ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Data Preserving By Anonymization Techniques for Collaborative Data Publishing

R. Indhumathi[1], S. Mohana Priya[2]

PG Scholar, Dept. of CSE, MIET Engg. College, Tamil Nadu, India[1]

Assistant Professor, Dept. of CSE, MIET Engg. College, Tamil Nadu, India[2]

Abstract— This paper mainly deals with the issue of privacy preserving in data mining while collaborating n number of parties and trying to maintain confidentiality of all data providers details while collaborating their database. Here two type of attacks are addressed "insider attack" and "outsider attack". In insider attack, the data providers use their own records and try to retrieve other data provider details. Formal protection model k-Anonymity, l-diversity, t-closeness are used to protect privacy. Here notion of m-privacy algorithm is used to maintain privacy and secure multiparty computation protocol can also be used for privacy preserving.

Keywords— Privacy preserving, Anonymization, SMC, Distributed data

I. INTRODUCTION

Privacy preserving is mainly used to prevent information disclosure. There are two type of information disclosure and they are Identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table, such that attacker can easily identified from the release table. Attribute disclosure occurs when new information about some individuals is revealed. Privacy preserving is different from conventional data security. Privacy preservation techniques are mainly used to reduce the leakage of formation about the particular individual while the data are shared and released to public.

The Anonymization process is carried out to change the data, before its being published to public. The two ways to achieve privacy are, first is to release limited data , so that personal information cannot be identified and second is to pre-compute heuristics and release them instead of any data.

Various Anonymization techniques are being used to maintain privacy and high data utility and they are generalization, suppression, anatomization, permutation and perturbation [1]. Most of the privacy preserving methods use generalization techniques. Various methods are used for Privacy Preserving Data Mining and they are Statistical methods which include Randomization methods, Swapping, Micro Aggregation and Synthetic data generation and the next method is Group based anonymization methods which include k-anonymity, l-diversity, t-closeness.

The classification of attribute in a table is given as key attributes, quasi-identifier (QI) and sensitive attributes. The key attribute is said to be the identifiers, which must be removed before publishing to public, since the attacker can easily identify the particular individual details. For example, consider a table 2 which is student table which is being released by a college by removing the identifiers.

II.THE K-ANONYMITY METHOD

The k-anonymity model requires that within any equivalence class of the micro data there are at least k records. K-anonymity requires each tuple in the published table to be indistinguishable from at least k-1 other tuples. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

cannot be distinguished from at least (k − 1) other records [2].

In the given table 1, student's details are provided such as Department, Age and Course.

Table1. Students Micro Data

	Department	Age	Course
1	ME	20	Mechanics
2	MME	21	Mechanics
3	ME	20	Mechanics
4	CHE	22	Algorithms
5	CHE	23	Psychology
6	CHM	22	Real Analysis
7	CSE	26	Algorithms
8	CSE	25	Algorithms
9	CSE	26	Mechanics

In Table 2 provides 3 equivalent class, here 3-anonymity by generalization is achieved.

Table2. 3-anonymous Students Micro Data

	Department	Age	Course
1	M*	[20-21]	Mechanics
2	M*	[20-21]	Mechanics
3	M*	[20-21]	Mechanics
4	CH*	[22-23]	Algorithms
5	CH*	[22-23]	Psychology
6	CH*	[22-23]	Real Analysis
7	CS*	[25-26]	Algorithms
8	CS*	[25-26]	Algorithms
9	CS*	[25-26]	Mechanics

K-anonymity cannot provide a safeguard against attribute disclosure. Various types of attacks are addressed in k-anonymity and they are homogeneity attack and the background knowledge attack. In table 2, the first equivalence class has courses as Mechanics, which is same for all students with in age (20-21). This type of attack is said to be homogeneity attack.

In the same way, in table 2 if a student is known who is doing CSE and he is not interested in mechanics, then it is easy to predict that particular student is from the third equivalent class with help of the background Knowledge of the particular person. This type of attack is considered background knowledge attack.

III. ℓ-DIVERSITY METHOD

ℓ-diversity is used to overcome the drawback of k-anonymity and tries to put constraints on minimum number of distinct values seen within an equivalence class for any sensitive attribute.

Definition 1 (The ℓ-diversity Principle): An equivalence class is said to have ℓ-diversity if there are at least ℓ " well -represented" values for the sensitive attribute. A table is said to have ℓ -diversity if every equivalence class of the table has ℓ-diversity [3].

The given table is said to be ℓ-diversified if every equivalence classes in the table contains at least ℓ well-represented sensitive attribute values. ℓ-diversity must guarantee that the SA value of a particular person cannot be identified unless the adversary has enough background knowledge to eliminate ℓ−1 SA values in the person's EC. Several measures were proposed to quantify the meaning of "well-represented" of ℓ-diversity. These include entropy ℓ-diversity [3], recursive (c,ℓ)-diversity [3] and simple ℓ-diversity.

