2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC), and IEEE Intl Conference on Intelligent Data and Security (IDS)

# A Comparative Study of Data Anonymization Techniques

Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim, Ramona Ramli
[1]Institute of Informatics and Computing in Energy,
Universiti Tenaga Nasional, 43000, Malaysia
[2]College of Computing & Informatics,
Universiti Tenaga Nasional, 43000, Malaysia
*suntherasvaran.murthy@gmail.com*, *asmidar@uniten.edu.my*, *fiza@uniten.edu.my*, *ramona@uniten.edu.my*

***Abstract*—In today's digital era, it is a very common practice for organizations to collect data from individual users. The collected data is then stored in multiple databases which contain personally identifiable information (PII). This may lead to a major source of privacy risk for the database. Various privacy preservation techniques have been proposed such as perturbation, anonymization and cryptographic. In this study, five anonymization techniques are compared using the same dataset. In addition to that, this study reviews the strengths and weaknesses of the different technique. In the evaluation of efficiency, suppression is found as the most efficient while swapping is in the last place. It is also revealed that swapping is the most resource-consuming technique while suppressing being less resource-consuming.**

***Keywords—privacy, security, privacy preservation.***

## I. Introduction

The incessant growth of stored data has brought increasing interest in data analysis due to the possibilities it can provide to business organizations. For example, collected data from e-commerce sites may profile clients based on their previous searches and purchases [1]. Data mining techniques applied to such data facilitate the extraction of knowledge to support a variety of domains.

As data mining becomes more pervasive, privacy concerns are increasing. The use of data containing personally identifiable information (PII) has to be restricted in order to protect individual privacy [2]. The practice of removing PII from collected data is extensively being discussed in a research society and also being practiced in many industries.

For analysis purposes, there is a need to limit disclosure risks to an acceptable level, while maintaining the information and hence retaining the utility [3]. For example, collected data from the smart meter may be required by an energy utility company for operational reasons [4]. There is also a growing use of consumption data to be released or shared externally [5]. Prior to using the data for further analysis or data sharing, the energy utility company must ensure that the data must not be attributable to a specific individual.

In the healthcare industry, invasion of patient privacy is a growing concern in the domain of big data analytics. Data anonymization prior to analytics is used to protect patient identity [6]. Still, sharing and publishing the data given the concern of privacy breach. Multiple data that has been published or shared with third parties can be combined to get a complete set of data. For example, medical information is determined by joining the patient data with public voter registration list or getting data from questioner given to public and combine with the published data.

Some researchers have found that with the date of birth, zip code, and gender, we can identify almost 60% of individual [7]. Those attributes can uniquely identify the majority of the population by performing a linkage attack and trying to infer the targeted victim's sensitive attribute values [8].

A number of techniques required to preserve the privacy also grows as privacy requirements raise, sometimes exponentially. The techniques such as classification, anonymization, association rule mining, clustering have been suggested in order to preserve privacy. In large data sets, the performance of techniques is essential while maintaining complexity [9].

In this paper, we compare several data anonymization techniques to allow data sharing with other organizations or entities, where appropriate technical controls may be imposed to reduce the risk of unauthorized disclosure of sensitive data. Fig. 1 illustrate a pictorial summary of anonymization concept.

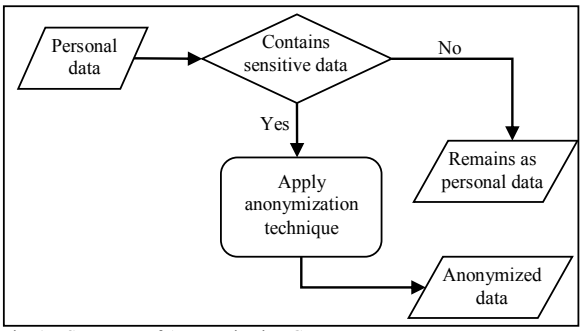


Fig. 1. Summary of Anonymization Concept

This paper is structured as follows: Section II reviews on several anonymization techniques. Section III explains the existing privacy-preserving techniques. Next, Section IV discusses privacy preservation in database fields. Lastly, Section V presents the conclusion and future works. The intention of this review is to compare five anonymization techniques in a situation where the data is about to be shared with other organizations or entities.

