Extended systematic clustering: Microdata protection by distributing semsitive values

Widodo[1], Wahyu Catur Wibowo[2], Eko K. Budiardjo[3], Harry T. Y. Achsan[4]
[1]Department of Informatics Education, Universitas Negeri Jakarta, Indonesia
[2,3,4]Faculty of Computer Science, University of Indonesia, Indonesia

Article Info

Article history:

Received Sep 5, 2019
Revised Dec 10, 2019
Accepted Feb 14, 2020

Keywords:

Diversity value
Extended systematic clustering
High-sensitive value
K-anonymity

ABSTRACT

Anonymity data for multiple sensitive attributes in microdata publishing is a growing field at present. This field has several models for anonymizing such as k-anonymity and l-diversity. Generalization and suppression became a common technique in anonymize data. But, the real problem in multiple sensitive attributes is sensitive value distribution. If sensitive values do not distribute evenly to each quasi identifier group, it is potentially revealed to sensitive value holder. This research investigated on how the high-sensitive values are distributed evenly into each group. We proposed a novel method/algorithm for distributing high-sensitive values when it forms groups. This method distributes high-sensitive values evenly and varies high-sensitive values in a group. We called our method as extended systematic clustering since it is an extension of systematic clustering method. Diversity metrics was used for evaluating our method. Experiment result showed our method outperformed systematic clustering with average diversity value 0.9719 while systematic clustering 0.3316.

Corresponding Author:

Widodo,
Department of Informatics Education,
Universitas Negeri Jakarta,
Jl. Rawamangun Muka, Jakarta, 13220, Indonesia.
Email: widodo@unj.ac.id

1. INTRODUCTION

Privacy is an important issue in publishing microdata table, while microdata contains information of individual dan identities data. An individual data covers three type of attributes that is called explicit identifier (EI), quasi identifier (QI), and sensitive attributes (SA) [1, 2]. EI is an attribute that contains an identifier such as name, employee number, or student identifier. Quasi identifier is two or more attributes which potentially become identifier when anonymity is conducted, while sensitive attribute is attribute that have a certain sensitivity value for person. In privacy preserving data publishing (PPDP), QI attributes are generalized or suppressed for obtaining anonymity table. Some records that the QI attributes cannot be distinguished formed quasi identifier groups. A table that contains some groups which each group has at least k records is called k-anonymity table [3-7].

Table 1 shows a simple example of microdata table. Name is an explicit identifier, Age and Zipcode are quasi identifiers, while disease is sensitive attribute. In k-anonymity, explicit identifier is removed and quasi identifier is generalized and/or suppressed. Table 2 exhibits k-anonymity model of Table 1. In each group, quasi identifiers are indistinguishable. As it is seen in Table 2, {age, zipcode} at group 1 contains {21-25, 1****} and three records are similar. Age is generalized so that in one group they cannot be distinguished while Zipcode is suppressed using '*'.

A sensitive attribute is attribute that have sensitive values specially to person who suffer a serious illness and the illness suffered tend to be embarrassing. In Table 1 and Table 2, disease attribute has flu, cancer, HIV, bronchitis, and diarrhea. To some people, HIV and cancer are embarrassed disease then someone who suffers from those could be shamed. The problem is when a group contains high sensitivity values (no low sensitive values), simply adversaries can guess by his/her background knowledge, that someone in such group is suffering disease with high-sensitive value.

Table 1. A microdata table

Name	Age	Zipcode	Disease
AA	21	15321	Flu
BB	23	17999	Cancer
CC	25	16330	HIV
DD	27	16200	HIV
EE	27	16200	HIV
FF	29	16200	HIV
GG	31	15217	Flu
HH	34	15219	Bronchitis
II	37	15211	Flu
JJ	38	15217	Diarrhea

Table 2. k-anonimity model

Group	Age	Zipcode	Disease
1	21-25	1****	Flu
1	21-25	1****	Cancer
1	21-25	1****	HIV
2	26-30	16200	HIV
2	26-30	16200	HIV
2	26-30	16200	HIV
3	31-40	1521*	Flu
3	31-40	1521*	Bronchitis
3	31-40	1521*	Flu
3	31-40	1521*	Diarrhea

We found that in previous research, distribution of sensitive values did not consider as a serious problem, as we describe in related works. Sensitive values distribution is an important aspect when quasi identifier groups are formed. If a group contains same level of sensitive values more over same sensitive values, then adversaries with his/her background knowledge can guess a sensitive value holder with high probalility. As depicted in Table 2, if adversaries know someone's age is 29, therefore adversaries know exactly he/she in group 2, then he can guess that FF has HIV because all sensitive values in group 2 are HIV. This research investigated on how sensitive values with high sensitivity is distributed evenly to each quasi identifier group. If sensitive values with high sensitivity is not distributed, it will be lot of stack of those in one group. This distribution should decrease the probability of guessing someone's disease (high sensitive disease).

