International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 23 (2018) pp. 16220-16225

Quadruple Cycle Secured Multiparty Sum Computation Protocol (QCSMC) for Privacy Preserving in Distributed Data Mining

S.Shunmugam[1] and Dr. R.K.Selvakumar[2]

[1]*Research Scholar, Center for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.*

[2]*Professor, Department. of Computer Science& Engineering, CVR College of Engineering, Hyderabad, India.*

Abstract

On the computer era, every organisation is flooded with enormous volume of data and efforts are being taken presently to extract useful hidden information in the common interest of the organisations. Since the availability of data are distributed geographically at different locations, mining of data in distributed environment without affecting the privacy of data of the parties is a big task. To overcome this cryptographic technique are used and Secure Multiparty computation is an important technique which allows different parties to combine the results of their individual data without sharing the data to others. Many protocols have evolved in SMC and we proposed a novel Quadruple Cycle Secured Multiparty Computation (QCSMC)protocol which involves Virtual Coordinator(VC). Our proposed architecture looks like Train Token Ring and computation involves the concept of Lagrange's Four-Square Theorem which is ensured zero leakage of data privacy of the parties and reduced communication and computation complexity.

Keywords: Secure multiparty computation, Four square theorems, Train token ring, Virtual coordinator

I. INTRODUCTION

With the advent of computers, most of the organisations have fully computerized and due to that enormous data are available in servers/data warehouses. These organisations are using data mining to extract hidden useful predictions/information for the development of the organisations. They have their data in different locations connected by network or different entities have common interest to explore the mining information for them is known as distributed data mining.

In the distributed data mining, the main concern is preserving the privacy of the data of the parties in the combined data mining. Though entities wish to extract mining hidden valuable information, hesitate to share the data due to privacy concerns. Since sharing of data may result in leakage of sensitive data such as customer shopping trends, patient's health condition which Is to be considered as violation. Even in some countries strict laws are in force against disclosure of personal information.

Despite all these limitations, distributed data mining is prominent in the last two decades. Cryptographic techniques play a vital role in privacy preservation in distributed data mining (PPDDM) [13]. PPDDM involves either trusted third party or Secure Multiparty Computation.

The parties in PPDDM are classified into three categories. They are

1. Honest party
2. Semi honest party
3. Corrupted/malicious party

Honest party is one who genuinely shares the data and not involved in any unethical activity during the mining. Semi honest party is one who obeys the rules during mining but having intentions/showing interest to know about other party information. Corrupted or malicious party is one who does not obey the rue and/ or doing unethical activities during the mining process. Hence PPDDM has to take care of honest party from getting desired result without though presence of semi honest and malicious party. The present work in PPDDM assumes the presence of semi honest and intruders in the network between the parties.

II. RELATED WORKS

Secure multi-party computation (SMC) is a subfield of cryptography. The main objective of SMC methods is to ensure the parties to perform collaborative computation with their inputs in such a way that each party in the group has to obtain the combined computation result and no party know about the private inputs of other parties [2] [10]. SMC has its origin in 1982 to solve the famous Two Millionaires Problem [7]. The Millionaires problem is stated as two millionaires wish to know who is richer but without disclosing their net worth [12]. A theoretical solution was suggested by Andrew C. Yao in his paper [1] in 1982. The solution provided by Yao was for semi honest. Since then SMC evolved rapidly with Multiparty and ensuring zero privacy leakage.

In 2009, Rashid Sheik et al. [4] proposed k-secure sum protocol in which data is segmented into k divisions (k-number of parties) and parties communicates to each other but possibility of collusion between neighbours was evident and in 2010 with little modifications they proposed modified ck-secure sum protocol[5]wherein data is segmented into k divisions(k-number of parties) and parties communicates to each other and changes their own position at each step. Then they proposed a Modified ck-sum protocol [6] wherein data was segmented into k divisions and data segments were

distributed Before computation has better privacy preservation but communication complexity was very high.