## II. Anonymization Techniques

When releasing anonymized data, the main concern is to prevent the sensitive information of the individuals from being disclosed. There are three types of information disclosure: identity, attribute, and inference.

Identity disclosure occurs when a guessable algorithm and insufficient anonymization are used that allow re-identification by linking a specific record in the anonymized data [10]. Furthermore, it will uncover the related sensitive

978-1-7281-0006-7/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataSecurity-HPSC-IDS.2019.00063

values within the dataset. For example, an anonymization process replacing 'A' with '1', 'B' with '2', 'C' with '3' and so on may lead to identity disclosure or re-identification.

Attribute disclosure appears when new data about an individual is exposed. For example, a dataset containing anonymized employee records of a particular department reveals that all employees above the age of 40 have received three months' salary bonus. If it is known that a particular employee is 43 years old and he is in the department, we then know that this individual has received the bonus, even if the individual's record cannot be differentiated from others in the anonymized employee records.

For inference disclosure, it may occur if the adversary obtains confidential information about an individual by correlating with different dataset. This situation happened when releasing anonymized data may also lead to indirect disclosure of other related sensitive data via inferences [11].

Anonymization techniques are frequently used to preserve privacy [12], which focused on the conversion of personal data into anonymized data to reduce the risk of unauthorized disclosure. Regardless of the techniques used, anonymization techniques are expected to reduce the original information in the dataset by some extent.

However, as the extent of anonymization increases, the utility of the data reduces [9]. Hence, the organization needs to choose suitable anonymization techniques to be applied in the dataset by anticipating the expected utility and the risk of de-identification.

The prominent method in anonymization is *k*-anonymity introduced by [13]. With *k*-anonymity, an original dataset containing sensitive information can be transformed so that an attacker unable to infer sensitive information associated with individuals [14]. A *k*-anonymized dataset has the property that each record is similar to at least another *k*-1 other records on the potentially identifying attributes.

In this review, we will discuss five anonymization techniques; suppression, generalization, swapping, masking and distortion. Let us consider an example of smart metering data of the individual to understand these five techniques. The original data is shown in Table I.

TABLE I. SAMPLE ORIGINAL DATA

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| 123 | Alice | 117, Jalan Kinrara 1 | 35400 | 434 |
| 234 | Bob | 14, Jalan Presint 9/1 | 51400 | 289 |
| 456 | John | 23, Jalan Amanah 5 | 81200 | 45 |
| 789 | Sarah | 89, Jalan Nuri 8 | 68100 | 872 |

### A. Generalization

Generalization is a process of replacing the value with less specific but semantically consistent value. This technique applies to the cell level, where some original values are maintained with additional confusion. This will increase stupefaction to the attacker to infer sensitive data.

However, generalization cannot be applied for all attributes. Based on the sample data in Table I, some of the attributes such as ID, name, and postcode are not suitable to generalize. For the address, the possible approach is to remove the house number and retain only the road name. While, for the consumption usage, the approach taken is to generalize into the range as shown in Table II.

TABLE II. ANONYMIZED DATA AFTER GENERALIZATION

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| 123 | Alice | Jalan Kinrara 1 | 35400 | 400 - 450 |
| 234 | Bob | Jalan Presint 9/1 | 51400 | 250 - 300 |
| 456 | John | Jalan Amanah 5 | 81200 | 1 - 50 |
| 789 | Sarah | Jalan Nuri 8 | 68100 | 850 - 900 |

In writing a program using object-oriented programming language, generalization technique unable to be coded dynamically. It must be hardcoded based on appropriate attributes. The different attribute might use a different way of generalization technique. For example, consumption usage can be generalized by defining a range value. Meanwhile, for address, the house number is removed. Because of the issues, the whole program cannot be written dynamically.

### B. Suppression

Suppression refers to the removal of an entire part of data (column or tuple) in a dataset by changing the value to one value that doesn't have meaning, example "****" replace the original data [7]. While, suppressions conceal information by deleting it, which records that exist in the original data are completely removed from the final output. This technique applies on column or tuples, hiding the column or tuples when needed.

