ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

# Survey on Anonymization Technique for Privacy Preserving Data Mining (PPDM)

Kiran Israni[1], Shalu Chopra[2]

ME Student, Department of I.T, V.E.S Institute of Technology, Mumbai, India [1]

Associate Professor, Department of I.T, V.E.S Institute of Technology, Mumbai, India [2]

**ABSTRACT**: Privacy preserving in data mining technique modifies the data in such a way that individual's sensitive information will be hidden and at the same time usability of information is not lost for mining purpose. Now a day's individuals are looking forward to their data privacy, major concern about data is non sensitive information may reveal their sensitive data. This paper provides survey on Anonymization technique of PPDM such as k anonymity using generalization and suppression, P sensitive k anonymity, (α, k) anonymity, l-diversity, m-invariance.

**KEYWORDS:** Privacy; privacy preserving technique; Anonymization; l - diversity; m-invariance.

## I. INTRODUCTION

Data mining is process of discovering interesting knowledge from large amount of data stored, either in data bases, data warehouses or other information repositories [1]. Privacy in data mining is information about individual should not be revealed after data mining operation. Now a day's individuals are looking forward to their data privacy, major concern is that non sensitive information may reveal their sensitive data.

In Privacy Preserving in Data Mining (PPDM) data is de identified before releasing it for data mining process to preserve privacy [2]. PPDM deals with tradeoff between utility of information and preserving privacy of information. Different PPDM techniques are anonymization, Randomization, perturbation, condensation and cryptography; they are briefed in section II and section III provides survey on anonymization technique of PPDM.

## II. RELATED WORK

PPDM techniques can be classified in to five categories: Anonymization, perturbation, randomization, condensation, cryptography [2].

In Anonymization identifier attributes are removed and quasi identifier values are changed with less specific value, which makes tuples appear similar, and it will be difficult to identify the sensitive attribute value. Anonymization is discussed in next section.
In Perturbation original attribute values are altered with some synthetic values and statistical results obtained from altered data does not differ much as compared to statistical results obtained from original data. It is done by adding noise, swapping data, multiplicative perturbation, rotation, projection perturbation [3].

In Randomization data is altered by using random noise which is generated by using probability distribution.
Condensation approach uses pseudo data rather than modified data.

Cryptography approach is used where data is distributed among multiple parties and they do not want to share the information for computing result [3].

## III. ANONYMIZATION

Anonymization technique is further classified into five types *k* anonymity using generalization and suppression, *p* sensitive *k* anonymity, (α, *k)* anonymity, *l* diversity, m-invariance.
Consider a private table consisting of few tuples. Each tuple consists of the following 4 types of attributes:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

• Identifier (ID): This can identify the individual directly such as name, contact number etc.

• Quasi-identifier (QID): This can be linked with publically available records to identify the individual such as age, gender, and pin code.

• Sensitive Attribute (SA): which person wants to hide from others such as salary, disease.

• Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA.

Before publishing data for data mining, the data is anonymized, identifiers are removed and quasi-identifiers are modified. So that from available data individual's identity and sensitive attribute values cannot be easily identified by others [4].

### *A. K ANONYMITY USING GENERALIZATION AND SUPPRESSION*

Anonymization is changing the value of quasi-identifier with less specific value that can be done by generalization or suppression.
Generalization is replacing attribute value with less specific value e.g. If age of person is 35 that can be generalized to <40. Suppression is replacing attribute value with special value e.g. Pin code of person is 400001 that can be suppressed as 40****. *K* anonymity using generalization and suppression is minimum *k* number of records will appear similar so that it will be difficult to identify the person [4].

Table I 2-anonymity [4], where quasi identifier are age; sex; zipcode

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 5 | Female | 12000 | HIV |
| 9 | Male | 14000 | Dyspepsia |
| 6 | Male | 18000 | Dyspepsia |
| 8 | Male | 19000 | Bronchitis |
| 12 | Female | 21000 | HIV |
| 15 | Female | 22000 | Cancer |
| 17 | Female | 26000 | Pneumonia |
| 19 | Male | 27000 | Gastritis |
| 21 | Female | 33000 | Flu |
| 24 | Female | 37000 | Pneumonia |

(a) Original Table

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [1,10] | People | 1**** | HIV |
| [1,10] | People | 1**** | Dyspepsia |
| [1,10] | People | 1**** | Dyspepsia |
| [1,10] | People | 1**** | Bronchitis |
| [11,20] | People | 2**** | HIV |
| [11,20] | People | 2**** | Cancer |
| [11,20] | People | 2**** | Pneumonia |
| [11,20] | People | 2**** | Gastritis |
| [21,60] | People | 3**** | Flu |
| [21,60] | People | 3**** | Pneumonia |

(b) 2-anonymous table

The table I (b) is 2 anonymize that is minimum 2 records will appear similar so it will be difficult for attacker to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

identify the person.

