Security in Cloud Computing Based on the Combination of Encryption and Association Rule Mining

[1*]Raghi K.R, [2]Dr. Arjun Paramarthalingam

[1*]Department of Computer Science and Engineering, Anna University Chennai, Tamil Nadu, India,[2]Assistant Professor, Department of Computer Science and Engineering, University College of Engineering,Villupuram, Tamil Nadu, India

***Abstract* — To preserve the privacy of data uploaded on the cloud, it is widely accepted to encrypt the data before uploading it. This leads to the challenge of data analysis, especially association rule mining while protecting data privacy. As one of the solutions, pailler encryption is presented allowing encrypted data processing without decryption. In particular, the twin-cloud structure is frequently applied in the privacy-preserving association rule mining schemes based on asymmetric homomorphic encryption, which contradicts the reality that most of the practical applications applied in the cloud environment. However, the existing related single cloud server schemes suffer from privacy leakage problems. To fill this gap in the literature, in this paper, we first present a universal secure multiplication protocol with the single cloud server using the garbled circuit and additive homomorphic encryption. Based on this multiplication protocol, we construct the inner product protocol, comparison protocol, frequent itemset protocol, and the final association rule mining protocol that is secure against privacy leakage.**

Keywords- Privacy preserving association rule mining; cloud security; encryption

I. INTRODUCTION

Data privacy and security in outsourced data in cloud got more significance in cloud computing applications. Because the outsourced database mayinclude sensitive information, it should be protected against the cloud server. Therefore, the encrypted data before being outsourced to the cloud. It is widely used data mining techniques in the cloud; the association rule mining analyzes the specific data of a company and the association of various information. Mostly, privacy reserving association mining techniques used to support data security [1-3]. Thus, the algorithms depend upon by adding the unnecessary items and it will hide the data frequency in the cloud. At the query processing time, the sensitive information of original data can be analyzed when both the data and the query are encrypted [3].

In this paper, we propose Apriori algorithm based association rule mining algorithm. This algorithm analysis the data generated in cloud environment. The association rule mining is performed over transaction databases [4]. The secure plaintext equality test protocol verifies the input ciphertexts are of identical values. By doing this the proposed approach provides the data security and privacy, while concealing query frequency.

The reminder of the paper is organized as follows. The related literature study associated with privacy-preserving mining techniques is discussed in section 2. The section 3 provides working information about Apriori algorithm based association rule mining algorithm. The experimental study and performance analysis are performed in section 4. The section 5 presented summary about this paper.

II. RELATED WORK

Security of data and privacy-preserving mining technique algorithms based study was performed by various authors [5-6] First, Wong et al [1] proposed a one-to-many item mapping that transforms transactions non-deterministically. However, there is a disadvantage that fake items are easily distinguished from the original data because the probability of fake items in thetransaction database is the same. Second, Giannotti et al. [2] proposed an association rule mining algorithm using *k*- anonymity. This algorithm adds fake transactions to the transaction database so that each item can have *k*-1 frequency. However, the original data can be exposed if faketransaction is known. Also, additional operations are neededto remove the frequency of fake transactions. Similarly, Xun etal. [3] proposed an association rule mining algorithm that supports *k*-anonymity on an encrypted database. This algorithm supports data protection and query protection by using Elgamal encryption system. However, it has an additional overhead for adding encrypted fake transactions.To compute the frequency of candidate set, it uses a conditional gate based on the binary array of ciphertext. However, the original data can be inferred if an attacker hassome knowledge about data frequency because it does not encrypt the data frequency in query processing.

III. PROPOSED ASSOCIATION RULE MINING ALGORITHM

A. System architecture

The typical types of adversaries are semi-honest and malicious [5]. The cloud is consider as the insider attackers, outside attacker have large number of authority than the adversaries. In Semi-honest adversarial model, the cloud follows the rules correctly, then it may try to obtain the more information not allowed. The cloud can deviate from protocol in the malicious model. We adopt a semi-honest adversarial model by following theearlier work [3]. The proposed system architecture is shown in the Figure 1.

The architecture consists of Owner of the data (*DO*), A-Cloud (C_A),B-Cloud (C_B), and Authorized User (*AU*). The database is owned by the data owner, and User is the service recipient who gainsaccesses to the cloud. The two cloud servers with two-party computation protocols perform computations securely on C_A and C_B,

The procedure for building systems is as follows.

