Proceedings of International Conference on Emerging Trends in Electronics & Telecommunications (ICETET) 29th - 31st May, 2015
KUALA LUMPUR, MALAYSIA (MT0CS003)

eISSN : 2349-0020
pISSN : 2394-4544

A Framework for Solving Identity Disclosure Problem in Collaborative Data Publishing

Guguloth Vijaya [1] , A. Devaki [2] and Dr. Shoban Babu Sriramoju [3]

[1 2] Research Scholar, Department of Computer Science and Engineering,
[1 2] Osmania University, Hyderabad, Telangana, India -500 007.
[3] Software Developer, [3] SR Group of Inistitutions, Warangal, Telangana, India – 506 001.
[1] gvijaya2005@gmail.com, [2] a.devaki2008@gmail.com, [3] babuack@yahoo.com

ABSTRACT

Identity disclosure has been a problem in data publishing. When it comes to data publishing in collaboration with multiple parties, this problem is more. There might be insider attacks meant for obtaining identity of objects in the given data sources provided by data providers. Sensitive details are to be protected when data is being published. Many anonymization algorithms came into existence. One such good algorithm is m-privacy proposed by Goryczka et al. In this paper we build a framework that solved identity disclosure problem in collaborative data publishing. We build a prototype application that uses m-Privacy concept which is applied to horizontally partitioned data and also set-valued data. Experimental results revealed that the prototype is useful for collaborative data publishing without allowing identity disclosure attacks.

Key Words – Data mining, privacy preserving data mining, collaborative data publishing.

I. INTRODUCTION

Data mining has been around for more than a decade to analyze data and make decisions which are useful to enterprises. Data mining is used in all domains. For instance banking, insurance, healthcare etc. are using data mining techniques to make expert decisions. In health care domain valuable information can be exchanged or shared among hospitals across the globe. This will allow collaborative computing as well. When data is published for mining activities, it is important to take care of privacy. The datasets in health care domain have data pertaining to patients. However, identity of the patients needs to be kept confidential. There are many possibilities that the identity is disclosed by the published data. Though the sensitive data is anonymized, it is possible that the data is matched with external data and establish identity of records. Therefore it is a challenging problem to achieve privacy preserving data publishing. Privacy preserving data analysis has been given much attention recently and many approaches as explored in [1], [2], [3] came into existence.

Collaborative data publishing is the approach in which multiple data providers share their data for data mining tasks. In this case the anoymized data is given by each provider. Still the insider attacks might be possible to establish identity of patients. Collaborative data publishing has been explored in [4], [5] and [6]. Another concept came into existence is known as secure multi-party computations [7], [8].

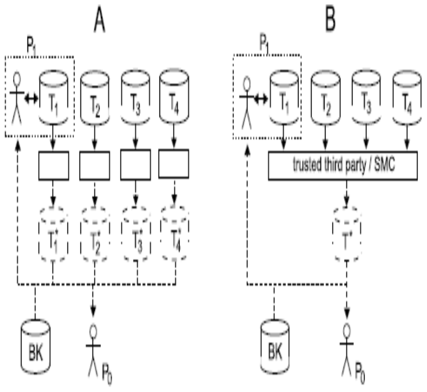
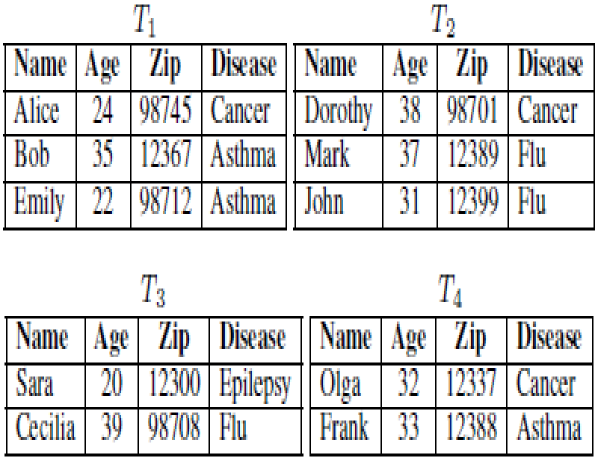


Figure 1 – Illustrates a scenario for distributed data publishing [9]

As can be seen in Figure 1, it is evident that (A) represents data publishing individually while (B) represents distributed or collaborative data publishing. The integrated data utility will not be lost in case of (B). Figure 2 shows the data tables which are considered as input for collaborative data publishing.



