Privacy Preserving by Anonymization Approach

Hetaswini J. Thathagar[1]
Computer Engineering
L.D. College of Engineering
Ahmedabad, India
Thathagarhetaswini93@gmail.com

Vimalkumar B. Vaghela
Computer Engineering
L.D. College of Engineering
Ahmedabad, India
vimalvaghela@gmail.com

***Abstract*—** Privacy Preserving takes more attention in data mining because now a days people registers every days on so many sites and gives their personal details like DOB, Zip code, etc. and from that any attacker can get sensitive data of individual person so here privacy is breached due to this problem Randomization, Anonymization, Cryptography, Partition, other many approaches are introduced from them each approach have their own limitations and Anonymization have Information Loss and In some case privacy also breached so for that this new approach introduce which will decrease Information Loss and increase privacy.

Keywords-*Big Data, Anonymization, Generalization, Suppression, k-anonymity*

*****

I. INTRODUCTION

Datasets that contain useful or sensitive information about people and which reveal private or sensitive data about individual people have always been in the focus of research. Researchers apply various methods to extract valuable information from data sets to understand people and make predictions for their future behavior. While legal systems may vary over each country it is depend on it' s legacy.

Specially, a specific type of personal data called sensitive data, e.g. ethnicity, religious affiliation, medical condition, can only be accessed, transferred or handled by entities explicitly stated in regulations, and with the consent of the data subject. Health care databases have an especially strict regulation because of the large number of sensitive data contained. For instance, pharmaceutical research must work with accurate data, but that retains all sensitive patient data as well, hence researchers working with such databases stumble very early in the legal limitations.

Records of health care databases hold sensitive information from which one may be able to reveal medical condition of a person. Medical conditions may relate to e.g. food consumption preferences, life expectancy, drug taking habits, and other personal strengths or weaknesses. In wrong hands, e.g. decisions over employments, or mortgages might depend on such information which would be very unethical to use, and it must be avoided at all costs.

On the other hand, health care databases also serve as the basis for better health care services, drug developments, and cost efficiency which also are in the focus of public interests. Therefore, before publishing any piece of information from the database, it has to go through an data privacy preserving procedure to hide sensitive data.

Hence researchers must be aware of the legal requirements, the methods applicable to meet these requirements and the level protection these techniques provide. here uniquely identifiers like name or id numbers removed before publishing so identify individual record become harder but what if some attacker have another details table of some persons like zip code, birth date, gender detail and he will join that table with this table and he can identify uniquely individual people and privacy will be breach. so, for that before publishing database data publisher apply these methods

II. LITERATURE SURVEY

There are so many techniques in privacy preserving but we only talk about some of them techniques these techniques are randomization, k-anonymization, l-diversity, t-closeness, cryptography, partition methods these techniques are very popular for research object and now a day's privacy is very important for individual people no matter if data of hospital, bank, insurance company or any other organization. Privacy is required so before getting original data openly these methods will be applied on original data for privacy preserving.One related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preserving methods data usability, information loss privacy protection of data is required parameter on basis these methods can be comparable and each have their own advantage and disadvantage in this paper our main goal is compare these methods on these parameters finalize which better in which scenario.

TABLE I LITERATURE REVIEW

Sr.No	Methods	Privacy Protection	Usability of Data	Information loss
1	Randomization	No	No	yes
2	Anonymity	yes	Yes	Very less
3	Horizontal partition	yes	Yes	yes
4	Vertical partition	yes	Yes	yes
5	Cryptography	yes	No	no

III. TYPES OF TECHNIQUES

A. *The randomizationmethod by distortion*

Randomized approach it's called distortion method.and distorted value will be store in centralized server.this method will distort original value and replace it with other distorted value that is why it is called distortion method and there are two types of this randomization method and they are **additive perturbation method** [1], **multiplicative perturbation** [3]. The additive perturbation method is also called value distortion method [1]. In this method a random value drawn from a given distribution, e.g., a uniform or normal distribution rule and this distorted value will replace with original value so by this approach we can save original value get revealed moreover there are some disadvantage like information loss and data utility decreased other thing is privacy get revealed in such cases like for e.g. In given below scenario clark age is 25 and after randomization remains 25 that means value 0 is added in original value that is called privacy breached no meaning of this approach.