Firstly from the original database, an Elgamal encryption key pair is generated. Then the information about public key and encrypted database are forwarded to cloud server C_A. the encryption key pair generated from Elgamal approach is forwarded to the cloud server C_B. At last, the query is encrypted by *AU* which is forwarded to C_A. Because the original data can be exposed using the plaintextequality test protocol [6]. This

privacy preserving plaintext equality test protocol (SPET) checks identity of input encrypted data from cloud servers C_A and C_B using Apriori algorithm.

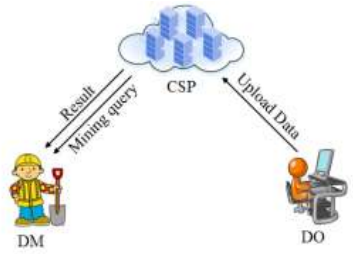

Figure 1. System Model of proposed work

We are committed to achieving privacy-preserving association rule mining with the single-cloud setting in the scenario where a miner submits the mining query to the cloud server that has collected a large set of encrypted transaction records from data owners. As shown in Fig. 1, the system model consists of a cloud service provider (CSP), data owner (DO), and data miner (DM). Specifically, the CSP in our system is considered to be honest but curious, namely the semi-honest model.

• Data Owner (DO): Data owners upload encrypted transaction records to the cloud server to perform association rule mining.

• Data Miner (DM): The data miner intends to mine potentially unknown association rules of the items by outsourcing mining queries to the cloud server.

• Cloud Service Provider (CSP): The cloud server receives the mining query from data miner, performs association rule mining on the encrypted transaction records uploaded by the data owner, and sends the mining result back to the data miner.

B. ATTACK MODEL

We mainly consider external attackers and internal attackers. For external attackers, the active attack method is relatively expensive, while the passive attack cost is extremely low and does not leave traces. Therefore, the main attack method of the external attacker we consider is the passive attack. Besides, internal attackers can collect intermediate results in the computation process. The ability of attacker A is defined as follows:

• A can obtain communication data between all entities by eavesdropping on the public channel.

• A can corrupt the CSP to try to obtain privacy or mining results of DO and DM.

• A can corrupt the DM and the CSP meanwhile to try to obtain the privacy of DO.

Note that, the CSP is not assumed to collude with the DO since this will directly lead to the privacy disclosure of DO. To resist the attacks defined in the above attack models, our protocol is supposed to protect the mining query and mining result privacy of DM and data privacy of DO.

C. Secure protocol

The proposed SPET protocol returns 1 if the plaintexts of two ciphers are equal and returns 0 otherwise. The working of SPET protocol is shown in Algorithm 1. First, C_B generates a composite number t and send E(t) to C_A(line 1~2). Second, C_A multiplies E(t) by E($cipher_1$) and E($cipher_2$), respectively.C_A sends g^{r1} and g^{r2} to C_B, where g^{r1} and g^{r2} represent the front of E($t \times cipher_1$) and that of E($t \times cipher_2$), respectively(line 3~6). Third, C_B returns $g^{r1} \times g^x$ and $g^{r2} \times g^x$, where x is the secret key(line 7~8). Fourth, C_A computes $\alpha = t \times m1 \times g = m1$(line 10). Finally C returns 1 if α

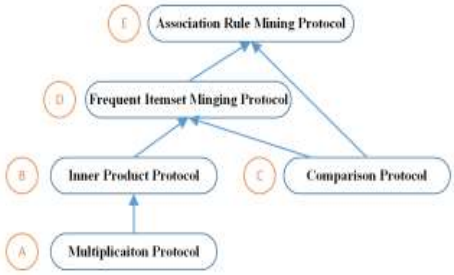

Figure 2. Relations of Proposed Protocols

D. PROTOCOLS

In this section, we construct a privacy-preserving association rule mining protocol under the single-cloud setting by invoking four protocols, namely the multiplication protocol, inner product protocol, comparison protocol, and frequent item set mining protocol, separately. In particular, the proposed multiplication protocol can be regarded as a general construction basing on the garbled circuit and any additive homomorphic encryption system. In this protocol, to facilitate understanding, we instantiate it by applying the Paillier cryptosystem. In the following discussion, we refer the encryption and decryption to the Paillier cryptosystem unless specified otherwise. Besides, to further clarify how the invoked four protocols support the final association rule mining protocol, we illustrate their relations in Fig. 2 .

