A Glance at Secure Multiparty Computation for Privacy Preserving Data Mining

Dr. Durgesh Kumar Mishra[1], Purnima Trivedi[2], Samiksha Shukla[3]
[1]Acropolis Institute of Technology and Research, Indore, India, durgeshmishra@rediffmail.com
[2]RKDF, Indore, India, mca_purnima@yahoo.co.in
[3]Christ University, Bangalore, Karnataka, India, smiksha_research@rediffmail.com

Abstract

In this paper, we provide an overview of the new and rapidly emerging research area of Secure Multiparty Computation (SMC). We also propose several existing as well as new SMC problems along with some solutions. It provides detailed overview of work done so far in this area and a brief evaluation and conclusion about SMC. SMC literally means: *Secure*- Safety concerns for data security and integrity of individual organizations. *Multiparty*- Involving multiple organizations/parties for Privacy Preserving Data Mining (PPDM). *Computations*- Any global operations. Therefore, keeping them together, it is a mechanism to provide collaborate computations of multiple organizations without revealing data of individual organization.

Keywords: Privacy, Security, Trusted Third Party, Secure Multiparty Computation.

I. Introduction

Internet and distributed computer architecture provides innumerable possibilities for collaborative/joint computations. SMC is a mechanism for privacy preserving data mining which is meant for joint computations in networked environment. It can be defined as, to provide computations among several diverse organizations in a safe or secure manner. With SMC, several parties can jointly perform some global computation on their private data without any loss of data security/privacy. It provides base for end-to-end secure multiparty protocol development.

Let $O_{1,}\ldots, O_n$ there be n organizations that wish to perform a joint computation C_1 on their private data. Since, computation is to be performed on private data, it is important requirement that this private data should not be accessible to any other organizations, i.e. if $D_{1,}\ldots, D_n$ be the data corresponding to n organizations and let D_i be data corresponding to i^th organization, then it is required for computation that, D_i should not be accessible to any D_j where $i \neq j$ *and* $j=1, 2\ldots n$. Therefore, each organization only gets the final results of joint computation without being aware of inputs involved and the computations made. This field is gaining youth interest; several research works in this area is being carried out by several scholars, in India and abroad. Privacy is vital concern for each organization and SMC protocols are a means to guarantee them in an easy manner.

For example, consider a patient has been ill for last 5 years. He has taken treatment from several doctors and sometimes he has been hospitalized too. If we wish to calculate the complete recovery time of that patient, it will be the joint sum of durations for which patient took treatment from each individual doctors and the duration for which he was hospitalized. Each doctor and hospitals maintain their patient's database. Now here joint computation is involved and this computation only provides recovery duration without revealing other information of any doctor's clinic or hospital databases.

The simplest and most general approach to solve such problem is to use the Trusted Third Party (TTP) atop of all organizations which perform all joint computations and maintain security. Therefore, no organization O_i can find out input from other organizations O_j, where $i \neq j$ and $j=1, 2\ldots, n$. Several subtle real world problems exist which can be viewed as SMC problems and can be solved by the use of SMC protocols. The SMC protocols may be based on any of the two paradigms [12]:

A. Real Model: Organizations run and use their own SMC protocols without the need for trusted third party.

B. Ideal model: Organizations rely on trusted third party for computations.

1. Literature Survey

Large amount of work has been done on SMC to provide secure joint computations among mutually distrusted entities. This computation can be anything like selective information sharing, arithmetic/relational operations, sorting, searching, hashing or other such operations. Yao presented the initial concept of SMC in the form of "Two Party Computations" [8]. Later, this was generalized to multiparty computation problems by Goldreich, who is prominent researcher who contributed a lot to SMC in the form of secure solutions for any

ISSN : 0975-3397

functionality [9]. Besides this, Agrawal *et al* provided fast and secure algorithms for mining association rules [10]; Atallah et al contributed to secure multiparty computation geometry, which are a base for routing and other network related problems [3]. Lindall *et al* provided cryptographic techniques and solutions for SMC [11]. Rebecca Wright provided some solutions to SMC and Privacy Preserving data mining through its PORTIA project [3]. Several problems and protocols to solve them have also been proposed by various eminent researchers which provide a clear view of SMC, their problems and solutions.

