Preserving data privacy in data market and obtain truthful data

Miss.Aishwarya Pratap Jadhav[1] Prof.ManojWakchaure[2]

student,DepartmentofComputerEngineering
AssistantProfessor,DepartmentofComputerEngineering,
AmrutvahiniCollegeofEngineering,Sangamner,MH,India[1]
AmrutvahiniCollegeofEngineering,Sangamner,MH,India[2]

Abstract

People are posting their data on different sites. Huge amount of data were collected daily by the different users. Research has been done on the accumulation, different kind of applications proceeded rapidly.
Whenever any user want to buy any online product, usually people will inform themselves by reading the different online reviews. Their might be chances that, those particular user is fake user and posting bogus data. Also security is also the main issue. This paper will introduce the TPDM mechanism and different privacvy preservation that are used technique will given. TPDM combines Truthfulness and Privacy Presrvation in Data Market. It will use encryption-decryption methodology for security. Also users identity is also get check whether its fake or real identity.

Keywords: *Truthful and untruthful data, security, Homomorphic encryption, data market, privacy protection*

I. INTRODUCTION

Data market is an online store where people can buy data. Data marketplaces typically offer various types of data for different markets and from different sources. Common types of data sold include business intelligence, advertising, demographics, personal information, research and market data. Data types can be mixed and structured in a variety of ways. Now a day large amount of data were present, many users share their private data. There are many open records systems were present. Due to which many users exchanged their data on the internet. For example, Facebook and twitter's API platforms were present, that collects personal social media data of many users. But, there is a security problem in those market-primarily based platforms, i.e., it is hard to assure that the data . To combine truthfulness and privateness maintenance in a data marketplace, there are main four challenges. The most important first design challenge is that to verifying the truthfulness of data which is collected from different users and gives confidentiality for the privacy of that data. Data collected from user must be valid user. The raw data is contributed by the data provider, while privacy confidentiality has a tendency to prevent them from learning these confidential contents. The next challenge comes from data processing, which makes verifying the truthfulness of facts series even more difficult. these days, increasingly records markets provide information services rather than giving raw data directly to the user. In this paper TPDM were introduced, Truthfulness and Privacy Preservation in Data Markets. The Fully homomorphic encryption technique used for providing the security to the data by applying specific operations i.e, addition and multiplication operation on ciphertext data. The third challenge is based on how to assure the accuracy of data processing, under the information asymmetry between the data users and the service provider due to data confidentiality. Mainly, to ensure data confidentiality against the data giver, the service provider can employ a conventional symmetric /asymmetric cryptosystem, and can let the data subscriber do encryption on their raw data to generate ciphertext. Unfortunately, a hidden problem arisen is that the data users fails to confirm the correctness and completeness of a again records carrier. And the last design challenge is the effectiveness need of data markets, the service provider need to be able to gather records from a big wide variety of data subscribers with low latency. Also, the service provider must be verify the data confidentiality.

II. LITERATURE SURVEY

T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su , They construct three major classification protocols that satisfy the privacy constraint: hyper plane decision, Naive Bayes, and decision trees.

Nowaday machine learning concepts were used in many fields, such as medical or genomics predictions, spam detection, face recognition, and financial predictions. There is privacy concern so that's why these classification protocols were introduced. These protocols to be combined with the AdaBoost . And protocols are efficient, also it taking milliseconds to a very few seconds to perform a classification when running on the real medical data sets. [2]

Jan Camenisch, Susan Hohenberger , Michael stergaard Pedersen, proposed a paper on the first batch verifier for messages. Furthermore they also propose a new signature scheme with very short signatures, for which batch verification for many users is also highly efficient. Although the new signature scheme which they proposed has some limitations, it is very efficient and still practical for some communication applications.[3]

Magdalena Balazinska, Bill Howe, and Dan Suciu, They discussed about It outline some of the key challenges that such markets face and also discussed the associated research issues that our community can help solve. Also they told the implications of the emerging cloud-based data markets on the database research community. Our community has a great opportunity in making a significant impact on these data markets, while solving exciting data management research challenges.[4]

