Risk of Re-identification Based on Euclidean distance in Anonymized Data *PWSCUP2015*

Satoshi Ito[1], Hiroaki Kikuchi[1]

Meiji University Graduate School, Tokyo, 164-8525 Japan,
cs172032@meiji.ac.jp, kikn@meiji.ac.jp

Abstract. We propose a new method to re-identify anonymized data by using Euclidean distance between the original record and the anonymized record and evaluate the accuracy of the proposed method. In order to clarify performance of several anonymization methods used in the competition of the *PWSCUP2015*, we examine each of single methods and attempt to estimate the accuracy of the combination of some methods.

1 Introduction

A personal identifiable information should be anonymized before it is purchased by the other companies who plan to use for other purpose. Re-identification is an action that attacker identifies an particular individual from the anonymized data. However, it is difficult to develop strong anonymization method that is not vulnerable against any re-identification method. After the de-identified data has been defined in a Japanese regulation in 2015 and many companies plan to use anonymization method. In order to study secure anonymization method and an evaluation of re-identification risk, the data anonymization competition *PWSCUP2015*[1] was held in October, 2015.

In this paper, we propose a new Re-identification method based on Euclidean distance that is good for an evaluation risk of the method. The existing re-identification methods use the quasi identifier (QI) of data or the sensitive attribute (SA) of data [2]. Several methods have been studied [2]. Among them, in this paper, we compare the four existing re-identification methods and our re-identification method and evaluate the security and the utility. In our analysis, we pay attention to the number of attributes used in anonymization method and to the processing time because these factors are significant in evaluating re-identification method. We use R to implement our algorithm and analyze our proposed re-identification method in terms of security and performance.

In our study, we analyze the some anonymized datasets that were submitted by some researchers to the *PWSCUP2015*. The statistics of datasets were given in [1] but the quantitative analysis has not done yet. Since these datasets were processed in hidden anonymization methods, the effect of single anonymization method was unclear. To identify the effect of single anonymization method with regards to utility and safety, we use a synthesized micro dataset [4]. In our experiment, we predict the anonymization methods used in processing the anonymized

datasets that were submitted to *PWSCUP2015*. We use these datasets to evaluate our proposed re-identification methods.

2 Utility and Security

2.1 The synthesized micro dataset

In the *PWSCUP2015*, the synthesized micro dataset that was developed by the National Statistics Center (NSC), the governmental institute in statistics, was used as a target dataset for anonymization. This dataset has 8333 records and 25 attributes and containing the annual expenditure of Japanese family in 2004. The attributes 1 - 13 are discrete values, e.g., a number of family and age and the attributes 14 - 25 are continuous values, e.g., a expenses in foods and a health care cost. We treat the attributes 1 - 13 as the QI and the attributes 14 - 25 as the SAs.

2.2 Utility and Security

In the *PWSCUP2015*, many evaluation indexes [1] are defined to evaluate the utility and the safety of anonymized datasets. Table 1 shows the details of these evaluation indexes (utility: $U1$,...,$U6$, security: $S1$,$S2$,$E1$,...,$E4$).

Table 1: Details of utility and security

	Name	Detail	Target
$U1$	meanMAE	mean absolute error of values of SA between original data and anonymized data	SA
$U2$	crossMean	mean absolute error of values between cross-tablation tables	QI,SA
$U3$	crossCnt	mean absolute error of values between cross-tablation tables	QI,SA
$U4$	corMAE	mean absolute error of values between correlation coefficients of SA	SA
$U5$	IL	mean absolute error of values between some values of anonymized data	SA
$U6$	nrow	defference of the number of records of anonymized data	rows
$S1$	k-anony	the minimum value of k (k-anonymization)	QI
$S2$	k-anonyMean	the mean value of k (k-anonymization)	QI
$E1$	identify-rand	descreption in subsection 2.3	QI
$E2$	identify-sa	descreption in subsection 2.3	QI,SA
$E3$	identify-sort	descreption in subsection 2.3	SA
$E4$	identify-sa21	descreption in subsection 2.3	SA

2.3 The existing anonymization methods

In our study, we compare our re-identification methods with the four existing re-identification methods that were used to evaluate safety of anonymized datasets in the competition. Table 2 and 3 show the sample original data X and the sample anonymized data B, respectively. The data B is an instance of anonymized

X. These data of four records have three attributes as QI and two as SAs. In our study, we denote the combination of attribute values, known as the QI, by a vector of attribute values. For example, the QI and SA of the first row of X are (2,1,1) and (100,100), respectively.

