DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5497

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PERFORMANCE EXPLANATION OF K-ANONYMIZATION ALGORITHMS FOR AVERAGE CLASS PARTITIONING METRIC

Deepak Narula
Research Scholar
Dept. of Computer Science & Applications,
KU, Kurukshetra, Haryana, India

Pardeep Kumar
Associate Professor
Dept. of Computer Science & Applications,
KU, Kurukshetra, Haryana, India

Shuchita Upadhyaya
Professor
Dept. of Computer Science & Applications,
KU, Kurukshetra, Haryana, India

Abstract **:**Nowadays besides the excessive use of technology peoples are unwilling to store their information because of privacy threat. In various domains data collection plays a significant role and beneficial in various fields such as Health care/Medical field etc. Some time collected data contains such sensitive data which is quiet personal and need not to be disclosed but if some of the information is revealed it may causes major risk. Privacy Protection data publishing (PPDP) works with an aim to give protection to an individual against identification risk and uses the process of data sanitization before publishing. Various techniques ensures the individuals identity to remain anonymous .In this research paper assessment of various k-anonymity algorithms have been made by keeping an objective in mind to determine that how well the equivalence classes have been formed when anonymization have been performed and data set is divided in to various classes. Moreover, proper investigation have been conducted in the direction of identifying the value of average equivalence class size on three publically available data sets with varying dimensions.

Keywords :Metrics, Equivalence Class, Privacy Preserving Data Publishing (PPDP), Quasi identifier (QID), American Time Use Survey (ATUS)

I. INTRODUCTION

Data Protection while publishing remains an important and crucial matter of concern for those who are using technology as with the advancement and huge use of technology large volume of data is collected. This security of such huge collected volume of data is always a major area of concern. Sometime this collected data contains sensitive information and the target of attacker is always to deduce the sensitive information. So, sometime peoples are unwilling to reveal their information digitally. Thus data publishing along with protection is always an active domain for research .Various techniques of anonymization are available in the literature, among them k-anonymizatin is a technique which is bases for all other techniques and widely used for data anonymization. It is based on the process of generalization and suppression. Datafly[1], Mondrian[2], Incognito[3] are the three algorithms which are based on k-anonymity and uses the concept of generalization and suppression. Moreover when any technique of anonymization is applied on dataset some losses to data will occur.

In this paper an evaluation of Datafly, Mondrian and Incognito anonymity algorithms have been done. Initial data is anonymized and further by applying the average equivalence class size metric values have been calculated on publically available different data sets to determine that how well the equivalence classes have been formed. A systematic analysis has been performed in the direction of determining the performance of three different algorithms and the effect of number of quasi attributes on the value of average equivalence class size. Besides this different characteristics of quasi attributes such as numeric, nonnumeric or their combination have been taken in the process of determining the result.

II. BACKGROUND AND RELATED WORK

With the advancement and use of technology huge volumes of data is accumulating day to day and due to such exponential growth of data which is an asset for today for the purpose of research .But this huge volumes of data leads to new challenges for publishing while protecting its privacy. As a result of that PPDP is an important and crucial area of research and collected data is an asset for today as it can be used for various purposes[4] [5]. Therefore handling vast collected data and providing security to individual remains a challenge for researcher and practitioners.

According to a typical scenario of PPDP the aim of attacker is to deduce the sensitive information from the stored data whereas the aim of PPDP is to anonymize the data before

publishing. The attacker deduce the information by linkage method on the bases of various attributes belonging to various categories .A variety of attributes in a relation are classified as key attributes, quasi-attributes, sensitive attributes and insensitive attributes.

Different anonymization models exist such as k-anonymity[6], l-diversity[7], t-closeness[8] etc. But the focus of this paper is only on k-anonymity model and its algorithms as it this model has been widely discussed in the literature and bases for the other.

