International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 17
July 2018

# Anonymity Preserving Privacy: An Overview

Radhika Tayal[1]
*NIU, Greater Noida*
radhikatayal@gmail.com

Dr. Rajdev Tiwari[2]
*GNIT, Greater Noida*
rajdevtiwari@yahoo.com

*Dr. Suryakant yadav*[3]
*NIU, Greater Noida*
suryakantyadav11@gmail.com

***Abstract:***

*In this paper, we can focus on all the different methods used by the researchers to disclose the identification disclosure of individuals. To prevent the identity of an individual many researchers convert the original data into anonymized dataset using different techniques namely K-anonymity, L-diversity and T-closeness. Many organizations & institutes use public data for their personal interest It leads to violation of data privacy of some individuals, there are many cases that even after removing private data, such as Name, Address, Individual privacy can be comprised by combining attributes from the database. These joined attributes are named as Quasi-identifier. Here we are addressing different techniques to preserve the privacy of an individual at the same time; we can use that data for finding trends, public benefits and more.*

***Keywords***

*Quasi Identifier, Sensitive Attribute, K-anonymity, L-Diversity, T-closeness, Privacy Preservation*

## 1. Introduction

In today's world due to the rapid improvement in the field of information technology, the World Wide Web (WWW), fast development of technology and cheap storage. All organizations are capable to store their data and used it for analysis purpose. For example hospital stores the patient information, supermarket stores the customer data, and government organization stores the information on voter registration, census and many more. As a result, we are able to generate terabytes of the data. Undoubtedly these data contain non-aggregate information of individuals. This data also contains some sensitive attribute, so that we can uniquely identify the identity of an individual. This is unacceptable with respect to user privacy. The main question arises "How we can effectively use this data by preserving the privacy of an individual at minimal cost and computation"? Privacy-preserving is only the method to limit the diffusion of individual personal data. Privacy can refer to an individual where nobody should know about any entity after performing data mining about a collection of entities. It mainly focuses to avoid the identification of an individual and the sensitive information cannot be brought to the attention of others.

## 2. Motivation

The main motivation behind this privacy preservation is that, we are able to collect the large amount of data due to both online and offline media. We are generating more data with digital sensors, devices, cameras, computers, and networks. We preserve more data since storage has become cheap and plentiful. However, despite in all the methods, a common understanding of what is meant by privacy is still missing, so we are trying to address this problem.

Especially in medical field, private information of an individual is collected, stored and processed in a variety of application domain. This private information has stored to provide the better quality service, improve lifestyle, provide data to researcher, Phramacetual Company and its application is crucial in many contexts. For example, drug companies and researchers may be interested in patient records for drug development. Researcher and doctor are interested to find the root cause of the disease. Healthcare people are interested to prevent medical errors and enhance patient care. Such additional usages of data are important and should certainly be supported. Yet, the privacy of the individuals to whom the data is related should be assured as well. To address the conflicting requirements of assuring privacy, at the same time supporting legitimate use of the data is the main concern. To solve this issue anonymization methods have emerged as an important tool to preserve individual privacy when releasing privacy sensitive data sets. Among all the user's privacies protection techniques K-Anonymity [1] has become a most prominent method for its easy implementation and extension. But all these methods and its extensions are suffered from a drawback that they decrease the utilization of anonymous datasets.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 17
July 2018

However, improving utilization of anonymous datasets is very important for many users. In this paper, we are addressing all the techniques used by the researchers to preserve the privacy of an individual.

## 3. Privacy Preserving Data Publishing (PPDP)

In recent years, PPDP has studied a lot about publicly release the data set with maximum anonymization and minimum loss of information. In PPDP researches study that, how we can publish the sensitive data so that it maintains the identity of an individual. Here, we can perform the operation in such a way the data has been released should be anonymized. At the same time, the data should be rich enough, so it can be used for data mining purpose [2]. Basically, PPDP consists of two phases namely data collection and data publication. I n data collection phase, the original data from record holders is retrieved by the data publisher. For example, the data publisher (hospital) collects the information from record owners and gives it medical center (Data recipient) for research and analysis purpose. Data can be collected online as well as offline media. While collecting the data, when the data recipient is willing to provide their personal information to data publisher, then a type of data publisher is called trusted data publisher. On the other hand, if the publisher is not reliable and tries to gain the confidential information of an individual, then type of data publisher is called untrusted data publisher. The main objective of the data recipient is to perform data mining to retrieve some useful information. In data publishing phase, the data retrieved by record holders in the data collection phase, is released to data recipient for analysis and mining purpose. In our thesis, we are mainly focused on how to publish the data of an individual with a maximum of anonymization and minimum loss of information. While publishing the data, we mainly focused on these two terms sensitive attribute and quasi identifier that are described below.

