International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

Implementation of Association Rule on Distributed Database in Data Mining

Srikanth Bethu[1], Suresh Mamidisetti[2], R Aruna Flarence[3]
[1,3]GRIET CSE Department, JNTU Hyderabad, Hyderabad, Telangana, India
[2]Department of Technical Education, Government Polytechnic College, Hyderabad, Telangana, India.

Abstract: The protocol we implemented consists of two way secure multi-party algorithms, in which first one computes the union of private subsets that each of the interacting players hold, and second type that tests the inclusion of an element held by one player in a subset held by another, and offers enhanced privacy with respect to the protocol. This protocol is used to calculate and simplify communication rounds, communication cost and computational cost.
Keywords: Cryptographic, Databases, Data mining, Multiparty computation.

INTRODUCTION

Privacy Preserving Data mining, in which the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold. In the first setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonym zing the data prior to its release. In the first setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonym zing the data prior to its release. The main approach in this context is to apply data perturbation. The idea is that. Computation and communication costs versus the number of transactions *N* the perturbed data can be used to infer general trends in the data, without revealing original record information. In the second setting, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners. This is a problem of secure multiparty computation. The usual approach here is cryptographic rather than probabilistic.
We compared the performance of two secure implementations of the FDM (Fast Distributed Mining Algorithm) algorithm[2] Section In the first implementation (denoted FDM-KC), we executed the unification step using Protocol UNIFI-KC, where the commutative cipher was 1024-bit RSA[6] in the second implementation (denoted FDM) we used our Protocol UNIFI, where the keyed-hash function was HMAC[4]. We tested the two implementations with respect to three measures:

1. Total computation time of the complete protocols (FDMKC and FDM) over all players. That measure includes the Apriori computation time, and the time to identify the globally *s*-frequent item sets, as described in later.
2. Total computation time of the unification protocols only (UNIFI-KC and UNIFI) over all players.
3. Total message size. We ran three experiment sets, where each set tested the dependence of the above measures on a different parameter: • *N* — the number of transactions in the unified database.

LITERATURE SURVEY

Existing System

Kantarcioglu and Clifton studied that problems and devised a protocol for its solution. The main part of the protocol is a sub-protocol for the secure computation of the union of private subsets that are held by the different players. (The private subset of a given player, as we explain below, includes the item sets that are s-frequent in his partial database. That is the most costly part of the protocol and its implementation relies upon cryptographic primitives such as commutative encryption, oblivious transfer, and hash functions. This is also the only part in the protocol in which the players may extract from their view of the protocol information on other databases, beyond what is implied by the final output and their own input. While such leakage of information renders the protocol not perfectly secure, the perimeter of the excess information is explicitly bounded and it is argued there that such information leakage is innocuous, whence acceptable from a practical point of view.

Disadvantages

- Insufficient security, simplicity and efficiency are not well in the databases, not sure in privacy in an existing system.
- While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players.

- Our protocol may leak is less sensitive than the excess information leaked by the protocol.

Proposed System

The protocol that we propose here computes a parameterized family of functions, which we call threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets.
Those are in fact general-purpose protocols that can be used in other contexts as well. Another problem of secure multiparty computation that we solve here as part of our discussion is the set inclusion problem; namely, the problem where Alice holds a private subset of some ground set, and Bob holds an element in the ground set, and they wish to determine whether Bob's element is within Alice's subset, without revealing to either of them information about the other party's input beyond the above described inclusion.

- **Advantages**
- We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency.
- The main ingredient in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players holds.

Different Approaches for recovery

- Privacy-Preserving Graph Algorithms In The Semi-Honest Model

AUTHORS: J. Brickell and V. Shmatikov
We consider scenarios in which two parties, each in possession of a graph, wish to compute some algorithm on their *joint graph* in a privacy-preserving manner, that is, without leaking any information about their inputs except that revealed by the algorithm's output.
Working in the standard secure multi-party computation paradigm, we present new algorithms for privacy-preserving computation of APSD (all pairs shortest distance) and SSSD[5] (single source shortest distance), as well as two new algorithms for privacy-preserving set union. Our algorithms are significantly more efficient than generic constructions. As in previous work on privacy-preserving data mining, we prove that our algorithms are secure provided the participants are "honest, but curious

- A Fast Distributed Algorithm For Mining Association Rules.

