A Secure Model for Medical Data Sharing

Wong Kok Seng[1,1],Myung Ho Kim[1], Rosli Besar[2], Fazly Salleh[2]

[1]Department of Computer, Soongsil University, 156-743 Sangdo-dong, Dongjak-Gu, Seoul, Korea
[2]Faculty of Engineering, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang Melaka Malaysia
[1]{kswong, kmh}@ssu.ac.kr, [2]{rosli, fazly.salleh.abas } @mmu.edu.my

Abstract. Hospitals within a Telemedicine system would like to share their private local database with other hospitals. However, they do not agree to keep a copy of their database into a central server. The central repository (data warehouse) model is not secure because too much control was granted to the central site. In order to fully utilize the distributed and heterogeneous resources, a secure and privacy preserved model should be used. We proposed a secure data sharing model to facilitate this problem.

Keywords: Data Sharing, Union Dataset, Privacy Preserving

1 Introduction

Medical databases are considered valuable to many parties including hospitals, practitioners, researchers, insurance company, and etc. Hospitals and practitioners used their patients' medical records to support their services, while researchers used these data to validate their research findings. Data sharing or information sharing is necessary for distributed systems, and much works have focused on designing a specific information sharing protocols [1]. However, the privacy of the shared data becoming a challenging issue. Privacy is addressed as preventing dissemination rather than integrating privacy constraints into the data sharing process [4].

In Telemedicine system, each collaborator (hospital) needs to share their private local database with other collaborators. The data sharing in healthcare industry is different from other domains. Medical data is useful, but also harmful to a patient if it's not accurate or real. The shared data received from other collaborators under the Telemedicine system can affects the decisions made by the practitioners. The level of trust among collaborators must be as high as possible in order to guarantee the mutual benefits for all parties.

Conventionally, the medical data sharing can be held by using the trusted third party [3, 8]. A copy of database from each collaborator will be gathered at one place to construct a central repository (data warehouse). Collaborator or data miner will then use the repository to mine their required data. Unfortunately, due to the security and privacy concern, many collaborators were afraid to share their patients' medical

[1] This work was supported by the Soongsil University Research Fund.

records in this way. Without the usage of centralized repository, how can each hospital securely share their private database while the privacy is preserved?

2 Related Research

Data sharing process can be considered as the backbone for many other operations. It is a necessary pre-requisite process for data integration and data mining operations. For data integration methods [5, 11], an effective data sharing technique can produce more accurate union set. At the same time, it is the basic requirement for any privacy preserving solutions.

The proliferation and misuse of medical data is now a subject of global interest. Cryptographic and security research communities paying a closed attention in medical data sharing. Techniques such as sovereign information sharing [2] have been proposed to solve the above mentioned problem. The database contents for each collaborator will not be revealed under this technique. Only the result for database operations (intersection, equijoin, aggregation, etc) can be learned by the collaborators.

Secure multi-party computation which was introduced in [12], is another techniques used to protect the collaborators from revealing additional information which is not needed to be shared. Under this approach, only the answer (output) to the query will be learned by the data miner, and nothing else can be revealed [6, 7].

Database operations such as union or intersection computation, equijoin, and aggregation are important operations that can be used to support the secure data sharing process. For example, intersection computation is used to find the common value for different distributed datasets while revealing only the intersection [9]. However, the computation of these database operations involve anonymity and security concerns [10]. Anonymity means the identity of the data owner shouldn't be identified.

3 Problem Definition

N hospitals, $\{H_a, H_b, ..., H_N\}$ with a private local database $\{D_a, D_b, ..., D_N\}$ respectively would like to share their database. Data miner with query (Q) spanning the tables in all databases to compute the answer to Q without revealing extra information apart from the query result.

Solution 1: Each hospital sends a copy of their entire database to the data miner. The data miner will process the Q on the local copy of the databases.

Solution 2: Each hospital sends a subset of their database to the data miner. The data miner will process the Q on the local copy of the subset databases.

Solution 3: Each hospital sends a copy of their entire database to the central server. The data miner will process the Q on the central repository.

Solution 4: Each hospital sends a subset of their database to the central server. The data miner will process the Q on the central site.

Solution 1 and 2 are not practical because the data miner needs a great cost to store all databases as well as the communication costs. In solution 3 and 4, the level of trust required is too high because too much additional information which is not needed to answer Q will be revealed.

4 Proposed Model

Based on the above definition, we present a solution which can be used as the model for secure data sharing. The following criterion gives a brief description of our proposed framework:

1. No central server is allowed to store databases from all collaborators. A temporary union dataset is used instead of central repository.
2. No collaborator can conclude how many collaborators contributed their dataset into the union dataset.
3. Prevent any collaborator to identify the owner of the datasets (anonymity concern).
4. Third party engine is required to form the union dataset and generate the results to the data miner.
5. No dataset owners can determine any origin records other that its own.
6. Two kinds of queries will be used in this model to facilitate the data sharing goals.
7. The proposed system is secure and privacy protected.

