Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.44 – 49

RESEARCH ARTICLE

A Novel Framework for Preventing Inference Attacks in Collaborative Data Publishing

Prof. L. Ramesh Babu[1], Mrs. M.Sandhya Rani[2], Shravya Channamadhavuni[3], Ramana Amrutha. M[4]

HOD, Department of IT, Bhoj Reddy Engineering College for women, Hyderabad, Andhra Pradesh, India[1]
Associate Professor, Bhoj Reddy Engineering College for women, Hyderabad, Andhra Pradesh, India[2]
Student, Department of IT, Bhoj Reddy Engineering College for women, Hyderabad, Andhra Pradesh, India[3]
Student, Department of IT, Bhoj Reddy Engineering College for women, Hyderabad, Andhra Pradesh, India[4]

laks.ramesh@gmail.com[1], sandhya_medi@yahoo.com[2], shravya.ch93@gmail.com[3], mungamuruamrutha@gmail.com[4]

ABSTRACT: Disclosure of sensitive data is the problem in collaborative data publishing. Collaborative data publishing involves multiple parties where data privacy is very important. There are number of threats to the privacy of data. For instance, there is possibility for insider attacks to obtain identity of real world objects. The data sources provided by multiple parties for collaborative publishing are to be protected from disclosure attacks. Moreover, the sensitive details are to be protected from being disclosed. To overcome this problem, many anonymization techniques came into existence. M-Privacy is one such good algorithm proposed by Goryczka et al. that provides dependable security to collaborative data publishing. In this paper, we propose a framework that that resolves the problem of identity disclosure in the context of collaborative publishing of data. We built a prototype application that demonstrates the proof of concept. Our empirical results are encouraging.

Index Terms – Data mining, privacy preserving data mining, collaborative data publishing

I. INTRODUCTION

Data mining is the process of discovering knowledge from databases. It has been around for many years for making well informed decisions by mining data and obtains business intelligence. Organizations of all fields need data mining techniques in order to get business intelligence mined from real world data sources. Banking, healthcare, insurance are some of the examples for the domains where data mining is widely used. Data can be published in collaborative fashion by collecting data from multiple data providers. For example, multiple hospitals or multiple banks can involve in collaborative data publishing. When data is provided for mining activities, it is essential to ensure that sensitive information is not disclosed. For instance identify of the patients in healthcare domain data needs to be protected. There are many attacks that can disclose identity of data being published. The attacks might be either from internal or external.

The data inference attacks can be made from the known data. The attack proceeds with matching of known data with unknown data in other data sources. Thus the identity disclosure takes place. In this context, it is essential to have mechanisms in place for protecting sensitive data from being disclosed. However, it is a challenging problem to achieve privacy preserving collaborative data publishing. Privacy preserving data analysis has been an important research area in data mining. Many researchers [1], [2], [3] contributed towards collaborative data publishing which is coupled with privacy preservation. Multiple data providers can give their data for collaborative data publishing which is part of data mining. However, each data provider is expected to give anonym zed data. However, insider attacks might be possible in the collaborative data publishing phenomenon. It has been explored in [6], [5] and [4]. Secure multi-party communications is another similar research area [8], [7].