Ms.Priyanka Jangade et al, [3] proposed a hybrid secure sum protocol which involves Trusted Third party, random numbers and data was segmented into 3 segments, This protocol was having less probability of data leakage and high communication complexity, Jyotirmayee Rautaray [8] come out with Distributed RK- Secure Sum Protocol in 2013 used bus topology and N-1 number of rounds and each party changed their position in each round, has high communication complexity,

In 2015, Israt Jahan et al.,[7] suggested Double Random Partitioned Model which involves trusted Third party only for transferring of data and random array size of each party was secret has comparatively low communication complexity. In 2015 Selva Ratna et al., proposed Two Phase Secured Multiparty Sum Computation Protocol (2PSMC) [11, 14] which has two cycles and each party break the data into two segments and random number at each party is used for encrypting, sites are arranged randomly and random number is subtracted in the second phase and the computation as well as communication complexity is low.

III. PROPOSED QCSMC PROTOCOL

In the proposed QCSMC protocol for segmenting data into four numbers with the concept of Four Square Theorem:

A. Lagrange's Four-Square Theorem:

Every positive integer can be expressed as a sum of four squares. Using this theorem, the data can be segmented into four numbers with the help of the following pseudo algorithms;

```
Algorithm(Iterative) – Lagrange Four Square
int[] FindFourSegmentation(int D)
{    int D1,D2,D3,D4;
     int M = square_ root (D);
         for (int p=0;i<=M; p++)
           for (int q=0;j<=M; q++)
             for (int r=0;k<=M; r++)
                for (int s=0;l<=M;s++)
                   if (p*p+ q*q+ r*r+ s*s==D){
                        print("found");
                       break;}
      D1 = p*p;
      D2 = q*q;
      D3 = r*r;
      D4 = s*s;
      return ([D1, D2, D3, D4]);
}
```

B. Train Token Ring Architecture:

The QCSMC protocol involves a Virtual Coordinator (VC) and not Trusted Third party for computation of SMC. The proposed algorithm QCSMC runs in Quadruple / Four cycles. Each site breaks the data block into four segments using the above-mentioned algorithm and shuffle the same. Also, each site generates a random number Ri which will be used for encrypting the sum at each site. The four cycles initiated with virtual coordinator and at the end of four cycles, VC obtains the required sum.

In railway signalling especially in India, the Train Token ring (A large one made of metal or bamboo with a small leather pouch attached to it. A written communication will be placed in the leather pouch.) Is a token — a physical object which a locomotive driver is required to have or see before entering onto a particular section of single track. Since the proposed Architecture resembles the QCSMC Architecture is named as Train Token Ring Architecture.

Fig 1. Train Token Ring Architecture

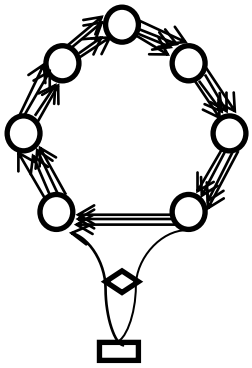

Fig 2 .QCSMC

3.1 First Cycle of QCSMC

Let us consider the number of sites in the PPDDM is N and N≥2.The data in each site D_i, $1 \le i \le N$ is partitioned into 4 segments as D_{ij}, $1 \le j \le 4$ using Four square theorem algorithm.

The VC initiates the computation of SMC and generates N random numbers R_i for each site and assigns the sign '+' or '-'for each site for random numbers as suggested in paper [9]. The random aggregator which is present in the VC sums the sign random numbers and kept aside the $\sum R_i$ as R.

In the first cycle, at site 1 the Value V_{11} is computed as

$$V_{11} = R_1 + D_{11} \quad (1)$$

And at each site for $2 \le i \le N$, $j = 1$

The partial sum V_{i1} is computed using the equation (2) as follows:

$$V_{ij} = V_{(i-1)\,j} + R_i + D_{i\,j} \quad (2)$$

After the partial sum computation at first cycle V_{N1} at the site N, then

$$V_{N1} = V_1 \quad (3)$$

3.2 Second Cycle of QCSMC

The second cycle continues after the partial sum obtained at Site N of first cycle and goes to site 1. In the second cycle, each site assigns an integer S_i (positive or negative). At site 1