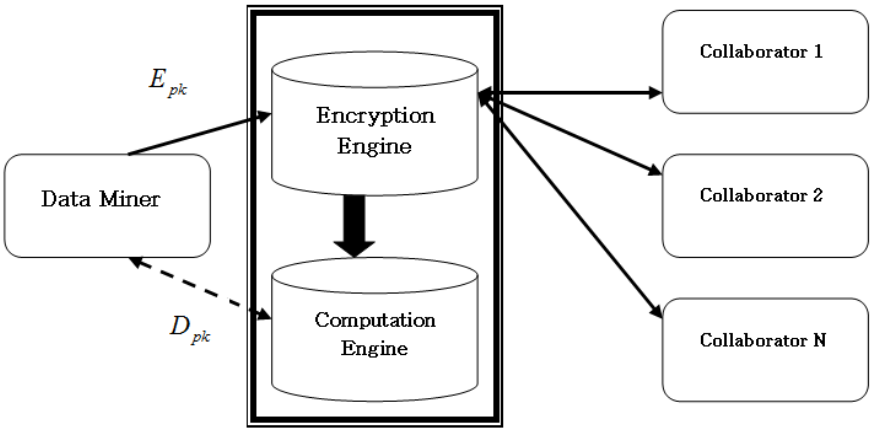

Fig. 1. General Architecture Design

The secure data sharing model contains the following five steps:

Step 1:
Data miner generates a pair of commutative encryption key-pair (E_{pk}, D_{pk}). Only encryption key, E_{pk} is sent to the encryption engine. The decryption key, D_{pk} will be sent and stored at the computation engine.

Step 2:
Encryption engine broadcasts the same encryption key, E_{pk} to all collaborators. A count, N will be generated to compute the frequency of successful delivered encryption keys. Primary queries will be sent to the collaborators once the connection has been established.

Step 3:
Each collaborator will encrypts the requested data with E_{pk} , and send the encrypted data ("ciphertext") to the encryption engine.

Step 4:
Encryption engine will make sure all collaborators responded to the query. It will send all received "ciphertext" to the computation engine after additional operations are performed on the encrypted data.

Step 5:
By using the private key (received from data miner), the computation engine decrypts the ciphertext and form the union dataset. Data miner uses the union dataset by sending the secondary queries.

4.1 Components

Data miner is one of the hospitals (collaborators) within the framework. All hospitals involved in the Telemedicine system can be the data miner when they need to use the shared resources. As the data miner, it must generate a pair of commutative cryptography key. The encryption key will be broadcast to other collaborators through encryption engine while the decryption key will be sent and use by the computation engine. Under this model, data miner will receive the requested dataset without knowing who the contributors are. At the same time, the number of sources will not be revealed.

Encryption engine is the only party which had the direct communication with data miner, computation engine, and all collaborators. It served as the central hub for all parties. Upon receiving the encryption key from the data miner, the encryption engine will broadcasts the key to all collaborators in order to establish an active connection. At this stage, the number of successful delivered key will be stored in a count, N. When the requested dataset is returned to the encryption engine, the encryption engine must make sure that the number of received encrypted dataset is equal to the count, N

before sending them to the computation engine. This is to ensure that all collaborators already responded to the queries and no other incoming dataset is in the queue. Encryption engine can learn the number of sources available, identity of data miner and all collaborators. At the same time, it can know the primary queries made by the data miner. However, it will not be informed about the contents of the requested data because all data were encrypted by the collaborators.

Computation engine plays an important role in this model. By holding the decryption key received from data miner, it responsible to decrypt all encrypted dataset received from the encryption engine. Decrypted dataset will be combined to form a union dataset. This union dataset is a temporary shared repository for a specific primary query. It will be used as the replacement for the central repository and support all secondary queries from the data miner.

Collaborator needs to response to the encryption engine to initiate an active connection. All collaborators will share their private local database in this framework by answering the primary queries made by the data miner. These primary queries were delivered by the encryption engine to each collaborator. Based on the primary queries, each of the collaborators will encrypt the requested dataset with the encryption key. If the requested dataset is not found, an empty dataset will be sent to prevent the encryption engine to know which collaborator does not contribute to the union dataset. In this model, no direct communication among collaborators. They do not know the number of collaborators involved as well as identity of each collaborator.

4.2 Communications between components

Communications among parties within the model need to be secured and restricted. This is to ensure that the privacy of all parties can be protected and extra information will not be concealed. In our model, the communications between parties have been restricted to one-way or two-way communication. Under one-way approach, only a single directly is allowed and no fall-back communication can be performed.

