www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 3 March, 2014 Page No. 4040-4043

A Survey on Privacy Preserving Homomorphic in Collaborative Data Publishing

R.Sharmila, Dr. A.V.K. Shanthi
Department of Computer Science,
Sathyabama University, Chennai, India,
Associate Professor, Sathyabama University Chennai, India,
rsharmila81@gmail.com
avks15@gmail.com

Abstract— **Security issues for software system ultimately concern relationships among social actors, stakeholders, system users, potential attackers and the software acting on their behalf. The tools and methodologies providing systematic guidance and support to the design process are much needed. Data in its original form typically contains sensitive information about individual, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as what types of data can be published. This paper presents A Survey on Privacy Preserving Homomorphic in Collaborative data Publishing that are found in the current literature. Problems in usage of privacy preserving areas are identified. Security areas that needed future research are presented.**

Keywords— Privacy Preserving, Homomorphic encryption, k-anonymity.

INTRODUCTION

Data mining techniques are becoming more and more important in decision support processes and in extracting hidden knowledge from massive amount of data. Successful applications of data mining have been demonstrated in marketing, business, medical analysis,banks, bioinformatics and scientific exploration, among others. Digital information collected by governments, corporations, and individuals has created tremendous opportunities for knowledge-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published. Publishing high dimensional data is part of the daily operations in commercial and public activity. The utmost task is to develop methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved. Collaborative data publishing is very common in real life.. There is an increasing need to share data for mutual benefits or for publishing to a third party. For example, banking sectors want to integrate their customer data for developing a system to provide better services for its customers. Banks do not want to indiscriminately disclose their data to each other for reason such as privacy protection and business competiveness. Security is to protect vital information while still allowing access to those who need it. and provide authentication and access control for resources. When the data is distributed among multiple data providers, adversaries can attempt to infer the information about data. The objective is to limit the disclosure risk to an acceptable level while maximizing the benefit.

In this paper, we transform the original data into some anonymous form to prevent from inferring its record owner's sensitive information with help of homomorphic authentication and triple des algorithm.

RELATED WORK

PRIVACY PRESERVING TECHNIQUES

K-anonymity: K-anonymity described by B.C.M.Fung [2] assumes every record is distinguishable with other k-1 record within the anonymity table in accordance with a set of QI attributes in a dataset. If the values of QI attributes are identical when compared with the other cords, a table is said to be K-anonymous. Ram Prasad Reddy and Raju a[9] proposed generalization or suppression can be made use to achieve the K-anonymity requirement. The classification of Attributes is done as Key attributes, quasi –identifiers(QI), and sensitive attributes. Key attribute which is represented as Name, Address, phone number which is uniquely identified and it always removed before release. Quasi - identifiers whose values when taken together can potentially identify an individual and may include, e.g., Zip-code, Birth date, and Gender. And the last one is Sensitive attributes that are considered sensitive, such as Disease and Salary.

Anatomization: Anatomization does not modify the quasi-identifier or the sensitive attribute, but dissociates the relationship between the two.

Perturbation: Perturbation has a long history in statistical disclosure control due to its simplicity, efficiency, and ability to preserve statistical information.

Attacks on k-Anonymity

Here two types of attacks are addressed and they are homogeneity attack and the background knowledge attack.

Homogeneity Attack: In this attack, all the values for a sensitive attribute within a group of k records are the same. Therefore, even though the data is k-anonymized, the value of the sensitive attribute for that group of k records can be predicted exactly.

Background Knowledge Attack: In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field. **Differential Privacy:** C. Dwork [1] proposed that no risk is incurred by joining statistical databases, finer analysis of realistic attacks. Dinur and Nissim drove the development of Gaussian noise instead of Laplacian, and they showed that this increase is essential. It is an unconditional privacy guarantee for statistical data release or data computations.

L-diversity: Fuad Ali Mohammed Al-Yarimi [11] proposed ,an equivalence class is said to have l-diversity if there are at least "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. The disadvantage is it fails to prevent attribute disclosure like in similarity attack.

Attacks in l-diversity

Skewness Attack: When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure.

Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. This leakage of sensitive information occurs because while l-diversity requirement ensures "diversity" of sensitive values in each group it does not take into account this mantical closeness of these values.

T-Closeness:

Ninghui Li and Tiancheng Li [8] proposed, Anonymization techniques are very simple and hence scalable in case of privacy preservation, they fail to efficiently prevent the records' critical values deduction against attacks. The distance between the distributions of a sensitive attribute in an anonymized group to that of the whole table should be less than 1 threshold. A table is said to have t-closeness if all equivalence classes have t-closeness. The t parameter in t-closeness enables one to tradeoff between utility and privacy. Variational distance can be defined to measure the distance between two probabilistic distribution P = (p1, p2, ..., pm), Q = (q1, q2, ..., qm), Here there are two probability distributions over the values and the distance between the two probability distributions to be dependent upon the ground distances among these values. This requirement leads us to the Earth Movers distance (EMD). Earth Mover Distance(EMD) measure for t-closeness requirement has the advantage of taking into consideration the semantic closeness of attribute values.