Research on distribution of sensitive attribute's values in anonymity is very rare. Some studies were part of larger research therefore its focus did not investigate on it. Study conducted by Liu et al [8] was one of some studies that focuses on distribution of sensitive values. Liu created an algorithm for distributing sensitive values in multiple sensitive attributes. Sensitive vales are categorized into highly sensitive, which having high sensitivity, and lowly sensitive value. Simply, he distributed a tuple based on number of sensitive values of each quasi identifier group. If a tuple has a highly sensitive value, candidates of quasi identifier group are those containing least high-sensitive values, otherwise all groups are candidate. They proved that their algorithm destroyed association among sensitive attributes, but explicitly the did not evaluate whether sensitive values are distributed evenly or not. A study by Zhang et al [9] also investigated on how to distribute sensitive values to balance and meet the diversity. Unfortunately, Zhang did not measure the diversity of sensitive values in the table, because his study concentrated to improve algorithm of individuation k-anonymity, therefore he focused in measuring the information loss. Susan and Christopher [10] also distributed sensitive attributes values. They applied advanced clustering algorithm (ACA) [11] to distribute and cluster relevant sensitive attributes. But, it was to partition attribute and slice vertically not to distribute highly sensitive values. A study by Ye et al [12] also distributed sensitive values but his study did not categorize sensitive value based on level of sensitivity. Hasan et al [13] investigated multiple independent data publishing. He set to distribute sensitive attributes value therefore group of quasi identifiers would not contain many same sensitive values. The levels of sensitive values were not considered. Distribution technique of sensitive attribute values following l-diversity [14] and t-closeness [15] is conducted in an investigation for privacy protecting microdata publication based on distribution of conditional probability and machine learning [16, 17]. The study was investigated for single sensitive attribute and multiple sensitive attributes. The experiment shows a good privacy guarantee and better data utility. Otherwise, they did not implicitly state the effect on distribution of sensitive values and did not categorized the types of distributed sensitive values.

A study by combining method of bucketization between anatomy and generalization resulted a better diversity of sensitive values in quasi identifier group [18]. Increasing of diversity ensure that threat to the disclosure in microdata table is reduced. Investigation by Man et al [19] found that l-diversity can lead to privacy leakage when the distribution of sensitive values is uneven. They grouped sensitivity rate of sensitive values and put those by distributing evenly on each quasi identifier group. Unfortunately, their work is still not proved by experiment because they still have some insufficient theory. A distributional model of sensitive values has been conducted for a main investigation [20]. This model is distributed evenly sensitive values based on highly sensitive values in the primary sensitive attribute. It is a simple distribution on sensitive values and experiment resulted high diversity of high sensitive values in quasi identifier groups.

Therefore, this research motivation is to distribute evenly high-sensitive values into quasi identifier groups and to improve the variety of high-sensitive values in a group. The variety of high-sensitive value can be reduced probability to guess and link a sensitive value to its holder. We called our method as Extended Systematic Clustering since it extends systemtic clustering method in grouping quasi identifiers [21, 22]. Systematic clustering is considered as an excellent method in grouping quasi identifers because its systematic way. We modify this method when it forms groups, sensitive value in tuple must be checked whether it is high-sensitive value or not. The contributions of this research are, (1) we proposed a novel algorithm for distributing high-sensitive values to each quasi identifier group, (2) we successfully implemented our method in multiple sensitive attributes, (3) we categorized sensitive values and set it into sensitive attribute categorization.

