

Privacy Preserving Updates for Anonymization Data Using ℓ -Diversity

Suyog Vilas Patil¹, Prof. K. B. Manwade²

¹ME (CSE) Student, Ashokrao Mane Group Of Institutions, Vathar, Kolhapur, Maharashtra, India

²Asst. Prof., Ashokrao Mane Group Of Institutions, Vathar, Kolhapur, Maharashtra, India

ABSTRACT

To prevent misuse of sensitive data by the unauthorized/application users and provide both privacy and security of the sensitive data. The privacy preservation mechanism is to protect the data from unauthorized user. Privacy Protection Mechanism (PPM) can satisfy privacy requirements such as k-anonymity and l-diversity. The privacy protection mechanism (PPM) is a general method used to transform the original data into some anonymous form to prevent from accessing owners sensitive information. PPM meets privacy requirement through k-anonymity it provides better privacy for the sensitive information which is to be shared. The privacy is achieved by the high accuracy of the user information. The ℓ -diversity method is an extension of the k-anonymity method, it is more efficient than the k-anonymity method. It avoids the attacks like background knowledge attack and others in k-anonymity method. In this paper we analyze ℓ -diversity method with different techniques.

Keywords: Privacy Protection Mechanism, k-anonymity, Access Control Mechanism, ℓ -diversity

I. INTRODUCTION

Data security is a very broad area that addresses many issues, like legal and ethical issues regarding the right to access certain information. The sensitive data is accessible to authorized users only. The database security is based on the Access Control Mechanism (ACM) and the Privacy Protection Mechanism (PPM). The privacy is achieved by the high accuracy of the user information. To protect data, the anonymization method is one of the best privacy protection mechanisms. The anonymization process will transform the sensitive information to some anonymized form using K-anonymity, ℓ -diversity. The PPM needs to satisfy an additional constraint namely the Imprecision bound for each selection predicate. The imprecision bound reduced the delaying for publishing data stream. The proposed system refers data anonymization using the ℓ -diversity. ℓ -diversity method reduces the granularity of representation of the data, ℓ -diversity

can still defend against background knowledge that is unknown to the data publisher. The ℓ -diversity method is an extension of the k-anonymity method. The k-anonymity is the anonymization techniques convert the sensitive information to some anonymized form using generalization and suppression. The proposed system uses ℓ -diversity method. ℓ -diversity is a form of group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation.

II. METHODS AND MATERIAL

1. Literature Review

Role based access control gives permission to the users to access data based on their role. For Relational data Nagabhushan, Arif Ghafoor, Zahid Pervaiz et al defines [2] selection predicates query technique is available to role while the privacy requirement is satisfy.

The stream data offers query processing over continues and sequencing data for data publishing and the windowing techniques generally emphasize on the streaming data. **T.Ghanem,A.Elmagarmid,P.Larsen and w.Aref et al.** proposed the predicate window query processing for streaming data [3].The access control uses Role based techniques to satisfy

To maintain the privacy of data it is need to minimize the imprecision of aggregate for all queries. The imprecision bound is a resulted value which determines the amount of imprecision that can be tolerated for each query. Privacy preserving mechanism needs to sum of false negative and false positive is less than imprecision bound. Zahid Pervaiz, Arif Ghafoor, Fellow, Walid G. Aref et.al proposed the Top Down Selection Mondrian (TDSM) used to minimize imprecision bound for rational data [2].

2. Attacks on k-Anonymity:

The homogeneity attack and the background knowledge attack, and we show how they can be used to compromise a k-anonymous dataset.

