ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

Survey on Privacy Preserving on Data Publishing and utility verification

[1] Manasi Dhage [2] Dr. A. N. Banubakode
[1][2] Department of Computer Engineering
Rajarshi Shahu College of Engineering
Pune, India

Abstract :-- **In recent, data sharing is an important task. This data publication is conducted by various organizations under some important rules and regulations. Such data is useful in various researches. But such original data contain sensitive or private information of data owners, and during such data publication privacy of individuals might be reveal to outsiders. Because of this, the need of privacy preserving arises, which is used to hide the individual's information. Sometimes the collaborative data publishing to multiple users is attacked by outsider or insider. To overcome the problems like, privacy, security, and data integrity during data publication, there is a need of privacy preserving and utility verified data publications approaches. This can by achieve by some data mining techniques. With this, there is another problem of data utility verification of published data as it require raw data, which is not allowed to reveal to users because of some privacy issues. This paper discuss the various issues of collaborative data publication and make study of some recent techniques based on privacy preserving data publication and utility verification with annonimization. Also discuss their respected advantages and limitations.**

Key Words: **-- Privacy preserving, annomization, data mining, information security.**

I. INTRODUCTION

The demand for gathering and sharing information is increment strongly because of the quick development of data. An large amount of information is utilized for investigation, insights and calculation to discover general pattern or guideline which is advantageous to social improvement and human advancement. In the interim, dangers emerge when enormous information accessible for the general population. For instance, individuals can burrow protection data by getting together sheltered appearing information, subsequently, there is an extraordinary plausibility uncovering people security. As per the study, roughly 87 % of the number of inhabitants in the United States can be uniquely distinguished by given dataset distributed for people in general. To stay away from this circumstance deteriorating, measures are taken by security division of numerous nations, for instance, declaring protection direction. The prerequisite for information distributer is that information to be published must fit for the predefined conditions. Distinguishing credit should be precluded from distributed dataset to ensure that individual's security can't be construed from dataset straightforwardly. Expelling identifier trait is only the readiness work of information handling, a few cleansing operations should be done further. Though, after information preparing, it might be diminish information utility significantly, while, information security did not get completely protected.

In face of the challenging risk, some researches havebeen proposed as a remedy of this uncomfortable circumstance, which focus at achieving the equalization of information utility and data security when publishing dataset. The continuous examination is called Privacy Preserving Data Publishing (PPDP). In the previous couple of years, specialists have responded to the call and undertaken a lot of examines. Numerous attainable methodologies are proposed for various security saving situation, which illuminate the issues in PPDP successfully. New strategies and hypothesis turn out consistently in experts' effort to complete privacy preserving.

Privacy Preserving Data Publishing:

Usually, the procedure of Privacy Preserving DataPublishing has two stages, information gathering and information distribute stage. It alludes to three sorts of parts in the process which are information proprietor, information distributer and information beneficiary. In

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

the information gathering stage, information distributer gathers dataset from information proprietor. At that point, in the information distributed stage, information distributer sends the prepared dataset to information beneficiary. It is important to say that raw dataset from information proprietor can't be specifically sent to information beneficiary. The dataset ought to be prepared by information distributer before being sent to information beneficiary.

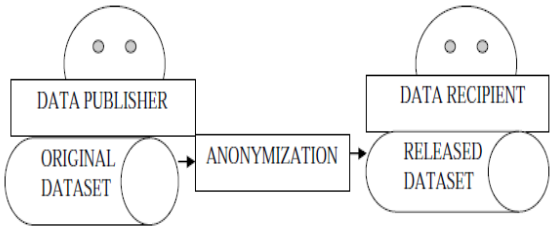

Fig 1: Privacy preserving data publishing

In this paper further we will see: Section II talks about related work studied till now on topic. Section III discuss existing system. Section IV describes proposed system and this paper is concluded in section V.

II. LITERATURE SURVEY

In this section discuss the existing method developed for privacy preserving of data. Now we discuss different methods developed by the researchers, the different methods are as follows:

K-Anonymity

The various phenomena arise when analyzing and publishing the data in high-dimensional space. K-Anonymity was a technique to hold up the blight. Speculation on K-Anonymity was connected to cover the careful estimation of an attribute [2]. The annoyance strategy on K-Anonymity was reasonable for total circulation of an individual than the inter-attribute relation of an individual. 2-anonymity and Gaussian cluster strategies proposed on K-Anonymity strategy, guarantee protection by assessing likelihood and assigning its value to zero. As per the authors view, this method tried to understand the probability distribution which would have maximum likelihood of its attributes. As per the authors, there would be a loss for high-dimensional data.

