Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 4): pp. 1164-1170
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.04.130

An Approach for data confidentiality and Logical data Security on data release by Multidimensional K-Anonymity Model

Abhijit J. Patankar, Research Scholar, VTU,Belagaum, Maharashtra, India, abhijitpatankarmail@gmail.com
Dr. Kotrappa Sirbi, Professor, Department of CSE, KLE society's Dr.M.S.S.C.E&T, Belgaum
Dr. Kshama V. Kulhalli, Principal,D.Y.Patil C.E.& T , Kolhapur, Maharashtra, India

ABSTRACT: In our daily routine life there is huge and significant increase in frequently collection of data having different varieties and it growing in huge quantity sometimes we need to share this information and Sharing this huge information is very useful for researchers, Marketing peoples and for business use as this sheared information might be useful to invent new things or to earn money but, Publishing the data may put person privacy on stake and protection of users privacy is a challenging task during micro data release, here we recommend a method which protects users privacy by applying Multidimensional K-Anonymity Method there are to methods single and multidimensional in multidimensional we need to identify level of K-Anonymity and apply Generalization and Suppression methods for users data privacy protection. Also we identify values of k , CAvg,CDM as Quality measures and plot the values of growing value of K to prove that Nearest Neighbor is effective technique for achieving data confidentiality and logical data security on data release.

Key words:- **CAvg, CDM**

I. INTRODUCTION

Once a database is generated for some specific information like critical medical disease, financial ups and downs etc. then it is sometimes necessary to share such data among public for the purpose of knowledge, research or for finding the facts and it is sometimes beneficial also in terms of earning money, publishing data also facilitate additional innovative analysis on the same datasets and to obtain feedback for improving data quality for ongoing efforts of data collection, sometimes there is policy of government to publish research data online for benefit of society ,[1-8]In future it is mandatory to all researchers to publish their data and test cases they did online. But problem with release of this kind of sensitive data is confidentiality of the person or group of persons is on stake due to linking of various such different records and which results in to disclosure of confidential information such as critical disease, financial status relationships etc. this information is identified with the help of three types of attacks on the released data Linking Attack, Background Knowledge Attack and Homogeneity Attack [10]

	Identifiers	**Non Sensitivity data**			**Sensitivity data**
No	**Name**	**Pincode**	**Age**	**Person Nationality**	**Disease**
1	Sachin	411033	28	Indian	Heart Disease
2	**James**	**13067**	**29**	**American**	**Heart Disease**
3	Vinod	411044	35	Indian	Viral Infection
4	Umeko	13067	36	Japanese	Cancer

Table 1.1 Published Data by Medical Agency

Name	**Zip**	**Age**	**Nationality**
John	13053	28	American
James	**13067**	**29**	**American**
Chris	13053	23	American

Table 1.2 Voter List Data

For Demonstrating Linking Attack Consider above two different tables in first Table there is data of Patients having Medical diseases and also in second table there is a Voter List published in general for the benefit of all peoples of society for voting Now we can correlate Sr No.2 Entry with Sr No.2 Entry of Medical Agency Published data and we came to know the sensitive Information about the person Bob by linking these two tables and conclude that Bob is having Disease of Heart this is example of Linking Attack

For any given set of data there are following attributes such as Identifiers like Name of the Person ,Quasi Identifiers such as Date of Birth, Sex, Zipcode which is common to all datasets and quasi identifiers are those identifiers which are used to re-identify the datasets and there is a sensitive data in dataset which need to be protected from public disclosure which is having very sensitive information like Disease , financial position etc. which need to be protected

ID	**quasi identifiers**			**Sensitive Data**
Name	Birthdate	Sex	Pincode	Disease
Andre	21/1/79	male	53715	Flu
Beth	10/1/81	female	55410	Hepatitis
Carol	1/10/44	female	90210	Brochitis
Dan	21/2/84	male	02174	Sprained Ankle
Ellen	19/4/72	female	02237	AIDS

Table:- 1.3 Different attributes in released dataset

ceR	**Country**	**Birth**	**Gen.**	**PIN**	**Problem**
T11	USA	1963	F	02141	Berating Problem
T12	USA	1965	M	02141	Pain in chest
T13	USA	1964	F	02138	obesity
T14	USA	1964	F	02138	chest pain
T15	Non-USA	1964	M	02138	chest pain

Table:-1.4 k-Anonymity Example where k=2

1.1 K-Anonymity details

It is a state that when record is released it should have at list k value which are anonymous and in such a way that it will hide the contents of actual data example given in Table 1.4 where value of k=2.

