International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 593 - 596

Providing Security in Collaborative Data Publishing from Heterogeneity Attack

Shalaka Ragit[1]
[1]Department of Computer Technology
Yashwantrao Chavan College of Engineering
Wanadongri Nagpur, 441110, India
shalakaragit01@gmail.com

Sagar Badhiye[2]
[2]Department of Computer Technology
Yashwantrao Chavan College of Engineering
Wanadongri Nagpur, 441110, India
sagarbadhiye@gmail.com

Abstract:-In Collaborative data publishing the data is distributed among multiple data providers or data owners. The main concern of collaborative data publishing is while publishing data preserving the individual's privacy. While publishing collaborative data to multiple data provider two types of problems are more likely to occur, first is outsider attack and second is insider attack. The attack, which is performed by people who is not data provider, is called as outsider attack. Whereas attack is performed by colluding data provider who may use their own data records to get the data records shared by other data providers, is called as outsider attack. Insider attack is performed by people who are data provider or data owner. In this paper to overcome the problem of such attacks in collaborative data publishing the encryption strategy can be used such as 3DES which provides individual's data protection by using three keys. Along with MD5 key generation mechanism.

Keywords:- *Collaborative, 3DES, MD5, Publishing & Privacy.*

*****

1. INTRODUCTION

Now a days there is great need of sharing data that contains the personal information in distributed system. Recently Privacy preserving for data analysis and data publishing has gain considerable attention as promising approaches for sharing data but also preserving individual privacy.[1] In Collaborative data publishing when the data is distributed among multiple data providers or data owners, data is anonymized by using two settings. First in which each provider anonymize the data independently (anonymize -and- aggregate, but first method results in potential loss of integrated data utility). Second more desirable way data is to anonymizes data from all providers as if they would come from one source (aggregate and anonymize), for this by using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations. The main goal of Collaborative data publishing is to provide anonymization to the data which has been integrated from multiple data owners without compromising the privacy of the individual records provided by other parties. It considers different types of malicious users and information they can use to perform attacks over individuals of different data owners. Fig 1 shows the working and how attack can be performed in collaborative data publishing in which a data recipient, e.g. provider 4, could be an attacker, who attempts to infer additional information about data provider's records using the published data and respective background knowledge (BK) which can be publicly available external data. By using the anonymization technique data is first modified and then released to the public. This process of anonymization is called as preserving privacy in collaborative data publishing.[1] The attributes of any dataset for privacy is basically classified as three types and they are Key attribute, quasi identifier and sensitive attribute. Where key attribute represents the unique identification of the individual, which is used to uniquely identify the individuals such as SSN(Social security number), name, ID. Quasi- identifier is actually a small segment of information which is not identifier but can be correlated with otherentities of information which is publicly available and find the key attribute. Example (DOB)date of birth, gender, which can be used link publicly available dataset or with other data. Now the last one is Sensitive attribute for example policy detail, salary and diseases.

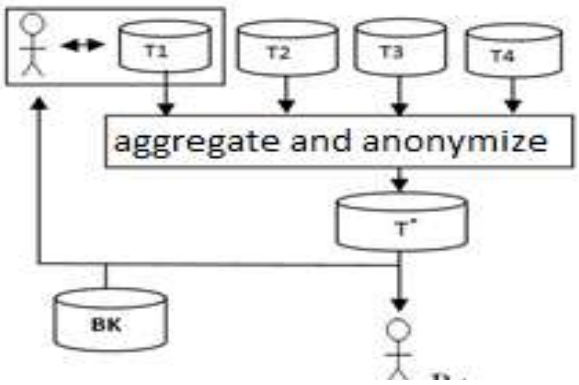

FIGURE 1: Collaborated data publishing

Here the most important goal is to provide a best anonymized view of the integrated data, which is to be published, which will be immune to attacks. Attacks are done by attackers, which can be insider or outsider. These attackers can be in any form such as single or a group of external and internal entities that wants to find out the private information of individuals data using attackers background knowledge as well as anonymized data, which is aggregated data received by the attacker. Privacy compromise is a main concern here. [1] One of the special cases of attack that can be considered here is when data

provider itself performs the attack after making the contribution. So to overcome such types of attacks data is anonymized and released. If suppose p1 is data provided by data provider, which is a subset of some record, it will be anonymized first, that is modified and then released. In this way neither data owner nor data provider could perform the attack.