TABLE II DATASET OF CREDIT CARD DETAILS

Name	Age	Sex	Zipcode	Disease
Smith	25	M	13001	Flu
Clark	25	M	12057	HIV
David	26	M	14599	Bronchitis
Ana	26	F	13001	Pneumonia
Rosy	27	F	17000	Hepatitis

TABLE III DATABASE WITH UNIFORMDISTORTION

Name	Age	Sex	Zipcode	Disease
Smith	27.8	M	13028	Flu
Clark	25	F	12060	HIV
David	14	M	14600	Bronchitis
Ana	23	M	13017	Pneumonia
Rosy	24.2	F	17011	Hepatitis

The additive perturbation method is adding some random number which is in given range to each records value and make new value range we can say this approach as distortion approach in which value will distort so privacy will preserve [11], for instance, when age in range <18 ,65> is randomized by adding a random value from the uniform distribution in range <-18,18>and the distorted value is equal to 0, then it is known that the true age is 18.one more thing is what if we need original data back? Then metrics spectral filtering method is defined for getting original data.[11] here is one problem if we apply this spectral filtering method then it will return back some error in result so this is less effective because of this reason information loss is also high. In the multiplicative perturbation this is second method for perturbation in this value will be place by some multiplicative value of original value[2].

B. *Anonymity*

A record in a data set consists of four types of attributes:

- Identifiers: These are the attributes which can directly identify an individual. So these attributes are removed before publishing the data. For e.g. name, social security number,etc.
- Quasi Identifiers: These are the attributes which cannot identify an individual directly but if they are linked with publicly available data then they can identify an individual easily. For e.g. zip code, age, sex,etc.
  An Equivalence Class is a set of records that have same value for all the quasi-identifier attributes.
- Sensitive Attributes: These are the attributes which an individual wants to hide from others. For e.g. disease, salary.

Linkingattack

Before publishing data, Identifiers (Name of the individual, Social Security Number) are removed. But still there are many attributes (quasi-identifiers) that if combined with external data can identify an individual easily. For e.g., suppose table 4 has to be released for further analysis.

TABLE IV DATABASE OF HOSPITAL

Name	Age	Sex	Zipcode	Disease
Smith	25	M	13001	Flu
Clark	25	M	12057	HIV
David	26	M	14599	Bronchitis
Ana	26	F	13001	Pneumonia
Rosy	27	F	17000	Hepatitis

Before releasing table 4, Identifiers (in this case Name attribute) must be remove from the table.

TABLE VAFTER REMOVING IDENTIFIERS FROM HOSPITAL DATASET

Age	Sex	Zip code	Disease
25	M	13001	Flu
25	M	12057	HIV
26	M	14599	Bronchitis
26	F	13001	Pneumonia
27	F	17000	Hepatitis

Now suppose there is an external data which is available to the attacker. Following table shows the external data which is a Voter Registration List available to the attacker.

TABLE VIVOTER REGISTRATION TABLE

Name	Age	Sex	Zip code
David	26	M	14599
Ana	26	F	13001
Clark	25	M	12057
Smith	25	M	13001
Rosy	27	F	17000

On joining Table 5 and Table 6, the attacker will get to know that Clark is suffering from HIV. So even after removing the identifiers, an individual can be re-identified with the help of data available publicly. Combining data of the released table with the data of the publicly available table is known as Linking Attack.