Homogeneity Attack:

K-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute. ataset. Homogeneity Attack: Alice and Bob are antagonistic neighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to discover what disease Bob is suffering from. Alice discovers the 4-anonymous table of current inpatient records published by the hospital (Figure 2), and so she knows that one of the records in this table contains Bob’s data. Since Alice is Bob’s neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob’s record number is 9, 10, 11, or 12. Now, all of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer. Note that such a situation is not uncommon. As a back of-the-envelope calculation, suppose we have a dataset containing 60,000 distinct tuples where the sensitive attribute can take 3 distinct values and is not correlated with the nonsensitive attributes. A 5-anonymization of this table will have around 12,000 groups² and, on average, 1 out of every 81 groups will have no diversity (the values for the sensitive attribute will all be the

same). Thus we should expect about 148 groups with no diversity. Therefore, information about 740 people would be compromised by a homogeneity attack. This suggests that in addition to k-anonymity, the sanitized table should also ensure “diversity” – all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes. Our next observation is that an adversary could use “background” knowledge to discover sensitive information.

Background Knowledge Attack:

K-Anonymity does not protect against attacks based on back- ground knowledge. Alice has a penfriend named Umeko who is admitted to the same hospital as Bob, and whose patient records also appear in the table shown in Figure 2. Alice knows that Umeko is a 21 yearold Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko’s information is contained in record number 1,2,3, or 4. Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is wellknown that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection. We have demonstrated (using the homogeneity and background knowledge attacks) that a k-anonymous table may disclose sensitive information. Since both of these attacks are plausible in real life, we need a stronger definition of privacy that takes into account diversity and background knowledge. This paper addresses this very issue.

3. Proposed Architecture

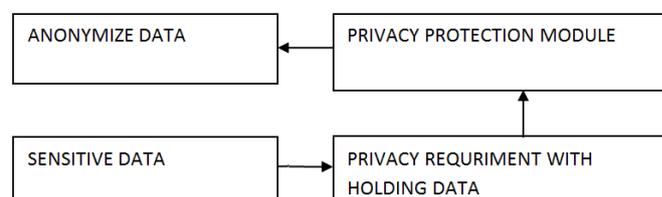


Figure 1. System architecture of Data privacy

The ℓ -Diversity Principle In this subsection we will derive the principle of ℓ -diversity in two ways. First, we will derive it in an ideal theoretical setting where it can be shown that the adversary’s background knowledge will not lead to a privacy breach. Then we will re-derive the ℓ -diversity principle from a more practical starting

point and show that even under less-than-ideal circumstances, ℓ -diversity can still defend against background knowledge that is unknown to the data publisher. Although the arguments in this subsection can be made precise, we will keep our discussion at an intuitive level for the sake of clarity.

The ℓ -diversity principle: Distinct ℓ -diversity and recursive ℓ -diversity.

1. Distinct ℓ -diversity: It captures the notion of well-represented groups due to the fact that entropy increases as frequencies become more uniform. We can also capture the role of background knowledge more explicitly with an alternate definition.

2. Recursive ℓ -diversity: For example, a clinic might be allowed to disclose that a patient has a “heart problem” because it is well known that most patients who visit the clinic have heart problems. It may also be allowed to disclose that “Medical Condition” = “Healthy” if this is not considered an invasion of privacy. At this point one may be tempted to remove tuples with non-sensitive “Medical Condition” values, publish them unaltered, and then create an ℓ -diverse version of the remaining dataset. In some cases this is acceptable. However, there are three important issues why the above suggestion may not be acceptable: the anonymity of the unaltered tuples, the privacy of the remaining tuples, and the utility of the resulting published data.

III. RESULTS AND DISCUSSION

Experimental Result:

Training Set

The training dataset of <http://api.census.gov//data.html> got from <https://www.census.gov.in/ftp/pub/DES/welcome.html>.

A census is the procedure of systematically acquiring and recording information about the members of a given population and business details of country.