1.1.1 How to create K-Anonymity:-

To create k-Anonymity from the original dataset we need to identify identical dataset values and apply following two methods in **Generalization we can r**eplace the actual value by a equally semantically but *little* specific value

Suppression Data not released at all, Can be viewed as first level of generalization in suppuration we can hide some common part of data item by special symbols like * , #,$ etc Table 1.9 and Table 1.10 shows the example of released micro data and Generalization and suppuration

1.1.2 Purpose of Multidimensional K-Anonymity

In Multidimensional K-Anonymity generalization of the quasi attributes values is based on not only on single attribute but based on other attributes of Quasi Identifiers For demonstrating what is Multidimensional K-Anonymity we will consider example of sample voter data and patient data and how to perform single dimensional and Multidimensional K-anonymity

Name	Age	Sex	Pin
Nitin	24	Male	411044
Ramesh	26	Male	411011
Suchita	30	Female	411033
Suresh	18	Male	411044
Meera	41	Female	411038

Table 1.5: Voter Registration Record

Person Age	Sex	Zip code	Illness Type
24	Male	53710	Flu
30	Female	53712	Hepatitis
26	Male	53711	Brochitis
18	Male	53710	Broken Arm
41	Female	53712	AIDS
28	Male	53711	Brochitis

Table1.6: Patient Data

Person Age	Sex	Zip code	Illness Type
25-28	Male	53710-11	Flu
25-28	Male	53710-11	BodayPain
25-28	Male	53710-11	Hypertension
25-28	Male	53710-11	High Sugar
25-28	Female	53712	Heart Disease
25-28	Female	53712	High BP

Table 1.7:Single Dimensional k-Anonymity

Person Age	Sex	Zip code	Illness type
25-26	Male	53710-11	Flu
25-26	Male	53710-11	Body Pain
27-28	Male	53710-11	Hypertension
27-28	Male	53710-11	High Sugar
25-27	Female	53712	High BP
25-27	Female	53712	Heart Disease

Table1.8: Multi-Dimensional k-Anonymity

Example of 4-Anonymity: As shown in table below original records and 4-anonymous data with generalization and suppression methods on the given dataset is as shown below

sr	Pincode	Sex	Nationality	Disease
1	411026	22	Indian	Leg Pain
2	411028	26	Japanese	Heart Attack
3	411028	22	Pakistani	High Fever
4	411029	21	Australian	High Fever
5	311053	52	Canadian	Sugar
6	311053	54	Chinese	Kidney Disease
7	311050	48	Australian	Dental Disease
8	311050	45	Indian	Dental Disease

Table 1.9:- Released Microdata

sr	Pin code	Sex	Nationality	Disease
1	4110**	<30	#	Leg Pain
2	4110**	<30	#	Heart Attack
3	4110**	<30	#	High Fever
4	4110**	<30	#	High Fever
5	31105*	≥40	#	Sugar
6	31105*	≥40	#	Kidney Disease
7	31105*	≥40	#	Dental Disease
8	31105*	≥40	#	Dental Disease

Table 1.10: 4-anonymous data with generalization and suppression

Here explicit identifiers are removed key- identifiers can be used to again identify individuals sensitive attributes (may not exist!) carry sensitive information

II. MATERIALS AND METHODS

In research on paper [1] all the concepts related to maintaining privacy are discussed no proper algorithm discussed here for preserving privacy on data release In reference paper[2] they discussed

