A Study of Correlation Impact on Privacy Preserving Data Mining

J. Hyma
GITAM University
Department of CSE,GIT

P.V.G.D. Prasad Reddy,
PhD
Andhra University
Department of CS & SE

A. Damodaram, PhD
JNTUH
Department of CSE

ABSTRACT
Data sharing is obvious in present day scenario of digital world, and when data is being shared among various application areas the sensitive data of the individuals is disclosed to the public. An evident awareness about this privacy violation has been created among the people now when compared to the earlier days and they are also showing a real concern towards their privacy in the technology enabled digital world. At one end several studies have been proved that privacy is a primary concern and also suggesting not to disclose too much of individual information, but at the other end people are disclosing their personal information knowingly or unknowingly through online surveys, social networks, online shopping sites, e-commerce, government agencies etc. This information sharing is obvious and it can't be unavoidable. Consequently several techniques have been proposed to protect privacy of the individual disclosed information, but still there is an immense need of new privacy preserving techniques that can equally accommodate with the proportionate expansion of the digital data. Existing privacy techniques applied on the data set assuming all the records are independently sampled, where as in the real world data set the correlations among the records is obvious and needs to be studied to achieve accurate privacy protection. This paper provides an overview of the development of privacy preserving models and the further enhancements to be carried out to accommodate with the diverse privacy requirements and data utilization along with the correlation study.

Keywords
Data Mining, Privacy Preserving Data Mining (PPDM), Correlation, Correlation constraints.

1. INTRODUCTION
In today's era of digital world attaining data and its storage is very simple. According to the web analytics, the volume of data being collected is approximately 2.5 Exabyte per day. Various data collection strategies being implemented to gather the data that is through online surveys, social networking sites, email, online shopping, government agencies, hospitals etc. This data being collected is the main unit of all statistical studies. Various researches being conducted in different fields, but in all these 'data' is the prerequisite, which is analyzed and interpreted to get useful information. Data Mining is one such process contains set of automated techniques used to extract hidden patterns from large collection of databases with the help of the modern computing devices. Applying this Data mining process in Marketing/Retailing, Banking/Crediting, Law enforcement, Researchers, Transportation, Medicine, and Insurance etc. improves effectiveness and decreases price [1]. Increasing data collection, processing and excessive data mining raises serious concern towards individual privacy violation [2]. There is always a trade-off between data utility and privacy preservation. The more we want the data utility, the lesser the privacy and vice versa. Privacy preserving data mining [3] is all about to retain the privacy with the preservation of the data mining. More simply, privacy is provided with the preservation of the data usefulness in data mining.

The basic review is done to get into the knowledge of privacy preserving data mining [3, 4, 5]. As a part of this study various techniques have been identified and studied that could be used to preserve the user's data sensitivity. The techniques namely Anonymization [6, 7, 8], Perturbation [9], Blocking based approach [10], Cryptographic techniques [11], Condensation approach [12, 13] have been studied. Each of these techniques had come up with their own advantages and disadvantages. Further researches are going on by adopting the basic idea behind these techniques and are also getting improvised day to day, so that they can easily fit into the current digital world scenario. In this direction the following study mainly focuses on impact of correlations on privacy protection. At present any real world data set are containing correlations inherently by nature. Ignoring these correlations among the data will not give accurate privacy protection. So there is an immense need of developing an efficient PPDM technique that could possibly gives enough of privacy levels with the proper attention towards correlations. This paper gives a theoretical approach to get an insight of correlation study in privacy protection.

The rest of the paper is organized as follows. In section 2 the existing Privacy Preserving Data Model is explained and the new Correlated Privacy Preserving Data Model with set of correlation constraints is proposed is section 3. Section 4 concludes the study and also future enhancements are specified.

2. PRIVACY PRESERVING DATA MODEL
A typical scenario of privacy preserving date model is shown in figure [1]. The data collectors are those who will collect the data from the record or data owners. These record owners may be an individual or an organization.

