Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 8, Issue 04, April-2021

A Novel Approach to Secure Data Sharing

Ms. Rekha Vishwakarma[1], Prof. Vijay Shelake[2]

[1]Department of Computer Engineering, YTCEM, Mumbai, India
[2]Department of Computer Engineering, YTCEM, Mumbai, India

Abstract — *Nowadays data sharing utilization is in research and development and many sectors like health, defense, machine learning, etc. Many institution or organizations have their policies to share data without loss of confidentiality of their data. Sharing of information across databases maintained by different organizations leads to an exchange of personal information about an individual. In practice, various statutory regulations and policies prevent the disclosure of such identifiers. So, to prevent the confidentiality of the data, it should share securely. Multiple techniques are implemented to protect the data from attack while sharing. K-anonymization, de-identification of personnel identifiers, noise addition for sharing data have become a topic of recent interest in many organizations due to their versatility and ability to share the data securely. A novel approach for data sharing has been worked out by combining the k-anonymization and addition of noise in the dataset. K-anonymization is applied on personnel identifiers and addition of random noise is applied to numerical data to secure the data for sharing. An application was developed to demonstrate the proposed concept.*

Keywords *- Data sharing, K-anonymization, De-identification, Linkage Attack, De-anonymization attack.*

I. INTRODUCTION

Organizations/ institutions collecting and maintaining data face difficulties in protecting the confidentiality of the data during its use and sharing. Datasets containing "micro-data", i.e. the information about different individuals, are increasingly becoming public both in response to open government laws and to support advanced research using data. Some datasets such as histories of health, individual preferences, purchases, and transactions, in general, are considered as private or sensitive information. Data sharing in today's world plays vital role in the advancement of technology, research, and development, defense, health sector, etc. Further, in today's digital era, data storage and sharing have become an integral part of individual lives. Storage and sharing of data have many benefits in all spectrums of human life.

Analyzing customer's data, personalized services can be provided, scientists can obtain data for scientific research more easily, government policy can be made and implemented more efficiently and effectively. With all the advantages of data sharing, the confidentiality of individual data should also be considered to be of paramount importance as described in[1][2] and the electronic version of the city's voters list can be purchased and used to re-identify the data. In addition to name and address, the list includes the gender and voters' birth date. Of these, people can be identified with the unique birth date, birth date and gender, birth date, and ZIP code. As per the literature, 97% of individuals are identifiable with the birth date and full postal code [1].In other literature, it is reported that approximately 87% of US Citizens can be uniquely identifieid through the combination of their personnel identifiers [2]. Thus, an attacker can get the user's privacy information if the dataset containing the above three attributes available directly. In the fields of medical research, hospitals share diagnostic/treatment records with relevant research institutions. The individual data is presented in the table format and has a firm structure: including age, name, zip, code, and type of disease. Even if the attribute ''Name'' is deleted, there is still a risk that personal privacy will be disclosed.

A hacker knowing an individual's personnel identifier attributes can conclude that a specific individual confidential diagnosis/treatment information of the patient. But if there are more than K (the value of K is large enough) individuals have the same age and zip code; the hacker will not be able to guess confidently. K-anonymous can ensure that at least k records have the same age and zip code. Many privacy protection techniques have been introduced. Among these techniques, k-anonymous and differential privacy which allow modification of data are the two most important privacy models. The k-anonymous method is straightforward to implement, and also the escape risk is measurable, therefore it is used widely [1],[3],[4]. However, owing to the rise of the attacker's background, the result of privacy protection is getting compromise [5]. However, the cost of implementing differential privacy procedures is very expensive, and with the increase of protection strength, the data availability will be badly damaged. To secure the privacy of the data, noise addition methods have been proposed, which are considered less expensive, easy to implement and damage of data availability will be minimum. In the present work, a noble approach using k-anonymization and adding random noise to secure the data sharing has been described

II. LITERATURE SURVEY

2.1. Noise addition method for securing data

Noise addition method to protect the addition of data noise functions by adding or multiplying confidential quantitative attributes with speculative or randomized data. A normal distribution with a zero mean and a standard deviation [6][7] is used to select the speculative value.

