



An Imaginative Approach on PPDM Through Zero-Knowledge Protocol

Ankita Shrivastava, U.Datta
C.S Dept., M.P.C.T, GWALIOR
India

ABSTRACT- In contemporary period, privacy-preserving data mining (PPDM) has been studied expansively, since the wide detonation of sensitive information on the internet. This paper walks around a use of privacy preserving data mining through cryptographic technique by using Zero knowledge protocol (ZKP). After survey of topical approaches that have been applied to knowledge hiding thread. Data mining services require accurate input data for their results to be evocative, but privacy concerns may require users to provide forged information. This paper is based on the computational and theoretical limits associated with privacy-preservation. Moreover, in several application areas such as person identification it is repeatedly required of entities to prove knowledge of some fact without enlightening this information. Such proof of knowledge are called Zero Knowledge Interactive Proofs (ZKIP) and involve interactions between two communicating parties the Prover and the Verifier as we shown in[1]. In a ZKIP, the Prover demonstrate the tenure of some information (e.g. authentication information) to the Verifier without revealing it. In this paper, we focus on the application of ZKIP protocols using PPDM. We studied well-established ZKIP protocols based on the ECC algorithm, but here we use RSA algorithm to provide sensitive information to the authorized user who proves his identification without revealing the actual knowledge. To the best of our knowledge, this is the first attempt of implementing and evaluating ZKIP protocols with emphasis on datamining technique in terms of PPDM. This work's results can be used from developers who wish to achieve certain levels of security and privacy in their applications.

Keywords- Privacy-preserving data mining (PPDM), Zero Knowledge Interactive Proofs(ZKIP), RSA.

I. INTRODUCTION

A number of algorithmic techniques have been analyzed for privacy-preserving data mining in which sensitive data is hide to the un-authorized person, only valid data to be shown. Here we introduce the amalgamation of ZKP (Zero knowledge protocol) and PPDM (Privacy-preserving data mining) on this database with the help of cryptographic technique as we discussed in[1]. To discover some information stored in the database the user must prove their identification weather they are authorized or not by giving some insinuation but not the tangible knowledge to the system. To best of my knowledge Data and knowledge hiding are two research advices that examine how the privacy of unprepared data, or information, can be maintained moreover before or after the course of mining the data. At the time of bear out identification of a user ZKP is used, then after applying the cryptographic algorithm (RSA) we use the PPDM algorithm to demonstrate the corresponded database. Generally, a zero-knowledge protocol allows a proof of the truth of an declaration, while expressing no information whatsoever about the declaration itself other than its real truth. Generally, such a protocol involves two entities, a prover and a verifier as we discussed in[2]. A zero-knowledge proof allows the prover to demonstrate knowledge of a secret while revealing no information whatsoever of use to the verifier in conveying this demonstration of knowledge to others. So here the user is prover and the system is verifier. The system verifies the identification of the user by demanding (insisting) the message and the key value for encrypting that message. Then system will encrypt the specified message by applying encrypting key value through RSA algorithm and turn out the encrypted message, after encryption the message is decrypted by applying decrypting key value to the encrypted message, if the original message will found after the entire process then it proves the user authorization then only system will display the corresponded existing database. If original message is not found after applying the whole process then it assumes that user is an un-authorized and system will not display any database. Here the user can proves his identification till five times. After five times system could not get value as it exceed the limit to get login. The prover's objective is to encourage the verifier about the truth of an assertion, e.g. the claimed knowledge of a secret. This data includes private and susceptible information like patient diseases, association structural details, bank account details etc. When data mining techniques are applied on these applications the private and sensitive information of the subjects will be exposed. However, it is necessary to share the information in such a way that the identities of the individuals are not revealed.