There are two type of attacks faced in ℓ-diversity and they are Skewness attack and Similarity attack. The attribute disclosure cannot be overcome in ℓ-diversity, but identity disclosure is successfully handled.

IV. DISTRIBUTED DATA PUBLISHING

The data are gathered from multiple users and they are collaborated [4] and two process can be carried out one is aggregation is done and then it is anonymized and another type is first the data are anonymized and then they are aggregated.

In figure 1(b), the Collaborative data publishing is carried out successfully with help of trusted third party (TTP) or

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Secure Multi-Party Computation (SMC) protocols, that guarantees that the information or data about particular individual is not disclosed anywhere, the privacy is maintained with help of SMC and there will be better data utility. Here it is assumed that the data providers are semi honest. So certain protocols are set and the all data providers accept that protocol and they continue the process.

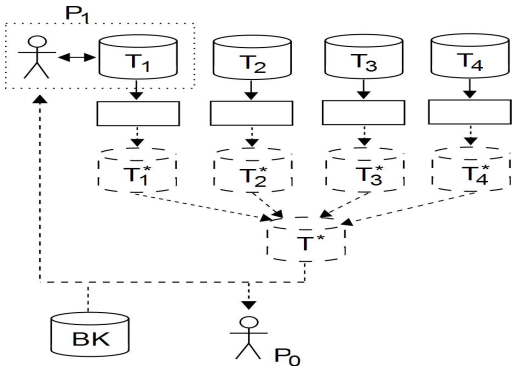

Figure 1(a). Anonymize-and-aggregate

In figure 1(a), the data providers are T1, T2, T3 and T4, here the data provider anonymize their own data and then they are aggregated and represented as T* and they are provided to the final user.

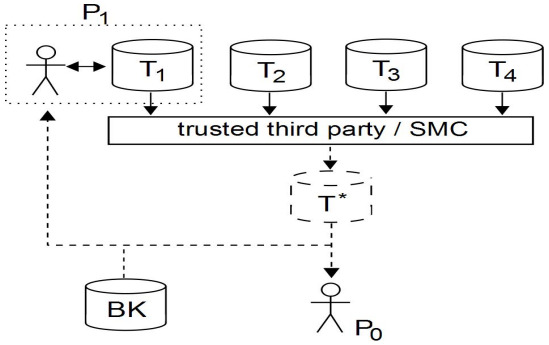

Figure 1(b). Aggregate-and-anonymize

In figure 1(b), the whole data is collected from the data providers and they are aggregated using trusted third party or SMC and then they are anonymized. In these two types of methods two types of attacks are faced and they are insider attack and outsider attack. If the attack is made by the data providers then they are treated as "insider attack" and if the attack is carried out by the outsider then that type of attack is said to be "outside attack". Here it is mainly focused on insider attack.

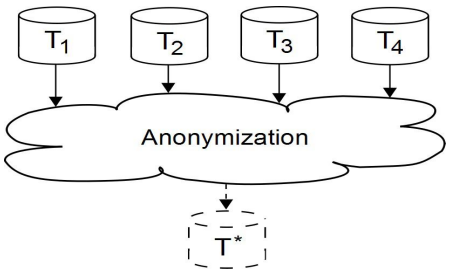

Figure 2.Collaborating 4 database of different providers

While collaborating data from different data providers, three types of algorithms are used here, to maintain privacy and they are

- The notion of *m*-privacy algorithm
- Heuristic algorithms
- Data provider-aware anonymization algorithm

a) Notion of *m*-privacy algorithm

In notion of m-privacy algorithm, main aim is to prevent data of an individual in anonymized table from m-adversaries. Where m-adversaries is a coalition of data users with *m* data providers cooperating to breach privacy of anonymized records. Here constraint C is set and privacy is checked against C for the data in anonymized data. M-privacy is defined with respect to privacy constraint C.

C holds the truthfulness of record level. Privacy is maintained for duplicate record too. For example if same record is provided from two different hospitals, then the particular individual detail can be easily identified with help of background knowledge, but it can be prevented with help of constraint C.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Monotonicity of privacy constraints is defined for a single equivalence group of records, i.e., a group of records that QI attributes share the same generalized values.

