A NOVEL FRAMEWORK FOR PRIVACY-PRESERVING DATA PUBLISHING WITH MULTIPLE SENSITIVE ATTRIBUTES

Saud Al-Otaibi[1] and lujain Al-Qurashi[2]

[1]Department of Information Systems, Umm Al-Qura University,
Makkah, Saudi Arabia
[2] Department of Computer Science and Engineering, Umm Al-Qura University,
Makkah, Saudi Arabia

ABSTRACT

The world is now experiencing a great technological revolution, as many fields have become dependent on it. The use of technology by members of society has become daily. Data is collected on individuals by using smart technology applications in hospitals or companies. These organizations are managed through databases that record data about their customers. The collected data may include sensitive data (e.g., personal data) that individuals do not want to disclose. In order to continue development , we sometimes need to publish this data for the purposes of research, statistical studies or decision-making. The publication of this data constitutes a threat to the privacy of the customer as it can be exploited by the intruder.

This research focuses on trying to provide Privacy-Preserving Data Publishing algorithm that preserves customer privacy with the possibility of publishing this data with less information loss.

KEYWORDS

trust Privacy-Preserving Data Publishing (PPDP), l-diversity

1. INTRODUCTION

The world is now experiencing a great technological revolution, as many fields have become dependent on it. The use of technology by members of society has become daily. The using of smart technology applications in smart cities leads to the collection of many data about individuals. The collected data may include sensitive data (e.g., personal data) that individuals do not want to disclose. This data is published to public for the purposes of research, statistical studies or decision-making. usually Privacy-Preserving Data Publishing by the traditional approach of removing direct identifying fields such as name or national number. This approach is not sufficient because it has been proven that the intruder can take advantage of Quasi identifiers (QID's) and reveal identity[1].

Figure 1 shows the stages of managing published data and that it passes through three stages: data collection, data storage, and data publishing . The intruder can obtain published data in data publishing stage to disclose identity .Therefore, the disclosure of published data threatens the privacy of the individual because it can be used to reveal identity .Hence the urgent need to use the anonymization techniques to apply it to the data before publishing it.

David C. Wyld et al. (Eds): SOEN, SIPP, PDTCA, ITE, CCSIT, NLPCL, DaKM, BIGML, AISC -2023
pp. 111-117, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131210

When the data is collected, it is published in the form of micro-data. Micro-data is a file made up of records 'n'. Each record consists of a number of attributes of individuals 'm'. For example, in the field of business, data about employees is collected to make a decision, as shown in Table 1. In this table, it is clear that there are different types of attributes: Identifying attributes, Quasi identifiers (QID's) and Sensitive Attributes (SA's). Identifying attributes directly reveals the identity of the individual and often this data is never published, that is, it is deleted before publication like name and ID. Quasi identifiers (QID's) indirectly help disclosure the attributes of an individual. Sensitive Attributes (SA's) is the personality attributes that the individual does not want to disclose like salary . Sensitive attributes are often multiple sensitive attributes.

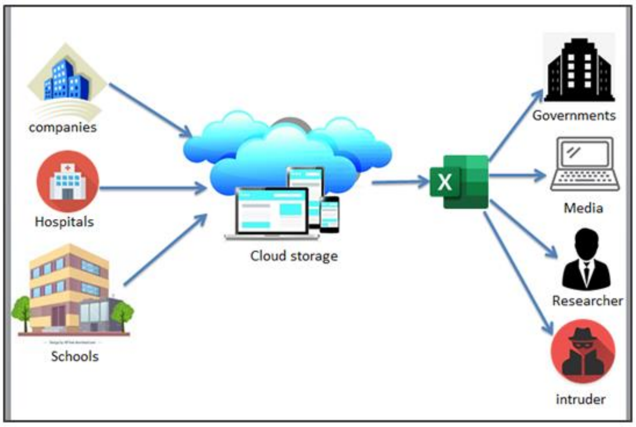
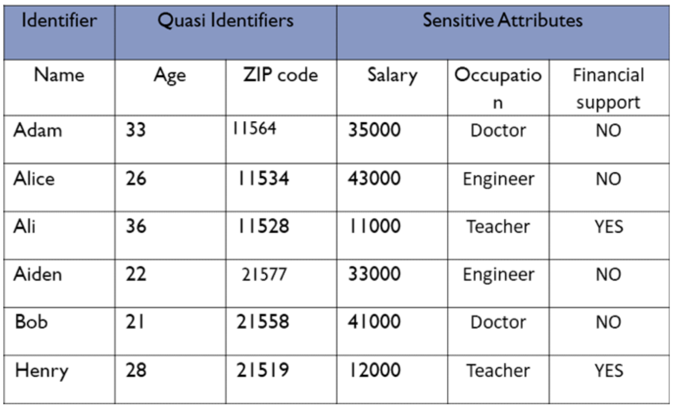


Figure 1. Data Publishing Stages

Table 1. Microdata table.



