International Journal of Scientific Research in Computer Science and Engineering

Review Paper

Vol.8, Issue.1, pp.58-62, February (2020)

E-ISSN: 2320-7639

Data Anonymization Techniques for Preserving Privacy in Public Release Data Model: A Technical Review

Arun Amaithi Rajan[1*], Anitha Amaithi Rajan[2]

[1]National University of Singapore, Singapore
[2]Francis Xavier Engineering College, Tirunelveli, India

**Corresponding Author: arunamaithirajan@gmail.com*

Available online at: www.isroset.org

Received: 02/Feb/2020, Accepted: 10/Feb/2020, Online: 28/Feb/2020

***Abstract*--** The protection of sensitive records is very necessary for a modern scenario. Lately, the informational index is accessible for open use for statistical analysis. In this situation increasingly sensitive information like medical records, nation resident's data, worker's compensation data and so on are affecting to a higher extent since we are giving our data to people in general. Thus, Data anonymization assumes significance in the present day to protect the open discharge of sensitive information. In this paper, we reviewed some anonymization techniques and proposed a simple anonymization technique which is the combination of synthetic data generation and pseudonymization approach which reduces attacks on sensitive facts.

***Keywords*--** Anonymization techniques, privacy-preserving algorithm, synthetic data generation and pseudonymization technique

I. INTRODUCTION

Data Anonymization is defined as a process by which personal data is irreversibly changed in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller or any other party. Data anonymization [2] allows the transfer of data across a boundary, such as between two departments within an office or between two offices, while reducing the risk of unintended disclosure, and in certain environments in a manner that permits evaluation and analytics post-anonymization. In the context of medical data, anonymized data refers to data from which the patient cannot be identified by the recipient of the data. Personal details ought to be removed, collectively with any other data which, in conjunction with other data held by or disclosed to the recipient, could identify the patient [10].

The main theme of this paper has synthetic data generation and pseudonymization. Synthetic data generation [4] is the system of producing data applicable to a given situation that is not obtained by direct measurement. Synthetic data are created to meet specific needs or certain conditions that may not be found in the original, real data. This could be useful when designing any type of system because the synthetic data are used as a simulation or as a theoretical value, situation, etc. Another use of synthetic data is to guard the privacy and confidentiality of authentic data which we took as a key usage for the proposed system. Pseudonymisation [5] is the personal records in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to assure that the personal data are not attributed to an identified or identifiable real person. Even anonymization and pseudonymization seem similar it has a difference based totally on re-identification risk. Pseudonymous data can still go through re-identification to link (through attribute) it to an individual (particular person's data) again while anonymous data cannot be re-identified. In this paper section 2 explains the related works on data anonymization techniques and section 3 analyzes its strengths and weaknesses, section 4 proposed a simple anonymization technique for the public release data model and section 5 concludes the paper and explains future work.

II. ANONYMIZATION TECHNIQUES FOR SENSITIVE DATA: REVIEW

In this section we did an elaborated literature survey of major anonymization techniques which helps us to come up with a simple anonymization technique for the public release model. We have discussed 8 major anonymization techniques in this section as follows: Attribute Suppression, Record suppression, Character masking, Pseudonymization, Generalization, Data perturbation, Synthetic data generation, and Aggregation. The sample dataset where each anonymization technique is going to be applied has been shown in Table 1.

Table 1. Sample dataset

A	B	C	D	E	F	G	H	I
Name	**Supervisor**	**Joining date**	**Age**	**Salary**	**Postal code**	**Height (cm)**	**Weight (kg)**	**Game**
Zing	Julia	Mar 1, 2019	25	7 LPA	600001	132	65	Cricket
Josh	Ganesh	Feb 3, 2018	35	13 LPA	621052	150	75	Table Tennis
Gary	Ganesh	Feb 5, 2018	22	6 LPA	627004	145	42	Cricket
Ram	Julia	Sep 6, 2017	41	22 LPA	651008	143	82	Basketball

A. Attribute Suppression

Attribute suppression refers to the deletion of an entire column of data in a dataset. We have to delete the attribute/attributes. If the overall structure of the dataset needs to be maintained, clear the data. Suppression has to be actual deletion not just hiding the column. It should be used when an attribute is not required in the anonymized dataset, or when the attribute cannot be suitably anonymized with other techniques. This technique should be applied at the beginning of the anonymization process. This is the strongest type of anonymization technique because there is no way of recovering any information back after anonymization [7]. After the elimination of the joining date attribute, the sample dataset changed as showed in Table 2.

Table 2. Sample dataset after Joining date attribute (C) suppressed

A	B		D - I
Name	**Supervisor**		**….**
Zing	Julia		….
Josh	Ganesh		…..
Gary	Ganesh		…..
Ram	Julia		…..

