IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Privacy Preserving Big Data Usings Combine Anonymous And Encryption Approach-Survey

Vidhi Desai

Information Technology Engineering Department, Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India

Available online at: www.ijcseonline.org

Accepted: 22/Nov/2018, Published: 30/Nov/2018

***Abstract*—** In today's world each individual wish that his private information is not revealed in some or the other way. Privacy preservation plays a vital role in preventing individual private data preserved from the praying eyes. Anonymization techniques enable publication of information which permit analysis and guarantee privacy of sensitive information in data against variety of attacks. The problem is that information loss and distortion are unavoidable by anonymization job. To reduce the distortion, this paper presents an efficient method that is based on deep anonymization detection. In the method, data publishers analyze the anonymization work, and determine if it is deep or light. If it is thought as deep anonymization, high information distortion is allowed when being distributed to a third party after anonymization. Otherwise, information distortion is kept as low as possible when anonymizing Big-Data to provide the receivers with more meaningful data. The decision for deep anonymization is done by considering a domain data characteristic, data receiver's purpose, and data criticality. Anonymization approaches are used to develop to reduce information loss or increase privacy protection. It aimed to give comparative evolution of the various algorithms. These algorithms are compared for efficiency (in terms of time) and utility loss. We analysis that paillier encryption is more efficient than other algorithms.

***Keywords*—**Privacy, Anonymization, encryption, Big Data

I. INTRODUCTION

Big data has brought revolution in the world of data analytics. Data which was discarded few years ago is now considered to be a powerful asset. Big data is now extensively used for knowledge discovery by normally all the sectors of society. Big data is produced by all of the digital processes and it is stored and shared on web. This poses a very serious security concern. Recently, data explosion incurs one important problem when the data are delivered to a third party. It is how to protect private information from attacker's record linkage or attribute linkage attack. To protect the private data of users, Samaritan and Sweeny have developed the basic theories of data anonymization called k-anonymity and l-diversity. Cryptography is another important aspect to information security in which confidentiality, data integrity and authentication are studied. Depending on complexity of mathematical algorithms, cryptosystems are divided into private-key cryptosystem and public-key cryptosystem. Both systems are controlled by keys. Public-key cryptosystem uses 2 keys; a public key for encrypting the information and a private key for decrypting it. Both keys can be known, but the information will be very difficult to reveal.

Although, the encryption work has been done on l-diversity and k-anonymity but it has not been done on t-closeness yet. So in this survey paper work is to be done on t-closeness.

TABLE I. COMPARISON BETWEEN METHOD

Anonymity Scheme	Description	Weakness/Attack
K-anonymity [1]	At least k number of redundant quasi-identifiers (QIDs) in the dataset; provides anonymity for k−1 individuals.	Homogeneity attack: If sensitive information is homogenous across each record, confidentiality can be compromised. Background attack: With background knowledge about an individual, sensitive information can be identified.
L-diversity [1]	Distribution of a sensitive attribute in each equivalence class has at least 1 "well-represented"	Similarity attack: An adversary can determine likely possibilities of sensitive information. Skewness attack: Sensitive information can be identified in specific parts of data, as distribution of the

	value.	sensitive information in the target data is significantly different than the sensitive information in the remainder of the data.
T-closeness [1]	The frequency distribution of sensitive attributes within each equivalence class should be "close" (t-close, where t is a fixed threshold value) to their distribution of the sensitive attributes in the entire dataset.	Lacks computational procedures to reach t closeness with minimum data utility loss. That is, data utility loss is likely when achieving for t-closeness

Section I contains the introduction of basic approach for weather forecasting. II contain the related works of basic literature papers. Section III contain the methodology and algorithms section IV explain the comparative study between different algorithms, Section V describes proposed system flow and its description and at last conclusion and future scope.

II. RELATED WORK

Tanashri Karle, Prof. Deepali Vora (2017) proposing this performing a fair comparison of anonymization algorithms is inherently a challenging task, since every proposed algorithm uses different settings and metrics. The performance of the algorithms might vary among different combinations of datasets and input parameters (e.g. an algorithm may work well in some experimental configurations and perform poorly in others). As a result, it is important to assess the algorithms by defining a common configuration which reflects parameters that are used in existing evaluations. Furthermore, a comparison requires the use of criteria that can be widely applicable to measure different aspects of the algorithms.