Identifier	Quasi Identifiers		Sensitive Attributes		
Name	Age	ZIP code	Salary	Occupation	Financial support
Adam	33	11564	35000	Doctor	NO
Alice	26	11534	43000	Engineer	NO
Ali	36	11528	11000	Teacher	YES
Aiden	22	21577	33000	Engineer	NO
Bob	21	21558	41000	Doctor	NO
Henry	28	21519	12000	Teacher	YES

2. CONTRIBUTIONS OF THE PAPER

- This study aims to provide an effective algorithm that forms a framework for Preserve the privacy of published data.
- In this study, many basic principles in the field of Privacy-Preserving Data Publishing will be discussed.
- In this study, there is discussion for many previous study about anonymization operations and privacy preserving algorithms.

3. ANONYMIZATION OPERATIONS

As mentioned previously, the published data is in the form of micro-data, which means that it is present in all its types (Identifying attributes, Quasi identifiers Attributes (QID's) and Sensitive Attributes (SA's) in single table. We need to separate the table by the anatomization.

Anatomization is a method of separating the main table into separate tables according to the type of data[1]. It separates the table vertically without distinguishing the data type. The outputs from the anatomization process are a table of sensitive data (ST) and a table of Quasi identifiers table(QT).

Slicing is used to horizontally to portion to form groups called tuple[2]. Slicing is used to apply l-diversity requirement and assign group IDs.l-diversity means every equivalence class should have been at least l well represented values for the sensitive attribute[3].

4. RELATED WORK

With the rapid development in the world, there is a need to use modern technology on a daily basis [4]. Government sectors and organizations collect a huge amount of data about individuals [5].

Therefore, there is a need to protect this data from disclosure. Several studies have been presented in the field of Privacy-Preserving Data Publishing (PPDP). All of these studies aim to provide privacy to the individual in an effective, safe and more reliable manner. PPDP contain two steps: data collection and data publishing phase [4].

In 2006 one of anonymization method of privacy-preserving Data publishing is Anatomy that is discussed in [6]. This method is used to hide identity. It is used in the process of generalization and suppression, as there is no modification to quasi-identifiers or sensitive attributes.In this method, the relationship between the quasi attributes and the sensitive attributes are separated. In 2009 While seeking to improve the data utility, taking into account privacy, ANGEL method appeared [7]. This method depends on the generalization and improve it.

In 2011 RATING algorithm is proposed create Attribute Table and an ID Table that are based on various sensitivity coefficients for different attributes[8].

In 2013 SLOMS proposed to be dealt with the multiple sensitive attributes where sensitive attributes are divided vertically into several tables and then grouped to meet the requirements of the I-diversity[9].The problem with this method is that there is a large percentage of information loss.

In 2016 Another method of anonymization method is Anatomization with Slicing .In this method, the anatomy and slicing techniques are combined. The anatomy method and slicing method produce the quasi-identifier table as the original form and divides the table horizontally and exchanges the sensitive attributes in each partition. The disadvantages of this methods generate more complex tables[10] .

In 2019 Raju et al proposed usethe distribution of sensitive values[11]. In this method, the distribution of sensitive values is used as a method to preserve privacy where there is similarity between them and group of quasi-identifier attributes.

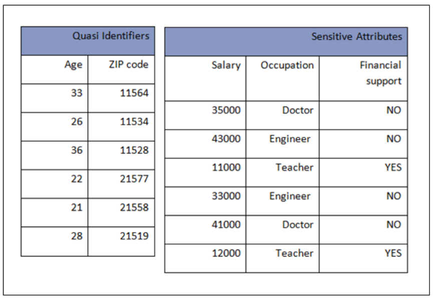
In 2020 ,[12] use special format of l-diversity that called (l, d) semantic diversity to Privacy-preserving data publishing with multiple sensitive attributes.

In 2021 Susan also proposed clustering method for Privacy-preserving Data Publishing With Multiple Sensitive Attributes [13].

