International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Performance Interpretation of *k*-Anonymization Algorithms for Discernibility Metric

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

[1*]Dept. of Computer Science and Applications, KU, Kurukshetra, India
[2] Dept. of Computer Science and Applications, KU, Kurukshetra, India
[3] Dept. of Computer Science and Applications, KU, Kurukshetra, India

**CorrespondingAuthor:dnarula123@yahoo.com*

Available online at: www.ijcseonline.org

Received: 19/Oct/2017, Revised: 31/Oct/2017, Accepted: 16/Nov/2017, Published: 30/Nov/2017

Abstract- Advancement in technology and web based activities has increased the size of data sets which may cause the risk of re-identification about individual's information. Multifarious techniques have been suggested for anonymizing the data sets. Aforesaid techniques ensure the individual's identity to remain anonymous. As a result of that, privacy preservation in the field of 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 discernibility that occurs due to anonymization. An experiment has been performed to determine the value of discernibility based on the type of attribute(s) on three publically available data sets that carries different dimensions.

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

I. INTRODUCTION

Protection of data besides privacy is always an important concern while handling public data sets. As a result, data protection along with its privacy is an active research domain in which various techniques of anonymization has been proposed to protect individual privacy. Moreover, the provided data sets to be anonymized is further used for analysis and during anonymization it is not only the selection of appropriate technique but also the parameter appropriateness is a matter of concern for various data utility components. *k*-anonymity is a technique which is widely used for data anonymity. This approach, anonymization is achieved using generalization and suppression. Different 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. Initial data is anonymized and further by applying the discernibility metric process its value have been calculated on different data sets to determine that how much tuples are indistinguishable and which algorithm is most suitable. Analysis have been done on various data sets to determine how these algorithms perform when characteristic of quasi attributes is taken into consideration and to check whether the value of discernibility depends upon number of quasi attributes .

II. BACKGROUND AND RELATED WORK

Due to rapid growth of web based activities, people are recording their activities online, hence size of data sets grow exponentially ever year [4]. Most of us are even unaware about the collection of continuously produced electronically data.

Such accumulated data is an important asset for today as it can be used for various purposes. But this huge collected data has brought new challenges for protection and privacy of people represented in these data sets. As a result, Privacy-Preserving Data Publishing (PPDP) is one of the areas of interest for researcher and practitioners. A typical scenario of PPDP is shown in the below figure

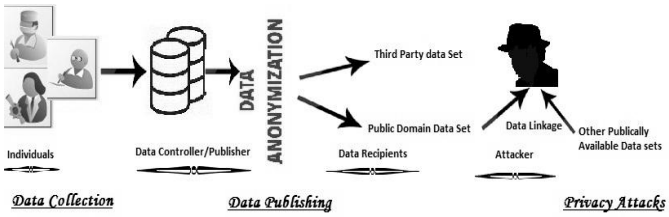

Figure 1. Overview about PPDP

Figure1 shows that the aim of PPDP is to modify the data by anonymization technique and also keeping its usefulness,

whereas the aim of attacker is to retrieve the useful personal information by data linkage method. These linkages have been done by quasi-attributes that exist in the relation. A variety of attributes in a relation are classified as key attributes, quasi-attributes, sensitive attributes and insensitive attributes.

There are numerous anonymization data models such as *k*-anonymity, *l*-diversity, *t*-closeness etc. This paper focuses only 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 [7].

k-anonymity This was the first model for data anonymization and base for the others . The formal definition of *k*-anonymity for relation is as[1,8]. "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

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

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

In Literature, variety 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 [9] named as *l*-diversity to solve *k*-anonymity problems. Further in year 2007 S. Venkatasubramaniam [10] 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 [11] to protect the relationship and identification to senstitive information. Bayardo and Agarwal [12] proposed another *k*-anonymity based optimal algorithm based on full generalization of table. However in literature various models have been introduced but they cannot go without *k*-anonymization. Thus three algorithms based on the principle of *k*-anonymization have been chosen namely: Datafly, Mondrian and Incognito.

In the discourse of computing the performance of various algorithms, different metrics exist in the literature such as generalized Information Loss , discernibility and average Equivalence class size in this study the value of discernibility has been calculated based on the characteristics of attributes. Whereas discernibility metric measures the number of tuples that are indistinguishable from each other. Also a discussions have been made for the selection of most appropriate algorithm for anonymization and to check whether the value of discernibility depends on quasi attributes, or not .

