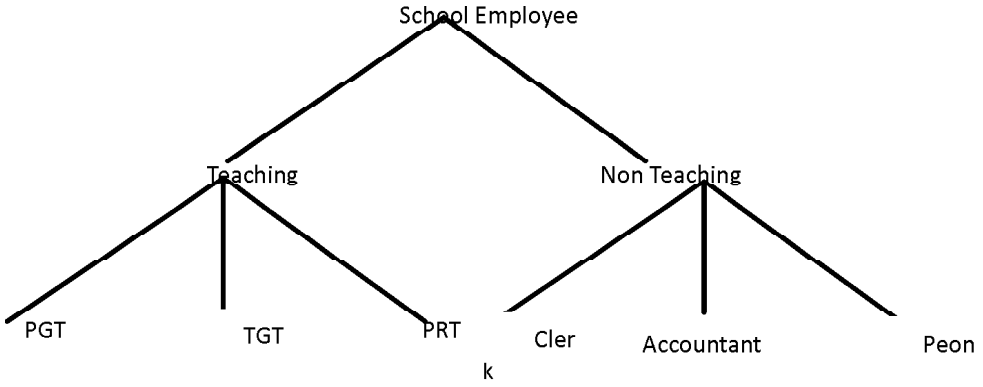

Figure 1: Example of generalization using hierarchical tree

2.1.2 Suppression This is another flavor of generalization . In this the original values of attribute generally quasi-attribute is replaced by special symbol(e.g. #,*) and makes the value of that attribute meaningless, e.g. in Table 3 the Zipcode of the city attribute is suppressed up to different levels and due to suppression at different levels will make data more anonymous.

Table 3
Example of Suppression

Zipcode		*Suppression*	
Level X	Level Y	Level Z	Level W
600231	60023*	6002**	600***
600221	60022*	6002**	600***
600210	60021*	6002**	600***

In Literature, number of algorithms have been proposed for implementing *k*-anonymity via the method of generalization and suppression for PPDP. Samarati and Sweeney[1] introduced the concept of *k*-anonymization. The *k*-anonymization is achieved by partitioning the domain of quasi attributes into set of intervals and by replacing the attributes with corresponding interval gap resulting set of at least *k*-1 tuples which are alike. Other model of anonymization was introduced by A. Machanavajjhala in 2006 [8] named as *l*-diversity to solve *k*-anonymity problems. Further in year 2007 S. Venkatasubramaniam [9] presents a model of t-closeness to overcome the possible attacks on l-diversity. An updated model of *k*-anonymity was proposed by J.Li and K.Wang [10] to protect the relationship and identification to senstitive information. Bayardo and Agarwal [11] proposed another *k*-anonymity based optimal algorithm based on full generalization of table. However in literature various models have been introduced but cannot go without *k*-anonymization. Thus three algorithms based on the principle of *k*-anonymization have been chosen namely: Datafly, Mondrian and Incognito.

In this study generalized information loss has been calculated based on the characteristics of attributes. Discussions have been made for selection of most appropriate algorithm for anonymization and to check whether general information loss depends on quasi attributes, or not.

III. K-ANONYMITY ALGORITHMS

In our evaluation analysis, following *k*-anonymity algorithms have been taken as these are based on the concept of generalization and suppression and widely cited in literature. Moreover, these algorithms are based on different tactics of anonymization. In this section a brief description about these algorithms is given:

3.1 **Datafly[1]** Data fly algorithm of anonymization is based on the concept of full domain generalization and also based on greedy heuristic algorithm approach. The data fly algorithm works by counting the frequency of similar tuples with respect to the attributes in Quasi-Id set and whether *k*-anonymity have been achieved or not .If it is not achieved further process of generalization and suppression is again applied on set of QI in table, At last process will be terminated resulting in an anonymized table in which *k*-anonymity is achieved.