The census data for business details of government is divided into rows and Colum. The Colum name is like:

Age	AAGE
Class of worker	ACLSWKR
Education	AHGA

In the Dataset each record maintains all Colum uses the space and commas to separate data. the data format are as follows:

For example,

```
record 1 --- 20, selfemployee, 10thgrade
record 2 --- 47, selfemployee, 12thgrade
```

Where,

20 is Age of the person Selfemployee is a class of worker, 12th grade is education of worker

Importing Training set



Training Dataset Collection

Training Dataset Window

id	AAGE	ACLSWKR	ADTIND	ADT	AHGA	AHRS	AHSCOL	AMARITL	AMJIND
1	73	Not in universe	0	0	High school gradu...	0	Not in universe	Widowed	Not in...
2	58	Self-employed-not incorporated	4	34	Some college but...	0	Not in universe	Divorced	Const...
3	18	Not in universe	0	0	10th grade	0	High school	Never ma...	Not in...
4	9	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
5	10	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
6	48	Private	40	10	Some college but...	1200	Not in universe	Married-cl...	Emple...
7	42	Private	34	3	Bachelors degree...	0	Not in universe	Married-cl...	Finan...
8	28	Private	4	40	High school gradu...	0	Not in universe	Never ma...	Const...
9	47	Local government	43	26	Some college but...	876	Not in universe	Married-cl...	Educa...
10	34	Private	4	37	Some college but...	0	Not in universe	Married-cl...	Const...
11	8	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
12	32	Not in universe	0	0	High school gradu...	0	Not in universe	Never ma...	Not in...
13	51	Private	4	34	Some college but...	0	Not in universe	Married-cl...	Const...
14	45	Private	37	31	High school gradu...	0	Not in universe	Divorced	Busin...
15	25	Private	24	12	Bachelors degree...	0	Not in universe	Never ma...	Manuf...
16	13	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
17	47	Private	39	36	Bachelors degree...	0	Not in universe	Never ma...	Perso...
18	39	Not in universe	0	0	10th grade	0	Not in universe	Married-cl...	Not in...
19	16	Not in universe	0	0	10th grade	0	High school	Never ma...	Not in...
20	35	Private	12	41	High school gradu...	0	Not in universe	Married-cl...	Manuf...
21	12	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
22	27	Self-employed-not incorporated	4	34	Some college but...	0	Not in universe	Married-cl...	Const...
23	55	Private	35	22	Some college but...	500	Not in universe	Married-cl...	Finan...
24	45	Private	45	12	Masters degree(M...	0	Not in universe	Married-cl...	Other...
25	55	Not in universe	0	0	Some college but...	0	Not in universe	Married-cl...	Not in...
26	2	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
27	1	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
28	37	Private	3	34	Some college but...	0	Not in universe	Married-cl...	Mining
29	4	Not in universe	0	0	Children	0	Not in universe	Never ma...	Not in...
30	37	Private	4	2	Bachelors degree...	0	Not in universe	Never ma...	Const...
31	63	Private	19	35	Less than 1st grade	0	Not in universe	Married-cl...	Manuf...

Training the data was imported form a file and arrange in specified rows and column.

Conversion of Dataset:

Importing Dataset

id	AAGE	ACLSWKR	ADTIND	ADT_	AHGA	AHRS	AHSCOL	AMARITL	AMJND
1	>=40	Not in universe	0	0	High school gradu	0	Not in universe	Widowed	Not in
2	>=40	Self-employed-not incorporated	4	34	Some college but	0	Not in universe	Divorced	Const.
3	<40	Not in universe	0	0	10th grade	0	Not in universe	Never ma.	Not in
4	<40	Not in universe	0	0	Children	0	Not in universe	Never ma.	Not in
5	<40	Not in universe	0	0	Children	0	Not in universe	Never ma.	Not in
6	>=40	Private	40	10	Some college but	1200	Not in universe	Never ma.	Not in
7	>=40	Private	34	3	Bachelors degree	0	Not in universe	Married-c.	Finan.
8	<40	Private	4	40	High school gradu	0	Not in universe	Never ma.	Not in
9	<40	Local government	43	26	Some college but	876	Not in universe	Married-c.	Educa.
10	<40	Private	4	37	Some college but	0	Not in universe	Never ma.	Not in
11	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
12	<40	Private	0	0	High school gradu	0	Not in universe	Never ma.	Not in
13	<40	Private	4	34	Some college but	0	Not in universe	Divorced	Busin.
14	<40	Private	37	31	High school gradu	0	Not in universe	Divorced	Busin.
15	<40	Private	24	12	Bachelors degree	0	Not in universe	Never ma.	Manuf.
16	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
17	<40	Private	39	36	Bachelors degree	0	Not in universe	Never ma.	Manuf.
18	<40	Private	0	0	10th grade	0	Not in universe	Married-c.	Not in
19	<40	Private	0	0	High school gradu	0	Not in universe	Never ma.	Not in
20	<40	Private	12	41	High school gradu	0	Not in universe	Married-c.	Manuf.
21	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
22	>=40	Self-employed-not incorporated	4	34	Some college but	0	Not in universe	Married-c.	Const.
23	>=40	Private	35	22	Some college but	500	Not in universe	Married-c.	Finan.
24	>=40	Private	45	12	Masters degree(M	0	Not in universe	Married-c.	Other.
25	>=40	Private	0	0	Some college but	0	Not in universe	Married-c.	Not in
26	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
27	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
28	>=40	Private	3	34	Some college but	0	Not in universe	Married-c.	Blngng
29	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
30	>=40	Private	4	2	Bachelors degree	0	Not in universe	Never ma.	Not in
31	>=40	Private	19	35	Less than 1st grade	0	Not in universe	Married-c.	Manuf.

Convert Dataset into I Distinct

id	AAGE	ACLSWKR	ADTIND	ADT_	AHGA	AHRS	AHSCOL	AMARITL	AMJND
1	>=40	Not in universe	0	0	High school gradu	0	Not in universe	Widowed	Not in
2	>=40	Self-employed-not incorporated	4	34	Some college but	0	Not in universe	Divorced	Const.
3	<40	Not in universe	0	0	10th grade	0	Not in universe	Never ma.	Not in
4	<40	Not in universe	0	0	Children	0	Not in universe	Never ma.	Not in
5	<40	Not in universe	0	0	Children	0	Not in universe	Never ma.	Not in
6	>=40	Private	40	10	Some college but	1200	Not in universe	Never ma.	Not in
7	>=40	Private	34	3	Bachelors degree	0	Not in universe	Married-c.	Finan.
8	<40	Private	4	40	High school gradu	0	Not in universe	Never ma.	Not in
9	<40	Local government	43	26	Some college but	876	Not in universe	Married-c.	Educa.
10	<40	Private	4	37	Some college but	0	Not in universe	Never ma.	Not in
11	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
12	<40	Private	0	0	High school gradu	0	Not in universe	Never ma.	Not in
13	<40	Private	4	34	Some college but	0	Not in universe	Divorced	Busin.
14	<40	Private	37	31	High school gradu	0	Not in universe	Divorced	Busin.
15	<40	Private	24	12	Bachelors degree	0	Not in universe	Never ma.	Manuf.
16	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
17	<40	Private	39	36	Bachelors degree	0	Not in universe	Never ma.	Manuf.
18	<40	Private	0	0	10th grade	0	Not in universe	Married-c.	Not in
19	<40	Private	0	0	High school gradu	0	Not in universe	Never ma.	Not in
20	<40	Private	12	41	High school gradu	0	Not in universe	Married-c.	Manuf.
21	<40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
22	>=40	Self-employed-not incorporated	4	34	Some college but	0	Not in universe	Married-c.	Const.
23	>=40	Private	35	22	Some college but	500	Not in universe	Married-c.	Finan.
24	>=40	Private	45	12	Masters degree(M	0	Not in universe	Married-c.	Other.
25	>=40	Private	0	0	Some college but	0	Not in universe	Married-c.	Not in
26	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
27	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
28	>=40	Private	3	34	Some college but	0	Not in universe	Married-c.	Blngng
29	>=40	Private	0	0	Children	0	Not in universe	Never ma.	Not in
30	>=40	Private	4	2	Bachelors degree	0	Not in universe	Never ma.	Not in
31	>=40	Private	19	35	Less than 1st grade	0	Not in universe	Married-c.	Manuf.

