Survey on Hybrid Anonymization using k-anonymity for Privacy Preserving in Data Mining

Binal Upadhyay[*], Sumitra Menaria[*],

[*]*Department of Computer Engineering, Parul University*
Limda, Waghodia Road, Vadodara, Gujarat 391760,India
[1] Email: binal1994upadhyay@gmail.com

Department of Computer Engineering, Parul University
Limda, Waghodia Road, Vadodara, Gujarat 391760,India
[2]Email: Sumitra.menaria@paruluniversity.ac.in

***Abstract*—**K-anonymity is the one of the popular privacy preserving model. In the data mining there is multiple technique is available k-anonymity is one of the technique which is used for the protecting privacy in the database. In this paper our main approach is hybrid anonymization. The main thing of this technique is that it is the mixing of two techniques. We introduce hybrid anonymization with hybrid generalization which is formed by not only generalization but also the data relocation. Data relocation serves trade-off between truthfulness and utility. Using the hybrid anonymization we maintain the privacy standard such as k-anonymity. In the previous research we find that k-anonymity is not good work with multiple sensitive data and there is more information loss occurs for that issue we use hybrid anonymization on multiple dataset. We show that our model can decrease the information loss in minimum time period.

***Keywords*—** *Data Mining, K-anonymity, Privacy Preserving, Hybrid Anonymization, Data relocation, Hybrid generalization*

*****

I. INTRODUCTION

K-anonymity is one of the well-known anonymizing approaches proposed by Samarati and Sweeney [16]. We can say that if a data complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least k-1 records whose data also appear in the data set. For the checking the k-anonymity requirements use the generalization and suppression methods are used for different datasets [1][16]. Now we conclude some basic thing about that how to we preserve the privacy in the k-anonymity model for that we using the PPDM techniques. There is many privacy preserving techniques available in the data mining. In all that technique k-anonymity is the one of PPDM technique What is the PPDM the basic thing is the extend traditional data mining technique to work with the modified data for hiding the sensitive attributes there are two approach is available one is the SMC and second is the Anonymization. In the secure multiparty computation it use the cryptographic approach for privacy preservation goal to create methods for parties to jointly compute a function over there inputs while keeping those input privates. In the anonymization some data is replace with the some modified related attribute from the overall data base.[2]There the highly popular two technique is available for the k-anonymity which is the generalization and suppression. One of the bigger drawback of the available generalization technique is that it is the manually generated domain hierarchy trees are required for every quasi-identifier attribute of the overall datasets before k-anonymity can be applied [3] hiding the data and that gives the most security for the datasets [1]. In the using hybrid anonymization we also used the hybrid generalization because of the draw back of the generalization method. In this our paper we try to achieve the incognito mode with k-anonymity because using the incognito mode with k-anonymity we take the more security from the previous work. Using the incognito mode we don't always put the value predict [11]. An anonymization must not only satisfy the underlying privacy requirement but it also preserve the data utility. If it is not work then it would be difficult to extract some useful data information from the anonymized data [4]. Using the anonymization via generalization at the cell level can proceed in two steps in the first all the records are divided into many groups like that each group contains at least k records. Then, the records in each group are generalized such that their values at each quasi-identifier are same means identical. To minimize the information loss incurred by the second step, the first step should place similar records with respect to the quasi-identifiers in the same group [12][3]. In the Anonmization we can say that the a person which is the nameless means the person hide the identity of him or herself just because of the privacy purpose the main thing of the privacy is the protect the people in the some competitive situation because they hide information

from the other people because some time they feels the embarrassing for other people to know about it. We call as the privacy is the limited accesses to a person and all the features related to a person [2]. In this paper we introduce k-anonymity with different attacks, Hybrid anonymization, PPDM techniques and incognito mode and the main techniques for the k-anonymity model and how it is work and which is the benefits of this technique and misuses of this technique. Over all of things that the how working hybrid technique with the privacy protecting technique.[14][15]

II. RELETED WORK

An easy Privacy protection technology is one type of firstly popular academic research which has many applications which are famous in many areas in recent years [2][16]. K-anonymity is one of the techniques which help in releasing a huge amount of data.

A. K-anonymity attacks.

The k-anonymity produces anonymized dataset in which each record is in-distinguishable from at least other k-1 records.

K-anonymity model performed on the basis of three attacks. First is Linking attack. Second is Background knowledge attack. And third is Homogeneity attack [6] [16].

