# Multilevel Privacy Preserving in Distributed Environment using Cryptographic Technique

Fuad Al-Yarimi, Sonajharia Minz

*Abstract*—**Data mining extracts pattern or knowledge from a large amount of data. In the applications that are based on information sharing, an additional challenge is faced; while dealing with data containing sensitive or private information. Common data mining techniques do not address this problem. Therefore, the knowledge extracted from such data may disclose patterns with sensitive or private information. This may put the privacy of individuals or groups, the business strategies and classified information at risk [1]. In the recent past years, Privacy Preserving Data Mining (PPDM) has attracted research interest with potential for wide applications. This paper considers all data repositories and one node for pattern extraction at distributed locations. Many techniques like anonymity, randomization and cryptography have been experimented with privacy preserving data mining . This paper considers information system based approach as not all attributes may store same level of sensitive data. Therefore, some attribute values may require higher degree of privacy preservation than some others. This paper explores the use of cryptography, namely DES algorithm for encrypted data sharing to achieve privacy preservation. The experiments have been carried out on the size of the secret key that is proportional to the level of sensitivity of the attribute for multilevel security. The RSA algorithm has been used to encrypt the DES algorithm key for enhancing secured communication.**

*Index Terms*— ***Cryptography, DES, Privacy Preserving Data Mining, RSA.***

## I. Introduction

Privacy Preserving Data Mining (PPDM) is one of the most recent of data mining research challenges. It refers to the area of data mining that attempts to protect sensitive information from disclosure. 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 [2, 3].

This paper attempts to consider, firstly, the information system based approach to address the problem of privacy preservation. Secondly, it explores the role of cryptography as a security technique, and finally, the outcome of the two previous objectives is to propose a novel approach for privacy preservation in data mining in a distributed environment.

The premise of the paper is that the data owner is unable to extract knowledge from a large repository of data including sensitive items. Therefore, the entire dataset does not comprise sensitive data values. The attributes values with high degree of sensitivity may require higher privacy preservation than the attribute values with less degree of private information. In the case of a trusted party as a data miner may be allowed to access the data in its original form. Yet, with increased data security threats across the communication channels, the private/sensitive data could require securing transformation from disclosure. Otherwise, the untrusted data miner could be provided with the encoded data for the protection of privacy or crucially sensitive information. Therefore, the paper explores the use of cryptographic techniques to achieve the objective of various scenarios.

A number of cryptographic techniques such as DES and RSA have emerged as efficient methods for secured data communication. [7] However, they have their respective drawbacks. RSA requires very high computational cost, especially when the data is large. Thus, DES, an efficient algorithm, has been considered as an alternative to encrypt various attribute values with multiple keys of variable sizes. To focus on the role of the cryptographic techniques for privacy preservation, this paper uses the DES algorithm. The DES algorithm employs a symmetric system for encryption with faster and more efficient outcome for a large dataset. RSA algorithm is proposed to be used for secured sharing of the secret key used by DES algorithm for encryption and decryption.

The experiments based on the implementation illustrate the proposed technique as a novel approach. The experiments also display the relations between security and privacy.

The next section briefly discusses the DES and RSA techniques used for two steps: Firstly, to encrypt the values of the private attributes and secondly to encrypt the secret key of the first technique. The section also discusses encryption key management and size of the cipher-text. In section 3, the information system based approach for multilevel privacy preservation is presented. Section 4 presents the proposed method for the distributed environment, followed by algorithmic representation of the method in section 5. The last section concludes and

Manuscript received March 14, 2012, 2012; revised April 17, 2012.

Fuad Ali Mohammed Al-Yarimi, School of Computer & Systems Science, Jawaharlal Nehru University, New Delhi 110067 – Mobile:0091-8826131975, India (email: frindh@gmail.com).

Sonajharia Minz, School of Computer & Systems Science Jawaharlal Nehru University, New Delhi 110067 – India, (email: sonaminz@mail.jnu.ac.in)

summarizes the paper.

## II. Cryptographic Techniques

In a distributed environment, in addition to the original parties there may be a third party as data miner. The third party could be a trusted party or an untrusted one. If the third party is a trusted party, then all the data owners send their data to the trusted party to apply data mining functions and send the outcome to the original parties. The objective of a cryptographic system is to design protocols that do not divulge any data except the data that is designated output. Secure multiparty computation is performed where each party has a part of the input needed to perform the computation [8]. However, at the end of the computation, the parties should have only learned the result of the computation.