AUTHORS: D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu
With the existence of many large transaction databases, the huge amounts of data, the high scalability of distributed systems, and the easy partitioning and distribution of a centralized database, it is important to investigate efficient methods for distributed mining of association rules. The study discloses some interesting relationships between locally large and globally large item sets and proposes an interesting distributed association rule mining algorithm, FDM (fast distributed mining of association rules), which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. A performance study shows that FDM has a superior performance over the direct application of a typical sequential algorithm. Further performance enhancement leads to a few variations of the algorithm

DESIGNING METHODOLOGY ANALYSIS

Figure 1 and Figure 2 explains the Input and Output designing of the Proposed sytem.

Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

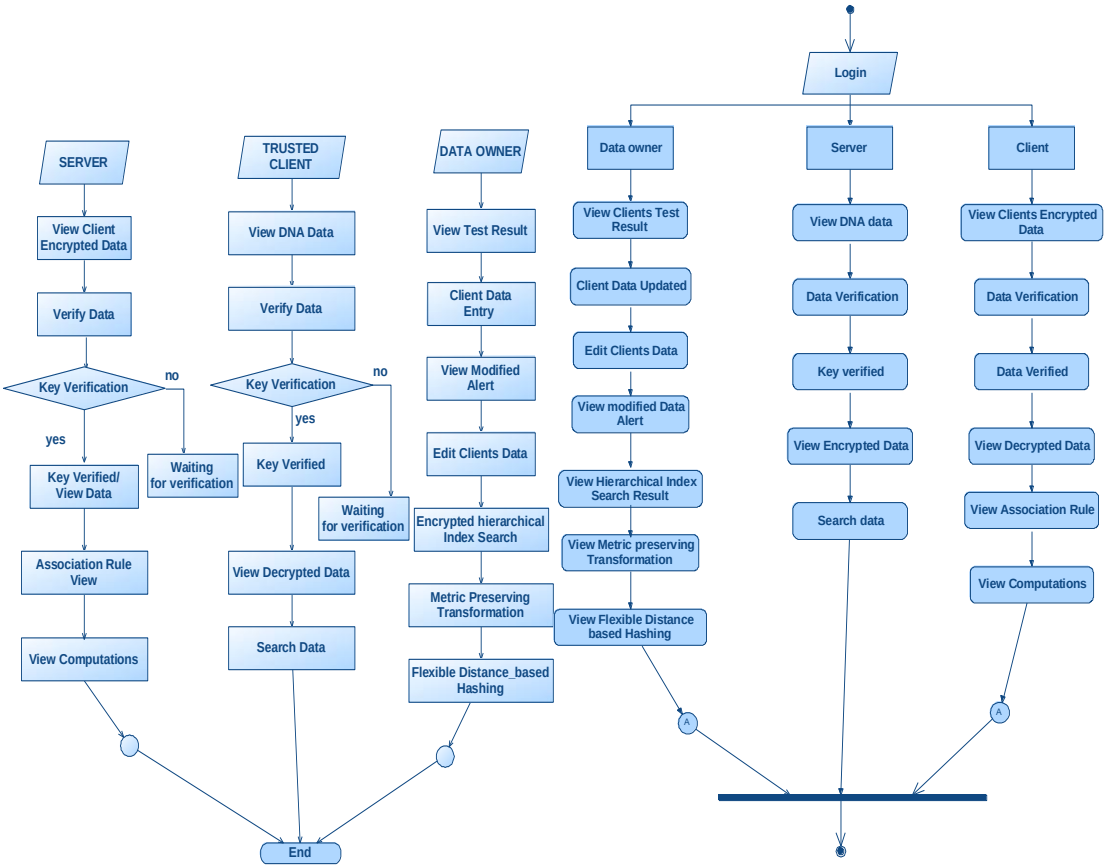

FIGURE 1. Architecture of Two-Way Secure Multi Party Algorithm (Proposed Algorithm)
FIGURE 2. Implementation Process of Proposed Algorithm

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

Select methods for presenting information. Create document, report, or other formats that contain information produced by the system.

