INVITED PAPER *Special Section on Cryptography and Information Security*

Study on Record Linkage of Anonymizied Data*

Hiroaki KIKUCHI[†a)], ***Member***, **Takayasu YAMAGUCHI**[††], ***Nonmember***, **Koki HAMADA**[†††], ***Member***, **Yuji YAMAOKA**[††††], **Hidenobu OGURI**[†††††], ***Nonmembers***, ***and*** **Jun SAKUMA**[††††††], ***Member***

SUMMARY Data anonymization is required before a big-data business can run effectively without compromising the privacy of personal information it uses. It is not trivial to choose the best algorithm to anonymize some given data securely for a given purpose. In accurately assessing the risk of data being compromised, there needs to be a balance between utility and security. Therefore, using common pseudo microdata, we propose a competition for the best anonymization and re-identification algorithm. The paper reported the result of the competition and the analysis on the effective of anonymization technique. The competition result reveals that there is a tradeoff between utility and security, and 20.9% records were re-identified in average.

key words: *data privacy, anonymization, re-identification risk, big data*

1. Introduction

The volume of digital data is growing exponentially. Many business organizations collect our personal data with the aims of sharing this data with partners and using data-mining algorithms to extract useful knowledge related to the behavior of customers and their preferences for goods. However, there is a risk that an individual could be identified from the big-data collection.

1.1 Attribute Estimation

Gong and Liu [5] proposed attribute inference attack integrating the social friendship structures of the target users and the user behaviors, called social-behavior-attribute network model, and demonstrated their proposed attack using a large-scale dataset collected from Google+ and Google play. They claimed that the inferring attribute becomes qualitatively different if attackers use both social structure and user behaviors. They demonstrated that their proposed attack correctly infer the cities a user lived in for 57% of the users and the success rate can be improved to over 90% if the attacker selectively attacks a half of the users.

Attribute inference could be real privacy threat in big-data business and the latest result implied that protecting the attribute estimation attempt is quite hard for attackers who motivated with learning partial knowledge of their target users. First, the attack can be possible using only publicly available information. The quantity of the open data increases year by year and the attacker would have access many resources. Second, the state-of-art technologies allows us to estimate more accurately. Third, the attribute inference is authorized in business perspective and the estimated data is not classified as personal data.

1.2 Record Linkage

Record linkage across over multiple databases is greatly required for medical applications including disease registries, electronic medical records, cohort suited and case control studies. However, record linkage using personally identifying information (PII) was already not done any more because of the reasons; sharing of PII is prohibited by regulations, potential privacy risks by a third party with knowledge, and a person's PII at two different databases will not be exactly the same.

Dewri, Ong, and Thurimella [6] addressed the record linkage issues by precomputing repetitive cryptographic operations of one-way hash function with commutative property, in conjunction with similarities of the sets of bigrams. They demonstrated their proposed method using dataset from North Carolina Voters Registration database obtained in 2012 consisting of 7,088,370 individuals. With parallelizing computations over a small number of hardware threads, and truncating large encryption outputs, they successfully performed record matching on datasets as large as 100,000 records in less than 10 minutes.

1.3 Anonymization Competition

To prevent data from being re-identified, many anonymization algorithms have been proposed, aiming to retain the utility of data that have been *anonymized*. Anonymization algorithms employ various operations, including *suppression* of attributes or records, *generalization* of values, re-

Manuscript received July 10, 2017.

[†]The author is with School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo, 164-8525 Japan.

[††]The author is with Research Laboratories, NTT DOCOMO, Inc., Yokosuka-shi, 239-8536 Japan.

[†††]The author is with NTT Secure Platform Laboratories, Musashino-shi, 180-8585 Japan.

[††††]The author is with FUJITSU LABORATORIES LTD., Kawasaki-shi, 211-8588 Japan.

[†††††]The author is with NIFTY Corporation, Tokyo, 169-8333 Japan.

[††††††]The author is with University of Tsukuba, Tsukuba-shi, 305-8577 Japan.

*A preliminary version of this paper was presented at Data Privacy Management and Security Assurance (DPM 2016).

a) E-mail: kikn@meiji.ac.jp

DOI: 10.1587/transfun.E101.A.19

placing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and micro-aggregation [2], [11].