SMC provides a transformational framework that systematically transforms a normal computation to SMC computation. According to the number of distinguishable inputs, we can classify the computations into single input and multi-input computation models. This advantage can also be viewed as a drawback as all the computations may not necessarily demand the same level of security. Therefore, there is a need to distinguish among the normal computations and the SMC. In this way, all computations won't incur the same overhead. Only SMC problems would cost the overhead and other computations can be carried out normally.

Moreover, if locality of reference is obeyed some results are requested again and again for further computation, caching such data can greatly improve the performance, but it must not hinder our security concerns.

2. SMC problems

SMC can be done in the form of database query, authorization or authentication validations, mathematical/relational computations, scientific computations, statistical computations or any geometrical operations. Several real world problems exists that can be viewed as SMC problems. We have listed many SMC problems [3] explored so far and provide some new SMC problems and their applications along with the solutions [1, 2, 4, 5]:

3.1 Privacy Preserving Co-operative Scientific Computations

Linear System of Equations: Let Alice has m private linear equations represented as $M_1*x=B_1$ and Bob has (*n-m*) private linear equations represented by $M_2*x=B_2$, where *x* is an n-dimensional vector. Alice and Bob wish to jointly find a vector '*X*' that could satisfy both Alice and Bob's equations.

Linear Least Square Problem: Let Alice has M_1 private linear equations represented by $M_1*x=B_1$ and Bob has M_2 private linear equations represented as $M_2*x=B_2$, where '*X*' is an n-dimensional vector and $M_1+M_2>n$. Since there are more conditions to be satisfied than the degree of freedom, it is possible that some of them may be violated. Therefore, we take the residual factor *r* such that *r* is kept as minimum as possible. The least square criterion is the use of Euclidean (least square) norm for size of *r*.

Linear Programming Problem: Let Alice has private linear system of equations represented as $M_1*x<=B_1$ and Bob has private linear system of equations represented as $M_2*x<=B_2$, where Alice has M1 linear equations in her system and Bob has M2 linear equations. We want to minimize $A_1*X_1+\ldots\ldots +A_n*X_n$ for known $A_1\ldots A_n$ and the solutions $X=(X_1\ldots X_n)$ should satisfy both Alice and Bob's requirements.

These problems are generally viewed as routing, planning, scheduling, assignment, design etc.

3.2. Privacy Preserving Database Query:

Database Query: Suppose Alice want to search a string q in Bob's database of strings $S=\{S_1,\ldots, S_n\}$ and it just want to return the result, without revealing the Bob's entire string database. The match could be exact or approximate match.

3.3. Privacy Preserving Intrusion Detection:

Profile Matching: Alice has a database of hacker's profile. Bob has recently traced a behavior of a person, whom he suspects a hacker. Now, if Bob wants to check whether his doubt is correct, he needs to check Alice's database. Alice's database needs to be protected because it contains hacker's related sensitive information. Therefore, when Bob enters the hacker's behavior and searches the Alice's database, he can't view his whole database, but instead, only gets the comparison results of the matching behavior.

Fraud Detection: Two major financial organizations want to cooperate in preventing fraudulent intrusions into their computing system, without sharing their data patterns, since their individual private database contains sensitive data.

3.4: Privacy Preserving Data Mining:

Classification: Alice has a private database D_1 and Bob has private database D_2. How can Alice and Bob build a decision tree based on $D_1\cup D_2$ without disclosing the contents of their private database to each other? Several algorithms like ID3, Gain Ratio, Gini Index and many other can be used for Decision Tree along with SMC protocols.