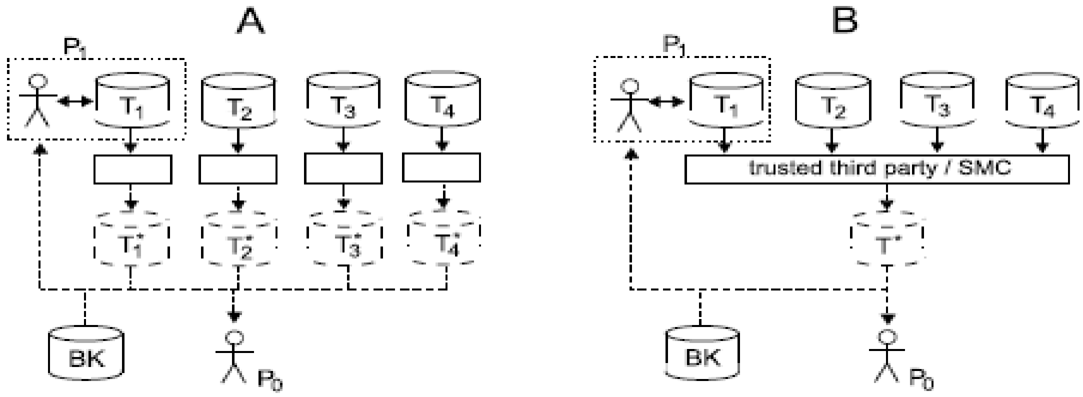
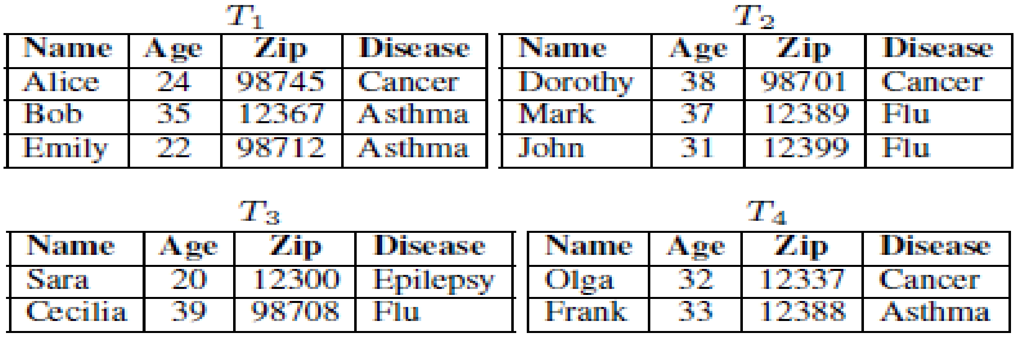


Figure 1 – Illustrates a scenario for distributed data publishing [9]

As can be seen in Figure 1, it is evident that (A) represents data publishing individually while (B) represents distributed or collaborative data publishing. With respect to B, there is no loss of integrated data utility. For collaborative data publishing sample data tables are provided in Figure 2.



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

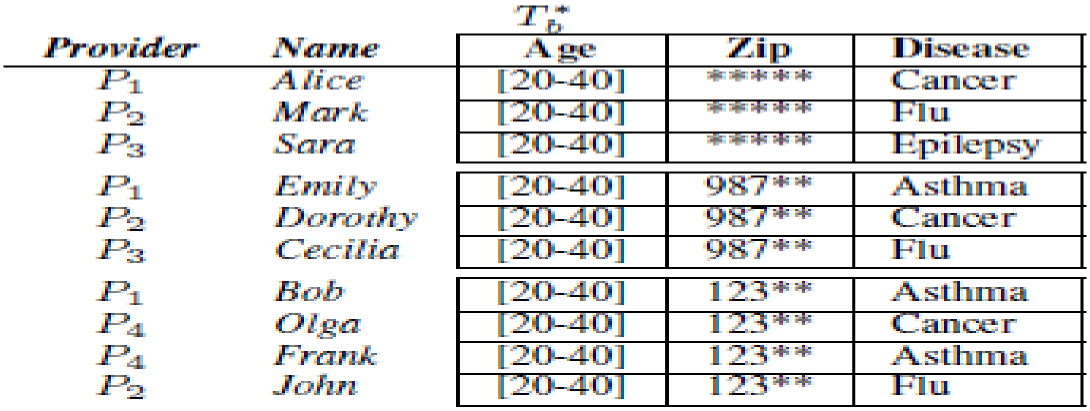
Name	Age	Zip	Disease
Sara	20	12300	Epilepsy
Cecilia	39	98708	Flu

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. Provider wise data is provided and there are sensitive columns that might be providing possible identity disclosure.



