# A Review on Privacy Violations of Data Anonymization

[1]Jignasha Kapadiya,[2] Jaimeel Shah
[1]Research scholar, [2]Associate Professor,
*Dept. of Computer Science & Engineering*
*Parul Institute of Engineering and Technology, Parul University, Vadodara, India*

***Abstract*-—** In today's world, privacy/anonymization of users data is of at most important. In this Study uses a clustering algorithm as a pre-process for privacy preserving methods to improve the diversity of anonymized data. 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). We review Paillier`s Encryption and application to privacy preserving computation outsourcing and secure system (e.g. Online voting). Our construction begins with a somewhat homomorphic encryption scheme that works when the function is the scheme's own decryption function. We will show how, anonymization and encryption works together for better privacy preserving in tabular data.

***Keywords—: K-anonymity, l-diversity, t-closeness asymmetric and symmetric***

## 1. INTRODUCTION

Now a days everything is online if we wants to login to the system at that time first we fill registration form in this from we fill our information's, and this data is in tabular form. And this data are public. So we use anonymization techniques to solve problem. Here anonymization provide privacy on some data so we can hide user's information to the other. E.g. medical data, salary data. In this types of data it contains all attributes, like attribute, qusi-idenfifed and sensitive information.

(1) Attributes is identified individuals. E.g. name address and so on.
(2) Quasi-identifiers are nothing but to potentially identify an individual. E.g. Zip-code, Gender, age birthdate etc.
(3) Sensitive information, it is nothing but it identified sensitive data like salary, diseases, Etc


**Figure 1: Data Anonymization Process**

## 2. RELATED WORK

In this paper [1] author has propose a new method called k-anonymity, Data fly and modern algorithm. Data fly algorithm is for synthetic dataset. Mondrian algorithm is for real dataset. And k-anonymity is provides privacy to make QIDs imprecise and therefore less informative. In this paper main focus of the study of privacy preserving technique using anonymization algorithm and detailed comparison of these two algorithms.

In this paper [2] author has Evaluation of Generalization Based K-Anonymization Algorithms. Here author describes about K- anonymization and different Algorithms likes Data fly, Samurai's, improved heuristic greedy, OLA and Flash. (i) Data fly Algorithm: generalization hierarchy is randomly climbed up and each node is checked for K-anonymity. If a node is not K-anonymous then it randomly checks its successor. (ii)Improved Heuristic Greedy Algorithm: - greedily finds a solution starting from zero level node. If a node does not satisfy K- anonymity it checks all its siblings. If no node among siblings satisfies K-anonymity then successors of a node having an identifier with most distinct values.

The process gets terminated

(iii) Samurai's Algorithm: it is binary search approach. It checks middle-level nodes of generalization hierarchy (IV) Optimal Lattice Anonymization (OLA) Algorithm: - This algorithm also starts from nodes at a h/2 level in generalization hierarchy. If a node is K- anonymous, (v) Flash Algorithm this algorithm performs two steps 1) find path 2) check Path. It first convert into binary and then perform all task so it is not important.

In this paper [3] author has described utility-preserving model. It provides Lower information loss and preserve data utility of the data. They examine the issue of health data utility after three anonymization techniques. By evaluating the utility loss of three important privacy preservation techniques with SVM and EMD, they show that today's privacy preservation techniques can significantly jeopardize the data utility due to the highly strict protection principles they impose.

In this paper [4] author has proposed a mainly two new methods (i) k-anonymity and bottom up approach, (ii) Multi-dimensional Sensitivity-based Anonymization Method. By using this methods it avoids the generalization. Reduce the iteration steps and time. Enhance the efficiency of the process. But method uses data sensitivity in its anonymization process and does not generalize the equivalent records, making it more methodical and more efficient.

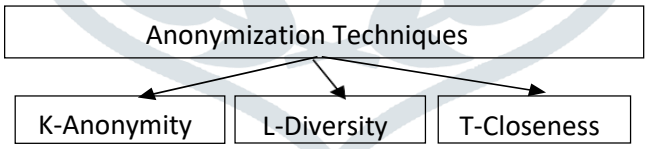
In this paper [5] author works on K- anonymity and also de-anonymization. In this work, they examined the trade-off between sensitivity and semantic of system logs. Since after a certain level of anonymization the semantic of system logs may be lost, keeping the semantic- less data is not the best practice.

In this paper [6] author has described different methods like, (i) Paillier cryptosystem (ii) Cryptography (iii) Data security (iv) Encryption In this paper, 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.