2.1.1. Additive Noise- Addictive noise was first published by Kim [8] with the general expression that

$$Z = X + \varepsilon \quad \ldots (1)$$

Where Z is the transformed data point, X is the original data point and ε is the random variable (noise) with a distribution $e \sim N(0, \sigma^2)$. This is then added to X. The X is then replaced with the Z for the data set to be published [9]. Random data is applied to confidential attributes with speculative noise to mask the distinguishing values, an example of increasing the GPA of a student by a decreasing percentage of 3.45 to 3.65 GPA [10].

2.1.2 Multiplicative Noise - Another type of speculative noise defined by Kim and Winkler[11] is multiplicative noise in which they explain that multiplicative noise is achieved by generating random numbers with a mean= 1, which is used as noise and multiplied to the original data set. A random number with a short Gaussian distribution, with a mean = 1 and a small variance, is multiplied by each data variable.

$$Y_j = X_j \varepsilon_j \quad \ldots (4)$$

Where Y refers to perturbed data; X refers to the original data; ε is the generated random variable (noise) with a normal distribution with mean μ and variance σ [11].

2.13. Logarithmic multiplicative noise - Another variant of multiplicative noise is defined by Kim and Winkler[11], in which a logarithmic modification is taken from the original data:

$$Y_j = \ln X_j \quad \ldots (5)$$

The random number (noise) is then generated and then added to the altered data [5]:

$$Z_j = Y_j + \varepsilon_j \ldots (6)$$

Where X is the original data; Y is the logarithmic altered data; Z is the logarithmic altered data with noise added to it; e^x is the exponential function used to calculate the antilog.

2.2. Data anonymization method for securing data

The data anonymity model categorizes the data attributes into three types: (i) Explicit Identifier attribute (EI), which are the attributes that can typically identify an individual, i.e. Social Security Number or name. (ii) Quasi Identifier attributes (QIDs) that are not considered private individual data and can be recognized as background knowledge by other people or can exist in other external available databases, i.e. age and zip code. QIDs can potentially identify the individual if taken from the published data and linked together with such available data. (iii) Sensitive Attributes (SAs) are the private and unknown sensitive individual attributes i.e. the Salary and Disease, which need to be prevented from being inferred and preserved against the different privacy disclosure attacks. Data anonymization rules do not publish EI, whereas QIDs may be masked using a certain disclosure control method, like generalization and/or suppression.

A popular approach for data anonymization is k-anonymity. An original data set containing personal health information can be transformed with k-anonymity so that it is difficult for an intruder to determine the identity of the people in that data set. A k-anonymized data set has the property that every record is analogous to a minimum of another k-1 other record on the doubtless identifying variables. For example, if k = 5 and therefore age and gender are the potentially identifying variables, then a k anonymized data set has a minimum of 5 records for any age and gender value combination. The most popular k-anonymity implementations use transformation techniques such as global generalization recoding and suppression[12] to distinguish every record during a k anonymized data set with a maximum likelihood of 1/k becoming recognized again. In practice, a knowledge custodian would choose the worth of k commensurate with the re-identification probability they're willing to tolerate a threshold risk. Higher k values mean a lower risk of re-identification, but also a greater distortion of information and hence greater loss of information in k

anonymization. In general, excessive anonymization will make recipients less usable for the disclosed data because some analysis becomes difficult or the analysis produces biased and incorrect results.

