Comparative Study of Greedy K-Member Clustering and Datafly Algorithm Performance

M. T. Adithia
Department of Informatics
Parahyangan Catholic University
Bandung, Indonesia

E. E. Zacharia
Department of Informatics
Parahyangan Catholic University
Bandung, Indonesia

***Abstract* - K-Anonymity becomes one of the privacy preserving data mining method to protect privacy of data when some data mining techniques are applied on it. Good k-Anonymity methods preserve the data privacy yet can still give a high data utility, which means the data analysis and mining applied on it can still give meaningful results. In this research, two k-Anonymity algorithms, namely the Datafly and Greedy k-Member Clustering algorithm, are compared through experiments. The experiments include the computation of information loss, the measurement of execution time, the comparation of matching cluster members between the original and anonymized data, and the computation of statistical properties. Based on the results the Greedy k-Member algorithm can describe the original data better and can still represent the attribute categories of the original data. The Datafly algorithm gives lower information loss as the value of k increases, and lower execution times.**

Keywords – Privacy; k-Anonymity; Datafly, Greedy k-Member Clustering; information loss

I. INTRODUCTION

Data mining is one method to get some information from data, involving several techniques, such as clustering and classification. However, data mining techniques may reveal some private information about the people related to the data. For example, in [1], shoppers' profile can be determined based on their shopping receipt, by using a classification technique, and in [6] the profile of high-school students are obtained by crawling their social media accounts.

Because of the needs to analyze data while the protecting the privacy, privacy-preserving data mining (PPDM) [2] is introduced. PPDM is designed to preserve privacy when data mining is used to analyzed data. There two approaches to reach the PPDM goal, namely, cryptographic and anonymization approach. When the cryptographic approach is used, data is encrypted first before it is published. However, when encrypted, data is not easy to be analyzed. It needs to be decrypted first. The other approach is by using the anonymization. Hence, the data is anonymized first, before it is published. This way, data can be analyzed without needing the decryption process.

There are several PPDM methods, such as, enciphering, *k*-Anonymity, *l*-Diversity, and *t*-Closeness. In this paper, the focus is on the *k*-Anonymity methods.

K-Anonymity [16] is a method to protect data privacy by developing a database record that cannot be recognized, because there are at least *k*-1 similar records in the database. *K*-anonymity involves several algorithms, for examples, Datafly, Incognito, Optimal Lattice, and Greedy *k*-Member Clustering algorithm.

In this research, two *k*-Anonymity algorithms, namely the Datafly [15][16] and Greedy *k*-Member Clustering [4] algorithm are compared through experiments. The experiments include the computation of information loss, the measurement of execution time, the comparation of matching cluster members between the original and anonymized data, and the computation of statistical properties.

II. RELATED WORKS

To prevent private information leaking, a method created to protect privacy by modifying the original information. However, the modification may cause the decrease of the information utility. The information may become inaccurate and when it is mined, the results are also inaccurate. Privacy-preserving data mining (PPDM) [2] is designed so that the original information modification can still protect private information, yet the information utility can still be maximized. This means, PPDM protects the data privacy, but still maintains the information utility when three mining aspects, namely, association, classification, and clustering [11] are applied. The higher the information utility, the better the data mining results.

In general, PPDM methods include two techniques: erasing or masking part of the information and enciphering the information. The first technique includes *k*-anonymity, classification, clustering, association rule, distributed privacy preservation, *l*-Diverse, randomization, taxonomy tree, condensation, and cryptographic [11]. The second technique includes cryptographic techniques to prevent the information confidentiality [5], however, these techniques are computationally expensive.

Some comparative studies on the anonymization techniques have been done. In [3], the performance of three types of anonymization techniques, namely, *k*-Anonymity, *t*-Closeness, and *l*-Diversity, are compared. The research concludes that as the number of attributes grows, the information loss increases. The *t*-Closeness technique gives less information loss compared to the other two techniques.

In [14], the execution time of three anonymization algorithms are compared, namely, Datafly [16], Incognito [8], and Samarati [13] algorithm. Based on this study, as the value of *k* increases, the execution time increases.

Another comparative study is done in [7], in which the performance of the Optimal Lattice Algorithm (OLA) [9] is compared the *k*-Anonymity algorithm. Both algorithms are improved in this study, and after that the information loss of both improved algorithms are compared. The study concludes that the information loss of the *k*-Anonymity algorithm is less than the OLA algorithm.

A comparative study to learn the impact of applying *k*-Anonymity algorithms on machine learning results is given in [17]. In this study, a k-Anonymity algorithm is applied on specific data, and some machine learning techniques, such as, ANN, logistic regression, and Naïve-Bayes classification, are applied on the results. The study shows that the k-Anonymity algorithm does not decrease the classification accuracy significantly.

III. K-ANONYMITY

K-Anonymity [16] is a method to protect data privacy by developing a database record that cannot be recognized, because there are at least k-1 similar records in the database. A group of k records which are similar, is called an equivalence class. The value of k in k-Anonymity is used to measure the privacy. As the value k increases, the more difficult to recognize a record in an equivalence class, because the probability to recognize a record is $\frac{1}{k}$. However, if the value of k is too high, the data utility gets lower since the generalization is high.

