ISSN No:-2456-2165

Use of Comprehensive Technique for Preserving Privacy in Data Mining

[1]Namrata Govind Ambekar, [2]Rajyalakshmi Jaiswal
[1]PG student Master of Engineering, [2] Prof., Assistant Professor Computer Engineering
Department LDCE, Ahmedabad, Gujarat

Abstract:- The world has entered into the digital age of information. Immersion in the field of information and technology comforts humanity but individual's privacy and security is deteriorates. The private data provided by the individual and various organizations at the time of using mobile phone internet for different purposes, which may contain individuals sensitive information cannot be disclosed to the anonymous person without applying the privacy-preserving technique on it. Nowadays, Preserving Privacy Data Mining (PPDM) has been studied rigorously because of the wide penetration of sensitive information on the internet. Many techniques have been proposed so far like K-anonymization, l-diversity, Randomization, Perturbation methods, and Cryptographic techniques designed for Preserving Privacy Data Mining (PPDM). There are some plus and minus points of every approach. The negative point constitutes a loss of data, reduction in the utility of data, lack of diversity of data, security issues likewise. In this research work, we are going to propose a "Comprehensive Technique" which works amongst existing algorithm by analyzing some work done in this field. We proposed a novel technique named "Clustering Based Anonymization by Assigning Weight to Each attribute", this k-means clustering algorithm is used with some of the alterations for anonymization of data. We are assigning feature weight manually so that distortion of data can be reduced. The main goal of the proposed model is to preserve privacy at the same time with minimum information loss.

I. INTRODUCTION

Recent hike in the information science have made easy for the collection and analysis of the data for collaboration purposes. Data can be collected from various organizations like credit card companies, online shopping habits, and hospital organization etc. collects individual sensitive information. Despite of advantages, there is also threat to the privacy and security to the individuals' private data. We would like to preserve the privacy of individual sensitive information during the process of data publishing. Various data owners such as banks, hospitals, insurance companies, credit card companies, educational institutes they itself anonymized their data values before releasing data to required clients or organization for analysis and other data mining activities. Data owner want a way to transfer data consisting sensitive information in secured way to protect personal sensitive data, therefore different techniques of privacy preservation have been developed such as Anonymization, Perturbation, Randomization, Distributed privacy preservation method, Cryptography methods etc. Nowadays Anonymization technique for privacy-preserving has drawn a lot of researcher's attention. In this paper, concern is to provide comprehensive privacy with the help of cluster based anonymization techniques by assigning weight to the attributes.

II. CLUSTERING IN ANONYMIZATION

We want to preserve the privacy of data during publishing. We came across the situation when data release to the other multinational companies for collaborative analysis purposes, and that table consist of individual sensitive information. Hence, we propose a new method for anonymizing data values is clustering based anonymization. Clustering is the process of organizing objects into groups of object that belongs to the same class. Where quasi-identifiers attributes values are firstly grouped into equivalent clustered and then cluster center are published. To ensure privacy of the data records, we impose the constraint that each cluster must contain no less than a pre-specified number of data records. This technique is more general since we have a much larger choice for cluster centers than k-Anonymity. Most cases, lot of information released without compromising privacy. Without disclosing single amount of the database records, we can assure the data release for analysis has minimum distortion and hence is more useful. Inherently clustering is the optimize solution for k- anonymization problem.

III. PROPOSED ALGORITHM

K- Anonymization approach have some lacuna such as l-diversity, t-closeness. To overcome these limitations we have to apply some modifications on k-anonymity which affects some factors like reduction in information, loss of data diversity, data mining efficiency downgraded likewise so we come up with a new approach which can preserve our data without any limitations. K-anonymity, other alternatives methods have approximation algorithms (NP-complete). Our algorithm is polynomial time algorithm. In our algorithm we follow k-mean clustering algorithm with some alteration for anonymizing data. In proposed clustering method, features weights are assigned manually, the data distortion can reduce. Clustering is the process of grouping a set of tuples having similar properties in a cluster and set of tuples having dissimilar properties in another cluster. This is the reason why the k-anonymity model can be addressed from the viewpoint of clustering. Main objective of the proposed approach is trying to preserve privacy and at the same time securing data.

- **Data utility -** Adversaries can learn from the published data. And main goal is to the preserve privacy and increase utility of the data by maintaining accuracy of data mining task of released records.
- **Privacy** – By clustering based anonymization we can provide privacy such a way that sensitive data disclosures sensitive data cannot be possible.
- **Information loss -** By assigning feature weight manually to each attribute hence loss of information and data distortion is also reduced.