It is not trivial to anonymize data so that the risk of re-identification is eliminated because there is a tradeoff between utility and security. If we alter the data sufficiently, the data can be secure against re-identification attempt. However, too much anonymized data loses the accuracy as well. Hence, we need to carefully determine the best algorithm for data anonymization that should be secure against re-identification risk without loss of data utility.

To address the issues in anonymization, we propose a data anonymization and re-identification competition using the common dataset [1] in 2015. We adopt the educational dataset "pseudo microdata", which was synthesized by a governmental agency, the National Statistics Center (NSTAC). This is based on real statistics about income and expenditure for Japanese households. To simplify our analysis, we assume there is a maximum-knowledge adversary who can access the original dataset before anonymization. This assumption makes our competition clear and simple. We have defined some utility measures, combined with some security measures in [1].

In this paper, we report the results of our competition from utility and security perspectives and examine the submitted anonymized data to find the best strategy for making data secure against re-identification. Our analysis includes the relationship between the utility measures and the effect of k-anonymization. We also found a simple permutation makes the data secure against re-identification without losing the accuracy of data. However, it is a sort of cheating and we need to prevent data processer from performing this type of permutation. The results of the competition provide useful knowledge related to data anonymization as well as evaluation of re-identification risk.

The remainder of the paper is organized as follows. In Sect. 2, we define fundamental definitions for data anonymization and re-identification in general. The set of attributes of the dataset is partitioned into two disjoint subsets, referred to as the quasi identifier (QI) and sensitive attribute (SA). We also introduce the idea of a maximum-knowledge adversary. In Sect. 3, we show the result of the competition and the distribution of utility measures. We also evaluate the re-identification techniques used in the competition, which provide the baseline of fundamental procedures. We conclude our study in Sect. 4.

2. Anonymization

2.1 Outline of the Competition

In October 21st, 2015, we held the first competition for data anonymization and re-identification, PWSCUP (Privacy Workshop CUP) 2015 "Ice and Fire"[†], in Nagasaki, Japan.

[†]The "ice" and "fire" mean anonymization and re-identification attempts, respectively.

Table 1 Competition data.

Phase	Preliminary	Final
Date	Aug. 24–Oct. 9, 2015	Oct. 21, 2015
De-identification	4 weeks	20 min
Re-identifying	2 weeks	60 min
Venue	(Online)	Nagasaki Brick Hall
Teams	17 (81 people)	12 (20 people)

It was organized by the SIG of Computer Security (CSEC) of the Information Processing Society Japan (IPSJ).

A total of 17 teams (of more than 80 people) participated in the competition. Most participants were privacy-technologies researchers from universities and industrial laboratories. Table 1 provides information about the competition.

2.2 Fundamental Definition

A *dataset* $\boldsymbol{X}$ consists of n *records*, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, of the form $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^m)$, defined in terms of m attributes, $X^1, \ldots, X^m$. Let I^X be a *record index sequence* for database X. For example, $I^X = (1, \ldots, n)$ is identity. We treat a dataset as containing *personal data* if some attributes are related to personal information such as name or postal address and are expressive enough to identify a particular subject.

Let $R(X)$ be a range of attribute X. Attributes have several types of value, such as continuous, categorical, or discrete. A continuous attribute such as payment has a range of integer values, or rational numbers. A categorical attribute has a finite range of symbols. Attributes can be classified into two classes: *static* attributes, such as name, sex, marital status, and postal code; and *dynamic* attributes, such as location, money balance, blood pressure, heart rate, and disease name. A set of attributes is known as *QID* if they link the records generated by a single user. Various properties to reduce the risk of re-identification from an anonymized dataset have been studied such as k-anonymity [18], [19] and ℓ-diversity [20]. The dynamic attributes often are referred to as *sensitive attributes* (SAs) because they may contain critical information that the user may wish to hide.

2.3 Anonymization

Many anonymization algorithms have been proposed to preserve privacy, while aiming to retain the utility of the data that have been *anonymized*. That is, the data are made less specific so that a particular individual cannot be identified. Anonymization algorithms employ various operations, including *suppression* of attributes or records, *generalization* of values, replacing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and micro-aggregation [2], [11].

In this paper, we use an anonymization as a general process possibly implemented by multiple algorithms, rather than by a particular algorithm.