3.2 Incognito algorithm [3] This algorithm works on the concept of full domain generalization and uses single dimensional method .It works by building a lattice based on generalization and traverse it by bottom up breadth first order and after traversing whole lattice returns anonymized table corresponding to the anonymized node. This algorithm finds all *k*-anonymous full domain generalization from which the "minimal" may be chosen according to any defined criteria.

3.3 Mondrian [2] This algorithm of *k*-anonymity is based on greedy multidimensional approach and works by partitioning the domain space recursively in to number of regions where each region contains at least *k*-records. This algorithm start its processing by selecting least specific value of the attribute in the QID. This also uses the attribute with widest ranges of values.

IV. DATA METRICS FOR *K*-ANONYMITY ALGORITHMS

Evaluation of anonymity algorithms is necessary to analyze as to which algorithm of anonymization is best suited for a particular type of data set. A brief description about these various metrics has been given and for evaluation purpose these have been implemented in Python.

4.1 Generalized Information Loss[12] This metric is used to calculate the amount of forfeiture incurred when an attribute is generalized. In the given metric for calculating the generalized loss, L_i and U_i are considered as lower and upper bounds of an attribute *i*. A cell entry for attribute *i* is generalized to an interval *i j* defined by lower bound L_{ij} and upper bound U_{ij} which are two end points, whereas the total information loss for an annonymized table is calculated as:

$$Genloss(T*) = \frac{1}{|T|*n} * \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i}$$

Whereas T* is anonymized table , |T| represents the cardinality of table, n represents the total number of attributes.

4.2 Discernibility Metric[11] This metric is used to calculate how a record is indistinguishable from the other available in a table T . In this a penalty is assigned to each record which is equal to the size of EQ to which it belongs. Moreover, if a record is suppressed, then assign a penalty equal to size of input table. The total DM for a table T is calculated as

$$\mathbf{DM(T*)} = \sum_{\forall E.Q.s.t|EQ| \geq k} |EQ|^2 \sum_{\forall E.Q.s.t.|EQ|} |T|*|EQ|$$

In the above defined formula T is actual table, |EQ| is size of equivalence class and T* is anonymized table

4.3 Average Equivalence class size Metric(C_{AVG}) This metric describes how well the creation of equivalence class size approaches the best case, where each record is generalized in an EQ of k record [2]. The total C_{AVG} score is calculated as

$$C_{AVG}(T^*) = \frac{|T|}{|EQs| * k}$$

Where T* is anonymized table, T is original table, |T| is cardinality of table T and|EQs| represents the total no of equivalence classes created and k is privacy requirement.

V. PROBLEM FORMULATION

In this paper, the problem is to identify which of the algorithm performs better as compared to other under various scenarios. The evaluation is based on various characteristics of attributes such as numeric, non-numeric or combination of both.

In this problem the input is taken to be three publically available data sets and the output will be generalized information loss after anonymizing the data set.

VI. DATA SETS USED FOR ASSESSMENT

In this section description about the datasets used in the comparison have been given.

6.1. Adult Data Set[13]

This is the first data set used for the purpose of evaluation. After removing the tuples with blank values the total numbers of attributes taken are nine whereas the total number of tuples used with this data set are 5411. The attributes considered for this data set are:

Adult = {Age, Sex, Race, Marital Status, Education, State, Qualification, Designation, Salary}

6.2. American Time Use Survey (ATUS) Data Set[13]

This is the second data set used for the purpose of evaluation. After deleting the tuples containing NULL values total number of attributes taken are five whereas the total number of tuples used with this data set are 56663. The attributes considered in this data set are:

ATUS = {Age, Region, Race, Marital Status, Qualification}

6.3. CUPS Data Set[13]

This is the third data set used for the purpose of evaluation. After removing the records with NULL values the total number of attributes taken is five whereas the total number of tuples used with this data set are 62414. The attributes considered in this data set are:

CUPS ={Zip Code, Age, Sex, Salary, Qualification}

VII. EXPERIMENTAL ANALYSIS

The goal of experiment is to make a comparison between three anonymization algorithms based on the model of *k*-anonymity and calculating general information loss by anonymizing the data using UTD software[15] and further data utility metric has been applied to calculate the value of generalized information loss. The data utility metric to calculate general information loss was implemented in Python language.