		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]

Figure 3 shows that insider attack is possible to match the data with external data which is available somewhere else in order to infer the data which is not known. In [5] secure multi-party communications is explored for privacy preserving. In other words it complements the privacy preserving approach. However, the SMC will not be able to detect attacks made by insiders. In order to overcome this problem, m-Privacy [9] was explored. The m-Privacy algorithm is able to preserve data and that is provided as part of multiparty collaborative publishing of data [9]. The algorithm is built in order to apply privacy, otherwise it results in data loss.

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

Data mining is used to mine data from relevant sources. Privacy preserving is an important aspect of data mining. Data when given for third party for mining, its privacy is very important otherwise it is possible to subject to inference attacks. To overcome this problem, many anonymizatoin solutions came into existence. For instance, k-anonymity [11], [10], l-diversity [12] and t-closeness [13]. Many researchers focused on privacy preserving approaches where unconditional pricy is guaranteed [1], [3], [14], [15], [16]. Privacy preserving techniques are common for many datasets. Various datasets can be used for privacy preserving data publishing with anonymization techniques apply. The anonymized data does not disclose sensitive information. There are many techniques that are used to improve anonymizaton in one way or other. With proper anonymization, adversaries will not be able to perform infereene or data disclosure attacks on such data mining operations.

As a matter of fact, recently Goryczka et al. [9] proposed a new privacy preserving algorithm known as m-Privacy which guarantees that information disclose attacks are not possible. The privacy constraint given is satisfied by the proposed model. Many heuristic algorithms are also came into existence. By applying anonymization techniques it is possible to protect data from malicious attacks. Moreover, it is essential in collaborative data publishing environment to be careful about the possible collaborative data publishing. M-Privacy is proved to provide security for data publishing activities. It also ensure that the m-Privacy concept can be achieved between than any other algorithm using m-privacy.

III. PROTOTYPE APPLICATION

This section provides the prototype application built by us. The application demonstrates the proof of concept pertaining to m-privacy [9] that facilitates privacy preserving data mining and data publishing with proposed anonymiztion technique. The environment used to build the application is a PC with 4 GB RAM, core 2 dual processor running Windows 7 operating system.