B. Record Suppression

Record suppression is the elimination of an entire record in a dataset. In contrast to other techniques, this technique disturbs multiple attributes at the same time. This will be used to remove outlier records which are unique or do not meet other criteria such as k-anonymity, and not to keep in the anonymized dataset. Elimination of a record can impact the dataset in terms of statistics such as average and median. This technique has to be applied either before or after some other technique has been done on the dataset such as aggregation [9].

C. Character Masking

Character masking is the transformation of the characters of a data value using a constant symbol such as "*" or "x". Masking is applied only to some characters in the attribute. Character masking differs depends upon the attribute type such as fixed masking and variable masking. When the data value is a string of characters and hiding part of it is adequate to provide the extent of anonymity required, replace the appropriate characters with a chosen symbol [6]. In sample dataset column F has been chosen for character masking and the result has been shown below in Table 3.

Table 3. Character masking on attribute Postal code (F)

Postal code
600***
621***
627***
651***

D. Pseudonymization

Pseudonymization is simply the replacement of recognizing data with made-up values. Pseudonyms can be irretrievable, where the original values are properly disposed and the pseudonymization was done in a non-repeatable fashion. Persistent pseudonyms allow linkage by using the same pseudonym values to represent the same individual across different datasets. On the other hand, different pseudonyms may be used to represent the same individual in different datasets to prevent the linking of the different datasets. Pseudonyms can also be deterministically or randomly generated. One way to do this is to pre-generate a list of made-up values, and randomly select from this list to replace each of the original values. The made-up values should be unique and should have no connection with the original values such that one can derive the original values from the pseudonyms. Persistent pseudonyms usually provide improved utility by maintaining referential integrity across datasets. Following table 4 shows the pseudonymization done on Name (A) attribute.

Table 4. Made up values for pseudonymization on Name (A) attribute

Name	Name (Pseudonymized)
Zing	A9876H
Josh	B1775G
Gary	V0946K
Ram	M5438R

If encryption is used, the encryption key should not be shared, and in fact must be strongly protected from unauthorized access, because a leak of such a key could result in a data breach by allowing the reversal of the encryption. The same applies to pseudo-random number generators, which require a seed. The security of any key used must be ensured like with any other type of encryption or reversible process [5].

E. Generalization

Generalization is a deliberate reduction in the accuracy of data such as converting a person's salary into a salary range. This technique is also called as recoding. It is used for values that can be generalized and still be useful for the intended purpose. We have to design appropriate data categories and rules for translating data. Table 5 shows the age and salary generalization [8].

Table 5. Age (D) and Salary (E) generalization

Age	Salary
20-30	< 10
30-40	< 20
20-30	< 10
40-50	< 25

F. Data Perturbation

Data perturbation is nothing but the values from the original dataset are altered to be slightly different. It is usually used for quasi-identifiers (typically numbers and dates) which may possibly be identifying when combined with other data sources, and slight changes in value are acceptable. This technique should not be used where data accuracy is crucial. We can use different data perturbation techniques includes rounding and adding random noise. The example in this section shows base-x rounding [1]. Following table 6 shows perturbation techniques chosen for sample dataset attributes height and weight.

Table 6. Perturbation techniques

Attribute	Perturbation Techniques
Height (in cm)	Base-10 rounding
Weight (in kg)	Base-5 rounding

The degree of perturbation should be proportionate to the range of values of the attribute. If the base is too small, the anonymization effect will be weaker, on the other hand, if the base is too large, the end values will be too different from the original and the utility of the dataset will likely be reduced. Following table 7 shows the example for data perturbation.

Table 7. Data perturbation on Height (G) and Weight (H)

Height (cm)	Weight (kg)
130	65
150	75
140	40
140	80

G. Synthetic Data Generation

This technique is slightly different as compared to the other techniques as it is mainly used to generate synthetic datasets directly and independently from the original data, instead of modifying the original dataset. When a large amount of data is required for system testing, but the actual data cannot be used and yet the data should be "realistic" in certain aspects, like format, relationships among attributes, etc [4].

First of all, we have to study the patterns from the original dataset and apply the patterns when creating synthetic dataset. The degree to which the patterns from the original dataset need to be replicated depends on how the anonymized dataset is to be used. The following example shows the synthetic data generation of the work-out data from the gym which has to be given to 3rd party and analyzed. To generate that synthetic dataset we need to get a statistical pattern on either a daily basis or a weekly basis. For example, employee A is going to the gym in the company daily. From the gym door access system, we can get in and out timings which will be sent to the 3rd party to analyze employee fitness. Table 8 shows the original date from the gym door access system.