Next, we show a rough overview of the protocols. Combined with the definition of association rule mining given in Sec. II, to compute the association rule of X =⇒ Y in the Association Rule Mining Protocol, we must compute out supp(X ∪ Y) and supp(X) which needs the join of Frequent Item set Mining Protocol, and to compare the ratio value of them with the given threshold $conf_{min}$ which needs to invoke the Comparison Protocol. Also, according to the definition of frequent item set mining given in Sec. II, to compute the support of a mining query in the transactions, we need to compute out the inner product of this query and each transaction record that requires Inner Product Protocol, and to compare the value with the support threshold $supp_{min}$ that requires Comparison Protocol. As defined, the transaction records and mining queries involved in the computation are represented as binary vectors. By splitting the inner product operation on vectors into multiplications and additions, we have that the Inner Product Protocol can invoke the Multiplication Protocol to compute the inner product value. Suppose that the DM runs the Paillier homomorphic encryption to generate its public/secret key pair (pk, sk) and publishes the publickeypk.

MULTIPLICATION PROTOCOL To clarify the applied symbols, we summarize the symbols used in the multiplication protocol to Table 4. Suppose that the DM owns bit x ∈ {0, 1} and the DO owns bit y ∈ {0, 1}, they interact with each other and the CSP to obtain the privacy-preserving multiplication result of x and y. Since x · y equals to the result of logical AND between x and y, we then combine the AND gate in the garbled circuit and homomorphic encryption to design our multiplication protocol as follows.

• Initialization of DM. With the public key pk generated using

Paillier homomorphic encryption, the DM generates three ciphertexts C1, C2, C3 of message 0 and one ciphertext C4 of message 1 as [0] = {C1, C2, C3}, [1] = C4, picks random numbers k_{a^0}, k_{a^1}, k_{b^0}, $k_{b^1} \in Z_{N^2}$

representing its and DO's choises of 0 and 1 respectively, and chooses a random secret number rA ∈ ZN and a public cryptographic hash function H : {0, 1} ∗ → {0, 1} N . The DM computes {C 0 i = $C_i \times r^N{}_A \bmod N_2$}i∈{1,2,3,4}.

It then generates the secret truth table as Table 5 and shuffles the order of the four secret values to obtain the disordered secret truth values T1, T2, T3, T4.

DM→DO. With the x ∈ {0, 1}, the DM selects k x a , computes {T 0 i = k x a ⊕ Ti}i=1,2,3,4, and sends k 0 b , k1 b , {T 0 i , C0 i }i=1,2,3,4 to the DO.

• DO→CSP. With the y ∈ {0, 1}, the DO chooses k y b , computes {k y b ⊕ T 0 i }i=1,2,3,4, and randomly shuffles the order of results to obtain {T 00 i }i∈{1,2,3,4}. It then sends {T 00 i }i∈{1,2,3,4} to CSP.

• CSP→DM. The CSP randomly chooses a number R ∈ ZN , generates four ciphertexts of R as [R] = CR,1, CR,2, CR,3, CR,4, and computes C 00 i = C 0 i × CR,i mod N 2 hi = H(C 00 i) T¯ i = T 00 i × CR,i mod N 2 i∈{1,2,3,4}

Algorithm 1. Secure plaintext equality test protocol

Input: E(*cipher*$_1$), E(*cipher*$_2$)

Output: if *cipher*$_1$= *cipher*$_2$ return $\alpha = 1$ else $\alpha = 0$

C$_B$

01: generate *t*(*t* is composite number)02: send E(*t*) to *C*$_A$

03: for(*i* = 0 to *k*)

04: for(*j* = 0 to *y.NumItem*) 05: r_1

04: E(*cipher*$_1$)*E(*t*) = (g^{r1x} , $t \times m_1 g^{r_2 x}$)

05: E(*cipher*$_2$)*E(*t*) = (g 06: , $t \times m_2 g$) send to g^{r1}, g^{r2} to *C*$_B$

07: calculate g^{r1x}, g^{r2x} ($x = secret$ ke)

08: send to g^{r1x}, g^{r2x} to *C*$_A$

09: receive g^{r1x}, g^{r2x} from *C*$_B$

10: calculate $\alpha = \frac{t \times m1g^{r1}}{t \times m_2 g^{r2x} g^{r1x}} \times g^{r2x}$

11: if α ==1, then return result = 1

12: else return result = 0

For association rule mining, we propose a privacy- preserving Apriori algorithm by using SPET protocol in cloud computing. The proposed algorithm consists of candidate set generation and frequency set calculation.

F. Candidate set generation

The candidate set generation step generates a candidate set containing many patterns, each of which has multiple items. The procedure of the candidate set generation step is as follows. First, one pattern pair <p_1, p_2> is selected in the *k*-1 frequent set, where p_1 and p_2 are different patterns.

Second, we perform a join operation between p_1's items and p_2's items, and insert the joined result into the candidate set, i.e., S_k, if the result consists of *k* items. Finally, we perform a join operation for all pairs except <p_1, p_2> and return S_k to C_A.

