Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 31-34

A Study of Privacy Preserving Using Anonymization Techniques

S. Renuka
Department of Computer Science, Government Arts College, Trichy, Tamil Nadu, India
E-Mail: spkumarrenu@gmail.com

Abstract **- Now a day's there is an extensive use of technology that has led to a massive increase in the amount of data that is generated. The analysis of such information will help the business and organization in various ways and also contributing beneficially to society in many different fields. As this data also contains the considerable amount of user-sensitive and private information, it will lead to the potential threats to the user's privacy if the data is published without applying any privacy preserving techniques to the data. This paper discusses the various anonymization techniques such as generalization and suppression which are used to preserve privacy during data publishing.**
Keywords: **Privacy, Anonymization, Suppression, Generalization, K-Anonymity**

I. INTRODUCTION

Many organizations like credit card companies, hospitals, and Government sectors collect and hold a large volume of data. These data has created a good opportunity for decision making. This information contains sensitive data about an individual like credit card details, ID number, disease etc. Privacy relates to safely disclosing the sensitive data without leaking the private and sensitive information regarding the authorized owner [14]. This sensitive information can be removed from the dataset without affecting the privacy of individuals. Some benefits of the information technologies are only possible through the collection and analysis of sensitive data. Preservation of privacy may be achieved by anonymizing the records before publishing. For example, a Hospital is connected with a research institution which collects the patient history in a research database, to guarantee the maximum privacy to each patient, the medical records only sends to the research database an anonymized version of the patient record. Thus the database used by the researchers is anonymous. Privacy-preserving has key research issue and receiving attention from the researcher [1] [2].

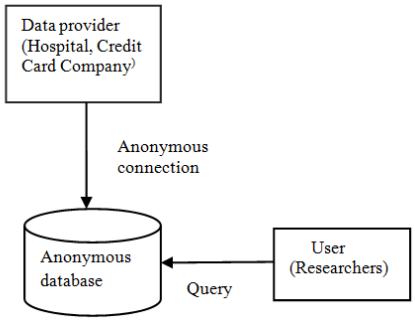

Fig.1 Anonymous Database

In fig.1 shows Data providers provide their data to Anonymous database system. The user can be treated as researchers who have access to anonymous database. The data provider's privacy is protected from these researchers. The communication between the data provider and database occurs through an anonymous connection, as provided by the protocols like Crowds [16] or Onion routing [15]. Anonymization is an approach for preserving the privacy of an individual. In this study, the various anonymization techniques are discussed in the following section.

II. DATA ANONYMIZATION

Anonymization is an important tool to preserve privacy when large volumes of data released the sensitive dataset [3][5]. It is the process of removal of personal identifying information and to preserve personal information owning the data or other parties. Anonymization means identifying information is removed from the original data to protect personal information using encryption. Non-anonymized version of data deleted from sender side after it is being sent to receiver side. There are many ways to perform data anonymization. We only focus on the k-anonymization approach [28]. An original dataset consists of four types of attributes.

TABLE I TYPES OF ATTRIBUTES

Set of Attribute	Description	Example
Explicit Identifiers (ID)	Used uniquely to Identify the Individuals	Name, ID Number, Mobile number
Quasi Identifiers (QID)	Potentially Identifies the Record Owners	Gender, Age, Zip code, DOB
Sensitive attribute(SA)	Person-specific sensitive Information (This attribute is useful for the purpose of data analysis)	Diseases, Income
Non Sensitive attribute(NSA)	All the attributes that do not fall into the previous categories are called a non sensitive attribute	

Table I shows set of attributes description with example. Before being published the table to others, the first step the table is to anonymized that is Explicit Identifiers are modified. A survey shows that approximately 87 percent of

USA citizens can be re-identified with the help of Quasi-Id like date of birth, gender, zip code [6].

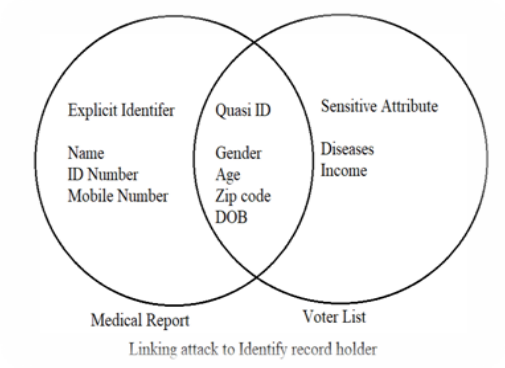

Fig. 2 Linking attack

Fig. 2 shows the person can be re-identified with the help of DOB, Zip code, Age, and Gender attributes when linked with the voter list database to the published medical report.