We do anonymization because data recipient could use receipt data and publicly available data such as Census dataset to perform attacks over the data owner's individual's privacy. We assume the data providers are semi-honest, which are commonly used in distributed computation setting. They can attempt to find out additional information about the data which is coming from other providers by analyzing the data which has been received during the anonymization. [1] In the social network or recommendation setting, a user may attempt to infer personal information about other users using the anonymized data or by using the recommendations which is given by some background knowledge and its own account information. Malicious users may collude by creating artificial accounts as in a shilling attack.

1.1 3DES algorithm

Triple DES uses a "key bundle" that comprises three DES keys, K1, K2 and K3, each of 56 bits (excluding parity bits).
The encryption algorithm is:
ciphertext = EK3(DK2(EK1(plaintext)))
I.e., DES encrypt with K1, DES decrypt with K2, then DES encrypt with K3.
Decryption is the reverse:
plaintext = DK1(EK2(DK3(ciphertext)))
I.e., decrypt with K3, encrypt with K2, then decrypt with K1.
Triple DES runs three times slower than DES, but is much more secure if used properly. The procedure for decrypting something is the same as the procedure for encryption, except it is executed in reverse. Like DES, data is encrypted and decrypted in 64-bit chunks. Although the input key for DES is 64 bits long, the actual key used by DES is only 56 bits in length. The least significant (right-most) bit in each byte is a parity bit, and should be set so that there are always an odd number of 1s in every byte. These parity bits are ignored, so only the seven most significant bits of each byte are used, resulting in a key length of 56 bits. This means that the effective key strength for Triple DES is actually 168 bits because each of the three keys contains 8 parity bits that are not used during the encryption process.

1.2 MD5

The MD5 message-digest algorithm is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications, and is also commonly used to verify data integrity.
MD5 is a one-way function; it is neither encryption nor encoding. It cannot be reversed other than by brute force attack.

2. RELATED WORK

In the field of collaborative data publishing most of the literature survey on privacy preserving data publishing considers different types of attacks which are likely to occur at that current scenario or time and provide different methods to overcome them. Most of them adopt weak methods to protect against specific types of attacks.
S. Goryczka et al [1] proposed m-privacy method which has provided a constraint to protect against privacy of individual.
M. Arafati et al [2] proposed a framework which is cloud based which makes the data provider to integrate data on consumer's demand dynamically in cloud.
S. KumaraSwamy et al [3] propose a model called Association Rule Sharing for Preservation privacy and providing efficiency to Collaborative Data Mining, to overcome the privacy problem this paper presents a (KDLPPDM) which is called as Key Distribution-Less Privacy Preserving Data Mining system. In which association rules is published which is generated by the parties is combined so that combined rule set is formed using the algorithm Commutative RSA.
K. Chen et al [4] provided a privacy protection for Multiparty Collaborative Mining by using method called as Gemetric Data Perturbation.
J. Yang et al [5] propose a modified data anonymization method of slicing which is based on Overlapping Slicing which uses the ides of fuzzy logic.
T. Li et al [6] another approach l-diversity prevent from attribute disclosure attacks, for this it requires that each equivalence group to contain at least l well-represented sensitive values, and another approach is t-closeness, this requires the distribution of a sensitive attribute in any equivalence group to be close to its distribution in the whole Population.
H. Zhang et al [7] present a suite of new techniques that make privacy-aware set-valued data publishing feasible on hybrid cloud. On data publishing phase, they propose a data partition technique, named extended quasi-identifier-partitioning (EQI-partitioning), which disassociates record terms that participate in identifying combinations.
H. Zhu et al [8] to publish anonymized table a new approach is proposed with ℓ- diversity which is against the attacker having publishing algorithm and some individual's sensitive values. This type of an approach replaces the sensitive attribute (SA) value of each record with values set which is consisted of the real SA value and several values which are random noise. The method can be applied to protect numerical and nominal SA; existing additive noise only is used to protect the numerical SA of anonymized table.
S. Kiruthika et al [9] in this the enhanced slicing models has been designed to overcome the drawbacks of slicing. The suppression slicing is done here by suppressing any one of the attribute value of the tuples and then slicing is performed. Here utility by suppressing only very few values are maintained with minimum loss and by random permutation privacy is maintained.
V. Rajalakshmi et al [10] in this paper an augmented Anonymization technique is introduced which has a better performance compared to the existing methods. In this the

data is altered by forming sub-clusters which is followed by an Isometric transformation.

