Survey on PPDP using Anonymization

[1]Shah Afreen B.Aziz, [2]Madhuri Kalidas Wavhal, [3]Yogita Ramdas Garje, [4]Jasmine Yusuf Mulla

Students
Modern Education Society's College of Engineering

***Abstract*: Data sets are available for almost all application domains. But cannot be made public due to the sensitivity of data. Due to the emergence of Big Data technology, it has become very important to provide large datasets, provided that the privacy of the data is preserved. Privacy preserving is the methodology of allowing the datasets to be used by various data mining algorithms, maintaining the confidentiality of the sensitive data. This paper will give a brief idea about what is Privacy Preserving Data Publishing (PPDP), an overview of various privacy models and focus mainly upon anonymization techniques. Various attacks that can occur on data and solution to those attacks are also covered in brief.**

Keywords: Privacy, Annonymization, Big Data, PPDP, Attacks

I. Introduction

Big data is term is used for very large data sets that have more varied and complex structure. Data generated from various sources are structured and unstructured. These characteristics usually correlated with additionally difficulties in storing, analyzing and applying further procedure and extracting results. However there is contradiction between data security and privacy of big data. The privacy of big data is very important. Big data analytics has a power to provide insights about the people, product and organization that are far and above what they know about themselves. In fairness this problem existed before big data, but it wasn't a huge risk until big data analytics gave us the tools and techniques to be highly accurate with our predictions.

Data privacy also called information privacy is the aspect of information technology that deals with the ability an organization has to determine what data in computer system can be shared with third parties. The latest research in this area is called Privacy Preserving Data Publishing (PPDP). Generally, the process of PPDP has two phases, data collection and data publish phase. In data collection phase data publisher collect dataset from data owner. Then, in data publish phase, data publisher sends the processed dataset to data recipient.

Currently, there are large number of data publishing models and methods which have been proposed for security of private data. For data privacy, there are various methods, namely k-anonymity, l-diversity, t-closeness. In k-anonymity, we can transform original data into some random identity so that it becomes difficult for the intruder to identify the original identity of an individual. For this, transformation techniques like generalization, suppression and global recoding are used. Generalization is a technique where in, the keywords or the characters in the fields which should be hidden are replaced by an asterisk (*). Once the generalization is done, we need to form groups of data with respective to certain conditions so that none of the data remains critical. For example, we can group our data set based on the age. Age greater than or less than certain limits. The correlating percentage between the data should be greater. The more is the percentage, the more it is difficult to identify the original data. With this it becomes difficult to recognize the dependent values hence, preserving the privacy. The next section covers the privacy attacks on the data, various anonymization models proposed and their drawbacks.

II. Literature Review

1. Privacy Preservation and Data Publishing

The privacy is one of the most critical characteristic of Information security system. Several efforts have been suggested to put forth many techniques for preserving the privacy of sensitive knowledge. Anonymization is one of the method of privacy preservation among several others like data perturbation, data swapping and cryptographic approach. Anonymization is removing iowner dentifiable information from data for privacy protection and using modified data for analysis purpose. The main aim of anonymization is to convert the sensitive data to a form from which owner's identity cannot be extracted or mined in any way.

In PPDP, the structure of data used by a publisher is of the following form:

RT(Explicit_identifier, Quasi_identifier, sensitive_attribute, non_sesnitive_attribute)

where,

Explicit_identifier: set of attributes that identify the owners directly,
Quasi_identifier: set of attributes that can identify the owners indirectly through linkages,
sensitive_attribute: sesnitive information about an owner, like the salary or health disease,
non_sesnitive_attribute: all the attributes not included in any of the above sets fall in this category

Such data format is converted into a different form by the publisher before publishing the data so as to protect privacy of the owners. Here, for the purpose of conversion, various anonymization privacy models are applied. The following will be the format of converted data:

RT'(Quasi_identifier, sensitive_attribute, non_sesnitive_attribute)

Here, Explicit_identifier is removed, Quasi_identifier is anonymised to satisfy privacy and ensure confidentiality.

2. Realted work on anonymization models

2.1 K-Anonymity

One of the way the privacy of owners of data can be violated is through the Record Linkage Attack. In the Record Linkage Attack some value x on quasi identifier Quasi_Id identifies a small number of records in the published table R. For example, using the job, age and sex attributes of a data table and linking it with an external database like voters database may result in identification of the owner information. To solve the problem of Record Linkage Attack, K-Anonymity privacy model is suggested (Samarthi and Sweeney 2001). K-Anonymization technique prevents joining attack by generalizing and suppressing portion of the dataset, so that it becomes nearly impossible to uniquely distinguish an individual from group of K features. In general, if a record of a table has a particular value for quasi-identifier then k-1 other records must also have the same value for the quasi-identifiers.