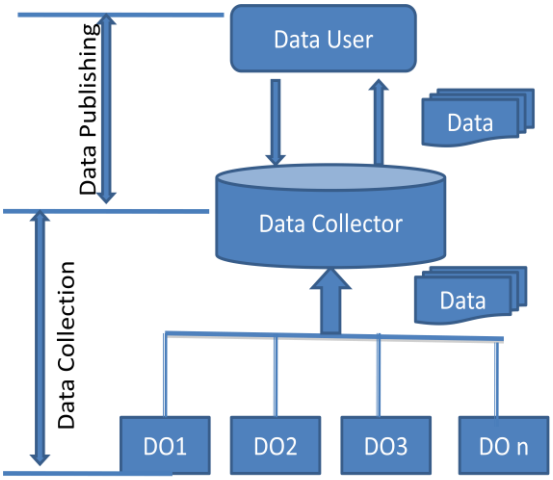


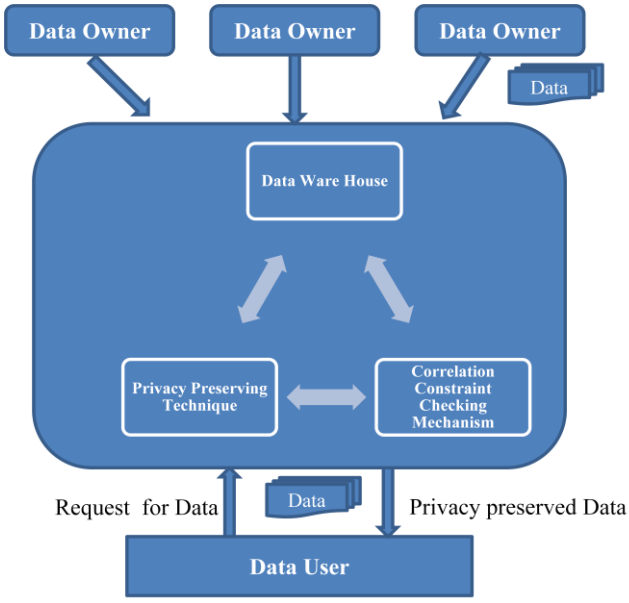
Figure 1. Privacy Preserving Data Model

A proper data preprocessing at the stage of data collecting is an implicit procedure that has to be carried out to make the data ready for further processing. In the next stage the data collector who is also known as data publisher will publish the required data to the data user or the data miner. There are set of protocols [14] proposed to control the disclosure among the entities at each layer. Data collection protocols [14] protect privacy during the data transmission from the data owner's layer to the data collection layer. Inference control protocols protect privacy during the data transmission from the data collector layer to the data mining or the data user layer, who will then, conducts the data mining. There is an obvious need to share the data globally among different organizations. Information sharing protocols protects privacy during the data transformation from one data user to another data user.

For instance, patients will provide the data to the hospitals for their treatment purpose and the hospital may release the data to any eternal medical center or to the researchers for any statistical survey purpose. Here the patients are the record owners and hospitals are the data collectors and researchers are the data miners or the data users. Every privacy preserving data model will have its own assumptions on data owners and data publishers and data users. These models will concentrate to protect privacy during data release, but now in present days with the vast availability of technology enabled devices data release is happening every second in the internet world. Hence we need much more efficient techniques to be implemented to process all these highly scalable data in a private manner.

3. IMPACT OF CORELATIONS ON PRIVACY

The existing privacy preserving data model will use various techniques as mentioned earlier to protect privacy. In these methods the data requested by the researcher is first modified using one of the privacy preserving technique in such a way that it protects the privacy and also preserves the data utility, but it will not consider the correlations among the data records. Existing correlations in the real time data is one of the important aspect has to be considered in PPDM phase. Studies have been proven that ignoring correlations and directly applying PPDM techniques may lead to privacy violation.