Devyani Patil, Dr Ramesh K. Mohapatra (2017) proposed that after studying and implementation of these (Data fly, Samarati's, Improved heuristic greedy, OLA and Flash) algorithms we can conclude that no algorithm outperforms independent of parameters such as suppression limit. Data fly algorithm has less execution time, information loss and DM value but it gives a local optimum solution. Samarati's algorithm outperforms for higher suppression limit and Flash algorithm outperforms for a larger value of K. SO we cannot say, that particular algorithm is best. Data publisher needs to know in prior the application of data being published which is not possible always. But this evaluation study will surely help to choose an appropriate algorithm with prior knowledge of an application an also it will be helpful for future study for the researchers.

Mohammed Al-Zobbi, Seyed Shahrestani, Chun Ruan(2016) proposed that the increased monitoring, processing and storage capabilities have lead to an explosive growth of big data. However, this is of value only when, for instance, through big data analytics, useful information can be securely extracted. This work presents some of the requirements of the anonymization process for implementation in the big data context to address part of the relevant privacy concerns. This is done through analysis of the contemporary anonymization approaches and identifying some of the reasons for their inefficiencies and potentials for high information loss. In particular, we show how the k-anonymity processes can be made more efficient by taking into account the increased proportion of equivalent records as a result of a high number of records in big data environments.

Sung-Bong Jang (2016) proposed that how to find an appropriate solution to reduce the information loss while protecting privacy when applying k-anonymity and l-diversity to Big Data. To solve the limitations, a method that is based on deep anoymization detection is presented. The future work are to embody the idea, implement the real system, and evaluate the system.

Alia K. Abdul Hassan (2015) proposed that two schemes are suggested for the purpose of avoiding insecurity problem of large size of keys, and made a reliable implementation of Paillier cryptosystem. In the first scheme a set of algorithms and functions were used, and in the other a pre-computation of some values that are required for repeated operations was adopted. The cryptosystem was executed without those two suggestions and its result was compared with those when executing the cryptosystem with the suggested schemes. The comparison proved that the suggested schemes are reliable ways to implement the Paillier cryptosystem with relatively long key size and with encryption time less than decryption time. The decryption process takes time longer than that needed by the encryption process because the mathematical operation in the decryption step is longer.

Jian Wang,Yongcheng Luo,Shuo Jiang,Jiajin Le (2009) proposed that with the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on

privacy preserving. In this paper we have firstly shown that a k-anonymity dataset permits strong attacks due to lack of diversity in the sensitive attributes. Then we have shown l-diversity, a framework that gives stronger privacy guarantees. K-anonymity and l-diversity have been studied widely as mechanisms for preventing re-identification attacks in microdata release. Then we have shown a simple and effective privacy preservation technology called Anatomy. At last we have shown a new privacy measure t-closeness. Besides, we also analyze the merits and shortcomings of these technologies.

III. METHODOLOGY

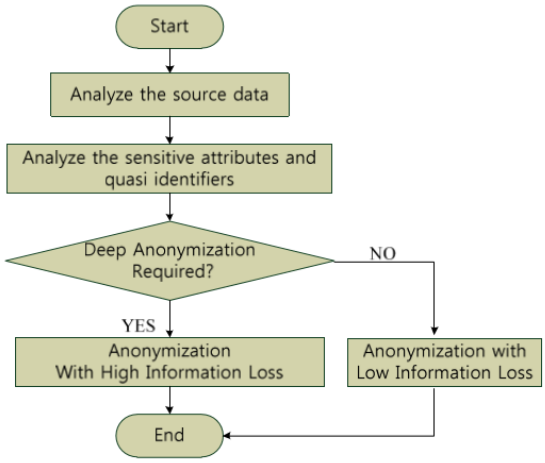

Figure 1. Basic Flow of anonymization

If we want to reduce the information distortion, we have to decrease down the privacy protection. However, if we raise the privacy, much information is distorted and lost. Hence, we need to find the optimal point which satisfy the privacy protection while keeping the data as meaningful as possible. To solve this problem, this paper present a method that is based on deep anonymization detection.

In this scheme, it determines which more important factor is between distortion and privacy protection before starting anonymization. This can be done by analyzing the quasi and sensitive identifiers of the original data with the help of some software tools. or example, if there are too many critical and sensitive information in the data as a result of analysis, privacy protection shall be considered as more important. In this case, it is required to strictly anonymize the original data because there are more frequencies for revealing private information.

First, it is determined by considering the data domain characteristic that represents what kind of data are contained in the original data. For example, suppose that we are going to anonymize medical data. The medical data contains much critical and sensitive information such as disease name, prescriptions, social security number, and treatment. For this data, we assign higher priority to privacy protection than information distortion. However, if the target data include the baseball players' records of past games, we don't have to consider the privacy protection because those data does not contain critical information. In our work, we assign different weight to anonymization work according to target data domain characteristic.