Definition 2.1: Let $\boldsymbol{Y}$ be an *anonymized dataset* generated from a dataset $\boldsymbol{X}$. The anonymized dataset $\boldsymbol{Y}$ comprises n'

$(n' \leq n)$ records, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n'}$ of the tuple of m' $(m' \leq m)$ attributes chosen from $\{X^1, \ldots, X^m\}$ of $\boldsymbol{X}$. A record $\boldsymbol{y}_j = (y_j^1, \ldots, y_j^{m'})$ of $\boldsymbol{Y}$ is de-identified from the corresponding record $\boldsymbol{x}_i$ of $\boldsymbol{X}$ such that $j = \pi(i)$, where π is a mapping

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n'\}.$$

The anonymizing processes such as sampling, swapping, or shuffling are represented by means of the *anonymized record index sequence* $I^Y = (i_1^Y, \ldots, i_{n'}^Y) = (\pi^{-1}(1), \ldots, \pi^{-1}(n'))$.

Note that π is not necessarily a surjection of I^X because some anonymization methods such as "top coding" suppresses some (unsafe) records that are likely to be identified. Therefore, the range of π is generally smaller than that of the domain.

Figure 1 illustrates a sample anonymized dataset. The figure shows how an anonymized process is specified by means of anonymized record index sequences I^Y. In the example, $I^Y = (4, 1, 2)$ where $\pi(4) = 1, \pi(1) = 2$, and the third record $\boldsymbol{x}_3$ has been dropped for some reason.

2.4 Re-Identification

A *re-identification* is a process that attempts to identify the record subject $\boldsymbol{x}_i$ from the anonymized record $\boldsymbol{y}_j$ based on some features of the original record. However, the term "re-identification" is ambiguous because some possible meanings have to be interpreted in context.

Consider the examples of privacy threads in the data anonymization in Fig. 2. The dataset $\boldsymbol{X}$ of two records with four attributes, "name," "year," "good," "payment," are anonymized as the lower table $\boldsymbol{Y}$, where names are replaced by pseudonyms, values are rounded, and the values "coffee" and "tea" are unified as a general "beverage." In the example, re-identification could happen via the following related events;

1. re-identifying the record subject "Kikuchi" from estimated record values,
2. linking two records by the pseudonym 1055 being the same owner,
3. estimating hidden attribute values,
4. exposing contact information about the subject and receiving an advertised message,
5. matching to another data resource.

Among these possible and related events, this paper adopts a strict definition of re-identification:

Definition 2.2 (Re-identification): Given an anonymized dataset $\boldsymbol{Y}$, an adversary estimates the record index sequence $I^E = (i_1^E, \ldots, i_{n'}^E) \cong I^Y$ by employing an algorithm E.

For the example in Fig. 1, the record index is estimated as $I^E = (3, 1, 2)$, for which the elements 1 and 2 are correctly estimated, but the first (3) is a false estimate.

2.5 The QI and SA Subsets

Attributes in a dataset $\boldsymbol{X}$ are partitioned into three subsets:

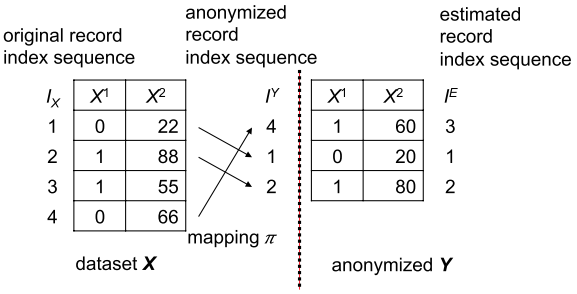


Fig. 1 Original, anonymized and estimated record index sequences.

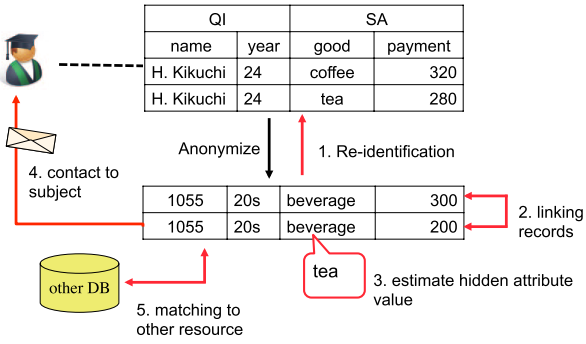


Fig. 2 Privacy threads and re-identification.

