M-Privacy for Collaborative Data Publishing

Jayashri K. Bhosle
M.E(CSIT) 2013-2015
M.B.E'sCOE Ambajogai
Maharashtra, India

Vanja R. Chirch
Assitant Professor
M.B.E's COE, Ambajogai
Maharashtra, India

ABSTRACT
More than one data provider collaborate to publish their data is considered here. m-privacy is a technique proposed to defend m-adversary during collaborative data publishing. M-privacy satisfies the privacy problem while publishing sensitive data. Apart from providing privacy to published data, it is also necessary to provide security between the data provider and third party/un-trusted server, to ensure this, Secure multiparty communication (SMC) protocol is used to provide secure data transfer from publisher and server. There were techniques such as k-anonymity, l-diversity, t-closeness, which were proposed to handle external attacks in data publishing, but none is published for considering internal attacks. This m-privacy is a technique, which considers internal attacks.

AIM: The goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties.

General Terms
Binary Algorithm, Heuristic Algorithm, K-Anoymity, t-closness

Keywords
Annoymization , Adversary, TTP, SMC.

1. INTRODUCTION
There is an increasing need for sharing data that contain personal information from distributed databases. For example, in the healthcare domain, a national agenda is to develop the Nationwide Health Information Network (NHIN) to share information among hospitals and other providers, and support appropriate use of health information beyond direct patient care with privacy protection. Privacy preserving data analysis and data publishing have received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. When the data are distributed among multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently, which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing, which anonymizes data from all providers as if they would come from one source, using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations.

2. LITERATURE SURVEY
- **M. Ashok Kumar, R. Nandhakumar**

Secure multi-party computation protocols for collaborative data publishing with m-privacy. All protocols are extensively analyzed and their security and efficiency are formally proved. Experiments on real-life datasets suggest that our approach achieves better or comparable utility and efficiency than existing and baseline algorithms while satisfying m-privacy. consider the collaborative data publishing problem for anonym zing horizontally partitioned data at multiple data providers. consider a new type of "insider attack" by colluding data providers who may use their own data records (a subset of the overall data) to infer the data records contributed by other data providers. The paper addresses this new threat, and makes several contributions. First, we introduce the notion of m-privacy, which guarantees that the anonym zed data satisfies a given privacy constraint against any group of up to m colluding data providers. Second, we present heuristic algorithms exploiting the monotonicity of privacy constraints for efficiently checking m-privacy given a group of records. Third, we present a data provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymzed data with efficiency.

- **Aseema Jana, Shubham Joshi**

Privacy takes an important role to secure the data from various probable attackers. For public advantage data need to be shared as required for Health care and researches, individual privacy is major concern regarding sensitive information. So while publishing such data, privacy should be conserved. Publishing collaborative data to multiple data provider's two types of problem occurs, first is outsider attack and second is insider attack. Outsider attack is by the people who are not data providers and insider attack is by colluding data provider who may use their own data records to understand the data records shared by other data providers.

Problem can be overcome by combining slicing techniques with m privacy techniques and addition of protocols as secure multiparty computation and trusted third party will increase the privacy of system effectively.

- **Priya V. Mundafale#1 Prof. GurudevSawarkar*2**

Data mining is the extraction of interesting patterns or knowledge from huge amount of data. With the explosive development in Internet, data storage and data processing echnologies, privacy preservation has been one of the greater concerns in data mining.

A number of methods and techniques have been developed for privacy preserving data mining. Privacy preserving data mining is an important issue in the areas of data mining and security on private data in the following scenario: Multiple parties, each having aprivate data set, want a group of people organized for a joint purpose rule mining without disclosing their private data to other parties. Because of the interactive nature among parties, developing a secure framework to achieve such a computation is both challenging and desirable. There is an increasing need for sharing data repositories containing personal information across multiple distributed, possibly untrusted, and privatedatabases.Such data sharing is subject to constraints imposed by privacy of data subjects as well as data confidentiality of institutions or data providers Developed a set of decentralized protocols that enable data sharing for horizontally partitioned databases.

- **Noman Mohammed and Benjamin C. M. Fung**

Sharing healthcare data has become a vital requirement in healthcare system management; however, inappropriate sharing and usage of healthcare data could threaten patients' privacy. Study the privacy concerns of sharing patient information between the Hong Kong Red Cross Blood Transfusion Service (BTS) and the public hospitals. We generalize their information and privacy requirements to the problems of centralized anonymization and distributed anonymization, and identify the major challenges that make traditional data anonymization methods not applicable. Furthermore, we propose a new privacy model called LKC-privacy to overcome the challenges and present two anonymization algorithms to achieve LKC-privacy in both the centralized and the distributed scenarios. Experiments on real-life data demonstrate that our anonymization algorithms can effectively retain the essential information in anonymous data for data analysis and is scalable for anonymizing large datasets.