IV. COMPARATIVE STUDY

Table I. Comparison between Feature Extraction Method

Encryption	Advantages	Disadvantages
DES	For encryption, DES uses the 56-bit key. Besides, there are 256 possible keys, which means a brute force attack will never have any impact.	The 56-bit key size is the biggest defect of DES. Chips to perform one million of DES encrypt or decrypt operations a second are available. DES cracking machine can search the entire key space in about 7 hours.
AES	It uses higher length key sizes such as 128, 192 and 256 bits for encryption. -more robust against hacking. -common security protocol.	It uses too simple algebraic structure. Every block is always encrypted in the same way.
MD5	MD5 message-digest algorithm is a widely used hash function producing a 128-bit hash value. It can still be used as a checksum to verify data integrity, but only against unintentional corruption.	In 2004 it was shown that MD5 is not collision-resistant. As such, MD5 is not suitable for applications like SSL certificates or digital signatures that rely on this property for digital security.
RSA	As computing power increases and more efficient factoring algorithms are discovered, the	The user should not worry if public key leak, but need to consider someone takes another's place by counterfeiting

	ability to factor larger and larger numbers also increases. Encryption strength is directly tied to key size.	published false public key. Complexity of the key creation.
PAILLIER ENCRYPTION	Suggested for the purpose of avoiding insecurity problem of large size of keys. Security is More. More than two algebraic use.	Complex implementation.

V. PROPOSED WORK

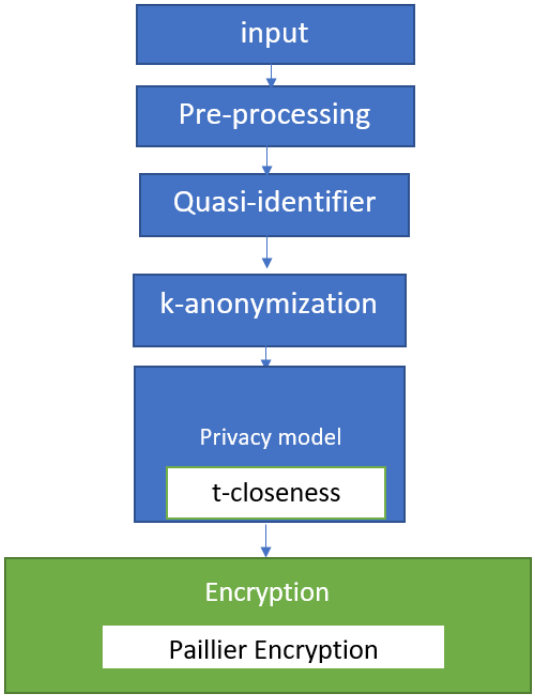

Figure 2: Proposed block

A. *Input:*

Input data is the large data set of structure data. It can be any finance, medical or other large dataset to be handled.

B. *Pre-processing:*

On this large dataset generalization and suppression methods are applied to normalize dataset.

C. *K-anonymization:*

K-anonymity is an efficient method to achieve privacy preservation before releasing data to other party or public. However, to obtain optimal k-anonymity is NP-hard.

D. *T-closeness:*

Some researchers found that the distributions of personal information which have the same level of diversity may provide very different levels of privacy. An equivalence class is said to have t-closeness if the distance between the distribution of Sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

E. *Paillier encryption:*

In Paillier was implemented in order to improving performance over the basic algorithms, with some ideas to find the algorithms with the best performance. The Paillier Cryptosystem is efficient encryption. Paillier Cryptosystem is worthwhile to consider both for the mathematics behind it, as well as for its potential real world applications.

VI. CONCLUSION

This paper shows comparative study of different algorithms and encryption techniques on basis of various parameters and for future analysis paillier encryption is more efficient than other algorithms as it can handle large dataset easily in less time complexity. A detailed study of Anonymization Techniques used in Privacy Preservation in Big Data is done. K-anonymity and its various k-anonymity operators are explained in detail. Performing a fair comparison of anonymization algorithms is inherently a challenging task, since every proposed algorithm uses different settings and metrics. The performance of the algorithms might vary among different combinations of datasets and input parameters. Homomorphic cryptosystems allow for the same level of privacy as any other cryptosystem, while also allowing for operations to be performed on the data without the need to see the actual data. We observe that Paillier scheme is always better than other scheme although, there is no surprise that RSA is the overall fastest, but Paillier scheme fastest probabilistic homomorphic scheme is faster than RSA in decryption because of finding r.