To address the issue of privacy security, a modern k-anonymous approach has been used that is distinct from traditional k-anonymous. In particular, by adding noises, numerical data achieves k-anonymous, and by using randomization, categorical data achieves k-anonymous. The drawback that at least k elements must have the same quasi identifier in the k anonymous dataset has been solved using the two above approaches.A two-step clustering method is used to divide the actual data set into equivalence classes, as the method of finding anonymous equivalence is very time-consuming. First, the actual data set is divided into many separate sub-datasets, and then the equivalence classes are created in the sub-datasets, substantially reducing the computational cost of identifying anonymous equivalence classes.

2.3. De-identification method for securing data

De Identification is intended to protect individual identity, making it difficult to discover if the data in a dataset is connected to a particular person while retaining some of the utility of the dataset for other purposes. When data have identifying information such as names, email addresses, geolocation information, or photographs, the conflict between the goals of data use and privacy protection can arise. De-identification attempts to resolve this conflict, allowing for some privacy-sensitive data that identifies individuals to be cleared, while allowing other useful information to remain. De Identification is therefore an essential technique that organizations can use to avoid the likelihood of privacy associated with the sharing, archiving, and even publishing of data containing personal information. De-identifying data at the time of collection or after minimal processing can reduce the costs associated with using and archiving data, by reducing the privacy risk associated with inadvertent release (i.e., a data breach).

De-identifying data that are shared can reduce the need for technical and policy controls. Therefore, De-identification can allow organizations to make the most use of data that could Otherwise not be possible.

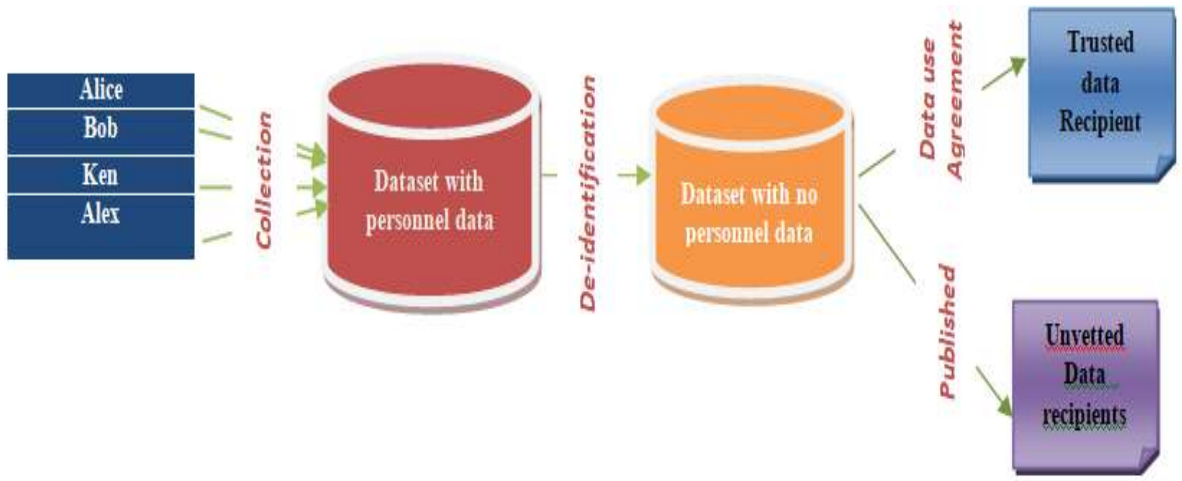

Figure 1: De-Identification Model [13]

The above figure provides an overview of the de-identification process. Data are collected from data subjects, the "persons to whom data refer." These personal data are integrated into a dataset containing personal information. De Identification generates a new dataset claimed to have no identifiable details that an entity can internally use this dataset instead of the original dataset to limit the risk of privacy. The dataset may also be provided to trusted data recipients who are bound by additional administrative controls such as data use agreements. Alternatively, by posting the de-identified data on the Internet, for example, the data may be made freely accessible to a greater number of data recipients.