A. Andersen et al [11] The combination of secure multi-party computations (SMC) algorithms, encryption, public key infrastructure (PKI), certificates, and a certificate authority (CA) is used to implement an infrastructure and a toolset for statistical analysis of electronic medical record (EMR) data.

R. Mahesh, T. Meyyappan[12] the authors propose a new method to preserve the privacy of individuals' sensitive data from record and attribute linkage attacks. In the proposed method, privacy preservation is achieved through generalization of quasi identifier by setting range values and record elimination. The proposed method is implemented and tested with various data sets.

Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian [13] shows that l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure.They propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). They choose to use the Earth Mover Distance measure for our t-closeness requirement. They discuss the rationale for t-closeness and illustrate its advantages through examples and experiments.

Yufei Tao , Xiaokui Xiao , Jiexing Li , Donghui Zhang [1 4]This paper deals with a new type of privacy threat, called "corruption", in anonymized data publication. Specifically, an adversary is said to have corrupted some individuals, if s/he has already obtained their sensitive values before consulting the released information. Conventional generalization may lead to severe privacy disclosure in the presence of corruption. Motivated by this, we advocate an alternative anonymization technique that integrates generalization with perturbation and stratified sampling. The integration provides strong privacy guarantees, even if an adversary has corrupted any number of individuals. They verify the effectiveness of the proposed technique through experiments with real data.

Adeel Anjum , Adnan Anjum [15] In order to make the semantic definitions more practical, the research community has focused its attention towards combining the practicalness of syntactic privacy with the strength of semantic approaches such that we may in the near future benefit from both research tracks.

3. CURRENT CHALLENGES

Current challenges in Collaborative data publishing are new types of attacks have not been studied yet which are more likely to occur in today's world. Such as insider and outsider attack. In insider attacks such as vampire attack and zombie attack. Where in vampire attack, attacker miss guides the user by creating multiple copies of the record or data. Where as in zombie attack, attacker continuously take data from file but displays net data as it was but when we open file it contains null data. In outsider attack we can consider Sybil attack and TCP-sync flood attack. In Sybil attack attacker makes fake identity of the user and request for data. And in TCP-sync flood attack too much of packets will be send into the network and because of which network gets jammed and no one can send message to anyone. As the traffic gets jammed attacker collects whole data from network in which data send by data publisher is also present. So to overcome such types of attacks a data encryption algorithm can be applied such as 3DES for providing protection mostly against outsider attacks and to protect against insider attacks we can specify or maintain the behaviour list of such insider attacks which are more likely to occur and every time match the published data with this behaviour list and in this way attacks can be detected and privacy can be provided in collaborative data publishing. Hongli Zhang et al [7] proposed a preservation of privacy over Hybrid Cloud for set valued data by applying data partition technique.

4. PROPOSED METHODOLOGY

To provide security in collaborative data publishing against above mentioned attacks we can use 3DES (Triple Data Encryption Algorithm) is symmetric block cipher, which encrypts data in a block of 64 bit by applying three keys first for encryption, second for decryption and third for encryption of same block. And key is generated using MD5 algorithm.

MD5 is a message digest algorithm which is cryptographic hash function. It is expressed in a 32 digit hexadecimal number as text format. It is basically applied for data integrity. It produces 128-bit hash value. MD5 consists of 64 of these operations, grouped in four rounds of 16 operations. Fig.2 shows the proposed anonymization area means how anonymization is being done with the help of above discussed algorithms.