Definition 2.2: (GENERALIZATION MONOTONICITY OF A PRIVACY CONSTRAINT [3], [6]) A privacy constraint *C is* generalization monotonic if and only if, for any two
Equivalence groups $A1(T)$ and $A1(T\_)$ that satisfy *C*, their union satisfies *C* as well $C(A1(T)) = true$ & $C(A1(T\_)) = true$
so, $C(A1(T) \cup A1(T)) = true$.

b) Heuristic algorithm

In heuristic algorithm *m*-privacy is efficiently checked with respect to an EG monotonic constraint. Then, it is modified to check *m*-privacy with respect to a non-EG monotonic constraint. The main aim for heuristic for EG monotonic privacy constraints is to search the adversaries with effective pruning, so that no need to check m-adversaries.

Here two types of pruning strategies are used and they are Downward pruning and Upward pruning. In Downward pruning approach, if a coalition does not maintain privacy, then the sub-coalition of m-adversaries is no need to be checked, since that too won't maintain privacy. In upward pruning process,
If the coalition is able to maintain privacy, then the super coalition will also maintain privacy.

The algorithm used here is Top-down algorithm and Bottom–up algorithm. The Top-down algorithm uses downward pruning strategies such that The topdown algorithm will check all $(n - 1)$-adversaries first, then smaller coalitions up to all *m*-adversaries and the Bottom-up algorithm uses upward pruning such that the bottom-up algorithm will check 0-adversary up to all *m*-adversaries.

By using these algorithms the time needed to check m-adversaries is saved. And the process is carried out fast.

c) Data provider-aware anonymization algorithm

The Data provider-aware anonymization algorithm is presented with adaptive *m*-privacy checking strategies to ensure high utility and *m*-privacy of anonymized data with efficiency. The above said algorithm is used on different condition, depending upon the data providers. The pruning strategies are selected according to the privacy and data utility and suitable algorithm is selected. Mostly top-down algorithm with downward pruning is used which reduces m-adversary check.

These are the three algorithm used in collaboration process to maintain privacy.

The above used algorithm runs with help of Trusted Third Party (TTP). The third party used here might be semi honest, and can't be trusted. To overcome this SMC protocol is used. Secure protocol verifies the privacy with respect to constraint C. SMC protocols are based on Shamir's secret sharing [7], encryption, and other secure schemas. SMC protocol uses bottom-up approach.

TTP can identify if duplicate record occurs from the data providers, but SMC protocol cannot detect the duplicate record. SMC [5] is mainly used to control the"insider attacker". The SMC uses two computation concepts and they are Ideal model and Real model paradigm.

V. M-ANONYMIZER

The process carried out in m-Anonymizer is explained with help of flow chart. These are the following steps followed:-

- Data from m providers are collected and collaborated.
- Next step is to identify the split point which is split horizontally until privacy is maintained.
- After doing splitting, the privacy constraint C is selected such a way that ensures privacy for all individual data. M privacy is checked with respect to the Constraint C.
- Next step is to check whether it is again split able, if it is possible then again the score is detected and the process from step 2 is again carried out.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

- Next step is finding the privacy fitness score, which quantifies the level of privacy fulfilment of the group and the most suitable algorithm, is selected.

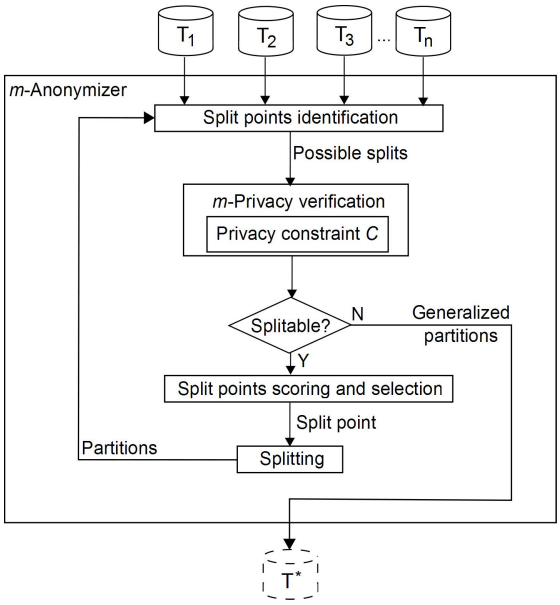

Figure 3.m-Anonymizer

- If it is not split able then the final anonymized table is finalized, which maintains privacy and data utility.