Data Anonymization is a technique which is used to preserve privacy when big data is published to third parties. Anonymization refers to hiding sensitive and private data of the users. Anonymization can make use of many techniques,

viz. generalization, suppression, perturbation, anatomization and permutation [4]. Mostly generalization and suppression techniques are used for anonymizing data because data anonymized using generalization and suppression still have high utility. So, this data can be used further by researches [5]. Generalization refers to replacing a value with more generic value. Suppression refers to hiding the value by not releasing it at all.Thevalueisreplacedbyaspecialcharacter,e.g.@,*. Generalization impacts all the tuples while suppression impacts a single tuple [6].

Privacymodels

Privacy models are approaches by which privacy can be maintained by releasing partial data or distorted data or generalized data or encrypted data for that different privacy models are available randomization approach, k anonymity approach, l diversity approach, partition approach, cryptography approach.

K-Anonymity: This privacy model is used to prevent Linking Attacks. Sweeney and Samarati proposed the k- anonymity principle [8]. According to k-anonymity principle, a tuple in the published data set is indistinguishable from k-1 other tuples in that data set. Therefore, an attacker who knows the values of quasi- identifier attributes of an individual is not able to distinguish his record from the k-1 other records [7]. k-Anonymity uses generalization and suppression techniques to hide the identity of an individual [5]. For e.g., Table 4 is 2-anonymous table, i.e., two tuples have same values in the quasi-identifier attributes (in this case, Age, Sex and ZipCode).

TABLE VII 2ANONYMOUS TABLE ON HOSPITAL DATASET

Age	Sex	Zip code	Disease
[20-40]	M	18***	HIV
[20-40]	M	18***	HIV
[41-50]	F	120**	Cancer
[41-50]	F	120**	Heart disease

Although k-anonymity can solve the problem of identity disclosure attack, it cannot solve the problem of attribute disclosure attack. For e.g., if the sensitive attribute lack diversity in values and attacker is only interested in knowing the value of sensitive attribute then the aim of attacker is achieved. This type of attack is known as Homogeneity Attack.

For e.g., if an attacker has Table 8 available as an external data, then he can link the table 4 and table 8 and can come to a conclusion that Andrew is suffering from HIV.

TABLE VIIIVOTER REGISTRATION TABLE

Name	Age	sex	Zip code
Andrew	31	M	18601
Clarke	27	M	18555
Rosy	49	F	12001
Ana	42	F	12456

Another kind of attack which k-anonymity cannot prevent is Background Attack. This model assumes that attacker has no additional background knowledge. Suppose, if attacker knows that Andrew is male and his zip code starts with 18table 7 and 8, attacker can conclude that Ana is suffering from a HIVdisease.

l-diversity

To prevent attribute disclosure attack, it was the next privacy model which was proposed. According to l-diversity model, an equivalence class must have l "well-represented" values for sensitive attributes. It is also known as Distinct l- diversity. For e.g., following table is 2-diverse, i.e., each equivalence class contains two distinct values for sensitiveattributes.

TABLE IX 2-DIVERSE TABLE

Age	Sex	Zip code	Disease
[21-30]	M	140**	Flu
[21-30]	M	140**	Bronchitis
[31-50]	F	17***	Pneumonia
[31-50]	F	17***	HIV

Distinct l-diversity model suffers from probabilistic inference attacks. For e.g. consider the following table.

TABLE XTABEL FOR PROBABILISTIC ATTACK

AGE	SEX	ZIP CODE	DISEASE
[21-30]	M	120**	HIV
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU
[21-30]	M	120**	PNEUMONIA
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU

There is only one equivalence class in table 10. The table is 3-diverse because it contains three distinct values in the sensitive attribute Disease. But five out of seven records contain Flu in the Disease attribute. So, an attacker can affirm that disease of target person is Flu with accuracy of 70%. l-diversity model is difficult to achieve. Moreover, this model also suffers from skewness and similarity attacks.

Consider following table for understanding the concept of skewness attack. There are two sensitive attributes Salary and Disease.