Based on [3], in general, *k*-Anonymity techniques are classified into two groups, namely one pass *k*-Means [10] and *k*-Member [4]. Whereas, according to [5], there are two techniques employed in *k*-anonymity, namely data generalization and suppression. The generalization technique changes a value with other more general value. The other technique is data suppression, which deletes parts of values or database records. These two techniques cause information loss. The higher the information loss, the lower the data utility. Hence, a good *k*-Anonymity method needs to balance the tradeoff between the information loss and the data utility.

Three attribute types are used in *k*-Anonymity, namely identifier, quasi identifier, and sensitive attributes. Identifier is an attribute that can directly identify someone; usually an identifier is unique for each person, for example, an identity card number. Quasi identifier is an attribute that can indirectly identify someone, for examples, post codes, and date of birth. Sensitive attribute is private information, such as, an illness and salary. The guidance from [11] is used to decide which attributes are identifier, quasi identifier, and sensitive attributes.

In this research, two algorithms are used, namely Datafly and Greedy *k*-Member Clustering. Datafly [15] [16] is one of the first *k*-Anonymity algorithms; it uses the combination of generalization, substitution, and suppression techniques. The generalization technique replaces a value with something less specific but semantically consistent value, and the suppression technique deletes a value. To generalize or suppress data, two trees are developed: domain generalization hierarchy and value generalization hierarchy. By using the Datafly algorithm, the data is anonymized according to the level of anonymity set by the user. More detailed description of these two algorithms are given in the following subsections.

A. *Datafly*

Datafly algorithm [15] [16] was was first developed to anonymized medical data. The algorithm works by performing generalization, substitution, and suppression, without losing essential information too much.

In general, Datafly algorithm consists of the following steps:

1. Develop the frequency table, in which each row contains the number of a unique record appears
2. Check all rows of the frequency table, if all unique records appear more than or equal *k*. If yes, go to Step 5. If no, go to Step 3.
3. Check if the frequency table is ready to be suppressed. If yes, go to Step 4. If no, go to Step 6.
4. Suppress the table.
5. Return the anonymized table.
6. Generalize the table and return to Step 1.

These steps are also given in Fig. 1.

B. *Greedy k-Member Clustering*

The Greedy *k*-Member Clustering (GKMC) [4] is a *k*-Anonymity algorithm which uses the idea of clustering to minimize information loss, so the quality of the data is maintained. With this algorithm, data records which are close to each other are put in the same group; each group contains at least *k* records. By clustering similar records first, the generalization becomes minimized.

The overview of GKMC algorithm steps are given as follows:

1. Develop clusters where each cluster contains one record or tuple.
2. Add tuples to nearest clusters, until each cluster contains *k* tuples. The distance between a tuple and a cluster is obtained by computing the information loss
3. Check if the number of tuples that do not belong to any clusters is less than *k*. If yes, go to Step 4. If no, go to Step 1.
4. Anonymized the quasi-identifier of each cluster.
5. Return the anonymized table.

These steps are also described in Fig. 2.

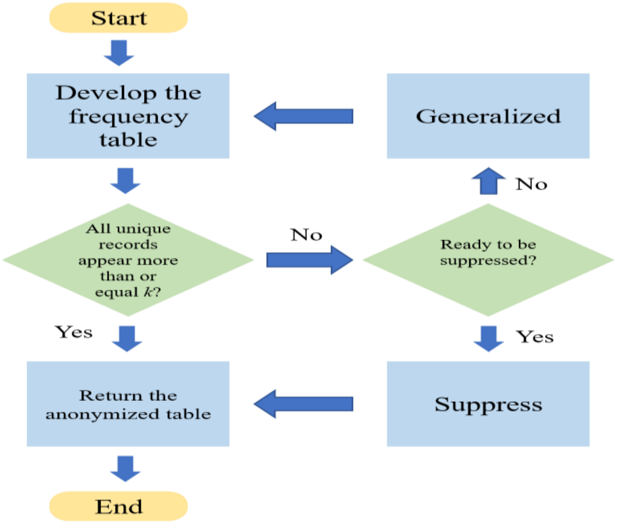
Start
Develop the frequency table
Generalized
No
All unique records appear more than or equal k?
No
Ready to be suppressed?
Yes
Yes
Return the anonymized table
Suppress
End

Fig. 1. Steps in Datafly Algorithm

IV. EXPERIMENTS

In this experiment, the ADULT database from UCI Machine Learning Repository is used. The ADULT database contains data of adult people such as age, occupation, level of education, and salary. The database consists of 48842 records and 15 attributes. For these experiments only 1000 records are used, which are chosen randomly from all records. The attributes used in the experiments are age, workclass, education, marital status, occupation, relationship, sex, race, hours-per-week, native country, and salary. The age and working hour attributes are numerical attributes, while the rests are categorical attributes. The salary attribute is used as a sensitive attribute when anonymizing data; the remaining attributes are used as quasi-identifier. The number of each attribute categories is given in Table I. For example, the number of categories of Sex is two, namely Male and Female.