Data Clustering: Alice has a private database D_1 and Bob has private database D_2. Alice and Bob want to jointly perform data clustering on $D_1\cup D_2$. This is primarily based on data clustering principle that tries to increase intraclass similarity and minimize interclass similarity.9*i. Mining Association Rules*: Let Alice has a private database D_1 and Bob has private database D_2. If Alice and Bob wish to jointly find the association rules from $D_1 \cup D_2$ without revealing the information from individual databases.

Data Generalization, Summarization and Characterization: Let Alice has a private database D_1 and Bob has private database D_2. If they wish to jointly perform data generalization, summarization or characterization on their combined database $D_1 \cup D_2$, then this problem becomes an SMC problem.

3.5. Privacy Preserving Geometric Computation:

Intersection: Let Alice has a private shape *a* and Bob has private shape *b*, if Alice and Bob want to find whether a and b intersect, then they need to share their database of shape coordinates to find whether they intersect

Point Inclusion Problem: Let Alice has a private shape *a* and Bob has private point *p*. Now, if Bob wants to know whether his private point p lies on shape boundary or inside or outside, then they need to jointly use both databases without revealing their individual information to each other.

Range Searching: Let Alice has a private range and Bob has *N* private points. Alice and Bob want to jointly find the number of points in the Alice's range; neither is willing to disclose their data to other party.

Closest Pair: Let Alice has *M* private points and Bob has *N* Private points in a plane. Alice and Bob want to jointly find the two points closest among (*M*+*N*) points, i.e. two points having their mutual distance smallest.

Convex Hull: Alice has *M* private points and Bob has *N* Private points in a plane respectively. They wish to find a convex hull from these (*M*+*N*) points.

3.6. Privacy Preserving Statistical Analysis:

Correlation and Regression: Let $D_1=(X_1,\ldots, X_n)$ be Alice's private dataset and $D_2=(Y_1,\ldots, Y_n)$ be Bob's private dataset. Alice and Bob wish to jointly find the following results:

Correlation Coefficient between x and y: Correlation coefficient is to be found between the private datasets D_1 and D_2 without revealing D_1 or D_2 to each other.

Regression Line: This helps to find the regression lines for D_1 and D_2 and perform regression analysis for future predictions.

3.7. Selection Problem: Let Alice and Bob have their own private databases. If they wish to apply any selection procedure on each other's databases, then such a process should not reveal their database knowledge to the other party.

3.8. Sorting Problem: Let Alice and Bob have their private databases and they jointly wish to sort their database without disclosing each others database.

3.9. Shortest Path Problem: Let Alice and Bob both have their location databases and they want to find the shortest path among the two locations a and b.

3.10. Privacy Preserving Polynomial Interpolation: Let Alice and Bob both have their databases and they want to interpolate against a polynomial.

3. SMC problems Proposed by Us

a. Let n research universities from various countries wish to discover some current research trends from their research databases without compromising the security of each individual database.

b. Let us consider that several shopkeepers of some general stores wish to find shopping trends of customers /buying patterns without revealing information about their databases.

c. Let us consider an Intelligence Department that considers database of fingerprints/ thumb impressions. Now, if some police station employee wishes to check a particular fingerprints, it must not be able to gain its complete access, instead, he should only get the test results.

d. Alternatively, if police wishes to check a particular person's identity from his thumb impression/signature, they can consult the bank database. Bank database only reveals the match results of thumb impression/signature.

e. Let us consider n hospitals situated in various different countries having their medical databases and patient's history stored on some remote database sites. If an insurance company wishes to verify the med claim of a particular person, he can get that patient's information from hospital's database, but the hospital's database is not completely provided, instead only the requested information is allowed to access.

f. Let all universities across the globe jointly wish to evaluate each other and then declare the top 20 universities of the world on the basis of their 5 year's academic records. They all want to preserve the privacy of their individual databases.

g. Let all doctor's team from several countries want to jointly find a remedy for a particular disease. All carry out research and studies and only reveal conclusions before each other without revealing the whole task.