II. STUDY OF DIFFERENT TECHNIQUES IN PRIVACY-PRESERVING DATA MINING

- A. **RANDOMIZATION METHOD:-** The randomization method is a procedure for privacy-preserving data mining in which noise is added to the data in order to disguise the attribute values of records. The noise added is adequately large so that individual record values cannot be well again. Therefore, techniques are designed to obtain aggregate distributions from the agitated records

B. **K-ANONYMITY**:- The k-anonymity model was developed to create indirect identification of records from public databases. The k-anonymous is just not to discover any k-anonymous data, instead to get one “superior” or even “excellent” according to various proven cost. This is because amalgamation of record attributes can be used accurately to identify individual records. According to k-anonymity method, it decreases the granularity of data representation with the use of techniques such as generalization and suppression. An important method for privacy de-identification is the method of k-anonymity. For example, if the identifications from the records are removed, attributes such as the birth date, name and zip-code can be used in order to uniquely identify the identities of the underlying records as shown below in tab.1

TABLE I: K-ANONYMOUS DATA

Age	Weight	Name
45	60	Celina
55	45	Jones
48	70	Mark C

(a) Original Data

Age	Weight	Name
[40-50]	[55-65]	Celina
[50-60]	[40-50]	Jones
[40-50]	[65-75]	Mark C

(b) K-Anonymous Data

C. **CRYPTOGRAPHIC APPROACH** :- In many cases, various parties may desire to distribute cumulative private data, without disclosing any sensitive information at their end . For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties. Also, we have a cryptographic approach i.e. Elliptic Curve Cryptography (ECC).

D. **ECC APPROACH**:- Elliptic Curve Cryptography (ECC) is an remarkable substitute to conventional public key cryptography, such as RSA. ECC is a preeminent candidate for accomplishment on controlled devices where the major computational possessions specifically speed, memory is limited and low-power wireless communication protocols are employed. Specifically because it conquers the same security levels with traditional cryptosystems using smaller parameter sizes as discussed in [2]. Moreover, in several application areas such as person identification and e-Voting, it is frequently required of entities to prove knowledge of some fact without revealing this knowledge. Hence, this proofs of knowledge are called Zero Knowledge Interactive Proofs (ZKIP) and involve communication between two communicating parties, the Prover and the Verifier as discussed in[2]. In a ZKIP, the Prover give an idea about the possession of some information like authentication information to the Verifier without divulging it. In this review paper, we bring to light on the application of ZKIP protocols on resource constrained devices. In this paper we study well-recognized ZKIP protocols proofs of the truth of an assertion, while conveying no information whatsoever regarding the declaration itself other than its real truth.

III. WHY WE NEED PPDM

Purposes in commercial domains acquire bulky datasets on individuals. This data consist of private and sensitive information specifically bank account details, Patient diseases, organization structural details etc. While data mining techniques are pertaining on these applications the private and sensitive information of the idea will be exposed. Privacy preserving data mining normally uses a variety of method to alter the original data or the data produced (computed, derived) using data mining technique as we discussed in[3].Nevertheless, it is essential to split the information in such a way that the identities of the individuals are not exposed. Therefore it is necessary to anonymize the data. Subsequently it include different techniques to anonymize the data. In addition also it has classification of PPDM algorithms.

