Data Utility Metrics for *k*-anonymization Algorithms

Deepak Narula
Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119
Email: dnarula123@yahoo.com

Pardeep Kumar
Associate Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119
Email: mittalkuk@gmail.com

Shuchita Upadhyaya
Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119
Email: Shuchita_bhasin@yahoo.com

Abstract

Privacy remains a major serious issue while publishing the data. In today's environment some organizations publish their data for the purpose of research. Although during publishing organization keeps a check on attributes which can uniquely identify the information for individual, yet sometimes published information may prove to be an asset for an attacker, which may retrieve sensitive information about an individual. As a result, data protection along with its privacy is an important domain of research. Multiple techniques of anonymization have been proposed to secure the privacy of an individual and to reduce information losses. Out of the provided anonymization techniques k-anonymity is one of the most popular technique which is based on the concept of generalization and puts a check on such type of attacks that helps to anonymize the data sets. In this paper first a brief discussion of various k-anonymity algorithms is presented. Further, various metrics have been discussed for verification of different data sets along with their illustration.

Keywords- Algorithms, Equivalence, *k*-anonymity, Metric, Privacy-Preserving Data Publishing(PPDP).

1 Introduction

Advances in the field of data storage, collection and inference techniques have enabled the formation of colossal database encloses personal information. This huge collected information always provides an opportunity for decision making. But this information in its native form may contain some personal sensitive information about an individual leakage of that creates a serious threat. As a result of that privacy preservation data publishing is always an area of interest for researchers [1].

Various anonymization algorithms have been proposed to anonymize the data but selection of the most appropriate technique is always a major issue. Moreover, verification of such anonymization technique with different data sets is always required to check the suitability of an algorithm for a particular data set. In literature different metrics have been proposed. In this paper a discussion on various data utility metrics has been done. This paper comprises of 5 sections as follows: section 2 provides relative background information along with required definition whereas section 3 provides various anonymity techniques, section 4 provides various data metrics for anonymiztion and shows how to apply these metrics on data sets. Finally, section 5 concludes and provides direction for further work.

2 Preliminaries

Development of various methods for secrecy of sensitive data is always a main intent of PPDP. There are various methods which are applied on data sets to anonymize it and to protect sensitive information. The operators which are frequently used for the purpose of anonymiztion are random perturbation, generalization and suppression etc.

2.1 Random Perturbation This is one of the natural method to anonymize numerical data by adding a random value to original data results to generate a new value as $Y^{'}=Y+R$ where R is a random value[2]. The method of randomization can also be described as follows. Consider a data set $A=\{a_1, a_2 ....a_n\}$ where a_i Ɛ A, when a noise component is added which are drawn independently and denoted by b_1, b_2........b_n. Thus new set of distorted records are denoted by $\{a_1+b_1 , a_2+b_2............a_n+b_n\}$.

2.2 Generalization This is another way used to anonymize the data by placing a substantive consistent value against less specific but semantically consistent value [3]. The process of generalization is achieved using hierarchical structure and associated attributes belonging to the nature of quasi-identifiers. In Fig. 1 a hierarchical structure for designation and age is given for school employees whereas Fig. 2 represents the hierarchical structure of pincode for different cities.