E. Frequent set calculation

The frequent set calculation step calculates the frequency of S_k, as shown in Algorithm 2. First, one pattern of S_k is selected (line 1~2). Second, the SPET protocol is performed between the items of the selected pattern and the items of the transaction. If the result of SPET protocol is 1, the number of the matched items(*match*) is incremented by 1 (line 3~8). Third, when *match* is equal to *k*, E(*x.sup*) is multiplied by *g*, where *g* is an arbitrary integer that is not included in a cyclic group of the encryption key (line 9~10). Fourth, the SPET protocol is performed between E(*x.sup*) and E(g^{minsup}). If the result of SPET is 1, the frequent attribute of x is included in the frequent set (line 11~14). Finally, the frequent set calculation for the remaining patterns of S_k is performed.

Algorithm 2. Frequent set calculation

Input: Candidate *k*-item set S_k

Output: Frequent set L_k

01: for(all $x \in S_k$)

02: for(all $y \in$ E(*T*))

03:match=

if(SPET(x_i, y_j)) *match*++

06: end for

07: end for

08: if(*match*==*k*){

19: enc_mul(*x.sup*, *g*)

10: if(SPET(*x.sup*, E(g^{minsup})) *x.freq* = true }11:

end for

12:

14: end for

15: return L_k

G. Proposed privacy-preserving mining *Apriori algorithm*

if(x.freq == true) LkUxThe proposed Apriori algorithm is shown in Algorithm 3.First, we set L_1 to 1-item sets which are received from the data owner (line 1). Second, we perform the candidate set generation algorithm of 4.1, called Candidate_set_generation (E(L_{k-1})), where E(L_{k-1}) represents the *k*-1 frequent set (line 4). Third, the frequency of S_k is calculated (line 6). Finally, if the *k* frequent set is no longer generated, the *k-1* frequent set is returned (line 5)

Algorithm 3. Proposed Apriori Algorithm

Input: Encrypted transaction database E(*T*)

Item set length *k*

Candidate pattern set S_k

Output: Frequency pattern set L_{k-1}

01: $L_1 = \{l_1,\ldots,l_n \mid *E(T)\}$

02: *k* = 2

03: while(TRUE)

04: E(S_k) = Candidate_set_generation(E(L_{k-1})) ={$c_1,\ldots,c_p | c \in k$ candidate set}05: if(E(S_k)=Ø) return E(L_{k-1}) to AU

06: E(L_k)=Frequent set calculation(E(T),E(S_k))

07:end While

IV. EXPERIMENTAL STUDY & PERFORMANCE ANALYSIS

The data security of the algorithm. Is the viewpoint of CA, the proposed algorithm encrypts the data frequency and the encrypted database consists of unidentifiable encrypted transactions. Because the Elgamal cryptosystem returns different ciphertexts for the same plaintext, there is no leakage of the original data. In the viewpoint of CB, the data cannot be exposed because the front of the ciphertext is not contained in the original data. Therefore, the proposed Apriori algorithm proves that it safe the data in semi model.

We evaluate the performance of the proposed Apriori algorithm, called S-ARM (Secure Association Rule Mining). The performance analysis was done under Intel Xeon E3-1220v3 3.10GHz, 32GB RAM. The proposed algorithm uses GMP library to represent a big integer in an Elgamal cryptosystem. The proposed algorithm is compared with the DP-ARM (Data Privacy Association Rule Mining) algorithm proposed by Xun et al.[3] because DP-ARM is the only existing algorithm to support both data privacy and query privacy. For performance analysis, we use the retail dataset collected from the Belgian market [7], and the performance measure of S-ARM and DP-ARM by varying the number of data. We also measure their performances by varying support changes (*minsup*) from 5% to 30% of data. Table 1 shows parameters for our performance analysis.

TABLE I. PARAMETERS FOR PERFORMANCE ANALYSIS

The Size of dataset	2k, 4k, 6k, 8k, 10k
Fake transaction ratio(Q)	50%, 100%
Minimum support	5%, 10%, 15%, 20%, 25%, 30%
Key Size	1024

A. Performance analysis varying the number of data

The performance result of S-ARM and DP-ARM by varying the number of data is shown in Figure 3. When *minsup* is 10% and Q is 50%, S-ARM shows 205% performance improvement on the average, compared with DP-ARM, and when Q is 100%., S-ARM shows 405% performance improvement. The reason is why S-ARM requires no additional operation for fake transactions unlikeDP-ARM. In addition, S-ARM requires no binary operationby using Elgamal cryptosystem through SPET protocol.