- Convey information about past activities, current status or projections of the Future.
- Signal important events, opportunities, problems, or warnings.

- Trigger an action.
- Confirm an action.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k*=1). Sometimes called the *k*-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

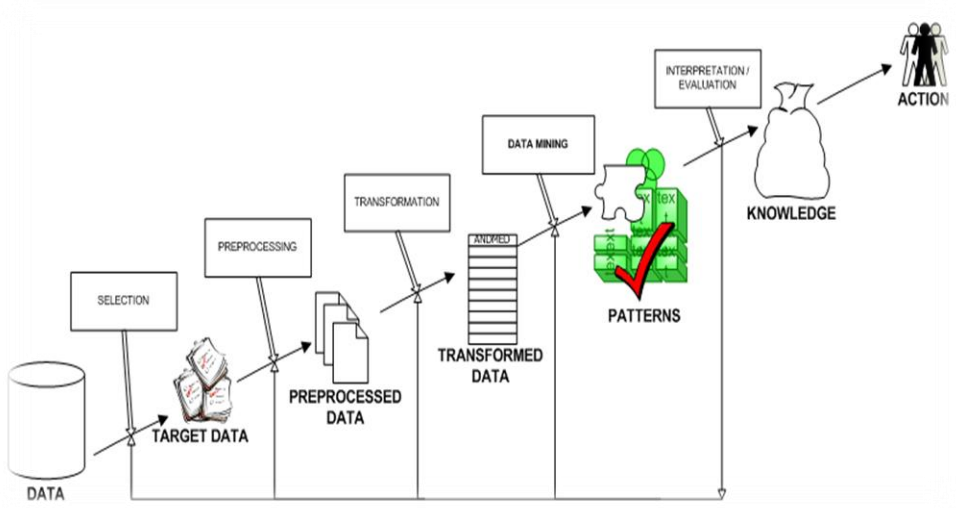

FIGURE 3. Relationship analysis

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