This can make the data unusable due to the data may result in the wrong outcome for the research. For example, some of the attributes such as name and ID number which is not considered important for research purpose but affect the privacy of the individual. The other attributes such as address, postcode and usage are not suitable to use suppression due to it will affect the usability of the data. Table III shows the anonymized data after suppression technique is applied.

TABLE III. ANONYMIZED DATA AFTER SUPPRESSION

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| *** | *** | 117, Jalan Kinrara 1 | 35400 | 434 |
| *** | *** | 14, Jalan Presint 9/1 | 51400 | 289 |
| *** | *** | 23, Jalan Amanah 5 | 81200 | 45 |
| *** | *** | 89, Jalan Nuri 8 | 68100 | 872 |

Table IV shows the anonymized data after generalization and suppression take place. The main advantage of generalization and suppression is viable. On the other hand, the replaced and deleted values lead to a high possibility of information loss.

TABLE IV. ANONYMIZED DATA AFTER GENERALIZATION AND SUPPRESSION

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| *** | *** | Jalan Kinrara 1 | 35400 | 434 |
| *** | *** | Jalan Presint 9/1 | 51400 | 289 |
| *** | *** | Jalan Amanah 5 | 81200 | 45 |
| *** | *** | Jalan Nuri 8 | 68100 | 872 |

### C. Distortion

Distortion refers to a process that will change the data to something else, that later can revert back by using the original data. For example, selected attributes can be used and will be published [15], as in:

$$Vd = Vu + Vr,$$

As shown in the equation, Vu is the actual data and added with Vr to form Vd which is a distorted value "0". If they need to find the original data Vu = Vd – Vr. Same goes to hashing if they need to find the person they need to hash the original data and compare with the hashed data. This will be useful in the healthcare industry to identify the person after the research is done. For the development, MD5 was used to hash the values. MD5 hash function will produce a 128-bit length hash value. This hash value is not reversible, means that its one-way process.

The only way to know which hash belongs to the original value is by hash the original value and compare with the hash value in the database. There are other methods to get the hash value by dictionary attack or brute force. Dictionary attack will compare provided hash value with a hash value already in the dictionary corresponding without unhashed values. Brute force works by hashing every possible values 0-9 or a-z one by one to get the desired hash value but its take more time break the hash. The time depends on the length of the value.

If the original contain upper/lower case letters and digit, that gives ((26 letter* 2 case) + 10 digits) = 72 characters. Now if the value is 5 characters longer than the possibility to get the values is 72^5 = 1934917632 as the length increase more time is required to break the hash. Even without including symbols, the total characters required is the attempt is 75. If symbols are included in the hash, it will much harder to crack. Still, it is advisable to use a different hash function other than MD5 hence this hash function considered cryptographically broken and unsuitable for further use.

### D. Swapping

Swapping is a process of rearranging variable within each column randomly [16]. In this example, the attribute name can be used to scramble the data within the same attribute. This technique also cannot be applied for all attributes due to the result of research might not be accurate. The main concern in this technique is the probability to get the same value as the original value due to the randomization process. The output after the swapping process is shown in Table V.

TABLE V. ANONYMIZED DATA AFTER SWAPPING

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| 123 | Bob | 117, Jalan Kinrara 1 | 35400 | 434 |
| 234 | Sarah | 14, Jalan Presint 9/1 | 51400 | 289 |
| 456 | Alice | 23, Jalan Amanah 5 | 81200 | 45 |
| 789 | John | 89, Jalan Nuri 8 | 68100 | 872 |

### E. Masking

Masking refers to a method of changing characters in the selected attribute(s) to a different character, making the variable inconceivable. Any numeric from 1-9 will be replaced with 1, and any lower case a-z will be replaced with z, and any upper case A-Z will be replaced with Z. The first character, number 0 and special character will be maintained as the original value [16]. The problem of masking is that it will consume more resource for checking and changing the value but in the end, the data is useless for research. Instate of masking we can use suppression that will not check the value but change all the value to some symbols while archiving the same result as masking with more efficiency.