3. EXISTING SYSTEM

Attacks by External Data Recipient Using Anonymized Data

- A data recipient, could be an attacker and attempts to infer additional information about the records using the published data and some background knowledge (BK) such as publicly available external data.
- Bayes-optimal privacy notion is used to protect against specific types of attacks by assuming limited background knowledge.
- For example, k-anonymity, prevents identity disclosure attacks by requiring each equivalence group, records with the same quasi-identifier values, to contain at least k records.
- Representative constraints that prevent attribute disclosure attacks include l-diversity, which requires each equivalence group to contain at least l "well-represented" sensitive values
- t-closeness, which requires the distribution of a sensitive attribute in any equivalence group to be close to its distribution in the whole population.
- Differential privacy publishes statistical data or computational results of data and gives unconditional privacy guarantees independent of attackers background knowledge.

Attacks by Data Providers Using Intermediate Results and Their Own Data

- The data providers are semihonest, commonly used in distributed computation setting. They can attempt to infer additional information about data coming from other providers by analyzing the data received during the anonymization.
- A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols can be used to guarantee there is no disclosure of intermediate information during the anonymization.

4. PROPOSED SYSTEM

Attacks by Data Providers Using Anonymized Data and Their Own Data

- Collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing a subset of records is considered.
- A data provider could be the data owner itself who is contributing its own records.
- Each provider has additional data knowledge of their own records, which can help with the attack. This issue can be further worsened when multiple data providers collude with each other.

5. RESULTS AND DISCUSSION

A new type of potential attackers in collaborative data publishing – a coalition of data providers, called m-adversary is considered. To prevent privacy disclosure by any m-adversary we showed that guaranteeing m-privacy is enough. Heuristic algorithms is presented exploiting equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy. We introduced also a provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymized data.

There are many remaining research questions. Defining a proper privacy fitness score for different privacy constraints is one of them. It also remains a question to address and model the data knowledge of data providers when data are distributed in a vertical or ad-hoc fashion. It would be also interesting to verify if our methods can be adapted to other kinds of data such as set-valued data.

6. 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 privacy preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

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

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

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

[12] 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.

[13] 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.

[14] D. Kifer, "Attacks on privacy and definetti's theorem," in Proc. of the 35th SIGMOD Intl. Conf. on Management of Data, 2009, pp. 127–138.

[15] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in Proc. of the 2011 Intl. Conf. on Management of Data, 2011, pp. 193–204.

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

[17] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-privacy for collaborative data publishing," Emory University, Tech. Rep., 2011.

[18] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in Proc. of the 32nd Intl. Conf. on Very Large Data Bases, 2006, pp. 139–150.

[19] G. Cormode, D. Srivastava, N. Li, and T. Li, "Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data," Proc. VLDB Endow., vol. 3, Sept. 2010.

[20] Y. Tao, X. Xiao, J. Li, and D. Zhang, "On anti-corruption privacy preserving publication," in Proc. of the 2008 IEEE 24th Intl. Conf. on Data Engineering, 2008, pp. 725–734.

[21] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in Proc. of the IFIP TC11 WG11.3 Eleventh Intl. Conf. on Database Security XI: Status and Prospects, 1998, pp. 356–381.

[22] 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.

[23] N. Mohammed, B. C. M. Fung, K. Wang, and P. C. K. Hung, "Privacypreserving data mashup," in Proc. of the 12th Intl. Conf. on Extending Database Technology, 2009, pp. 228–239.

[24] S. Zhong, Z. Yang, and R. N. Wright, "Privacy-enhancing kanonymization of customer data," in Proc. of the 24th ACM SIGMODSIGACT-SIGART Symposium on Principles of Database Systems, 2005, pp. 139–147.

[25] P. Jurczyk and L. Xiong, "Distributed anonymization: Achieving privacy for both data subjects and data providers," in DBSec, 2009, pp. 191–207.

[26] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan, "Computational differential privacy," in Advances in Cryptology – CRYPTO 2009, ser. Lecture Notes in Computer Science, vol. 5677, 2009, pp. 126–142.