ℓ - diversity is a form of group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. It captures the notion of well-represented groups due to the fact that entropy increases as frequencies become more uniform

Convert Dataset into I- Recursive

id	AAGE	ACLSWKR	ADTIND	ADT_	AHGA	AHRS	AHSCOL	AMARITL	AMJND
1	>=40	Not in universe	*	0	0	0	0	0	0
2	>=40	Self-employed-not incorporated	4	0	0	0	0	0	0
3	<40	Not in universe	*	0	0	0	0	0	0
4	<40	Not in universe	*	0	0	0	0	0	0
5	<40	Not in universe	*	0	0	0	0	0	0
6	>=40	Private	40	0	0	0	0	0	0
7	>=40	Private	34	3	0	0	0	0	0
8	<40	Private	4	0	0	0	0	0	0
9	>=40	Local government	43	0	0	0	0	0	0
10	<40	Private	4	0	0	0	0	0	0
11	<40	Not in universe	*	0	0	0	0	0	0
12	<40	Not in universe	*	0	0	0	0	0	0
13	>=40	Private	4	0	0	0	0	0	0
14	>=40	Private	*	0	0	0	0	0	0
15	>=40	Private	24	0	0	0	0	0	0
16	<40	Not in universe	*	0	0	0	0	0	0
17	>=40	Private	39	0	0	0	0	0	0
18	<40	Not in universe	*	0	0	0	0	0	0
19	<40	Not in universe	*	0	0	0	0	0	0

VI. Implementation Steps:

There are following techniques and algorithm in proposed methodology:

1. Data Collection: The proposed system stores the sensitive data in the form of DataStream. Accessing a DataStream is concerned with extracting knowledge represented in non-stopping, continues and ordered sequence of information.
2. Privacy Protection Mechanism: Privacy protection mechanism includes data anonymization. The data anonymization is the process that transforming sensitive data to some anonymized form. The proposed system used ℓ -diversity, it is techniques better than k-anonymity. It has strong background knowledge and maintains lack of diversity. The approach for preserving privacy is based on data anonymization.

IV. REFERENCES

- [1] Zahid Pervaiz, Arif Ghafoor, Walid G. Aref, "Precision-Bounded Access Control Using Sliding-Window Query Views for Privacy-Preserving Data Streams", IEEE Trans. Knowl. Data Eng., July 2015.
- [2] Z.Pervaiz, W.G.Aref, A.Ghafoor, and N. Prabhu, "Accuracy constrained privacy-preserving access control mechanism for relational data", IEEE Trans. Knowl. Data Eng., April 2014.
- [3] T. Ghanem, A. Elmagarmid, P. Larson, and W. Aref, "Supporting views in data stream management systems," ACM Trans. Database Syst., 2010.

- [4] J. Cao, B. Carminati, E. Ferrari, and K. Tan, "Castle: Continuously anonymizing data streams," *IEEE Trans. Dependable Secure Comput.* May/Jun. 2011.
- [5] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. IEEE Int. Conf. Data Eng. Workshop Privacy-Preserving Data Publication Anal.*, 2013.
- [6] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, "Continuous privacy preserving publishing of data streams," in *Proc. 12th Int. Conf. Extending Database Technol.: Adv. Database Technol.*, 2009.
- [7] Sai Wu, Xiaoli Wang, Sheng Wang, Zhenjie Zhang, "k-Anonymity for crowd sourcing Database" *IEEE Trans. Knowl. Data Eng.*, sept 2014.