5. METHODOLOGY

This study seeks to provide algorithm for Privacy-Preserving Data Publishing. At the beginning, we will explain the inputs and outputs of this algorithm. The inputs will be collected by companies or government agencies, either for the purpose of statistical studies or others. This data is either sensitive , Quasi-identifier data or identifier data. The inputs deals as micro data. identifier data is usually removed before publishing. .The outputs of this algorithm are anonymized tables. This algorithm will be detailed in the form of steps. Inputs, outputs and the goal of each step will be explained. The first step is to apply anatomy method in order to split the micro-data table vertically to separate sensitive data table ST and Quasi Identifiers data table QT as in Figure 2.The input: micro-data table,outputs:ST and QT and objective: to separate sensitive data from Quasi Identifiers data. The second step is to create dependency table(DT) where each sensitive attribute is assigned a SID and determines its dependence on the remaining sensitive attributes as shown in Figure 3. Input: ST, output:DT and objective:Create a dependency table to use in step 3. The third step is to create the sensitive tables STi separately, depending on DT that resulting from the second step, as shown in Figure 4.Inputs:DT, outputsST1,ST2 ...STi and objective: to be used in building anonymized tables. The fourth step is to slice each STi horizontally so as to satisfy l-diversity requirement and assign group IDs as shown in Figure 5 .The Inputs:STi, outputs: STi with The slicing that meets the l-diversity requirement, and objective:To provide tables that contribute to creating an anonymized table. The fifth step is to slice each STi horizontally that contain numerical value so as to satisfy l-diversity requirement and assign NIDs as shown in Figure 6 .The Inputs:STi that contains numerical value , outputs: NSTi with The slicing that meets the l-diversity requirement, and objective:To provide tables that contribute to creating an anonymized table. The Sixth step is to construct the anonymized table (AT) from QT in the original form and the SA that are mapped to their respective tables (STi,NSTi) as shown in Figure 6 .The Inputs:STi ,NSTi and QT , outputs: anonymized table AT and objective:Configure an anonymous table so that it can be published without worrying about identity detection.

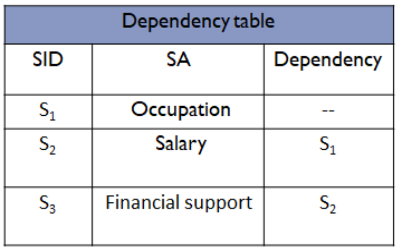
The following tables are published STi, NSTI, QT and AT without worrying about identity detection.



Quasi Identifiers	
Age	ZIP code
33	11564
26	11534
36	11528
22	21577
21	21558
28	21519

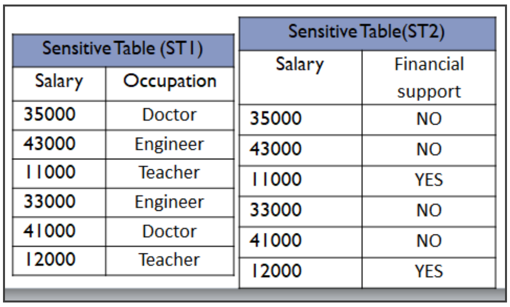
Sensitive Attributes		
Salary	Occupation	Financial support
35000	Doctor	NO
43000	Engineer	NO
11000	Teacher	YES
33000	Engineer	NO
41000	Doctor	NO
12000	Teacher	YES

Figure 2. Separating QT and ST



Dependency table		
SID	SA	Dependency
S_1	Occupation	--
S_2	Salary	S_1
S_3	Financial support	S_2

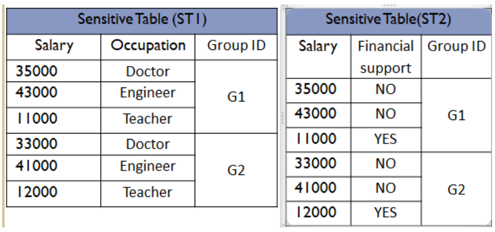
Figure 3. Dependency Table



Sensitive Table (ST1)	
Salary	Occupation
35000	Doctor
43000	Engineer
11000	Teacher
33000	Engineer
41000	Doctor
12000	Teacher

Sensitive Table(ST2)	
Salary	Financial support
35000	NO
43000	NO
11000	YES
33000	NO
41000	NO
12000	YES

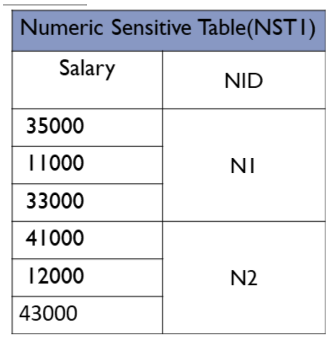
Figure 4.Sensetive Tables (STi)



Sensitive Table (ST1)		
Salary	Occupation	Group ID
35000	Doctor	G1
43000	Engineer	
11000	Teacher	
33000	Doctor	G2
41000	Engineer	
12000	Teacher	