TABLE I. NUMBER OF ATTRIBUTE CATEGORIES ON THE ORIGINAL DATA

Attributes	Original data
Age	69
Workclass	7
Education	16
Marital status	6
Occupation	13
Relationship	6
Race	5
Sex	2
Hours per week	59
Native country	28

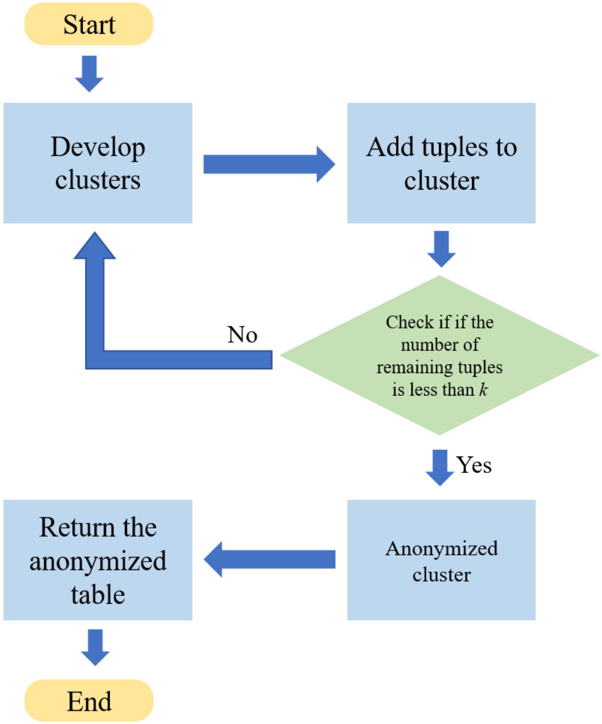
Start
Develop clusters
Add tuples to cluster
Check if if the number of remaining tuples is less than k
No
Yes
Anonymized cluster
Return the anonymized table
End

Fig. 2. Steps in the GKMC Algorithm

There are three experiments conducted. The first, second, and third experiment uses all attributes, only numerical attributes, and only categorical attributes, respectively. For each experiment, different values of k are used, namely, 20, 30, 40, 50, 60, 70, 80, 90, and 100.

On each experiment, the following things are observed:

- The information loss as the value of k increases.
- The execution time as the value of k increases. The execution time is measured to determine which algorithm is better to be applied in a big data environment.
- The statistical properties of the anonymized data, such as the number of attribute categories, and the mode of each attribute.
- The number of matching tuples between the anonymized and original data tables when the records are clustered. This experiment is done to investigate the impact of k-Anonymity algorithms on data mining technique results.

The information loss is calculated by using the formula as given in [4]. The information loss in this paper is the sum of information loss of each GKMC cluster. Since there is no cluster in the Datafly algorithm, it is assumed that a cluster is represented by a frequency row. A frequency row is a group of records having the same quasi identifier values.

To conduct the experiments, a software implementing the Datafly and GKMC algorithm is developed. The software performs the anonymization based on the algorithm chosen and the input data. The software also saves the execution time and computes the information loss. The results are saved to a file, to be analyzed further. The description of this software is given in Fig. 3.

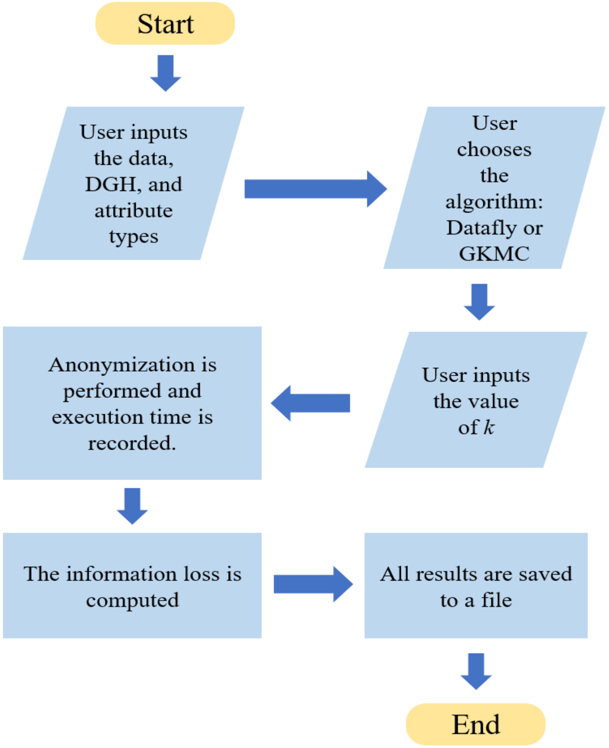
Start
User inputs the data, DGH, and attribute types
User chooses the algorithm: Datafly or GKMC
User inputs the value of k
Anonymization is performed and execution time is recorded.
The information loss is computed
All results are saved to a file
End

Fig. 3. The description of how the anonymization software works

The statistical properties are computed based on the anonymized data, by using the Microsoft Excel software. The clustering is done by using Weka, also based on the anonymized data resulted.