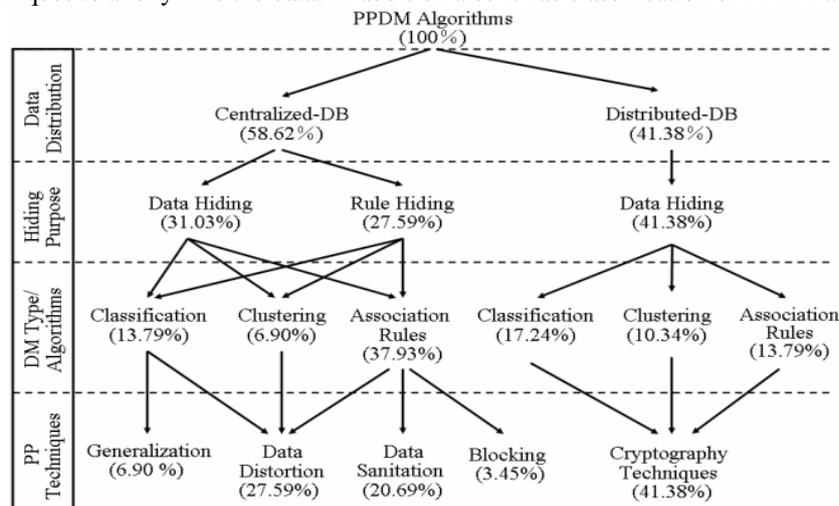


Fig. 1 Classification of PPDM algorithms

IV. PRIVACY PRESERVING DATA MINING (PPDM)

Privacy Preserving Data Mining (PPDM) is solitary the extreme brand new concept of data mining research challenges. It express area of data mining those attempts to defend sensitive information from disclosure. The difficulty with data mining output is to reveal some information, which is measured to be private and personal. Unproblematic access to such personal data causes risk to individual privacy. The tangible apprehension of people is that their private information should not be distorted after the prospect without their knowledge. The factual threat is that once information is at liberty, it will be unfeasible to prevent misuse. There has been rising apprehension about the chance of wrong use with personal information at the back of the user without the acquaintance of actual data holder. The general description of privacy in the cryptographic technique confine the information that is disclose by the distributed calculation to the information that can be studied from the selected output of the calculation[4]. Privacy preserving data mining technique provides new course to resolve this problem. PPDM offers legal data mining results without exploring the core data values. The advantage of data mining can have the benefit of, without compromising the privacy of individuals. The original data is adapted or use a process in such a way that private data and private knowledge stay private yet after the mining process. The key purpose of privacy preserving data mining is intend to efficient framework and algorithms that can dig out appropriate knowledge from huge amount of data without disclosing any sensitive information. In the case of a trusted party as a data miner may possibly allowed to obtain the data in its original structure. However, through improved data security threats to the communication channel, the personal/sensitive data might require safe transformation from disclosure. Else, the untrusted data miner might be supply with the encoded data for the defense of privacy or remarkably sensitive information.

V. AN OVERVIEW OF ZERO KNOWLEDGE PROTOCOLS

In general, a zero-knowledge protocol allocate a proof of the truth of an declaration, while transmission no information at all about the declaration itself other than its real fact. Generally, such a protocol involve two individuals, a prover and a verifier. The zero-knowledge proof allows the prover to exhibit knowledge of a secret although revealing no information whatsoever make use of the verifier in assigning expression of knowledge to others.

The zero-knowledge protocols to be talk about are example of interactive proof systems and non-interactive proof systems. In the beginning, a prover and a verifier exchange several messages (challenges and responses), usually dependent on random numbers which remain secret subsequently the prover sends only one message as shown in[2]. In both steps the prover's aim is to convince the verifier about the truth of an assertion, say knowledge of a secret. The verifier also accept or discard the proof. The function of verification is normally executed by a selected centralized group (verification team). This authentication process is *quick* and *safe*, that is, it does not disclose any information about data owner as discussed in[5]. Zero-knowledge proof have to comply with the assets of completeness and security. Hence, verification is complete if given an honest prover and an honest verifier, the protocol accomplish with vast probability and sound if the probability of a fraudulent prover to complete the proof successfully is insignificant. We can take typical example of zero-knowledge proof is recognized as Alibaba's cave problem. As we discussed this story in review also, Annie has exposed the secret word to open a magic door in a cave. The cave is shaped like a ring, with the doorway on one side and the magic door blocking the opposite side, as shown in Figure 2. The left path from the entrance is marked as A and the right B. John states with the intention that he will pay her for the secret, but not until he will guaranteed that she actually knows it. Annie claims that she will tell the secret, but not until she receives the money.