Table 2 Statistics for the NSTAC pseudo microdata [10].

Dataset	# of records	# of QIs	# of SAs	
			Expenditure	Income
	n	m		
Full	32,027	14	149	34
Simple	8,333	14	11	N/A

(1) a direct identifier such as a name and social security number (2) a QI subset comprising a combination of attributes, such as sex, age and zip code, identifies unique individual, and (3) other attributes that contains sensitive attributes (SA) such as disease and religion.

2.6 Common Dataset

The NSTAC "pseudo microdata" is a dataset of family income and expenditure in Japan, which was synthesized in 2012 by NSTAC for educational purposes in schools [10]. The dataset comprises 59,400 records, each representing the income and expenditure for a family (including 5,002 single-person households), in 2004. The statistical features of the real data were preserved in the NSTAC pseudo microdata under the assumption that the occurrence of values in all continuous attributes are (logarithmic) normally distributed.

Table 2 shows some fundamental statistics for the NSTAC pseudo microdata. The simple dataset comprises $n = 8,333$ records, and is a subset of the full dataset, being limited to four-person household with at least one is employed. The dataset involves $m = 25$ attributes, where the first 14 attributes describe features of the household, such as numbers of people, the number employed, and the age and

Table 3 Measures for utility ($U_1, \ldots, U_6$) and for security S_1, S_2, $E_1, \ldots, E_4$.

No.	measures	meaning	target
U_1	meanMAE	Error of means for all SAs	SA
U_2	crossMean	Error of mean of some SAs for some QIs	QI
U_3	crossCnt	Error of record counts for some QIs	QI
U_4	corMAE	Error of correlations of all pair of SAs	SA
U_5	IL	Error of all values of all records	QI, SA
U_6	nrow	Number of records	N/A
S_1	k-anony	k-anonymity (minimum k)	QI
S_2	k-anonyMean	k-anonymity (mean k)	QI
E_1	IdRand	re-id by a random guess in a subset of records with QIs	Qi
E_2	IdSA	re-id by searching in a subset of records with QIs	QI, SA
E_3	Sort	re-id by sorting for sum of values of SAs	SA
E_4	SA21	re-id by searching all records for 21th SA	SA

sex breakdown. These are classified as QIs. The remaining 11 attributes are treated as SAs that indicate monthly subtotals of expenditure for items such as foods, accommodation, lightning and heating, clothes, medical treatment, and travel.

2.7 Players and Rules

The following types of players compete to demonstrate their skills.

1. An anonymizing player ("ice")
   performs anonymization of given pseudo microdata $\boldsymbol{X}$, while aiming to make re-identification impossible, and submits the anonymized data $\boldsymbol{Y}$ and the corresponding record index I^Y to the judge. Note that $\boldsymbol{Y}$ is available to everyone, while I^Y is sent only to the judge.
2. An adversary ("fire"),
   attempts to estimate the most likely mapping between the original data $\boldsymbol{X}$ and the anonymized data $\boldsymbol{Y}$, submitting the estimated record index I^E for algorithm E to the judge.
3. The judge,
   receives the anonymized data $\boldsymbol{Y}$ and publishes measures in terms of utility of the data and the security against risk of re-identification using the sample algorithms. The judge evaluates the re-identification ratios of the estimated record index I^E for each adversary and declares the winner.

Our competition involves the following tasks.

1. Data anonymization.
   Given an original $\boldsymbol{X}$ (of NSTAT pseudo microdata), perform data anonymization and submit the anonymized data $\boldsymbol{Y}$ and the corresponding record index I^Y. A player is allowed to submit at most three different anonymized datasets for the original data $\boldsymbol{X}$. The player whose anonymized data is the most useful and the most secure against any re-identification attacks will be the winner for this task.
2. Re-identification.
   Given some anonymized data $\boldsymbol{Y}$s, estimate the process of data anonymization and submit the estimated record index I^E. An adversary is allowed to submit only one estimated record index I^E for each $\boldsymbol{Y}$. The adversary who performs the most accurate re-identification with the highest ratio[†] will be the winner for this task.