This was the first model for data anonymization and base for the others. The formal definition of k-anonymity for relation is as[1,6]. "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 Records for School Employees[9]

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 Anonymized table (k=3) for School Employees[9]

Sno	EQ	QID			Sensitive Attribute
		Designatio	Age	Pin	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

There are a variety of methods which are suggested in literature for the implementation of k-anonymity using the method of generalization and suppression. Samarati and Sweeney[1] acquainted with the concept of k-anonymization. In [10] Xuyun Zhang et al. have given the concept of providing security and privacy to the intermediate data sets. Whereas an amended model of k-anonymity was proposed by J.Li et al. [11] for protecting the relationship and identification of sensitive information. Bayardo et al.[12] has given another k-anonymity based optimal algorithm also based on full generalization of table. Mohammad Reza Zare [13] aims on providing privacy over data publishing under the concept of privacy data utilization and prevention of disclosure of individual identity.

However in literature a variety of anonymization methods have been given but k-anonymity is the base for all. In this paper three algorithms have been taken namely: Datafly, Mondrian and Incognito these are based on the concept of k-anonymity.

Moreover for determining the performance of different algorithms, various metrics are available in the literature such as generalized information loss , value of discernibility and average equivalence class size. Work has been already performed to calculate and compare the performance of various algorithms to calculate generalized information loss[14] and value of discernibilty[9] by the researcher. In this work the value of average equivalence class size has been calculated based on the characteristics of attributes. Further a systematic comparison has been given to select the most appropriate algorithm for anonymization and to verify whether the average equivalence class size value depends on number of quasi attributes or not .

III. AVERAGE EQUIVALENCE CLASS SIZE METRICS FOR k-ANONYMITY ALGORITHMS

To select the most appropriate anonymization algorithm from the set of available algorithms a systematic assessment is needed. Moreover, a concise elucidation about average equivalence class size metric has been given and for evaluation purpose these have been implemented in Python .

3.1 AVERAGE EQUIVALENCE CLASS SIZE METRIC

This metric describes how well the formation of equivalence class size approaches to the best case, where each record is generalized in an EQ of k record [2][15]. The total C_{AVG} score is calculated as

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

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

To calculate the value of this metric Table 2 will be considered which shows two equivalence classes, the C_{AVG} value will be $\frac{6}{2*3} = 1$

IV. PROBLEM FORMULATION

The aim of this paper is to deduce the performance of different algorithms under different circumstances that include different characteristics of quasi attributes. Moreover, publically available different datasets are taken as source of input and output is the value of average equivalence class size.

V. VARIOUS DATA SETS USED FOR EVALUATION

This section contains information about different datasets used for evaluation.

5.1. Adult Data Set[16]

Firstly the process of anonymization is applied on adult data set and then the value of average equivalence class size is determined. For analyzing the value of average equivalence class size total no of tuples taken are 5411 with 9 attributes. The list of attributes considered are:

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

5.2. American Time Use Survey (ATUS) Data Set[16]

The process of anonymization is applied on ATUS data set further the value of average equivalence class size is

determined. Moreover, for analyzing the value of average equivalence class size total no of tuples taken are 56663 with five attributes. The attributes considered in this data set are:
ATUS = {Age, Region, Race, Marital Status, Qualification}
5.3 CUPS Data Set [16]

The third data set used for analysis is CUPS for its processing first the process of anonymization is applied then the value of average equivalence class size is determined. For analyzing the value of average equivalence class size total no of tuples taken are 62414 with five attributes. The attributes considered in this data set are:
CUPS = {Zip Code, Age, Sex, Salary, Qualification}

VI. EXPERIMENTAL ANALYSIS

The objective of experiment is to produce a comparison between three different anonymization algorithms based on the concept of k-anonymity and determining the value of equivalence class size by anonymizing the data using UTD software[17] and further data utility metric has been applied to determine the value of equivalence class size. The data utility metric to calculate the value of average equivalence class size was implemented in Python language.

6.1 Average Equivalence Class Size for Adult data set : To deduce the value of average equivalence class size process of anonymization and evaluation have been performed on adult data set containing 5411 records after the process of data sanitization where the value of k is taken as 300. Table 3 shows the evaluation upshot considering different characteristics of quasi attributes such as Age (numeric), Marital Status (Non numeric), Qualification(Non numeric).