We define a re-identifying rate as a fraction of the number of correctly identified records out of the number of whole records of original data.

Table 2: Sample original data X

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

Table 3: Sample anonymized data B

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	2	280	210
1	1	2	390	520

Identify-rand($E1$) This method searches the records that have same vector of QI with a target record of B from X and chooses randomly one as an identified record. For example, the first record of B has the same vector of QI (2,1,1) to the first and the second records of X, so we choose randomly one record from two records of X.

Identify-sa($E2$) This method searches the records that have same vector of QI to a target record of B from X and chooses the closest record that has the smallest distance of specific SA from these records. For example, the first record of B has the vector of QI (2,1,1) which is the same vector of the first and the second records of X, so we choose the first record as identified record because this record has the nearest value of SA1.

Identify-sort($E3$) This method sorts records of X and B in an ascending order of the sum of values in SAs and chooses the identified record with the same rank to the target record. For example, sorting records of X in an ascending order of sum of SA1+SA2 gives the order of 1,3,2,4. The sorted records of B in an ascending order of are the exactly same to the order of records of X. In this case, the re-identification of Identify-sort is completely successful, i.e., the re-identify rate is 1.0.

Identify-sa21($E4$) This method identifies records by only single attribute values of specified attribute SA without taking into account of QI. For example, the 2nd record of B is identified as the second record of X because this record has value 220 of SA1, which is the nearest to the target record with 220 in X.

3 Re-identification method by using the Euclidean distance

3.1 Identify-euc

In our proposed method, we search the records that have same vector of QI with a target record of B from X and identify record based on the Euclidean distance $D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$ between SA of X and SA of B. For example, the first record of B has vector of QI (2,1,1) which is the same to the first and the second records of X. Given the vector of SA of records 1,2 of X $a_1 = (100, 100)$ and $a_2 = (200, 400)$ and the vector $b_1 = (110, 90)$ of B, the Euclidean distances are

$$D(a_1, b_1) = 14.142 < 322.8 = D(a_2, b_1).$$

Therefore, we estimate that the first record 1 of X is identical to the first record of B, namely, these are owned by the same person.

3.2 *EUC*1 and *EUC*2

Our proposed method works in principle but has limited use. The problem happens when missing record. Suppose that B is the anonymized data that is processed only attribute of QI so that the values of attributes of QI of X completely match with B. But if the attributes of QI are altered, the values of attributes of QI of X does not match any record in B. Table 4 shows a sample anonymized data D. Note D is the data that all values of QI3 of X are modified to 1.

When we identify 3rd and 4th records of D according to our re-identification method, an issue arises because X has no record that has the vector of QI (1,1,1). For address the missing-record problem, we propose two variations of our method, say $EUC1$ and $EUC2$. Table 5 shows the details of these methods and we show algorithms and examples of these methods in Algorithm 1 and 2.

Table 4: Sample anonymized data D(or E)

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500

Table 5: $EUC1$ and $EUC2$

$EUC1$	If vectors of QI don't match, give up identifiyng the record
$EUC2$	If vectors of QI don't match, identify the record from all records of original data

Algorithm of *EUC*1

1. Input: X (original data), B (anonymized data), n (the number of records of B), q (the vector of QI that we use for re-identification), s (the vector of SA)

2. Make index f from key (the values of q) and value (the indexes of records that have value of key).
3. Find the all records of X that have the same vector of QI to the record i of B and calculate the Euclidean distances $D(a_j, b_i)$ between each of these records of X and the record i of B by values of s. Set the identified record of X that has the nearest distances to the record i of B.
4. If there is no record with the same values of attributes of QI of X to the B, set i-th the record of X as the identified i-th record of B (give up identifying record i of B).
5. Repeat the step 3 and step 4 for record 1,.., i of B and output the identified record index.