There are two of of implementing k-anonymity: Generalization and Suppression. In generalization, the categorical attributes are replaced by a general value selected from the heirarchy. For example, student, teacher, housewife,etc can be generalized as person or human whereas dogs, cats, etc can be generalized and replaced with the value animals. In case of numerical attributes, the values are replaced by range of values. Various ways in which generalization can be applied are Full-Domain Generalization(K. LeFevre, D. J. DeWitt, and R. Ramakrishnan 2005), Sibling Generalization (K. LeFevre, D. J. DeWitt, and R. Ramakrishnan 2005),Sub-Tree Generalization (V. S. Iyengar 2002) , Cell Generalization (J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu 2006) and Multi-Dimensional Generalization scheme (K. LeFevre, D. J. DeWitt, and R. Ramakrishnan 2006).

Suppression also anonymizes the values of quasi-identifiers by suppressing the values with one of the following methods: Record suppression (V. S. Iyengar 2002) refers to suppressing, Value suppression (K. Wang, B. C. M. Fung, and P. S. Yu 2007) and Cell suppression (K. Wang, B. C. M. Fung, and P. S. Yu 2007). K-anonymity model was further improvised by (X-Y)-anonymity (K. Wang and B. C. M. Fung 2006) and MultiRelational k-anonymity (M. Ercan Nergiz, C. Clifton, and A. Erhan Nergiz 2007) techniques. Considering two distinct attribute sets x and y, (x-y)-anonymity specifies that every value on the attributes of x is associated to at least k distinct values on attributes of set y. MuliRelational k-anonymity technique works with more than two relational tables together.

Even though k-anonymity solves the problem of linkage attack, it does not deal with the Homogeneity Attack and background knowledge attack. Homogeneity Attack occurs when a group of quasi-identifier has same values of sensitive attributes, then the owner identification becomes easy through linkage. In background knowledge attack, the intruder know some background information about the data owner.

2.2 L-Diversity

Another type of attack on the data that can occur is the Atribute Linkage Attack. In Attribute Linkage Attack, the sensitive values of the owner can be identified by linking the groups of records to which the owner belongs. To solve this problem as well as to overcome the drawbacks of k-anonymity, a privacy model is suggested by Machnavajjhala et al. known as l-diversity (A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam 2006). The principle of l-diversity is that each group of quasi-identifiers must contain atleast l sensitive values of the sensitive attribute set. L-diversity suffers from 2 types of attacks, namely, Similarity attack occurs when all the values of sensitive attribute set semantically have the same meaning or belong to same category and Skewness attack occurs when the intruder can still derive the sensitive information based of the frequency distribution of the values.

2.3 (k,e)-Anonymity

The k-anonymity focuses mainly on categorical attributes. For the purpose of numerical attributes like age and salary, the concept of (k,e)-anonymity was suggested (Q. Zhang, N. Koudas, D. Srivastava, and T. Yu 2007). It works on the principle that the records of the table are partitioned into groups, in which each group has atleast k distinct sensitive values with a range of e. The only drawback of this approach is that it ignores some subrange S. Some of the sensitive attributes might fall continuously in this subrange S, thus giving a possibility of identification of sensitive information by the intruder.

2.4 LKC- Privacy

When the number of attributes in the quasi-identifier set is more, a huge data need to be suppressed or generalized in order to achieve k-anonymity. This results in significant information loss. Such that might not give best result in mining informative knowledge, which violated the property of privacy preservation which says that, even after applying privacy models on the dataset, it must not lead to data loss that need to be used for data mining purpose. To deal with this problem a new anonymization model called LKC- Privacy was proposed (N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. K. Lee 2009). In this model it is assumed that the intruder has background knowledge of atmost L values of quasi-identifier attributes. So LKC-privacy model takes care that any combination of quasi-identifier attributes of length L in the table is shared by at least K records, with the confidence that there are atmost C sensitive values in the record. The values of L,K, and C can be either specified by the owner of the data or the publisher.