V. RESULTS AND ANALYSIS

The experiment results are given in the following subsections.

A. Information Loss

Table II and Table III shows some examples of the anonymization results by using the Datafly and GKMC algorithm, respectively, when all attributes are included, and with *k*=20. Table II shows that, with Datafly algorithm, more records are generalized, compared to those on Table III. This is because, with the GKMC algorithm, records are clustered first, so the generalization can be minimized.

TABLE II. THE EXAMPLES OF DATA ANONYMIZATION RESULTS BY USING THE DATAFLY ALGORITHM, WITH K=20, AND ALL ATTRIBUTES ARE USED

Age	Workclass	Education	Marital-status	Class
*****	*****	*****	Married	<= 50K
*****	*****	*****	Not-married	<= 50K
*****	*****	*****	Not-married	<= 50K
*****	*****	*****	Not-married	<= 50K
*****	*****	*****	Married	<= 50K
*****	*****	*****	Not-married	<= 50K
*****	*****	*****	Married	<= 50K
*****	*****	*****	Married	<= 50K
*****	*****	*****	Not-married	<= 50K

TABLE III. THE EXAMPLES OF DATA ANONYMIZATION RESULTS BY USING THE GREEDY K-MEMBER CLUSTERING ALGORITHM, WITH K=20, AND ALL ATTRIBUTES ARE USED

Age	Workclass	Education	Marital-status	Class
41-50	Private	HS-grad	Married-civ-spouse	<= 50K
41-50	*****	Associates	Not-married	<= 50K
11-20	*****	Higher-edu	Not-married	<= 50K
41-50	*****	Associates	Not-married	<= 50K
41-50	*****	Higher-edu	*****	<= 50K
41-50	*****	Associates	Not-married	<= 50K
41-50	*****	Associates	Married-civ-spouse	<= 50K
41-50	*****	HS-grad	Married-civ-spouse	<= 50K
11-20	*****	Associates	Never-married	<= 50K

As shown in Fig. 4, when all attributes are used, as the value of *k* increases, the information loss of the anonymization results obtained by using the Datafly algorithm decreases. The opposite happens when the GKMC algorithm is used, because as the cluster size grows, more generalization happens, hence the increase of the information loss. When the value of *k* is low, 20-30, the information loss of the GKMC algorithm is lower than those of the Datafly algorithm. The GKMC algorithm makes cluster of size *k*, so the cluster can have members that are outliers, hence causing more generalization. Other than that, the Datafly algorithm deletes records that cannot satisfy the *k*-Anonymity requirements, which yields a better information loss as the value of *k* increases.

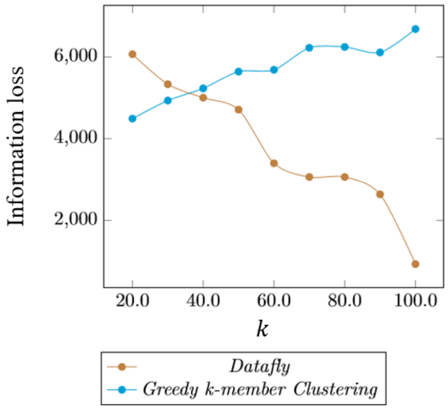

Fig. 4. Experimental Results in Which All Attributes are Used

Table IV shows some examples of the anonymization results with *k*=20, when the Datafly algorithm is used, and with only numerical attributes included.

TABLE IV. THE EXAMPLES OF DATA ANONYMIZATION RESULTS BY USING THE DATAFLY ALGORITHM, WITH K=20, AND ONLY NUMERICAL ATTRIBUTES ARE USED

Hours-per-week	Age
1-50	41-50
1-50	31-40
1-50	41-50
1-50	21-30
1-50	51-60
1-50	31-40
1-50	31-40

In general, the same information loss results obtained when only numerical attributes are used, as can be seen in Fig. 5. However, the information loss is a lot lower, compared to when all attributes are used. When the GKMC algorithm is used, the increase of the information loss is almost linear. The GKMC algorithm performs better when the value of *k* is low, 20-50. As the value of *k* increases, the Datafly algorithm performs better.

Table V shows some examples of the anonymization results with *k*=20, when the GKMC algorithm is used, and with only categorical attributes included.

TABLE V. THE EXAMPLES OF DATA ANONYMIZATION RESULTS BY USING THE GREEDY K-MEMBER CLUSTERING ALGORITHM, WITH K=20, AND ONLY CATEGORICAL ATTRIBUTES ARE USED

Workclass	Education	Occupation	Race	Class
*****	Associates	*****	People	<= 50K
Private	Higher-edu	Other-service	People	<= 50K
Government	Higher-edu	White-collar	People	<= 50K
Private	Associates	Blue-collar	People	<= 50K
Private	HS-grad	Other-service	People	<= 50K
*****	Associates	*****	White	<= 50K
*****	Associates	*****	People	<= 50K
*****	Associates	*****	People	<= 50K
*****	Associates	*****	People	<= 50K