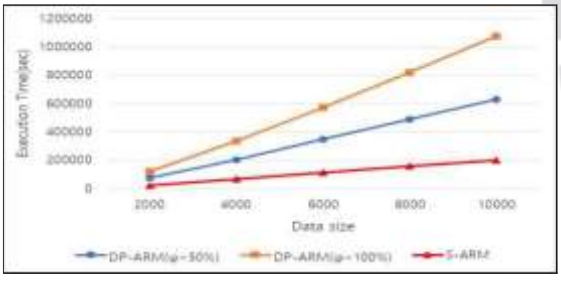

Figure 3. Performance result by varying the number of data

B. Performance analysis varying minsup

The performance result of S-ARM and DP-ARM according to *minsup* is shown in Figure 4. When the number

of data is 10,000 and Q is 50%, S-ARM shows 216% performance improvement on the average, compared with DP-ARM, and when Q is 100%, S-ARM shows 429% performance improvement on the average. The reason is why S-ARM does not require any additional operation for the faketransactions unlike DP-ARM.

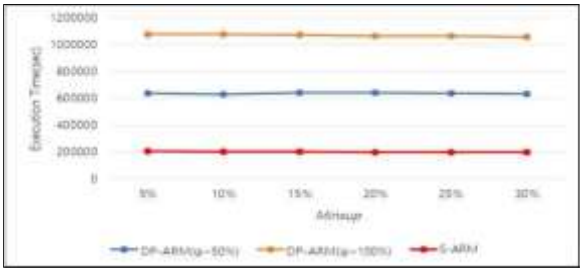

Figure 4. Performance result by varying minsup

V CONCLUSIONS AND FUTURE WORK

In this paper, we propose a privacy-preserving association rule mining protocol on encrypted cloud data under a single cloud server. To realize the security under the single-cloud setting, we construct the multiplication protocol which is the base of association rule mining protocol using garbled circuit and additive homomorphic encryption system. Compared with the state-of-art works, our protocol reduces the number of cloud servers and enhances the flexibility of cloud outsourcing association rule mining. In future work, we will focus on further improving the efficiency of our scheme.

REFERENCES

[1] Wong, Wai Kit, et al. "Security in outsourcing of association rule mining." Proceedings of the 33rd International conference on Very large data bases. VLDB Endowment, 2007.

[2] Giannotti, Fosca, et al. "Privacy-preserving mining of association rules from outsourced transaction databases." IEEE Systems Journal 7.3 (2013): 385-395.

[3] Yi, Xun, et al. "Privacy-preserving association rule mining in cloud computing." Proceedings of the 10th ACM symposium on information, computer and communications security. ACM, 2015.

[4] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th International conference on Very large data bases, VLDB. Vol. 1215. 1994.

[5] Kim, Hyeong-Jin, Hyeong-Il Kim, and Jae-Woo Chang. "A Privacy- Preserving kNN Classification Algorithm Using Yao's Garbled Circuit on Cloud Computing.", 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), 2017.

[6] Jakobsson, Markus, and Ari Juels. "Mix and match: Secure function evaluation via ciphertexts." International Conference on the Theory and Application of Cryptology and Information Security. Springer, Berlin, Heidelberg, 2000.

[7] Brijs, Tom. "Retail market basket data set." Workshop on Frequent Itemset Mining Implementations (FIMI'03). 2003.

[8] [12] H. Kim, H. Kim, and J. Chang, "A privacy-preserving knn classification algorithm using yao's garbled circuit on cloud computing," in 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), Honolulu, HI,

USA, June 25-30, 2017, G. C. Fox, Ed. IEEE Computer Society, 2017, pp. 766–769.

[9] L. Li, R. Lu, K. R. Choo, A. Datta, and J. Shao, "Privacy-preservingoutsourced association rule mining on vertically partitioned databases, "IEEE Trans. Inf. Forensics Secur., vol. 11, no. 8, pp. 1847–1861, 2016.

[10] J. Wu, N. Mu, X. Lei, J. Le, D. Zhang, and X. Liao, "Secedmo: Enabling efficient data mining with strong privacy protection in cloud computing," IEEE Trans. Cloud Comput., 2019.

[11] S. Patil and S. Joshi, "Improved privacy preservation of personal health records via tokenization," International Journal of Pure and Applied Mathematics, vol. 118, no. 18, pp. 3035–3045, 2018.

[12] S. Patil and S. Joshi, "Demystifying user data privacy in the world of iot," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 10, pp. 4412–4418, 2019.

[13] S. Patil, S. Joshi, and D. Patil, "Enhanced privacy preservation using anonymization in iot-enabled smart homes," Smart Intelligent Computing and Applications, pp. 439–454, 2020.