7.1 General Information Loss for Adult data set : Anonymization and evaluation have been done to calculate general information loss on the basis of different attributes with varying characteristics' such as numeric, non numeric or their combination. For evaluation, total number of records considered are 5411 and value of *k* is 300. Table 4 shows the result of evaluation on the basis of three different algorithms with different attribute such as Age(numeric),Marital Status(Non numeric),Qualification(Non numeric).

Table 4
Result of General Information loss for Adult data set

Algorithm/No of QI	*Age*	*Marital Status*	*Age, Marital Status*	*Age, Marital Status, Qualification*
Data Fly	0.238594695	0.432144397	0.441894722	0.270048389
Mondrian	0.110750373	0.165650219	0.251833519	0.144373462
Incognito	0.451645047	0.432144397	0.441894722	0.472876961

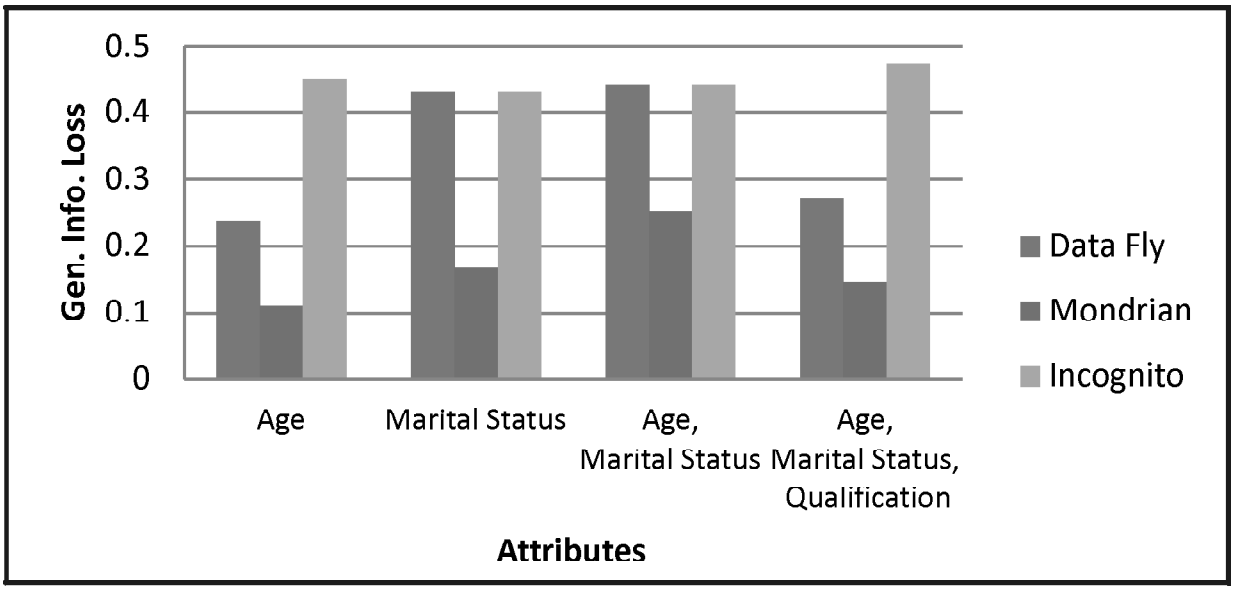

Figure 2: Comparative analysis of the three algorithms for Adult data set

It has been observed from Fig. 2 that Mondrian outperforms in all cases when anonymization have been made on numeric attribute (Age) or nonnumeric type attribute(Marital Status) or combination of both types of attributes(Age, Marital Status). It has also been observed that general information loss is minimum when anonymization has been performed using either numeric or non numeric single attribute but while using the combination of both type of attributes the information loss.