### *A. DES Encryption Package*

The Data Encryption Standard (DES) encryption package was developed by IBM. DES is one of the most tested algorithms, and no major drawbacks. In the original DES algorithm, a 56-bit long key is used to generate a pseudorandom stream of 1's and 0's. This stream is mixed with the data, and then fed back into itself, resulting in a very scrambled stream of the same data. Changing the key yields another unique stream. The decryption process is designed to generate the same random pattern, starting at the same point as the sender and reversing the technique to arrive at plaintext data. It is necessary that both the sender and the receiver have the same key. This inherently authenticates one to the other. Data integrity is preserved because an attacker has no idea of where to begin in order to break the encryption.

### *B. RSA Public Key Technique*

The RSA encryption technique uses the principle that if two large prime numbers, ‘e’ for encryption and ‘d’ for decryption are multiplied, the resulting number ‘n’ is hard for factorization back into the two original numbers, i.e.

n = e d.

A data sent using *n* as the key can only be decrypted by the receiver possessing either *e* or *d*. As *e* never leaves the transmitter and *d* never leaves the receiver, and factoring *n* into ‘*e*’ or ‘*d*’ is not possible in any reasonable length of time, the system is secure. Since *n* appears in public, it is called the public key. The private key *d*, remains with the receiver, and *e* with the sender. This RSA encryption method is known as an asymmetric cryptosystem, as for a unique decryption key *d* there is a unique encryption key *e* for a given *n*. RSA is also a two key system, since either of the two private keys *e* or *d* can unlock the system if revealed. As *n* can be released safely to the public, it is practical to build a public key directory. It simplifies the major worry of private key distribution. The unique value of a public key *n* can be assigned to an individual or a company and used for authentication [4].

Encryption-Decryption

Key management is a major consideration in an environment requiring secure communication. Like any other asset, keys must be generated, distributed and accounted for. Key distribution can be simplified by using a public key system such as RSA. [4] Since the public key for any computer or individual need not be secure, a public key directory allows a computer or individual to send data simply. In a private key system, the keys must be the same at both ends and therefore authentication is assured intrinsically. A private key system such as DES is generally less expensive and incurs less processing intensity than a public key system. The ideal combination is to use the public key system to distribute private or secret keys, and then use the faster secret key system to transmit data. In cases where distributing keys to individuals is a major concern, or where it is necessary to authenticate previously unknown individuals, a public key algorithm like RSA should be considered. In cases where security and speed are paramount, or where it is relatively easy to distribute keys to known individuals, a private key algorithm should be considered. They are generally more secure, faster and cheaper than public key algorithms.

RSA can encrypt data blocks that are shorter than the key length, and also that it is much slower than a symmetric encryption like DES. Therefore, the advantages of the two techniques are combined in the following steps as a general approach proposed for encryption of the sensitive data values:

1. A random key (secret key) of the required length for DES is generated. (The length of the key depends on the level of security that is needed.)
2. The designated data is then encrypted using the DES and using the corresponding secret key.
3. The secret key is encrypted using the RSA public key.
4. Both the outputs from 2 and 3 are then sent to the trusted third party.

The trusted third party will do the flowing steps to then decrypt the data before applying the data mining algorithms:

1. Decrypt and retrieve the DES secret key using the private key, applying RSA algorithm.
2. Decrypt the data using that DES secret key.

As mentioned earlier, the length of the secret key generated by DES algorithm is related to the level of sensitivity of the attribute value. Though, not all attributes may be at the same level of sensitivity. In such cases, many secret keys with different lengths are required to be generated for multilevel privacy preserving.

Practically, it has been observed that the number of private attribute in a dataset may not be more than 15% out of the total size of attributes. Therefore, to encrypt the private attributes as opposed to the entire dataset depends on only the vertical partition of the dataset. It also reduces the time required for the encryption process. This would also have an impact on the time of communication for the pre-processed encrypted data.