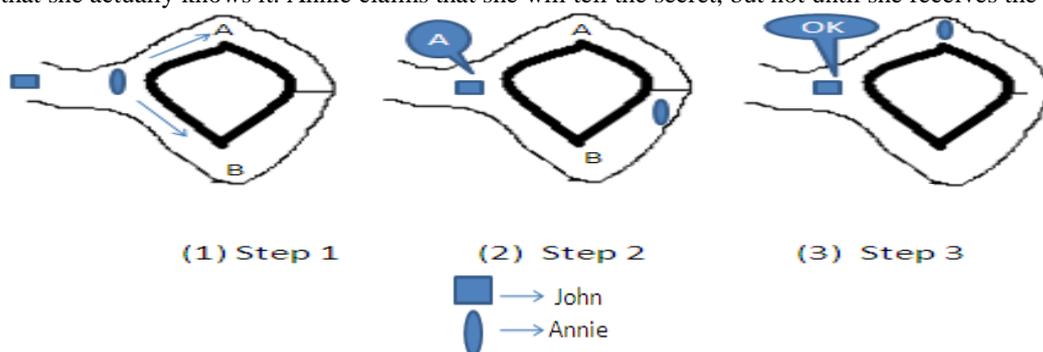


Fig. 2 Alibaba's Cave Problem.

John remain outside the cave as Annie goes inside as we already discussed in[1]. Annie at random takes either path A or B inside the cave later John gets enter into the cave and shouts the name of the path he desire from her to use to return either A or B, choosen at random Annie make use of the secret word if looked-for to open the magic door. The above steps are repeated n times until John get sure that Annie knows the secret word. At the moment, assume that Annie does not know the secret word. As John chooses path A or B at random, Annie has a 1/2 possibility of cheating at one round. If the above steps are repeated for many rounds, Annie's chance of successfully anticipating all of John's requests would become vanishingly small. Therefore, if Annie reliably become visible at the exit according to John names, then he can conclude that she is very likely to know the secret word. In this paper we drawn attention to zero-knowledge protocols based on the scheme by which Annie can prove that she knows the magic word without disclosing it to John. The steps of this method are now described below.

- John remains outside the cave as Annie goes inside the cave.
- Annie at random takes either path A or B inside the cave.
- John enters the cave and shouts the name of the path he wants her to use to return either A or B, selected at random.
- Annie will apply the secret word if needed to open the magic door.
- The above steps(1 to 4) are repetitive n times until John are convinced that Annie knows the secret word.

VI. PROPOSED WORK

In this paper we pooled the concept of ppdm with zero knowledge protocol. In which user gives insinuation to the system and pretending authorized user for accessing private data. As we proposed here, user has to go through some check points by this method user must prove there authenticity by giving key value which is not actual value for the database access. Here Prover demonstrates the possession of some information (e.g. authentication information) to the Verifier without disclosing it. If user is authorized can easily access the original data. If user is un-authorized data will not reveal. So here we take two types of database, Original and PPDM data. Original database is open for authenticated person by proving authenticity, PPDM data is open for user who is authorized but not the actual user. And data will not display for un-authorized user. Also it will not take more than five values to be entered as key. We put here some limit upto five times the user can enter the value. After that system will not take value as it display access denied. Now we taking here two types of database, original and PPDM.

Lets take some example regarding proposed approach:-

1) *USE OF ZKIP (Zero Knowledge Interactive Proofs)*:- Assume we encrypt the message "coming at morning". The first thing we done here is to change the message into a numeric format. Each letter is represent by an ascii character,thus it can be accomplished somewhat easily. Therefore we are not going in depth to convert strings to numbers or vice-versa, but it can be done very simply. Here, we convert the string into bit(array), and then the bit(array) to a large number. This can be easily reversed to get back the original string by giving the large number.Using this method, "coming at morning" becomes 88 (let). So $m=88$.

- Key Generation for original database

Now, we pick two large primes, p and q . These numbers must be random and also not too close to each other. Here, the numbers that we generate using RSA Algorithm:

- $P = 17$
- $q = 11$

With these two large numbers, we can calculate n and $\phi(n)$

$$n = p * q = 17 * 11 = 187$$

- $\phi(n) = (p - 1) * (q - 1) = 16 * 10 = 160$
- e - the public key is 7 such that $1 < e < \phi(n)$ and e and n are co-prime. Let $e = 7$
- d - the private key is 23 such that compute a value for d such that $(d * e) \% \phi(n) = 1$. Therefore $d = 23 [(23 * 7) \% 160 = 1]$
- Public key is $(e, n) => (7, 187)$
- Private key is $(d, n) => (23, 187)$

The encryption of message $(m) = 88$ is $c = 88^7 \text{ mod } 187 = 11$. Now our message "coming at morning" is encrypted and assumed message is 11. We can easily decrypt the message through the key value of d . For e.g. The decryption of $c = 11$ is $m = 11^{23} \text{ mod } 187 = 88$. Now it is prove that key value which the user has passed is correct as it matched with the existing value of decryption. Here it will produce the original database shown below table 2. which is associated to the decryption value.