Figure[2]. Correlated Privacy Preserving Data Model

In this study various correlation constraints are identified that needs to be considered along with the privacy preserving processing techniques.

3.1 CORRELATION CONSTRAINTS

3.1.1 What is correlation

Correlation measures the strength of association between two variables. That means, consider any two random samples then "Correlation" studies the coupled behavior associated with each other. The correlation coefficient value ranges from -1 to 1. If the value is positive then the variables are positively correlated and if it is negative then the variables are negatively correlated and if it is 0 there is no correlation. These correlations have to be handled carefully to acquire the desired privacy level. When we impose these correlations constraints on the data, it determines a new privacy level. We have defined various correlation constraints, varies from simple correlation constraints to stronger correlation constraints. These are narrated with the synthetic data set given in Table 1 and nature of attributes in Table 2 and a 2-anonymous data in Table 3. In Table 3 to get the anonymity the disease HIV is generalized to Infectious Disease leaving the Drug attribute as it is and the inherent correlation between Disease and Drug attributes would violate privacy protection. That means if the adversary have the knowledge of drugs used for diseases then he can directly predict the patient's disease with a high probability. Another correlation study we can make is the employees staying in the same area having age range of 40 suffering with TB. Here the correlation among the records is helping adversary to extract additional knowledge beyond the privacy protection. This violation would be large in large scale data bases. The following correlation constraints are used to analyze the existing correlations and to assign privacy levels accordingly.

Table 1: Patient Data Set

PID	NAME	AGE	ZCODE	OCCUPATION	DISEASE	DRUG
1	A	40	530011	Employee	TB	Rifampin

2	B	28	530012	Self Employed	Fever	Paracetamol
3	C	25	530012	Student	HIV	antiretroviral
4	D	24	530014	Student	HIV	antiretroviral
5	E	40	530011	Employee	HIV	antiretroviral
6	F	29	520011	Employee	Fever	Paracetamol
7	G	45	520010	Retired Employee	TB	Rifampin

Table 2: Attribute nature

ATTRIBUTE	NATURE
Patient ID (PID), Name	Personally Identifiable Data (PID)
Age, Zipcode, Occupation	Quasi Identifiers
Disease	Sensitive Attribute
Drug	Normal

Table 3: 2-Anonymous Table

AGE	ZCODE	OCCUPATION	DISEASE	DRUG
4*	5300**	Any Employee	TB	Rifampin
4*	5300**	Any Employee	TB	Rifampin
2*	5300**	Any Employee	Fever	Paracetamol
2*	5200**	Any Employee	Fever	Paracetamol
2*	5300**	Any	Infectious Disease	Antiretroviral
2*	5300**	Any	Infectious Disease	Antiretroviral
**	5300**	Any	Infectious Disease	Antiretroviral

3.1.2 Simple correlation constraints

These constraints are useful to assign the privacy level to a relation or attribute or to a record depending upon the existing correlated data. The entire dataset can be analyzed to identify the correlations among the records and attributes. The data owner can fix a threshold level to the correlated data, and accordingly he can assign the privacy level.

3.1.3 Value-Based correlation constraints

These constraints are specified according to the context or value present in the data. Suppose if there is a patient suffering with Flu, and his immediate relatives' information existing in the data, and the data releasing of the patient suffering with Flu, helps the adversary to predict the health condition of his relatives with some probability. So in this example, the value of the disease attribute gives the impact on the correlation among the people. When we consider a normal fever and Flu values in the disease attribute they should result in different privacy levels with their correlations.

3.1.4 Attribute-Based correlation constraints

These constraints are helpful to obtain the privacy level when the attributes are taken together. That means it gives the association or correlations between the attributes. For example, if we take drug and disease together, which has internal dependency and revealing one record with both the values, and another record with only the drug value, allows the adversary to obtain the disease information of the second record also.