Table 1**.** Hospital Patient Database

| DOB | Sex | Zipcode | Disease |
|---|---|---|---|
| 1/21/76 | Male | 53715 | Heart Disease |
| 4/13/86 | Female | 53715 | Hepatitis |
| 2/28/76 | Male | 53703 | Bronchitis |
| 1/21/76 | Male | 53703 | Broken Arm |
| 4/13/86 | Female | 53706 | Flu |
| 2/28/76 | Female | 53706 | Hang Nail |

### C. Cipher-text with DES

To calculate the size of the cipher-text produced by the block-cipher encryption, the following information is needed:

1. Length of the plaintext value
2. Encryption block size
3. Padding information (if padding is used)

In the most generic case, the size of the cipher-text can be calculated as:

CipherText = PlainText + Block - (PlainText MOD Block)

Where CipherText, PlainText and Block indicate the sizes of the cipher-text, plain-text and encryption block respectively. Basically, the resulting cipher-text size is computed as the size of the plaintext extended to the next block. If the size of the plaintext is an exact multiple of the block size then padding is used and with the addition of one extra block containing padding information.

For example, if a nine-digit Social Security Number (SSN) is to be encrypted using the Rijndael encryption algorithm with the 128-bit (16-byte) block size and PKCS #7 padding. For the purpose of the illustration, the dashes are removed from the SSN value before the encryption, so that 123-45-6789 becomes 123456789, and the value is treated as a string, not as a number. If the digits in the SSN are defined as ASCII characters, the size of the cipher-text is calculated as below:

CipherText = 9 + 16 - (9 MOD 16) = 9 + 16 - 9 = 16 (bytes)

Alternately, let a 16-digit credit card number (defined as a 16-character ASCII string) needs to be encrypted. Here, the size of the plain-text value is the exact multiple of the block size therefore; an extra block containing padding information is appended to the cipher-text. Thus, the size of the cipher-text will be:

CipherText = 16 + 16 - (16 MOD 16) = 16 + 16 - 0 = 32 (bytes)

## III. Multilevel Privacy Preserving System

The information system as presented in the table 1 is an array of vectors describing each data sample. Some information related to the persons may be very general, some private and others sensitive. The first category pertains to the values of the attributes that do not need to be anonymized. The second category pertains to the values of the attributes called quasi-identifier, i.e. those that can be exploited to link to a private attribute, e.g. marital status, sex, working hours of individuals, whether they suffer from hypertension, etc. The attributes DOB, Zipcode, and Sex could constitute the quasi-identifier. The third category corresponds to the set of sensitive attributes, e.g. Diseases. This kind of information requires to be protected from dissemination. Therefore, such attibute values may be transformed so that no individual is identifiable, and is subjected to undue treatment.

In view of the above considerations, the attributes describing a data row in a dataset are categorized in the following three types:

1. **Identity attributes:** the attributes that identify user e.g. number, name, etc.
2. **Quasi-identifier attributes:** the attributes that can be used in conjunction with public records in order to uniquely identify the records e.g. zip code, phone number, etc.
3. **Private attributes:** the attributes that contains private and sensitive information which must not be disclosed to unauthorized persons, e.g. disease, credit card numbers, salary etc [5].

### A. Information System

The dataset as depicted in table 1 can also be viewed as an information system. Let $\mathrm{I} = (\mathrm{U}, \mathrm{A}, \mathrm{V})$ be an information system with U representing the universe of samples, A the set of attributes and V the values of all the attributes [9]. Consider a partition of A = {$A_1$, $A_2$, $A_3$}. Let $A_1$, $A_2$ and $A_3$ be subsets of the set of attributes A, of the type identifiers, quasi-identifier, and private respectively. i.e.

$$A = A_1 \cup A_2 \cup A_3$$

such that,

$$A_1 \cap A_2 = A_2 \cap A_3 = A_1 \cap A_3 = A_1 \cap A_2 \cap A_3 = \emptyset$$

Let $f$ be a function mapping and ordered pair (x, a) ∈ U×A to an element in V. i.e.

$$f: \mathrm{U}\times\mathrm{A} \rightarrow \mathrm{V}$$

$$f(\mathrm{x}, \mathrm{a}) \in \mathrm{V}, \forall \mathrm{x} \in \mathrm{U} \text{ and } \forall \mathrm{a} \in \mathrm{A}\}$$

Let $V_a$, called the domain of a ∈A be the set of all values of the attribute a. Then,

$$\mathrm{V} = \cup_{\forall\, \mathrm{a}\epsilon\mathrm{A}} \mathrm{V}_{\mathrm{a}}.$$

Therefore, corresponding a partition A = {$A_1$, $A_2$, $A_3$}, a partition of the domains of all the attributes may be obtained,

$$\mathrm{V} = \{\mathrm{V}_{A_1}, \mathrm{V}_{A_2}, \mathrm{V}_{A_3}\}$$

i.e. $\forall \mathrm{a} \in \mathrm{A}_1$, $f(\mathrm{x}, \mathrm{a}) \in \mathrm{V}_{A_1}$ or $\mathrm{V}_{A_1} = \cup_{\forall\, \mathrm{a}\epsilon A_1} \mathrm{V}_{\mathrm{a}}$

Similarly, let the subsets $\mathrm{V}_{A_2}$ and $\mathrm{V}_{A_3}$ be the blocks of the partition of V corresponding the blocks $A_2$ and $A_3$.