### *B. P SENSITIVE K ANONYMITY*

K anonymity protects against identity disclosure, but it fails to protect against attribute disclosure.

T. Truta [6] presented *p* sensitive *k* anonymity, group of tuples which satisfies *k* anonymity in that group number of distinct sensitive attributes will be at least p times.

In table I (b) assume last 2 tuples have same disease then if the attacker knows that age of person for which he is searching is between 21 to 60 and zip code is 3***** then attacker will easily know persons disease.

Consider the data record from Table II that satisfies 3-anonymity property with respect to Age, Zip Code and Sex. To find the value of *p*, T. Truta [6] have analyze each group with identical values for all key attributes. The first group (the first three tuples) has two different illnesses, and only one income, therefore the value of *p* is 1. This data satisfies 1-sensitive 3-anonymity property. when $p = 1$ that means all the group of records will have same value for sensitive attribute If the first record would have a different value for income (such as 40,000) then both groups would have two different illnesses and two different incomes, and the value of p would be 2.

Table II *p* sensitive *k*-anonymity [6]

| Age | Zipcode | Sex | Illness | Income |
|---|---|---|---|---|
| 20 | 43102 | F | AIDS | 50,000 |
| 20 | 43102 | F | AIDS | 50,000 |
| 20 | 43102 | F | Diabetes | 50,000 |
| 30 | 43102 | M | Diabetes | 30,000 |
| 30 | 43102 | M | Diabetes | 40,000 |
| 30 | 43102 | M | Heart Disease | 30,000 |
| 30 | 43102 | M | Heart Disease | 40,000 |

From the definition of *p*-sensitive *k*-anonymity property it is observed that *p* is always less than or equal to *k*.

From the above examples, the following two conclusions are drawn:

- To avoid the possibility of identity disclosure, a given data records must satisfy *k*-anonymity with *k* greater than or equal to 2.

- To avoid the possibility of attribute disclosure, a given data records must have *p*-sensitive *k* anonymity with *p* greater than or equal to 2.

### *C. (α, k ) ANONYMITY*

Suppose if quasi identifier for some tuples will have unique value and attacker has background knowledge about person whose quasi identifier value is unique then he can easily get information about that person.

Table III Raw medical data set [7]

| Job | Birth | Postcode | Illness |
|---|---|---|---|
| Clerk | 1975 | 4350 | HIV |
| Manger | 1955 | 4350 | Flu |
| Clerk | 1955 | 5432 | Flu |
| Factory worker | 1955 | 5432 | Fever |
| Factory worker | 1975 | 4350 | Flu |
| Technical supporter | 1940 | 4350 | Fever |

Table IV (0.5, 2) - anonymous table of Table III by full domain generalization [7]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

| Job | Birth | Postcode | Illness |
|---|---|---|---|
| * | * | 4350 | HIV |
| * | * | 4350 | Flu |
| * | * | 5432 | Flu |
| * | * | 5432 | Fever |
| * | * | 4350 | Flu |
| * | * | 4350 | Fever |

Table V (0.5, 2) - anonymous data set of Table III by local recoding [7]

| Job | Birth | Postcode | Illness |
|---|---|---|---|
| White – collar | * | 4350 | HIV |
| White – collar | * | 4350 | Flu |
| * | 1955 | 5432 | Flu |
| * | 1955 | 5432 | Fever |
| Blue – collar | * | 4350 | Flu |
| Blue – collar | * | 4350 | Fever |

In R. Wong [7], they defined a term equivalence class, that is, if Q be the quasi-identifier (QID). An equivalence class set, is QID-EC, for the same QID value of a table with respect to Q is a collection of all tuples in the table containing identical values of Q.