In this paper [7] author has proposed about how to find an appropriate solution to reduce the information loss while protecting privacy when applying k-anonymity and l- diversity to Big Data.

## 3. METHODOLOGY

**Figure 1.1 Data anonymization Types**




Data anonymization technique: there are mainly three types of anonymization technique:

**[1] K-Anonymity:**

Organizations today are entrusted with personal information that they use to serve customers and improve decision making capabilities, but a lot of the value in the data still goes untapped[1]. This data could be invaluable to third party researchers and analysts in answering questions ranging from town planning to fight cancer, there will be impact on the protection against the privacy of individuals.

Data owners want a way to transform a dataset containing highly sensitive information into a privacy-preserving that can be shared with anyone from researchers to corporate partners[1-4]. Increasingly however, there have been cases of companies releasing datasets which they believed anonymized.

This introduction looks at *k*-anonymity, a privacy model commonly applied to protect the data subjects' privacy in data sharing scenarios, and the guarantees that *k*-anonymity can provide when used to anonymize data [4]. In many privacy- preserving systems, the end goal is anonymity for the data subjects [2].Consider the example below:

TABLE 1: ORIGINAL DATA

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Pin Code | Age | Country | Condition |
| 1 | 390002 | 38 | China | Heart Disease |
| 2 | 390041 | 39 | USA | Heart Disease |
| 3 | 390041 | 31 | Canada | TB |
| 4 | 390002 | 33 | USA | TB |
| 5 | 480063 | 60 | India | Cancer |
| 6 | 480063 | 65 | China | Heart Disease |
| 7 | 480023 | 57 | USA | TB |
| 8 | 480023 | 59 | USA | TB |
| 9 | 390002 | 41 | USA | Cancer |
| 10 | 390002 | 47 | India | Cancer |
| 11 | 390041 | 46 | Canada | Cancer |
| 12 | 390041 | 45 | USA | Cancer |

TABLE 2: ANONYMIZE DATA

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Pin Code | Ag e | Country | Condition |
| 1 | 390*** | <40 | * | Heart Disease |
| 2 | 390*** | <40 | * | Heart Disease |
| 3 | 390*** | <40 | * | TB |
| 4 | 390*** | <40 | * | TB |
| 5 | 4800** | ≥50 | * | Cancer |
| 6 | 4800** | ≥50 | * | Heart Disease |
| 7 | 4800** | ≥50 | * | TB |
| 8 | 4800** | ≥50 | * | TB |
| 9 | 390*** | 4* | * | Cancer |
| 10 | 390*** | 4* | * | Cancer |
| 11 | 390*** | 4* | * | Cancer |
| 12 | 390*** | 4* | * | Cancer |

**[2] L-diversity:**

L-diversity anonymity guaranty different values for each group's sensitive attributes. Thus, an attack can recognize a user's sensitive information[1-4]. L-Diversity offers preservation of privacy Requirement of well representation of sensitive data. The k-anonymity algorithms can be acclimatized to calculate L-divers tables. The limitations of k-anonymity approach are resolved by L-Diversity. L-diversity may be difficult and unnecessary to achieve.

**Example 1.** Suppose that the original data has only one sensitive attribute: the test result for a particular result. It takes two values: pass and fail. Further suppose that, there are 10000 records, with 99% of them being pass, and only 1% being fail. Then the two values have very different degrees of sensitivity. One would not mind being known to be tested pass, because then one is the same as 99% of the population, but one would not want to be known/considered to be tested fail. In this case, 2-diversity is unnecessary for an equivalence class that contains only records that are pass. In order to have a distinct2-diverse table, there can be at most 10000×1% = 100 equivalence classes and the information loss would be large. Also observe that because the entropy of the sensitive attribute in the overall table is very small, if one uses entropy l- diversity, l must be set to a small value.L-diversity is insufficient to prevent attribute disclosure.

Below we present two attacks on l- diversity. Skewness Attack: When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure. Consider again[8].

**Example 2.** Suppose that one equivalence class has an equal number of pass records and fail records. It satisfies distinct 2-diversity, entropy 2- diversity, and any recursive (c, 2)-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population. Now consider an equivalence class that has 49 fail records and only 1 pass record. It would be distinct 2diverse and has higher entropy than the overall table (and thus satisfies any Entropy l-diversity that one can impose), even though anyone in the equivalence class would be considered 98% fail, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 fail and 49 pass records, even though the two classes present very different levels of privacy risks. Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information [8].

**[3] t-closeness:**

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a thresh hold t[10]. A table is said to have t- closeness if all equivalence classes have t- closeness.