Let |A| = k. Then the multidimensional database U is primarily an array of k-dimensional vectors corresponding to a set of attributes. A partition of I is considered to be the set of subsystems corresponding to each block of the partition of A = {$A_1$, $A_2$, $A_3$}, say, $\mathrm{I}_{A_1}, \mathrm{I}_{A_2}$ and $\mathrm{I}_{A_3}$, i.e.

$$\mathrm{I} = \langle \mathrm{I}_{A_1} | \mathrm{I}_{A_2} | \mathrm{I}_{A_3} \rangle$$

where each $\mathrm{I}_{A_i} = (\mathrm{U}, \mathrm{A}_{\mathrm{i}}, \mathrm{V}_{\mathrm{A}_{\mathrm{i}}})$ for i∈ {1, 2, 3}.

In case of datasets described by attributes that may be quasi-identifier and private in their characteristics, there may be two main scenarios requiring privacy preserving:

**Scenario 1**: When the data owner in a company has a database containing some sensitive information and needs to send it to the technical department for analysis and when the technical department is not authorized to disclose the sensitive data.

**Scenario 2**: Outsource analysis to a third party requiring communication. In this case, we have two sub-cases:

Case 2a: The third party is a trusted one. He/she has authority to access the original/raw data. No need for pre-processing between owners of data and third party, but communication of data requires security. Therefore, sensitive attributes must be encrypted.

Case 2b: The third party is not a trusted one. The data miner cannot be allowed to access the original (sensitive) attribute values. Therefore, pre-processing techniques be applied to the data before communication. Data mining is done on the anonymized data.

The following Table specifies the requirements of security and a combination of security and privacy techniques if data mining is outsourced with respect to the location of the data.

Table 2. Privacy-Security Requirements

| Outsourced Analysis by Third party<br>Data location | Trusted | Untrusted |
|---|---|---|
| Central | Secure Com | Privacy & Secure Communication |
| Distributed | Secure Com | Privacy & Secure Communication |

In Privacy Preserving System scenario 2, mentioned above, all the data owners need to send their data to the third party for the data mining process.

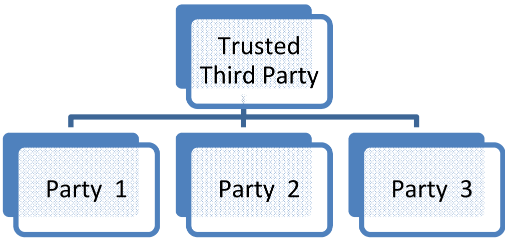

Figure 1: Distributed Data Base with Trusted Third Party Scenario

In this scenario, secure communication needs to be ensured. If the third party is a trusted party, then the data needs to be encrypted corresponding to various levels of privacy preservation before sending it to the analyser. The third party is required to decrypt it before applying the data mining process on decrypted data. In case of privacy preservation required for only a part of the data description corresponding to some attributes, encryption of the corresponding domains may be considered. It may further be considered that a number of cryptography techniques may be applied to various attribute oriented partitions of the dataset to ensure higher security. For the cryptosystem as discussed in section 2, the size of the private key as used by DES algorithm is related to the level of privacy required and the RSA technique is employed for communicating multiple private keys of multiple sizes.

In scenario 2b, the third party may not be a trusted party; both the privacy and secured communication need to be considered. The concept of privacy is used because there is a need for data anonymity before encrypting the data and sending it to the third party. In this case the third party may decrypt the data but he would not be able to access the real data because it is anonymized.

### ***B.*** *Proposed Approach:*

Practically, in order to achieve the optimal efficiency, the symmetric key algorithms and public key cryptography algorithms are always combined together. That is, using asymmetric key cryptosystem to encrypt the confidential information needed to be sent, while using the RSA asymmetric key cryptosystem to send the DES secret key, the required amount of data security can be achieved. This takes advantage of both the kinds of cryptography, namely, high-speed DES and RSA key management mechanism.