h. Let us consider Airlines Company that has a reservation database for each country. If a person wishes to make a reservation from city A located in country A to a city b located in country B, then we need to consult each intermediate countries databases. These databases provide only the queried details without disclosing their whole reservation database.

i. Let a social organization providing funds to large number of charitable trusts located in different countries. These charitable trusts can query the organization to check whether the requested fund has been issued or not, but cannot see the organization's whole database.

j. Several websites provides ocean of knowledge and contains authentication information. Whenever, we do e-shopping /e-commerce, the authentication database first validates us as an authenticated user and then when it comes to payment, our account number/credit card number is checked for correctness in the bank database and if transaction successfully completes, then only item is said to be purchased. In this case, authentication only checks the individual person's identity and bank's database check the card number only and other authentication and the bank database is kept confidential.

II. 4) SMC PROBLEM SOLUTIONS

All the SMC protocols proposed so far make use of any one of the 2 approaches [2] as shown in fig 1:

a. *Cryptographic Approach*: In this, the input from several parties in received in encrypted form by TTP.

b. *Randomization Approach*: In this the input from several parties is first concatenated/ associated with a random number, in order to keep it secure.

Several proposed solutions to SMC problems include The Oblivious Protocol, 1-Out Of N Oblivious Protocol, Zero Knowledge Proof, Oblivious Evaluation Of Polynomials, Secret Matching, Threshold Cryptography, Yao' Millionaire Protocol etc.. [1, 4, 5].

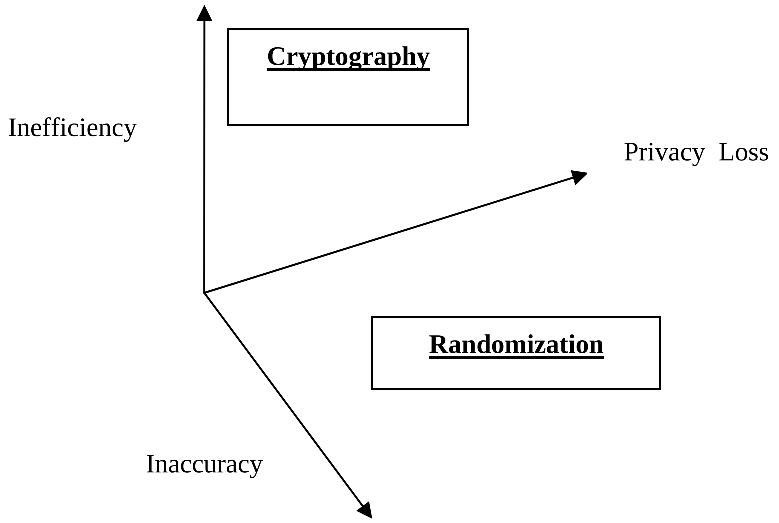

Figure 1. Kinds of SMC Solutions

Some other techniques like Secure Sum, Secure Set Union, Secure Size of Set intersection, Scalar Product, EM clustering etc can also be employed along with above mentioned approaches to find the SMC solutions [6]. Another mechanism can be to introduce an additional layer in between the organizations and the TTP that is known as *Anonymous Layer* [1]. This anonymous layer may be corresponding to each organization or can span multiple organizations as shown in fig.3.

