Survey Result on Privacy Preserving Techniques in Data Publishing

S.Deebika
PG Student, Computer Science and Engineering, Vivekananda College of Engineering for Women, Namakkal India

A.Sathyapriya
Assistant professor, Computer Science and Engineering, Vivekananda College of Engineering for Women, Namakkal, India

S.K.Kiruba
Assistant professor, Computer Science and Engineering, Vivekananda College of Engineering for Women, Namakkal, India

Abstract—Data Mining which is sometimes also called as Knowledge Discovery Data (KDD) is the process of extracting useful pattern from large databases into useful information. Data mining consist of many specific areas they are Clustering, Classification, Privacy, Text mining, Visual data mining, Web mining. In this, brief about Data Security. Privacy preserving in data publishing is most important research area in data security field. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy and analyzed many techniques used for data privacy and their reward and short comes are noted very well. Some of the privacy techniques such as k-anonymity, l-diversity, differential privacy. As finally, proposed a (n,t)-Closeness technique for preserve the sensitive information.

Keywords: - Privacy Preservation, Data publishing, k- Anonymization, Data Security, PPDP, (n,t)-closeness.

I. INTRODUCTION

Privacy means anonymize the data. It can be done means by Removing "personally identifying information" (PII)such as Name, Social Security number, phone number, email, address… etc anything that identifies the person directly so it can be anonymized.

A. Basic form of privacy-preserving data

In the most basic form of privacy-preserving data publishing (PPDP) [3], the data holder has a table of the form: D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number SSN), containing information that explicitly identifies record owners, Quasi Identifier is a set of attributes that could potentially identify record owners, Sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.

II. BACKGROUND

Most works assume that the four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner. Large amount of person-specific data has been collected in recent years both by governments and by private entities. Data and knowledge [6] extracted by data mining techniques represent a key asset to the society. Analysing trends and patterns. Formulating public policies Laws and regulations require that some collected data must be made public. For example, Census data. Motivation of Privacy [2] is to publicly release statistical information about a dataset without compromising the privacy of any individual. The main requirement of the Privacy data is anything that can be learned about a respondent from a statistical database should be learnable without access to the database. Reduce the knowledge gain of joining the database. Require that the probability distribution on the public results is essentially the same independent of whether any individual options in to, or options out of the dataset.

A. Data Collection and Data Publishing

A typical scenario of data collection and publishing is described in Fig.1. In the data collection phase, the data holder collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data holder releases the collected data to a data miner or the public, called the data recipient, who will then conduct data mining on the published data.

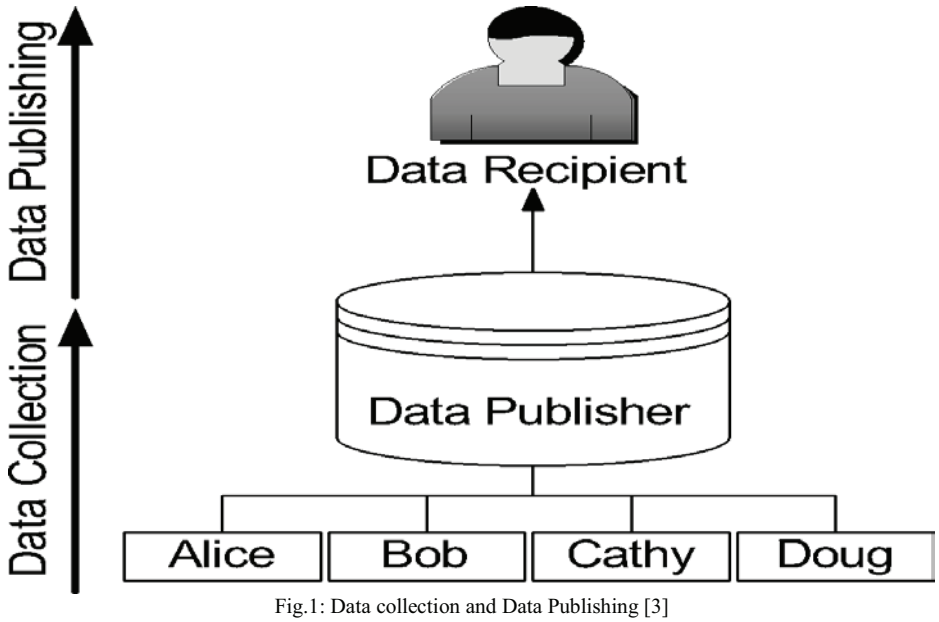

Fig.1: Data collection and Data Publishing [3]

III. VARIOUS APPROACHES ON PRIVACY PRESERVATION

Privacy preserving paradigms have been established: k-anonymity [4], which prevents identification of individual records in the data, and l-diversity [12], which prevents the association of an individual record with a sensitive attribute value.Differencial privacy [1] guarantees privacy even if an attacker knows all but one record and so on. In this we explain detail summary on this principles. Let us see below.