TABLE VI. ANONYMIZED DATA AFTER MASKING

| ID | Name | Address | Postcode | Usage (kW) |
|---|---|---|---|---|
| 123 | Alice | 117, Jalan Kinrara 1 | 3XXXX | 434 |
| 234 | Bob | 14, Jalan Presint 9/1 | 5XXXX | 289 |
| 456 | John | 23, Jalan Amanah 5 | 8XXXX | 45 |
| 789 | Sarah | 89, Jalan Nuri 8 | 6XXXX | 872 |

## III. PRIVACY PRESERVATION IN DATABASE APPLICATIONS USING ANONYMIZATION TECHNIQUES

Table VIII shows the best use of technique(s) based on the attribute in a database application. Some of the technique(s) is/are suitable for a specific attribute which depends on the nature of data itself. For example, we can generalize Postcode but not a Telephone number.

TABLE VII. ANONYMIZATION TECHNIQUES USED IN DATABASE APPLICATION

| Attribute name | Anonymization technique(s) best to use | Anonymization technique(s) can be used |
|---|---|---|
| Id | Used By The System as Primary Key | |
| Name | Swapping, Distortion, Suppression | Swapping, Distortion, Suppression Masking |
| Street Name 1 | Distortion, Suppression | Swapping, Distortion, Suppression Masking |
| Street Name 2 | Distortion Suppression | Swapping, Distortion, Suppression Masking |
| Telephone Number | Distortion, Suppression, Masking | Distortion, Suppression, Masking |
| Postcode | Generalization | Generalization, Distortion, Suppression, Masking |
| Meter Number | Distortion, Suppression | Distortion, Suppression, Masking |
| Age | Generalization, Distortion | Generalization, Distortion, Masking |
| Date Of Birth | Generalization, Suppression | Generalization, Suppression, Distortion |
| Town | Distortion, Suppression | Distortion, Suppression, Masking |
| Usage | Distortion, Suppression Masking | Swapping, Distortion, Suppression Masking |

Table VIII showed the attributes name versus the anonymization techniques reviewed in this paper. The result showed that for each attribute, the suitable anonymization techniques to apply to preserve privacy. From this table, we can summarize that most methods ideal for all types of data is the Masking and the Distortion techniques. As for Masking, this is due to its simplicity in converting numeric and text using symbols. However as mentioned in section (E), Masking will make data useless for research and a better technique to apply will be the Suppression. From the table, we can identify that almost all attributes can utilize the suppression technique too except for Age attribute. Result obtained from Suppression will be the same as the result of the Masking technique with more efficiency. Distortion works by adding the noise to the original data which make

original data become unrecognized and later can easily be revert back to the original data by removing the noise. This table also indicates most of the sensitive data in AMI can be anonymized with many techniques and we can hybrid these techniques to suit the needs of an application.

TABLE VIII. ATTRIBUTES VS ANONYMIZATION TECHNIQUES

| Attribute name | Anonymization technique(s) | | | | |
|---|---|---|---|---|---|
| | S1 | S2 | D | M | G |
| Name | √ | √ | √ | √ | |
| Street Name 1 | √ | √ | √ | √ | |
| Street Name 2 | √ | √ | √ | √ | |
| Telephone Number | √ | √ | √ | √ | |
| Postcode | | √ | √ | √ | √ |
| Meter Number | | √ | √ | √ | |
| Age | | | √ | √ | √ |
| Date of Birth | | √ | √ | √ | √ |
| Town | | √ | √ | √ | |
| Usage | √ | √ | √ | √ | |

*Legend: S1-Swapping, S2-Suppression, D-Distortion, M-Masking, G-Generalization*

## IV. CONCLUSION AND FUTURE WORKS

In smart grid data management, high-frequency usage data collected by AMI often contains sensitive information about the end consumers. When such data is shared by the utilities with external stakeholders, consumer privacy is at risk. In this paper, we compare several anonymization techniques for privacy preservation that can be used by organizations to anonymize their data. We have presented examples based on data anonymization techniques.