Table IV contains two QID-EC's. The first two and the last two tuples form a one QID-EC, because these tuples contain identical values of Q. Similarly, the third and the fourth record form the second QID-EC.

To achieve *k* Anonymization job attribute is generalized to lower level in Table III. *k* value is 2 because third and fourth tuple is having same postcode.

R. Wong [7], have proposed (α, *k)*-anonymity model, in which each EC of anonymized set should not have value of sensitive attribute more than α frequency.

They have shown two generalizations schemes: global recoding and local recoding. In recording all values of an attribute are generalized. E.g. In table III all values in attribute job are generalized i.e. clerk, manager, factory worker are generalized by using *.

Drawback of global generalization is it may lose more information as compared with local recoding because it suffers from over-generalization.

In **local recoding**, values may be generalized to different levels in the domain.

Table V is (0.5, 2) - anonymous table by local recoding, i.e. for one QID-EC job attribute is generalized and for other QIID-EC birth attribute is generalized.

***D. L DIVERSITY***

Machanavajjhala [8], proposed l-diversity technique which combines *k* Anonymization and diversify the sensitive attribute value in Equivalence class.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

Table VI Inpatient data [8]

| | Non – Sensitive | | | Sensitive |
|---|---|---|---|---|
| | ***Zip Code*** | ***Age*** | ***Nationality*** | ***Condition*** |
| 1 | 13053 | 28 | Russian | Heart disease |
| 2 | 13068 | 29 | American | Heart disease |
| 3 | 13068 | 21 | Japanese | Viral infection |
| 4 | 13053 | 23 | American | Viral infection |
| 5 | 14853 | 50 | Indian | Cancer |
| 6 | 14853 | 55 | Russian | Heart disease |
| 7 | 14850 | 47 | American | Viral infection |
| 8 | 14850 | 49 | American | Viral infection |
| 9 | 13053 | 31 | American | Cancer |
| 10 | 13053 | 37 | Indian | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer |
| 12 | 13068 | 35 | American | Cancer |

Table VII 4-anonymous inpatient data [8]

| | Non – Sensitive | | | Sensitive |
|---|---|---|---|---|
| | **Zipcode** | **Age** | **Nationality** | **Condition** |
| 1 | 130** | <30 | * | Heart Disease |
| 2 | 130** | <30 | * | Heart Disease |
| 3 | 130** | <30 | * | Viral Infection |
| 4 | 130** | <30 | * | Viral Infection |
| 5 | 1485* | ≥40 | * | Cancer |
| 6 | 1485* | ≥40 | * | Heart Disease |
| 7 | 1485* | ≥40 | * | Viral Infection |
| 8 | 1485* | ≥40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

Consider the records in table VI, table VII is 4 anonymous, if attacker know that Bob age is 31 year old American from zip code 13053 then attacker can easily observe that Bob is having cancer. But if records are diversified as in table VIII it will be difficult to identify sensitive attribute value.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Table VIII 3-diverse inpatient data [8]

| | Non – sensitive | | | Sensitive |
|---|---|---|---|---|
| | *Zipcode* | *Age* | *Nationality* | *Condition* |
| 1 | 1305* | ≤ 40 | * | Heart Disease |
| 2 | 1305* | ≤ 40 | * | Viral Infection |
| 3 | 1305* | ≤ 40 | * | Cancer |
| 4 | 1305* | ≤ 40 | * | Cancer |
| 5 | 1485* | > 40 | * | Cancer |
| 6 | 1485* | > 40 | * | Heart Disease |
| 7 | 1485* | > 40 | * | Viral Infection |
| 8 | 1485* | > 40 | * | Viral Infection |
| 9 | 1306* | ≤ 40 | * | Heart Disease |
| 10 | 1306* | ≤ 40 | * | Viral Infection |
| 11 | 1306* | ≤ 40 | * | Cancer |
| 12 | 1306* | ≤ 40 | * | Cancer |

### *E. m-INVARIANCE*

X. K. Xiao [9], proposed generalization principle called m-invariance. When records are added to or deleted from the original data then how to protect privacy of individual. Previous methods discussed are for one time publication of data. Those methods do not support re publication of data after new tuples are inserted or deleted from records.