T_1

Name	Age	Zip	Disease
Alice	24	98745	Cancer
Bob	35	12367	Asthma
Emily	22	98712	Asthma

T_2

Name	Age	Zip	Disease
Dorothy	38	98701	Cancer
Mark	37	12389	Flu
John	31	12399	Flu

T_3

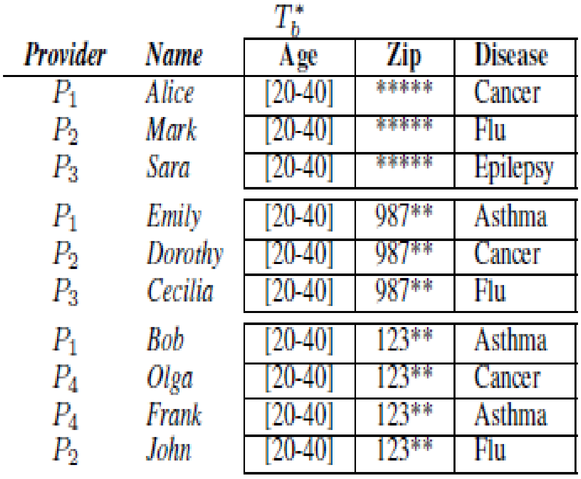
Name	Age	Zip	Disease
Sara	20	12300	Epilepsy
Cecilia	39	98708	Flu

T_4

Name	Age	Zip	Disease
Olga	32	12337	Cancer
Frank	33	12388	Asthma

Figure 2 – Sample input data [9]

As seen in Figure 2, the data is multiple tables. This data is considered for processing. The m-Privacy results as explored in [9] are presented in Figure 3. The data is given provider wise and there are sensitive columns in the table that can disclose identity of patients.



T_b^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Figure 3 – Provider wise data [9]

As seen in Figure 3, it can be understood that insider can match the data with some external data available in order to infer the identity of the records. Secure multi-party protocols explored n [5] can be used in order to preserve privacy. However it cannot detect the insider attacks made by providers. To overcome this problem m-Privacy [9] approach came into existence. In m-Privacy al algorithm was built in order to preserve privacy of shared data by multiple parties for collaborative data publishing.

In this paper we build a prototype application that uses the concepts of m-Privacy in order to achieve privacy preserving collaborative publishing. Healthcare domain is considered as an example for the application. The remainder of the paper is structured as follows. Section II provides review of literature on anonymIzation techniques and privacy preserving data publishing and distributed data publishing. Section III provides details about the prototype application. Section IV presents experimental results while section V concludes the paper.

II. RELATED WORK

Privacy preserving data mining has been around for many years. This kind of mining is meant for mining data without disclosing identity of the records in the data source. Many anonymization techniques came into existence. They include k-anonymity [10], [11], l-diversity [12] and t-closeness [13]. An unconditional privacy guarantee approach is differential privacy [1], [3], [14], [15], [16]. These privacy preserving techniques operate on various data sets and ensure that the data is published for mining with anonymization. The anonymized data is the data where sensitive fields are not disclosed. It does mean that identity of the records cannot be disclosed. All these techniques have improved the art of anonymization to some extent. With increased anonymization it is difficult for adversaries to perform identity disclosure attacks on data given for data mining.