DkA-Datafly Wei Jiang and Chris Clifton proposed by [4],[5] generates precise k-anonymous dataset and, it is a bottom-up algorithm that it computes k-anonymous data by substituting a specific value with a more generalized value. DkA is very general in a sense that any centralized k-anonymization protocol can be used to compute locally k-anonymous data, and its structure can be effectively adopted to create a secure two-party k-anonymization protocol from an insecure centralized k-anonymization algorithm. DPP2GA two party protocol is proven to preserve the k anonymity privacy constraint. It is a generic protocol that can be used to compute locally k-anonymous data. This approach may produce more precise data, and also introduce additional inference problems that may violate k-anonymity with respect to individual parties.

Dependency vulnerability analysis: Lin Liu and Eric Yu [7] proposed, Dependency vulnerability analysis aims at identifying the vulnerable points in the dependency network. Dependency relationships bring vulnerabilities to the system and the depending actor . Potential attackers may exploit these vulnerabilities to actually attack the system, so that their malicious intents can be served. Dependency modeling allows a more specific vulnerability analysis because the potential failure of each dependency can be traced to a depender and to its dependers.

SSMCDM TECHNIQUES

Balamurugan, Bhuvana and Pandian [10] proposed Session based Secured Multiparty Collaborative Data Computation. The framework of Session based Secured Multiparty Collaborative Data Computation comprises of participants, trusted third party, rule generation for private-sharable data, multisessions and instance generation. The participants are the users authenticated to involve in the data mining process and the trusted third party center is a common trust centre which provides session authentication to all participants. Rule generation deals with the formation of private data rules coined by the participants to preserve their privacy information . Here, multi-session component that maintains more number of sessions for one or more participants in parallel performing various data mining tasks is considered The instance generation is dealt with, that cohesively handles both data dependency and independency rules to mine the authenticated sharable data of respective participant in any given session. In SSMCDM framework, participant A requests the session key from the trusted center. The trusted center checks the session key status from the participant B after the session key verification and the session key is issued to participant A. Once, if the session key is generated for participant A, the privacy data rules are defined for participant A. Based on these, security is achieved during multiparty computation. Once a rule is generated, a share key is given to each participant so that it knows which participant wants to communicate with which other participant. Normally, share key is different for every participant and share key indicates which participants can communicate with each other. In SSMCDM framework, participant A can communicate with participant B and that communication takes place only between participant A and participant B. At that time
participant A cannot communicate with the participant C whereas participant C can share data with participant D.

MERGE–THEN–HIDE TECHNIQUES

Shyue – Liang Wang , Ting - Zheng Lai [15] proposed, Preserving output privacy on multiple/distributed data sets ,Trusting Third Party Approach (MTH) has been proposed. Given a set of horizontally partition data sets, $D1,D2, . . . , Dn$, that are owned by non-trusting collaborative parties, if a trusting third party existed, a simple solution to publish jointed but sanitized data set is to submit all data sets to this trusted third party. The third party hides the specified association rules then publishes the sanitized data set, ($D1 + . . . + Dn$)', if a trusting third party did not exist, an alternative is to hide designated rules in each data set independently, ($D\_$ 1, . . .,$D\_$ n). The sanitized data sets are then submitted to the third party. The third party then merges the individually sanitized data sets and publishes the results, ($D\_$ 1 + . . . + $D\_$ n), this is referred to as *Hide-Then-Merge* (*HTM*) approach or non-trusting-third-party approach.

SECURE MULTIPARTY COMPUTATION

Indhumathi[13] proposed Secure Multiparty Computation. It is mainly used to control the "insider attacker". The data providers are considered to be semi honest and they may try to verify the private record of other data provider, so to control this SMC is used. Thus, two important requirements on any secure computation protocol are privacy and correctness. The SMC problems use two computation concepts: - Ideal model and Real model paradigm.
In ideal model a Trusted Third Party (TTP) is used, which accepts inputs from all the parties, evaluates the common function and sends result of the computation to the parties. If the TTP is honest, then the parties can know the result only.
In real model, there is no third party, instead all the parties agree on some protocol which allows them to evaluate the function while preserving privacy of individual inputs. Secure computational Protocol for computation of sum of individual parties preserving privacy of their inputs. The protocol allows parties to break their data inputs into segments and distributing these segments among parties before computation.

CRYPTOGRAPHIC TECHNIQUES

HOMORPHIC ENCRPTION: Yehuda Lindell [6] proposed Homomorphic Encryption scheme, it is an encryption scheme which allows certain algebraic operations to be carried out on the encrypted plaintext, by applying an efficient operation to the corresponding ciphertext. In Homomorphic encryption schemes: if , the message space is a ring). There exists an efficient algorithm +pk whose input is the public key of the encryption scheme and two ciphertexts, and whose output is Epk(m1)+pk Epk(m2) = Epk(m1 +m2). It is easy to compute, given the public key alone, the encryption of the sum of the plaintexts of two ciphertexts. There is also an efficient algorithm +pk, whose input consists of the public key of the encryption scheme, a ciphertext, and a constant c in the ring, and whose output is c.pkEpk(m) = Epk(c.m). Decryption requires a single exponentiation.