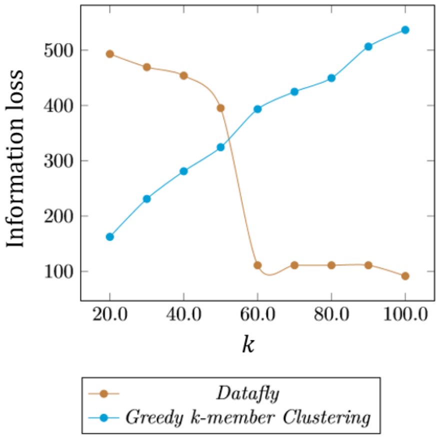

Fig. 5. Experimental Results in Which Only Numerical Attributes are Used

Fig. 6 shows that in general, also the same information loss results obtained when only categorical attributes are used. Almost the same results are shown when all attributes are used, however the information loss is about 90% lower compared to when all attributes are used.

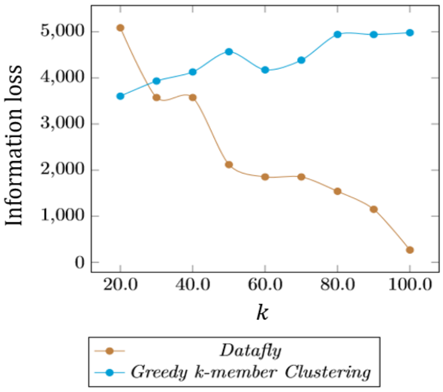

Fig. 6. Experimental Results in Which Only Categorical Attributes are Used

B. Execution Time

The execution time of both algorithms are measured. The measurement is conducted by using a computer with Intel(R) Core(TM) i7-4720HQ @ 2.6 GHz dan 4GB RAM

The execution time of both algorithm as the value of *k* increases is given in Fig. 7. The Datafly algorithm requires less time than the GKMC algorithm to perform the anonymization. This is because the complexity of the GKMC algorithm is higher than the Datafly algorithm.

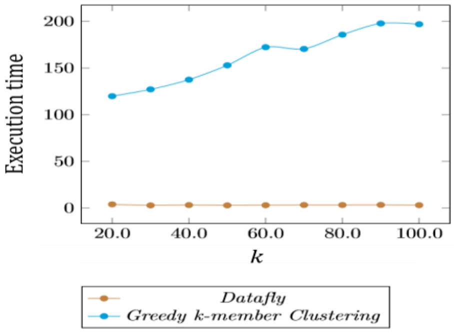

Fig. 7. The Execution Time When All Attributes are Used

The same results shown when only numerical or categorical attributes are used; the Datafly algorithm requires less time to perform the anonymization, compared to the GKMC algorithm. One factor that influence the execution time is the number of attributes included. When more attributes are included, more execution time is needed. See Fig. 8.

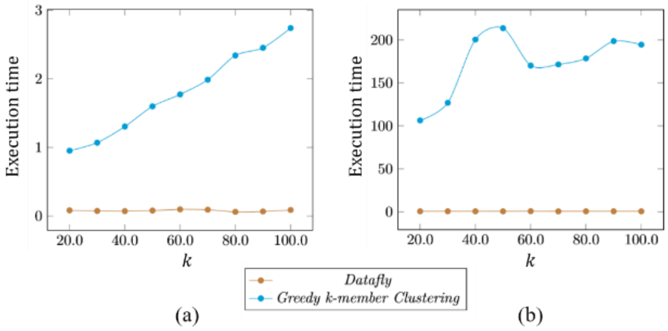

Fig. 8. Execution Time When (a) Only Numerical and (b) Only Categorical Attributes are Used

C. Statistical Properties

Based on the previous experiments regarding the information loss, the following values of *k* are used to compute some statistical properties of the anonymized data. The values of *k* used for the next experiment are *k*=40, *k*=50, and *k*=30, when all attributes, only numerical attributes, and only categorical attributes are included, respectively. The chosen values give almost the same information loss for each of the three experiments previously described.

The first computation done in this part is determining mode of each attributes of the original and anonymized data, when all attributes are included in the anonymization process. The results are given in Table VI. Based on the table, the GKMC algorithm gives matching results on five out of 11 attributes, namely workclass, marital status, relationship, hours per week, and class. Whereas the Datafly algorithm only gives four matching results. In this case, the GKMC algorithm performs better.

TABLE VI. EACH ATTRIBUTES MODE OF THE ORIGINAL DATA AND ANONYMIZED DATA

Attributes	Original data	Datafly	GKMC
Age	34	1-50	11-20
Workclass	Private	*****	Private
Education	HS-grad	*****	Associates
Marital status	Married-civ-spouse	Married	Married-civ-spouse
Occupation	Exec-managerial	*****	Blue-collar
Relationship	Husband	*****	Husband
Race	White	People	People
Sex	Male	Male	Human
Hours per week	40	1-50	1-50
Native country	United States	World	Asia
Class	<=50K	<=50K	<=50K

The number of categories for each attributes of anonymized data resulted from the Datafly and GKMC algorithm are also computed. The results are given in Table

VII, respectively. Table VII shows that the GKMC algorithm in general preserves more attributes categories compared to the Datafly algorithm. The number of categories for each attribute of the original data is given on **Error! Reference source not found.** For example, for the workclass attributes, the Datafly algorithm suppresses all the categories, while the GKMC algorithm still preserves two out of seven categories. This means, the GKMC algorithm can describe the original data better.