with its advantages and drawbacks ,In reference paper[3] a unique model for discussing privacy is mentioned .In Efficient Multi-dimensional Suppression for k-anonymity[4] different suppression methods are discussed. In reference paper [5] performances of various K-Anonymity Methods are discussed and approaches are highlighted. Mondrian multidimensional K-anonymity[6] is a very popular approach for studying where single and multi-dimensional Anonymity procedures are explained with proper example. The WEKA data mining software: an update[7] all the updates related to WEKA tool for data analysis are discussed and we are using this tool for doing analysis. In refrence [8] in this paper Multidimensional K-Anonymity using Divide and conquer method was explained and comparative results with other methods also discussed. In reference paper [9] explains the concept and paper[10] is a IEEE based paper and explains how minimum loss of information is possible during Anonymization. In reference [11] explains how privacy is preserved in the data mining process and discovery of knowledge. In reference [12]giving comparative study of different data mining tools and advantages and drawbacks of using them. In reference [13] explains use of UI repository and ML concepts, In reference paper [14] explains generalization and suppression methods of Anonymization, In reference paper[15] explains privacy constraints for Relational data. In reference paper [16] explains concept of Optimized Data de-Identification.

2.1 PROBLEM DEFINITION:

Here we are recommending an Approach using nearest neighbor method for achieving released data confidentiality and Logical data Security using Multidimensional K-Anonymity Model which will give better results as compare to other existing methods

III. PROPOSED ARCHITECTURE

3.1 Architecture of the System

For Non-Relational data if we consider the results and compare with our proposed Nearest Neighbour method then we can prove that we can develop a model using Nearest Neighbour method for obtaining Quality results and our proposed algorithm will work for improving anonymity and logical data Security. In Proposed System Architecture data preprocessing is done and resultant updated and modified data is taken for Consideration later we check whether divisible multidimensional is processed or not then if possible divide multi -dimensional to single dimensional dataset else we do the Anonymization using nearest neighbor method and release the records with Anonymized data.

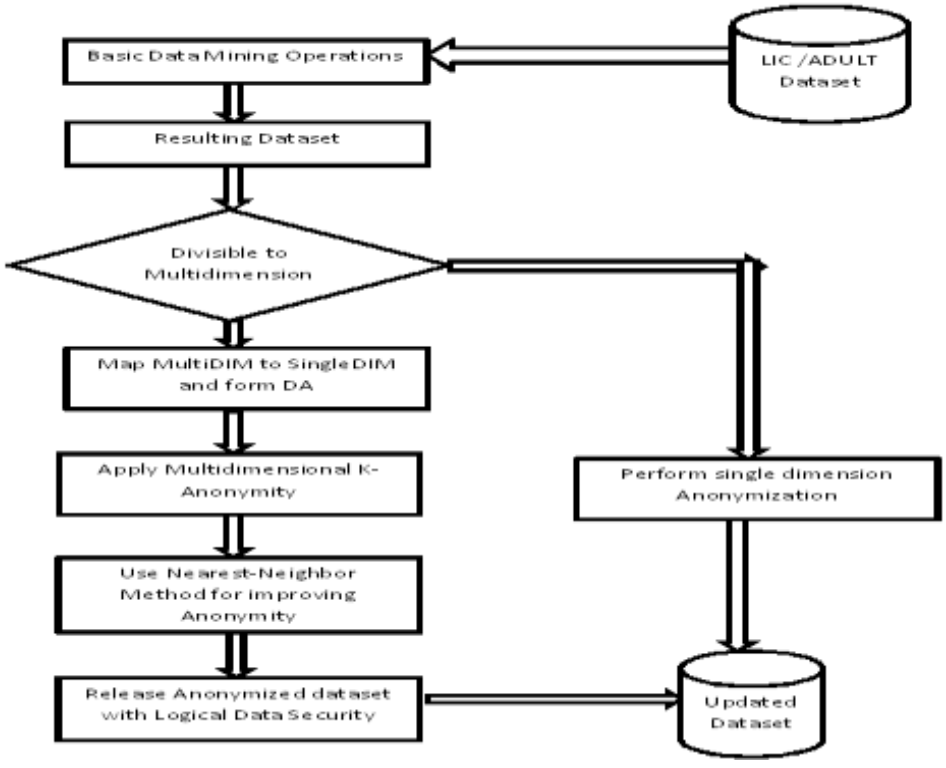

Fig 1. System Architecture

3.2 Proposed Algorithm

In proposed Algorithm we calculate two Quality variables INPUTD and OUTPUTD for calculating degree of Anonymity we first do multi to single dimensional mapping then calculate Dimensional Array DA and generate Anonymized record set based on value of K

