A preliminary phase on anatomizing multiple sensitive attribute by determining main sensitive attribute

W Widodo1,* and A Wahyudin2

1 Department of Informatics Education, Universitas Negeri Jakarta Jl. Rawamangun Muka, Jakarta, Indonesia
2 Department of Computer Science, Universitas Pendidikan Indonesia, Bandung, Indonesia

*widodo@unj.ac.id

Abstract. This paper aims to determine the main sensitive attribute in anatomy with multiple sensitive attributes. Privacy in published data becomes important and interesting research recently. Many works have been conducted on keeping published data privately. K-anonymity was proposed as the first model in this study. Yet, this model still has some drawbacks. A better model called anatomy is proposed to fix k-anonymity. Many studies on anatomy are performed on a single sensitive attribute, while in the real world, a microdata table always contains multiple sensitive attributes. This work handles problems on how to publish data in private mode by rearranging the anatomy model. We adopted the method in determining main sensitive attribute. We conducted it in the preliminary phase and its result is a prepared model of the main sensitive attribute for anatomy. The result success in ensuring one sensitive attribute as a main sensitive attribute.

1. Introduction

Process in published data using anatomy for multiple sensitive attributes can be started by determining main sensitive attribute. Related to this, the main resource are data. Data also play an important role in any field of research. For certain reason, data also can be shared to other parties. When data is shared or published it need to private some important and sensitive information. A field which studies privacy in published data is Privacy-Preserving Data Publishing (PPDP). Many methods have been developed for handling privacy problems. It is started by a method called k-anonymity [1-2], l-diversity [3], also p-sensitive [4]. All those basic methods are run by hiding or generalizing quasi identifier attributes. Quasi identifier attributes are type of attribute that can be potentially a key when it is generalized or suppressed [5]. Another method that segregate quasi identifier and sensitive attributes which are called anatomy [6]. Anatomy has been studied and developed in many works. However, most of them are conducted for a single sensitive attribute. In the real world, most of the data contain multiple-sensitive attributes, therefore it is necessary to investigate deeply how anatomy works on it.

Many works have been conducted in the anatomy model, and a few discussed in multiple sensitive attributes. A work in anatomy for achieving a p-sensitive model has been conducted [7]. They attempted to have p-sensitive conditions by using anatomy techniques. However, they performed in a single sensitive attribute. Then, a study that combined anatomy and slicing method is conducted [8]. This work used microdata with multiple sensitive attributes, but they were not set which one is the main sensitive
attribute. Data anonymization then enhanced the anatomy model into permutation anonymization [9]. Information can be retained more efficiently compared to conventional anatomy. They utilized microdata with single sensitive attribute. An excellent work called ANGELMS was performed for multiple sensitive attributes [10]. Anatomy is the method they used, but unfortunately, they still did not consider which main sensitive attribute. This work is an extended work of ANGEL [11]. Multiple sensitive attributes are also used in recent works [12-14], but they did not use the anatomy method and even not set the main sensitive attribute. Work is done in multiple sensitive attributes and set the main sensitive attribute called primary sensitive attribute [15]. However, it was also run for the p-sensitive model, not anatomy.

Anatomy in single sensitive attribute is a simple form. First, we have to understand data that can be processed in anatomy is a microdata table, not a macrodata table. A Microdata is a table contains individual data, while macrodata is aggregative [16]. Microdata usually has two types of attributes, quasi identifier attribute, and sensitive attribute. Quasi identifier is a set of attributes that can be generalized to disguise records in a group. A sensitive attribute is an attribute that contains sensitive values. Both attributes in anatomy then separated each other. Its purpose is to break the correlation between them and maintain privacy. This method is considered better than k-anonymity which is only generalizing quasi identifier attributes for having privacy. The anatomy model is illustrated in the tables below. We describe the original table in Table 1, the k-anonymity table in Table 2, quasi identifier table in Table 3 and a sensitive table in Table 4.

**Table 1.** Original microdata table.

| Name | Age | Gender | Zipcode | Disease |
|------|-----|--------|---------|---------|
| 22   | M   | 11000  |         | pneumonia |
| 26   | M   | 13000  |         | dyspepsia |
| 34   | M   | 59000  |         | dyspepsia |
| 56   | M   | 12000  |         | pneumonia |
| 61   | F   | 54000  |         | flu |
| 66   | F   | 26000  |         | gastritis |
| 66   | F   | 26000  |         | flu |
| 69   | F   | 30000  |         | bronchitis |

Table 1 shows the original microdata table, consists of quasi identifier attributes (*age, gender, zipcode*) and sensitive attribute (*disease*). *Name* is identifier attribute and this attribute will be hidden when it is anatomized.