TABLE VII. NUMBER OF ATTRIBUTE CATEGORIES ON THE ANONYMIZED DATA

Attributes	Datafly	GKMC
Age	1	2
Workclass	0	2
Education	0	3
Marital status	2	4
Occupation	0	4
Relationship	0	5
Race	1	2
Sex	2	3
Hours per week	2	1
Native country	1	4

The second computation of mode is done to the original and anonymized data when only numerical and only categorical attributes used. The results are given in Table VIII and Table IX. When only numerical attributes are used, the Datafly works better than the GKMC algorithm; all results match the original data results. When only categorical attributes are used. The GKMC works better than the Datafly algorithm since it gives four out of nine results matched to the original data results.

TABLE VIII. ATTRIBUTES MODE OF THE ORIGINAL DATA

Attributes	Original Data
Age	34
Workclass	Private
Education	HS-grad
Marital status	Married-civ-spouse
Occupation	Exec-managerial
Relationship	Husband
Race	White
Sex	Male
Hours per week	40
Native country	United States
Class	<=50K

TABLE IX. ATTRIBUTES MODE OF ANONYMIZED DATA WHEN ONLY NUMERICAL AND CATEGORICAL DATA IS USED

Attributes	Only Numerical		Only Categorical	
	Datafly	GKMC	Datafly	GKMC
Age	31-40	41-50		
Workclass			*****	Private
Education			Associates	Higher-edu
Marital status			Married	Married-civ-spouse
Occupation			*****	Blue-collar
Relationship			Parent	Husband
Race			People	People
Sex			Male	Human
Hours per week	1-50	1-50		
Native country			America	Asia
Class			<=50K	<=50K

The number of each categories for each attribute is also computed on the original and anonymized data, when only numerical and only categorical attributes are included. See Table X and Table XI. Table XI shows that the GKMC algorithm also preserves the category better in this case, compared to the Datafly algorithm. This also means that the GKMC algorithm describes the original data better.

TABLE X. NUMBER OF ATTRIBUTE CATEGORIES ON THE ORIGINAL DATA

Attributes	Original Data
Age	69
Workclass	7
Education	16
Marital status	6
Occupation	13
Relationship	6
Race	5
Sex	2
Hours per week	59
Native country	28
Class	2

TABLE XI. NUMBER OF ATTRIBUTE CATEGORIES ON THE ANONYMIZED DATA WHEN ONLY NUMERICAL AND CATEGORICAL ATTRIBUTES IS USED

Attributes	Only Numerical		Only Categorical	
	Datafly	GKMC	Datafly	GKMC
Age	5	8		
Workclass			0	4
Education			1	5
Marital status			2	4
Occupation			0	7
Relationship			3	8
Race			1	2
Sex			2	3
Hours per week	1	4		
Native country			1	4
Class			2	2

The number of attribute categories of the anonymized data resulted when all, only numerical, and only categorical attributes are used, is also compared. The results are shown in Table XII and

Table *XIII***Error! Reference source not found.**. The tables show that, when the numerical and categorical attributes are processed separately, more attributes category can be preserved, especially when the GKMC algorithm is used for the anonymization process. For example, when all attributes are used, two out of 69 categories of age can be preserved. When only numerical attributes are used, eight out of 69 categories of age can be preserved. The overall results show that when the process is separated, about 55% of the number of attribute categories increases, compared to when all attributes are processed at once. The rest of the number of attribute categories remains the same; they do not decrease.

TABLE XII. COMPARISON OF THE NUMBER OF EACH ATTRIBUTE CATEGORIES

Attributes	All Attributes	
	Datafly	GKMC
Age	1	2
Workclass	0	2
Education	0	3
Marital status	2	4
Occupation	0	4
Relationship	0	5
Race	1	2
Sex	2	3
Hours per week	2	1
Native country	1	4
Class	2	2

TABLE XIII. COMPARISON OF THE NUMBER OF EACH ATTRIBUTE CATEGORIES WHEN ONLY NUMERICAL AND CATEGORICAL DATA IS USED

Attributes	Only Numerical		Only Categorical	
	Datafly	GKMC	Datafly	GKMC
Age	5	8		
Workclass			0	4
Education			1	5
Marital status			2	4
Occupation			0	7
Relationship			3	8
Race			1	2
Sex			2	3
Hours per week	1	4		
Native country			1	4
Class			2	2

D. Clustering Results

In this experiment, the original data and the anonymized data from both algorithms are clustered, by using *k*-Means clustering algorithm in Weka. The anonymization methods are applied to data with all attributes, only numerical attributes, and only the categorical attributes. The clusters resulted from the anonymized data are then compared to those from the original data, to check the number of matching members between the clusters of the original and anonymized data. The value of *k* is increased, to observe the impact to the results. These are all done by using another software. The software compared two clustering results based on the attributes chosen. The description on how the software works is given in Fig. 9.