Function MULTI_Anonymize (pointSet , INPUTD , k)

Input: LIC OR ADULT Dataset, multi-dimensional locate as INPUTD, degree of anonymization is k.
Output: Resulting l sets after applying anonymization algorithm named OUTPUTD

```
D←
        If ( INPUTD !  DIVISIBLE to MULTI_DIM)
            OUTPUTD← INPUTD
        Else
            For (x = 0 ; x ＜D ; x++ )

            End for

            DA ← choose_DimArray ( Pro1 , PPA1 )
            For ( i = 0 ; i ＜ DA1.length of Array ; i++ )
                If ( Resultant data set= dividable Multi dimension )
                    Apply K-Anonymity on Quasi Identifiers using Multiple dimensions
                else If ( Resultant dataset = dividable single dimension )
                    Apply K-Anonymity on Quasi Identifiers using Single dimensions
                Break from loop
                    Else if ( i = DA1 .length of array – 1 )
                        Return INPUTD→ results OUTPUTD
                    if end
                for end
                    Return Anonymization (Recordset, k)
            End if
Return OUTPUD
```

IV. WORK SCOPE

4.1 Major Areas

In deciding scope of the system various different areas are included as follows

- System will provide logical data security by implementing k-Anonymity
- System will focus on how personal information identification will be avoided
- System will prevent from different attacks after data anonymizaton for maintain confidentiality
- System will implement anonymity before data release which will protect data privacy and data will be also useful for researchers and data scientists for analysis

V. WORK OUTCOME

In the work outcome has shown different values which are plotted for Quality Measures as follows
If we consider CAVG value in fig 5.1 Increase in k CAVG gradually increasealso in Fig 5.2 Increase in K, CDM value for Median and NN are almost same.In Fig 5.3 CDM value are stable after increase in k after some interval ,In Fig 5. 4 Increase in K, CDM value increase and stable after threshold

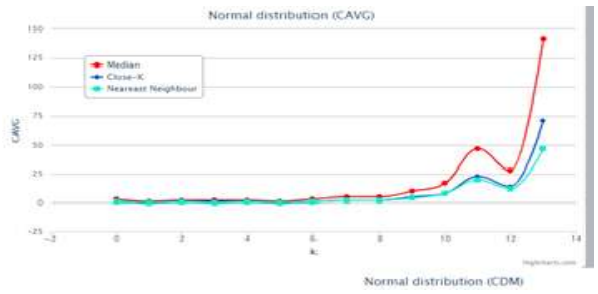

Fig 5.1 Increase in k CAVG gradually increase

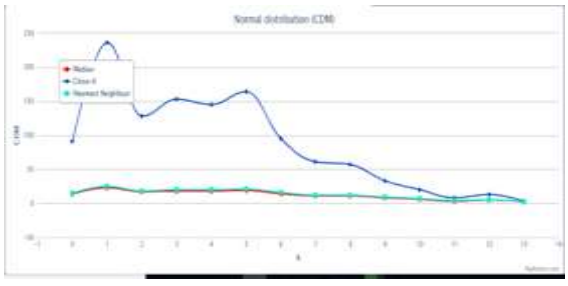

Fig 5.2 Increase in K, CDM value for Median and NN are almost same

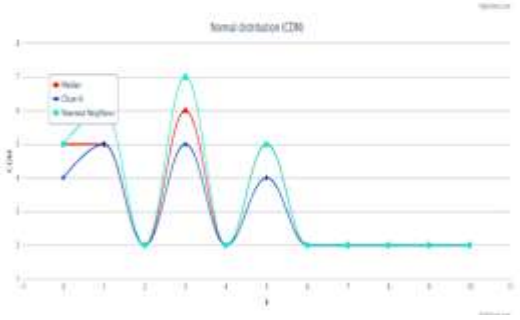

Fig 5.3 CDM value are stable after increase in k after some interval

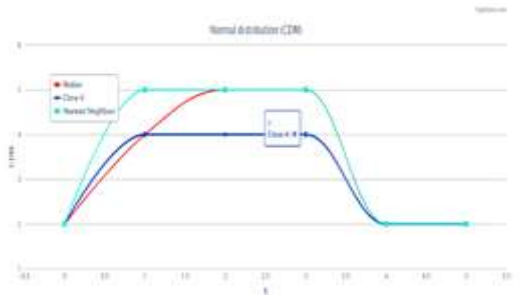

Fig 5. 4 Increase in K, CDM value increase and stable after threshold

5.1 Comparative Study of different methods of K-Anonymity

Following are different methods MEDIAN, CLOSE-K and NEAREST NEIGHBOUR

In Fig 5.5 and Fig 5.6 CAVG is having Maximum Value after increase in value of k so its better method than Median and close K Methods

