ISSN:0975-9646

A Novel Approach to Preserve the Privacy of Data

Gayathri T.[1], A. Viji Amutha Mary[2]

M.E Computer Science and Engineering[1], Faculty of Computer Science and Engineering[2],
Sathyabama University
Chennai-600119, India

***ABSTRACT*-A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to conserve privacy before data are published. The data owner before publishing the data first perform changes to the data into n number of copies based on the access privilege and publishes. In that the high privilege data contains a smaller amount noise and low privilege data contains more additional noise. But the formal attackers have possibility to make diversity attack to rebuild the data using non-linear techniques. To overcome the problem, a novel approach has been introduced where the data owner after performing modification finds the noise of every copy and compare. If the noise is similar means there is no privacy in the modified data and the attacker have possibility to reconstruct the information. So the data owner adds further noise until there is no similarity.**

Keywords: Access privilege, Non-linear Techniques, Random perturbation

I. INTRODUCTION

We live in a networked environment which is experiencing rapid growth in the amount of person-specific data available. Some Organizations need to make their data available to public which may be useful to others for research purposes. For example, Patients medical records in a hospital may be used for data analyzers to build a classification model using patients' epoch, smoking habits and obesity condition to predict their life time or to study the characteristics of various diseases. On the other hand, data publishers are often prohibited by law from revealing any person-specific information that compromises an individual's privacy.

Mining data sets that include information about people in a population is a great way of intellect about properties of that population. Applications include observing the effects of treatments on disease, dealing with disease outbreaks. Aside from this useful information, such data sets also include sensitive information like the disease of an individual, the salary, etc. for the reason that of this, the goal of privacy-preserving data publishing is to make the most of the "good utility" while limiting the ability of an opponent to identify specific individuals and learn their sensitive information from the data set.

II. RELATED ARTICLES

Xiaokui Xiao et al. (2006) proposed a new generalization framework based on the concept of personalized anonymity [1]. This technique performs the minimum generalization for satisfying everybody's requirements, and thus, gathers the major amount of data from the micro data and results in generalized tables that permit accurate aggregate analysis. Personalized anonymity represents a new generalization framework. Personalized anonymity carries out a careful theoretical study that leads to valuable insight into the behavior of alternative solutions.

The Generalization framework is implemented by making use of quasi-identifier attributes and K-anonymity. In particular, a table is k-anonymous if the QI values of each tuple are identical to those of at least k −1 other tuples. K-anonymity cannot guarantee privacy protection if an individual may correspond to multiple tuples in the micro data. For instance, the greedy algorithm presented in the paper is not optimal, that is it does not necessarily achieve the lowest information loss.

Ada Wai-Chee Fu et al. (2005) proposed the concept regarding the mining of frequent patterns [2] (item sets) where data is resided in multiple sites. In the mining of frequent patterns privacy of individual parties will not be exposed when data mining techniques are applied to a large collection of data about the parties. The mining of frequent patterns has significance not only in itself but also for other data mining tasks such as mining of association rules for the data, correlations, sequences, classifiers and clusters for the related data.

The mining of frequent patterns addresses the problem of privacy-preserving frequent pattern mining in a schema across two dimension sites. Here it is considered that sites are not trusted and they are semi-honest. The solution to this concept is based on semi-join and also it does not consider any thing about data encryption. By using a star schema it is possible to make use of semi-join for privacy preservation.

Kun Liu et al. (2006), explores the possibility of using multiplicative random projection matrices [3] for privacy preserving distributed data mining. Here the problem is directly related to many other data-mining problems such as clustering the similar data, principal component analysis, and classification of data.

Privacy preserving distributed data mining proves that, after perturbation, the distance-related statistical properties of the original data are still well maintained without exposing the dimensionality and the exact data values. The random projection-based technique may be even more powerful when used with some other geometric transformation techniques like scaling, translation, and rotation. Combining this with SMC-based techniques offers another interesting direction.

Daniel Kifer et al. (2006), proposed Log-linear models and logistic regression models [4], which are the popular techniques for analyzing tabular data. They provide a compact and interpretable representation of high dimensional probability distributions. The K-Anonymity and l-diversity are weaker privacy definitions they do not protect against adversaries with arbitrary amounts of Background knowledge but they provide considerably more utility.

Bernardo A. Huberman et al. (1999), introduced a number of new techniques for finding members of groups sharing similar preferences [5],while also allowing reputations to be built and updated. In this a count value is used, for all the members who are having the similar information simply the count value will be incremented which enables outsiders difficult to guess the specific person. Another application designed consists of removal of disincentive associated with issuing recommendations.

Xiaokui Xiao et al. (2009), proposed the Random perturbation [6], which is a popular method of producing anonymized data for privacy preserving data mining. It is simple provides with the strong privacy protection, and allows effective Mining of a large variety of data patterns. Random perturbation mainly focuses on a detailed random perturbation procedure, which refers to as uniform perturbation.