2.8 Security: *re-id*

Let $\boldsymbol{Y}$ and $I^Y = (i_1^Y, \ldots, i_{n'}^Y)$ be some anonymized data and their corresponding anonymized record index sequence. Let $I^E = (i_1^E, \ldots, i_{n'}^E)$ be the estimated record index sequence of $\boldsymbol{Y}$ using re-identification algorithm E. The *re-identification ratio* of E is defined as

$$\text{re-id}^E(I^Y, I^E) = \frac{\left|\{j \in \{1, \ldots, n'\} | i_j^Y = i_j^E\}\right|}{n'}.$$

Note that re-id$(I^Y, I^Y) = 1.0$.

We denote the re-identification ratios for the three algorithms, Sort, IdRand, and IdSA by re-id^{Sort}, re-id^{IdRand}, and re-id^{IdSA}.

2.9 Definition of Utility and Security

Table 3 is the list of measures in terms of utility ($U_1 \ldots, U_5$) and security ($S_1, S_2, E_1, \ldots, E_4$). We show the general meaning of these measures with the target attribute (SA or QI).

2.10 The Cheating Permutation

A malicious anonymizing player could cheat in the competition as follows.

The player lets $\boldsymbol{Y} = \boldsymbol{X}$ but submits I^Y as some permutation has been done. For example, $\boldsymbol{y}_i = \boldsymbol{x}_i$ for all $i = 1, \ldots, n$, and

$$I^Y = (n, 1, 2, \ldots, n-1).$$

could be submitted.

This cheating does not modify any record in $\boldsymbol{Y}$ and the utility is therefore maximized. In addition, the re-identification ratio could be minimized because most algorithms estimate the permutation of the record as $I^E = (1, 2, \ldots, n)$, which results in re-id$(E) = 0$. However, this

[†]The adversary can be an anonymizing player, but Re-identification of data anonymized by that person is not counted in the ratio.

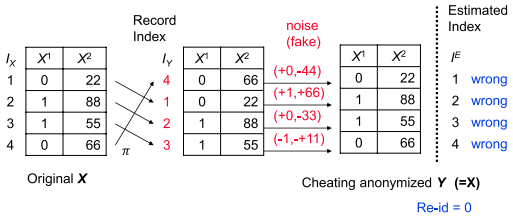


Fig. 3 The cheating permutation.

data anonymization is clearly meaningless in practice.

Example of the cheating is illustrated in Fig. 3, where the anonymized record $\boldsymbol{Y} = \boldsymbol{X}$ but the record index $|_Y$ is inconsistent with $\boldsymbol{Y}$. That results in making all records to be identified wrongly.

To prevent players from cheating in this way, we introduce the following detection algorithm on the submitted I^Y and reduce the security correspondingly.

(Re-identification AYA) Let $\boldsymbol{Y}$ and I^Y be the anonymized data of some dataset $\boldsymbol{X}$ and the record index sequence $(i_1^Y, \ldots, i_{n'}^Y)$. With an estimated record index sequence I^E in an arbitrary algorithm, the re-identification algorithm AYA outputs

$$i_i^{AYA} = \begin{cases} i_{i^Y} & \text{if } |X_{i_{i^E}} - Y_i| < |X_{i_{i^Y}} - Y_i|, \\ i^E & \text{otherwise} \end{cases}$$

3. Competition Result

3.1 The Evaluation System

We developed a system for our competition that provides a web interface for downloading the original data, submitting and withdrawing anonymized data, and submitting the estimated record index sequence.

Table 4 shows the technical specifications of our implementation.

3.2 Competition Result

Table 5 shows the top 10 anonymized data for the competition involving the measures $U_1, \ldots, U_6$ for utility, S_1, S_2 for k-anonymity, and $E_1, \ldots, E_4, E_{AYA}$ for re-identification ratios. For reference purpose, we show the utility and the security scores for the original dataset $\boldsymbol{X}$ in Table 6. As we expected, the utilities are almost zero (no error) and the re-id ratios are close to 1.0 (no security).

The anonymized data are ranked by the sum of utilities and the security against all re-identification techniques. For example, the 5th data preserves higher utilities, while the security is not so good, i.e., most records were re-identified with re-id of 0.92. On the other hand, some anonymized data e.g. 4th, 6th and 7th focused on its security rather than utilities. The first and the second data, submitted by the same team (02), balanced both scores very well and succeeded that most records were not identified by $E_1, \ldots, E_4$ as almost zero.

Note that score S_1 indicates if the data is altered so that

Table 4 Competition system.