[1] Linking attack:

TABLE I
MEDICAL DATABASE

Age	Gender	Zip Code	Disease
25	Female	2213	Cancer
29	Male	2214	Ulcer
45	Female	2210	Flu
49	Male	2217	Herat Problem

TABLE II
VOTER DATABASE

Name	Age	Gender	Zip Code
Riana	25	2213	Cancer
Harry	29	2214	Ulcer
Charlie	45	2210	Flu
Alice	49	2217	Heart Problem

In the table 1 and table 2 we so the one is the medical database and another is voter database. Voter database is publically available database that can reveal the identity of individual. Now in the Linking attack. Now here we can easily find that the Riana has a Cancer problem.

[2] Homogeneity attack:

In the Homogeneity attack all the sensitive values in each equivalence class are identical. In such case, even though the data is anonymized, the sensitive value of an individual we can be predicated the value [6].

TABLE III

3 ANONYMIZED MEDICAL DATABASE

Equivalence Class	Age	Gender	Zip Code	Disease
1	2*	Person	221*	Heart Problem
	2*	Person	221*	Heart Problem
2	4*	Person	221*	Ulcer
	4*	Person	221*	Cancer

TABLE IV

HOMOGENEITY ATTACK

Riana	
Zip code	Age
2213	25

In this attack table 3 shows the 3-anonymizezd medical database and table 4 shows the homogeneity attack. Now we gives the example how find the data from the anonymized data.

Here we can say that Riana and the Harry are hostile neighbours. One day Riana falls ill and she is taken by ambulance to the hospital. Having seen the ambulance, Harry sets out to find what aliment Riana suffering from. For that he fined 2-Acknolegement table which is published by that hospital and he find that acknowledgment table and he is her neighbour so he know she is American women and her postal code is 2213 and she is 25 year old. Now he check that table and he found that Riana's record number is 1 and he found that all of those patients have the same medical condition along these lines harry conclude that Riana has cancer.

[3] Background knowledge attack.

In the background knowledge attack, the sensitive attribute can be identified based on the association between one or more quasi identifier attributes. For example, Machanavajjhala, Kifer, Gehrke, and Venkitasubramaniam (2007) showed that "knowing that heart attacks occur at a reduced rate in Japanese patients could be used to narrow the range of values for a sensitive attribute of a patient's disease" [6].

. TABLE V
BACKGROUND KNOWLEDGE ATTACK

Charlie	
Zip code	Age
2210	45

From the table 3 and table 4 we show one is the 3-anonymized database and another table is example of background knowledge attack. For the background knowledge attack we take the one example like previous example. Here harry is the pen companion named Charlie who is admitted in the same hospital where Riana is admitted. And whose patient record also shows up in the table which is demonstrated in the table 3. Harry know that Charlie is 45 year old Japanese female who currently lives in postal code 2210 and her record available in the record no 3.Harry know that Japanese have an too low incidence of the heart disease. Along these lines harry concludes with close certainty that Charlie has ulcer problem.

B. Different PPDM Techniques

In the previous work there is many PPDM techniques are available. [7]. K-anonymity is the one of the best PPDM technique. PPDM has multiple techniques but there is popular technique is available in table which has some merits and demerits shows in the table. Here PPDM has two area first is data modification based technique and other is SMC based technique [7][2]. In the data modification we can change, delete, add, the data which is called as the data modification means we can easily modified the data. On the other hand in the SMC it create the parities to jointly compute a function over that input while keeping those input is being private. In the simple way we can say that a set of parties with some private inputs [7] [14].

TABLE VI
MERITS AND DEMERITS OF THE PPDM TECHNIQUES [7]

Techniques	Merits	Demerits
Perturbation	Different attributes are preserved like that is separate. It has high data utility.	Privacy preservation is very less. If we want to reconstruct the original data that is not possible.
Condensation	It is good performed with the stream data sets.	There is large amount of information loss occur.
Anonymization	There is individual privacy is maintained.	Use linking attack. Heavy information loss occurs.
Differential Privacy	Accuracy of results and improved utility.	There is problem is that Scalability level is still a question
Evolutionary Algorithms	It is more secure and effective.	High uncertainty.
SMC	Accuracy of results Effective. Transformed data are exact and more protected.	Complicated when more than two parties are involved. And it is more Expensive.

III. ALGORITHMS

The k-anonymity model work with some algorithm and gives the different type of result. K-anonymity is the one most privacy preservation model that provide the protection on the anonymized data sets. Some algorithms are emphasized here.

