A Study on Privacy Preserving Techniques for Data Access in Cloud and IoT

1st Sampath Yelchuri
Department of Computer Science
Bowling Green State University
Bowling Green, Ohio - 43402
yvenkat@bgsu.edu

***Abstract*—In this paper, our primary focus is to perform a study of privacy-preserving techniques that are used for secured data access in the current emerging technologies i.e., cloud and IoT. Nowadays, most of the applications rely on the services offered by the IoT and cloud. This immense usage of cloud and IoT device(s) results in producing a huge amount of data. Since many users or client store this data in the cloud, it is very important to provide security and privacy to the data. Since cloud and IoT device(s) share the internet as a common medium for their communication, the process of accessing the data has a key role in maintaining the privacy of the user(s). There are many data mining techniques which help in maintaining the privacy of the confidential data. Here we mainly concentrate on the working mechanism of two privacy-preserving techniques. They are k-anonymity and l-diversity.**

***Index Terms*—Privacy, Internet of Things (IoT) and cloud**

I. Introduction

In the real world, centralized computing systems like cloud, fog or edge computing brought enormous changes to the current technologies. The phrase Cloud Computing is characterized as a virtual data storage center that offers easy and ad-hoc access to the data or information that is stored in it [1]. It likewise gives many elastic computing resources to its client(s) or user(s) with the help of various techniques like virtualization, distributed computing and so on. A cloud computing service virtually offers unlimited features such as storage, processing power, privacy resources, and many more. Cloud computing provides an infinite computation power with a shared pool of resources that can be easily accessed over the network. Client(s) or user(s) communicate with the cloud in their everyday life knowingly or unknowingly.

Fig. 1. User Authentication to Cloud using FDS [6].

We all utilize various resources of the cloud by paying for the services offered by the cloud [2]. The cloud offers services to its user(s) or clients(s) in three different models as shown in figure 1. They are SaaS, IaaS and PaaS.

The IaaS provides all its client(s) or user(s) many virtual infrastructure components like computational power, storage capacity for storing their data and much more [7]. The PaaS provides all the cloud user(s) or client(s) a virtual platform to develop, test, run and deploy their applications which could be used by the public [7]. And SaaS provides all the cloud user(s) or client(s) an opportunity to view and access various application(s) or software(s) that are developed and deployed by various people around the world [7]. Since we see that this virtual platform offers many services to various organizations like Google, Microsoft, Facebook, etc. which typically store the information of different user(s) who use these applications. And the cyber attacker is the one who can view the user(s) sensitive data over the network with the help of security attacks.

The term Internet of Things (IoT) is generally regarded as a small processing device that can store a small amount of data or information and has a reasonable computational power as well. The newly emerging 5G technology helps various IoT devices to work effectively as this enables the scope for the IoT devices to communicate or share data at very high speeds when compared with all other networks which are available till date. Although the IoT devices have only a small amount of storage capacity and computational power when compared with the cloud, these devices can use the services offered by the cloud to increase the ability to work efficiently and effectively. That implies services that are offered by the cloud can be easily obtained by the IoT devices [3].

The incorporation of both IoT and cloud helps many people and organizations around the globe to accomplish various tasks effectively and quickly. Since all the IoT device(s) are capable to produce huge amounts of data and it is highly impossible to store and process this entire data in the same device, the cloud offers the IoT device(s) a platform to store and compute the entire data in it. In order to make this happen, there must be an exchange or transmission of data from the IoT device(s) and cloud or vice versa [8]. The

internet is the only common medium through which this exchange or transmission of data can be made. So, as the data moves over different systems there is a high possibility of numerous attacker(s) or an intruder(s) who may have access to the data which is moving on a network. Hence, it is very important to maintain privacy and security for the data. Nowadays, for many organizations, privacy and security of the data are key challenges in the IoT and cloud.

There are several techniques that are used to maintain privacy for the data or information that is being transmitted from IoT devices and cloud and vice versa. One of these techniques is the Fine-Grained Privacy Preserving Query (FGPQ) scheme maintains privacy for the mobile customer(s). This scheme comparatively decreases computational overheads and provides high data transmission rates [4]. To overcome the problems in multi-cloud broadcast in IoT, multi-cloud based outsourced Attribute-Based Encryption (ABE) is used. By compelling the interaction between various clouds, these mechanisms significantly decrease the computational overhead at the resource-constrained IoT device [5].