7.2 General Information Loss for ATUS data set : Anonymization and evaluation have been done to calculate general information loss on the basis of different attributes with varying characteristics' such as numeric, non numeric or their combination. For evaluation, total number of records considered is 56663 and value of *k* is 300. Table 5 shows the result of evaluation on the basis of three different algorithms with different attribute such as Age(numeric), Race (Non numeric),Marital Status(Non numeric).

Table 5
Result of General Information loss for ATUS data set

Algorithm/No of QI	*Age*	*Race*	*Age, Race*	*Age, Race, Marital Status*
Data Fly	0.256995876	0.478026225	0.49549729	0.58176919
Mondrian	0.017757263	0.07778268	0.058168965	0.172796215
Incognito	0.256995876	0.478026225	0.49549729	0.568615616

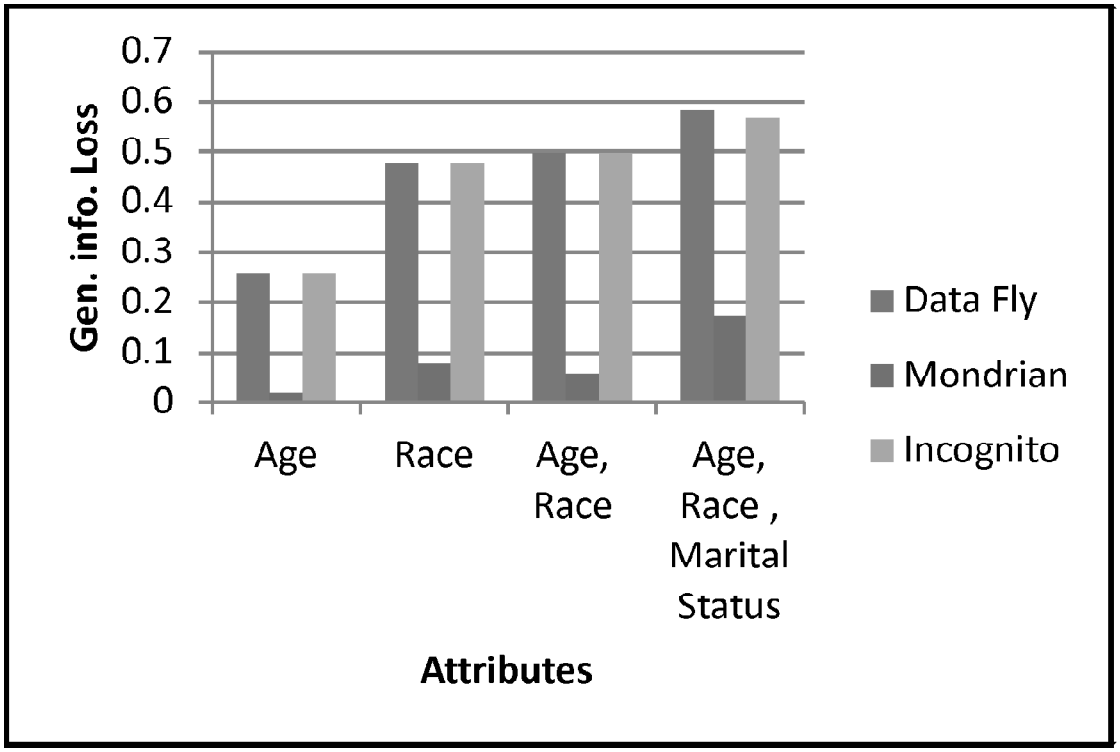

Figure 3: Comparative analysis of the three algorithms for ATUS data set

From the Fig. 3 it has been observed that information loss is minimum in case of numeric single attribute. General information loss for Mondrian is less as compared with other two algorithms. Moreover, Datafly and Incognito produces almost similar results. Information loss is growing along with no of attributes as well as their characteristics.