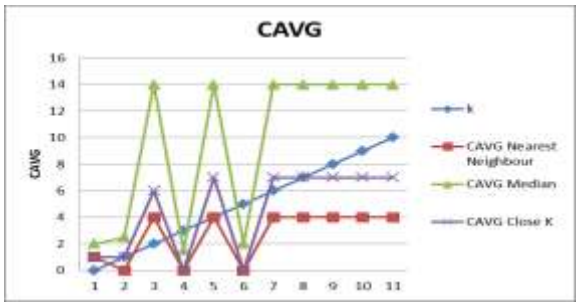

Fig 5.5 CAVG with Increase in K

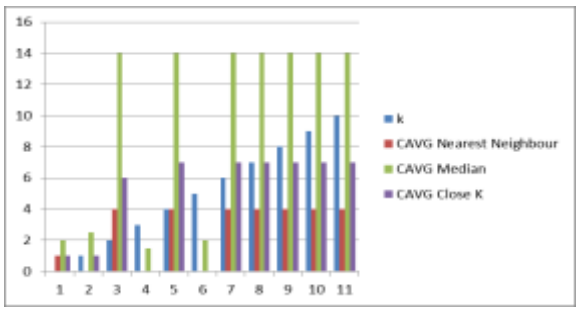

Fig 5.6

Similarly ,In Fig.5.7 and fig.5.8 value of CDM is high during increase of K and later we get constant value soNearest Neighbour is better method than Median and close-K

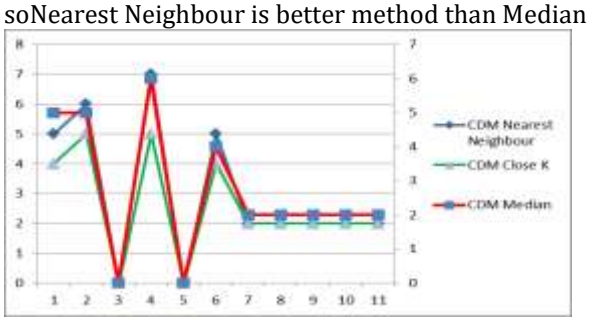

Fig 5.7 CDM with Increase in K

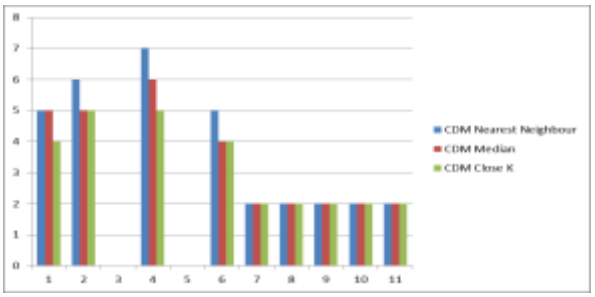

Fig 5.8

VI. CONCLUSION

We have discussed about the basic concepts privacy preservation during the process of data mining we have explained in detail about single and Multidimensional K-Anonymity. And how to achieve logical data security using multidimensional K-Anonymity. when we have a database which we cannot form dividable multidimensional. Also different features of establishing logical data security. We have also discussed about how these features have been implemented in sample non-relational databases. We have also given our proposed solution for achieving multidimensional k-Anonymity. We conclude that for released data confidentiality and Logical data Security Multidimensional K-Anonymity Model with Nearest Neighbor method will be the effective solution. The future work can be we can establish a web/cloud based service where any database without anonymity is accepted and provide anonymized database with logical data security.