Sever	Nifty Cloud (R)
CPU	Intel(R) Xeon(R) 3.00GHz
Memory	4 GB
R	R version 3.2.0
Java	Java(TM) SRE 1.8.0_45
Ruby	ruby 2.2.2p95
Python	Python 2.7.10

k-anonymity is satisfied. For instance, 4th, 6th, and 10th data guarantee that there are at lest $k = 3$ records for any combinations of QIs. The 7th satisfies $k = 5$ degree of anonymity.

3.3 Evaluation of Utility Measures

We show the distribution of utility measures $U_1, \ldots, U_6$ for anonymized data (ordered in the measures) in Figs.4, 5, 6, 7, and 8. In the figure, utility measures are normalized as mean of 0 and variance of 1.0.

Utility measures vary widely. We observed that some utilities U_1, U_2, U_3 (meanMAE, cross) are skewed at the bottom, while utility U_5 (IL) is increasing monotonically with rank. Since U_5 is defined for all values in dataset, it is interpreted as the total degree of altering. Hence, we noticed that in the dataset records were very carefully altered so that the utility measures $(U_1, \ldots, U_4)$ were lost.

Several strategies for anonymization were taken. From the observation of Fig. 9, where utility measures U_1, U_3, U_5 are plotted in the order of U_5, we find four large peaks. At the first one (very left), we think that only QI attributes were altered without chaining any SA because measure U_3 of QI is high. On the contrary, the third peak (around id 16) shows the evidence that SAs were altered well without changing any QI because measures U_1 of SA is high. In this way, we see that a variety of data anonymization strategies were attempted in the competition.

3.4 Tradeoff between Utility and Security

We find a tradeoff between utility and security for the set of anonymized data in Fig. 10, where the 24 submitted data are scattered over the space of the maximum re-id ratio (Y axis) and the representative utility measure U_5 (X axis).

The security against re-identification is maximized in the cost of utility loss. For example, the cluster of anonymized data plotted at the right bottom is of high security and low utility. On the other hand, the left-top cluster of anonymized data preserves the property of the original data $\boldsymbol{X}$ accurately but they are vulnerable against any re-identification attempt. We indicate the top 5 anonymized data with their ranks, which are plotted slightly lower than the tradeoff between utility and security. The secret techniques might be applied to the top data in order to optimize processing for both security and utility perspectives.

Table 5 Utilities and Re-id scores of the top 10 anonymized data $\boldsymbol{Y}$.

rank	team ID	U_1	U_2	U_3	U_4	U_5	U_6	S_1	S_2	E_1	E_2	E_3	E_4	E_{AYA}	Max E_i
1	02	0.00	0.00	0.00	0.09	0.01	0.00	1.00	13.71	0.00	0.00	0.00	0.00	0.03	0.03
2	02	0.00	0.00	0.00	0.09	0.01	0.00	1.00	13.68	0.00	0.00	0.00	0.00	0.08	0.08
3	01	0.00	0.00	0.00	0.07	0.02	0.00	1.00	36.07	0.00	0.02	0.00	0.00	0.36	0.36
4	02	0.00	4321.75	1.54	0.03	0.01	0.00	3.00	36.07	0.00	0.02	0.01	0.00	0.21	0.30
5	10	0.00	0.00	0.00	0.00	0.03	0.00	1.00	36.07	0.00	0.08	0.08	0.01	0.92	0.92
6	15	0.00	31400.95	0.99	0.00	0.02	0.00	3.00	4.86	0.19	0.24	0.25	0.05	0.57	0.57
7	07	0.00	46944.41	2.16	0.00	0.02	0.00	5.00	89.60	0.00	0.00	0.00	0.00	0.63	0.63
8	10	0.00	0.00	0.00	0.00	0.03	0.00	1.00	36.07	0.00	0.07	0.07	0.01	0.93	0.93
9	10	0.00	0.00	0.00	0.00	0.03	0.00	1.00	36.07	0.00	0.07	0.07	0.01	0.93	0.93
10	15	0.00	31572.91	1.01	0.00	0.02	0.00	3.00	4.91	0.20	0.24	0.25	0.05	0.63	0.63

Table 6 Scores of the original dataset $\boldsymbol{X}$.

