Privacy-Preserving Data Publishing for Horizontally Partitioned Databases

Pawel Jurczyk
Department of Math&CS
Emory University
Atlanta, GA, USA
pjurczy@emory.edu

Li Xiong
Department of Math&CS
Emory University
Atlanta, GA, USA
lxiong@emory.edu

ABSTRACT

There is an increasing need for sharing data repositories containing personal information across multiple distributed, possibly untrusted, and private databases. Such data sharing is subject to constraints imposed by privacy of data subjects as well as data confidentiality of institutions or data providers. We developed a set of decentralized protocols that enable data sharing for horizontally partitioned databases given these constraints. Our approach includes a distributed anonymization protocol that allows independent data providers to build a virtual anonymized database, and a distributed querying protocol that allows clients to query the virtual database.

Categories and Subject Descriptors

H.2 [**DATABASE MANAGEMENT**]: General, Systems

General Terms

Algorithms, Experimentation, Design, Security

1. INTRODUCTION

Current information technology enables many organizations to collect, store, and use various types of information about individuals in large repositories. Government and organizations increasingly recognize the critical value in sharing such a wealth of information across multiple distributed, private, and possibly untrusted databases. An example is the Shared Pathology Informatics Network[1] initiative by the National Cancer Institute that attempts to provide a search interfaces for electronic databases at institutions across the country to locate human specimens and associated clinical and pathologic data needed for cancer research. However, personal health information is protected under regulations such as the Health Insurance Portability and Accountability Act[2]. In addition, institutions may not want to reveal their private databases to each other for various reasons.

These scenarios can be generalized into the problem of privacy preserving data publishing for multiple distributed databases where multiple *data custodians* need to publish an anonymized and integrated view of the data that does not contain individually identifiable information. Such data sharing is subject to two constraints. The first constraint is the privacy of the individuals or in general the data subject (such as the patients). The second is the data confidentiality of the data custodians (such as the institutions). Given a query spanning multiple databases, query results should not contain individually identifiable information. In addition, institutions should not reveal their databases to each other apart from the query results.

Existing and potential solutions. Privacy preserving data publishing for a single database has been extensively studied in recent years. A large body of work contributes to data anonymization that transforms a dataset to meet a privacy principle such as k-anonymity using techniques such as generalization or suppression (removal) so that it does not contain individually identifiable information (e.g. [2]).

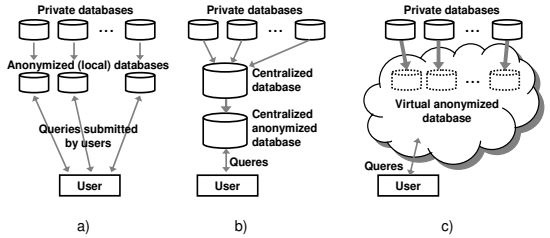

Figure 1: Architectures for privacy preserving data publishing

There are a number of potential approaches one may apply to enable privacy preserving data publishing for distributed databases. A naive approach is for each data custodian to perform data anonymization independently as shown in Fig. 1a. Data recipients or clients can then query the individual anonymized databases or an integrated view of them. One main drawback of this approach is that data is anonymized before the integration and hence will cause the data utility to suffer. In addition, individual databases reveal their ownership of the anonymized data.

An alternative approach assumes an existence of third party that can be trusted by each of the data owners as shown in Fig. 1b. In this scenario, data owners send their data to this trusted third party where data integration and anonymization are performed. Then, clients can query the centralized database. However, finding such a trusted third party is not always feasible. Compromise of the server by hackers could lead to a complete privacy loss for all participating parties.

Contributions. We propose a distributed data anonymization approach as illustrated in Fig. 1c. In this approach,

[1]http://www.cancerdiagnosis.nci.nih.gov/spin/
[2]http://www.hhs.gov/ocr/hipaa/

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

data owners participate in distributed protocols to produce a *virtual* integrated and anonymized database which can be then queried by clients. The anonymized data still resides at individual databases and the integration and anonymization of the data is performed through the distributed protocols.