Dan Boneh, Matthew Franklin proposed a paper on fully functional identity-based encryption scheme (IBE). The system is based on the bilinear maps between groups. In this paper identity-based encryption scheme is introduced. The security of the system is a natural analogue of the computational Diffie-Hellman assumption. The limitation of this system is Revocation for private key is not present. [5]

Seung Hyun Seo, Mohamed Nabeel, Xiaoyu Ding, proposed a paper on An Efficient Certificateless Encryption for Secure Data Sharing in Public System storage clouds. The Safely share responsive data and information in public system storage clouds. Move forward towards the effectiveness. Additionally has downside that Network Connections Dependency furthermore Cost is more this calculation utilized is public key encryption algorithms.[6]

Ricardo Mendes and Joaqo P. Vilela these fellows proposed a paper and gives the survey on the most relevant PPDM techniques from the literature and the measured used to evaluate such techniques and represents typical applications of PPDM methods in relevant fields. The paradigm were introduced also known as Privacy-Preserving Data Mining (PPDM). In this survey, an overview of data mining methods that are applicable to PPDM of large quantities of information is provided.[7].

III. SYSTEM ARCHITECTURE

In the proposed research work to design and implement a system which is first efficient secure scheme for the data markets, which simultaneously garantees the data accuracy and confidentiality preservation. The TPDM is structured internally in a way of Encode -then-Sign, using fully homomorphic encryption technique and identity based signature. The service provider must be Collect the true data and process that data. In online datasets two different datasets were downloaded- 1) Real datasets, 2) Fake datasets.

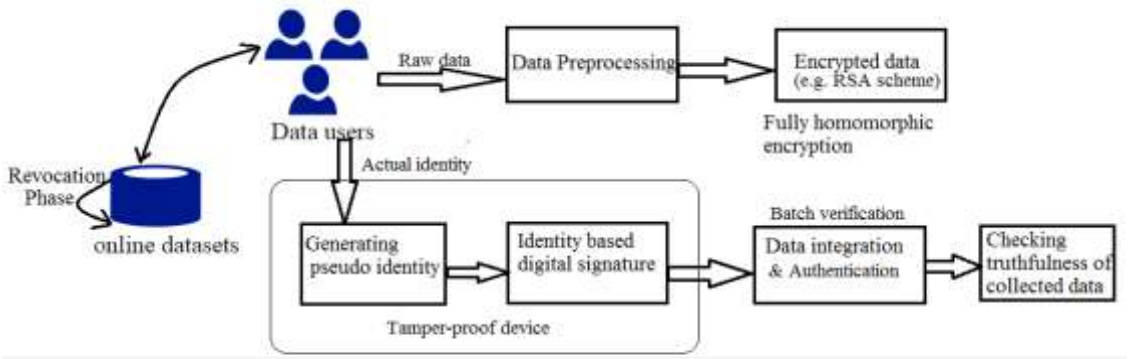

Fig. 1. System architecture

TPDM mainly consists of total 5 phases:
1. Initialization
2. Key generation
3. Data submission
4. Data preprocessing and verification
5. Revocation and Tracing

Phases illustration:

1. Initialization:
In the initial phase the data users and data Subscriber should make registration on registration center. The registration center gives the parameters for a fully homomorphic encryption system: a private key SK, a public key PK, an data encoding method E(.), and a decode method D(.). For activatation of the tamper-proof device, which is implemented by using specific form of software and hardware and each and every registered data subscriber is assigned with a one REAL identity RID and one password PW.

2. Key generation
To provides data confidentiality, we gives fully Homomorphic encryption . Which uses addition and multiplicative operation on data and encrypt that given data.

3. Data submission
Two-layer batch verification is considered in this phase which is conducted by both the service provider and the data users. Data preprocessing and signatures aggregation done by the service provider. And outcome verification conducted by the data users

• **Batch Verification:** After constructing the encrypted data, we can allow each data subscriber digitally sign their encrypted raw data. But, digital certificates has incurs significant communication overhead. To handel these problem, we gives an identity-based signature scheme.
• **Space Construction:** The main problem is that how to enable the data user to verify the validnesses of user's signatures, while maintaining data confidentiality. Hence the public key encryption method is apply for the construction of the encrypted text, and then service provider has to decode/decrypt that data using decryption algorithm and then process the data. The fully homomorphic cryptosystem for is used for encryption and decryption.