Recently in [9] Goryczka et al. introduced the notion of m-Privacy that guarantees privacy and avoids privacy disclosure. Moreover the m-Privacy works well for collaborative data publishing. The anonymzied data satisfies given privacy constraint. Heuristic algorithms were presented by them for checking m-privacy efficiently. The provider-aware algorithm provided by them was able to anonymize data for collaborative data publishing. M-privacy is able to provide better utility for privacy preserving collaborative data publishing.

III. PROTOTYPE APPLICATION

We built a prototype application that demonstrates the concepts introduced in m-privacy [9]. The application works with both horizontally partitioned data and also the data with set of values. The environment used to build the application is a PC with 4 GB RAM, core 2 dual processor running Windows 7 operating system.

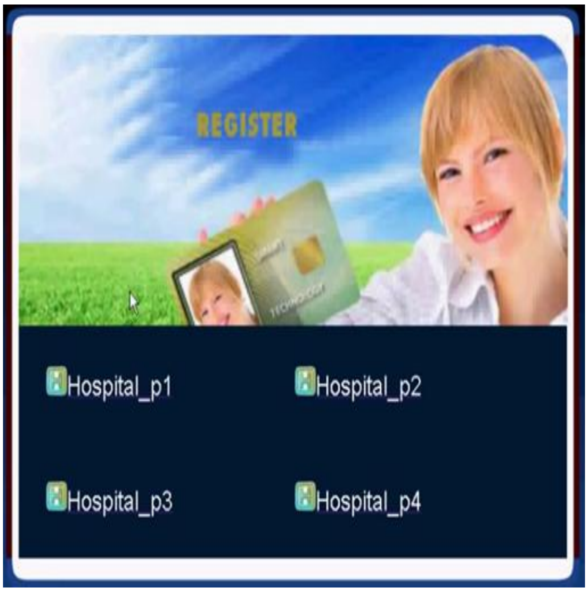
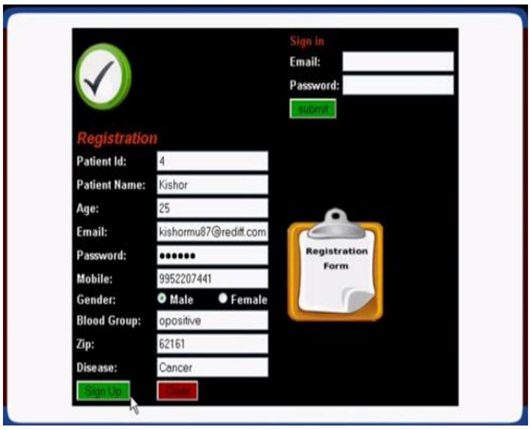


Figure 4 – UI for registration of four data providers

As can be seen in Figure 4, it is evident that there is provision for four data providers for demonstrating privacy preserving collaborative data publishing. The four data providers register themselves and provide various data that is used for publishing after applying m-privacy.



Registration
Patient Id: 4
Patient Name: Kishor
Age: 25
Email: kishormu87@rediff.com
Password: ••••••
Mobile: 9952207441
Gender: ● Male ○ Female
Blood Group: opositive
Zip: 62161
Disease: Cancer

Figure 5 – UI for patient registration

Each data provider application allows registration of new patients. This will help in generating synthetic dataset that can help in testing the concept of privacy preserving collaborative data mining. The UI in figure 5 shows form that captures all details of patients including the disease with which he/she is suffering.

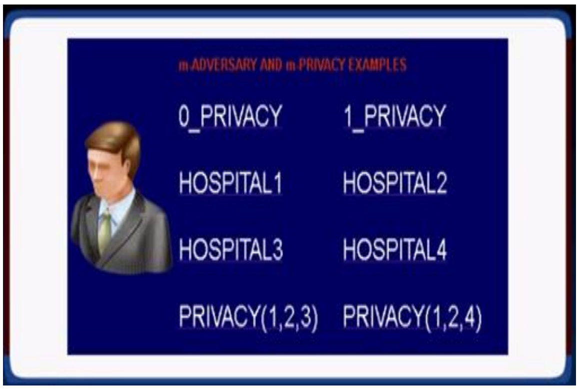


Figure 6 – Provision for collaborative data publishing

As shown in Figure 6, the prototype application has provision for applying m-privacy for the data provided by four hospitals. The m-privacy concept is applied on the collaborative data and the publishing ensures that the identity of the records is not disclosed. It also ensures that no one can launch collusion attack or identity disclosure attack on the data.