TABLE XI DATASET OF HOSPITAL PATIENT

Age	Zip code	Salary	Disease
24	12889	2k	Gastric ulcer
26	12110	3k	Gastritis
28	12005	4k	Stomach cancer
31	15601	6k	Flu
33	15666	7k	Bronchitis
37	15689	9k	Cancer
43	19123	11k	Heart disease
45	19765	12k	Gastritis
49	19303	14k	Pneumonia

Following table is 3-diverse version of the above table.

TABLE XII 3-DIVERSION VERSION OF TABLE XI

Age	Zip code	Salary	Disease
[21-30]	12***	2k	Gastric ulcer
[21-30]	12***	3k	Gastritis
[21-30]	12***	4k	Stomach cancer
[31-40]	156**	6k	Flu
[31-40]	156**	7k	Bronchitis
[31-40]	156**	9k	Cancer
[41-50]	19***	11k	Heart disease
[41-50]	19***	12k	Gastritis
[41-50]	19***	14k	Pneumonia

Now suppose that attacker knows that Alice has low salary (2K-4K). Then he can conclude that Alice is suffering from some stomach disease. This is known as Similarity Attack because there is some kind of similarities in the values of sensitive attribute Disease [9].

Now consider following table to understand concept of skewness attack. Suppose there are 1,00,000 records of a virus and that virus attacks only 1% of the population. Third equivalence class consists of equal number of positive and negative records. In other words, everyone in that class has 50% chance of having the class which is much higher than the real distribution[9].

TABLE XIII HOSPITAL DATASET FOR SKEWNESS PROBLEM

Age	Zip code	Salary	Disease
[11-20]	12***	2k	Negative
[11-20]	12***	3k	Negative
[21-30]	156**	4k	Negative
[21-30]	156**	6k	Negative
[31-40]	19***	7k	Negative
[31-40]	19***	9k	Positive
[41-50]	170**	11k	Negative
…	…	…	…
[81-90]	170**	12k	Negative

C. Cryptographic

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties.

Some Cryptography Methods:

1. Secure Multiparty Computation All these methods are almost based on a unique encryption protocol called as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining made up of a set of protected sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product [19]. Safe, Secure, Trust-worthy Communication complexity grows exponentially with n.

2. Public-key cryptosystems (asymmetric ciphers) A public-key (asymmetric key) algorithm uses two separate keys: a public key and a private key. The public key is used to encrypt the data and only the private key can decrypt the data. A form of this type of encryption is called RSA (discussed below), and is widely used for secured websites that carry sensitive data such as username and passwords, and credit card numbers. In asymmetric or public key, cryptography there is no need for exchanging keys, thus eliminating the key distribution problem. The primary advantage of public-key cryptography is increased security the private keys do not ever need to be transmitted or revealed to anyone. Can provide digital signatures that can be repudiated A disadvantage of using public-key cryptography for there are popular secret-key encryption methods which are significantly faster than any currently available public-key encryption method.

E. Distributed Privacy-Preserving Data Mining

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy preserving datamining.

1.Semi-honest Adversaries**:** Participants Alice and Bob are curious and attempt to learn from the information received by them during the protocol, but do not deviate from the protocol themselves. In many situations, this may be considered a realistic model of adversarialbehavior.

2.Malicious Adversaries**:** Alice and Bob may vary from the protocol, and may send sophisticated inputs to one another to learn from the information received from eachother.

1.Horizontal Partitioning

In this case, the different sites may have different sets of records containing the same attributes.

TABLE XIV PART 1 HOSPITAL DATASET

Age	Sex	Zip code	Disease
25	M	13001	Flu
25	M	12057	HIV

TABLE XV PART 2 HOSPITAL DATASET

Age	Sex	Zip code	Disease
26	M	14599	Bronchitis
26	F	13001	Pneumonia
27	F	17000	Hepatitis

2.Vertical Partitioning

In this case, the different sites may have different attributes of the same sets of records.