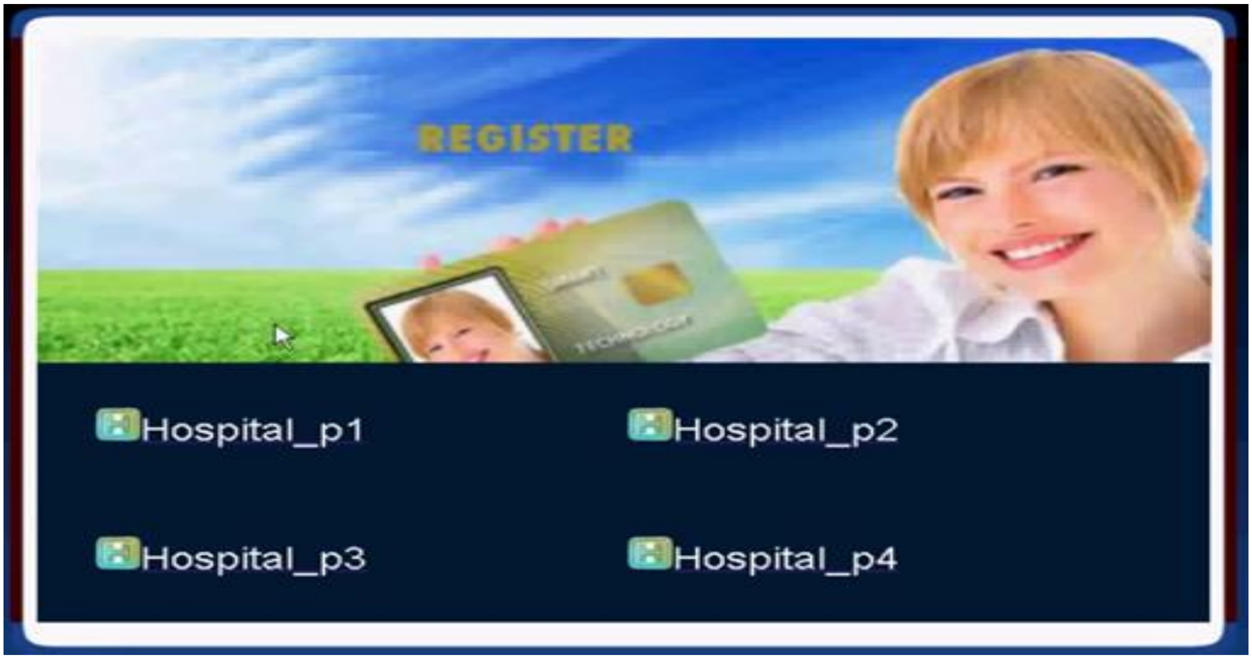
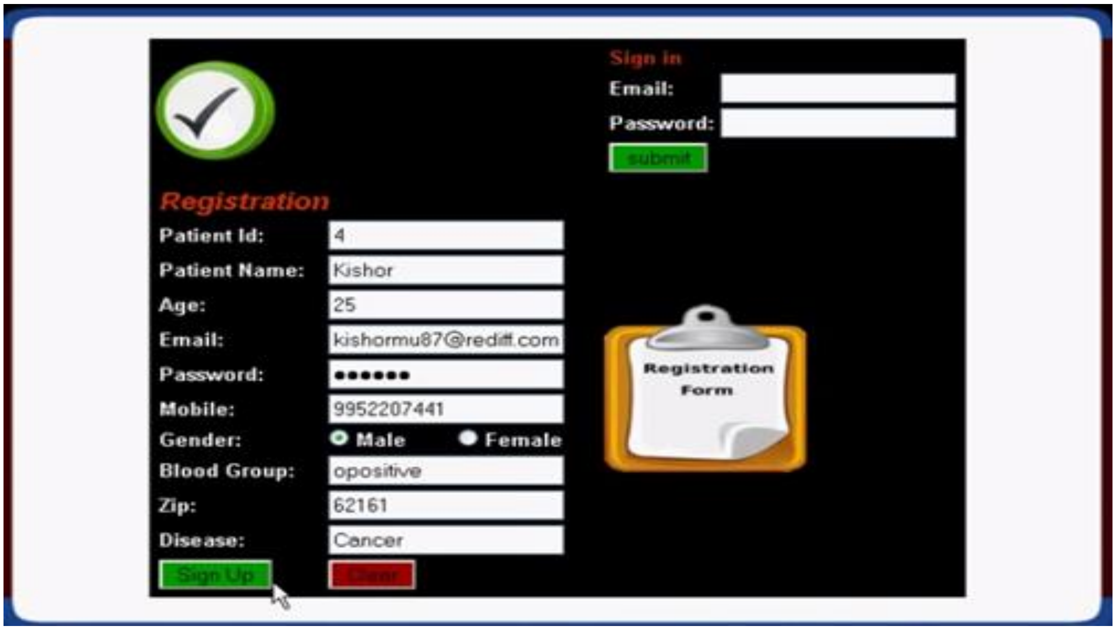


Figure 4 – UI for registration of four data providers

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 mc-privacy.