In this research paper, we survey various privacy-preserving mechanisms used data access in the cloud and IoT. This paper is organized in the following manner: section II will discuss the existing privacy-preserving techniques like k-anonymity and l-diversity which are already implemented in the real world. In section III, we will get deep into the working or flow of both the techniques. The coming section IV will discuss an experiment and its results on a sample data set where we can understand how the privacy of data or information is maintained. Then section V projects some of the advantages and limitations of both k-anonymity and l-diversity techniques and section VI discuss regarding the future scope and conclude the paper.

II. Related Work

Numerous methods and mechanisms have been proposed by researchers for data privacy and security in the IoT and cloud. In this section, we discuss some brief reviews of various contributions that have been made by different researchers.

Tie Qui et al. [9] have contributed by designing a framework to maintain privacy and security of face identification and resolution. In the IoT and fog computing framework, face recognition and resolution mechanisms were facing some privacy and security issues. The objective behind designing this framework was to enhance the computational speed and usage of less bandwidth. In order to implement this mechanism and overcome the issues, different schemes like data encryption, session key agreement and data integrity checking were proposed. The proposed schemes also meet the demand for data confidentiality, integrity, and availability.

Hui Li [10] and their group have presented a secure and privacy-preserving scheme for the data which is generated from the body sensor(s). The main objective of introducing this scheme was to overcome the threats from various interfaces that have the capability to capture and transmit the data which is being generated by different body sensor(s). Using this scheme, the sensitive data which is generated from various body sensor(s) is transformed from linear fashion to multi-dimensional data. For this transformation of the data, they have used the Weighted Euclidean Distance Contrast algorithm. Now, each data that is present in the multi-dimensional format is translated into a numeric digit and then transmitted to the cloud.

L. Wang et al. [11] have proposed a fine-grained data access scheme that uses a ciphertext updating scheme for data transmitting from cloud to IoT device(s) and vice versa. Using multiple policies of the attribute-based encryption scheme, initially the user(s) or client(s) sensitive data is encrypted and then transmitting the data into the cloud. The decryption of the sensitive data only happens when the attributes of the user(s) or client(s) satisfy the access policy. The main advantage of this scheme is that it takes less time for encryption and decryption processes.

Wang et al. [12] have proposed a Hierarchal Attribute-Based Encryption (HABE) framework which enables privacy, secrecy, and fine-grained data access control. In this scheme, every time the client(s) or user(s) desires to access the data on the cloud, the IoT device renews the access policy of the user(s) by updating the key that updates the policy of the user on the cloud. This process of renewing the updated access policy on the cloud server is achieved by transferring the updated key which will not reveal the sensitive data to the cloud server. The decryption of the data is only done when the IoT device(s) and the attributes of the application satisfy the access policy.

X. Li et al. [13] in this paper discuss different anonymity models that are used for maintaining the privacy of the data. The main objective of this paper is to provide privacy for the data of the mobile user(s) with the help of three different models which are known as (k, l, p)-anonymity. Here using this scheme, they have proposed a new algorithm which is known as dynamically structure minimum anonymous sets (DSMAS) that are helpful in hiding the user(s) identity, position, and critical data of a user who is typically changing his/her position from time to time.

III. Methodology

Sharing of information or data over a network plays an important role for research scientists, data mining mechanisms, protection of customer(s) data and many more which are very essential to our societies. As technology progresses, figure 2 represents the experimental growth of data collections. This data collection includes huge amounts

of the user(s) sensitive data that can be gathered from various IoT devices(s). To handle these large chunks of sensitive data, data pullers that are operated autonomously face a challenging task of maintaining the privacy of the data over the network. One of the most difficult tasks would be ”which data can be released in which way” [14]. There are various data mining techniques that can be used and implemented in the real-time systems to enhance the privacy of the data over a network.