GEOMETRIC DATA PERTURBATION

Geometric perturbation has shown to be an effective perturbation method in single-party privacy preserving data publishing. Keke Chen [12] proposed the geometric perturbation approach to multiparty privacy-preserving collaborative mining. The main objective is to securely unify the perturbations used by different participants without much loss of privacy guarantee and data utility. He designed three protocols and analyzed the features and the cost of each protocol, they are simple protocol, negotiation protocol, and space adaptation protocol. In the simple protocol, the data providers use the same randomly generated perturbation to perturb data. There are two issues in this protocol, the first one is , how to securely generate the same random perturbation in each site, while preventing the curious service provider knowing the unified perturbation, and the second one is, how to prevent privacy breach caused by curious data providers. The negotiation protocol aims at improving the overall privacy for all data providers. Some data providers may not be satisfied with the randomly generated perturbation in the simple protocol in terms of privacy guarantee. In the negotiation protocol, each data provider has a chance to review the candidate perturbation and vote for or against the candidate. In the negotiation protocol, the perturbed data has to be encrypted before distribution. To maintain different version of perturbed datasets results in additional cost. And the last one is the space adaptation (SA) protocol, which inherits the convenience of distributing data in the single-party scenario, while also reduces the cost of communication, encryption and maintenance.

CF USING SVD

Ibrahim yakut and Huseyin polat [14] proposed, Singular value decomposition (SVD) based Collaborative filtering (CF) that offer reliable and accurate predictions when they own large enough data. CF systems usually operate on existing databases, such as ratings for items collected from many users. Data collected for CF purposes might be split between different parties even competing online vendors. This partition might be horizontally or vertically. In horizontal partitioning, different online vendors hold disjoint sets of users preferences for the same items, while in vertical partitioning, they own disjoint sets of items' ratings collected from the same users. The objective is to achieve SVD-based CF tasks from partitioned data between different parties with privacy. Data holders that want to integrate their data should not be able to find out the true rating values and the rated/unrated items in each other's databases, but in the data sharing the identity of the products and customers should be established across the data holders databases. CF systems produce recommendations calculated with privacy concerns referrals to many users in real time. During an online interaction, users get referrals from CF systems. Online computation time for providing recommendations should be small enough so that many users can obtain referrals without wasting too much time.

DISCUSSION

To sum up all, Privacy Preserving is proven for single and two party protocol. Security is very low, so that the users are afraid of uploading the data in the distributed servers. No proper mechanism was implemented to audit the data that are stored in the distributed servers. Multiple checks of privacy constraint leads to worst case in Time complexity. To improve the efficiency of security, public auditing process was implemented. Extensive security and performance attracts everyone view towards it.

REFERENCES

[1] C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008,pp. 1–19.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*,vol. 42, pp. 14:1–14:53, June 2010.

[3] Privacy Protected Spatial Query Processing for Advanced Location Based Services Wei-Shinn Ku · Yu Chen · Roger Zimmermann© Springer Science+ Business Media, LLC. 2008

[4] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in *Data and Applications Security XIX*, ser. Lecture Notes in Computer Science, 2005, vol. 3654, pp. 924–924.

[5] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," *VLDB J.*, vol. 15, no. 4, pp. 316–333, 2006.

[6] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy preserving data mining," *The Journal of Privacy and Confidentiality*,vol. 1, no. 1, pp. 59–98, 2009.

[7] Security and Privacy Requirements Analysis within a Social Setting Lin Liu Eric Yu John Mylopolos Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4

[8] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and ldiversity," in In Proc. of IEEE 23rd Intl. Conf. on Data Engineering *(ICDE)*, 2007.

[9] Ram Prasad Reddy and Raju, "A Dynamic Programming Approach for Privacy Preserving Collaborative Data Publishing",Internation Journal of computer applications,volume22-no.4,May 2011.

[10] Balamurugan and Bhuvana, "Privacy preserved Collaborative Secure Multiparty DataMining", Journal of computer science 8 (6)-872-878,2012

[11] Fuad Ali Mohammed Al – Yarimi and Sonajharia Minz "Data Privacy in Data Engineering, the Privacy Preserving Models and Techniques in Data Mining and Data Publishing: Contemporary Affirmation of the Recent Literature"International Journal of Computer Applications , December 2012.

[12] Keke ,"Privacy –preserving Multiparty collaborative Mining with Geometric Data Perturbation" Member, IEEE, and Ling Liu,Senior Member, IEEE.

[13] R.Indhumathi and S.Mohana "Data Preserving Techniques for Collaborative Data Publishing"International journal of Engineering Research & Technology-2013.

[14] Ibrahim Yakut and Huseyin Polat, "Privacy Preserving SVD-Based Collaborative Filtering on Partitioned Data", Department of Computer Engineering, Turkey.

[15] Liang Wang,Ting-Zheng Lai and Tzung-Pei Hong,"Hiding Collaborative recommendation association rules on horizontally partitioned data ", Intelligent Data
Analysis 14 (2010)47-67.