4. Data preprocessing and verification:
The fully homomorphic encryption properties used. Also the data users able to verify the accuracy of data processing. Under the different circumstances the data users should knows his/ her plaintext. Verification done for checking the truthfulness of data.

5. Revocation and Tracing:
In any user misbehave the that particular user get revocate and do not have authority to access the data, modify the data.

IV. ALGORITHMS

I I. RSA Algorithm

INPUT: Required modulus bit length, k
OUTPUT: An RSA key pair ((N,e),d)
1. Select a value of e from 3,5,17,257,655373,
2. repeat
3. p ← genprime(k/2)
4. until (pmode)≠1
5. repeat
6. q ← genprime(k - k/2)
7. until (qmode)≠1

8. N ← pq
9. φ (n) ← φ (p)* φ (q)↔(p-1)(q-1) // 'φ' Euler's totient function.
10. e ← 1<e< φ (n)
11. d ← ≡ e (mod φ(N))
12**.** return (N,e,d)

A. Encryption:
Sender does the following:
1. Obtains the public key (n,e).
2. Represents the plaintext message as a positive integer m with 1<m<n
3. Computes the ciphertext c=m^e modn.
4. Sends the ciphertext c .

B. Decryption :
1. Person A recovers m from c by exploitation his or herprivate key exponent, d, by the computation m cd (mod n).
2. Consider m, Person A will recover the first original message M by reversing the padding scheme.

This procedure works since c me (mod n), cd (me)d (mod n), cd mde (mod n).
By the symmetry property of mods we have that mde mde (mod n).
Since de = 1 + k(n), we can write
mde m1 + k(n) (mod n), mde m(mk)(n) (mod n), mde m (mod n).

B. l- Depth tracing Algorithm [1]

Initialization: S = {δ1,δ2,…..δn}, head = 1, tail = n, limit = l,
whitelist = Ø, blacklist = Ø, resubmitlist = Ø
1.Function l-DEPTH-TRACING(S, head, tail,limit)
2. if │whitelist│ + │blacklist│ = n or limit = 0 then
3. return
4. else if CHECK-VALID(S, head, tail) = true then
5. ADD-TO-WHITELIST(head, tail)
6.else if head = tail then //Single signature verification
7.ADD-TO-BLACKLIST(head, tail)
8. else // Batch signatures verification from δhead to δtail
9. mid = │ head+tail│
10. l-DEPTH-TRACING(S,head, mid, limit -1)
11. l-DEPTH-TRACING(S,mid + 1, tail, limit - 1)

V. MATHEMATICAL MODEL

A. Mapping diagram

A function is a relationship that pairs each input with exactly one output. A function can be represented by ordered pairs or a mapping diagram. A mapping diagram consists of two parallel columns. The first column represents the domain of a function f , and the other column for its range. Lines or arrows are drawn from domain to range, to represent the relation between any two elements.

• S be the whole system S= {I, P,O}

Where,
I =Input
P =Procedure
O =Output
user u= {data distributor, service provider, data consumer}

• I= {I0, I1, I2, I3, I4}

I0 = real id of user
I1 =user's raw data
I2 =public/private key
I3 =Enter comments
I4 =Service provider activities

• P ={P0, P1, P2, P3, P4, P5, P6, P7,P8}

P0 = Login to registration server
P1 =create pseudo identity
P2 =Data encryption(AES algorithm used)
P3 =Identity base signature generation
P4 =Data preprocessing
P5 =Check data integrity and authenticity
P6 =Check truthfulness of data
P7 = Product review in data market
P8 =Revoke user

• O= {O0,O1,O2}

O0 =provide comments
O1 =collection of truthful data
O2 =Feedback to market place
O3 =revocation of user