3.1.5 Event Based correlation constraints

We can represent an event as a query occurrence. In a Query interaction model the events can be correlated with each other. For example the event is the number of HIV patients in a particular hospital is released then the next event number of male and female count of HIV patients should come with new privacy level. Each successive events may be correlated and need to be analyzed carefully to protect the user data.

3.1.6 Personalized Vs Universal correlation constraints

In the existing PPDM Model, the original record owners are completely isolated from the privacy preservation mechanism. Once the data is given to the data publisher assuming him trustworthy, the data owner is no where concerned with the further processing, and even he does not know where and with whom the data is and how it is being used. Then the data publisher applies a universal privacy requirement to achieve privacy. At one end the privacy is an individual choice of disclosure, the record owner's heterogeneous choice must be considered rather than the universal approach. At the other end as our data is correlated relying completely on the owner's choice also not advisable. Hence it is proposed that considering either of these approaches alone will lead to privacy violation. So we need to get a balanced mechanism to impose this constraint.

4. CONCLUSION AND FUTURE WORK

In this paper a new correlated based privacy preserving data mining model is proposed. The study has shown the impact of correlations on privacy violations. Various types of correlations that have to be considered are explained with real time examples. A strong conclusion has been made that the privacy level that imposed on the data with correlations and without correlations is not yet same and it also needs an accurate correlation constraint mechanism to protect the privacy. A theoretical approach of correlation study has been proposed and as a future work correlation constraint mechanism with the right amount of privacy will be implemented and compared with the existing methodologies.

5. REFERENCES

[1] Bharat Bhushan Agarwal and Sumit Prakash Tayal, "Data Mining and Data Warehousing", Laxmi Publications Ltd, 2009.

[2] LjiljanaBrankovic and Vladimir Estivill-Castro, "Privacy Issues In Knowledge Discovery and Data Mining", In Proceedings of Australian Institute of Computer Ethics Conference, Melbourne, Victoria, Australia, July 1999

[3] Agarwal, R, and Srikanth ,R, "Privacy Preserving Data Mining" Proceeding of Special Interest Group on Management of Data, pp 439-450, 2000

[4] Jian Wang, Yongcheng Lou, Yen Zhao, Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Applications, pp.111-114, 2009.

[5] Anita A. Parmar, Udai Pratap Rao, "Blocking Based approach for Classification Rule Hiding to Preserve the Privacy in Database", International Symposium on Computer Science and Society (ISCCS), pp.323-326, 2011.

[6] L. Sweeney, "K-anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570, 2002.

[7] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation", IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[8] S.Vijayarani, A.Tamilarasi, M.Sampoorna, "Analysis of Privacy Preserving K Anonymity Methods and Techniques", Proceedings of the International Conference on communication and Computational Intelligence, pp.540-545, December 2010.

[9] Xiaolin Zhang, Hongjing Bi, "Research on Privacy Preserving Classification Data Mining Based on Random Perturbation", International Conference on Information Networking and Automation (ICINA), pp.173-178, 2010.

[10] Jinfei Liu, Jun Luo, and Joshua Zhexue Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity Requirements", International conference on Data Mining Workshops, pp.666-670, 2011.

[11] Vassilios S. Verykios, Elisa Bertino, "State-of-the-art in Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Data (SIGMOD) Record, Vol. 33, No. 1,pp.50-57, 2004.

[12] Haisheng Li, "Study of Privacy Preserving Data Mining", Third International Symposium on Intelligent Information Technology and Security Informatics, pp.700-703, 2010.

[13] Charu C. Aggarwal, Philip S. Yu, "A condensation approach to privacy preserving data mining", International Conference on Extending Database Technology (EDBT), pp. 183-199, 2004.

[14] Nan Zhang; Wei Zhao, "Privacy-Preserving Data Mining Systems," in Computer , vol.40, no.4, pp.52-58, April 2007