IV. EXPERIMENAL RESULTS

The proposed application demonstrates m-privacy concept with four data providers. The data collected from the data providers is collectively anonymized before using it. The anonymization is done using m-privacy. The m-privacy guarantees that the identity disclosure attack is prevented. The results of the experiments are presented in Fig 7.

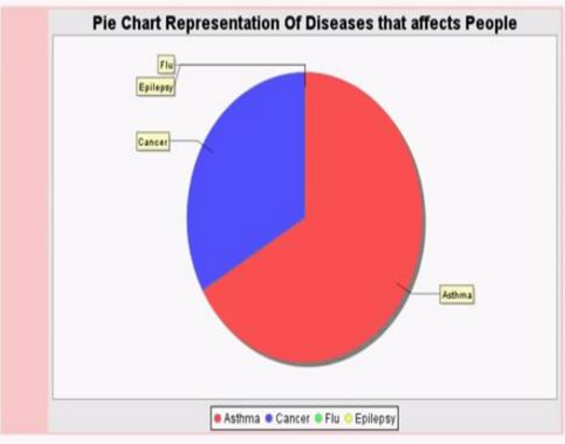


Figure 7 – Results showing disease distribution

As shown in Figure 7, it is evident that the disease distribution is mined and presented in privacy preserving fashion. The result of mining is presented which shows that Asthma is prevailing in the society as more number of people is affected by this. Besides the prototype application also ensures that the data that has been published for mining does not disclose identity information.

IV. CONCLUSIONS AND FUTURE WORK

In this paper we studied privacy preserving collaborative data publishing. The notion of m-privacy proposed in [9] has been studied used it in our prototype implementation. Moreover, we adapted it for set-valued data. The prototype application is built to support multiple data providers and collaborative data publishing with privacy constraints. The application is built with user-friendly interface. The application demonstrates the privacy preserving data publishing of data taken from multiple data providers. The empirical results revealed that the application is very useful in the real world. The application works with horizontally partitioned data and also the data with set of values. In future it can be improved to work with other kind of data such as uncertain data.

REFERENCES

[1] C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, pp. 14:1–14:53, June 2010.

[3] C. Dwork, "A firm foundation for private data analysis," Commun. ACM, vol. 54, pp. 86–95, January 2011.

[4] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 4, pp. 18:1–18:33, October 2010.

[5] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in Data and Applications Security XIX, ser. Lecture Notes in Computer Science, 2005, vol. 3654, pp. 924–924.

[6] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

[7] O. Goldreich, Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press, 2004.

[8] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacypreserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

[9] Slawomir Goryczka, Li Xiong and Benjamin C. M. Fung, "m-Privacy for Collaborative Data Publishing", IEEE transactions on knowledge and data engineering vol:pp no:99 year 2013, p1-10.

[10] L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzz., vol. 10, no. 5, pp. 557–570, 2002.

[11] P. Samarati, "Protecting respondents' identities in microdata release," IEEE T. Knowl. Data En., vol. 13, no. 6, pp. 1010–1027, 2001.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in ICDE, 2006, p. 24.

[13] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and ldiversity," in In Proc. of IEEE 23rd Intl. Conf. on Data Engineering (ICDE), 2007.

[14] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in ICDE, 2006.

<

[15] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, "Identifying attack models for secure recommendation," in In Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.

[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k-anonymity," in Proc. of the 2005 ACM SIGMOD Intl.Conf. on Management of Data, 2005, pp. 49–60.