TABLE XVI PART 1 HOSPITAL DATASET

Age	Sex
25	M
25	M
26	M
26	F
27	F

TABLE XVII PART 2 HOSPITAL DATASET

Zip code	Disease
13001	Flu
12057	HIV
14599	Bronchitis
13001	Pneumonia
17000	Hepatitis

IV. LIMITATIONS OF PRIVACY

Many privacy-preserving data-mining methods are inherently limited by the curse of dimensionality in the presence of public information. For example, the technique in analyses the k-anonymity method in the presence of increasing dimensionality. The curse of dimensionality becomes especially important. when adversaries may have considerable background information, as a result of which the boundary between pseudo-identifiers and sensitive attributes may become blurred. This is generally true, since adversaries may be familiar with the subject of interest and may have greater information about them than what is publicly available. This is also the motivation for techniques such as diversity in which background knowledge can be used to make further privacy attacks. Thus, the data loses its utility for the purpose of data mining algorithms. The broad intuition behind the result in is that when attributes are generalized into wide ranges, the combination of a large number of generalized attributes is so sparsely populated, that even two anonymity becomes increasinglyunlikely.

V. APPLICATION

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope.

VI. CONCLUSION

In this paper, a study of the broad areas of privacy-preserving data mining and the underlying algorithms is done. The broad areas of classification include Privacy-preserving data publishing, Privacy-Preserving Applications, Utility Issues, Distributed Privacy, cryptography and adversarial collaboration are analyzed.

A variety of data modification techniques such as randomization and k-anonymity based techniques has been studied and analyzed based on their activities. A complete study is done on for distributed privacy-preserving mining, and the methods for handling horizontally and vertically partitioned data. issue of downgrading the effectiveness of data mining and data management applications such as association rule mining, classification, and query processing. The limitation of privacy preserving as the increase in the dimension also analyzed and application which can employ the privacy algorithm is alsostudied. In this study paper, a brief overview of privacy preservation in big data publishing is presented. Privacy of an individual should be preserved before big data is published to a third party because big data also consists of user-specific information. Description of the three main privacy models, namely, k-anonymity, l-diversity is given in this study paper. Then algorithms which are used to implement k-anonymity model are also explained.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, SIGMOD Conference, pages 439–450. ACM, 2000.

[2] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy preserving olap.In SIGMOD ’05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 251–262, New York, NY, USA, 2005. ACM.

[3] Jay J. Kim, Jay J. Kim,William E.Winkler, andWilliam E.Winkler. Multiplicative noise for masking continuous data. Technical report, Statistical Research Division, US Bureau of the Census, Washington D.C, 2003.

[4] Mehmood, Abid, et al. "Protection of big data privacy." IEEE access 4 (2016):1821-1834.

[5] Zhang, Xuyun, et al. "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud." IEEE Transactions on Parallel and Distributed Systems 25.2 (2014): 363-373

[6] Zhu, Yan, and Lin Peng. "Study on k-anonymity models of sharing medical information." Service Systems and Service Management, 2007 International Conference on. IEEE, 2007. [5] Russom, Yohannes. Privacy preserving for Big Data Analysis. MS thesis. University of Stavanger, Norway, 2013.

[7] Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker. "Privacy-preserving big data publishing."Proceedings of the 27th International Conference on Scientific and Statistical Database Management. ACM, 2015.

[8] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Mondrian multidimensional k- anonymity." Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, 2006.

[9] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k- anonymity and l-diversity." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[10] Keke Chen and Ling Liu. Privacy preserving data classification with rotation perturbation. In ICDM, pages 589–592. IEEE Computer Society, 2005.

[11] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In ICDM, pages 99–106. IEEE Computer Society, 2003.

[12] Kun Liu, Chris Giannella, and Hillol Kargupta. An attacker’s view of distance preserving maps for privacy preserving data mining. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, PKDD, volume 4213 of Lecture Notes in Computer Science, pages 297–308. Springer, 2006.

[13] IJIRST –International Journal for Innovative Research in Science & Technology| Volume 2 | Issue 11 | April 2016 ISSN (online): 2349-6010