However, in this paper, the use of DES & RSA algorithms for privacy preserving data mining in the distributed environment has been proposed for sharing the data with a trusted third party (TTP) for analysis and pattern extraction. The proposed approach offers a framework for each data owner to know only his input, the set of private keys, the public key and the outcome of knowledge extraction. The TTP would be in possession of all the sets of public keys corresponding to all the data owners to successfully decrypt the data. After executing some data mining algorithms the TTP would send the data mining results to the data owners. If the extracted patterns still disclose some sensitive information, then the patterns are encrypted using the private key set by the data owner and the TTP with the same DES & RSA algorithms. However, this is observed to be faster than data communicated as the size of knowledge is significantly less than the size of data communicated between the two parties.

## IV. Multilevel-Privacy Preserving Algorithm (MPPA):

Most real application domains either possess data as distributed locations or have typical distributed database applications [10,11,12]. Therefore, the algorithmic steps for the proposed multilevel privacy preserving data mining system in a distributed environment are presented as below:

**Step 1:** A pair of keys; public and private, are generated using RSA algorithm by the trusted third party. The private one is kept a secret and the public key is published. The size of private and public key depends on the level of secrecy that we want.

**Step 2:** A pair of keys; public and private, are generated using RSA algorithm by every other parties. **Step 3:** a. The data is encrypted using the secret key in DES algorithm by all parties.

b. The secret key is encrypted using the trusted third party's (TTP) public key.

c. The encrypted data and the encrypted secret key are sent to the TTP.

**Step 4:** a. Receive the encrypted data and key by TTP.

b. TTP will use his private key to encrypt the secret key then use the secret key to decrypt the data.

c. Doing the data mining functions by TTP.

**Step 5:** TTP will send the result of the data mining functions to all parties in an encrypted form or in a normal form depends on the knowledge itself, If the extracting results still disclose some sensitive information TTP will encrypted again before send it to the other parties using DES & RSA algorithm as same described in steps 3-4 above.

The complexity of the algorithm is not affected by the complexity RSA algorithm. The complexity of the algorithm is related to the complexity of DES algorithm, the number of attributes and various sizes of the private keys. Therefore, the complexity of the proposed algorithm depends on the complexity of DES algorithm which in turn depends on the length of the DES key.

## V. CONCLUSION

The use of DES and RSA algorithms pre-processes the data required for privacy preservation has been proposed and discussed in the paper. The data is encrypted before communicating with the trusted third party who applies data mining functions. The paper proposes a novel approach to secure sensitive and private information in the data with multiple levels of privacy for data mining in a distributed environment.

## REFERENCES

[1] Xiaodan Wu, Chao-Hsien Chu, Yunfeng Wang, Fengli Liu, and Dianmin Yue "Privacy Preserving Data Mining Research: Current Status and Key Issues" Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 762–772, 2007. Springer-Verlag Berlin Heidelberg 2007.

[2] Alexandre Evfimievski "Privacy-Preserving Data Mining" 2009, IGI Global.

[3] Jian Wang Yongcheng "A Survey on Privacy Preserving Data Mining". 2009 First International Workshop on Database Technology and Applications.

[4] Agrawal, R. and Srikant, R, "Privacy-preserving data mining", In Proc. SIGMOD00, 2000, pp. 439-450.

[5] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J, "Privacy preserving mining of association rules", In Proc. KDD02, 2002, pp. 217-228.

[6] G. K Gupta, "Introduction to Data Mining with Case Studies" second edition 2011.

[7] Xin Zhou, Xiaofei Tang "Research and Implementation of RSA Algorithm for Encryption and Decryption " 2011 The 6th International Forum on Strategic Technology.

[8] Ahmed K. Purdue, Amit P "Privacy-Preserving Data Mining Models and Algorithms" Advances in database systems" Volume 34 Series Editors

[9] Z. Pawlak, "Rough Set Theory and its Applications", Journal of Telecommunications and Information Technology, Vol.3 2002

[10] Benny Pinkas "Cryptographic techniques for privacy preserving Data mining". SIGKDD Explorations. Volume 4, Issue 2.

[11] 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.

[12] P.Kamakshi, A.Vinaya Babu "Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data". Jornal of computing , volume 2, Issue 4, April 2010, Isbn 21519617

[13] Charu C. Aggarwal and Philip S. "On Static and Dynamic Methods for Condensation-Based Privacy-Preserving Data Mining". ACM Transactions on Database Systems, Vol. 33, March 2008 .

[14] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002, pp. 557-570.