A. K-Anonymity

K-Anonymity [4] means if the information for each person contained in the release cannot be recognized from at least k-1 individuals whose information also appears in the release. For e.g., if you try to identify a man from a release, but the only information you have is his birth date and gender. There are k people meet the requirement. This is said to be k-Anonymity. The database is assumed to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. K-Anonymity thus prevents definite database linkages. K-Anonymity assures that the data released is perfect. K-anonymity has two techniques: generalization and suppression [8]. To protect respondents' identity when releasing micro data, data holders are often eliminate or encrypt explicit identifiers, such as names and security numbers. K-anonymity does not provide guarantee of anonymity in de-identifying of data. Every released information often contains other data, for example: birth date, sex, and ZIP code that can be publicly available and it cannot intend for release. One of the talented concepts in micro data protection is k-anonymity [9]. One of the interesting aspects of k-anonymity is its association with protection techniques that preserve the honesty of the data. The assurance given by k-anonymity is that no information can be linked to groups of less than k individuals. In high-dimensional data Generalization for k-anonymity wounded considerable amount of information.

Limitations of k-anonymity

Here some Limitations of k-anonymity [4] are: (1) it does not cover given individual in the database, (2) it reveals individuals sensitive attributes, (3) it does not protect background knowledge attack, (4) k-anonymization algorithm can violate privacy, (5) it is not applicable to high-dimensional data without complete loss of utility, and (6) If more than one special methods are published it required the dataset is anonym zed.

B. l-Diversity

To resolves the drawbacks of anonymity Machanavajjhala et al. [12] proposed the diversity principle called l-Diversity, to prevent attribute linkage.

L-diversity principle: A q-block is l-diverse if contains at least l 'well represented" values for the sensitive attribute S. A table is l-diverse if every q-block is l-diverse. ℓ-Diversity provides privacy preserving even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main idea of ℓ -diversity is the requirement that the values of the sensitive attributes are well-represented in each group. The l-diversity can be of two approaches. First, Distinct l-diversity Each equivalence class has at least *l* well-represented sensitive values Second, Entropy *l*-diversity Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough. It means the entropy of the distribution of sensitive values in each equivalence class is at least log (l).Sometimes this may be too restrictive. When some values are very common, the entropy of the entire table may be very low. This leads to the less conservative notion of l- diversity. Declare if you have a group of k different records that can share all particular quasi-identifier. An attacker cannot identify the individual based on the quasi-identifier. But what if the value they're interested in, (e.g. the individual's medical diagnosis) is the same for every value in the group. The distribution of target values within a group is referred to as "l-diversity"[1]. It resolves the short comes of k-anonymity.

Limitation of l-Diversity

Doesn't prevent the probabilistic inference attacks which tends to be more intuitive to the human data publisher. For Eg.In one equivalent class, there are ten tulles. In the "Disease" area, one of them is "Cancer", one is "Heart Disease" and the remaining eight are "Flu". This satisfies 3-diversity, but the attacker can still affirm that the target person's disease is "Flu" with the accuracy of 80%.

C. ϵ-Differential privacy

ϵ-Differential privacy [2] compares the risk with and without the record owner's data in the table. It does not prevent linkage, but it assures that new records won't make any difference on discovery that could not have been figured out before.

D. Differential privacy

Next we move on to the differential privacy [7].The risk to my privacy should not substantially increase as a result of participating in a statistical database. Fig.2 shows ratio bounded by using differential privacy. Let us see the definition first, A randomized function K gives ε-differential privacy if for all values of DB, DB 'differing in a single element, and all S in Range (*K).*It can be defined

$$\frac{[\Pr K(DB) \text{ in } S]}{Pr[K(DB') \text{ in } S]} \leq e^{\varepsilon} \sim (1+\varepsilon)$$

No perceptible risk is incurred by joining DB. Any info adversary can obtain, it could obtain without Me (my data).There are two models: Interactive and Non-Interactive [10].Interactive: Multiple Queries, Adaptively Chosen. Non-Interactive: Data are sanitized and released.

For f: $D \rightarrow R^k$, the sensitivity of f is

$$\Delta f = \max_{D1D2} \| f(D_1) - f(D_2) \|_1$$

For all D_1, D_2 differing in at most one element.

Captures how great a difference must be hidden by the additive noise.

Limitations of Differential privacy:

It does not preserve data truthfulness at the record level. It does not prevent Record linkage and table linkage.