The main contribution of this paper lies in the techniques presented that accommodate a different type of data. From Table VII and Table VIII, the research showed that many techniques can be applied to any data, and which methods that suitable for what type of data. The strength and weaknesses of every technique have been discussed, and it is a guide to those to choose which techniques are best to apply in their works.

However, there is no "one size fits all" countermeasure for organizations to protect the privacy of personal data. Additional approaches should be taken into consideration when deciding on the use of suitable anonymization technique(s) such as the appropriate anonymization level, type of data, impact towards organization after considering risk management, and the utility required from the anonymized data. After all, we can use suitable anonymization technique(s) in order to prevent external stakeholders from obtaining identifiable data, while still enabling them to perform their respective functions.

## ACKNOWLEDGMENT

Work presented in this paper forms part of the research on Efficient Cryptographic-Based Technique in Industries Practicing Big Data, which was funded by Universiti Tenaga Nasional Internal Research Grant scheme (UNIIG).

## REFERENCES

[1] M. J. Silva, P. Rijo, and A. Francisco, "Evaluating the Impact of Anonymization on Large Interaction Network Datasets," *Proc. First Int. Work. Priv. Secuirty Big Data - PSBD '14*, pp. 3–10, 2014.

[2] M. Nithya and T. Sheela, "A Comparative Study on Privacy Preserving Datamining Techniques," *Int. J. Mod. Eng. Res.*, vol. 4, no. 7, pp. 11–14, 2014.

[3] E. Poovammal and M. Ponnavaikko, "APPT: A privacy preserving transformation tool for micro data release," *Proc. 1st Amrita ACM-W Celebr. Women Comput. India*, p. 29:1--29:8, 2010.

[4] C. Efthymiou and G. Kalogridis, "Smart Grid Privacy via Anonymization of Smart Metering Data," *Smart Grid Commun. (SmartGridComm), 2010 First IEEE Int. Conf.*, pp. 238–243, 2010.

[5] L. Yang, H. Xue, and F. Li, "Privacy-preserving data sharing in Smart Grid systems," *2014 IEEE Int. Conf. Smart Grid Commun.*, pp. 878–883, 2014.

[6] H. Kupwade Patil and R. Seshadri, "Big Data Security and Privacy Issues in Healthcare," *2014 IEEE Int. Congr. Big Data*, pp. 762–765, 2014.

[7] R. Xiangmin and Y. Jing, "Research on privacy protection based on K-anonymity," in *2010 International Conference on Biomedical Engineering and Computer Science*, 2010.

[8] I. Ozalp, "Privacy-Preserving Publishing of Hierarchical Data," *ACM Ref. Format Ismet Ozalp ACM Trans. Priv. Secur*, vol. 19, no. 29, pp. 1–28, 2016.

[9] Personal Data Protection Commission Singapore, "Guide to Basic Data Anonymisation Techniques," 2018.

[10] A. Kumar, M. Gyanchandani, and P. Jain, "A comparative review of privacy preservation techniques in data publishing," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 1027–1032.

[11] J.-H. Hoepman, "Privacy Design Strategies (extended abstract)," *SEC 2014 ICT Syst. Secur. Priv. Prot.*, vol. 428, no. 10532, pp. 446–459, 2014.

[12] V. Muntés-Mulero and J. Nin, "Privacy and anonymization for very large datasets," *Proceeding 18th ACM Conf. Inf. Knowl. Manag. - CIKM '09*, pp. 2117–2118, 2009.

[13] L. Sweeny, "k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1," *Int. J. Uncertainty, Puzziness Knowledge-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[14] N. Maheshwarkar, K. Pathak, and N. S. Choudhari, "K-anonymity Model for Multiple Sensitive Attributes," *Spec. Issue Int. J. Comput. Appl.*, no. 10, pp. 51–56, 2012.

[15] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," *ACM Trans. Database Syst.*, vol. 10, no. 3, pp. 395–411, 1985.

[16] F. Abdul Rahim, A. A. Bakar, S. Yussof, R. Ramli, R. Ismail, and B. M. Yusof, "Privacy Preserving Technique for Smart Metering Data : A Preliminary Result," *Adv. Sci. Lett.*, vol. 24, pp. 1839–1842, 2018.