2. RESEARCH METHOD

We conducted this investigation on microdata with multiple sensitive attributes. Briefly, this research consists of three steps, those are pre-processing, processing, and post processing. Pre-processing is to prepare and adjust the data as our method needs. Processing is to run extended systematic clustering using the prepared and adjusted data. We also run a base line method for comparing the resut. Post processing is to evaluate the result of experiment in processing steps. General step of our method is described in Figure 1. The data and those three steps are detailly explained in next sub session.

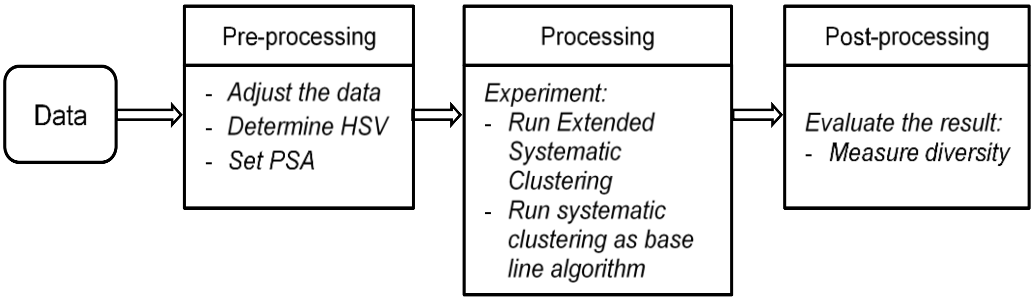

Figure 1. General step of methodology

2.1. Data

We used Adult datasets from UCI machine learning repository. Adult datasets are chosen because this data sets have characteristic as a microdata table. This data contains 32561 records and 14 attributes, then after missing values are removed it remains 30718. We adjusted data as our method needs into only 6 attributes.

2.2. Pre-processing

After downloading adult dataset from UCI machine learning repository, then we set the data to satisfy requirement for microdata. Only six relevant attributes were taken from adult dataset. Structure of adult dataset we used is shown in Table 3. The table consists of three quasi identifier attributes and three sensitive attributes.

Table 3. Structure of microdata table used

No	Attributes	Number of Unique Value	Role
1	Age	72	Quasi Identifier
2	Sex	2	Quasi Identifier
3	MaritalStaus	7	Quasi Identifier
4	Education	16	Sensitive Attribute
5	WorkClass	7	Sensitive Attribute
6	Occupation	14	Sensitive Attribute

We set primary sensitive attribute (PSA) as key in distributing sensitive value [20]. Before we determined which attribute is PSA, two terms are explained first. Sensitive values are categorized into two:
a. High-sensitive value (HSV), is values of sensitive attribute containing high sensitivity, such as HIV and cancer,
b. Low sensitive value (LSV), is values of sensitive attribute containing low sensitivity, such as flu

A sensitive attribute is set as PSA if there are more HSV than other sensitive attributes. PSA is used as a base in distributing HSV, while others are adjusted. It means, when HSV is distributed, its distribution is focused on HSV in PSA and when HSV in PSA is completely distributed, the rest HSV in other sensitive attributes is distributed.

2.3. Processing (algorithm)

In this step, a process in sensitive attribute distribution is conducted. Systematic clustering is adopted and adapted in our proposed method. Systematic clustering is a method in anonymizing table using generalization and/or suppression in k-anonymity and l-diversity [21, 22]. The method is called extended systematic clustering, since its basis method come from systematic clustering. Systematic clustering aims to create quasi identifier efficiently. This efficiency is measured by its information loss. Systematic clustering is deemed as a good method with minimum information loss because it is developed systematically.

Extended systematic clustering aims to distribute evenly HSV in PSA into each quasi identifier group in anonymized microdata table. This method is begun by sorting the tuples based on numerical quasi identifier. Then, number of quasi identifier group is determined by dividing total of records by k, where k is parameter in k-anonymity. Figure 1 shows algorithm of extended systematic clustering for multiple sensitive attributes. The process c=n/k forms cluster C based on quasi identifier, that is $C = \{C_{q1}, C_{q2}, \dots, C_{qp}\}$ where $C_{qi} \cap C_{qj} = \emptyset$, and for all $i \neq j = 1,2,\dots,p$, and $C_{qi} \in T,\ \left|C_{qi}\right| \geq k$.