Table 3 Upshot values of average equivalence class size for Adult data set

Algorithm/ No of QI	Age	Marital Status	Age, Marital Status	Age, Marital Status, Qualification
Data Fly	4.508333	9.018333	4.509167	9.018333333
Mondrian	2.004074	4.509167	1.387436	1.503055556
Incognito	9.018333	9.018333	4.509167	4.509166667

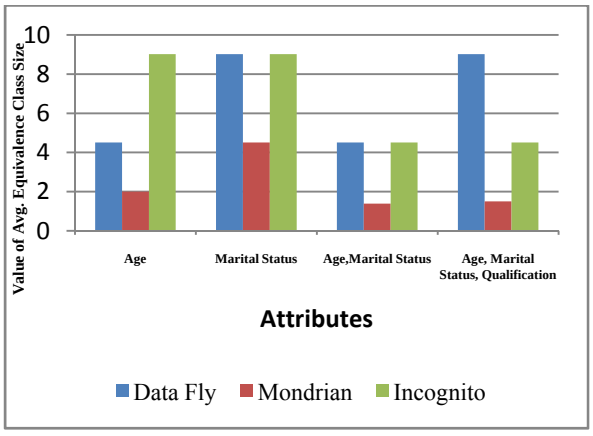

Figure 1: Comparative analysis of the three algorithms for Adult data set

It has been observed from Fig. 1 that the performance of Mondrian is consistent among all the cases under varying characteristics of quasi attributes .Moreover, best case with all anonymization algorithms will occur when anonymization have been performed with the combination of numeric and non numeric attribute.

6.2 Average Equivalence class size for ATUS data set :
For deducing the value of average equivalence class size process of anonymization and evaluation have been performed on ATUS data set containing 56663 records after the process of data sanitization where the value of k is taken as 300. Table 4 shows the evaluation upshot considering different characterstics of quasi attributes such as Age(numeric), Race (Non numeric),Marital Status(Non numeric).

Table 4 Upshot values of average equivalence class size for ATUS data set

Algorithm/ No of QI	Age	Race	Age, Race	Age, Race ,Marital Status
Data Fly	47.21917	62.95889	31.47944	31.47944444
Mondrian	2.951198	37.77533	2.303374	2.122209738
Incognito	47.21917	62.95889	31.47944	20.9862963

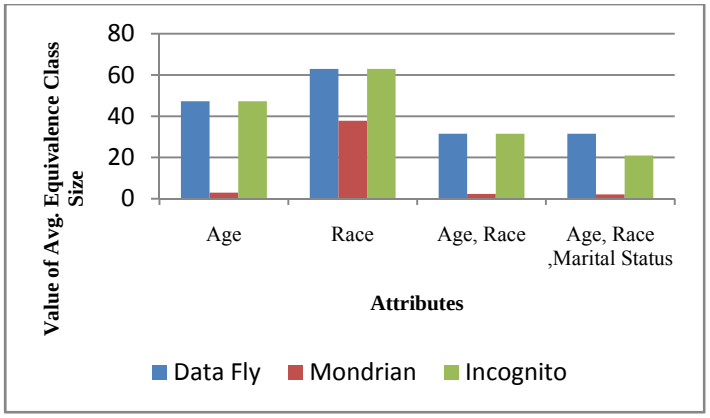

Figure 2: Comparative analysis of the three algorithms for ATUS data set

From the Fig. 2 it has been observed that performance of Mondrian is outstanding among all three algorithms where as the number of quasi attributes carrying different characteristics. Moreover while increase in the number of quasi attributes for anonymization the value of average equivalence class size reduces and the best case occurs when anonymization have been performed with the combination of numeric and non-numeric attribute.
6.3 Average equivalence class size for CUPS data set : In third case for calculating the value of average equivalence class size process of anonymization and evaluation have been performed on cups data set containing 62414 records after the process of data sanitization where the value of k is taken as 300. Table 5 shows the evaluation upshot considering different characteristics of quasi attributes such as Age(numeric),Qualification (Non numeric),Sex(Non numeric).