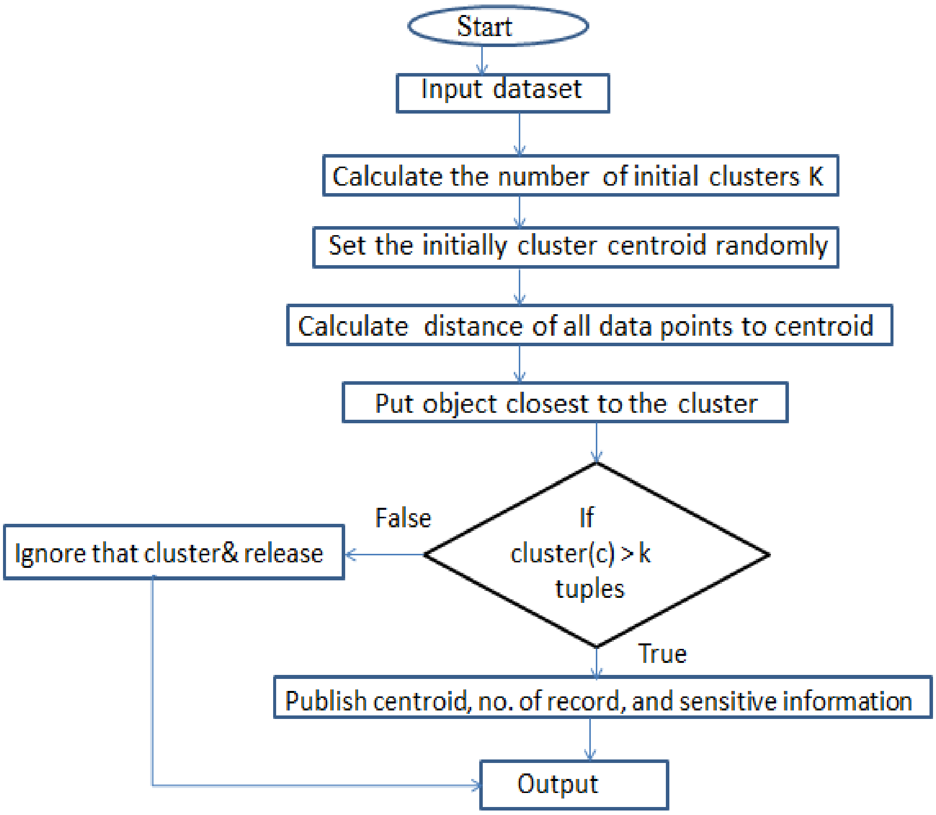

Start
Input dataset
Calculate the number of initial clusters K
Set the initially cluster centroid randomly
Calculate distance of all data points to centroid
Put object closest to the cluster
If cluster(c) > k tuples
False
Ignore that cluster& release
True
Publish centroid, no. of record, and sensitive information
Output

Figure 1: Workflow of Proposed Algorithm

Input: Database D contains M numbers of records in which each record has N quasi-identifier attributes and value of k is the anonymity level.

Output: Anonymized table

Algorithm:

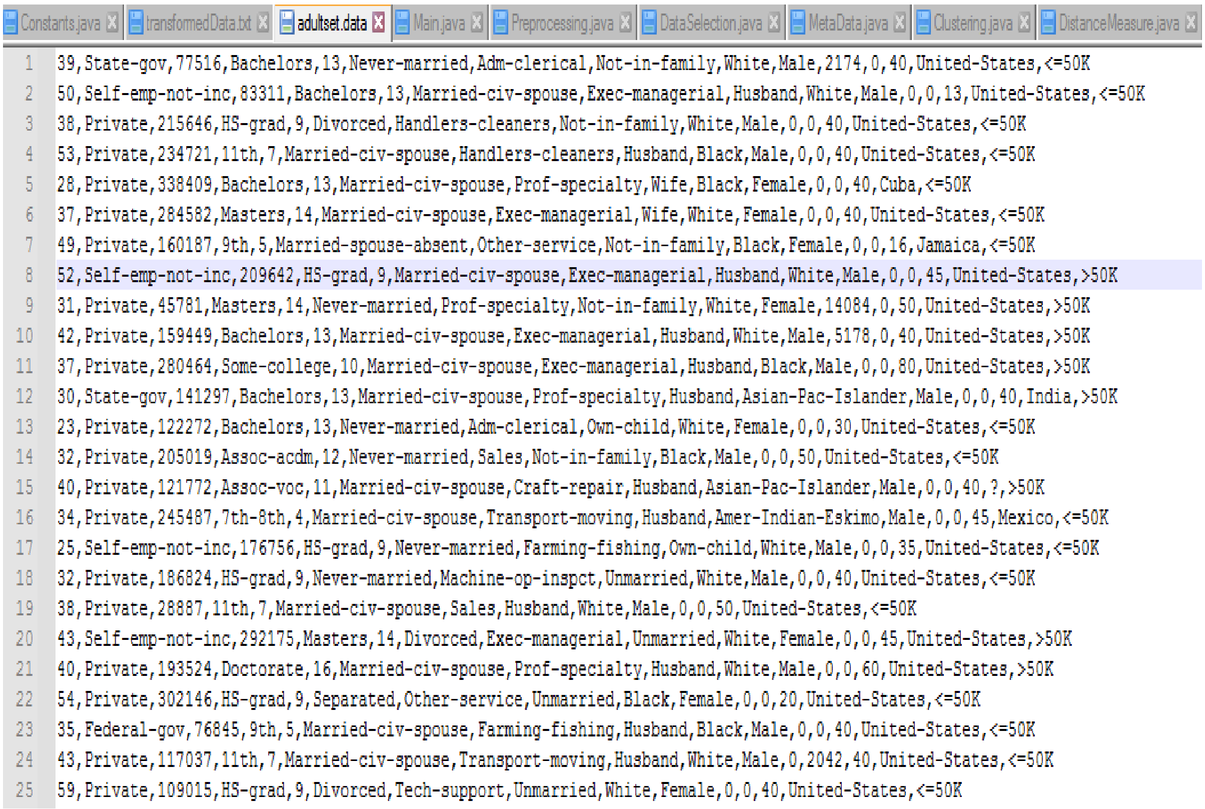
```
1. Assign waits for every attribute (sum of weights should be one).
2. Calculate number of initial clusters (#clusters(c) = M / k).
3. Assign initial random centroids for every cluster.
4. for every tuple
- calculate distance (di) between every cluster centroid.
- assign that tuple to minimum distance cluster.
- update centroid (by taking average).
5. for every cluster
- If cluster contains more than k tuples.
- Publish that centroid, number of records and sensitive information.
- Otherwise ignore that cluster
```

Complexity: M×c

IV. EXPERIMENTAL SETUP

This experiment uses ADULT data from the UCI Machine Learning Repository for testing, which consider as standard for k- anonymization. The ADULT dataset holds 32561 records and 15 attributes. Out of them, we recollect only attributes Age, Gender, fnlwt, Occupation, Marital-status, and Race. The attributes Age and fnlwt are numeric attributes, and Race, Gender, Marital-status and Occupation are the categorical attributes. The attribute Occupation is reserved as a sensitive attribute in the dataset. The research will be executed in Java with JDK 1.7 in a system constructed with Intel core i3 processor, 500GB hard disk,4 GB RAM, and Notepad++ for writing program.