REFERENCES

[1] Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557570, 2002

[2] Turban and J.E. Aronaon. Decision support Systems and Intelligent Systems, Prentice-Hall, New Jersey, USA, 2001

[3] P. P. de Wolf, J.M.Gouweleeuw, P. Kooiman, L. Wil-lenborg, Reflections on PRAM. Statistical data protection, proceedings of the conference, Lisbon, 1998.

[4] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In Proc. of ICDE-2005, 2005 [13] Stephen Lee Hansen and Sumitra Mukherjee. A Polynomial Algorithm for Optimal Microaggregation

[5] Matthias Schmid1 and Hans Schneeweiss, 2005, The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study
[6] M. Lindh and J. Nolin, "Information We Collect: Surveillance and Privacy in the Implementation of Google Apps for Education," European Educational Research J., vol. 15, no. 6, 2016, pp. 644–663.
[7] A. Narayanan and V. Shmatikov, "Myths and Fallacies of 'Personally Identifiable Information'," Comm. ACM, vol. 53, no. 6, 2010, pp. 24–26.
[8] M. Barbaro, T. Zeller, and S. Hansell, "A Face Is Exposed for AOL Searcher no. 4417749," The New York Times, vol. 9, no. 2008, 9 August 2006; www.nytimes.com/2006/08/09/technology/09aol.html
[9] M. Jensen, "Challenges of Privacy Protection in Big Data Analytics," IEEE Int'l Congress on Big Data (Big Data Congress), 2013; doi.org/10.1109/BigData.Congress.2013.39.
[10] ISO/IEC 27040, Information Technology – Security Techniques – Storage, standard ISO/IEC 27040, Int'l Organization for Standardization, 2015; www.iso.org/standard/44404.html.
[11] 2016 Data Breach Investigations Report, report, Verizon, 2016; www.verizonenterprise.com/resources/reports/rp_DBIR_2016_Report_en_xg.pdf.
[12] Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557570, 2002
[13] Turban and J.E. Aronaon. Decision support Systems and Intelligent Systems, Prentice-Hall, New Jersey, USA, 2001
[14] P. P. de Wolf, J.M.Gouweleeuw, P. Kooiman, L. Wil-lenborg, Reflections on PRAM. Statistical data protection, proceedings of the conference, Lisbon, 1998.
[15] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In Proc. of ICDE-2005, 2005 [13] Stephen Lee Hansen and Sumitra Mukherjee. A Polynomial Algorithm for Optimal Microaggregation
[16] Matthias Schmid1 and Hans Schneeweiss, 2005, The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study
[17] X. Zhang, C. Liu, S. Nepal, C. Yang, J. Chen, "Privacy Preservation over Big Data in Cloud Systems," Security, Privacy and Trust in Cloud Systems, pp 239-257, Springer.
[18] J. Sedayao, Enhancing cloud security using data anonymization, White Paper, Intel Coporation.
[19] Top Ten Big Data Security and Privacy Challenges, Technical report, Cloud Security Alliance, November 2012
[20] J. Salido, "Differential privacy for everyone," White Paper, Microsoft Coproration, 2012.
[21] Big Data Privacy Preservation, Ericsson Labs, http://labs.ericsson.com/blog/privacy-preservation-in-big-data-analytics.
[22] O. Heffetz and K. Ligett, Privacy and data-based research, NBER Working Paper, September 2013.
[23] M. V. Dijk, A. Juels, "On the impossibility of cryptog-raphy alone for privacy-preserving cloud computing," Proceed-ings of the 5th USENIX conference on Hot topics in security, August 10, 2010, pp.1-8.
[24] F. H. Cate, V. M. Schnberger, "Notice and Consent in a World of Big Data," Microsoft Global Privacy Summit Summary Report and Outcomes, Nov 2012.
[25] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression," Technical report, SRI International, 1998.
[26] Y. Lindell and B. Pinkas. Privacy preserving data mining. In CRYPTO, 2000.
[27] M. Naor and B. Pinkas. Computationally secure obliv-ious transfer. Journal of Cryptology, 2005

Authors Profile

Miss.Vidhi Desai have pursued Bachelors of Engineering from laxmi Institute of Technology, and currentlypursuing Masters of Engineering in Systems and Network Security from Sardar Vallabhbhai Patel Institute of Technology. She had publish paper in international journal like IEEE.