**Table 2.** Microdata table in k-anonymity.

| Age     | Gender | Zipcode     | Disease |
|---------|--------|-------------|---------|
| [21 – 60] | M     | [10001 – 60000] | pneumonia |
| [21 – 60] | M     | [10001 – 60000] | dyspepsia |
| [21 – 60] | M     | [10001 – 60000] | dyspepsia |
| [21 – 60] | M     | [10001 – 60000] | pneumonia |
| [61 – 70] | F     | [10001 – 60000] | Flu |
| [61 – 70] | F     | [10001 – 60000] | gastritis |
| [61 – 70] | F     | [10001 – 60000] | Flu |
| [61 – 70] | F     | [10001 – 60000] | bronchitis |

Table 2 shows an anonymized table using k-anonymity which is divided into 2 groups, the first group is represented by record 1 to 4, and the second group is represented by record 5 to 8. Each group has the same quasi identifier attributes that have been generalized. Generalization of *Age* is {*[21-60]* and *[61-70]*} and disease is {*10001 – 60000*}. 

---

2
Table 3. Quasi identifier table.

| Age | Gender | Zipcode | GroupID |
|-----|--------|---------|---------|
| 22  | M      | 11000   | 1       |
| 26  | M      | 13000   | 1       |
| 34  | M      | 59000   | 1       |
| 56  | M      | 12000   | 1       |
| 61  | F      | 54000   | 2       |
| 66  | F      | 26000   | 2       |
| 66  | F      | 26000   | 2       |
| 69  | F      | 30000   | 2       |

Table 4. Sensitive table.

| GroupID | Disease | Count |
|---------|---------|-------|
| 1       | dyspepsia | 2     |
| 1       | pneumonia | 2     |
| 2       | bronchitis | 1     |
| 2       | Flu       | 2     |
| 2       | Gastritis | 1     |

Table 3 and Table 4 are the result of anatomizing process, it is segregated into two tables, Table 3 is quasi identifier table and Table 4 is a sensitive table. Both tables are connected by GroupID which represents the identity of the quasi identifier group from the k-anonymity table. Table 4 also being added count attribute which represents the number of certain sensitive values in its record.

This paper has one main contribution, it is how to determine the main sensitive attribute from microdata with multiple sensitive attributes for anonymizing table. This is necessary since the main sensitive attribute becomes a basis in distributing sensitive table.

2. Methodology
Anatomy is a method for separating quasi identifier attributes from sensitive attribute. Therefore, it extends a table into two. If the number of sensitive attributes is more than one, it is possible that form two or more tables. In this study, we focus on the preliminary phase of anatomizing, not all processes. The first phase of multiple sensitive anatomizing is to determine the main sensitive attribute. This is useful for configuring and distributing its sensitive values. A method in Wibowo is adopted in this study [15]. This method for determining main sensitive attribute is utilized in p-sensitive and called primary sensitive attribute. Figure 1 below exhibits the method.
Figure 1 shows the methodology we used that is adopted from Wibowo [15]. First, sensitive attributes should be separated from quasi identifier attributes. Second, in each sensitive attribute, *High-sensitive value* (HSV) should be assigned. HSV is values of a sensitive attribute which have high sensitivity. It is user-defined. Then, number of HSV in each sensitive attribute is counted. Sensitive attribute that contains most HSV is a Primary Sensitive Attribute (PSA). If there is more than one sensitive attributes have most HSV, then one which has more vary of sensitive value is set as PSA. PSA is a main sensitive attribute that will be an attribute which is used to configure and distribute sensitive value when the anatomizing process is performed.

3. Results and discussion
Due to this study is a preliminary phase of our research, we have not experimented yet. In this section, we discuss our proposed method to determine the main sensitive attribute using method in Budiardjo and Wibowo [12] and Wibowo [15]. We create a simple dummy table for our discussion as shown in Table 5 below. The table is a sensitive table that has been separated from quasi identifier attributes.

|   | SAt_1 | SAt_2 | SAt_3 |
|---|-------|-------|-------|
| X | A     | P     |       |
| X | A     | P     |       |
| Y | A     | Q     |       |
| Y | A     | Q     |       |
| Z | B     | R     |       |
| Z | B     | S     |       |

Table 5 shows a dummy sensitive table for our discussion. There are three sensitive attributes, *SAt_1*, *SAt_2*, and *SAt_3*. *SAt_1* has three sensitive values, *X*, *Y*, and *Z*. *SAt_2* has two sensitive values, *A* and *B*, while *SAt_3* has four sensitive values, *P*, *Q*, *R*, and *S*. By user-defined, HSV in *SAt_1* are *X* and *Y*, HSV in *SAt_2* is *A*, and HSV in *SAt_3* is *P*. It is clearly seen that *SAt_3* must be removed from PSA candidate since this attribute only has 2 HSV. *SAt_1* and *SAt_2* has 4 HSV each. However, *SAt_1* is chosen as PSA because *SAt_1* has more vary in sensitive value. They have 2, *X* and *Y*, while *SAt_2* only 1, that is *A*. Hence, *SAt_1* is a PSA automatically this sensitive attribute is determined as main sensitive attribute. This attribute will be an attribute basis when anatomizing process being performed. We do not discuss the anatomizing process since this research is still not finished yet and this is a preliminary phase for assigning main sensitive attribute.