## 4. Basic Terminology

**Table 1 2 -Anonymous Patient Data**

| ZIP | AGE | DISEASE |
|---|---|---|
| 500* | 2* | Stomach Cancer |
| 500* | 2* | Gastric Ulcer |
| 501* | >40 | Flu |
| 501* | >40 | Gastritis |
| 502* | 3* | Stomach Cancer |
| 502* | 3* | Stomach Cancer |

1. Attribute Identifiers: Let T= $\{t_1, t_2 ... t_m\}$ be a table that contains the information of an individuals. Each table contains a set of attributes A = $\{A_1, A_2 ... A_n\}$. Here we defined three type of attributes in A, named as explicit identifiers, quasi-identifiers, and sensitive identifiers.
2. Explicit identifier: An attribute $A_i$ is labeled as, explicit identifier, if it can be used to uniquely identify an individual. For examples, social security number and name are defined as a sensitive attribute. To preserve the privacy of the published data we assume that the explicit identifier attributes undergo a transformation process such as randomization. In table 2 names is the explicit attribute.
3. Quasi-identifiers: A set of attributes $\{A_1, A_2 ... A_n\}$ of a table T is called a quasi-identifier set, if these attributes can be linked with external data to uniquely identify at least one individual in the general population. It is assumed that generally domain experts defined the quasi-identifier based upon the specific knowledge of the domain. For example combination of all these attributes (Age, ZIP) may use to determine an individual record from the table to his/her medical problem as shown in Table 1.
4. Sensitive-identifier: An attribute that contains extremely personal information. Such as salary, disease state, etc. In other words, we can say that the sub-class of quasi-identifier is defined as a sensitive identifier. As shown in Table 1 disease is the sensitive attribute.
5. Frequency Set: Let Q = $\{A_1, A_2 ... A_q\}$ be a subset of A. The frequency set of table T with respect to Q is a mapping from each unique combination of values $\{\{V_1, V_2 ... V_q\}$ of Q in T (the value groups) to the total number of the tuple's in T with these values of Q. In other words, the frequency set of T with respect to Q stores the set of counts of each unique combination of values of Q in T.
6. Generalization: The basic idea of generalization is to re-identify quasi-identifier attributes of table T. In other words, we can say that quasi-identifier is replaced by less specific but semantically consistent values. For examples, generalization includes generalizing zip code values by replacing the last digit with a wildcard (i.e. *). Original ZIP codes {50032, 50039} can be generalized to 5003*, Date of birth to generalized to year only while hiding month and date value.
7. Suppression: Suppression is similar to generalization but in this value of quasi-identifier value is completely hidden for the table. Suppression is mainly classified into three types.

(i). Record Level: When the complete entry of a record from the table is eliminated or suppressed.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 17
July 2018

(ii). Value Level: All instance or records of a particular value in the table are suppressed.
(iii). Cell Level: Some of the records for a given value are suppressed in a table.

## 5. Anonymization Techniques

### 5.1. K-anonymity

The initial definition of anonymization is given by Sweeney, namely k-anonymity [8]. K-anonymity model was first described by Sweeney and Samarati and later expanded by Sweeney [9] in the context of data table releases. It was the primary model proposed for anonymization and it is the base from which further expansion has been developed. According to the Sweeney, K-anonymity is defined as a process, in which each row in the database is identical with at least (k-1) other rows. At this point, the database is said to be k-anonymous. The definition of k-anonymity is as follow.

Let RT be the relational table of attribute $\{A_1, A_2 \dots A_N\}$ and $QI\_{RT}$ be the quasi-identifier associated with it. The relational table RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI\_{RT}]$ appears with at least k occurrences in $RT[QI\_{RT}]$. It can also be trivially proven that if the released data RT satisfies k-anonymity with respect to the quasi-identifier $QI_{PT}$, then the combination of the released data RT and the external sources on which $QI_{PT}$ was based, cannot link on $QI_{PT}$ or a subset of its attributes to match fewer individuals.

#### 5.1.1. Attack on K-anonymity

**Table 2 Voter Information of john**

| John's Information | |
|---|---|
| ZIP | AGE |
| 50037 | 37 |

As k-anonymity requires each tuple in (the multistep) T [QI] to appear ≥ k times but does not say anything about the sensitive attribute. In this type of attack, an attacker gains some information about his sensitive attribute from the released table, even though attacker is not able to link the victim with any individually published record [8]. As shown in Table 1 and 2, attacker can find that all the male having age 30 whose lives in particular areas are suffering from Stomach Cancer. So (ZIP, male, 30) attribute is having confidence 100 \% Stomach Cancer by this information. It found that family suffers from Stomach Cancer. To prevent from attribute linkage attack Machanavjjhala [3] proposed a technique named L-Diversity.