VI. EXPERIMENTS

i. m-Privacy verification runtime for different algorithms vs *m*

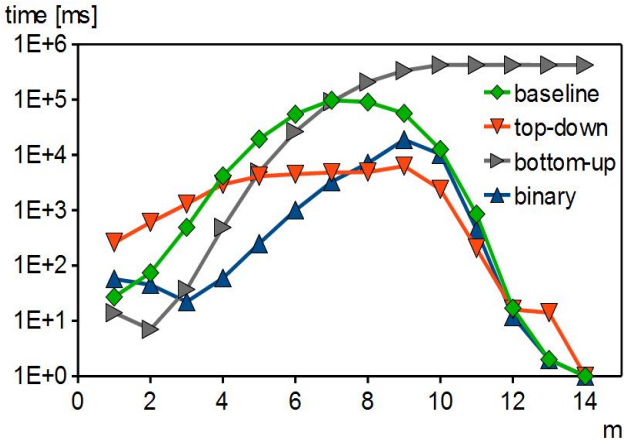

Figure 4(a). Average privacy fitness score per Provider = 0.8

In this experiment, we compare *m*-privacy verification heuristics against different attack powers, and different number of data providers. Fig. 4(a) shows computation time with varying *m* and *nG* for all heuristics.

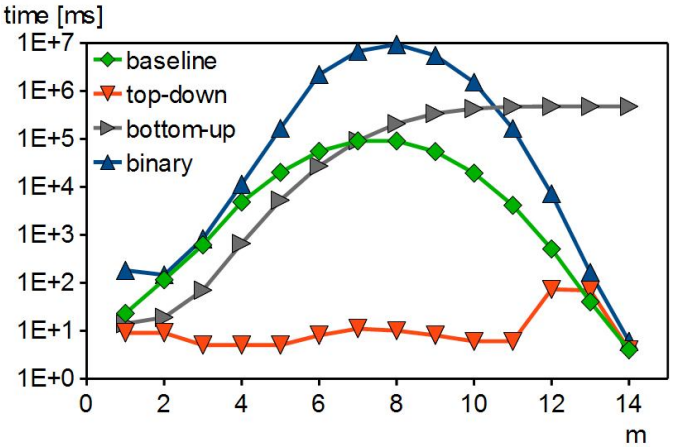

Figure 4(b).Average privacy fitness score per provider = 2.3

ii. *m*-Anonymizer runtime and query error for different anonymizers vs number of data records.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

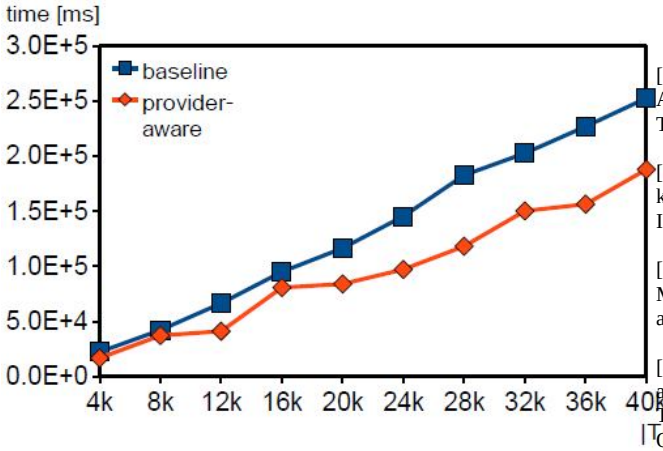

Figure 4(c).Computation time vs. *m* and the number of providers

In Fig. 4(c) shows the estimated computation time with varying *m* for both approaches. In addition, both approaches have comparable computation times with negligible differences.

VII. CONCLUSION

While collaborating the data of different data provider, two types of attacks are identified and they are insider attack and outside attack. The insider attack occurs from the data providers and attacker will be within the data provider, here it mainly deals with insider attack and how the individual detail is preserved from the attacker. And also data utility is also increased. Mainly three algorithms are used and they are notion of m-privacy, heuristic algorithm and adaptive provider aware algorithm. And the best method followed is to use SMC (secure multiparty computation) protocol, which is better than the TTP process.

VIII. FUTURE ENHANCEMENT

Here the privacy is preserved only when there are limited numbers of data providers (45,222), but when the data providers increases, the privacy is not protected against m-adversaries. So work is carried out to maintain high data utility and to protect the privacy of the individual data of the data provider from adversaries.

REFERENCES

[1].Karthikeyan.B,Manikandan. G,Vaithiyanathan. V," A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING", Journal of Theoretical and Applied Information Technology,2011,Vol. 32 No.2.

[2].Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. IEEE Symp. On Security and Privacy, 1998.

[3].AshwinMachanavajjhala, Johannes Gehrke, DanielKifer, Muthuramakrishnan Venkita subramaniam , ℓ-diversity: privacy beyond k-anonymity, IEEE International Conference on Data Engineering, 2006, p. 24.

[4].N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee,"Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowl. Discovery from Data, vol. 4,no. 4, pp. 18:1–18:33, October 2010.

[5].Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

[6] N. Li and T. Li, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007.

[7] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11,pp. 612–613, nov 1979.