7.3. General Information Loss for CUPS data set

Again anonymization and evaluation have been done to calculate general information loss on the basis of different attributes with varying characteristics' such as numeric, non numeric or their combination. For evaluation, total number of records considered is 62414 and value of k is 300. Table 6 shows the result of evaluation on the basis of three different algorithms with different attribute such as Age (numeric), Qualification (Non numeric), Sex (Non numeric).

Table 6
Result of General Information loss for CUPS data set

Algorithm/No of QI	*Age*	*Qualification*	*Age, Sex*	*Age, Qualification*	*Age, Sex, Qualification*
Data Fly	0.271223517	0.001057455	0.135611758	0.341724594	0.227816396
Mondrian	0.017173317	0.26197648	0.480139515	0.152148698	0.430588098
Incognito	0.271223517	0.149075528	0.135611758	0.341724594	0.227816396

From Fig. 4 it has can be observed that Mondrian outperforms when data set is anonymized only on numeric single attribute but when a new attribute of non numeric category is added for anonymization and the domain set does not contain multiple distinct values Mondrian does not performs well. As in case of sex attribute when domain set contains only two values and applied along with another attribute, information loss is on higher side with Mondrian. Moreover the average general information loss in Datafly and Incognito is almost similar.

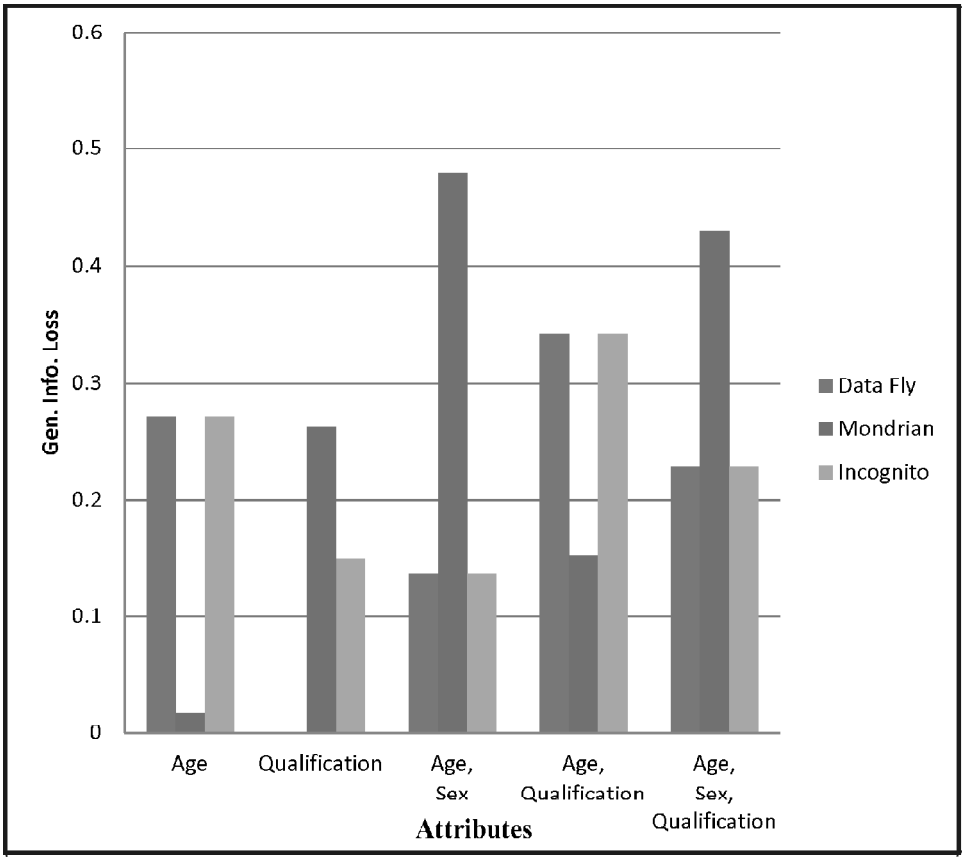

Figure 4: Comparative analysis of the three algorithms for CUPS data set

VIII. CONCLUSIONS

Many methods have been proposed for anonymizing the data sets and to preserve privacy while publishing till date. This paper provides a comprehensive analysis for different data sets with different dimensions and characteristics. It can be concluded that none of the anonymization algorithms always performs better to give consistent results with every data set and general information loss does not depend upon no of quasi attributes. Moreover, general performance of Mondrian is better than the Datafly and Incognito. General information loss in case of Incognito algorithm is more than the others and therefore the performance of incognito is not as good as Datafly or Mondrian. Moreover, if anonymization has been performed on the basis of attribute with small distinct domain set then Mondrian does not perform as good as other methods. So, there is a scope of enhancement in the all the methods so that minimum information loss occurs. In future analysis of anonymization based on more data utility metrics will be discussed.