Consider a hospital releases patients' records quarterly, but each publication includes only the results of diagnoses in the 6 months preceding the publication time. Table IX
(a) shows the data for the first release, at which time the hospital publishes the generalized relation in Table IX (b). The data at the second release is presented in Table X (a).

Table IX Data and its generalization at first release [9]

| Name | Age | Zip | Disease |
|---|---|---|---|
| Bob | 21 | 12000 | Dyspepsia |
| Alice | 22 | 14000 | Bronchitis |
| Andy | 24 | 18000 | Flu |
| David | 23 | 25000 | Gastritis |
| Gary | 41 | 20000 | Flu |
| Helen | 36 | 27000 | Gastritis |
| Jane | 37 | 33000 | Dyspepsia |
| Ken | 40 | 35000 | Flu |
| Linda | 43 | 26000 | Gastritis |
| Paul | 52 | 33000 | Dyspepsia |
| Steve | 56 | 34000 | Gastritis |

(a)Data

| G.ID | Age | Zip | Disease |
|---|---|---|---|
| 1 | [21,22] | [12k,14k] | Dyspepsia |
| 1 | [21,22] | [12k,14k] | Bronchitis |
| 2 | [23,34] | [18k,25k] | Flu |
| 2 | [23,24] | [18k,25k] | Gastritis |
| 3 | [36,41] | [20k,27k] | Flu |
| 3 | [36,41] | [20k,27k] | Gastritis |
| 4 | [37,43] | [26k,35k] | Dyspepsia |
| 4 | [37,43] | [26k,35k] | Flu |
| 4 | [37,43] | [26k,35k] | Gastritis |
| 5 | [52,56] | [33k,34k] | Dyspepsia |
| 5 | [52,56] | [33k,34k] | Gastritis |

(b) Generalization

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

Table X Data and its generalization at the second release [9]

| Name | Age | Zip | Disease |
|---|---|---|---|
| Bob | 21 | 12000 | Dyspepsia |
| David | 23 | 25000 | Gastritis |
| Emptily | 25 | 21000 | Flu |
| Jane | 37 | 33000 | Dyspepsia |
| Linda | 43 | 26000 | Gastritis |
| Gary | 41 | 20000 | Flu |
| Mary | 46 | 30000 | Gastritis |
| Ray | 54 | 31000 | Dyspepsia |
| Steve | 56 | 34000 | Gastritis |
| Tom | 60 | 44000 | Gastritis |
| Vince | 65 | 36000 | Flu |

(a) Data

| G.ID | Age | Zip. | Disease |
|---|---|---|---|
| 1 | [21,23] | [12k,25k] | Dyspepsia |
| 1 | [21,23] | [12k,25k] | Gastritis |
| 2 | [25,43] | [21k,33k] | Flu |
| 2 | [25,43] | [21k,33k] | Dyspepsia |
| 2 | [25,43] | [21k,33k] | Gastritis |
| 3 | [41,46] | [20k,30k] | Flu |
| 3 | [41,46] | [20k,30k] | Gastritis |
| 4 | [54,56] | [31k,34k] | Dyspepsia |
| 4 | [54,56] | [31k,34k] | Gastritis |
| 5 | [60,65] | [36k,44k] | Gastritis |
| 5 | [60,65] | [36k,44k] | Flu |

(b) Generalization

The tuples of Alice, Andy, Helen, Ken, and Paul have been deleted (as they describe diagnoses over six months ago), while 5 new tuples have been inserted. Accordingly, the hospital publishes the generalized relation in Table X (b).

Both Tables IX (b) and X (b) are 2- anonymous and 2-diverse; an attacker can still identify the disease of a patient, by correlating between the two tables.

Assume, an attacker who has Bob's age and Zipcode, and knows that Bob has a record in both Tables IX (b) and X (b) i.e., Bob was admitted for treatment, within 6 months before both tables are published. From Table X (b), attacker finds out that Bob's disease must be either *dyspepsia* or *gastritis*. By combining information from both tables, the attacker correctly identifies Bob's real disease *dyspepsia*.

To overcome this issue they suggested deleted tuples should not be removed from published data, but the problem with this will be number of tuples in published data will go on increasing.