$$V_{12} = V_1 + S_1 + D_{12} \quad (4)$$

And at each site for $2 \le i \le N$, $j = 2$

$$V_{ij} = V_{(i-1)j} + S_i + D_{ij} \quad (5)$$

After the partial sum computation at second cycle V_{N2} at the site N, then

$$V_{N2} = V_2 \quad (6)$$

3.3 Third Cycle of QCSMC

The Third cycle continues after the partial sum obtained at Site N of second cycle and goes to site 1. In the third cycle, at each site the assigned number is subtracted. At site 1

$$V_{13} = V_2 - S_1 + D_{13} \quad (7)$$

And at each site for $2 \le i \le N$, $j = 3$

$$V_{ij} = V_{(i-1)j} - S_i + D_{ij} \quad (8)$$

After the partial sum computation at third cycle V_{N3} at the site N, then

$$V_{N3} = V_3 \quad (9)$$

3.4 Fourth Cycle of QCSMC

The fourth cycle continues after the partial sum obtained at Site N of third cycle and goes to site 1. In the fourth cycle, at site 1

$$V_{14} = V_3 + D_{14} \quad (10)$$

And at each site for $2 \le i \le N$, $j = 4$

$$V_{ij} = V_{(i-1)j} + D_{ij} \quad (11)$$

After the partial sum computation at fourth cycle V_{N4} at the site N, then

$$V_{N4} = V_4 \quad (12)$$

After fourth cycle site N, the process reaches the VC where the Value R stored in the Random aggregator is subtracted from V_4 to arrive at the resultant secure sum D by the VC.

$$D = V_4 - R \quad (13)$$

Thus the frequency of a particular item set is computed in semi honest distributed environment with the result obtained at each site without affecting the privacy of data at other sites using the above equations (1), (2),…(13), ensures zero leakage of privacy of data between the sites.

IV. ALGORITHM FOR QCSMC PROTOCOL

1. *At each site data is split into 4 segments D_{ij} using "Four Square Algorithm"*
2. *VC initiates with generation of random number R_i and sign for each site*
3. *Random aggregator computes the sum of random numbers with sign and kept aside the value R*
4. *Calculate $V_{11} = R_1 + D_{11}$*
5. *for $2 \le i \le N$, $j = 1$,*
   Calculate $V_{ij} = V_{(i-1)j} + R_i + D_{ij}$
   //partial sum at the end of first cycle
6. *$V_1 = V_{N1}$*
   //In the second cycle an integer S_i (positive or //negative) is added by each site
7. *Calculate $V_{12} = V_1 + S_1 + D_{12}$*
8. *for $2 \le i \le N$, $j = 2$,*
   Calculate $V_{ij} = V_{(i-1)j} + S_i + D_{ij}$
   //partial sum at the end of second cycle
9. *$V_2 = V_{N2}$*
   //In the third cycle S_i is subtracted by each site
10. *Calculate $V_{13} = V_2 - S_1 + D_{13}$*
11. *for $2 \le i \le N$, $j = 3$,*
    Calculate $V_{ij} = V_{(i-1)j} - S_i + D_{ij}$
    //Partial sum at the end of third cycle
    $V_3 = V_{N3}$
12. *Calculate $V_{14} = V_3 + D_{14}$*
13. *for $2 \le i \le N$, $j = 4$, Calculate $V_{ij} = V_{(i-1)j} + D_{ij}$*
14. *//Partial sum at the end of third cycle*
    $V_4 = V_{N4}$
15. *//At VC ,resultant secure sum D is*
    $D = V_4 - R$

V. DEMONSTRATION OF QCSMC PROTOCOL

Let us consider 5 sites involved in PPDDM A, B, C, D, E holding values 53, 24, 45, 38, 29. The values are split into

four squares each and the Random numbers and site assigned numbers are shown in Table 5.1.

Table 5.1 QCSMC Illustration

Sites	**A**	**B**	**C**	**D**	**E**
Value	53	24	45	38	29
Four squares	36 0 16 1	4 0 4 16	0 25 4 16	9 0 25 4	16 4 9 0
Random Number	+61	-41	-23	+72	+29
Site Assigned Number	-40	33	-21	-8	19