TABLE II AN ORIGINAL DATASET

Name	Gender	Zip code	Age	Disease
Reena	Female	444805	45	TB
Ram Nivas	Male	424806	46	Diabetes
Kavitha	Female	424809	58	Fever
Neha	Female	444806	65	Cancer

Table II shows an example. In this table Name is key attribute, disease is sensitive attribute, gender, zip code, and age are quasi-identifier.

TABLE III DATASET AFTER REMOVING IDENTIFICATION

Gender	Zip code	Age	Disease
Female	444805	45	TB
Male	424806	46	Diabetes
Female	424809	58	Fever
Female	444806	65	Cancer

In table III, explicit attribute as name removed and released it. However, removing identity attribute is not enough as an attacker may have knowledge of the individual in publish table. An attacker can gather this information from personal knowledge or from public databases such as voter registration list.

TABLE IV VOTERS REGISTRATION LIST

Name	Gender	Zip code	Age
Reena	Female	444805	45
Ram Nivas	Male	424806	46
Kavitha	Female	424809	58
Neha	Female	444806	65

From Table III and Table IV one can deduce that Neha has cancer. According to this survey, the data owner is linked with the publicly available database like voter list and re-identified with the help of Quasi-identifiers for this linking attack [7].Therefore it is required to apply anonymization techniques so that individual privacy is not linked while data is being published.

III. DATA ANONYMIZATION TECHNIQUES

There are many numbers of techniques for anonymizing the data before it is published. The popular techniques are suppression and generalization [4].

A. *Suppression:* In suppression method the value of the Quasi-Identifiers is completely hidden by using a special symbol like *** before the dataset [8].

1. *Tuple Suppression (TS)*: Suppression is performed the complete entry of a record from the table is suppressed. It removes the whole tuple.
2. *Attribute Suppression (AS)*: Suppression is performed all instance of a particular table is suppressed. It performed at the level of the column.
3. *Cell Suppression (CS)*: Suppression is performed in some of the records for given value are suppressed in a table. It performed at the level of the single cell.

TABLE V ORIGINAL DATASET

Birth date	Sex	Zip code
21/1/80	Male	53715
10/1/80	Female	55410
21/2/83	Male	02274
19/4/82	Male	02237

TABLE VI SUPPRESSED DATA WITH K=2

Birth date	Sex	Zip code
*/1/80	Person	5****
*/1/80	Person	5****
//8*	Male	022**
//8*	Male	022**

As shown in table V which contains the original database having Birth date, Sex, Zip code attributes. Table VI shows a suppression based k-anonymization with k=2 means, a tuple in the dataset set is indistinguishable from k-1 other tuples in that data set.

B. Generalization

Generalization is a popular data anonymization technique [9]. It modifies the quasi-identifier value to some generalized values of specific description. For example in fig.3, the parent node of Professional is more general term of engineer and lawyer. And the parent node Artist is more general term of writer and dancer. The root node of Job-any

represent most general term of professional and artist attribute.

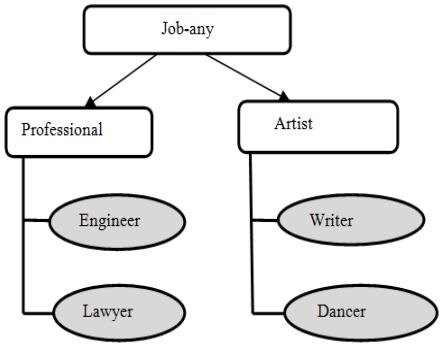

Fig.3 Generalization tree for the attribute job

1. Full Domain Generalization

In Full Domain Generalization scheme [10] while generalizing, for all records and for any quasi-identifier should be generalized at the same level in a given tree structure. For e.g., If a class of {Writer, dancer} are generalized to Artist. But full domain generalization leads to a major distortion of the dataset

2. Sub Tree Generalization

In Sub tree Generalization scheme [11], when any nodes other than the leaf node, all its child values are generalized or none is generalized. For e.g., if all writer and Dancer is generalized to the artist but Lawyer and Engineer may be generalized can retain its specific value at leaf level.

3. Sibling Generalization

In this generalization scheme [12] mostly works similarly as Sub tree Generalization but in this, some sibling can remain un generalized. For e.g., if Lawyer is generalized to professional then Engineer may remain ungeneralized.

4. Cell Generalization

All the generalization schemes [13] give more distortion. In Cell Generalization scheme is a value is generalized in one record the same value for the same attribute in another record may be un generalized. It causes least data distortion as compared to other generalization technique. For e.g., if Writer is generalized to artist and writer in another record may remain un generalized.