TABLE II
FOR $e = 7 \Rightarrow$ ORIGINAL DATA

ID	PATIENT NAME	PATIENT DISEASE	DOCTOR	AGE	CITY	GENDER
1	John	Diabetes	Johnson N	28	San Jose	M
2	Jones L	Cholera	Randall T	35	Dallas	M
3	Leape LL	Brain cancer	Mant D	36	Austin	F
4	Cullen DJ	Headache	Bates DW	45	San Francisco	F

So here we used the concept of ZKIP as user has to prove his/her authenticity and system has to verified user's legitimacy without demanding the actual key value (d).

Now,

2) *USE OF PPDM (Privacy Preserving Data Mining)*:- As we already said we use here amalgamation of PPDM concept with ZKIP. After using ZKIP here we discuss the PPDM as it display valid but anonymize data for the user who is not owner but the valid user. So for user who is authorized but not actual(owner) can access this database which does not effect data sensitivity because of privacy preserving. The main purpose of privacy preserving data mining is to develop efficient frameworks and algorithms that can extract relevant knowledge from a large amount of data without disclosure of any sensitive information. So it produce the valid but potential data which can use further for knowledge extract or business strategies.

Here we take the same message “coming at morning”, i.e. $m = 88$. Also, we take the same value for p and q .

But for e we take the different value that is to say $e = 23$,

- e - the public key is 23 such that $1 < e < \phi(n)$ and e and n are co-prime. Let $e = 23$
- d - the private key is 7 such that compute a value for d such that $(d * e) \% \phi(n) = 1$. Therefore $d = 7 [(7 * 23) \% 160 = 1]$
- Public key is $(e, n) => (23, 187)$
- Private key is $(d, n) => (7, 187)$

The encryption of message $(m) = 88$ is $c = 88^{23} \bmod 187 = 11$. Now our message “coming at morning” is encrypted and assumed message is 11. We can easily decrypt the message through the key value of d . For e.g. The decryption of $c = 11$ is $m = 11^7 \bmod 187 = 88$. Now it is prove that key value which the user has passed is correct as it matched with the existing value of decryption. Here it will produce the ppdm database shown below table 3. which is associated to the decryption value. As it will not display the patient identity (*name, actual age etc.*)

TABLE III
FOR $e = 23 \Rightarrow$ PPDM DATA

ID	PATIENT DISEASE	DOCTOR	AGE	GENDER
1	Diabetes	Johnson N	20-30	M
2	Cholera	Randall T	30-40	M
3	Brain cancer	Mant D	35-40	F
4	Headache	Bates DW	40-50	F

Therefore our proposed approach is highly secured and confidential. Confidential in provision of patient data thus if user does not want to share his/her disease or identity with any one therefore our proposed approach will anonymize (*de-identify*) the sensitive data. As it will hide the sensitivity of data and turns into the valid data for knowledge discovery.

VII. FUTURE WORK

Cryptographic Techniques for privacy preserving in data mining get astounding results. The approach presents here, indicates securing sensitive data and knowledge from cruel users. Our approach is better way to apply datamining techniques with security that covers our logical paradigm from others. Currently we are using RSA algorithm for data encryption. In future we can use SHA or advanced algorithm for encryption and also we can include the simulation result to show the efficiency and performance. Our approach can be expanded by integrate the additional Algorithms for identification and classification of sensitive and non sensitive data can build our PPDM, a complete competent system for datamining that will work in the direction of mining based information with minimal privacy violate.