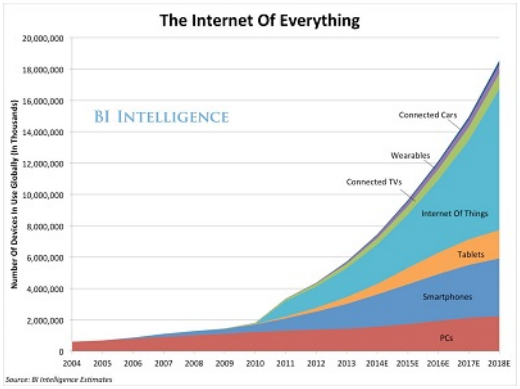

Fig. 2. Exponential growth of data produced through various sources [15]

There are many data or information privacy issues that can be handled when the information is shared among the third-party agencies through the network. The below figure 3 represents a pictorial representation of Bob and Alice who share information over the network where Alice is not a trusted party. Here Bob can share his/her information to Alice where this third party can get some personal information of Bob.

Fig. 3. Two entities share their data over a network [30]

Now in-order to, prevent Alice from extracting or knowing the personal information of Bob we have various privacy-preserving techniques which helps in preventing the exploitation of the sensitive information. Basically, all the privacy-preserving techniques fall into two specific categories as mentioned below:

1. **Data Transformation Techniques.**

2. **Cryptographic Protocol Techniques.**

Cryptographic Protocol Techniques

These techniques follow the approach of performing some cryptographic protocol primitives which help in encoding and decoding of the information which is being shared among various entities. From figure 4, we can observe that Bob sends the information to Alice using various protocols that use cryptographic primitives [30].

Data Transformation Techniques

These are the techniques which are helpful in transforming the data in the datasets which are stored in various data centers. In the above case, Bob uses these transformation techniques on the datasets before sending it to Alice. Now, Alice who is a third-party entity cannot view the detailed information of Bob since the information is has been transformed from its original state. Below figure 5 represents a pictorial representation of Bob sharing the data or information to Alice using the transformation techniques [30].

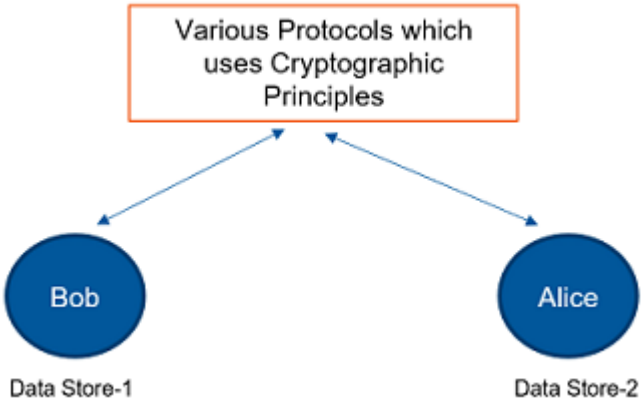

Fig. 4. Two entities use Cryptographic Protocol Techniques [30]

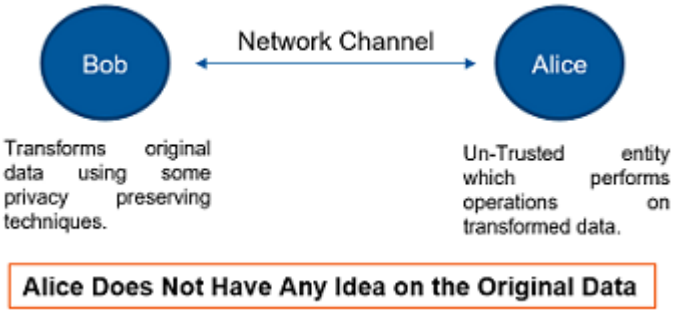

Fig. 5. Two entities use Transformation Techniques [30]

Generally, data server(s) or cloud store entire in the form of datasets. A dataset is a representation of collection rows and columns. Anonymization is defined as a method that makes a deviation of a record to be unidentifiable in a large group of datasets. Each data set in a server consists of at least three different attributes [16]. They are as follows:

Unique Identifiers: These are the attributes that act like a primary key that is helpful in distinguishing an individual

(such as SSN, Driving License).

Sensitive Identifiers: These are the attributes that represent the confidential or sensitive information of an individual in the dataset (such as Disease information).

Quasi Identifiers: These are the attributes that act like a foreign key that is helpful in creating a relationship among different datasets.

So, here we discuss two techniques namely k-anonymity and l-diversity which were proposed by Sweeney [17] and Machanavajjhala et al. [18]. Lets us now investigate the approach that the k-Anonymization technique follows to protect the privacy of the user(s) sensitive data.