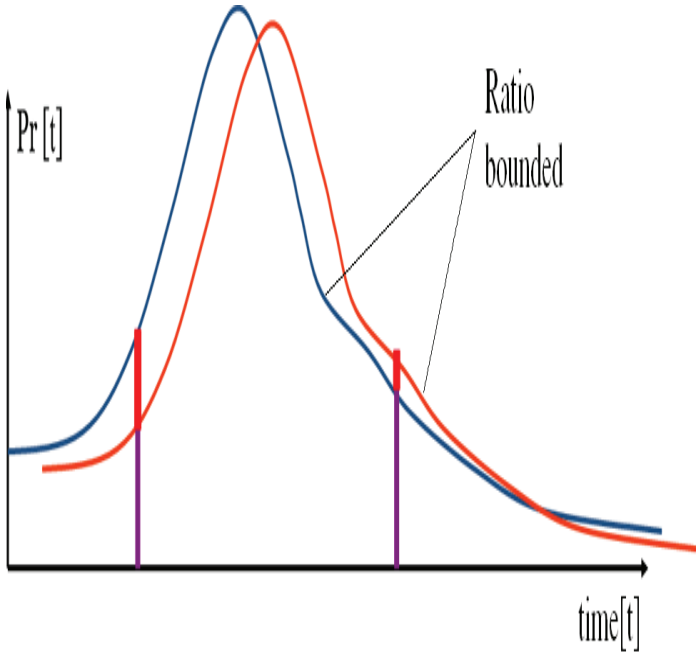

Fig.2 ratio bounded in differential privacy

E. t-closeness

K-anonymity prevents identity disclosure but not attribute disclosure. To solve that problem l-diversity requires that each eq. class has at least l values for each sensitive attribute. But, l-diversity has some limitations t-closeness [11] requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table. Privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the multiple sensitive attribute value of an individual.

We present the base model t-closeness, which requires that the distribution of a multiple sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We then propose a more flexible privacy model called on t-closeness that offers higher utility.

Limitations of T-closeness

There is no computational procedure to enforce t-closeness [11] followed in. There is effective way till now of combining with generalizations and suppressions [8] or slicing [13]. Lost co-relation between different attributes. This is because each attribute is generalized separately and so we lose their dependence on each other.
Utility of data is damaged if we use very small t.(And small t will result in increase in computational time.

IV. COMPARISON OF PRIVACY TECHNIQUES

The below TABLE. I show the different privacy models with their advantage and disadvantage

TABLE. I COMPARISON OF PRIVACY TECHNIQUES

Privacy models	**Advantages**	**Disadvantages**
K-anonymity	K-Anonymity assures that the data released is perfect.	it does not cover given individual in the database.
l-diversity	ℓ-Diversity provides privacy preserving even when the data publisher does not know what kind of knowledge is possessed by the adversary.	Doesn't prevent the probabilistic inference attacks which tend to be more intuitive to the human data publisher.
Differential privacy	Any info adversary can obtain, it could obtain without Me (my data).	It does not preserve data truthfulness at the record level.
t-closeness	More flexible privacy model called on t-closeness that offers higher utility.	There is no computational procedure to enforce t-closeness.

V. CONCLUSION

From the above analysis of privacy preserving in data publishing there are many complexity of privacy problems. It shows that K-anonymity, l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new notion of privacy called "(n,t)-closeness."It helps to preserve the data from attackers. It is better than the previous approaches. By using this, ideal global knowledge and background attackers are prevented.

REFERENCES

[1] Arik Friedman, Assaf Schuster, Data Mining with Differential Privacy. KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys, 42(4), December 2010.

[3] Benjamin C.M. Fung, Ke Wang, Rui Chen, Philip S. Yu. Privacy-Preserving Data Publishing A Survey on Recent Development

[4] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality,"Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.

[5] Dwork, C., Nissim, K.: Privacy-Preserving Data mining on Vertically Partitioned Databases. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 528–544.Springer, Heidelberg (2004)

[6] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.generalisation

[7] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan. Computational differential privacy. In Advances in Cryptology — Crypto 2009, volume 5677 of LNCS, pages126–142. Springer, 2009.

[8] L. Sweeney (2002). Achieving k-anonymity privacyprotection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol 10(5), pp. 571–588

[9] L. Sweeney. Guaranteeing anonymity when sharing medical data, the Datafly system.Proceedings, Journal of the American Medical Informatics Association. Washington,DC: Hanley & Belfus, Inc., 2000

[10] L. Sweeney, "Differencial privacy: a model for protecting privacy",International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002, pp. 557-570

[11] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-anonymity and l-Diversity", in proc. of icde, 2007, pp.106-115

[12] machanavajjhala, a., gehrke, j., kifer, d., and venkitasubramaniam, m. [2006-2007]. l-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering(ICDE).
[13] Tiancheng Li, Ninghui Li, Senior Member, IEEE, JiaZhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" proc.IEEE transactions on knowledge and data engineering, vol. 24, no. 3, march 2012