### 5.2. L-Diversity

The L-diversity model (Distinct, Entropy, Recursive) [3, 4, 5] is an extension of the k-anonymity model which diminishes the granularity of data representation utilizing methods including generalization and suppression in a way that each equivalence class of publicly released table has at least *l* different values for each sensitive attribute. In other words, we can say that sensitive attributes must be "diverse" within each quasi-identifier equivalence class. The L-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting the corresponding sensitive values that were generalized or suppressed, particularly when the sensor values in a group exhibit homogeneity. The L-diversity model includes the promotion of intra-group diversity for sensitive values in the anonymization mechanism. The problem with this method is that it depends upon the range of sensitive attribute. I want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. Also, L-diversity method is subject to skewness and similarity attack [1] and thus can't prevent attribute disclosure.

L-diversity requires that each equivalence class of publicly released table has at least L different values for each sensitive attribute and the released table satisfies L-diverse property if for all qid group.

$$C = \sum P(qid; s)\log\big(P(qid, s)\big) \geq \log(1)$$

Here S is a sensitive attribute; P (qid, s) is a part of records whose sensitive value is *s* for the total records whose equivalence class is group denoted by qid. The more uniformly distributed sensitive values in each equivalence class group qid higher will be the entropy of sensitive attribute. So higher value of entropy in the released table, lesser is the chance of probabilistic attack, a higher value of threshold *l* increases its privacy and lesser is the information gained by an attacker from a released table.

#### 5.2.1. Similarity Attack on L-Diversity

The major drawback of L-Diversity is that it cannot differentiate between the equivalence classes.

Equivalence class 1: 49 HIV+ and 1 HIV.

Equivalence class 2: 1 HIV+ and 49 HIV.

L-diversity does not consider semantics of sensitive values. For example ss shown in the Table \ref{Similarity Attack on L-Diversity}, we know that bob has Zip code = 500032 and age = 27. So we can conclude that bob's has some stomach-related disease. To overcome this problem "Li et al. ICDE '07 [7] proposed a technique named *T-closeness*.

### 5.3. T-closeness

According to him the distance between the distribution of a sensitive attribute in the equivalence class and the distribution of the attribute in the whole table is no more than a threshold of t. In other words, we can say that the distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database. It is a further improvement of an l-diversity group based anonymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation. This reduction is a trade-off that results in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model(Equal/Hierarchical distance) \cite [6, 7] extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

An equivalence class is said to have t-closeness if the distance between the conveyance of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. The main advantage of t-closeness is that it intercepts attribute disclosure. The problem lies int-closeness is that as size and variety of data increases, the odds of re-identification too increases. The brute-force approach that examines each possible partition of the table to find the optimal solution takes n O(n) mO(1) = 2O (nlogn) mO(1) time. We first improve this bound to single exponential in n (Note that it cannot be improved to polynomial unless P = NP) [7].

## 6. Conclusion

By study different anonymity methods, we conclude that by using different anonymization techniques, we can prevents the privacy of an individuals from some strong attacks by apply different techniques in the sensitive attributes. At the same time privacy and utility are duals of each other. As we can protect the data from breaching the privacy of an individual, the utility of the data can decrease. We concluded that T-closeness method could lead to stronger notions of anonymity and to notions which can measure the effectiveness of introducing dummy data or dummy queries to further enhance the security of personal data. By using different techniques on the record of an individual have higher disclosure risks than most other records. Such records could be the outliers in the dataset, and they may severely degrade the quality of anonymization. Users can then eliminate these sensitive records with high disclosure risks.

## 1. References

[i] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* , 571-588.

[ii] Kulldorff, M. (2001 ). Prospective time periodic geographical disease surveillance using a scan statistics. *Royal Statistical Society* , 61-72.

[iii] Machanavajjhala, J. G. (2006). L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)* , 24-28.

[iv] A. Ton, M. S. (27, April 2015). *evolving focus in Internet security*. Retrieved from http://www.ericsson.com/research-blog/data-knowledge/big-data-privacy-preservation/2015

[v] Benjamin C. M. Fung, K. W.-C. (2010). *Introduction to privacy preserving data publishing concepts and techniques.* London: Champan & Hall/CRC press.

[vi] Lawrence K. Saul, Y. W. (2004). *Advances in Neural Information Processing Systems 17.* Vancouver: MIT Press.

[vii] N. Li, T. L. (2007). T-Closeness: Privacy Beyond k-Anonymity and L -Diversity. *IEEE 23rd International Conference on Data Engineering, Istanbul* , 106-115.

[viii] P. Samarati, L. S. (1988). *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.*

[ix] Sweeney, L. (2002). K-anonymity, a model for protecting privacy. *International Journal Uncertain Fuzz* , 557-570.