In determining main sensitive value, we use distribution of HSV model in primary sensitive attribute in Budiardjo and Wibowo [12] and Wibowo [15]. The method is to count numbers of HSV in each sensitive attribute. The sensitive attribute that has most HSV is set as primary sensitive attribute. It is reasonable due to this attribute is used to be a basis when distribution of sensitive values is performed. This distribution ensures each quasi identifier attribute have equal number of sensitive values.

4. Conclusion
As we described previously, this work is a preliminary phase for the anatomy method with multiple sensitive attributes. Therefore, we do not experiment with our whole model due to it is not finished yet. In this preliminary phase, we have adopted a method for determining the main sensitive attribute in p-sensitive. This method can run well and our work can be continued to the next phase. A main sensitive attribute can be determined by counting HSV in each sensitive attribute. HSV is user-defined. The attribute that most contains HSV is assigned as a main sensitive attribute. We called it a Primary sensitive attribute. The next phase, as our next work is using this PSA as a basis to form anonymity with multiple sensitive attributes. It can be also for distributing sensitive values when anatomy is processed.
References

[1] Wagner I and Eckhoff D 2018 Technical Privacy Metrics: A Systematic Survey ACM Computing Survey 51(3) 57
[2] Arava K and Lingamgunta S 2020 Adaptive k-Anonymity Approach for Privacy Preserving in Cloud Arab. J. Sci Eng. 45 2425–2432
[3] Mehta B B and Rao U P 2019 Improved l-Diversity: Scalable Anonymization Approach for Privacy Preserving Big Data Publishing Journal of King Saud University-Computer and Information Sciences xxx(xxxx) xxx
[4] Truta T M and Vinay B 2006 Privacy protection: p-sensitive k-anonymity property 22nd International Conference on Data Engineering Workshops (ICDEW'06) (IEEE) pp 94-94
[5] M.K. Kundawal, K. Chatterjee, and A. Singh 2018 An Improved privacy preservation technique in health-cloud. ICT Express 5 (3) pp. 167–172.
[6] C. Liu, S. Chen, S. Zhou, J. Guan, and Y. Ma 2019 A novel privacy preserving method for data publication. Information Sciences 501 421-435.
[7] Sun X, Wang H, Li J and Ross D 2009 Achieving p-sensitive k-anonymity via anatomy 2009 IEEE International Conference on e-Business Engineering (IEEE) pp 199-205
[8] Susan V S and Christopher T 2016 Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes SpringerPlus 5(1) 964
[9] He X, Xiao Y, Li Y, Wang Q, Wang W and Shi B 2011 Permutation anonymization: Improving anatomy for privacy preservation in data publication Pacific-Asia Conference on Knowledge Discovery and Data Mining (Berlin, Heidelberg: Springer) pp 111-123
[10] Luo F, Han J, Lu J and Peng H 2013 ANGELMS: a privacy preserving data publishing framework for microdata with multiple sensitive attributes 2013 IEEE Third International Conference on Information Science and Technology (ICIST) (IEEE) pp 393-398
[11] Tao Y, Chen H, Xiao X, Zhou S and Zhang D 2009 Angel: Enhancing the utility of generalization for privacy preserving publication IEEE transactions on knowledge and data engineering 21(7) 1073-1087
[12] Budiardjo E K and Wibowo W C 2019 Privacy Preserving Data Publishing with Multiple Sensitive Attributes based on Overlapped Slicing Information 10(12) 362
[13] Khan R, Tao X, Anjum A, Sajjad H, Khan A and Amiri F 2020 Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity Wireless Communications and Mobile Computing 2020 1-18
[14] Han J, Luo F, Lu J and Peng H 2013 SLOMS: A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Microdata JSW 8(12) 3096-3104
[15] Wibowo W C 2018 A distributional model of sensitive values on p-sensitive in multiple sensitive attributes 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS) (IEEE) pp 1-5
[16] Can O 2018 Personalised anonymity for microdata release IET Information Security 12(4) 341-347