Benjamin C. M. Fung et al.(2008), proposed an efficient anonymization algorithm to thwart the attacks in the model of continuous data publishing [7]. In practical applications, data is published continuously as new data turn up, the same data may be changed differently for a different purpose or a different recipient. In such cases, even when all new released data are properly *k*-anonymized, the vagueness of an entity may be inadvertently compromised if recipient cross-examines all the releases received or colludes with other recipients. In the paper, we systematically characterize the correspondence attacks. Here each release contains the new data as well as previously collected data. Finally, proved that detection and the anonymization methods are extendable to deal with multiple releases and other privacy requirements.

III. SUMMARY OF EXISTING SYSTEM

Consider the case of mining healthcare data for detection of bioterrorism which may require studying the clinical records and pharmacy transactions data of certain off-the-shelf drugs. However, grouping such dissimilar data sets belonging to different parties may violate the privacy laws. Although health organizations are permitted to release information as long as the identifiers like name, SSN, address, etc., are not involved, it is not considered safe enough since reconstruction may be possible for linking different public data sets to categorize the original subjects. This requires a well designed technique that pays cautious concentration to hide privacy-sensitive information, while hiding the intrinsic statistical dependencies which are important for data mining applications.

The feasible drawback of adding noise makes one be unsure about the chances of using multiplicative noise for protecting the privacy of the data. Mainly there are two basic forms of multiplicative noise. One is to multiply each data aspect by an arbitrary number that has a reduced Gaussian distribution with mean one and tiny variance. The further one is to take a logarithmic conversion of the data first, add predefined Multivariate Gaussian noise, and take the antilog of the noised data. In practice, the formal method is good if the data owner only wants to make little changes to the original data; the other method assures privileged security than the first one but contains the data utility in the log scale, which is difficult to assess.

A possible problem of conventional additive and multiplicative perturbation is that each data element is perturbed autonomously; therefore the pair-wise comparison of records is not guaranteed to be maintained. Additive and multiplicative perturbation usually deals with numeric data only.

In the existing system, the data owner initially converts the data into n number of copies by adding some amount of noise to the original data. Then the data owner finds noise of every copy using every linear technique. Then compares one noise to other noise, and finds if the noise is similar the data owner adds additional noise to that data, and then publishes. The issue of existing system is, here the author proposed the solution only to overcome the diversity attacks which are performed using linear techniques. There may be chances that authorized user can make diversity attacks using non-linear techniques. The privacy of every data is less in this case.

IV. PROPOSED SYSTEM

In the proposed system constructed a network is constructed for preserving the privacy of individual data in case of Hospital Patient records, which is highly useful for the researches in performing some calculations using patients' epoch, smoking habits and obesity condition to predict their life time or to study the characteristics of various diseases. At the same time the privacy of individual also we need to be protected as nobody wants to reveal their Diseases history to the public. So here we had done this by using partial hiding of data by adding some amount of Noise to the original

To overcome the problem in the existing system where the authorized user can make diversity attacks using non-linear techniques we have trained our system itself with the non-linear techniques and by using sequential Generation algorithm. So by this the authorized attackers cannot perform diversity attacks in any way. This technique before publishing the data to the server find the noise level between the existing (available) and publishing (new data) data.

If it is similar that is if it is within the prescribed threshold value which makes it possible to guess for the outsider then the server requests the data owner to add some additional noise to avoid the similarity. Then again after adding the additional noise also it checks with the existing noise level and finds similarity; it will continue this until there are no similarities between the original and publishing data. By using this privacy of every data is increased. Based on this approach the data owner prevents noise similarity. Now it is not possible for the data consumer to make any diversity attacks using either linear or non-linear techniques.

Below figure1 shows the functional architecture of how the privacy preserving can be accomplished by making use of non-linear techniques in the proposed system.

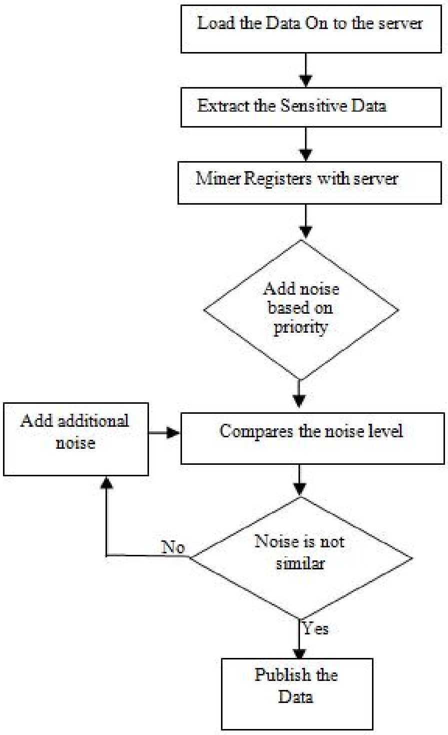

Load the Data On to the server
Extract the Sensitive Data
Miner Registers with server
Add noise based on priority
Add additional noise
Compares the noise level
No
Noise is not similar
Yes
Publish the Data

Fig 1: Functional Architecture

V. Implementation Details

The implementation part contains the detailed informatioı regarding how the network for preserving the privacy of dat: is developed. The developed detailed system frameworl contains a server, the Users who may be either a data owne who uploads the data or may be a consumer who makes us of the published data.