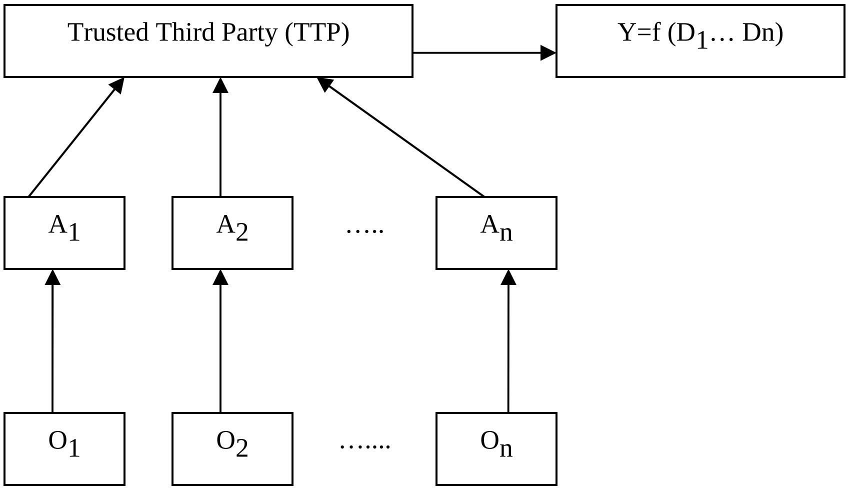

Fig 3: The SMC Architecture using Anonymous Layer

6. Conclusion and Future Works:

This paper bring several SMC problems and their solutions to light such as database queries, intrusion detection, geometric computation, Statistical Analysis and Scientific Computation. Researches are still underway to get efficient solutions to all the SMC problems and as the scope of the SMC are growing wider and wider, this area is gaining a lot of interest and attention. With widespread use of computers, proliferation of sensitive and private data is very important. The main aim of this paper is to divert the attention of the people who work even in other computation areas to view computation problems as SMC problems and suggest solutions for the same.

III. 7. References:

[1] D.K. Mishra and M. Chandwani,"Anonymity Enabled Secure Multiparty Computation for Indian BPO". In Proceeding of the IEEE Tencon 2007: International conference on Intelligent Information Communication Technologies for Better Human Life, Taipei, Taiwan on 29 Oct. - 02 Nov. 2007, pp. 52-56.

[2] Rebecca Wright, “Progress on the PORTIA Project in Privacy Preserving Data Mining,” A data surveillance and privacy protection workshop held on 3rd June 2008.

[3] Wenliang Du and Mikhail J. Atallah,”Secure Multiparty Computation Problems and their Applications: A review and Open Problems,” Tech. Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001..

[4] Jaideep Vaidya and Chris Clifton, “Leveraging the ‘multi’ in Secure Multiparty Computation,” WPES’03 October 30, 2003, Washington, DC, USA, ACM Transaction 2003, pp120-128.

[5] Andrew C. Yao,”Protocols for Secure Computations”, In Proc. 23rd IEEE Symposium on the Foundation of Computer Science (FOCS), IEEE 1982, pp 160-164.

[6] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, Michael Y. Zhu, ”Tools for Privacy Preserving Data Mining”. international conference on knowledge discovery and data mining, Vol. 4, No. 2, 2002, pp. 1-8.

[7] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parsiliti Provenza, Yucel Saygin, Yannis Theodoridis, “State-of-The-Art in Privacy Preserving Data Mining”, SIGMOD Record, Vol. 33, No. 1, March 2004.

[8] Y.C.Yao, “How Generate and Exchange Secrets”. In proceedings of the IEEE Symposium on Foundation of Computer Science IEEE, 1986, Pages 162-167.

[9] O.Goldreich, “Secure Multiparty Computation”, September 1998 (Working draft) Online available on: http://www.wisdom.weizmann.ac.il/~oded/pp.html.

[10] R.Agrawal and R.Srikant, “Fast Algorithms for Mining Association Rules”, in the proceedings of the 20th International Conference on Very Large Databases (VLDB), Santiago, Chile, September 12-15 1994.

[11] Y.Lindell and B. Pinkas, “Privacy Preserving Data Mining”. In advances in Cryptography-CRYPTO-2000, pp 36-54, Springer-Verlag, August 24 2000.

[12] Y.Lindell, IBM T J Watson “Tutorial on Secure Multiparty Computation”, available on wesite:- http://www.cs.biu.ac.il/~lindell/research-statements/tutorials-secure-computation.ppt