Table 5 Upshot values of average equivalence class size for CUPS data set

Algorithm/ No of QI	Age	Qualification	Age, Sex	Age, Qualification	Age,Sex, Qualification
Data Fly	52.011667	13.002917	26.005833	13.00291667	52.011667
Mondrian	3.715119	41.60933	3.302328	1.926358025	3.715119
Incognito	52.01167	52.01167	26.00583	13.00291667	52.01167

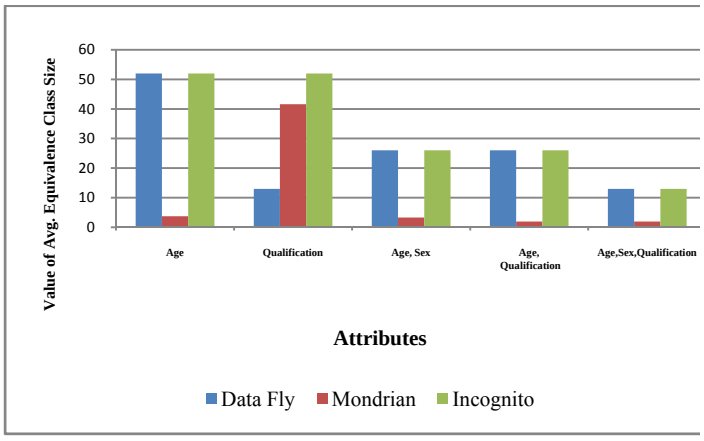

Figure 3: Comparative analysis of the three algorithms for CUPS data set

From Fig. 3 it has been deduced that the performance of Mondrian is outstanding in all the cases except when anonymization has been made on single character quasi attribute. In such case performance of datafly is good .Moreover, the value of average equivalence class size decreases and going to produce best case when anonymization have been performed on the bases of numeric and non-numeric attribute and domain set of non numeric have multiple values.

VII. CONCLUSIONS

In current period of time various techniques have been proposed for publishing the data while keeping the privacy of data .This paper provides a detailed analysis of different data sets with varying size and characteristics and it has been deduced that none of algorithm produces consistent results. Moreover, keeping in view the general performance Mondrian outperforms among all three algorithms and it has been concluded that the formation of equivalence class size approaches the best case when anonymization have been performed with the combination of quasi attributes .Moreover, It has been also concluded that Mondrian outperforms when domain set of an attribute contains multiple different values. So, there is a scope of enhancement of methods that formalizes equivalence classes of the best case.

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, page 25,2006.

[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 1-59593-060-4/05/06.

[4] 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.

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

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

[7] 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.

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

[9] Deepak Narula , Pardeep Kumar, Shuchita Upadhyaya , "Performance Evaluation Of k-Anonymization Algorithms for Generalized Information loss", International Journal of Computer Science and Engineering, Vol-5(2017), Issue 11(2017), ISSN 2347693,PP74-78,Nov-2017

[10] X Zhang, C Liu, S Nepal, S Pandey, J Chen, " A Privacy Leakage Upper Bound Constraint-Based Approach for Cost Effective Privacy Preserving Of Intermediate sets in Cloud" IEEE Transactions on Parallel and Distributed Systems 24 (6), 1192-1202

[11] R.CW. Wong, J. Li,a.WC.Fu, and Ke. Wang.(α,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, Philadelphia, PA, 2006.

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

[13] Mohammad Reza Zare Mirakabad, Aman Jantan, Diversity versus anonymity for privacy preservation Conference Paper · September 2008 DOI: 10.1109/ITSIM.2008.4632044 · Conference: Information Technology, 2008. IT Sim 2008. International Symposium on, Volume: 3

[14] Deepak Narula , Pardeep Kumar, Shuchita Upadhyaya , "Performance Evaluation Of k-Anonymization Algorithms for Generalized Information loss", International Journal of Control Theory and Applications ,Vol-9(2016), ISSN 0974-5572 Issue 40(2016), PP 227-235

[15] Vanessa Ayala Rivera, Patrick McDonagh, " A Systematic Comparison and Evaluation of k-anonymization algorithms for practitioners", Transactions on data privacy Volume 7: 337-378,2014

[16] ataSource:https://drive.google.com/open?id=0B1QMEQlbBZ9zMy1LU0FEaXprem8

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