IV. K-ANONYMITY

Thus k-anonymous model makes sure that individual cannot be identified by external linking attacks. This k-anonymity is achieved by using Generalization and Suppression. Both the generalization and suppression can be applied globally the same type of transformation is done to all items of the dataset. When both applied locally the transformation is done to the specific transaction of the dataset. A dataset is k-anonymous (k>1) if each record in the dataset is indistinguishable from at least (k-1) another record within the same dataset.

TABLE VII DIAGNOSIS DATA SET

Explicit Identifier	**Quasi Identifier**			**Sensitive Attribute**
ID	Sex	Age	Zip code	Disease
1	M	20	13000	Flu
2	M	24	13500	HIV+
3	F	26	16500	Fever
4	F	28	16400	Cancer

TABLE VIII 2-ANONYMOUS VIEW OF TABLE VII

Explicit Identifier	**Quasi Identifier**			**Sensitive Attribute**
ID	Sex	Age	Zip code	Disease
1	M	[20-24]	13*00	Flu
2	M	[20-24]	13*00	HIV+
3	F	[26-28]	16*00	Fever
4	F	[26-28]	16*00	Cancer

V. CONCLUSION

In this paper, a study of anonymization based privacy-preserving is presented. It leads to further research to develop a new privacy preserving technique. The major goal of this study is to understand the various existing anonymization techniques and finding out an effective anonymization algorithm for the case of multiple sensitive attributes.

REFERENCES

[1] D. Agrawal and C.C. Aggrawal, "On the design and quantification of privacy preserving data algorithm," *In proceedings of the twentieth ACM PUDS, 2001*.

[2] M. Kantarcioglau, J. Jin, and C. Clito, "When do data mining result violate privacy?" *In proceeding of the tenth ACN SIGKDD 2004.*

[3] B. Fung, K. Wang, R. chen, and P.S Yu, "Privacy preserve data publishing: A Survey of recent development," *ACM computer survey (CSUR)*, Vol. 42, No. 4, pp. 14, 2010.

[4] A. Meyerson and R. Williams, "On the Complexity of Optimal k-anonymity," *In proceeding of twenty-third ACM SIGMOD-SIGACT-SIGART symposium on principles of database system*, pp. 223-228, ACM, 2004.

[5] Disha Dubli and D.K. Yadavi, "Secure Techniques of Data Anonymization for Privacy Preservation," *International Journal of Advanced Research in Computer Science*, Vol. 8, No. 5, May-June. 2017.

[6] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based System* Vol. 10, No. 05, pp. 571-588, 2002.

[7] J. Goldberger and T. Tassa, "Efficient anonymizations with enhanced utility," *In Data mining workshops, 2009. ICDMW'09. IEEE international conference on IEEE*, pp. 106-113, 2009.

[8] Kinjal Parmar and Vinita Shah, "A Review on Data Anonymization in Privacy Preserving Data Mining," *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)* Vol. 5, No. 2, February. 2016.

[9] D. Samarati and L. Sweeney, "Protecting privacy when disclosing information: Anonymity and its enforcement through generalization and suppression," *SRI International, SRI-CSL-98-04, 1998*.

[10] L. Sweeney, "Achieving k-anonymity Privacy Protection using generalization and suppression, "*International Journal of Uncertainty, Fuzziness and Knowledge-based system*, Vol. 10, No. 05, pp. 571-588, 2002.

[11] S.K. Adusumalli and V.V. Kumari, "Attribute based anonymity for preserving privacy," *In Advances in Computing and Communications*, pp. 572-579, Springer, 2011.

[12] B. Bercic and C. George, "Identifying Personal Data Using Relational Database Design Principles," *International Journal of Law and Information Technology*, Vol. 17, No. 3, pp. 233-251, 2009.

[13] K. El Emam, F.K. Dankar, R. Issa, E. Jonker, D. Amyot, E. cogo, J.P. Corrivean, M. Walker, S. Chowdhway, R. vaillancourt, *et al*., "A Globally Optimal K-anonymity method for the de-identification of health data," *Journal of the American medical informatics Association*, Vol.16, No.5, pp.670-682, 2009.

[14] E. Bertino, and R. Sandhu, "Database security - Concepts, approaches and challenges", *IEEE Transactions on Dependable and Secure Computing*, Vol. 2, No.1, pp. 2–19, 2005

[15] M. Reed, P. Syverson, and D. Goldschlag, "Anonymous connections and Onion routing", *IEEE Journal of Selected Areas in Communications*, Vol. 16, No.4, pp. 482–494, 1998

[16] M.K. Reiter, and A. Rubin, "Crowds: anonymity with Web transactions", *ACM Transactions on Information and System Security (TISSEC)*, Vol.1, No.1, pp.66–92, 1998.