Table 8. Original date from the gym door access system

User	Date	Time In	Time out
A	Jan 1	8:20	9:10
A	Jan 2	5:30	6:30
A	Jan 3	5:45	7:00
A	Jan 4	4:40	6:00
A	Jan 5	5:00	7:00

From the original dataset, we have to create a synthetic dataset based on getting statistical value on some attributes. Here the average time spent by user A for a week has been calculated and generated the new data set based on that.

Table 9. Average time spent by user A in a week

Time In	Time out	Time Spent
8:20	9:10	0.83 hr
5:30	6:30	1 hr
5:45	7:00	1.25 hr
4:40	6:00	1.33 hr
5:00	7:00	2 hr
Average (per week)		1.282 hr

We can sample the data and generate a synthetic dataset as shown in table 10 using any one of the synthetic data generation techniques available [ref]. After generating the synthetic dataset statistical value should not be changed in the generated dataset.

Table 10. Generated dataset

User	Date	Time In	Time out
A	Jan 15	2:40	4:00
A	Jan 16	5:40	6:45
A	Jan 17	7:45	8:30
A	Jan 18	3:20	5:50
A	Jan 19	9:15	10:00

When applying this technique, outliers may need added attention. For testing purposes, outliers are often very valuable, but outliers in the synthetic data may also indicate certain outliers within the original dataset. It is therefore suggested to create outliers in synthetic data intentionally and independent of the original data. This method is usually used for data analysis.

H. Data Aggregation

Data aggregation is converting a dataset from a list of records to summarized values. It is used when individual records are not required and aggregated data is sufficient for the purpose. Aggregation may need to be applied in combination with suppression [3]. Some attributes may need to be detached, as they contain details that cannot be aggregated, and new attributes might need to be added. The sample dataset has been aggregated based on a Game attribute (Table 11).

Table 11. Aggregation based on Game (I)

Game	No of Employees
Cricket	2
Table Tennis	1
Basket Ball	1
Total	4

III. ANALYSIS OF DATA ANONYMIZATION TECHNIQUES

Each anonymization technique has its strengths and weaknesses. In Table 12, we tabulated all the technique's strengths and limitations moreover analyses each technique over the two attacking [11] criteria which are,

A- Is it still possible to link together various data sets associated with the same person?
B- Is it still possible to deduce information associated with an individual person?

Table 12. Analysis of different Anonymization Techniques [**Legend: X-** Probably Not √- Yes]

No	*Anonymization Technique*	*Strengths*	*Weakness and Limitations*	*A*	*B*
1	*Attribute Suppression*	No way to recover info from such attribute	Not suitable for all attributes	X	X
2	*Record Suppression*	Very useful in removing outlier records	May impact dataset's accuracy in statistical analysis	X	√
3	*Character Masking*	Easy to implement	Inference attack can be done on the length of masked characters	√	√
4	*Pseudonymization*	Powerful technique	Double coding takes more storage, power and expensive	√	√
5	*Generalization*	Fit for moderate size k More generalization gives effect (Address → Town/City)	If k is too low, easy to identify If k is too large, accuracy will be affected Inference attack can be done on unique records	√	√
6	*Data Perturbation*	Easy to implement	Should not be used where accuracy is important	√	X
7	*Synthetic Data Generation*	Perfect for public release model (Statistical analysis → Duplicate dataset)	The outlier is needed for accuracy	X	X
8	*Aggregation*	Linking dataset attack cannot be done	Attributes need to be removed when they cannot be aggregated (May affect precision)	√	√

Above analysis shows that some anonymization techniques are vulnerable to either A or B or both A and B. After reviewing these techniques we proposed a new technique which is not vulnerable to A or B that is described in the following section.

IV. PROPOSED SIMPLE TECHNIQUE FOR PUBLIC RELEASE DATA MODEL

We proposed a technique that has synthetic data generation followed by pseudonymization or character masking which is suitable for the public release data model. Synthetic data generation has the goal to generate datasets that are structurally and statistically similar to the real data but that are,

1. Obviously synthetic.
2. Offer strong privacy guarantees to prevent adversaries from extracting any sensitive information.

Firstly we have to analyze the data types, distributions and correlations of the attributes in the original dataset then output a data summary. Secondly, add Laplace noise to the distributions to preserve privacy then samples from the summary computed by the previous step and outputs synthetic data. The above suggested synthetic data generation is epsilon differentially private. After synthetic data set has been created some attributes can be anonymized using either pseudonymization or character masking concerning the attribute type. The flow of this technique has been shown in Figure 1.