Example 1 of *EUC*1 X, B: sample data of table 2,3, q=1,2,3, s=4,5

Because of QA q=1,2,3, we make index f from attributes 1 to 3 of B. table 6 shows f in this case. When we identify the first record of B (b_1), the first and the second records of X have the same vector of QI (2,1,1). Therefore, we calculate the Euclidean distances $D(a_1, b_1)$ and $D(a_2, b_1)$ and estimate that b_1 is the identified record of X that has nearest distances to b_1. We repeat these steps for b_2, b_3, b_4 and output the identified index.

Example 2 of *EUC*1 X, D: sample data of table 2,4, q=1,2,3, s=4,5

Because of QA q=1,2,3, we make index f from attributes 1 to 3 of D. Table 7 shows f in this data. When we identify the record 1 of D (d_1), the first and the second records have the same vector of QI (2,1,1)of X (a_1, a_2). Therefore, we calculate the Euclidean distances $D(a_1, d_1)$ and $D(a_2, d_1)$ and identify the record of X that has nearest distances to d_1. We repeat these steps for d_2, d_3, d_4 and output identified index. However, X has no record that has the same vector of QI (1,1,1) to d_3 and d_4. Therefore, we compromise identifying d_3 and d_4 and estimate that a_3 and a_4 are d_3 and d_4.

Table 6: f in example 1 of $EUC1$

key	value
(2,1,1)	1,2
(1,1,2)	3,4

Table 7: f in example 2 of $EUC1$

key	value
(2,1,1)	1,2
(1,1,1)	3,4

Algorithm of *EUC*2

1. Same as step 1,...,3 of $EUC1$
2. If the values of specified attributes of QI of X, say b_j, doesn't match with any record of B, calculate the Euclidean distances between b_j and all records of X and identify the record of X that has the nearest distances to b_j.
3. Repeat the step 3 and step 4 for records of B and output identified index.

Example 1 of *EUC2* X, D: sample data of table 2,4, q=1,2,3, s=4,5

Because of QI as q=1,2,3, we make index f from attributes 1 - 3 of D. In this case, f is same as table 7. For the first record of D, d_1, the first and the second records, a_1, a_2, of X have the same vector of QI (2,1,1). Therefore, we calculate the Euclidean distances $D(a_1, d_1)$ and $D(a_2, d_1)$ to identify d_1 the record of X that has the nearest distances to d_1. We repeat these steps for d_2, d_3, d_4 and output the identified indexes. However, X has no record for vector of QI (1,1,1) as d_3, d_4. Therefore, we calculate the Euclid distances between d_3,d_4 and all records of X and determine the closest record index.

4 Evaluation

In this section, we analyze the anonymized datasets of *PWSCUP2015* in the following points.

- The effect of single anonymization method
- Evaluation of single anonymization method by *PWSCUP2015*
- Evaluation of our re-identification method (identify-euc)

4.1 The anonymized datasets of *PWSCUP2015*

We evaluate the datasets, $D_1, .., D_{12}$, submitted to *PWSCUP2015* by five teams (include top 3) and the anonymized synthesized micro dataset. Table 8 shows the details of these datasets.

4.2 The anonymized datasets by a single method

The anonymized data, $D_1, .., D_{12}$, was performedly the combining multiple anonymization methods. Therefore, the effects of single anonymization method are unknown. So, we study the effect of a single anonymization method, which is used to process $D_1, ..., D_{12}$, by using a sample data that was processed by these methods.

We have eight sample anonymized data, say $D_a, .., D_h$, produced by each single anonymization method. The original data is data that has randomly sampled 100 records from the synthesized micro dataset. Table 9 shows the statistics of $D_a, .., D_h$. We predict anonymization methods that are used to generate $D_1, .., D_{12}$ by comparing the utility and the security of $D_a, .., D_h$.