VIII. CONCLUSION

The increasingly ability to identify and collect large amounts of data, evaluate the data using data mining process and conclusion on the results gives potential advantage to organizations. The theory of zero-knowledge proof is used to assist quick, scattered, reliable however secure public verification. Although, such repositories also include private and sensitive information but by making the data public which may be personal information can cause major damage to data holder. Therefore, anonymous environment should be proposed where any sensitive information can not be disclose. Thus there is need to discover and scatter the databases, without compromising with the privacy of the individual’s data. Therefore, in this paper zero knowledge protocol looks useful thought for de-identification our secret information without any loss of knowledge to data miner. This protocol is latest concept in cryptography techniques. A zero-knowledge proof allow to disclose knowledge of a secret without enlightening information whatsoever use by the verifier to assign this expression of knowledge to others. This is the first attempt of merging the concept of ZKIP protocols with PPDM emphasis in terms of privacy with security according to best of my acquaintance. This approach can be used by developers who wish to achieve certain levels of security and privacy in their applications. We can attain better privacy because the result achieved is in anxious form, so the privacy of original data will obtain valid data mining result.

REFERENCES

- [1] Ankita Shrivastava, U.Dutta," *An Emblematic Study of Different Techniques in PPDm* " International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 8, August 2013 ISSN: 2277 128X .
- [2] Ioannis Chatzigiannakis, Apostolos Pyrgelis, Paul G. Spirakis, Yannis C. Stamatiou "*Elliptic Curve Based Zero Knowledge Proofs and Their Applicability on Resource Constrained Devices*" University of Patras Greece, arXiv: 1107.1626v1 [cs.CR] 8 Jul 2011.
- [3] Kiran P, S Sathish Kumar and Dr Kavya "*A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining*",An International Journal (ACIJ), Vol.3, No.2, March 2012.
- [4] Anand Sharma and Vibha Ojha"*Implementation of Cryptography for Privacy Preserving Data Mining*" International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010.
- [5] Debasri Saha and Susmita Sur-Kolay, IEEE,"*Secure Public Verification of IP Marks in FPGA Design Through a Zero-Knowledge Protocol*" IEEE Transactions on very large scale Integration (VLSI) Systems, Vol. 20, No.10, October 2012.
- [6] Lambodar Jena, Ramakrushna Swain," *A Comparative Study on Privacy Preserving Association Rule Mining Algorithms*" International Journal of Internet Computing, Volume-I, Issue 1, 2011.
- [7] Umesh Kumar Singh, Bhupendra Kumar Pandya, Keerti Dixit "*An Overview on Privacy Preserving Data Mining Methodologies*" International Journal of Engineering Trends and Technology- Sep to Oct Issue 2011 ISSN: 2231-5381.
- [8]. Ashraf El-Sisi," *Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database*",The International Arab Journal of Information Technology, Vol. 7, No. 2, April 2010
- [9] Benny Pinkas "*Cryptographic techniques for privacy-preserving data mining*" Published in SIGKDD Explorations, Volume 4, Issue 2, January 10, 2003.
- [10] Archana Tomar, Vineet Richhariya, Mahendra Ku. Mishra," *A Improved Privacy Preserving Algorithm Using Association Rule Mining in Centralized Database*",International Journal of Advanced Technology & Engineering Research (IJATER) ISSN NO: 2250-3536 Volume 2, Issue 2, March 2012.
- [11] Nathani sushma, Priyanka Kanaparthi," *Multidimensional Techniques for Privacy Preservation in Datasets*" International Journal of Computer Science and technology (IJCSST) Vol. 2, Issue 4, Oct- Dec.2011.
- [12] P.Kamakshi , Dr.A.Vinaya Babu," *Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data*" Journal of Computing, Volume 2, Issue 4, April 2010.
- [13] Fuad Al-Yarimi, Sonajharia Minz," *Multilevel Privacy Preserving in Distributed Environment using Cryptographic Technique*"Proceedings of the World Congress on Engineering 2012 Vol I WCE 2012, July 4 - 6, 2012, London.
- [14] Muthu Lakshmi and Dr. K Sandhya Rani," *Privacy Preserving Association Rule Mining Without Trusted Party for Horizontally Partitioned Databases*" International Journal of Data Mining & Knowledge Management Process (IJDMP), Vol.2, No.2, March 2012.