From Figure 2, it is clearly seen microdata table as an input. Step 1 and 2 describe an early step after data is set and adjusted to satisfy as microdata. The data must be sort first based on its numeric quasi identifier. Then, a step to determine number of quasi identifier or class C is done. Step 3 to step 7 implements basic systematic clustering to distribute records into each class for satisfying k-anonymiity. Step 8 checks the record to distribute is HSV or not in its PSA. If in PSA is not HSV, then check in all LSVs. When this record is HSV, put it into the group/class, if it is not HSV then skip. Step 9 and 10 are looping step. Step 10 ensures that each group in microdata has variety and maximum number of HSV is *k-1*. This step describes that HSV is distributed well. The last step ensures microdata table at least in k-anonymity. If any groups do not in k-anonymity, then it should be exchanged to meet this requirement. This exchange must be performed from the closest group. It can be HSV and LSV, HSV and HSV, or LSV and LSV depends on variety condition in a group.

```
Input     : Microdata Table (T)
Output : Privacy Table with Multiple Sensitive Attributes(T')

1.  Sort all records (n) by their quasi identifiers
2.  Let C=int[n/k], k is parameter of k-anonymity
3.  Get randomly k distinct records $r_1,r_2,\dots,r_k$ from first 1 to k
4.  Let $C_{ij}$ is $j^{th}$ element of $i^{th}$ cluster
5.  For j=1 to k
6.  For i=1 to C
7.  Let $C_{i1}=(r_i+(c-1)k)^{th}$
8.  If $r_i$ not HSV then skip
9.  Next i
10. Next k
11. Check in each iteration max(HSV in QI)=k-1
12. CHECK FOR ANONYMITY LEVEL (K-ANONYMITY AND P-SENSITIVE)
```

Figure 2. Extended systematic clustering

To form a group (cluster), if n is the number of tuples, c is the number of clusters, k is parameter in k-anonymity, and R is a collection of tuples with $R = \{r_1, r_2, \dots, r_n\}$, then C can be formed $(r_i + k)^{th}, (r_i +$

$2k)^{th}, \dots, (r_i + (c-1)k)^{th}$, for one iteration, each group is filled by a record. In the next iteration $(r_j + k)^{th}, (r_j + 2k)^{th}, \dots, (r_j + (c-1)k)^{th}$ where $r_i \neq r_j$. Each time a record is entered $(r_i + k)^{th}$, starting with the second iteration, the HSV should be checked on each *C* to avoid accumulation of HSV in a group, therefore if *i* is the i^{th} iteration, then the maximum number of HSV at the time of iteration *(i=2,3, ..., n)* in each C is (*k-1*).

2.4. Post processing

Post processing is a step for evaluating the result. As mention in previous section, this work aims to investigate on how to distribute evenly HSV into each quasi identifier group. Its result should be a group with more varies of HSV and distributive HSV in a table. A group with 2 different HSV is better than a group with 2 similar HSV. A metrics called diversity metrics is used for measuring it [14, 23-25].

$$d = \sum -p_i \log_2 p_i \tag{1}$$

In (1), *d* denotes diversity values, whle p_i denotes probability of HSV lies in a quasi identifier group. Probability of HSV explains the number of HSV variety occurrence in a group. The higher value of *d*, the more vary HSV in a group. Total of *d* value indicates the vary of HSV in a microdata table. The higher value denotes the higher vary in the table and indicates a good diversity of HSV. It is shown in (2).

$$td = \frac{\sum d}{n} \tag{2}$$

In (2) shows average of diversity value in a microdata table. *td* is table diversity, $\sum d$ is total of diversity value, while n is number of tuples. The higher *td*'s value denotes the higher vary in the table.

3. RESULTS AND DISCUSSION

In this section, the results of research are represented and explained. We also discuss it as well we obtain the result. The experiment result is performed our proposed method, extended systematic clustering and compared with systematic clustering as base line method. Before we run our method, first we determined PSA for being a basis in HSV distribution. Three attributes that is decided as sensitive attributes are education, workclass, and occupation. The number of HSV in three sensitive attributes are education 7755, workclass 2541, and occupation 1370 respectively. This result leads education to PSA. Then, result of our method is compared with systematic clustering due to systematic clustering still a good method in forming group in anonymizing data. The experiment is performed in k-anonymity model with k values from 3 to 23. Result of two methods are shown in Figure 3. Detail values of our experiment are shown in Table 4.