III. *k*-ANONYMITY ALGORITHMS

In our evaluation analysis, subsequent *k*-anonymity algorithms have been taken. Moreover, these algorithms are based on different tactics of anonymization. In this section a brief description about these algorithms is provided:

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. DISCERNIBILITY DATA METRICS FOR *k*-ANONYMITY ALGORITHMS

Evaluation of anonymity algorithms is necessary to analyze as to which algorithm of anonymization is best suited. A brief description about discernibility metric has been given and for evaluation purpose these have been implemented in Python .

4.1 Discernibility Metric[12] 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

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

In the above defined formula T is actual table, |EQ| is size of equivalence class and T* is anonymized table.
For e.g. from the table 2. The value of discernibility metric is 18 as table contains two equivalence classes and both the classes satisfying the value of k and of size 3 each i.e. |EQ|=3 thus DM=$(3)^2$+$(3)^2$=18.

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 value of discernibility 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[14]

Firstly Adult data set is used to calculate the value of discernibility . The evaluation was done on 5411 tuples with nine attributes after removing the tuples with blank values from the original data set. 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[14]

This is the second data set used for the purpose of evaluation. In this data set total number of tuples taken are 56663 with five attributes after deleting the records containing NULL values. The attributes considered in this data set are:
ATUS = {Age, Region, Race, Marital Status, Qualification}

6.3 CUPS Data Set[14]

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 discernibility by anonymizing the data using UTD software[16] and further data utility metric has been applied to calculate the value of discernibility. The data utility metric to calculate discernibility was implemented in Python language.

7.1 Discernibility for Adult data set

Anonymization and evaluation have been done to calculate the value of discernibility on the basis of different attributes with varying characteristics' such as numeric, non numeric or their combination. For calculating the value of discernibility metric, total number of records considered are 5411 and value of *k* is 300. Table 3 shows the result of evaluation on the basis of three different algorithms with different attribute such as Age(numeric), MaritalStatus(Non numeric),Qualification(Non numeric).

Table 3 Result of discernibility for Adult data set

Algorithm/ No of QI	Age	Marital Status	Age, Marital Status	Age, Marital Status, Qualification
Data Fly	11409831	14695573	10137021	2376071
Mondrian	3441301	10177581	2311577	144019
Incognito	19801145	14695573	10137021	2875245

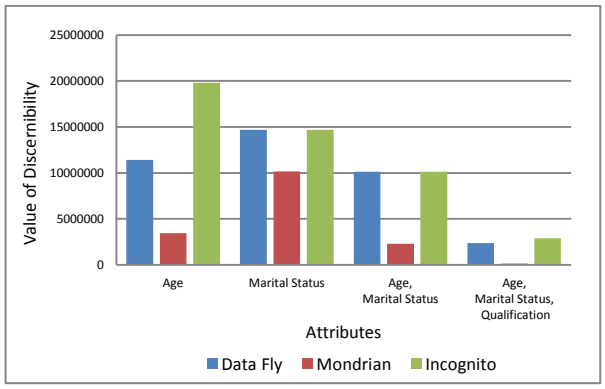

Figure 2: Comparative analysis of the three algorithms for Adult data set

It has been observed from Figure 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) whereas result produced by datafly is marginal good as compared with incognito. It has also been observed that discernibility is minimum when anonymization has been performed with a combination of numeric and non numeric attribute. Moreover, the value of discernibility decreases with increase in the number of attributes for anonymization .

7.2 Discernibility for ATUS data set

Anonymization and evaluation have been done to calculate discernibility 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 4 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 4 Result of discernibility for ATUS data set

Algorithm/ No of QI	Age	Race	Age, Race	Age, Race ,Marital Status
Data Fly	1028322257	2389955961	1461693267	716511535
Mondrian	60704599	2322775237	47680211	43698515
Incognito	1028322257	2389955961	1461693267	866038249