REFERENCES

[1] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5):571{588, 2002}.

[2] Kristen LeFevre, David J. DeWitt. Mondrian Multidimensional K-Anonymity, In proceeding of 22nd International Conference on Data Engineering, ICDE'06, pp25,2006.

[3] Kristen LeFevre, David J. DeWitt,Raghu Ramakrishnan. Incognito: Efficient Full-Domain K-Anonymity, SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA Copyright 2005 ACM 1-59593-060,May,2006.

[4] Zaman, A N K and Obimbo, Charlie on Privacy Preserving Data Publishing: A classification Persepctive, International Journal of Advanced Computer Science and Applications, Vol 5, No 914, pp 129-134, 2014.

[5] Lamba S. and Abbas S. Qamar, Model for Privacy Preserving of Sensitive Data, International Journal of Technical Research and Applications, Vol 1, e-ISSN:2320-8163,pp 07-11, July-August, 2013.

[6] J. Soria-Comas , J. Domingo-Ferrer, D. Sanchez, and S. Martýnez. Improving the Utility of Differentially Private Data Releases via k-Anonymity. In Proceedings of the 12th IEEE International Conference on Trust , Security and Privacy in Computing and Communications, TRUSTCOM '13, pp 372–379, 2013.

[7] P. Samarati. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Engineering, 13(6), 2001.

[8] A.Machanavajjhala , J. Gehrke ,D.Kifer,and M. Venkitasubramaniam."l-diversity:Privacy beyond k-anonymity".In proc. Of the 22nd IEEE International Conference on Data Engineering (ICDE),Atlanta, GA, 2006.

[9] N.Li,T. Li., t-closeness:Privacy beyond k-anonymity and l- diversity. Proc of 21st IEEE Internatinal Con ference on Data Engineering ICDE), Istanbul, Turkey, April 2007.

[10] R.CW. Wong, J. Li,a.WC.Fu, and Ke. Wang.(á,k)- Anonymity: An Enhanced k-Anonymity Model For Privacy Preserving Data Publishing, In Proceeeding of 12th International Conference on Knowledge Discovery and Data Mining pp754-759, 2006.

[11] Bayardo, R. J. and Agrawal, R., "Data Privacy Through Optimal k-Anonymization", In Proceedings of the 21st International Conference on Data Engineering, ICDE '05, pp 217–228, 2005.

[12] Nergiz, M. E. and Clifton, C. "Thoughts on k-Anonymization", Data and Knowledge Engineering, 63(3):pp622–645, 2007.

[13] *https://drive.google.com/open?id=0B1QMEQlbBZ9zMy1LU0FEaXprem8*

[14] Manjusha S. Mirashe, Kapil N. Hande,Survey on Efficient Technique for Annonymized Microdata Preservation,International Journal of Emerging and Development, 2015, Vol.2, Issue 5, ISSN 2249-6149, pp 97-103, March,2015.

[15] UTD Anonymization Toolbox. *http://cs.utdallas.edu/dspl/cgi-bin/toolbox/*