Sensitive Table(ST2)		
Salary	Financial support	Group ID
35000	NO	G1
43000	NO	
11000	YES	
33000	NO	G2
41000	NO	
12000	YES	

Figure 5. STi After Apply l-diversity



Numeric Sensitive Table(NST1)	
Salary	NID
35000	N1
11000	
33000	
41000	N2
12000	
43000	

Figure 6. Numeric Sensetive Table(NSTi)

6. CONCLUSIONS

Privacy threats increase dramatically on publishable data, which contains sensitive data that must be preserved and anonymised. This paper discusses a proposed algorithm to preserve the privacy of published data.It also provides an effective model against privacy threats for the published data.As part of future work it can be demonstrated how this algorithm contributes to reduce the information losing and execution time and increase diversity degree .

REFERENCES

[1] Xiao, X., & Tao, Y, “Anatomy: Simple and effective privacy preservation”, In Proceedings of the 32nd international conference on Very large data bases , pp. 139-150,2006.

[2] Li, Tiancheng, Ninghui Li, Jian Zhang, and Ian Molloy. "Slicing: A new approach for privacy preserving data publishing." IEEE transactions on knowledge and data engineering vol. 24, issue 3 pp. 561-574, 2010.

[3] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M, “l-diversity: Privacy beyond k-anonymity”, ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, issue 1, 2007.

[4] R. Liu, “Privacy-Preserving Data Publishing,” ICDE Workshops, 2010.

[5] E. Ismagilova, L. Hughes, Y. K. Dwivedi, K. R. Raman, “Smart cities: Advances in research—An information systems perspective,” International Journal of Information Management ,pp. 88–100,2019.

[6] Xiao, X., & Tao, Y, “Anatomy: Simple and effective privacy preservation”, In Proceedings of the 32nd international conference on Very large data bases , pp. 139-150, 2006.

[7] Tao Y, Chen H, Xiao X, Zhou S, Zhang D,”Angel: enhancing the utility of generalization for privacy preserving publication”, IEEE Trans Knowl Data Eng, vol. 21, issue 7,pp.1073–1087,2009.

[8] Liu J, Luo J, Huang JZ, “Rating: privacy preservation for multiple attributes with different sensitivity requirements”, In: IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp 666–673, 2011.

[9] Han J, Luo F, Lu J, Peng H, “SLOMS: a privacy preserving data publishing method for multiple sensitive attributes microdata”, JSW , vol.8,issue 12,pp. 3096–3104, 2013.

[10] Susan, V. Shyamala, and T. Christopher, "Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes." SpringerPlus, vol.5 issue 1, pp. 1-21, 2016.

[11] N.V.S.L. Raju, M.N. Seetaramanath, and P.S. Rao, "A Novel Dynamic KCi-Slice Publishing Prototype for Retaining Privacy and Utilityof Multiple Sensitive Attributes," International Journal of Information Technology and Computer Science, vol. 4, pp. 18-32, 2019.

[12] K. Oishi,Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: Privacy considering distance between values of sensitive attribute," Computers & Security, vol. 94, no. 101823, 2020.

[13] V. S. Susan, "An Anonymization Approach for Dynamic Dataset with Multiple Sensitive Attributes," Intelligent Computing and Applications, vol. 1172, pp. 731-739, 2021.

AUTHORS

Dr. Saud S. Alotaibi is an assistant professor of Computer Science at the Umm Al-Qura University, Makkah, Saudi Arabia. He received his Bachelor of Computer Science degree from King Abdul Aziz University, 2000. Dr. Saud started his career as Assistant Lecturer in July 2001 at Umm Al-Qura University, Makkah, Saudi Arabia. He then earned his Master degree in Computer Science from King Fahd University, Dhahran, May 2008. After that, he worked as a Deputy of IT-Center for E-Government and Application Services, January 2009, in Umm Al-Qura University. Under Dr. Charles Anderson supervision, Saud completed his Ph.D. degrees in Computer Science from Colorado State University in Fort Collins, US, August 2015. Nowadays, Dr. Saud is working right now with the Deanship of Information technology to improve the IT services that are provided to the Umm Al-Qura University. He can be contacted at ssotaibi (dot) uqu (dot) edu (dot) sa.

Lujain O. AlQurashi is a student at Umm Al-Qura University, Saudi Arabia, Department of Computer Science and Engineering. She obtained a Bachelor’s degree inComputer Science from King Abdulaziz University in 2017 and joined Umm Al-Qura University to obtain a master’s degree. She can be contacted at: s43980133@st.uqu.edu.sa