K-Anonymity Model

Sweeney [17] in her research explained the approach of this model by taking a real-time scenario where the identity of an entity or a person was not likely to be uniquely distinguishable from a group of health record data-sets which ware stored privately by hospital management. According to Sweeney [17], a k-anonymity model always preserves the privacy of particular record in a data set by maintaining some kind of ambiguity for an attacker or intruder from identifying the identity of the user(s) with the help of different information present in the records of the data-sets. The main property of the k-anonymity model, the sensitive data of every entity or a person which is carried out into the public space cannot be identified from the remaining n-1 entity records present in the datasets [16]. The k-anonymity model is very effective against the linking attacks. The k-anonymity model follows the two internal approaches to Generalization and Suppression [16].

Generalization: The concept of Generalization states that the actual information which is stored in various records of the datasets is transformed from its original values into generalized expression values which are responsible for creating ambiguity for the attackers to fetch the identity of the user(s) [16]. This generalization is formed with the help of Taxonomy Tree as shown in figure 6.

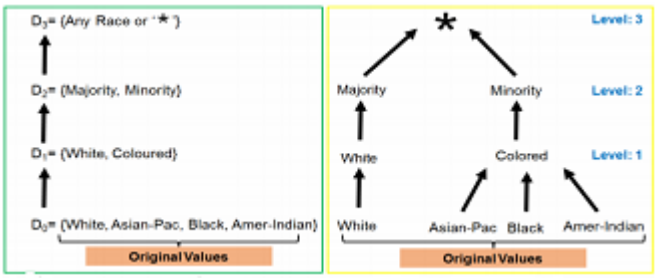

Fig. 6. Taxonomy Tree for the Generalization of the data in the data-set records [19]

Suppression: The term Suppression is defined as the process replacing the original information or values in the datasets with the data or information which is obtained by generalization step [16].

Either of these two processes can be carried out in the global or local scope of a dataset. When applied at global scope, the transformation is carried out for all the records which exist in the datasets. Whereas, when applied on a local scope the effect of the transformation is carried for each and every transaction that tries to access the information from the records of data-sets [17] An another model has been introduced by R.C Wong [20] which is the extended version of the k-anonymity model with the integration of Quasi Identifier.

L-Diversity Model

Machanavajjhala [18] et al, in their discussions has explained the functionality of the l-diversity model. As we have discussed previously each table, or a dataset is a collection of a different set of attributes or identifiers which are as follows:

a. Sensitive Attributes

b. Non-Sensitive Attributes

c. Explicit Identifiers

d. Quasi Identifiers

Generally, each set of attributes or identifiers can be classified into different classes. Each class is said to be l-diversified if and only if each attribute which is sensitive towards the identity of an individual is designated with at least 'l' well-represented values [21]. So, the representation of these values can be done in three different ways. They are as follows:

Distinct *l*-diversity: This type of l-diversity class represents a collection of unique values in a dataset which reduces the amount of redundant data in a table [21]. Due to the redundancy of the data in a table, it is possible for an attacker to infer the identity of an individual by making different probabilities.

Entropy *l*-diversity: Entropy is defined as a function which states that each class said to have entropy l-diversity if and only if the dataset has at least a value of log (l) entropy [22].

Recursive (c, *l*)-diversity: The main purpose of this way of representing the l-diversified values is that the class does not represent the values of the dataset which are being used more repeatedly [22].

IV. Experiment and Results

So, here in this section, we discuss an experiment that we have conducted using a sample data of which represents a kind of employee database of a software organization. To move forward with this experiment, we make use of the methodology and an application which was developed by N. Li et al. [23].

Basically, the main objective of performing this experiment is to analyze the anonymity of the information stored in various datasets in the form of tables. And the main idea behind performing this task is to represent a dataset with sensitive data in such a fashion that the attackers cannot extract sensitive data out of it.

Initially, we have considered a database with a table which stores information regarding all the employees working in a software organization. Some of the information includes the attributes like Employee Name, Employee ID, and Department ID and etc as shown in figure 7. So, if a third-party agency that is authorized to view the information on a network, then we make use of the two models i.e. k-anonymity and l-diversity of representing the data in the employee database.