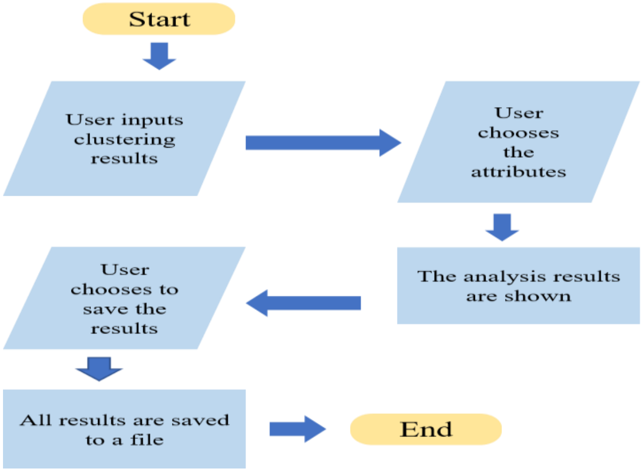
Start
User inputs clustering results
User chooses the attributes
The analysis results are shown
User chooses to save the results
All results are saved to a file
End

Fig. 9. The description on how the clustering result analysis is done

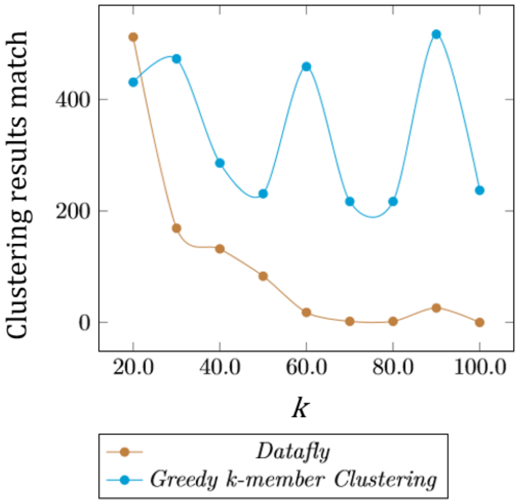

Fig. 10. Clustering Results Match When All Attributes are Used

Fig. 10 shows the results when all attributes are used in the experiment. In general, the matching members of clusters from GKMC algorithm results are higher. It means, the GKMC algorithm gives more similar results with the original data, when the anonymized data is clustered. It means, the GKMC algorithm preserves the data utility better. As the value of *k* increases, the number of matching members of the Datafly algorithm decreases. For the GKMC algorithm results, the value of *k* does not give any specific patterns.

When only numerical or categorical attributes are used, the same conclusion is also shown, that in general GKMC algorithm gives better results than the Datafly algorithms. See Fig. 11 and Fig. 12. Only, when only numerical attributes are used, the number of matching cluster members between the two algorithms does not significantly different.

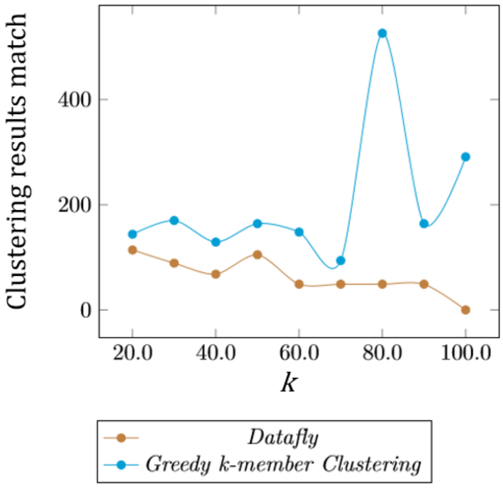

Fig. 11. Clustering Results Match When Only Numerical Attributes are Used

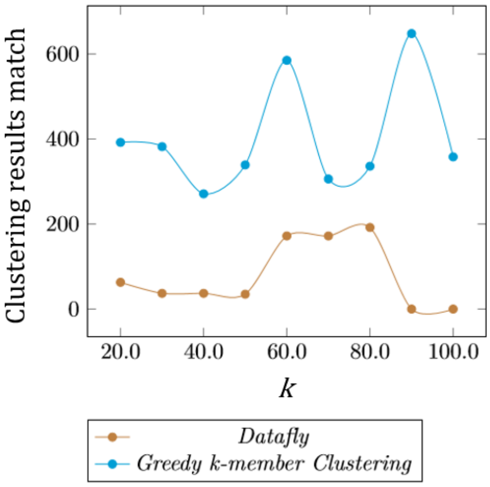

Fig. 12. Clustering Results Match When Only Categorical Attributes are Used

VI.CONCLUSIONS AND FUTURE WORKS

Based on the experiments conducted, several conclusions can be taken. The first conclusion is, based on the information loss obtained, the Datafly algorithm performs better than the GKMC algorithm. The information loss decreases as the value of *k* increases, since the Datafly algorithm deletes records that cannot satisfy the *k*-Anonymity requirements.