ℓ-Diversity

Data around an individual couldn't be distributed without uncovering delicate attribute [3].

K-Anonymization was insufficient to secure the information which incorporate homogeneity attack and background learning attack. ℓ-Diversity method portrayed that sensitive attribute would have at most same recurrence. For instance, with positive disclosure, if Alice needs to find Bob, Alice would decide Bob with high-likelihood appropriation. The negative exposure would happen when an adversary could accurately dispense with some conceivable estimation of the delicate traits. There could be a minimum distinction between the earlier conviction and back conviction.

T-Closeness

Anil Prakash, RavindarMogili found that K-Anonymity and ℓ - Diversity was not used to avert quality revelation [4]. ℓ-Diversity would have very much represented sensitive characteristic esteem that was allotted just with certain number of restrictions. t-closeness has been proposed to depict the appropriation of sensitive characteristic with comparability class. Earth Mover Distance was used to gauge the separation between the two probabilistic distributions. Conjunction has been proposed to consolidate machine learning and factual investigation. Closeness among the segments was diminished by aggregation.

Km Anonymity

Km Anonymity has been proposed for an anonymize value-based database [5]. Km Anonymity go for ensure the database against an enemy who knows about m items in the transaction. The speculation was utilized to keep up the set esteemed information. For any exchange on K-1 records, other indistinguishable transaction would likewise show up. Km secrecy has been presented by means of top down nearby speculation procedure to record the quantity of exchange records. The segment based methodology was utilized to gathering (parcel) the comparative items in a top-down way. The km secrecy model would keep protection breaks raised from an enemy who might discovered m things in an transaction database.

Distributed K-Anonymity framework (DKA)

The gathering of information from various locales can't be shared specifically. The key step was to anonymize

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

the data in order to generalize a specific value [6]. A protected 2-party structure was intended for multiparty calculation that has been utilized to join the dataset from various sites. Appropriated K-Anonymity prevent recognizable proof of an individual by make utilization of worldwide Anonymization in the encrypted form. DKA give a protected structure between two parties. Two parties would agree on Global Anonymization algorithm that could produce local Anonymous dataset. Additionally, DKA give a protected dispersed protocol which would require that two parties could commonly semi-honest. Still the exchange off amongst utility and capability of information was misused in DKA.

K-Anonymity Clustering

Among various clustering methods, hierarchal clustering was mostly used to achieve KAnonymity. Weighted Feature C-means Clustering [WFC] utilized to diminish the data deformation. WFC segment all records into proportionality class and would combine the class utilizing class merging mechanism [7]. The numerical values of quasi identifier were used to evaluate the Weighted Feature C-means Clustering technique. The authors also try to provide the dissimilarity evaluated approach which would take different types of feature values for class merging mechanism.

Paper Name	Proposed work	Advantages	Disadvantage
Privacy-Preserving Utility Verification of the Data Published by Non-interactive Differentially Private Mechanisms	Propose a privacy-preserving utility verification mechanism based upon cryptographic technique for DiffPart-a differentially private scheme designed for set-valued data.	Improve the security and efficiency of the system	Association rule mining over huge data may increase the execution time.
On k-Anonymity and the Curse of Dimensionality	View the k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes.	the curse of high dimensionality also applies to the problem of privacy preserving data mining.	loss for high-dimensional data
ℓ-Diversity : Privacy Beyond K-Anonymity	show with two simple attacks that a k-anonymized dataset has some subtle, but severe privacy problems.	ℓ-diversity is practical and can be implemented efficiently	Limited till this work

III. EXISTING SYSTEM

A privacy-preserving utility verification mechanism for DiffPart, a differentially private anonymization algorithm designed for set-valued data. DiffPartperturbs the frequencies of the records based on a context-free taxonomy tree and no items in the original data are generalized. This proposal solves the challenge to verify the utility of the published data based on the encrypted frequencies of the original data records instead of their plain values. As a result, it can protect the original data from the verifying parties (i.e., the data users) because they cannot learn whether or how many times a specific record appears in the raw dataset without knowing its real frequency.