Table XI Remedying critical absence with counterfeits [9]

| Name | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| Bob | 1 | [21,22] | [12k,14k] | Dyspepsia |
| C1 | 1 | [21,22] | [12k,14k] | Bronchitis |
| David | 2 | [23,25] | [21k,25k] | Gastritis |
| Emptily | 2 | [23,25] | [21k,25k] | Flu |
| Jane | 3 | [37,43] | [26k,33k] | Dyspepsia |
| C2 | 3 | [37,43] | [26k,33k] | Flu |
| Linda | 3 | [37,43] | [26k,33k] | Gastritis |
| Gary | 4 | [41,46] | [20k,30k] | Flu |
| Mary | 4 | [41,46] | [20k,30k] | Gastritis |
| Ray | 5 | [54,56] | [31k,34k] | Dyspepsia |
| Steve | 5 | [54,56] | [31k,34k] | Gastritis |
| Tom | 6 | [60,65] | [36k,44k] | Gastritis |
| Vince | 6 | [60,65] | [36k,44k] | Flu |

(a) Data with counterfeits

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

| Group-ID | Count |
|---|---|
| 1 | 1 |
| 3 | 1 |

(b) Published counterfeit statistics

In [9] they have done integration of m-invariance and counterfeited generalization. In counterfeited generalization c1 and c2 generalized tuples added to republishing data for protecting privacy and m-invariance forces the group of tuples will have same values for sensitive attribute values in both the publications of data.
Table XI (a) involves a generalized tuples for every row in Table X (a), along with two counterfeit tuples c1 and c2; The 13 tuples are partitioned into six QI groups. Table XI (b) indicates that a counterfeit is placed in QI groups 1 and 3, respectively. From an attackers point, a counterfeit tuple is indistinguishable from the other tuples in the QI group (that contains the counterfeit).suppose if attacker has background knowledge about Bob. Then also the group has the same set of sensitive values {dyspepsia, bronchitis}. Therefore, the attacker will not get to know about the Bobs disease

## IV. CONCLUSION

There are various PPDM techniques such as anonymization, perturbation, randomization, condensation, cryptography. In this paper we have reviewed anonymization technique of PPDM such as *k* anonymity using generalization and suppression, *p* sensitive *k* anonymity, (α, *k)* anonymity, l diversity, m-invariance. Except m-invariance all other techniques are for static data, and m-invariance is used for dynamic data. We have provided all the techniques by considering data table and after applying technique how the data table will look like. We hope review provided in this paper will be helpful.

## REFERENCES

[1] J. Han and M. Kamber , "*Data Mining: Concepts and Techniques*", 2nd,the Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.
[2] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects," in *Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT)*, Nov. 2012, pp. 26_32.
[3] Charu C. Aggarwal, Philip S. Yu," A General Survey of Privacy-Preserving Data Mining Models and Algorithms", in springer ISBN 978-0-387-70991-8, e-ISBN 978-0-387-70992-5,DOI 10.1007/978-0-387-70992-5.
[4] LEI XU, CHUNXIAO JIANG, JIAN WANG,JIAN YUAN,AND YONG REN, "Information Security in Big Data: Privacy and Data Mining". Date of publication October 9, 2014, date of current version October 20, 2014. *Digital Object Identifier 10.1109/ACCESS.2014.2362522*.
[5] LATANYA SWEENEY," ACHIEVING *k*-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; pp 571- 588.
[6] T Truta, B. Vinay, "*Privacy Protection: p-Sensitive k-Anonymity Property"*, In Proceedings of the 22nd International Conference on Data Engineering Workshops, 94-103, 2006.
[7] R. Wong, Y. Liu, J. Yin, Z. Huang, Ada Wai-Chee Fu, and Jian Pei,"(α, *k)*-anonymity Based Privacy Preservation by Lossy join", APWeb/WAIM'07 Proceedings of the joint 9th Asia-Pacific web and 8th International Conference on Web-age Information Management Conference (2007) .
[8] A. MACHANAVAJJHALA, D. KIFER,J. GEHRKE,M. VENKITASUBRAMANIAM," ℓ-Diversity: Privacy Beyond k-Anonymity ", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1,Issue 1, March 2007 Pages 1–47.
[9] X. K. Xiao, Y.F. Tao, "M-invariance*: Towards Privacy Preserving Re-Publication of Dynamic Datasets"*, Proceedings of the ACM Conference on Management of Data (SIGMOD), 689-700, 2007.