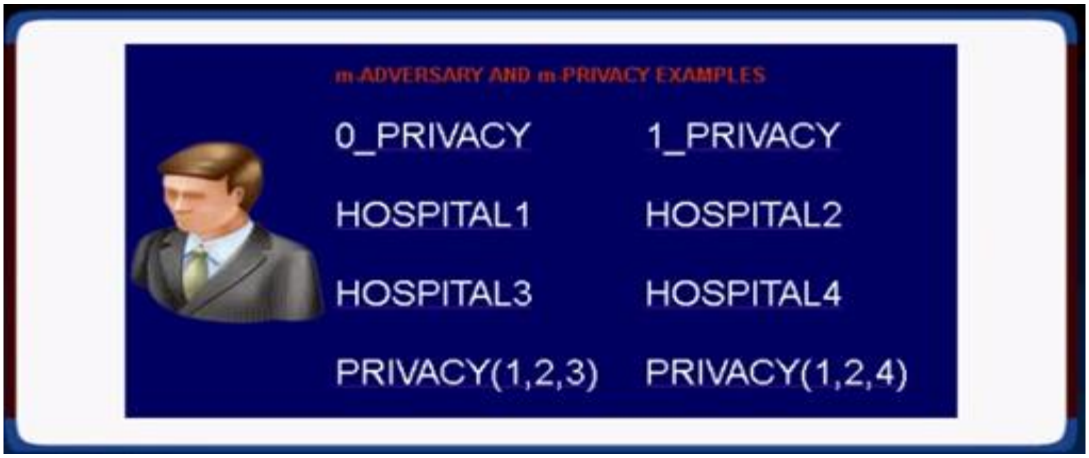
Sign in
Email:
Password:

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.



m-ADVERSARY AND m-PRIVACY EXAMPLES

0_PRIVACY 1_PRIVACY

HOSPITAL1 HOSPITAL2

HOSPITAL3 HOSPITAL4

PRIVACY(1,2,3) PRIVACY(1,2,4)

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 Figure 7.

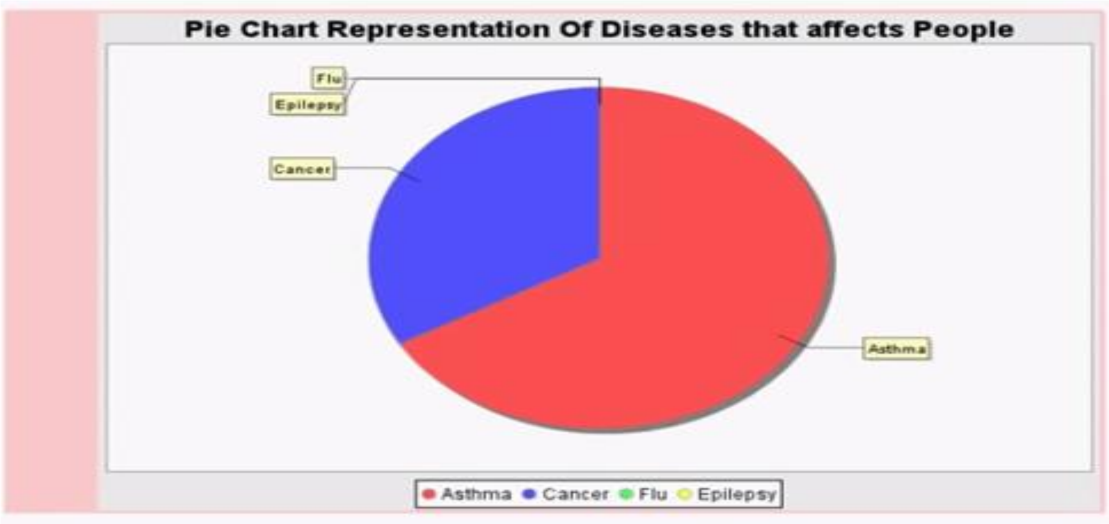


Figure 7 – Results showing disease distribution

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.

V. CONCLUSIONS AND FUTURE WORK

Privacy of data plays an important role in privacy preserving data mining in collaborative environment. The notion of privacy has been applied by the researchers in [9] which provide data publishing in secure environment. Set-valued data has been adapted in our prototype application. M-Privacy plays an important role in providing security to data. The prototype application built by us is meant for mining data which has been provided by multiple parties that are invoked in data sharing. User-friendly application has been built by using privacy preserving data mining techniques. The application we built can demonstrate the proof of concept. The application can be extended further in future to have more flexible and scalable solution to the problem of collaborative data publishing. The prototype works with both vertical and horizontal 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.