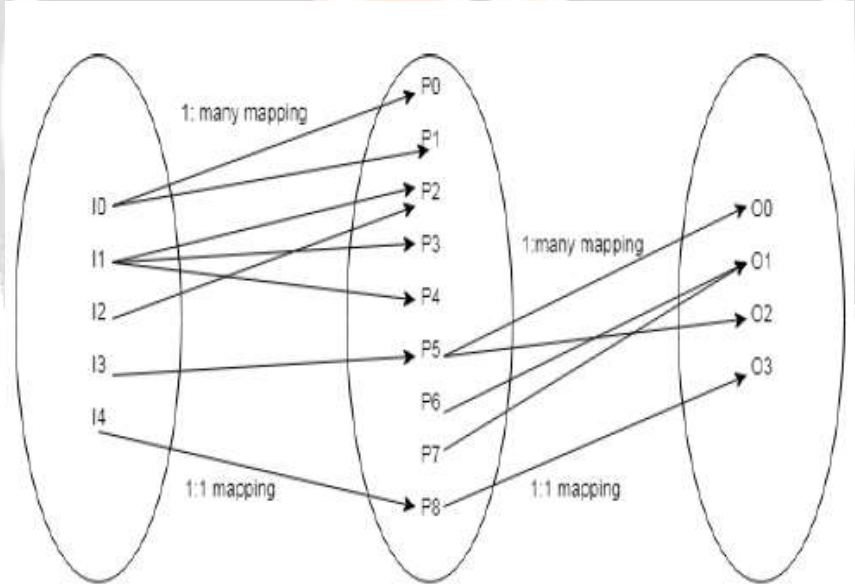

Fig. 2. Function mapping

VI. CONCLUSION

The TPDM mechanism for truthfulness for data and providing the confidentiality preservation fir that data, the data subscribers have to truthfully submit their own data. Except, the service provider must collect true data . In TPDM, the data contribution have truthfully submit their own data, but cannot impersonate others. Moreover, both the personally identifiable information and the sensitive raw data of data subscribers are well protected. Service providers is forced to truthfully collect & process data. Both the personally identifiable information the sensitive raw data of data subscribers are well protected.

VII.REFERENCES

[1] Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Xiaofeng Gao and Guihai Chen " Achieving Data Truthfulness and Privacy Preservation in Data Markets "" IEEE Transactions on Knowledge and Data Engineering (2018 Early Access).Study on the ISCX Dataset." Data Intelligence and Security (ICDIS), 2018 1st International Conference on.IEEE, 2018.

[2] T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su, AccountTrade: accountable protocols for big data trading against dishonest consumers, in INFOCOM, 2017.

[3] J. Camenisch, S. Hohenberger, and M. Ø. Pedersen, "Batch verification of short signatures," Journal of Cryptology, vol. 25, no. 4, pp. 723–747, 2012.

[4] M. Balazinska, B. Howe, and D. Suciu, Senior Members, IEEE "Data markets in the cloud: An opportunity for the database community," Vol. 4, no. 12, pp. 1482–1485, 2011.

[5] Dan Boneh, Matthew Franklin, Fellow,"Identity-based encryption from the weil pairing," in CRYPTO, 2001.

[6] Seung-Hyun Seo, Member, IEEE, Mohamed Nabeel, Member, IEEE, Xiaoyu Ding, Student Member, IEEE, and Elisa Bertino, Fellow,An
Efficient Certificateless Encryption for Secure Data Sharing in Public system storage clouds,Vol.25, No.9, PP.2107.

[7] Ricardo Mendes, Student member, And Joaqo P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications", Vol. 5, 2017

[8] Z. Zheng, Member, IEEE, Y. Peng, Member, IEEE, F. Wu, S. Tang,Member, IEEE, and G. Chen, Member, IEEE "Trading data in the crowd: Profit-driven data acquisition for mobile crowdsensing," IEEE Journal on Selected Areas in Communications, vol. 35, no. 2, pp. 486–501, 2017..

[9] M. Barbaro, T. Zeller, and S. Hansell,Fellows ―A face is exposed for AOL searcher no. 4417749, N Y Times, August 9, 2006.

[10] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in INFOCOM, 2010.

[11] Wakchaure M. A., Sane S. S. ,An Algorithm for Discrimination Prevention in Data Mining: Implementation Statistics and Analysis. In2018 International Conference On Advances in Communication and Computing Technology (ICACCT) 2018 Feb 8 (pp. 403-409). IEEE.

[12] Shitole M, Wakchaure M. A. ,Survey: Techniques Of Data Mining For Clinical Decision Support System. Vol-2 Issue-1 IJARIIE-ISSN (O)-2395-4396.;1571.