Different from DiffPart, DiffGenmay generalize the attribute values before perturbing the frequency of each record. Information losses are caused by both the generalization and the perturbation. These two kinds of information losses are measured separately by distinct utility metrics. The analysis shows that the utility verification for generalization operations can be carried out with only the published data. As a result, this verification does not need any protection.

IV. PROPOSED SYSTEM

Horizontal Partitioning: To increase the processing time of frequent patterns generation, we divide the process horizontally. A partition is a division of a logical database or its constituent elements into distinct independent parts. Database partitioning is normally done for manageability, performance or availability reasons, as for load balancing. Horizontal partitioning involves putting different rows into different tables.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

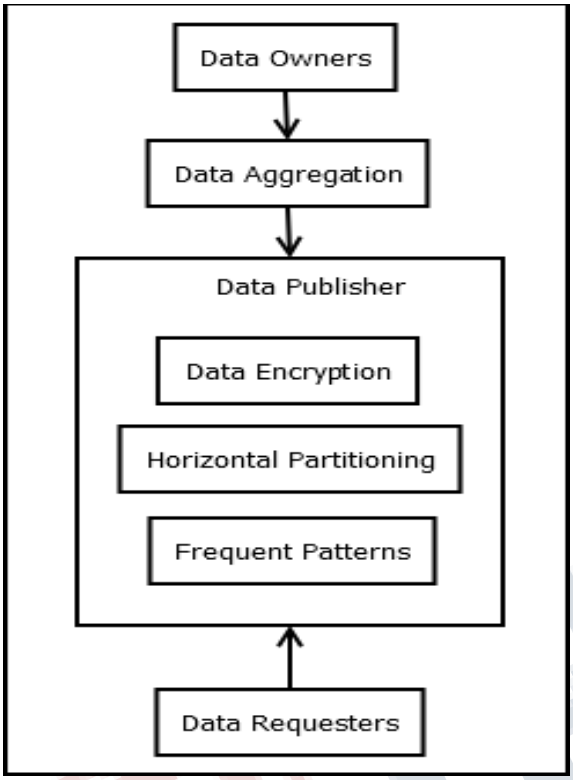

Fig 2: Proposed system architecture

V. CONCLUSION

In this survey, we analyze several recent anonymization techniques to maintain the privacy in collaborative data publishing with various data mining techniques. Due to the huge amount of information, it is necessary to preserve the Privacy. Number of anonymization techniques are developed in recent but still has some limitations which are also discussed here to improve the privacy and utility verification of data publishing.

REFERENCES

1) Jingyu Hua, An Tang, Yixin Fang, Zhenyu Shen, and Sheng Zhong, ``Privacy-Preserving Utility Verification of the Data Published by Non-interactive Differentially Private Mechanisms", IEEE Transactions on Information Forensic and Security.

2) Charu C. Aggarwal, (2005), ''On k-Anonymity and the Curse of Dimensionality", Proceedingsof the 31st VLDB Conference, Trondheim, Norway, pp.901-909

3) AshwinMachanavajjhala, Daniel Kifer,JohannesGehrke, MuthuramakrishnanVenkitaSubramanian, (2006),"ℓ-Diversity : Privacy Beyond K-Anonymity", Proc.Internationalconference on Data Engineering.(ICDE),pp.24.

4) Anil Prakash, RavindarMogili ,(2012),''Privacy Preservation Measure using t-closeness withcombined l-diversity and k-anonymity", International Journal of Advanced Research in ComputerScience and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33

5) Yeye He, Jeffery Naughton .F, (2009), "Anonymization of Set Valued Data via Top DownLocal Generalization", Proc. International Conference on Very Large Databases (VLDB), pp.934-945.

6) Wei Jiang, Chris Clifton, (2006), "A secure distributed framework for achieving kanonymity",the VLDB Journal, Vol.15, No.4, pp.316-333.

7) Chuang-Cheng Chiu, ChiehYuan Tsai, (2007),"A k Anonymity Clustering method forEffective Data Privacy Preservation", Springer journal on Verlag Berlin Heidelberg, pp.88-99.