2. DISTRIBUTED ANONYMIZATION

We assume that the data are split horizontally among n sites ($n > 2$) and each site owns a private database d_i. In addition, the *quasi-identifier* of each local database is uniform among all the sites. The sites engage in a distributed anonymization protocol where each site produces a local anonymized dataset a_i and their union forms a virtual database that is guaranteed to be k-anonymous. Note that a_i is not required to be k-anonymous by itself. When users query the virtual database, each individual database executes the query on a_i and then engage in a distributed querying protocol to assemble the results that are guaranteed to be k-anonymous.

Protocol Structure. The proposed protocols are designed to run over a decentralized network. The protocol structure is presented in Figure 2. Nodes are mapped to a ring topology randomly. We assume that each node knows its predecessor and successor. Each node has a local computation module that executes its part of the protocol independently and passes the computation result along the ring.

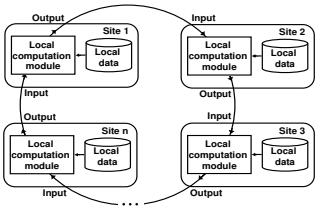

Figure 2: Protocol Structure

Distributed Anonymization Protocol. Our distributed algorithm for anonymization is based on the Mondrian algorithm [2] that uses greedy recursive partitioning of the (multidimensional) quasi-identifer domain space to satisfy k-anonymity. It recursively chooses the split attribute with the largest normalized range of values, and (for continuous or ordinal attributes) partitions the data around the median value of the split attribute. This process is repeated until no allowable split remains, meaning that a particular region cannot be further divided without violating the anonymity constraint, or constraints imposed by value generalization hierarchies.

The key idea for the distributed anonymization protocol is to use a set of secure atomic multi-party protocols to realize the Mondrain method for the distributed setting. We assume a leading site is selected for the protocol. The steps performed at the leading site are similar to the centralized Mondrian method. It chooses the best attribute (with largest spread) for split. Next, it performs the split and recursively checks whether further split of new subsets is possible. When selecting the best split attribute, the leading site needs to have the knowledge of the ranges of values of each attribute with respect to data items located at all sites. So a secure min/max protocol is used to compute the minimum and maximum value of each attribute across the databases. The leading site can then compute the range of each attribute and select the best attribute with largest range. When finding the split point for partitioning, a secure median protocol is used to find the median value of the attribute with respect to the data across the databases. Finally, when determining whether a partition can be further split, a secure sum protocol is used to count the total number of tuples of the partition across the databases.

Distributed Querying Protocol. The above protocol enables set of nodes to produce a virtual k-anonymous database based on the union of the data horizontally split among the nodes. At the end of the protocol, the local anonymized datasets are not necessary k-anonymized. However, the union of datasets forms the virtual database and is guaranteed to satisfy k-anonymity requirement. Our approach also includes a distributed querying protocol for this virtual database. When a query is received, each database runs the query against its local randomized dataset and the results are then unioned and included in the response. As we require that ownership of items is not revealed during the querying phase, we propose a novel and efficient secure set union protocol for this purpose.

The protocol works as follows. In the initialization or leader selection round, a leader node is selected. In the main protocol round, the leader node i generates a random set r and adds its local subset x_i to this random set. Then it passes its intermediate result to the node $i + 1$. Starting from this point, each node j adds its local subset x_j to the intermediate result and passes the result to node $j+1$. When node i receives the result from its predecessor, the set union can be found by removing random items r from this set.

3. SUMMARY

We have presented a distributed anonymization approach for privacy-preserving data publishing for horizontally partitioned databases. It includes a distributed anonymization protocol based on the Mondrian partitioning method and a distributed querying protocol for securely computing the union of data tuples partitioned among n nodes. Our protocols are not based on expensive cryptographic primitives, rather, we leverage the inherent anonymity of the multiple number of participants of the protocol and utilize randomization approaches to achieve minimal information disclosure and minimal overhead. Our experimental results also show that the approach achieves data utility comparable to a centralized approach. For a detailed description of the protocols, and formal analysis and experimental evaluations of the proposed solution, please refer to [1].

Acknowledgement

The work is partially supported by an Emory URC and an Emory ITSC grant. We would like to thank Kristen Lefevre for providing the implementation of Mondrian algorithm and the anonymous reviewers for their valuable comments.

4. REFERENCES

[1] P. Jurczyk and L. Xiong. Privacy-preserving data publishing for horizontally partitioned databases. Technical Report TR-2008-013, Emory University, Math&CS Dept., 2008.

[2] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.