REFERENCES

[1]. Preserving Privacy during Big Data Publishing using K-Anonymity Model – A SurveyDivya SadhwaniDr. Sanjay SilakariMr. Uday ChourasiaInternational Journal of Advanced Research in Computer ScienceVolume 8, No. 5, May June 2017ISSN No. 0976-5697 pp 801-810

[2]. Liu, Kai-Cheng & Kuo, Chuan-Wei &Liao, Wen-Chiuan & Wang, Pang-Chieh. (2018). "Optimized de- Identification of data UsingMultidimensional k-Anonymity ".1610-1614. 10.1109/TrustCom/BigDataSE.2018.00235

[3]. Pingshui Wang, Jiandong Wang1, Xinfeng Zhu1, Jian Jiang, "Research on Privacy Preserving Data Mining", International Conference on Biological and Biomedical Sciences *Advances in Biomedical Engineering, Vol.9*, pp.251-257, 2012. haru C. Aggarwal, "*A General survey of Privacy-Preserving Data Mining Models And Algorithms*", IEEE, pp 11-52,2008.

[4]. Sweeney L., "*K-anonymity: A Model for protecting privacy*", International Journal of Uncertainty, Fuzziness and Knowledge based system, 10(5), pp.557-570, 2002.

[5]. Slava kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira,"Efficient Multi-dimensional Suppression for k-anonymity", IEEE transaction, pp1-14,2009.

[6]. Kshitij Pathak, Nidhi Maheshwarkar,"*Performance issues of various K-anonymity Strategies*", IJCTEE ,ISSN:2231-2307, Vol.1,Issue 2, pp18-22,2011.

[7]. LeFevre K, DeWitt D J, Ramakrishnan R,"Mondrian multidimensional K- anonymity", IEEE International Conference on Data Engineering(ICDE06), Atlanta, GA, USA, pp1- 11,April 2006.

[8]. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, v.11 n.1, pp10- 18june 2009 [doi:10.1145/1656274.1656278].

[9]. Qian Wang, Cong Xu, Min Sun, " Multi-dimensional K-anonymity based on Mapping for Protecting Privacy", Journal of Software, Vol. 6, No. 10, pp1937-1947,October 2011 .

[10]. Qian Wang, Cong Xu, Min Sun, " Protecting Privacy by Multi-dimensional K- anonymity", Journal of Software, Vol. 7, No. 8, August 2012,pp1873-1880.

[11]. Yongbin Yuan, Jing Yang, Sheng Lan " Approach of P-sensitive k-anonymity Based on Nearest Neighborhood Search in Privacy Preserving", Journal of Information and Computational Science 9: 5(May 2012) ,pp1385-1393.

[12]. Gionis A, Tassa T., "k-Anonymization with Minimal Loss of Information", Knowledge and Data Engineering, IEEE Transactions, pp. 206-219, 2009.

[13]. Madhan Subramaniam, Senthil R., "An Analysis on Preservation of Privacy in DataMining", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010,pp1696-1699.

[14]. Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N., "A Comparison Study between Data Mining Tools over some Classification Methods", International Journal of Advanced Computer Science and Applications, December 2011.

[15]. C. Blake, C. Merz., "UCI repository of machine learning databases", 1998. http://www.ics.uci.edu/mlearn/M1Repository.html.

[16]. Samarati, Latanya Sweeney, "Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression", IEEE Transactions on Knowledge and Data Engineering, 2001

[17]. Accuracy-Constrained Privacy- Preserving Access Control Mechanism for Relational Data Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE TRANSACTIONS ON KNOWLEDGE AND DATAENGINEERING,VOL.26,NO. 4, APRIL 2014.

[19]. Privacy Preserving Data Publishing Basedon *k*-Anonymity by Categorization of Sensitive Values, Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini ChaudharyInternational Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014ISSN pp 2229-5518

[20]. Survey on Anonymization using k-anonymity for Privacy Preserving in Data MiningBinal Upadhyay*, Dr. amit ganatra2 International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March-2017 ISSN 2229-5518 pp 376-380

[21] Clustering Based K-anonymity Algorithm for Privacy PreservationSang Ni1, Mengbo Xie1, Quan Qian1;2International Journal of Network Security, Vol.19, No.6, PP.1062-1071, Nov. 2017 (DOI: 10.6633/IJNS.201711.19(6).23) pp 1062-1071.