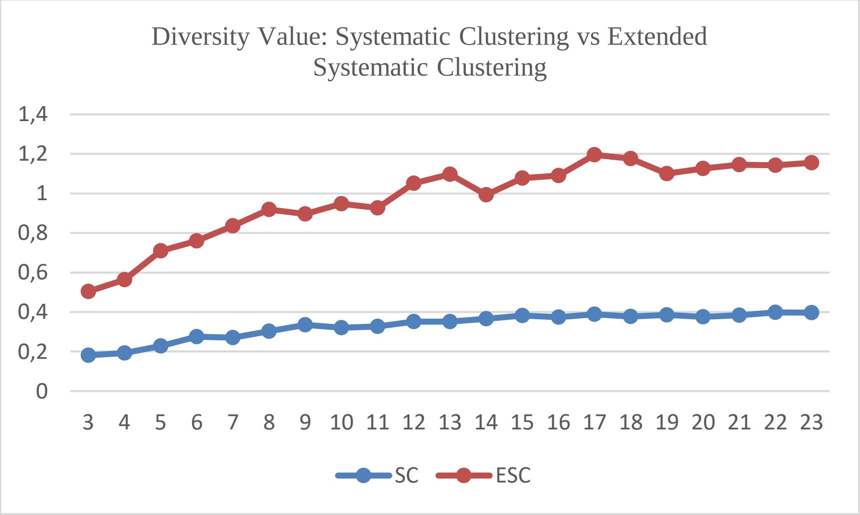

Figure 3. Result of diversity value: systematic clustering vs. extended systematic clustering

Table 4. Detail diversity values

K	3	4	5	6	7	8	9	10	11	12	13
SC	0.18	0.19	0.23	0.27	0.27	0.3	0.34	0.32	0.33	0.35	0.35
ESC	0.5	0.56	0.71	0.76	0.84	0.92	0.9	0.95	0.93	1.05	1.1
k	14	15	16	17	18	19	20	21	22	23	
SC	0.37	0.38	0.37	0.39	0.38	0.38	0.38	0/38	0.4	0.4	
ESC	0.99	1.08	1.09	1.19	1.18	1.1	1.13	1.15	1.14	1.15	

Figure 3 shows the experiment result. Comparison with systematic clustering shows that our proposed method is outperformed systematic clustering. From k=3 until k=23, in each point, extended systematic clustering always has higher values than systematic clustering. This result indicates that extended systematic clustering has more diversity in HSV and better performance. When k=3, systematic clustering has 0.181 in diversity value, while extended systematic clustering has 0.504. The increament of k value pushes diversity value to increase. Increament of extended systematic clustering is obviously seen higher than systematic clustering. The cause that systematic clustering does not achieve better performance is this method focuses in clustering a group with minimum information loss, it does not focus in distributing and varying HSV. Since its difference is higher enough, then the average of diversity value ensures the superiority of extended systematic clustering. Table 5 is explained it.

Table 5. Average of diversity value

Method	td
Extended Syst. Clustering	0.9719
Systematic clustering	0.3316

Table 5 describe the result of both method in evaluation using diversity metrics. The difference between them is more than 0.6. This high difference also indicates the higher diversity in each quasi identifier group. This higher diversity leads this table into higher model than k-anonymity. This can be satisfied l-diversity or p-sensitive, though this needs to be proofed in next investigation.

The result of experiment shows our algorithm obtain good high of HSV in a microdata table. This high diversity indicates a very good distribution of HSV. As we mrention in previous section, this investigation aims to distribute evenly HSV into each quasi identifier group, and the result shows that our purpose of this research is achieved. The distribution of sensitive values is important since it affects in protecting a sensitive value from a disclosure. Table 6 shows Disease as a sensitive attribute has only HIV values. If adversaries have a background knowledge that someone in this group with age 28, then he can directly guess 100 percent that this one has HIV. It is different from Table 7.

Table 6. A group with no diversities in sensitive attribute

Group	Age	Zipcode	Disease
2	26-30	16200	HIV
2	26-30	16200	HIV
2	26-30	16200	HIV

Table 7. A group with diversity in sensitive attribute

Group	Age	Zipcode	Disease
2	26-30	16200	Flu
2	26-30	16200	Cancer
2	26-30	16200	HIV

Table 7 exhibits a group in microdata table with diversity in HSV. There are two HSV, that is Flu and HIV. Another one sensitive value, Flu, is not an HSV but LSV because this value does not tend to be disgraceful to someone. In this circumstance, adversaries have difficulty in guessing someone’s disease since the variation of values in sensitive attributes. Adversaries can not ensure some one has cancer, HIV, or Flu.