2.5 t-Closeness

As seen earlier, l-diversity suffers from skewness attack and similarity attack. To overcome these attacks, t-closeness privacy model has been proposed (Ninghui. Li, T. Li, and S. Venkatasubramanian 2007). A new term "Information gain" is coined by the authors, which is the difference in the belief of the intruder about the distribution of sensitive values in the table probably before and after anonymization. This information gain should not be more than a specified threshold t, and hence the name, t-closeness. It uses Earth Mover Distance (EMD) function to measure the information gain and the specified threshold is achieved by distributing the sensitive values in the same manner in our table's quasi-identifier set as well as the publicly available databases.

2.6 Utility Specification Technique

PPDP based utility specification (Hongwei Tian and Weining Zhang 2013) provides a way to specify certain constraints or requirements that need to be maintained before applying privacy to the data. It helps to meet various requirement of various mining applications that may need some sensitive atrributes as it is. So this model helps to preserve cetain attributes and their values while other attributes are anonymized. A nested ordered list is used to specify the utility requirements. The drawback of this model is maintaing accuracy is a difficult task as the sensitive attributes are not uniformely distributed then it reduces data mining accuracy. Besides that the relationship between quasi-identifiers and sensitive attributes is still present which may lead to breach of privacy.

2.7 Slicing

As the name suggests, slicing (Tiancheng Li, Ninghui Li, Jian Zhang and Molloy 2012) technique partitions the data table horizontally and vertically. In vertical partioning, groups of colums are created, which consists of subset of attributes which are highly associated with each other. Horizontal partitioning is performed with the help of bucketization method, in which groups or buckets are formed based on the quasi-identifier values. Then, permutation is performed on the column values within each bucket. This helps to eliminate the linking between different columns. This leads to preservation of association of attributes within each column, eradicating the associations across columns.

2.8 Multiple Sensitive Bucketization

Multiple Sensitive Bucketization (MSB) (Jianmin han, Fangwei Luo, Jianfeng Lu and Hao Peng 2013) approach is mainly designed for focusing on multiple sensitive attributes rather than single attributes. MSB is suitable for 2-3 attributes only. As the number of sensitive attributes increases, the suppression ratio also increases, which results in significant data loss. To overcome this issue SLOMS (SLicing On Multiple Sensitive) approach is proposed. In this vertical partioning is performed by partitoning the various sensitive attributes into different tables and a single quasi-identifier table. Further the records of each table is partioned into equivalence classes. On each of the sensitive attribute table slicing and bucketization is performed to satisfy l-diversity privacy model. The quasi-identifier table is generalised to meet k-anonymity privacy model. The only drawback of this approach is that, if for further mining tasks the relationship of sensitive attributes matters, then the approach fails as it partitions sensitive attributes into different tables.

3. Comparison among anonymization models

	Attacks resolved	**Data Quality**	**Drawbacks**
k- anonymity	Link Disclosure, Record Tracing	Reduced due to suppression	Homogeneity Attack and background knowledge attack
l-Diversity	Attribute Disclosure	Reduces quality as L increases	Suffers Skewness attack, simmilarity attack
t-closeness	Distribution Disclosure	Provides good quality	identity disclosure.
(k,e)-Anonymity	Link and Attribute Disclosure	Good quality for numeric attributes	Works only for numeric attributes, high information loss
LKC privacy	Information loss	Depends on user specified thresholds l,k,c	Fails if attacker's knowledge is more than threshold
Slicing	Attribute Linkage	Comparative good data quality	Partitioning is responsible for quality

III. Conclusion

Information sharing is becoming an indispensable part of the real world in every field for individuals as well as in organizations, privacy preserving data publishing is getting more and more attention from all over the world, which provides an essential guarantee for information sharing. To put it simply, the role of the PPDP is to transform the original information or dataset from one form or state to the other form or state so as to avoid privacy disclosure and withstand diverse attacks.

This paper presents a survey for most of the common attacks techniques for anonymization-based PPDP and explains their effects on Data Privacy. k-anonymity is used for security of respondents identity and decreases linking attack in the case of homogeneity attack a simple k-anonymity model fails and we need a concept which prevents from this attack and solution for this is l-diversity. All tuples are arranged in well-represented form and adversary will divert to l places.

l-diversity limits in case of background knowledge attack because no one predicts knowledge level of an adversary. It is observed that using generalization and suppression we also apply these techniques on those attributes which don't need this extent of privacy and this leads to reducing the precision of publishing table.

Generalization with suppression is also the causes of data lose because suppression doesn't release values which are not suited for k-factor. Future works in this field can include defining a new privacy measure along with l-diversity for multiple sensitive attributes and we will focus on other techniques which are used to achieve k-anonymity without suppression because suppression leads to reduce the precision of publishing table.