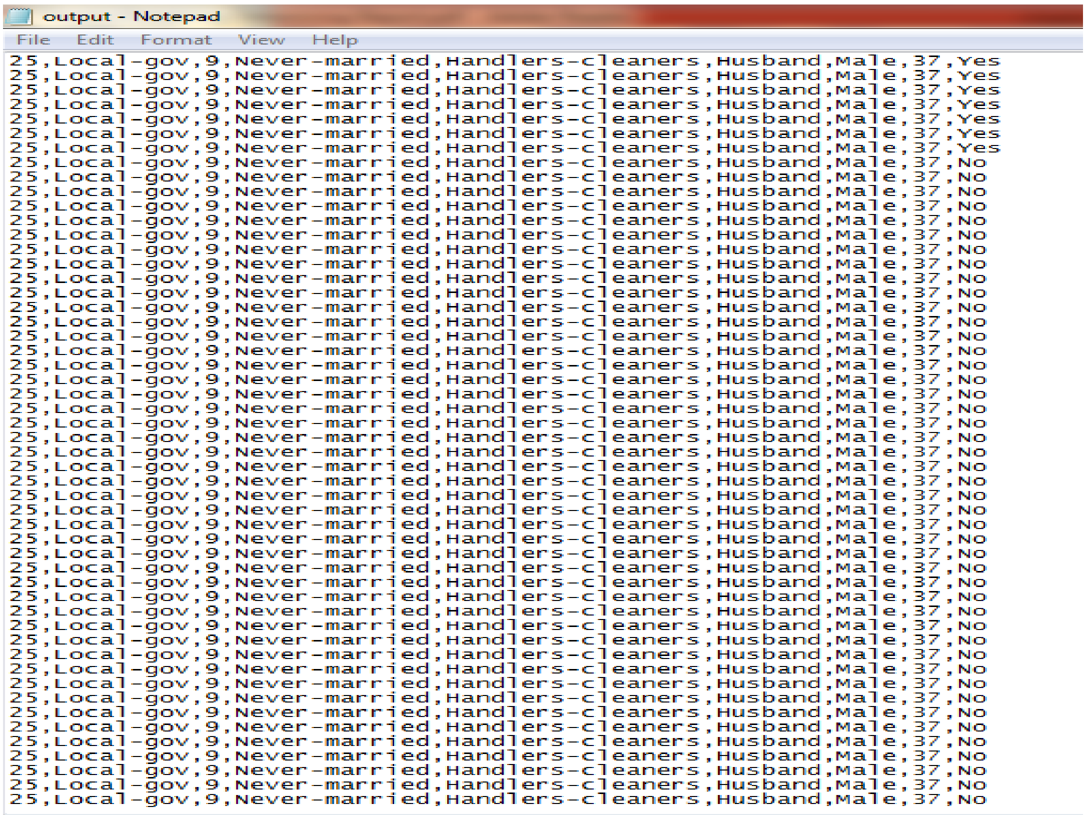
INPUT: Adult Dataset



```
39,State-gov,77516,Bachelors,13,Never-married,Adm-clerical,Not-in-family,White,Male,2174,0,40,United-States,<=50K
50,Self-emp-not-inc,83311,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,13,United-States,<=50K
38,Private,215646,HS-grad,9,Divorced,Handlers-cleaners,Not-in-family,White,Male,0,0,40,United-States,<=50K
53,Private,234721,11th,7,Married-civ-spouse,Handlers-cleaners,Husband,Black,Male,0,0,40,United-States,<=50K
28,Private,338409,Bachelors,13,Married-civ-spouse,Prof-specialty,Wife,Black,Female,0,0,40,Cuba,<=50K
37,Private,284582,Masters,14,Married-civ-spouse,Exec-managerial,Wife,White,Female,0,0,40,United-States,<=50K
49,Private,160187,9th,5,Married-spouse-absent,Other-service,Not-in-family,Black,Female,0,0,16,Jamaica,<=50K
52,Self-emp-not-inc,209642,HS-grad,9,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,45,United-States,>50K
31,Private,45781,Masters,14,Never-married,Prof-specialty,Not-in-family,White,Female,14084,0,50,United-States,>50K
42,Private,159449,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,5178,0,40,United-States,>50K
37,Private,280464,Some-college,10,Married-civ-spouse,Exec-managerial,Husband,Black,Male,0,0,80,United-States,>50K
30,State-gov,141297,Bachelors,13,Married-civ-spouse,Prof-specialty,Husband,Asian-Pac-Islander,Male,0,0,40,India,>50K
23,Private,122272,Bachelors,13,Never-married,Adm-clerical,Own-child,White,Female,0,0,30,United-States,<=50K
32,Private,205019,Assoc-acdm,12,Never-married,Sales,Not-in-family,Black,Male,0,0,50,United-States,<=50K
40,Private,121772,Assoc-voc,11,Married-civ-spouse,Craft-repair,Husband,Asian-Pac-Islander,Male,0,0,40,?,>50K
34,Private,245487,7th-8th,4,Married-civ-spouse,Transport-moving,Husband,Amer-Indian-Eskimo,Male,0,0,45,Mexico,<=50K
25,Self-emp-not-inc,176756,HS-grad,9,Never-married,Farming-fishing,Own-child,White,Male,0,0,35,United-States,<=50K
32,Private,186824,HS-grad,9,Never-married,Machine-op-inspct,Unmarried,White,Male,0,0,40,United-States,<=50K
38,Private,28887,11th,7,Married-civ-spouse,Sales,Husband,White,Male,0,0,50,United-States,<=50K
43,Self-emp-not-inc,292175,Masters,14,Divorced,Exec-managerial,Unmarried,White,Female,0,0,45,United-States,>50K
40,Private,193524,Doctorate,16,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,60,United-States,>50K
54,Private,302146,HS-grad,9,Separated,Other-service,Unmarried,Black,Female,0,0,20,United-States,<=50K
35,Federal-gov,76845,9th,5,Married-civ-spouse,Farming-fishing,Husband,Black,Male,0,0,40,United-States,<=50K
43,Private,117037,11th,7,Married-civ-spouse,Transport-moving,Husband,White,Male,0,2042,40,United-States,<=50K
59,Private,109015,HS-grad,9,Divorced,Tech-support,Unmarried,White,Female,0,0,40,United-States,<=50K
```