Our method also ensures that in a quasi identifier group, the maximum of HSV is (*k*-1), while k is parameter in k-anonymity. Its purpose is to minimize probability of adversaries in guessing HSV. Result of experiment shows extended systematic clustering outpermed systematic clustering in diversity of HSV. Therefore, exended systematic clustering has lower probability for adversaries in guessing sensitive values which is linked to identities.

4. CONCLUSION

Anonymizing table has an important role when a sensitive data is published. The problem in anonymizing data using k-anonymity is still not pay attention in distributing high-sensitive value in sensitive attributes. This iworks proposed a new method/algorithm in distributing high-sensitive values in microdata with multiple sensitive attributes. The method is inspired by systematic clustering when formed quasi identifier group, therefore we called extended systematic clustering. We also introduced sensitive values as high-sensitive values and low sensitive values. Our method aims to distribute evenly HSV into each quasi identifier group. We also performed to vary HSV in a group. The experiment result shows our method is outperformed systematic clustering. Extended systematic clustering has its superiority when it is measured using diversity metrics. This result leads extended systematic clustering as a better method when our purpose is to distribute HSV evenly into each group.

Extended systematic clustering shows better performance than systematic clustering, even its difference value is high. However, it is conducted in k-anonymity model and not investigated yet in higher privacy model. Therefore, our future work is to investigate this method in p-sensitive and l-diversity. This is necessary because the higher privacy model brings data to better privacy.

ACKNOWLEDGEMENTS

Part of this research was supported by PTUPT Grants from Directorate of Research and Community Service, General Directorate of Strengthening Research and Development, Ministry of Research, Technology, and Higher Education of Republic of Indonesia.

REFERENCES

[1] Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu, “Privacy Preserving Data Publishing: A Survey of Recent Development,” *ACM Computing Surveys*; vol. 42, no. 4, pp. 1-53, 2010.

[2] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, “Microdata Protection,” *Secure Data Management in Decentralized Systems*; Springer, vol. 33, pp. 291-321, 2007.

[3] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570, 2002.

[4] S. Ni, M. Xie, Q. Qian, “Clustering Based K-anonymity Algorithm for Privacy Preservation,” *International Journal of Network Security*; vol. 19, no. 6, pp. 1062-1071, 2017.

[5] Z. El Ouazzani, H. El Bakkali, “A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k,” *Procedia Computer Science*, vol. 127, pp. 53-59, 2018.

[6] Luo Yongcheng, Le Jiajin and Wang Jian “Survey of Anonymity Techniques for Privacy Preserving”, *Proc. of CSIT 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009),* IACSIT Press, Singapore, vol. 1, pp. 248-252, 2009.

[7] Y. Pan, X. Zhu, T. Chen, “Research on Privacy Preserving on K-Anonymity,” *Journal of Software*, vol. 7, no. 7, pp. 1649-1656, 2012.

[8] F. Liu, Y. Jia and W. Han, "A New K-anonymity Algorithm towards Multiple Sensitive Attributes," *2012 IEEE 12th International Conference on Computer and Information Technology*, Chengdu, pp. 768-772, 2012.

[9] L. Zhang, J. Xuan, R. Si, R. Wang, “An Improved Algorithm of Individuation K-Anonymity for Multiple Sensitive Attributes,” *Wireless Personal Communications*, vol. 95, no. 3, pp. 2003-2020, 2017.

[10] V. S. Susan, T. Christopher, “Anatomisation with slicing: A new privacy preservation approach for multiple sensitive attributes,” *SpringerPlus*, vol. 5, no. 964, 2016.

[11] V. S. Susan, T. Christopher, “Advanced cluster-based attribute slicing: a new approach for privacy preservation,” *Proceedings of the International Conference on Soft Computing Systems*. Springer, New Delhi, pp 205-213, 2016.

[12] Y. Ye, Y. Liu, C. Wang, D. Lv, J. Feng, “Decomposition: Privacy Preservation for Multiple Sensitive Attributes,” *International Conference on Database Systems for Advanced Applications*, Springer, Berlin, Heidelberg, pp. 486-490, 2009.