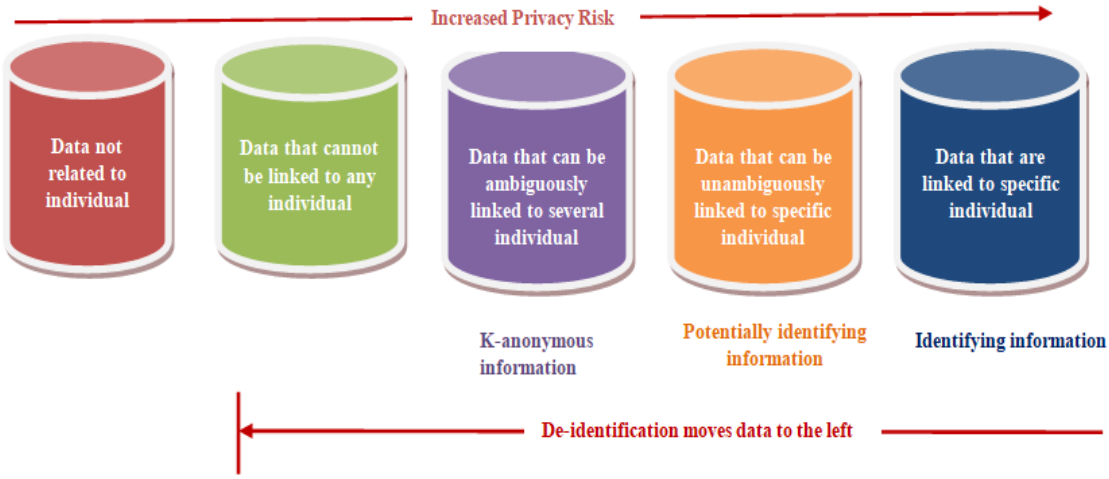

Figure 2. Data Identifiability Spectrum[13]

III. PROPOSED METHOD

In this research, combining k-Anonymity and random noise addition to secure data while sharing. Firstly we implemented-Anonymity generalization and suppression method on personnel identifier attributes such as name, age, etc. The next addition of random noise is applied to key attribute data. There are two cases involved in the addition of random noise

- Additive noise is added
- Multiplicative with additive noise added

The above approach will secure data from linkage and deanonymization attacks and also improve the data quality after modification of data. Based on the concept an application was developed.Below figure will explain the process of proposed system.

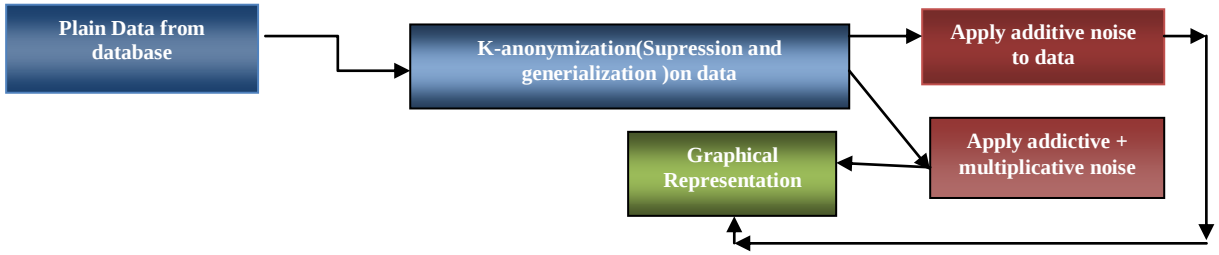

Figure 3. Proposed System

IV. RESULT AND DISCUSSION

In proposed system, we have some random data to test. In the random data there are 10 records in database. In the record there are personnel identifiers and some sensitive fields. Records contain both numerical and alpha numerical data. In this novel approach the k-anonymization is applied on personnel identifier such as name and age. On name column suppression method of k-anonymization is applied and on age field generalization method of k-anonymization is applied. After this step random noise is added to salary field as there are considering it as a sensitive data on which the researcher have main concern.