A. Network deployment and user authentication

This phase contains details about network deployment an user authentication. This network contains three types o users. One is researchers, second is doctors and third is nurse The controller of all users is server. Then the data owner an consumer make registration with server and ge authentication certificates.

B. Database train system

In this phase Data owner uses nonlinear techniques to avoid diversity attacks. The data owner generates noise using sequential generation algorithm. Add generated noise in every level input data. Input the noise data to non-linear techniques to find the similarity level. If the noise is similar

-Adds additional Noise which is generated using sequential generation algorithm.

-If noise is not similar data owner publishes the data.

C. Modify new data Copy

If the noise is similar then the data owner generates noise using sequential generation algorithm. Then add the generated noise in every level input data. Input the noisy data to nonlinear techniques. This technique finds the noise and predicts the similarity. Based on the similarity the data owner adds additional noise to prevent the similarity. Again inputs the data and find noise. If the noise is not similar then the data owner uploads modified data.

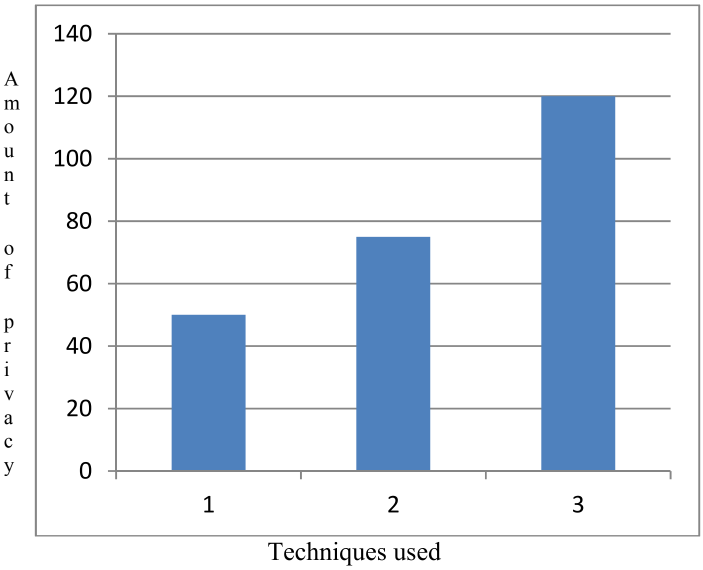
D. Process of user and server.

In this phase the user sends request for particular data to server. Then the server verifies the details and sent the required file to user based on priority level. Here the user who is may be researcher or doctor or nurse receives the data

VI. Performance Evalution

The proposed system contains the detailed information regarding how the network for preserving the privacy of data is developed. Here we train the system to overcome from both linear and non-linear attacks, where as the existing systems deals with only linear attacks by making use of additive and multiplicative perturbation. So by this we have increased the individual data privacy more than in the existing system. Now after checking with the existing noise level until there is no similarity we can increase the privacy to greater extent.

The following graph shows the comparison among the methods mentioned above in the existing and proposed systems.



From the graph it is shown that after training the database with non-linear techniques can obtain a better trade-off for accomplishing a better privacy on the sensitive data.

1-Additive perturbation

2-Multiplicative perturbation

3-Non Linear techniques

CONCLUSION

The key challenge lies in preventing the data miners from combining copies at different trust levels to jointly reconstruct the original data more accurate than what is allowed by the data owner. We address this dispute by properly comparing the noise across copies at different trust levels. Our solution provides maximum prevention from reconstructing noise copy to original copy. In the future, we have planned to improve the security of data transmission.

ACKNOWLEDGMENT

I would like to thank Dr.B.Bharathi, Head of the Department, Department of Computer Science and Engineering and Ms.A.Viji Amutha Mary for her encouragement and support.

REFERENCES

[1]. X. Xiao and Y. Tao, “Personalized Privacy Preservation,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2006.

[2]. A.W.-C. Fu, R.C.-W. Wong, and K. Wang, “Privacy-Preserving Frequent Pattern Mining across Private Databases,” Proc. IEEE Fifth Int’l Conf. Data Mining, 2005.

[3]. K. Liu, H. Kargupta, and J. Ryan, “Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 92-106, Jan. 2006.

[4]. D. Kifer and J.E. Gehrke, “Injecting Utility Into Anonymized Datasets,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2006.

[5]. B.A. Huberman, M. Franklin, and T. Hogg, “Enhancing Privacy and Trust in Electronic Communities,” Proc. First ACM Conf. Electronic Commerce, pp. 78-86, Nov. 1999

[6]. X. Xiao, Y. Tao, and M. Chen, “Optimal Random Perturbation at Multiple Privacy Levels,” Proc. Int’l Conf. Very Large Data Bases, 2009.

[7]. B. Fung, K. Wang, A. Fu, and J. Pei, “Anonymity for Continuous Data Publishing,” Proc. Int’l Conf. Extending Database Technology(EDBT), 2008.