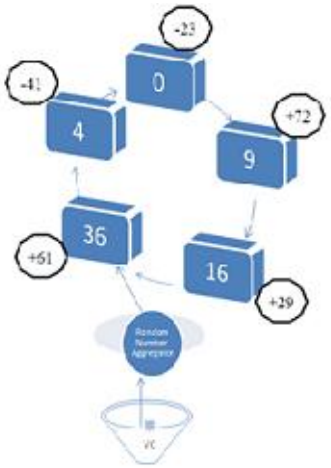

Fig 3. Cycle 1

At the end of Cycle 1, Partial Sum is
36+61+4-41+0-23+9+72+16+29 =163

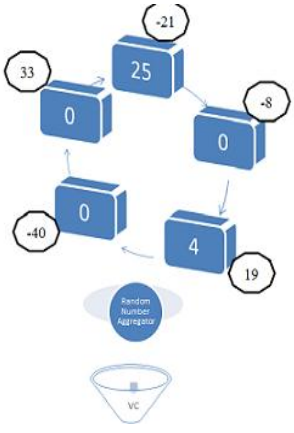

Fig 4. Cycle 2

At the end of Cycle 2, Partial Sum is
163+0-40+0+33+25-21+0-8+4+19 =175

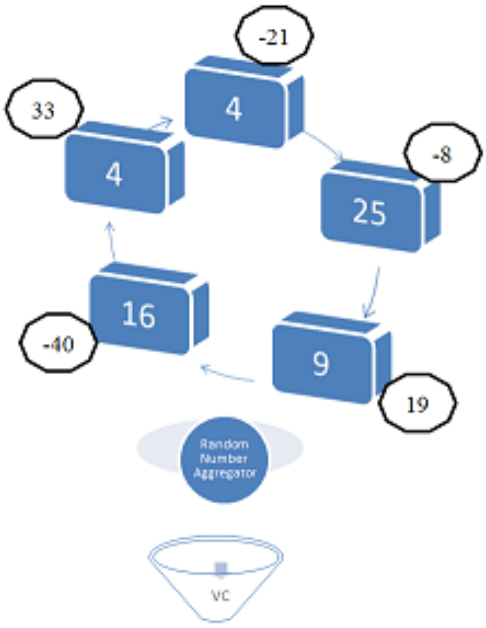

Fig 5. Cycle 3

At the end of Cycle 3, Partial Sum is
175+16-(-40) +4-33+4-(-21) +25-(-8) +9-19 =250

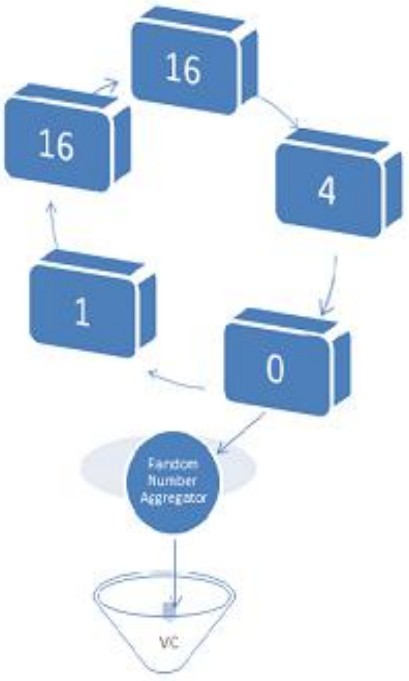

Fig 6. Cycle 4

At the end of Cycle 4, Partial Sum is
250+1+16+16+4+0 =287

At VC, resultant Partial Sum is
287-98 =189

VI. PERFORMANCE ANALYSIS OF QCSMC PROTOCOL

In the proposed QCSMC protocol, at each site the data is split into four segments using the four-square number generators and the same is shuffled once. Even the sites do not know about the four segments. At VC the random numbers for each site is generated and assigns sign for each site. Hence in the first cycle neither the site nor the VC know the value added to first segment at each site. In the second cycle positive or negative integer is added to second segment. In the third cycle the sites subtract the assigned integer with the third segment value at sites. In the fourth cycle, only fourth segment at each site is added. At the end of fourth cycle the Random sum in the Aggregator is subtracted to get the required secured sum.