A. Hybrid anonymization Algorithms

In hybrid anonymization there is available algorithm is described here,

[1] Single dimensional hybrid k-anonymization
[2] Multi-dimensional hybrid k-anonymization

In the first single dimensional hybrid k-anonymization algorithms each addressed a different adversary .This algorithm is based on the optimal single dimensional anonymization algorithm, it improves incognito mode by searching the space into the hybrid generalization [4]. There is available the deterministic S-hybrid in this case the relocation tuple two phases after the overall relocation the algorithm check if the maximum number of data allowed relocations has been exceed. If in the some cases apply algorithm doing the roll back on the last relocation and returns to the known k-anonymity hybrid generalization generated as so far. In the other hand in randomized s-hybrid can attempt to reverse engineering and available algorithm create non k-anonymous sub groups. In the most of cases time and size of these sets are large enough to create probability space of the available size for this algorithm [4]. Now another remaining algorithm is statistical S-hybrid in this statistical adversary known use the as the known distribution of the tuples identify article relocation. It should be noted as even if the large α settings like α-hybrid anonymity which is a strict privacy definition.

Now we come to the next algorithm which is the multidimensional hybrid k-anonymization

algorithm for that used in the three adversaries. This algorithm is based on the Mondrian LeFevre et al. (2006) [4]. It is used for the hybrid creation and in the generalization. In the hybrid creation it adapting the mainly M-hybrid algorithm so that it gives the protection against statistical algorithm-aware adversaries involves modifying the creation of the hybrid function [4]. In this case we can say that the goal of this paper is to create that type of anonymous datasets where the predictive performance of the classifier trained on the anonymous data set is as similar and s possible to depend on the performance of the classifier [3].

B. MAGE

This algorithm mainly used for the mixed data. MAGE is called as the micro-aggregation. We can say as that neither generalization nor micro-aggregation anonymized mixed data very effectively. MAGE used the mean vector of numerical data to preserve more semantics data from the generalization. MAGE also uses some generalization values to replace the categorical data. MAGE can also adopt the two different categorical attributes for the anonymized mixed data. TSCKA algorithm is also used for the mixed data. For the local recording algorithm can generate better quality for the anonymous dataset however their computing complexity is general high, and it is especially good for the large data set. It is the used for the improve process of the performance. [8][9].

The KACA algorithm is a bottom-top clustering algorithm. The time complexity of KACA is $O(n^2)$.KACA is less efficient than TSCKA. TSCKA gives the better result and more efficient and there is also a decrees the overall information loss. Information loss is calculate using the how much value increase of k. That means Information loss increasing with the k-increase [8]. We can say that the MAGE method retains more semantics for mixed data then micro aggregation and generalization method [8].

C. KAMP

In the KAMP algorithm two pattern are include which are,

[1] Generalization
[2] Suppression

Basically in case of generalization there we will predict the value because in that value put in the one range like one person who age 35 in that case in the generalization we put the value in range like 33-36 like that and in the suppression the value of attribute are replace with some special value like "*" for that we get one example age with value [39] is generalized as [3*].In that case we don't find the exact value that gives high privacy. For the generalization and suppression we take example in the table [9] [10]. In this generalization and suppression both are example is shown in figure which are expressed here. Here suppression is better than the generalization method.

TABLE VII
EXAMPLE OF GENERALIZATION AND SUPPRESSION

#	Zip	Age	Nationality	Condition
1	130**	< 35	*	Heart Disease
2	130**	< 35	*	Viral Infection
3	130**	< 35	*	Flu

Generalization Suppression

D. The greedy algorithm

In the case of the greedy algorithm is that is the instead of striving to build a k-regular generalization graph over the data at once, we can do one thing is that we can set the data in a sequence of k distinct iterations and adding a whole single assignment to the graph under the construction at each iteration. We can say that this algorithm is designed to achieve optimum solution for a given problem in the data set. We can say that in greedy algorithm approach, whole the decisions are made from the given solution from the given domain. As being greedy, the closest solution of that seems to the main aspects is this algorithm provide an optimum solution which is chosen. Greedy algorithms trying to find a localized optimum solution for the data set, which are may eventually lead to globally optimized solutions. However, generally greedy algorithms do not provide globally optimized solutions. We find a greedy algorithm have not time complexity by the experiment we find that the data utility is gain up 41% and also gives the efficiency advantages [19]. By the result we show that greedy algorithm works for the practical values of the k used which is used in the real world settings and also find that linear-time–

 ISSN: 2321-8169
350 – 355

back-tracking for the greedy process which does not affect the complexity O(n^3) which is the iteration complexity. We find that overall complexity of the iteration is O(kn2). We show the advantages apply on the time efficiency [15].