S No	ID	Employee Name	Age	Dept.
1	1134	John	31	DEV
2	1156	Mike	27	SPT
3	6457	Alex	45	TST
4	6424	Maxwell	42	DEV
5	6445	Martin	41	TST
6	2146	Ian	29	TST
7	2156	Lauren	34	DEV

Fig. 7. Employee database of software organization [23]

So, considering this as our master dataset we first perform the generalization and suppression of data which is present in the table. Here we have considered Employee Name as the unique and sensitive attribute and ID, Age and Dept as quasi-identifiers.

When the above table was generalized, the below figure 8 represents the way that the information in the table being generalized and suppressed.

S No	ID	Age	Dept.
1	11**	31	DEV
2	11**	27	SPT
3	64**	4*	TST
4	64**	4*	DEV
5	64**	4*	TST
6	21**	29	TST
7	21**	34	DEV

Fig. 8. Generalized and suppressed information of Employee Database [23]

So, once most of the columns are generalized and suppressed, then we have grouped or clustered the columns. This is done due to the reason that we can observe in the above figure 8 that the Age attribute still has some values which are directly or independently visible to the attackers. Hence, we form a different cluster(s) or group(s) in the table with the help of Age, Dept. attributes and replace the actual Age values with some generalized expression using the comparison symbols as shown in figure 9.

S No	ID	Age	Dept.
1	11**	>30	DEV
2	11**	<30	SPT
3	64**	4*	TST
4	64**	4*	DEV
5	64**	4*	TST
6	21**	<30	TST
7	21**	>30	DEV

Fig. 9. Clustered information of generalized and suppressed Employee Database [23]

So, here we can observe the privacy level of the data or information which is stored in the Employee database. Therefore, now we can publish the data to the third-party agencies that can perform some sort of analysis according to their requirement which is presented in the form figure 9 without knowing the identity of the individual.

V. Advantages and Limitations

In this section, we mainly focus on various advantages and limitations of both k-anonymity and l-diversity.

Advantages of K-Anonymity

- This technique is helpful in preserving the exposure of the character by forming conditional links to various datasets which have data under 'k'. This kind of preserving approach avoids the attacker from linking up to the sensitive information of a user from various other data sources [24].
- Using this kind of approach is very efficient in terms of cost when compared with designing an algorithmic approach which takes a lot of time for the development [25].
- Some of the well-known web-services like Incognito and Mondrian make use of this approach effectively for publishing the data across networks [26].

Limitations of K-Anonymity

As we know that any approach or technique in the world of security is not complete. Even this approach has got its own flaws where it could not handle some attacks like unsorted

matching and temporal attacks. Here we discuss an attack as mentioned below that had some exceptions:

- **Homogeneity Attack:** Suppose let us think that there are two persons 'X' and 'Y' who are neighbors to each other. Now, let us say that person 'X' has the information of person 'Y' age and he/she location of stay which has the same zip-code. So, now it becomes very simple for person 'X' to decode the sensitive information of person 'Y' if the values of the Sensitive Attribute remain same for all the records. Thus, we can observe that this technique needs to be enhanced more by using the quasi-identifiers [27].

Advantages of L-Diversity

The below mentioned below are some of the benefits by which the limitations of the K-Anonymity can be resolved:

- This kind of technique helps to prevent disclosing of an attribute [28].
- The L-Diversity technique is more efficient than K-Anonymity [29].
- The main advantage of this technique is that it makes the sensitive data to be represented in a more distributed manner [27].

Limitations of L-Diversity

The main disadvantage of this technique is that it does not produce effective results when there are huge similarities in the sensitive attributes [29].

VI. Conclusion

The main goal behind using k-anonymity and l-diversity techniques is to maintain the secrecy and privacy of the customer(s) data which is being produced through various IoT device(s) and is stored in the cloud. This paper is completely a survey on how the two data mining techniques i.e. k-anonymity and l-diversity enhance the privacy of sensitive information in the data centers. In future, we would also work on the various other privacy-preserving techniques which play a crucial role in maintaining the privacy while sharing the user(s) information with an un-trusted party. We would also work on the second category of privacy-preserving techniques which is Cryptographic Protocol Techniques and make a comparative study between both Transformation and Cryptographic Protocol Techniques.

References

[1] National Institute of Standards and Technology, NIST, 30-Jan-2019. [Online]. Available: https://www.nist.gov/. [Accessed: 11-Feb-2019].