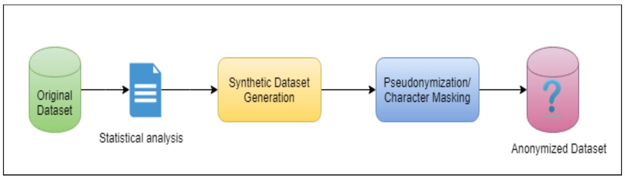
Original Dataset
Statistical analysis
Synthetic Dataset Generation
Pseudonymization/ Character Masking
?
Anonymized Dataset

Figure 1. A simple technique to preserve privacy in the public release data model

We applied the proposed technique on Table 9 which produces the result as follows in Table 13. Here User attribute has been pseudonymized after the synthetic data generation.

Table 13. Generated dataset using a proposed technique

User	Date	Time In	Time out
65789	Jan 15	2:40	4:00
65789	Jan 16	5:40	6:45
65789	Jan 17	7:45	8:30
65789	Jan 18	3:20	5:50
65789	Jan 19	9:15	10:00

In the above-suggested technique,

1. An attacker can analyze only statistics from this anonymized dataset.
2. If an attacker has any individual’s data, he cannot match that with this anonymized dataset.
3. Good for the public release model.
   Thus, the resulted dataset is not vulnerable to either A or B.

V. CONCLUSION AND FUTURE WORK

This paper reviewed major anonymization techniques and analyzed those techniques against the two attacking models. We proposed a simple anonymization technique for the public release model which has the flow of synthetic data generation using statistical analysis followed by pseudonymization or character masking. In the future, the length of the masking characters can be hashed and added as an extra character in the character masking technique because hash cannot be converted back since hashing is one way, it will be more secure too. Random pseudonym function can be used if the pseudonym function is deterministic then it will be easy to attack in pseudonymization. Removing outliers can prevent an attack in generalization. Generally, datasets that are distributed for statistical analysis should be encrypted to prevent common attacks.

REFERENCES

[1] H. Kargupta, S. Datta, Q.Wang, and K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques,” In Proceedings of the International Conference on Data Mining (ICDM), pp. 99-106, 2003.

[2] Kavita Rodiya and Parmeet Gill, “A Review on Anonymization Techniques for privacy preserving data publishing,” IJRET: International Journal of Research in Engineering and Technology, November 2015.

[3] Disha Dubli and D.K Yadav, “Secure Techniques of Data Anonymization for Privacy Preservation,” International Journal of Advanced Research in Computer Science, Vol. 8, Issue. 05, pp. 1694-1698, 2017.

[4] Surendra .H, Dr. Mohan .H .S, “A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing,” International Journal of Scientific & Technology Research, Vol. 6, Issue. 03, pp. 95-101, March 2017.

[5] White Paper on Pseudonymization Drafted by the Data Protection Focus Group for the Safety, Protection, and Trust Platform for Society and Businesses in Connection with the 2017 Digital Summit.

[6] Ajayi, Olusola Olajide, Adebiyi, Temidayo Olarewaju, “Application of Data Masking in Achieving Information Privacy,” IOSR Journal of Engineering (IOSRJEN), Vol. 04, Issue. 02, pp. 13-21, 2014.

[7] Thakkar A., Bhatti A.A., Vasa J., “Correlation Based Anonymization Using Generalization and Suppression for Disclosure Problems,” In Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing, Vol. 320, pp. 581-592, 2015.

[8] Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian, “A Survey of Privacy Preserving Data Publishing using Generalization and Suppression,” Applied Mathematics & Information Sciences. Vol. 8, Issue. 03, pp. 1103-1116, 2014.

[9] Anisha Tiwari1, Minu Choudhary, “A Review on K-Anonymization Techniques,” Scholars Journal of Engineering and Technology (SJET), Vol. 5, Issue. 06, pp. 238-245, 2017.

[10] Tanashri Karle and Prof Deepali Vora, “Privacy Preservation in Big Data Using Anonymization Techniques,” In International Conference on Data Management Analytics and Innovation (ICDMAI), pp. 340-343, 2017.

[11] Ramesh Bandaru , Rao S Basavala "Information Leakage through Social Networking Websites leads to Lack of Privacy and Identity Theft Security Issues." International Journal of Scientific Research in Computer Science and Engineering (IJSRCSE), Vol 1, Issue. 03, pp. 1-7, 2013.

Authors Profile

Mr. Arun Amaithi Rajan recieved his Bachelor of Engineering in Computer Science and Engineering from College of Engineering Guindy, Anna University, India in 2018 and Master of Computing from National University of Singapore in year 2019. His main research interests are in Cryptography Algorithms, Privacy preserving mechanisms and Trusted computing.

Mrs Anitha Amaithi Rajan received her Bachelor of Engineering in Biomedical Engineering from PSNA College of Engineering and Technology , Affiliated to Anna University, India in year 2015 and Master of Engineering in Computer Science and Engineering from Francis Xavier Engineering College, Affiliated to Anna University, India in year 2018. Her main research intersts are lies in Networks, Security principles and Computational biology.