RESULTS ANALYSIS

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

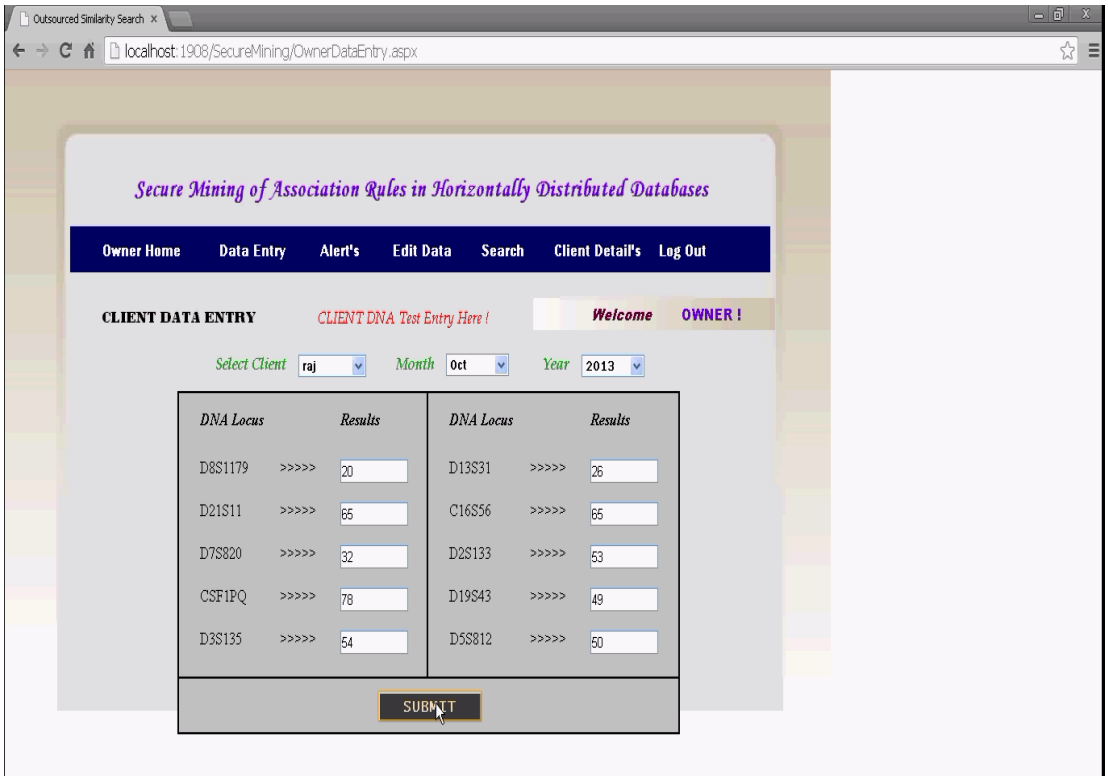

FIGURE 4. Client data test entry page

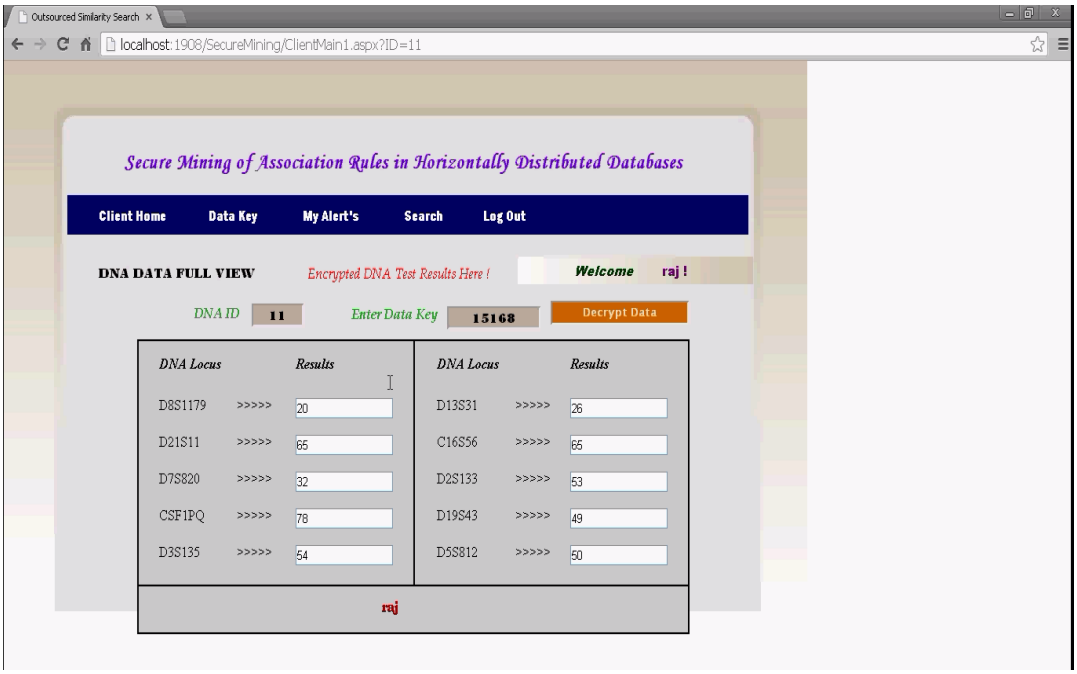

FIGURE 5. Encrypted Client test data

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

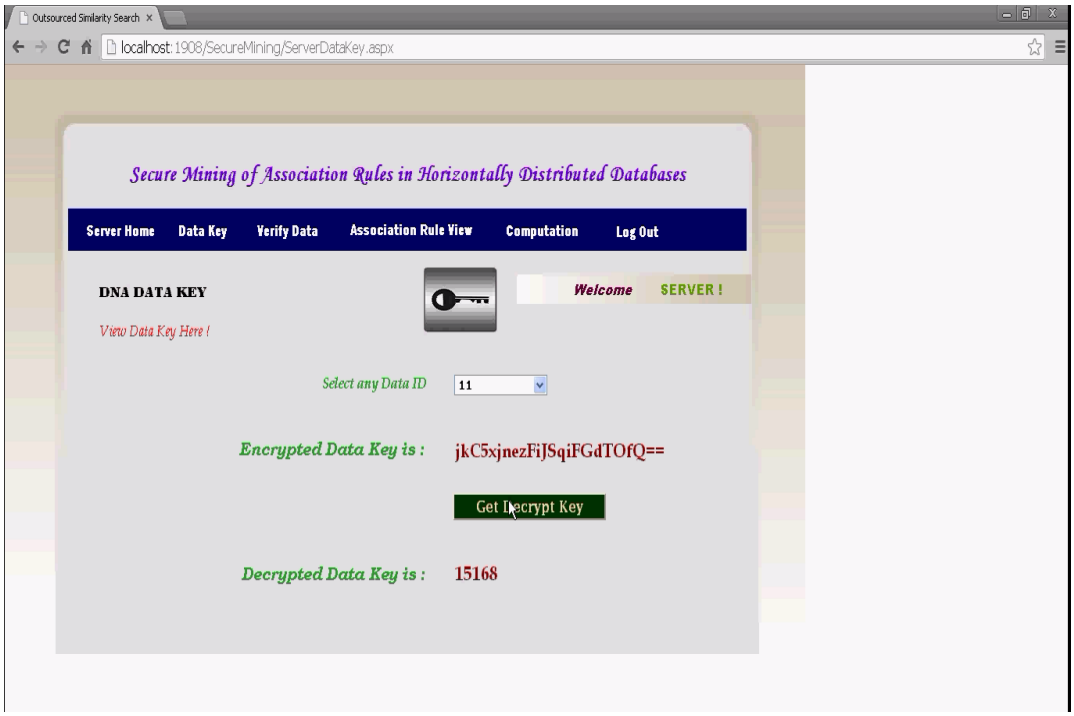

FIGURE 6. Encryption and Decryption

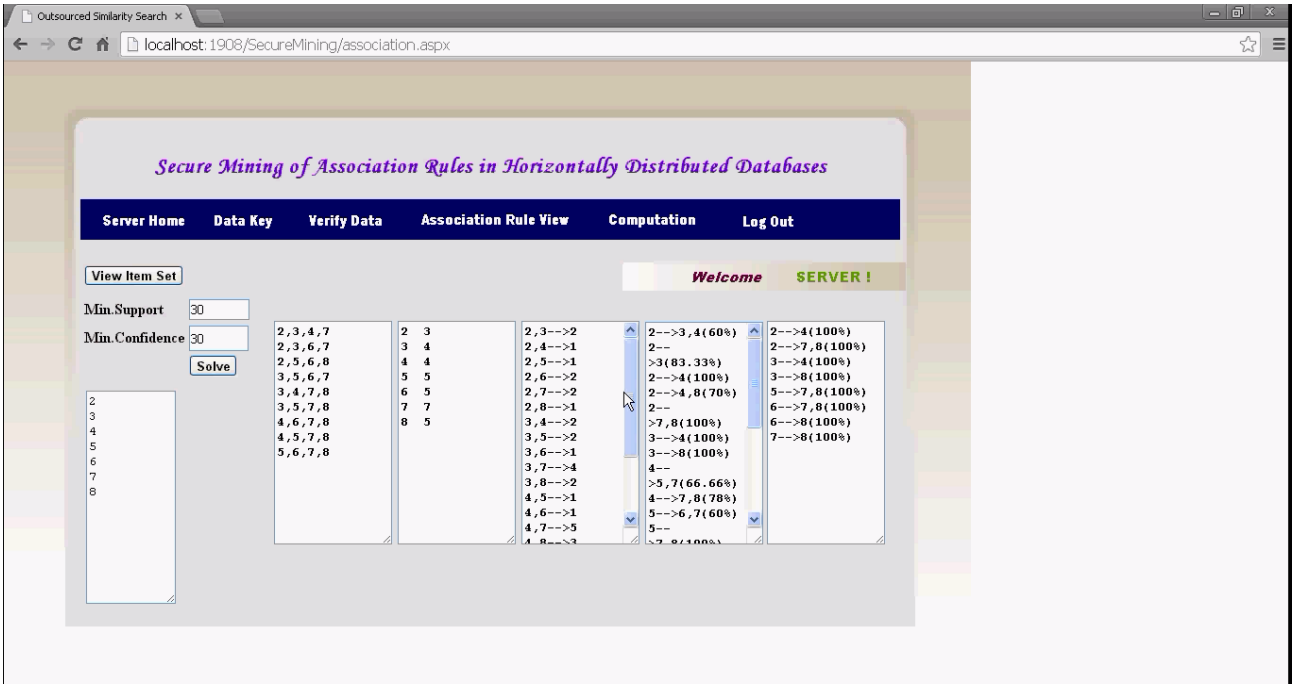

FIGURE 7. Partition set with support and confidence