Addition of random noise to SAs In this method, we add random noise to values of SAs of the original data. The column B of table 3 is an example of anonymized X by this method. If we anonymize the original data by this method, the utilities defined for SAs $U1, U2, U4, U5$, must be reduced and the securities, related to the target SAs $E3, E4$, must be improved.

Table 8: Details of data of *PWS-CUP2015*

Name	Team	Rank
D_1, D_2	T_1	
D_3, D_4	T_2	2
D_5, D_6	T_3	
D_7, D_8, D_9	T_4	1
D_{10}, D_{11}, D_{12}	T_5	3

Table 9: Details of test data

Name	Method	Target
D_a	K-anonymization	QI
D_b	Addition noise to SA	SA
D_c	Yamaoka-anonymization	ID
D_d	Unification QI 1	QI
D_e	Unification QI 2	QI
D_f	Averaging SA	SA
D_g	Swapping QI	SA
D_h	Deleting records	records

Replacement of values of QI by the Unified Value In this method, we replace the some attribute values of QI by a particular value. The column D of table 4 is an example of the anonymized X in this method. If we anonymize the original data by this method, we can increase securities of data without decreasing utilities. But if we modify the attributes of QI that are used to quantify the utility, the utilities of data must decrease. In the *PWSCUP2015*, some utilities ($U2, U3$) are defined with attributes QI1,.., QI6. The columns D and E are attribute used as QI and not QI, respectively.

Replacement of values of SA by the Average value In this method, we compute the average of all SA attributes if the vector of QI is same. Table 10 shows the anonymized X in this method, say F. In the anonymized data in the method some utilities, $U4, U5$, decrease and some securities, $E3, E4$, increase. In this case, F is classified into two groups by means of the vector of QI.

Swapping records for QI In this method, we swap record values of SA attributes of SA among the records having the same vector of QI. Table 11 shows an example of the anonymized X in the method, say G. In group 1 of the first and the second records, the values of SA1 are swapped and the values of SA2 are swapped in the group 2. Because swapping is performed within groups, the average value and the utilities, $U2, U3$, don't change at all. If we anonymize data in the method, some utilities, such as correlation coefficient ($U4, U5$), must decrease and some safeties , $E2, E3, E4$, must increase.

Table 10: Sample anonymized data F

Group	QI1	QI2	QI3	SA1	SA2
1	2	1	1	150	250
1	2	1	1	150	250
2	1	1	2	350	350
2	1	1	2	350	350

Table 11: Sample anonymized data G

Group	QI1	QI2	QI3	SA1	SA2
1	2	1	1	200	100
1	2	1	1	100	400
2	1	1	2	300	500
2	1	1	2	400	200

Records suppression In this method, we suppress some records. If we anonymize data in the method, most utilities, $U1, U2, U3, U5, U6$, decrease and some securities, $E3, E4$, increase. Note that this method is not used by any team in the submitted data in the *PWSCUP2015*.

K-anonymization, Cheating-anonymization If we anonymize data so that k-anonymity [3], some utilities, $U2, U3$, decrease and some securities, $S1, S2, E1, E2$, increase. The cheating-anonymization [2], which is replaced by index only without modifying data, must decrease some utility $U5$ but increase some securities, $E1, .., E4$.

4.3 Expected effects

Generally, an anonymized data decrease utilities and increase security. Table 12 shows the result of our prediction of effects for the anonymization methods mentioned in the Section 4.2. We label with " positive " and " slightly " and " negative " and " - " in this table. In the column of utilities, " negative " means " significantly decrease " and " slightly " means " slightly decrease " and " - " means unchanged. In the column of $S1$ and $S2$, " positive " means " increase " and " - " means " unchanged ". In the column of $E1, ..., E4$, " positive " means resilient against this method and " slightly " means slightly resilient against this method and " negative " means vulnerable to this method.