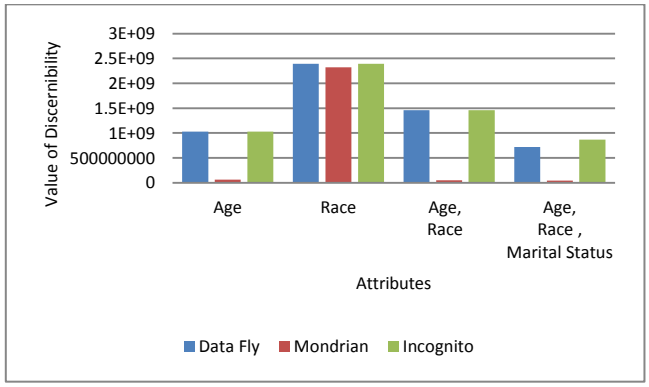

Figure 3: Comparative analysis of the three algorithms for ATUS data set

From the Figure 3 it has been observed that Mondrian outperforms in all cases except the case of single character attribute. The result produce by Datafly and Incognito algorithms are almost equal. Moreover, the value of discernibility decreases with increase in the number of attributes for anonymization .

7.3 Discernibility for CUPS data set

Again anonymization and evaluation have been done to calculate discernibility 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 5 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 5 Result of discernibility for CUPS data set

Algorithm/ No of QI	Age	Qualification	Age, Sex	Age, Qualification	Age, Sex, Qualification
Data Fly	137096367	754109190	692869444	891391266	450051540
Mondrian	85847612	1121197738	82186576	41592726	41184292
Incognito	1370963670	1479157842	692869444	891391266	450051540

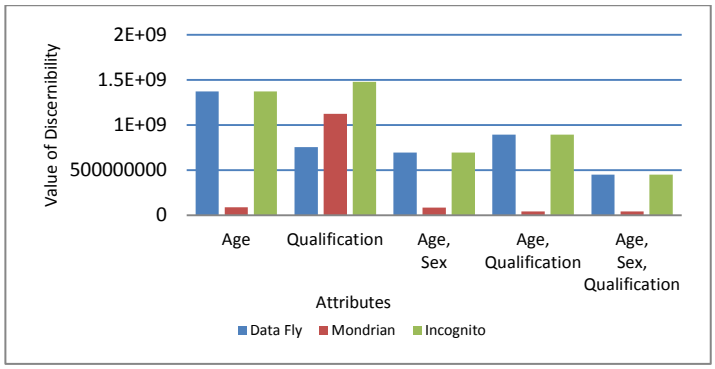

Figure 4: Comparative analysis of the three algorithms for CUPS data set

From Figure 4 it has been observed the performance of Mondrian is better than other two algorithms whereas datafly and incognito performs almost equal. It has also been observed that discernibility is minimum when anonymization has been performed with a combination of numeric and non numeric attribute. Moreover, the value of discernibility decreases with increase in the number of attributes for anonymization .

VIII. CONCLUSIONS

In present spell, many techniques have been proposed by various researchers for anonymizing the data sets and to preserve privacy while publishing. This paper provides an extensive analysis for different data sets with different dimensions and characteristics with reference to that it can be derived that none of the anonymization algorithms always performs to give consistent results with every types of attribute, and the value of discernibilty depends upon number of quasi attributes. Moreover, general performance of Mondrian is better than the Datafly and Incognito. Discernibility in case of Incognito algorithm is more than the other and on comparing Incognito with Datafly the performance of Datafly is better than Incognito. Moreover, It has been interpreted that the value of discernibility decreases as number of attributes increases for anonymization. Furthermore, if anonymization has been performed on the basis of attribute with small distinct domain set then Mondrian does not perform to give good results. So, there is a

scope of enhancement of methods that provides minimum information loss.

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 Lefebvre, David J. DeWitt. *"Mondrian Multidimensional K-Anonymity"* , In proceeding of 22nd International Conference on Data Engineering, ICDE'06, page 25,2006.

[3] Kristen Lefebvre, 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] J. Gantz and D. Reinsel. *"The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"*. Technical report, IDC, sponsored by EMC, December, 2012.

[5] M.Manigandan and K. Aravind Kumar, "*Secure Multiparty Protocol for Distributed Mining of Association Rules*", International Journal of Scientific Research in Computer Science and Engineering, Vol.3, Issue.1, pp.6-10, 2015.

[6] 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.

[7] J. Soria-Comas , J. Domingo-Ferrer, D. Sanchez, and S. Martınez, *"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, pages 372–379, 2013.

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

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

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

[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] Nergiz, M. E. and Clifton, C. *"Thoughts on k-Anonymization"*, Data and Knowledge Engineering, 63(3):622–645, 2007.

[14]DataSource:https://drive.google.com/open?id=0B1QMEQlbBZ9zMy1LU0FEaXprem8

[15] 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.

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