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

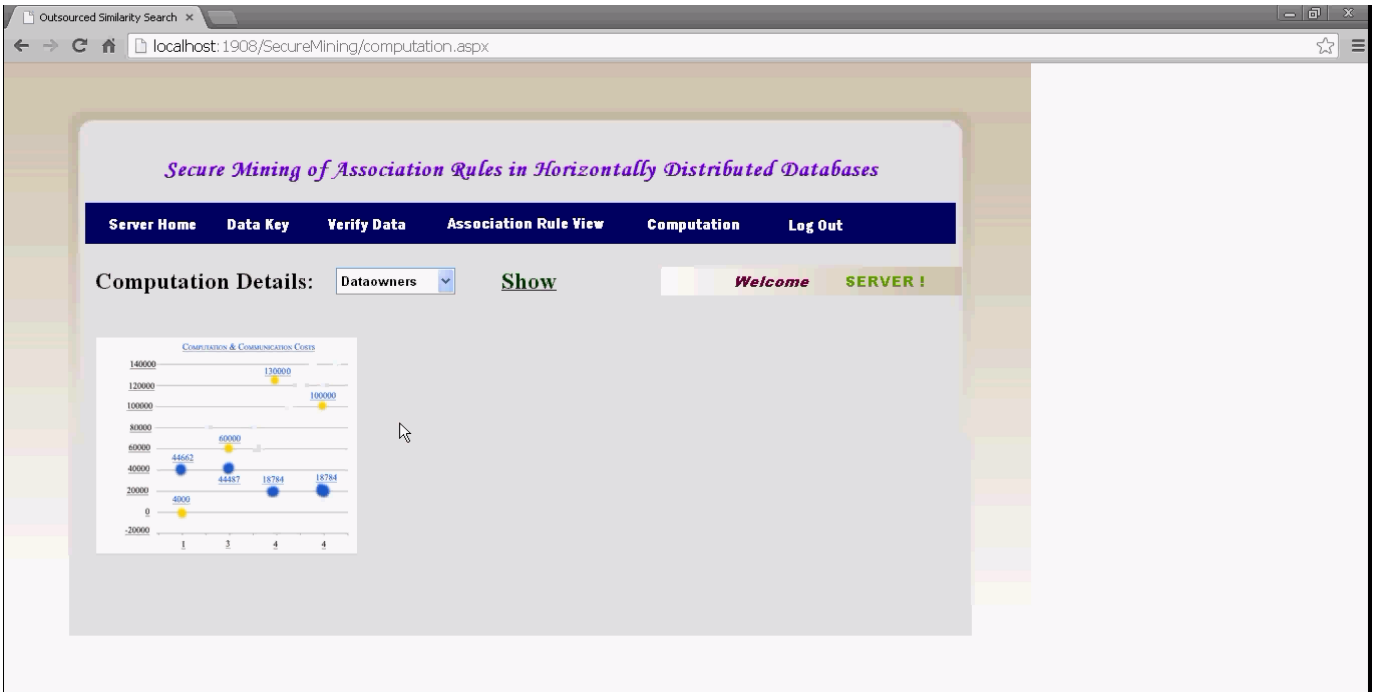

FIGURE 8. Implementation of Computation process

FIGURE 9. Implementation of Computation process

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 176-183 Jan 2018-June 2018 www.csjournals.com

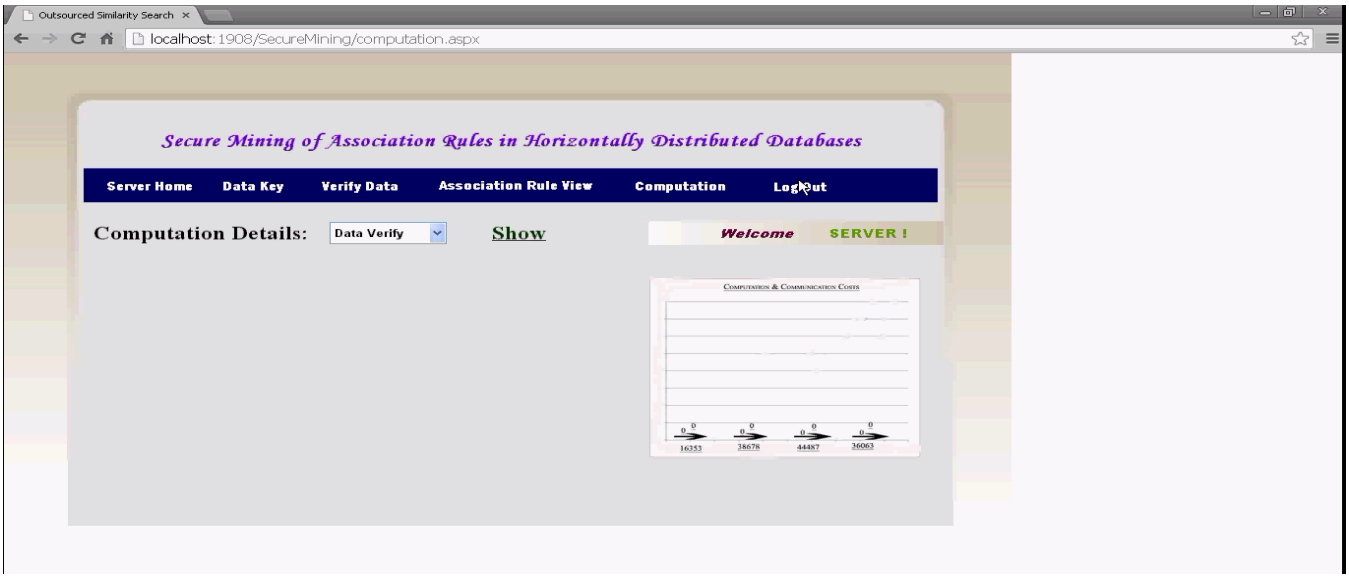

FIGURE 10. Implementation of Computation process

CONCLUSION

We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol [10] in terms of privacy and efficiency. Those protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.
[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.
[3] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In *STOC*, pages 503–513, 1990.
[4] M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication. In *Crypto*, pages 1–15, 1996.
[5] A. Ben-David, N. Nisan, and B. Pinkas. FairplayMP - A system for secure multi-party computation. In *CCS*, pages 257–266, 2008.
[6] J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secret secret. In *Crypto*, pages 251–260, 1986.
[7] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In *ASIACRYPT*, pages 236–252, 2005.
[8] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In *PDIS*, pages 31– 42, 1996.
[9] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. *IEEE Trans. Knowl. Data Eng.*, 8(6):911–922, 1996.
T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theo*