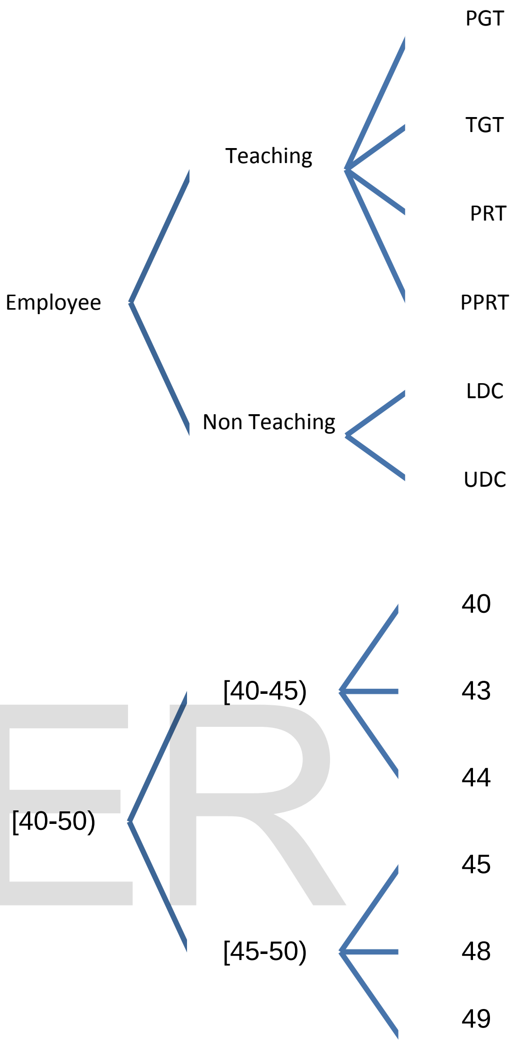

Fig 1 Hierarchical Structure for designation and age

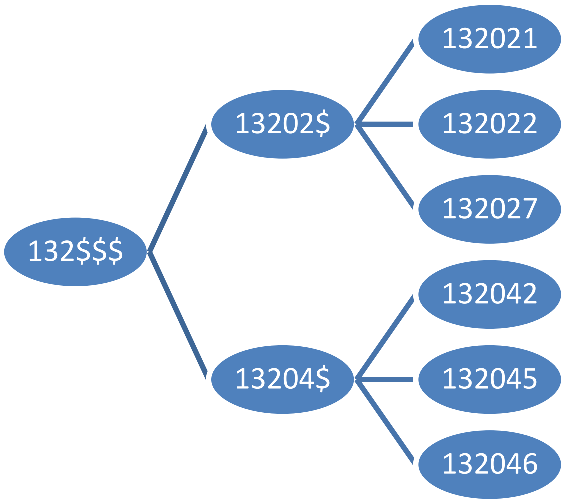

Fig. 2 Hierarchical Structure for Pincode

2.3 Suppression This is a process in which original values of the attributes are replaced with some special symbols such as (@,$ etc.) which makes the value meaningless. This process is applied on quasi-attributes as shown in Table 1

Table 1. Suppression up to different levels

Pincode	Suppression		
Level 0	Level 1	Level 2	Level 3
900231	90023$	9002$$	900$$$
900221	90022$	9002$$	900$$$
900210	90021$	9002$$	900$$$

3 **Anonymity Algorithms**

Various anonymity algorithms exist in the literature which are used to anonymize the data. A brief discussion has been given on various k-anonymity algorithms as this is a fundamental principle of privacy.

***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 a table is as [4][5]. "A table T is k-anonymous with respect to Quasi-Identifiers Qi (Q_1,……., Q_d) if every unique tuple (q_1,….q_d) in the projection of T on Q_1,….Q_d occurs at least k times". For example Table 2 represents the original table containing data about school employees where as Table 3 represents the anonymized data with k=3.

Table 2 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	Durgesh	PPRT	45	132045	34000
6	Raghav	PGT	43	132027	55000

Table 3 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

3.1 **Datafly** Datafly algorithm of anonymization is based on the concept of full domain generalization and also based on greedy heuristic algorithm approach [5]. The data fly algorithm works by counting the frequency over the quasi identifiers and generalize the attributes which have most distinct values until k-anonymity is not satisfied. One of the problem that is associated with data fly algorithm is that it generates all values associated with an attribute and suppress all values within a tuple. This algorithm produces generalization which satisfies k-anonymization but may not produce k-minimal distortions.

3.2 **Incognito algorithm** This algorithm works on the concept of full domain generalization and uses single dimensional method [3]. 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** 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 [6]. This algorithm start it’s processing by selecting least specific value of the attribute in the QID. This also uses the attribute with widest ranges of values. Moreover, quasi identifiers are represented by spatial representation.