As the second conclusion, the Datafly algorithm has lower execution time because the complexity of the algorithm is lower than those of the GKMC algorithm. The execution time of both algorithms decreases as the number of attributes decreases.

In the next experiment, the mode, and the number of categories of each attribute are computed. Based on this experiment, the last conclusion is the GKMC algorithm describes the data better than the Datafly algorithm since more attributes of the anonymized data resulted by the algorithm, have the same mode as the original data. Furthermore, the GKMC algorithm also preserves more attribute categories, compared to the Datafly algorithm. The computation results also show that, when the anonymization process between the numerical and categorical attributes are separated, the number of attribute categories increases.

Since the results indicate that GKMC algorithm preserved more original data information than the Datafly algorithm, as the future works, a different information loss formula should be used, to investigate whether the information loss results can match the statistical properties.

The last experiment is related to the number of matching cluster members between the anonymized and the original data. Based on this experiment, the third conclusion can be drawn, namely, the GKMC algorithm performs better than the Datafly algorithm since in general the number of matching cluster members between the original and anonymized data is higher for all cases. This shows that the GKMC algorithm can describe the data better within a cluster. However, there is no pattern found how the value of *k* impacts the results.

The future work is to investigate within cluster statistical properties and compare the results with the original data. This way, it can be found out whether the clusters obtained from the anonymized data are similar to those of the original data.

The last future work is to perform classification on the original and anonymized data and compare the results. The classification accuracy will also be computed to investigate how good the classification is.

ACKNOWLEDGMENT

This research is funded by the Hibah Penelitian Dasar Unggulan Perguruan Tinggi from the Indonesian Ministry of Research, Technology, and Higher Education.

REFERENCES

[1] M.T. Adithia and E. Yudhistira, “Data Mining Based Privacy Attack Through Paper Traces”, Proceedings of the International Conference on Sustainable Engineering and Creative Computing, Bandung, Indonesia, August 20-22. 2019.

[2] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, “A Comprehensive Review on Privacy Preserving Data mining”, SpringerPlus, No. 4, 2015

[3] D. K. Arora, D. Bansal, and S. Sofat, “Comparative Analysis of Anonymization Techniques”, International Journal of Electronic and Electrical Engineering, Vol. 7, No. 8, pp. 773-778, 2014.

[4] J. W. Byun, A. Kamra, E. Bertino, and N. Li, “Efficient *k*-Anonymization Using Clustering Techniques”, Proceedings of the 12th International Conference on Database Systems for Advanced Applications, Bangkok, Thailand, April 9-12, 2007.

[5] V. Ciriani, S. D. C. Vimercati, S. Foresti, and P. Samarati, “Privacy-Preserving Data Mining”, Edited C. C. Aggarwal and P. S. Yu, Springer, New York, pp. 105–136, 2008.

[6] R. Dey, Y. Ding, and K. W. Ross,”The High-School Profiling Attack: How Online Privacy Laws Can Cctually Increase Minors’ Risk”, Proceedings of the 2015 ACM on Conference on Online Social Networks, California, USA, November 2-3, 2015.

[7] Z. F. Fei, D. Li. Feng, W. Kun, and L. Yang, “Study on Privacy Protection Algorithm Based on K-Anonymity”, Physics Procedia, vol. 33, 483-490, 2012.

[8] K. LeFevre, J. D. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-Domain *k*-Anonymity”, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, USA, June 14-16, 2005.

[9] T. Li. and N. Li, ”Optimal *k*-Anonymity with Flexible Generalization Schemes through Bottom-Up Searching”, Proceedings of the 6th IEEE International Conference on Data Mining Workshops, Hong Kong, China, pp. 518–523, 2006.

[10] J. Lin and M. Wei, “An Efficient Clustering Method for *k*-Anonymization”, Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, Nantes, France, pp. 26-35, 2008.

[11] E. McCallister, T. Grance, and K. Scarfone, “Recommendations of the National Institute of Standards and Technology”, NIST Special Publication 800-122, Gaithersburg, 2010.

[12] Sachan, D. Roy, and P. V. Arun, “Advances in Computing and Information Technology”, Edited N. Meghanathan, D. Nagamalai, and N. Chaki, Springer, New York, vol. 3, pp. 119–128, 2013.

[13] P. Samarati, “Protecting Respondents’ Identities in Microdata Release”, IEEE Transactions on Knowledge and Data Engineering, Vol. 13(6), pp. 1010–1027, 2001.

[14] M. S. Simi, K. S. Nayaki, M. S. Elayidom, “An Extensive Study on Data Anonymization Algorithms Based on *k*-Anonymity”, Proceedings of the International Conference on Materials, Alloys and Experimental Mechanics, India, July 3-4, 2017.

[15] L. Sweeney, “Achieving *k*-Anonymity Privacy Protection Using Generalization and Suppression”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10 (5), pp. 571-588, 2002.

[16] L. Sweeney, “*k*-Anonymity: a Model for Protecting Privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10 (7), pp. 557-570, 2002.

[17] H. Wimmer and L. Powell, “A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms”, International Journal of Advanced Computer Science and Applications, vol. 5, pp. 155-160, 2014.