[2] Rajaraman, V. (2014). Cloud computing. Resonance, 19(3), 242-258. doi:10.1007/S12045-014-0030-1.

[3] Liu, Y., Dong, B., Guo, B., Yang, J., Combination of cloud computing and net of things (IOT) in medical observance systems, International Journal of Hybrid info Technology, 2015.

[4] Yang, Xue, Fan Yin, and Xiaohu Tang.A Fine-Grained and Privacy-Preserving Query Scheme for Fog Computing-Enhanced Location-Based Service, Sensors, vol. 17, no. 7, p. 1611, 2017.

[5] L. Yang, A. Humayed, and F. Li, A multi-cloud based privacy-preserving data publishing scheme for the internet of things, Proceedings of the 32nd Annual Conference on Computer Security Applications - ACSAC 16, 2016.

[6] Nexiiblog, SERVICE MODELS IN CLOUD COMPUTING, NexiiLabs Blog. [Online]. Available: http://nexiilabs.com/blog/service-models-in-cloud-computing/. [Accessed: 15-Feb-2019].

[7] Samuel J. Morales, Myupdate Studio. [Online]. Available: https://myupdatestudio.com/cloud-computing-the-ins-and-outs/. [Accessed: 19-Feb-2019].

[8] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, Security and Privacy Preservation Scheme of Face Identification and Resolution Framework Using Fog Computing in Internet of Things, IEEE Internet of Things Journal, vol. 4, no. 5, pp. 11431155, 2017.

[9] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, Security and Privacy Preservation Scheme of Face Identification and Resolution Framework Using Fog Computing in Internet of Things, IEEE Internet of Things Journal, vol. 4, no. 5, pp. 11431155, 2017.

[10] H. Zhu, L. Gao, and H. Li, Secure and Privacy-Preserving Body Sensor Data Collection and Query Scheme, Sensors, vol. 16, no. 2, p. 179, 2016.

[11] Q. Huang, Y. Yang, and L. Wang, Correction to Secure Data Access Control With Ciphertext Update and Computation Outsourcing in Fog Computing for Internet of Things, IEEE Access, vol. 6, pp. 1724517245, 2018.

[12] Q. Huang, L. Wang, and Y. Yang, DECENT: Secure and fine-grained data access control with policy updating for constrained IoT devices, World Wide Web, vol. 21, no. 1, pp. 151167, 2017.

[13] C. Piao and X. Li, Privacy Preserving-Based Recommendation Service Model of Mobile Commerce and Anonimity Algorithm, 2015 IEEE 12th International Conference on e-Business Engineering, 2015.

[14] S.Vijayarani, A.Tamilarasi, and M.Sampoorna, Analysis of Privacy Preserving K-Anonymity Methods and Techniques, International Conference on Communication and Computational Intelligence, pp. 540545, Dec. 2010.

[15] T. Danova, THE INTERNET OF EVERYTHING: 2014 [SLIDE DECK], Business Insider, 24-Oct-2014. [Online]. Available: https://www.businessinsider.com/the-internet-of-everything-2014-slide-deck-sai-2014-2. [Accessed: 01-Mar-2019].

[16] S.Dhanalakshmi and P.S.Ahammed Shahz Khamar, Data Preservation Using Anonymization Based Privacy Preserving Techniques A Review, IOSR Journal of Computer Engineering, pp. 1821.

[17] L. Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557570, 2002.

[18] Machanavajjhala, A, Kifer, D, Gehrke, J, and Venkitasubramaniam, M, l-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), 2007.

[19] A. Majeed, F. Ullah, and S. Lee, Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data, Sensors, vol. 17, no. 5, p. 1059, 2017.

[20] Wong, R. C. W., Li, J., Fu, A. W. C., and Wang, K., (, k) -anonymity: an enhanced k-anonymity model for privacy preserving data publishing, 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 754759, 2006.

[21] Keerthana Rajendran, Manoj Jayabalan, and Muhammad Ehsan Rana, A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data, IJCSNS International Journal of Computer Science and Network Security, vol. 17, no. 12, 2017.

[22] J. Han, H. Yu, and J. Yu, An improved l-diversity model for numerical sensitive attributes, 2008 Third International Conference on Communications and Networking in China, 2008.

[23] N. Li, T. Li, and S. Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, 2007 IEEE 23rd International Conference on Data Engineering, 2007.