[13] A. S. M. Hasan, Q. Jiang, H. Chen, S. Wang, “A New Approach to Privacy-Preserving Multiple Independent Data Publishing,” *Applied Science*, vol. 8, no. 5, 2018.

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, “*L*-diversity: Privacy beyond *k*-anonymity,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, 2007.

[15] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 106-115.

[16] C. Liu, S. Chen, S. Zhou, J. Guan, Y. Ma, “A novel privacy preserving method for data publication,” *Information Sciences*, vol. 501, pp.421-435, 2019.

[17] J. Vasa, P. Modi, “Review of Different Privacy Preserving Techniques in PPDP,” *International Journal of Engineering Trends and Technology (IJETT)*, vol. 59, no. 5, pp. 223-227, 2018.

[18] R. Saeed and A. Rauf, "Anatomization through generalization (AG): A hybrid privacy-preserving approach to prevent membership, identity and semantic similarity disclosure attacks," *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, pp. 1-7, 2018.

[19] N. Man, X. Li, K. Wang, “A Privacy Protection Model Based On K-Anonymity,” *2017 International Conference Advanced Engineering and Technology Research (AETR 2017)*. Atlantis Press, 2018.

[20] Widodo and W. C. Wibowo, "A Distributional Model of Sensitive Values on p-Sensitive in Multiple Sensitive Attributes," *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, pp. 1-5, 2018.
[21] M. E. Kabir, H. Wang, E. Bertino, “Efficient systematic clustering method for k-anonymization,” *Acta Informatica*, Springer, vol. 48, no. 1, pp. 51-66, 2011.
[22] Md Enamul Kabir, Hua Wang, Elisa Bertino, Yunxiang Chi, “Systematic clustering method for l-diversity model,” *Conferences in Research and Practice in Information Technology (CRPIT)*, Australian Computer Society Inc., Brisbane, Australia, vol. 104, pp. 18-22, 2010.
[23] F. Kohlmayer, F. Prasser, K. A. Kuhn, “The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss,” *Journal of biomedical informatics*, vol. 58, pp. 37-48, December 2015.
[24] I. Wagner, D. Eckhoff, “Technical Privacy Metrics: A Systematic Survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no 3, article 57, pp. 1-38, 2018.
[25] A. V. D. M. Kayem, C. T. Vester, C. Meinel, “Automated k- anonymization and l-diversity for shared data privacy,” *International Conference on Database and Expert Systems Applications*. Springer, Cham, pp. 105-120, 2016.

BIOGRAPHIES OF AUTHORS

Widodo is an Assistant Professor at Department of Informatics Education of Universitas Negeri Jakarta. Widodo earned a PhD in Computer Science from University of Indonesia in January 2020, MSc in Computer Science from University of Indonesia in 2004, and BSc in Information Systems from Gunadarma University in 1999. His research interests include Privacy Preserving Data Publishing, Natural Language Processing, Classification and Clustering.

Wahyu Catur Wibowo is an Associate Professor at Faculty of Computer Science, University of Indonesia. His obtained his PhD from Royal Melbourne Institute of Technology (RMIT) in 2003, MSc in Computer Science from Indiana University, and BSc in Informatics from Bandung Institute of Technology (ITB). Dr. Wahyu C. Wibowo has research interests in Information Engineering, Data Science, Natural Language Processing, Data Mining and Knowledge Management. He has published more than 30 articles in international journal and conference.

Eko Kuswardono Budiardjo is a Professor of Computer Science at Faculty of Computer Science University of Indonesia. He earned his PhD in Computer Science from University of Indonesia in 2007, MSc in Computer Science from University of New Brunswick, Canada, BSc in Electronics Engineering from Bandung Institute of Technology (ITB). Professor Budiardjo’s research interests lie in Software Engineering, Requirements Engineering, and Information Systems. His publications can be found in many journals and international conferences. He also has many experiences in Software Development.

Harry Tursulistyono Yani Achsan is a PhD Student of Computer Science at Faculty of Computer Science University Indonesia. He obtained MSc in Computer Science from University of Indonesia and BSc in Nuclear Engineering from Universitas Gadjah Mada (UGM). His research interests are Data Mining, Scientometrics, and Web Mining.