EMD (Earth mover separation) obliges that those separation between those two probabilistic circulations will be subordinate upon those ground distances around the values of a trait. The primary point for EMD will be that it has the ability should catch those semantic separation between values. The separation between two probabilistic circulations might be measured utilizing world Mover's separation (EMD). EMD obliges that the separation the middle of the two probabilistic circulations will a chance to be indigent upon those ground distances "around those values about a trait.

## 4. COMPARATIVE STUDY

TABLE 3: ANONYMIZATION METHODS

| Sr. No. | METHOD | PROS | CONS |
|---|---|---|---|
| 1. | K-Anonymization[1- 4] | K-anonymity model is simple, intuitive, and well- understood.<br><br>This protects respondents 'identities while releasing truthful information. | k-Anonymity does not provide privacy if Sensitive values in an equivalence class lack diversity<br><br>K-anonymity is difficult to achieve before all data are collected in one trusted place |
| 2. | L-<br><br>Diversity[3,4] | l-diversity works one step ahead of k anonymity in preventing attribute disclosure | l-diversity is more difficult to achieve and also it is not able to provide sufficient protection for privacy |
| 3. | T-Closeness | T-Closeness solve the problem of l- diversity and k-anonymity | with minimum data utility loss |

TABLE 4: PROTECTION METHODS

| Methods | Advantages | Disadvantages |
|---|---|---|
| **Asymmetric** | | |
| (1) AES[12] | AES is more secure.<br>Support Large Key | Complex Implementation.<br>Protocol Support not Provide. |
| (2) DES[12] | Easy to implement.<br>Gives batter Performance. | DES is Less secure.<br>Don't Support Large Key. |
| **Symmetric** | | |
| (1) RSA[12] | High Performance.<br>Fast Process. | Low security because fix algebraic use. |
| (2) Paillers [12] | Security is More.<br>More than two algebraic use. | Complex Implementation |

## CONCLUSION

In review paper we have studied different Privacy Violations of Data Anonymization techniques and encryption techniques. Comparative Study shows that how these methods are different from each other in terms of security, performance and complexity. In future there are combination of these two approach anonymization and encryption is process give rise to security.

## REFERENCES

[1] Tanashri Karle, prof, Dipali vora (2017). PRIVACY PRESERVATION IN BIG DATA USING ANONYMIZATION TECHNIQUES. 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)

[2] Al-Zobbi, M. Shahrestani S. and Ruan, C. (2016). Sensitivity-Based Anonymization of Big Data. 2016 IEEE 41st Conference on Local Computer Networks Workshops (LCN Workshops).

[3] Goswami, P. and Madan, S. (2017). Privacy preserving data publishing and data anonymization approaches: A review. 2017 International Conference on Computing, Communication and Automation (ICCCA).

[4] Jang, S. (2017). A study of performance enhancement in big data anonymization. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT).

5] Karle, T. and Vora, D. (2017). PRIVACY preservation in big data using anonymization techniques. 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI).

6] Shamsi, J. and Khojaye, M. (2018). Understanding Privacy Violations in Big Data Systems. IT Professional, 20(3), pp.73-81.

[7] Alia K. Abdul Hassan . (2014) Reliable Implementation of Paillier Cryptosystem. IJAP, Vol. 10, No. 4, October-December 2014, pp. 27- 29

[8] Ashwin Machanavajjhala, Johannes Gehrke Daniel Kifer, ℓ-Diversity: Privacy Beyond k-Anonymity

[9] Devyani Patil, Dr. Ramesh K. Mohapatra, Dr. Korra Sathya Babu, (2017) Evaluation of Generalization

Based K-Anonymization Algorithms. 2017 IEEE 3rd International Conference on Sensing, Signal Processing and Security (ICSSS)

[10] Dr. Puneet Goswam, Ms. Suman Madan, (2017) Privacy Preserving Data Publishing and Data Anonymization Approaches: A Review. 2017 International Conference on Compu ting, Communication and Automation (ICCCA)

[11] Keerthana Rajendran, manoj Jayabalan, Muhammad Ehsan Rana. A Study on k- anonymity, l-diversity, t-closeness techniques focusing Medical Data (2017), IJCSNS international journal of computer science and network security, vol.17 no.12, December 2017

[12] Ritu Tripathi, Sanjay Agrawal.(2014) Comparative Study of Symmetric and Asymmetric Cryptography Techniques. International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 1, Issue 6, June 2014. ISSN 2348 - 4853