OUTPUT:

- **Anonymity level:** 3
- **Relevant Attributes:** 9/15
- **Output data:** contains 90% of data



```
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,Yes
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
25,Local-gov,9,Never-married,Handlers-cleaners,Husband,Male,37,No
```

V. EVALUATION PARAMETERS

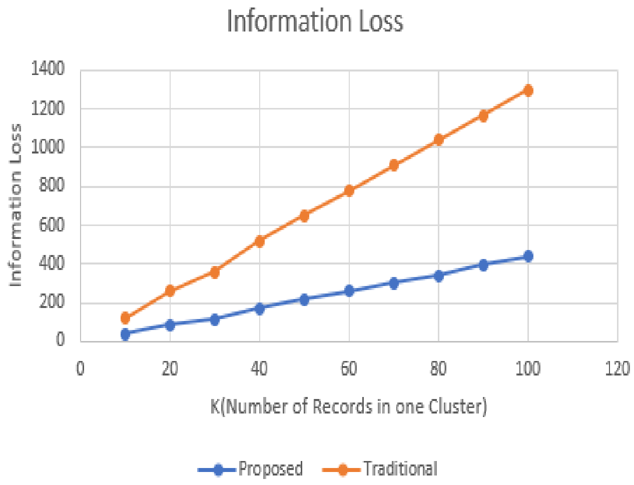
Comparative parameters of proposed work are as follows:

Proposed Algorithm contributes to increase following parameters:

- Information loss
- Data Diversity reduction
- Utility of data
- Privacy and Security

Evaluation:

- **Information Loss:**



VI. CONCLUSION

In recent day's releasing data about individuals without disclosing personal information is a crucial problem. Most of the privacy-preserving techniques consist of some types of data alteration by reducing the uniqueness by publishing the data. This results in a loss of effectiveness of mining algorithms. The most common data transforming techniques are randomization perturbation, generalization, suppression, anatomization, Permutation, Swapping, anonymity model like L-diversity, t- closeness, distributed privacy preservation, etc. A big challenge that must be considering in any technique of the privacy-preserving should focus on the efficiency of the data mining algorithm to extract the relevant information from the dataset even after anonymization. Thus the goal of this research is secure data through different optimization techniques. In this research work, we have implemented the clustering based anonymization by assigning weight to each attributes. This technique is free from almost all attack and its output does not depend on input size of the database.

REFERENCES

[1]. Pierangela Samarati, Latanya Sweeney : “k-anonymity and its enforcement through generalization and suppression”.

[2]. Qian Wang, Zhiwei Xu and Shengzhi Qu : “An enhanced K-anonumity model against homogeneity attack”,2011.

[3]. Dan Zhu, Xiao-Bai Li, Shuning Wu : “Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining”. Decision Support Systems. December 2009, Volume 48, Pages 133-140.

[4]. Heiko Paulheim : “Exploiting linked open data as background knowledge in data mining”.

[5]. Debasis Mohapatra1, Dr. Manas Ranjan Patra : “Analysis of k-Anonymity for Homogeneity Attack”, International Journal of Advances in Computer Science and Technology, 2014.

[6]. Tiancheng Li, Ninghui Li : “Mining background knowledge for data anonymization”.

[7]. Ashwin Machanacajjhala, Johnes Gehrke : “Daniel Kifer privacy beyond k-anonymity”.

[8]. Md. Zahidul Islam and Ljiljana Brankovic : “A Framework for Privacy Preserving Classification in Data Mining”, 2004, Australian Computer Society.

[9]. Arik Friedman and Assaf Schuster : “Data Mining with Differential Privacy”.

[10]. Jim Dowd, Shouhuai Xu, and Weining Zhang : “Privacy-Preserving Decision Tree Mining Based on Random Substitutions”.

[11]. Shucheng Yu, Cong Wang, Kui Ren, “Attribute Based Data Sharing with Attribute Revocation”,ASIACCS’10 April 13–16, 2010, Beijing, China.

[12]. Batya Kenig, Tamir Tassa, “A Practical Approximation Algorithm for Optimal k-Anonymity”.

[13]. Alexandre Evfimievski, “Randomization in Privacy Preserving Data Mining”

[14]. Stanley Robson de Medeiros Oliveira, “Data Transformation For Privacy-Preserving Data Mining”, PhD thesis, University of Alberta, 2005.

[15]. Nan Zhang, "Privacy preserving data mining", Phd Thesis, Texas A&M University

[16]. archive.ics.uci.edu/ml/datasets/Adult