E. Comparison table for algorithm

TABLE VIII
COMPARISON [2][8][4][15]

Algorithm	Parameters		
	Complexity	Efficiency	Time cost
KACA	O (n^2)	Low	High
KAMP	O$(logn)$	High	Low
Hybrid anonymization	O$(n\ logn)$	Need Improve	High
Greedy Decision tree	O$(n\ logn)$	High	Low

IV EXPERIMENTAL RESULT

A. Comparison graph

Here we take the different data for the check the information loss and time cost. And result is shown in figure.

TABLE IX
Example of generalization and suppression

Different Data	Time Needed	Information Loss
10000	0.3 sec	3.999
100000	0.3 sec	0.1
30000	0.3 sec	3.315
50000	0.3 sec	1.917
500000	0.5 sec	0.005

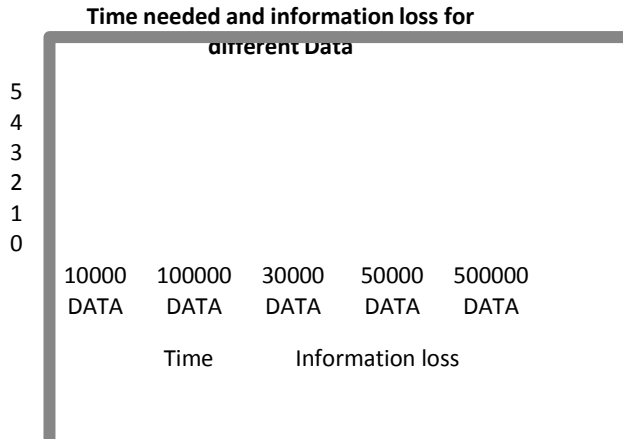

Fig.1 Graph of time needed and information loss for data.

Here show in figure there is different data is available. Here we can perform one the different data sets.

First of all we check for the 10000 data and here it found that in the o.3 sec there is 3.999 information occurs it performers is the basis on the using outliers and without outliers. Now on the 100000 data it takes the also 0.3 sec but we improve the data so there is less information loss occurs which is the 0.1. Now we perform on the 30000 data it also take the 0.3 sec and it information loss is 3.315. On the next point we check for 50000 data that takes the same time then the others the time is 0.3 sec on that data the range of information loss is 1.917. Now on the last we maximize the data from the original data sets and it became the wider as 500000 data and it will take the time is 0.5 sec and the information loss is 000.5 which is lesser then the others.

B.Line chart for data utility

We can say that It is a more important issue for utility of data privacy protection. We can say that in order to hide sensitive information, false information should insert the database, or block data values. Although sample Techniques do not modify the information stored in the database, but that, since their information is incomplete, still reduces data utility. More changes to the database, less data utility of the database. So estimated parameters of data utility is data information loss applied privacy protection. Of course, the estimate of information loss related with the specific data mining.

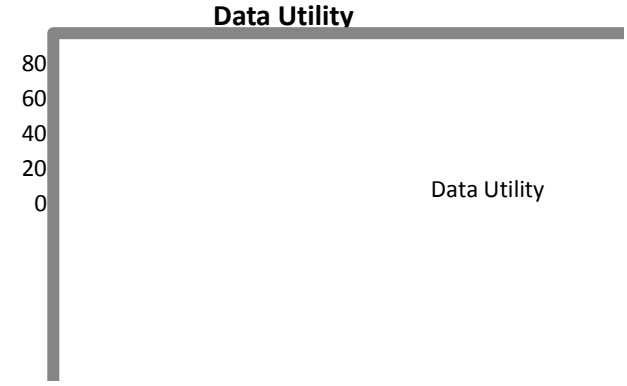

Fig. 2 A sample line graph for the data utility.

Here line graph shows that the all the different algorithm's data utility graph from that graph we find the hybrid anonymization have the highest data

utility and KAMP have lowest data utility and MAGE and k-anonymous decision tree need to improve that data utility that we find hybrid Anonmization is the better in the performance using the hybrid anonymization we get the better result in k-anonymity and also in the multi-dimensional k-anonymization.[8][4][15].

IV. Conclusions

Finally, I conclude that privacy preserving in data mining ids the main aspect to provide the privacy. Privacy is necessary to protect people in competitive situations. Using the k-anonymity with anonymization and using the suppression and generalization method for the more secure database it provide the security to the different type of datasets. K-anonymity is important privacy preserving model for the data mining. We also show the complexity, time cost, efficiency and complexity of our experiments. Privacy in data stream mining, Efficiency and minimum computation cost in distributed PPDM, Privacy and accuracy with minimal loss.