Table 12: Expected Result

	K-ano	Add noise	Cheating	Unification 1	Unification 2	Averaging	Swapping	Deleting
$U1$	-	slightly	-	-	-	-	-	negative
$U2$	negative	slightly	-	-	negative	-	-	negative
$U3$	negative	slightly	-	-	negative	-	-	negative
$U4$	-	slightly	-	-	-	negative	negative	negative
$U5$	-	slightly	negative	-	-	negative	negative	negative
$U6$	-	-	-	-	-	-	-	negative
$S1$	positive	-	-	-	-	-	-	-
$S2$	positive	-	-	positive	positive	-	-	-
$E1$	slightly	vulnerable	positive	slightly	slightly	vulnerable	vulnerable	vulnerable
$E2$	slightly	vulnerable	positive	slightly	slightly	vulnerable	slightly	vulnerable
$E3$	vulnerable	slightly	positive	vulnerable	vulnerable	positive	positive	vulnerable
$E4$	vulnerable	slightly	positive	vulnerable	vulnerable	positive	slightly	vulnerable
EUC	slightly	vulnerable	positive	slightly	slightly	vulnerable	positive	vulnerable

4.4 The Result of evaluation

Evaluation and prediction anonymization method Table 13 shows the utilities and the securities of $D_a, ..., D_h$. In column of " original " shows evaluation of the original data. The values of $E4$ are low in this case because the

re-identification rate of SA21 is often 0 and these records couldn't be identified correctly. In the dataset of 100 records, the re-identification rate of SA21 of 76 records are 0 and the maximum value of $E4$ is 0.24.

Table 14 shows the utilities and the securities of $D_1, .., D_{12}$ and Table 15 shows the evaluation and the prediction of anonymization methods for $D_1, ..., D_{12}$. Combining some anonymization methods, the anonymized data has mixed properties of some anonymization methods. For example, D_{10} is anonymized data by k-anonymization (D_a) and replacing by the average SA (D_f) and so it has combined properties of D_a with D_f (shown in table 12,15). The sample data X, B, C, D, F, G (shown in Table 3) are not exactly same as $D_a, .., D_h$ but roughly preserving same properties of the anonymized by the same method. For example, B and D_b are anonymized by adding random noise to SAs.

Table 13: Utility and Security of test data

	Original	D_a	D_b	D_c	D_d	D_e	D_f	D_g	D_h
$U1$	0	0	46.225	0	0	0	0	0	295.731
$U2$	0	38837.9	7808.7	0	0	15135.7	104.9	209.7	1094.4
$U3$	0	5.833	0	0	0	2	0	0	0.097
$U4$	0	0	0.02	0	0	0	0	0	0.049
$U5$	0	0	0.016	0.12	0	0	0	0	0
$U6$	0	0	0	0	0	0	0	0	10
$S1$	1	3	1	1	1	1	1	1	1
$S2$	1.031	7.692	1.031	1.031	1.053	1.053	1.031	1.031	1.034
$E1$	0	0.13	0.99	0	0.07	0.11	0.94	1	1
$E2$	0	0.17	1	0	1	1	1	1	1
$E3$	0	1	0.54	0	1	1	1	0.91	0.067
$E4$	0.24	0.24	0.22	0	0.24	0.24	0.24	0.24	0.089
$EUC1$	0	0.13	1	0	0.07	0.11	1	1	1
$EUC2$	0	0.17	1	0	1	1	1	1	1

4.5 Comparative $EUC1$ and the existing methods

In the comparison of $EUC1$ and the four existing methods, $E1, .., E4$, we use the synthesized micro dataset (original data) and $D_1, ..., D_{12}$ (anonymized data).

Table 16 shows the re-identification rate of $E1, ..., E4$ and $EUC1$. The red values show the most efficient re-identification method for specific anonymized data. Our proposed method, $EUC1$, has more best efficient values (5) than others $E1, ..., E4$ (maximum 3).