However, size of domain space grows exponentially with the number of target attributes.

4 Data Metrics for *k*-anonymity algorithms

Evaluation of anonymity algorithms is necessary to analyze that which algorithm of anonymization algorithm is best suited for a particular type of data set. In this section a discussion has been made on various general purpose metrics and how these can be applied on data sets to calculate generalized information loss, how a record is indistinguishable from the other and how an equivalence class approaches to the best case.

4.1 Generalized Information Loss This metric is used to calculate the amount of forfeiture incurred when a specific attribute is generalized. In the given metric [7] for calculating the generalized loss L_i and U_i be the lower and upper bounds of an attribute i. A cell entry for attribute i is generalized to an interval ij defined by lower bound L_{ij} and upper bound U_{ij} these are two end points, whereas the total information loss for an annonymized table is calculate as:

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

Whereas |T| represents the cardinality of table, n represents the total number of attributes U_{ij} , L_{ij} represents the upper and lower bound of cap points and u_i and l_i represents upper and lower bound for attributes i.

To illustrate the concept of generalized Information loss metric Table 3, Fig.1 and 2 for designation, age and pincode will be taken, and for quasi attributes containing non numeric data a numeric value in range is assigned against each attribute’s value(for e.g. PGT is mapped with 1, TGT is mapped with 2 and so on). So for the attribute designation, which is a non numeric, the Gen Loss for cells with value Teaching is $\frac{(4-1)}{(6-1)} = \frac{3}{5}$. For age, a numeric attribute the GenLoss for [45-50) will be calculated as $\frac{(49-45)}{(49-40)} = \frac{4}{9}$ and similarly for Pincode the GenLoss will be calculated as for 13204$ as $\frac{(6-4)}{(6-1)} = \frac{2}{5}$ and thus the total GenLoss for the whole table is $\frac{1}{6*3} * \frac{78}{9} = \frac{13}{27}$

4.2 Discenibility Metric This metric is used to calculate that how a record is indistinguishable from the other available in a table T [8]. 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

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

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

To illustrate the concept of Discenibility Metric consider Table 3, as in table both classes of size 3 each, thus the total value for this will be $3^2+3^2=18$.

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 [9]. 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 3 will be considered which shows two equivalence classes, the C_{AVG} value will be $\frac{6}{2*3} = 1$

5 Conclusion and Future work

Data in its original form contains sensitive data therefore anonymization is applied but during the process of anonymization various information losses may occurs. In this paper, various verification approaches have been discussed to check the losses incurred while annonymizing the data set. In addition to this, various data utility metrics have also been discussed which helps to identify how efficient is a particular anonymization technique is. These metrics also helps in identifying that how much information losses may occur during an anonymization process.

In future, these metrics will be used to check and to investigate the performance of various k-anonymity algorithms on different publically available data sets.

References

[1] Gantz, J. and Reinsel, D., "*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.

[2] Aggarwal, Charu C. and Philip S. Yu.A., "*General Survey of Privacy-Preserving Data Mining Models and Algorithms*", Volume 34 of the series advances in Database Systems pp 11-52,2008.

[3] LevFevre, K.., J.Dewitt, David and Raghu, R., "*Incognito: Efficient Full-Domain k-anonymity*". In Proceeding of ACM SIGMOD,pp 49-60, New York,2005.

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

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

[6] LevFevre, K., J. Dewitt, David and Ramakrishnan, R., "*Mondrian Multidimensional K-Anonymity*", In Proceedings of the 22nd International Conference on Data Engineering, ICDE '06, page 25, 2006.

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

[8] 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.

[9] LeFevre, Kristen, J. DeWitt, David, "*Workload-Aware Anonymization*", KDD'06, August 20–23, Philadelphia, Pennsylvania, USA. Copyright 2006 ACM 1-59593-339-5/06/0008, 2006.