References

[1] Slava Kisilevich, Lior Rokach, Yuval Elovici, Member, IEEE, and Bracha Shapira "Efficient Multidimensional Suppression for K-Anonymity" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 3, MARCH 2010

[2] Xinjun Qi , Mingkui Zong School of Technology,Harbin University An Overview of Privacy Preserving Data Mining, 2011 International Conference on Environmental Science and Engineering (ICESE 2011)

[3] Jun-Lin Lin Department of Information Management Yuan Ze University, Taiwan jun@saturn.yzu.edu.tw Meng-Cheng Wei Department of Information Management Yuan Ze University, Taiwan mongcheng@gmail.com Chih-Wen Li Department of Information Management Yuan Ze University, Taiwan s966219@saturn.yzu.edu.tw
Kuo-Chiang Hsieh Department of Information Management Yuan Ze University, Taiwan s956210@saturn.yzu.edu.tw A Hybrid Method for k-Anonymization 2008 IEEE Asia-Pacific Services Computing Conference.

[4] Mehmet Ercan Nergiz, Muhammed Zahit Gok, Hybrid k-anoymity, ScienceDirect, COMPUTERS & SECURITY 44 (2014) 51-63 journal homepage: www.elsevier.com/locqte/cose

[5] Qi Jia∗, Linke Guo∗, Zhanpeng Jin∗, Yuguang Fang† ∗Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA †Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA Email: {qjia1, lguo, zjin}@binghamton.edu, fang@ece.ufl.edu Privacy-preserving Data Classification and Similarity Evaluation for Distributed Systems 2016 IEEE 36th International Conference on Distributed Computing Systems.

[6] Arik Friedman · Ran Wolff · Assaf Schuster Providing k-anonymity in data mining Received: 30 September 2005 / Revised: 24 May 2006 / Accepted: 2 August 2006 / Published online: 10 January 2007 © Springer-Verlag 2007.

[7] G. Arumugam Senior Professor and Head, Department of Computer Science Madurai Kamaraj University Madurai, Tamilnadu, India. V. Jane Varamani Sulekha Research Scholar, Department of Computer Science Madurai Kamaraj University Madurai, Tamilnadu, India. IMR based Anonymization for Privacy Preservation in Data Mining.

[8] Jaimain Han, Jaun Yu, Yuchang Mo, Jianfeng Lu, Huawen Liu, MAGE: A Semantics retaining k-anonymization method for mixed data. journal homepage: www.elsevier.com/locate/knosys 0950-7051/$ - see front matter 2013 Elsevier B.V. All rights reserved.

[9] Chia-Hao Hsu Department of Electrical Engineering National Chung HsingUniversityTaichung,Taiwan,R.O.C.Email:w100064001@mail.nchu.edu.tw Hsiao-Ping Tsai Department of Electrical Engineering National Chung Hsing University Taichung, Taiwan, R.O.C. Email: KAMP: Preserving k-anonymity for Combinations of Patterns 2013 IEEE 14th International Conference on Mobile Data Management.

[10] T. Pranav Bhat, C. Karthik∗ and K. Chandrasekaran Department of Computer Science and Engineering, NITK Surathkal 575 025, Karnataka, India A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory T. Pranav Bhat, C. Karthik∗ and K. Chandrasekaran Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015).

[11] Fran Casino, Joshep Domingo-ferrer, constantinos, Domenec puing, Agusti solans, A k-anonymous approach to privacy preserving collaborative filltwering Journel pf computer and system science 2014.

[12] Aggarwal, C., Yu, P. S. A condensation approach to privacy preserving data mining. In proceedings of International Conference on Extending Database Technology (EDBT), pp.183–199, 2004. 746.

[13] Kun Liu, Chris Giannella, and Hillol Kargupta . A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods. Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, Springer, (2008)..

[14] Arik Friedman · Ran Wolff · Assaf Schuster Providing k-anonymity in data mining Received: 30 September 2005 / Revised: 24 May 2006 / Accepted: 2 August 2006 / Published online: 10 January 2007 © Springer-Verlag 2007.

[15] Matthew Andrews Bell Labs, Murray Hill, NJ a Gordon Wilfong Bell Labs, Murray Hill, NJ Lisa Zhang Bell Labs, Murray Hill, NJ Analysis of k-Anonymity Algorithms for Streaming Location Data The Third International Workshop on Security and Privacy in Big Data (BigSecurity 2015).