4.6 Discussion

The proposed method performs better than any other method. The reason of this result is because our method evaluates more attributes than other $E1, .., E4$

Table 14: Utility and Security of data of *PWSCUP2015*

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	0	0	0	0	0	0	0	0	0	0	0	0
$U2$	58340.87	0	31572.91	31400.95	0	0	4321.75	0	0	65093.42	52975.02	46100.64
$U3$	18.6	0	1.01	0.99	0	0	1.54	0	0	7.28	2.97	1.85
$U4$	0	0.01	0	0	0.07	0.07	0.03	0.09	0.09	0.15	0.11	0.11
$U5$	0	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02
$U6$	0	0	0	0	0	0	0	0	0	0	0	0
$S1$	1	1	3	3	1	1	3	1	1	41	8	4
$S2$	2.66	1.88	4.91	4.86	36.07	36.07	36.07	13.71	13.68	106.83	42.3	31.09
$E1$	0.03	0.65	0.2	0.19	0	0	0	0	0	0.01	0.02	0.02
$E2$	0.82	0.65	0.24	0.24	0.02	0.02	0.02	0	0	0.01	0.02	0.02
$E3$	1	0	0.25	0.25	0	0	0.01	0	0	0	0	0
$E4$	0.19	0	0.05	0.05	0	0	0	0	0	0	0	0
$EUC1$	0.3	0.48	0.21	0.21	0.07	0.07	0.88	0	0	0	0.01	0.01

does. For example, $E2$, identify-sa, attacks based on a specific single attribute of SA, so the method fails to identify records correctly if the values of other attribute are significantly changed. On the other hands, $EUC1$ does identify records correctly even if more than many attributes were significantly changed.

But our method requires longer processing time than $E2$. The $E2$ is the similar algorithm as $EUC2$ and hence performs similar. We tried to compare $EUC2$ and $E2$ but we gave up because of the time limitation. In the case of D_1 (k-anonymization), $E3$ (identify-sort) has more values than $EUC1$.

We note that this result assumes the $D_1, ..., D_{12}$ were anonymized for *PWSCUP2015* so these data may be resilient to $E1, .., E4$. Therefore, $EUC1$ improves accuracy of any existing methods.

4.7 Evaluation the ability of our method

The synthesized micro dataset has 13 attributes of QI. The performance of our method depends on how many attribute to compute.

If the number of attributes of QI to use for our method, i.e., $|q|$, increase, the amount of computation decreases in the reduced cost of re-identify rate because the method spends too much in processing QI. Figure 1,...,4 show the the number of attributes of QI to use for our method, $|s|$, the processing time and, the re-identification rate in terms of number of attributes, $|q|$. For example, in Figure 1 , the more $|q|$ is, the less time to calculation is. In Figure 2, the more $|q|$ is, the higher rate of re-identify is given.

5 Conclusions

We have studied our re-identification method and showed the comparison to the existing methods by using the anonymized data submitted to the *PWSCUP2015*

Table 15: Evaluation and Prediction of data of *PWSCUP2015*

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	-	-	-	-	-	-	-	-	-	-	-	-
$U2$	negative	-	negative	negative	-	-	negative	-	-	negative	negative	negative
$U3$	negative	-	slightly	slightly	-	-	slightly	-	-	negative	negative	slightly
$U4$	-	slightly	-	slightly	slightly	slightly	slightly	slightly	slightly	negative	negative	negative
$U5$	-	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly
$U6$	-	-	-	-	-	-	-	-	-	-	-	-
$S1$	-	-	slightly	slightly	-	-	slightly	-	-	positive	positive	slightly
$S2$	-	-	slightly	slightly	slightly	positive	positive	positive	positive	positive	positive	positive
$E1$	slightly	negative	negative	negative	positive	positive	positive	positive	positive	slightly	slightly	slightly
$E2$	negative	negative	negative	negative	slightly	slightly	slightly	positive	positive	slightly	slightly	slightly
$E3$	negative	positive	negative	negative	positive	positive	positive	positive	positive	positive	positive	positive
$E4$	negative	positive	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$EUC1$	negative	negative	negative	negative	slightly	slightly	negative	positive	positive	positive	positive	positive
D_a	-	-	×	×	-	-	×	-	-	×	×	×
D_b	-	-	-	-	-	-	-	-	-	-	-	-
D_c	-	-	-	-	×	×	-	×	×	-	-	-
D_d	-	-	-	-	×	×	×	-	-	-	-	-
D_e	×	-	-	-	-	-	-	×	×	-	-	-
D_f	-	×	-	-	-	-	-	-	-	×	×	×
D_g	-	-	×	×	-	-	×	×	×	-	-	-
D_h	-	-	-	-	-	-	-	-	-	-	-	-