Data miner is one of the collaborators (hospitals) involved in the Telemedicine system. It is the party who makes the query (request) for dataset and will receive the expected results or outputs at the end. The communication between data miner and encryption engine is restricted to one-way communication. Encryption engine is not allowed to response to the data miner in any form.

Encryption engine is responsible to broadcast the encryption key to all collaborators. The connection between the two parties must be established before the queries from data miner can be made. Two-way direct connection will be established. Encryption engine will sends the received encrypted data to the computation engine through the one-way communication. Computation engine is not allowed to communicate with the encryption engine.

At the initial stage, the data miner will sends the decryption key to the computation engine. Once the requested data have been arrived, two-way communications will be established in order for the results to be sent back to the data miner. Further operations (such as data mining process) can be performed directly from the temporary union dataset by the data miner. There is no direct communication between data miner and other collaborators. The communication is established indirectly by using the en-

cryption and computation engine. There is no interaction between collaborators (except data miner) involved in this model. They do not have any communication directly or indirectly via any channel.

5 Discussion

In our model, there are two types of queries made by the data miner. Primary query is sent to the collaborators via encryption engine. The answers for the primary query (Q_p) will be used to form the union dataset ($d_1, d_2, ..., d_n$) in the computation engine. Secondary query (Q_s) will be sent to the union dataset for further operations. Data miner will not gain any extra knowledge despite of the answer to its queries.

Algorithm 1: Primary Query

1: *Input*:
2: Q_p: Primary Query [key(s), dataset]
3: *Output*:
4: *Answer P*: Encrypted (dataset)
5: *Procedure*:
6: **for** i =1; i $\leq$ $|D_a, D_b, ..., D_N|$; i++ **do**
7: Search Q_p from $D_i = \{ D_a, D_b, ..., D_N \}$
8: Add encrypted[i] to P
9: **end** for

Algorithm 2: Secondary Query

1: *Input*:
2: Q_s: Secondary Query [process]
3: *Output*:
4: *Answer S*: Secure sum, Intersection, Union, Equijoin, etc
5: *Procedure*:
7: Search Q_s from $\{ d_1, d_2, ..., d_n \}$
8: return S

Encryption and computation engine used in this model can prevent the identity of the data miner and other collaborators being revealed to each other. The privacy of the shared data will still preserved even thought one of the engines is compromised.

6 Conclusion

By using the proposed model, hospitals within the Telemedicine system do not need to construct a central repository to share their local databases. They only need to answer to the queries made by the data miner and contribute requested dataset into the union dataset. The protocols for database integration such as intersection and equijoin can be applied using this model.

References

1. Agrawal, R., Evfimievski, A., Srikant, R.: Information Sharing across Private Databases. In: 22 nd ACM SIGMOD International Conference on Management of Data, pp. 86-97. ACM Press (2003)
2. Agrawal, R., Terzi, E.: On honesty in sovereign information sharing. In: 10th International Conference on Extending Database Technology, pp.240-256, Germany (2006)
3. Ajmani, S., Morris, R., Liskov, B.: A trusted third-party computation service. Technical Report, MIT-LCS-TR-847, MIT (2001)
4. Clifton, C., Kantarcioglu, M., Doan, A., Schadow, G., Vaidya, J., Elmagarmid, A.K., Suciu, D.: Privacy preserving data integration and sharing. In: 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge, pp. 19-26, Paris (2004)
5. Dayal, U., Hwang, H.Y.: View definition and generalization for database integration in a multidatabase system. vol. 10(6), pp. 628-645. IEEE Transaction (1984)
6. Goldreich, O..: The foundations of cryptography, vol 2. Cambridge University Press, 2004
7. Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game - a completeness theorem for protocols with honest majority. In: 19th ACM Conference on Theory of computing, pp.218-229. ACM Press, (1987)
8. Jefferies, N., Mitchell, C., Walker, M.: A proposed architecture for trusted third party services. In: Cryptography Policy and Algorithms Conference 1995. LNCS, vol 1029, pp98-104. Springer, Verlag (1995)
9. Naor, M., Pinkas, B.: Oblivious transfer and polynomial evaluation. In: 31st ACM Symposium on Theory of Computing, pp.245-254, Atlanta (1999)
10. Stefan, B., Sebastian, O.: Secure set union and bag union computation for guaranteeing anonymity of distrustful participants. Journal of Software, vol 3(1), pp. 9-17. Academy (2008)
11. Wiederhold, G.: Intelligent integration of information. In: ACM SIGMOD Conference on Management of Data, pp. 434-437, Washington (1993)
12. Yao, A.C.: How to generate and exchange secrets. In: 27th Annual Symposium on Foundations of computer science, pp.162-167. IEEE Press (1986)