Now, these data are secure to share. Users may use the salary data for their analysis, but at the same time they are not aware of the actual data, therefore the novel approach, including k-anonymity and addition of noise, can provide an effective means of ensuring important data sharing. In the proposed system, two methods are combined to make data sharing more secure.

The disclosure risks from which a dataset is to be protected can be specified by categorizing the attributes of the input dataset into different types such as

- Identifying attributes that are associated with a high risk of re-identification. Names are truncate, other attributes;
- Quasi-identifying attributes are anonymized to avoid attacks. They will be transformed. Typical examples are gender, date of birth, and ZIP codes.

- Sensitive attributes encode properties with which individuals are not willing to be linked as such, which could be of interest to an attacker and, if disclosed, could cause harm to data subjects. They will be kept unmodified but may be subject to further constraints, such as other methods. Typical examples are diagnoses.

- Insensitive attributes are not associated with privacy risks. They will be kept unmodified. Once the discloser risk has been identified, the proposed methodology for securing the data can be implemented so that the confidentiality of the data is not compromised.

In the proposed method, statistical consideration has been kept in mind while implementing the noise addition. The idea of multiple types of privacy is highlighted that can be implemented while sharing the data. In the proposed method, actual data is generalized or anonymize which will finally give only the information of a general aspect of the data but not the specific information in the data to the third party. Following figure shows the salary variation in graph with random noise addition method and random noise with addition and multiplicative method.

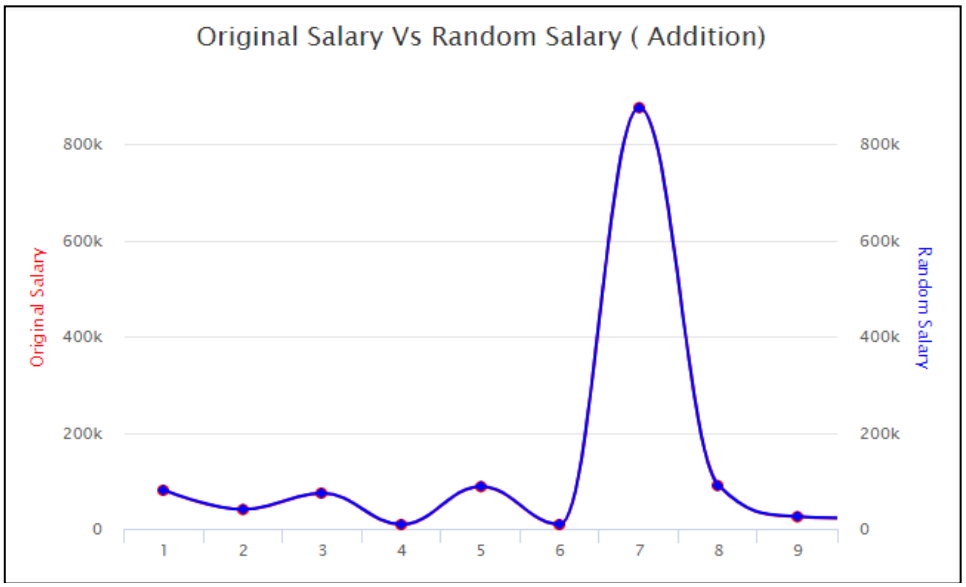

Figure 4. Results of the normal distribution of Original Data and Random Salary (Addition) of applying random noise to the salary field