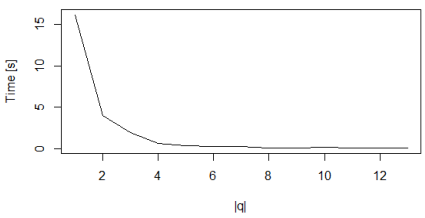

Fig. 1: Calculation time for $|q|$

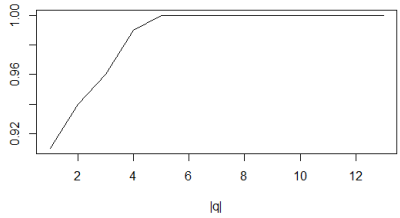

Fig. 2: Re-identification rate for $|q|$

and have analyzed properties of single method by using sampled data that are anonymized by single method.

As the result, we found that if we combine some anonymization methods, the anonymized data has combined properties of them. We also found that our proposed method performs similar to the existing method for the anonymized data that were submitted to *PWSCUP2015* and that there is no significant difference between our re-identification method identify-euc ($EUC1$) and the existing methods.

We plan to improve our method and develop new anonymization methods for future works.

Acknowledgements We thank *NSC* for synthesized micro data and thank all participants of *PWSCUP2015* for the anonymized data to study.

Table 16: The rate of re-identify of existing methods and our method

Anonymized data	id-rand	id-sa	id-sort	id-sa21	EUC1
D_1	0.033	0.824	*1.000	0.186	0.301
D_2	0.649	*0.651	0.001	0.002	0.478
D_3	0.199	0.241	*0.248	0.051	0.207
D_4	0.189	0.240	*0.253	0.045	0.211
D_5	0.000	0.022	0.000	0.000	*0.074
D_6	0.000	0.022	0.000	0.000	*0.074
D_7	0.002	0.022	0.009	0.001	*0.876
D_8	0.000	0.000	0.000	0.000	*0.001
D_9	0.000	0.000	0.000	0.000	*0.002
D_{10}	0.006	*0.007	0.000	0.000	0.004
D_{11}	*0.018	0.016	0.000	0.000	0.008
D_{12}	*0.021	*0.021	0.000	0.000	0.008
avearge	0.093	0.172	0.126	0.024	*0.187
standard deviation	0.174	0.258	0.268	0.050	0.243
best score	2	3	3	0	5

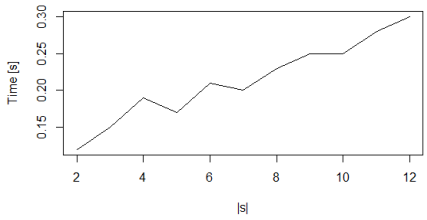

Fig. 3: Calculation time for $|s|$

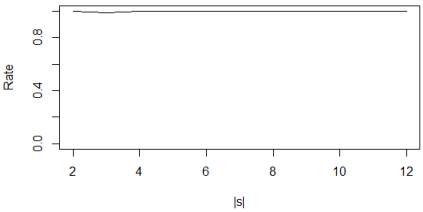

Fig. 4: Re-identification rate for $|s|$

References

1. H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "What is the Best Anonymization Method? - a Study from the Data Anonymization Competition Pwscup 2015", Data Privacy Management Security Assurance (DPM2016), LNCS 9963, pp. 230 - 237, 2016.
2. H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization," 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, pp. 1035-1042, 2016.
3. Latanya Sweeny, "k-anonymity", *Int. J. of Uncertainty, Fuzziness & Knowledge-Based System*, Vol. 10, pp. 571-588, 2002.
4. H. Akiyama, K. Yamaguchi, S. Ito, N. Hoshino, T. Goto, "Usage and Development of Educational Pseudo Micro-data –Sampled from national survey of family income and expenditure in 2004 –", Technical Report of the National Statistics Center (NSTAC), 16, pp. 1-43, 2012 (in Japanese).