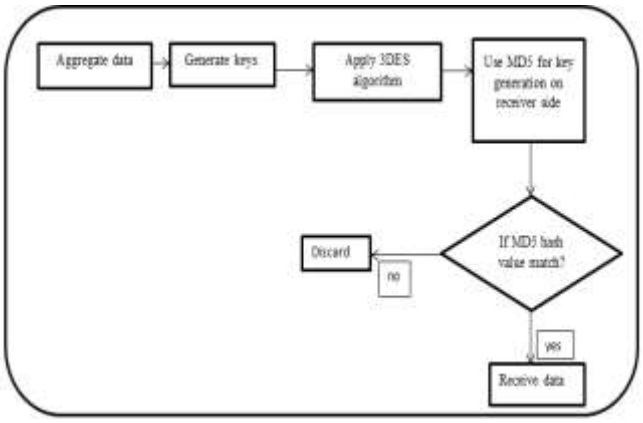

FIGURE 2: Workflow of proposed methodology

5. EXPERIMENTAL RESULT

For implementation of proposed methodology first the dataset was collected and created. First diabetic patient's dataset and other is heart attack patient's dataset. These two datasets are the dataset of publisher. The publically available dataset is census dataset which will be used as background knowledge. We have done the aggregation of data from two datasets. After this attack is performed on this data using the background knowledge, that is census dataset.

Client server connectivity is done and file transfer is done from client to server. Now the file which contains the aggregated data is encrypted using the random generated

key and send to the requester. The requester uses the received key and decrypts the received text file. As the data is send in encrypted form so the insider attack performed is not success full and aggregated data is safely send.

6. CONCLUSIONS

In this paper, a survey of different ways of providing privacy to the individual's data in collaborative data publishing has been discussed. We considered a new type of potential attackers in collaborative data publishing. The different techniques has been used to find the attacks which are most likely to occur at that time and methods has been provided to overcome them and they provide protection against specific types of attacks. This method has provided the best anonymization as compared to other methods. As data is send in encrypted format and key is also safely send, collaborated data is not possible to be decrypted by the attacker.

REFERENCES

[1] S. Goryczka, L. Xiong, and B.C. M. Fung, "m-Privacy for Collaborative Data Publishing", IEEE, 2014.
[2] M. Arafati, G.G. Dagher, B.C. M. Fung, Patrick C. K. Hung,"D-Mash: A Framework for Privacy- Preserving Data-as-a-Service Mashups", IEEE, 2014.
[3] S. Swam, S. Manjula, K. R. Venugopal, S. Iyengar, L. Patnaik, "Association Rule Sharing Model for Privacy Preservation and Collaborative Data Mining Efficiency", IEEE, 2014.
[4] K. Chen, L.Liu,"Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation", IEEE, 2009.
[5] J.Yang, Z. Liu , yangyue ,J.Zhang, "A Data Anonymous Method Based on Overlapping Slicing", IEEE, 2014.
[6] T. Li, Li, J. Zhang, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE, 2012
[7] H. Zhang, Z. Zhou, L.Ye, X. Du,"Towards Privacy Preserving Publishing of Set-valued Data on Hybrid Cloud", IEEE, 2015.
[8] H. Zhu, S. Tian, M. Xie, M. Yang, "Preserving Privacy for Sensitive Values of Individuals in Data Publishing Based on a New Additive Noise Approach", IEEE, 2014.
[9] S. Kiruthika, Dr. M. Raseen,"Enhanced Slicing Models For Preserving Privacy In Data Publication", IEEE, 2013.
[10] V. Rajalakshmi and G. S. AnandhaMala,"Data Anonymization Using Augmented Rotation of Sub-Clusters for Privacy Preservation in Data Mining", IEEE, 2013.
[11] A. Andersen, "An implementation of Secure multi-party computations to preserve privacy when processing EMR data", IEEE 2013.
[12] R. Mahesh, T. Meyyappan, "Anonymization Technique through Record Elimination to Preserve Privacy of Published Data", Pattern Recognition, Informatics and Mobile Engineering, February 21-22.
[13] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and -Diversity", 2007 IEEE.
[14] Yufei Tao , Xiaokui Xiao , Jiexing Li , Donghui Zhang , "On Anti-Corruption Privacy Preserving Publication".
[15] Adeel Anjum , Adnan Anjum, "Differentially Private K-anonymity", 12th International Conference on Frontiers of Information Technology.

\

Authors : Shalaka ragit, Sagar Badhiye