The main advantages of the protocol are

a) The Trusted third party is not involved and VC plays the role of coordinator only and VC is not having chance to know about the data of the sites neither partially nor fully.

b) The collusion between neighbours is completely avoided by splitting the data into four segments and involving Random number and assignment of integer at each site.

c) The sites are not interchanging its position in each cycle minimises the computation and communication complexity.

d) Guarantees zero leakage of privacy of data by the sites and collusion between the neighbours.

e) Since the number of cycles is 4 and the number of computation is N at each cycle only, both the computation and communication complexity between sites themselves is O (n) only. But communication complexity with respect to VC by sites is O (1) only. Hence the computation and communication complexity of the QCSMC protocol is considered as very low compared to the existing protocols.

f) For this protocol, the number of sites can be greater than or equal to 2.

VII. CONCLUSION

In this paper, we have proposed a novel SMC protocol which ensures zero leakage of privacy of data with low computation and communication complexity. The computation involves only Virtual coordinator and third party which the parties in the PPDDM may prefer for data mining computation. Hence, this protocol in many ways better than the other existing protocols. In the future, the protocol will be further strengthened to have minimal computation and communication complexity for using in PPDDM.

REFERENCES

[1] A.C. Yao, "Protocol for secure computations," in proceedings of the 23rd annual IEEE symposium on foundation of computer science, pp. 160-164, Nov. 1982

[2] Chris Clifton, Murat Kantarcioglu and Xiaodong Lin, Michael Y. Zhu," Tools for Privacy Preserving Distributed Data Mining", ACM New York, NY, USA, ISSN: 1931-0145 EISSN: 1931-0153, Volume 4 Issue 2, December 2002.

[3] Ms. Priyanka Jangde, Mr. Gajendra Singh Chandel and D. K. Mishra, "Hybrid Technique For Secure Sum Protocol," World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741, Vol. 1, No. 5, pp. 198-201, 2011.

[4] R. Sheikh, B. Kumar and D. K. Mishra, "Privacy - Preserving k- Secure Sum Protocol," in the International Journal of Computer Science and Information Security, USA, Vol. 6, No. 2, pp. 184-188, Nov. 2009.

[5] R. Sheikh, B. Kumar and D. K. Mishra, "Changing Neighbors k-Secure Sum Protocol for Secure Multi-Party Computation," in International Journal of Computer Science and Information Security, Vol. 7, No. 1, pp. 239-243, USA, Jan. 2010.

[6] R. Sheikh, B. Kumar and D. K. Mishra, "A Modified ck-Sum Protocol for Multi-Party Computation," Journal of Computing, USA, Vol. 2, Issue 2, pp. 62-66, Feb. 2010.

[7] Israt Jahan, Nure Naushin Sharmy et al," Design of a Secure Sum Protocol using Trusted Third Party System for Secure Multi-Party Computations", 2015 6th International Conference on Information and Communication Systems (ICICS), 2015.

[8] Jyotirmayee Rautaray and Raghvendra Kumar," Distributed RK- Secure Sum Protocol for Privacy Preserving", IOSR Journal of Computer Engineering (IOSR-JCE), e- ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 1, Jan. - Feb. 2013.

[9] N V Muthu lakshmi and Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1), 2012, 3176 – 3182.

[10] Mayur B Tank and Tushar A Champaneria," Privacy Preserving Distributed Data Mining", IJSRD – International Journal for Scientific Research & Development, Vol. 3, Issue 04, 2015.

[11] Selva Rathna and Dr.T. Karthikeyan," Two Phase Secured Multiparty Sum Computation Protocol (2PSMC) for Privacy preserving data mining", International Journal Of Engineering And Computer

Science ISSN: 2319-7242 Volume 4 Issue 4 April 2015, Page No. 11453-11456.

[12] Lu, Yunmei, "Privacy Preserving Data Mining For Horizontally Distributed Medical Data Analysis", Dissertation, Georgia State University, 2016.

[13] S.Shunmugam, Dr. R.K. Selvakumar and P.Kavitha,"A Comprehensive Survey on Privacy-preserving Distributed Data Mining",pp. 895-900, IJRECE VOL. 6 ISSUE 3,ULY - SEPTEMBER, 2018.

[14] S.Shunmugam, Dr. R.K. Selvakumar and P.Kavitha,"A Virtual Coordinator based Privacy-Preserved Distributed Data mining Using Association Rule",International Journal of Pure and Applied Mathematics,Volume 119 No. 16, 1535-1540,2018.