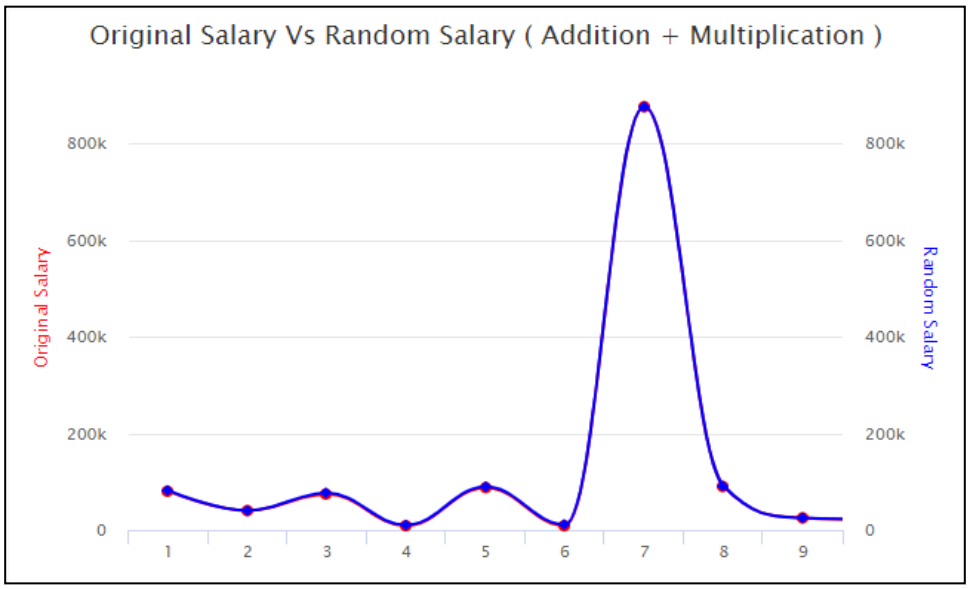

Figure 5. Results of the normal distribution of Original Data and Random Salary (Addition + Multiplication) of applying random noise to the salary field

V. CONCLUSION

A noble data sharing approach has been suggested and implementation has been developed. Not only will the alpha or string-based attributes, but even the numerical attributes of the data be anonymized by the proposed method. Research show data anonymization can be enhanced concerning the disparity in data forms, providing a good alternative strategy for better anonymization.

REFERENCES

[1] L. Sweeney. Weaving technology and policy together to maintain confidentiality. J. Law Med. Ethics, 25(2–3):98–110, 1997.

[2] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst, 10 (5); 571588, 2002.

[3] Dwork C. Differential privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. (2006) https://doi.org/10.1007/11787006_1

[4] T. Ma, J. Jia, Y. Xue, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, ''Protection of location privacy for moving KNN queries in social networks,'' Appl. Soft Comput. 66, 525–1532, 2018.

[5] T.Ma,W.Shao,Y.Hao,andJ.Cao,Graph classification basedon graph set reconstruction and graph kernel feature reduction, Neurocomputing, 296, 33–45, 2018.

[6] MdZahidul Islam, Privacy Preservation in Data Mining Through Noise Addition, PhD Thesis, School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, New South Wales 2308, Australia, November 2007

[7] KadampurM. A, Somayajulu D.V.L.N., A Noise Addition Scheme in Decision Tree for, Privacy Preserving Data Mining, Journal of Computing, 2(1), 2151-9617,2010.

[8] Jay Kim, A method for limiting disclosure in microdatabased random noise and transformation, Proceedings of the Survey Research Methods, American Statistical Association, Pages 370-374, 1986.

[9] J. Domingo-Ferrer, F. Sebé, and J. Castellà-Roca, On the security of noise addition for privacy in statistical databases," Privacy in Statistical Databases, Vol. 3050, p. 519. Springer Berlin / Heidelberg, 2004

[10] J. Domingo-Ferrer and V. Torra (Eds.), On the security of noise addition for privacy in statistical databases, LNCS 3050, pp. 149–161, Springer-Verlag Berlin Heidelberg 2004.

[11] Jay J. Kim and William E. Winkler, Multiplicative noise for masking continuous data, Research Report Series, Statistics (#2003-01), Statistical Research Division, U.S. Bureau of the Census. Washington D.C. 20233.2003

[12] Samarati, Pierangela; Sweeney, Latanya Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Carnegie Mellon University. Journal contribution; 2018:

[13] https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