team ID	U_1	U_2	U_3	U_4	U_5	U_6	S_1	S_2	E_1	E_2	E_3	E_4	Max E_i
$\boldsymbol{X}$	0.00	0.00	0.00	0.09	0.01	0.00	1.00	1.88	0.65	1.0	1.0	1.0	1.0

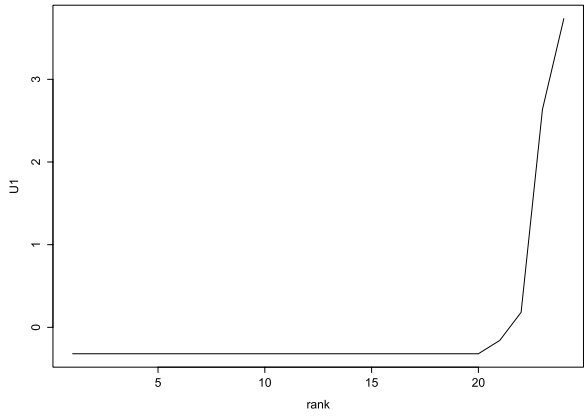


Fig. 4 Distribution of utility U_1.

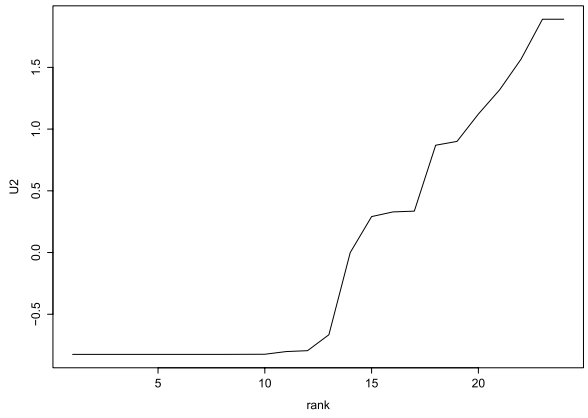


Fig. 5 Distribution of utility U_2.

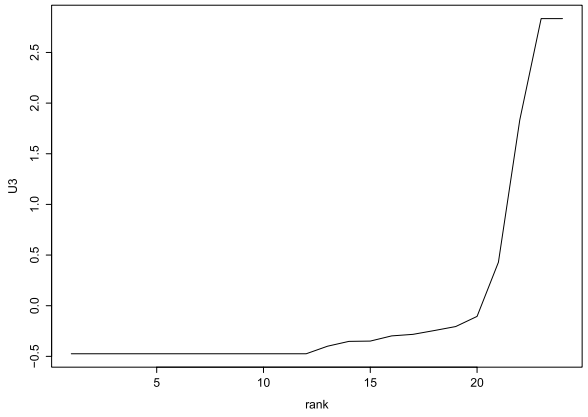


Fig. 6 Distribution of utility U_3.

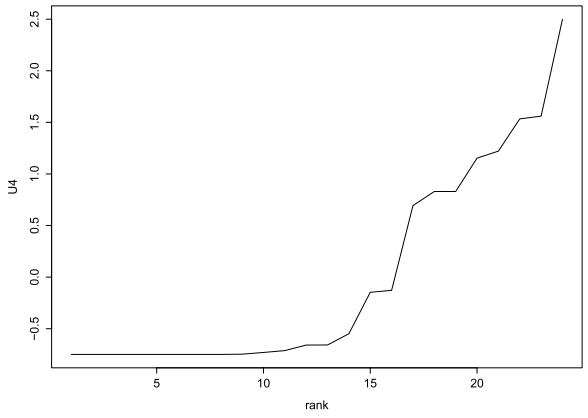


Fig. 7 Distribution of utility U_4.

3.5 Evaluation of Re-Identifications Technique

Table 7 shows the ranking of 13 teams for re-identification measure in the order of the total number of successfully re-identified records. In the competition, the players were allowed to submit the estimated record index once per anonymized data and they did not have to estimate all data. Hence, some team carefully chose their victim data that looked easy to be re-identified. For example, the 5th ranked team achieved the highest re-id ratio, 51.4%, by performing only 8 anonymized data. The first ranked team tried to re-identify as many data as they could and won the highest score of 51,628 out of 174,993 records. The re-id ratio of 29.5% is smaller than the 5th team.

Figure 11 shows the scatter plot of the scores of 13 teams. Note that the plot are scattered over the space of the re-identification probabilities (ratio) vertically and the total

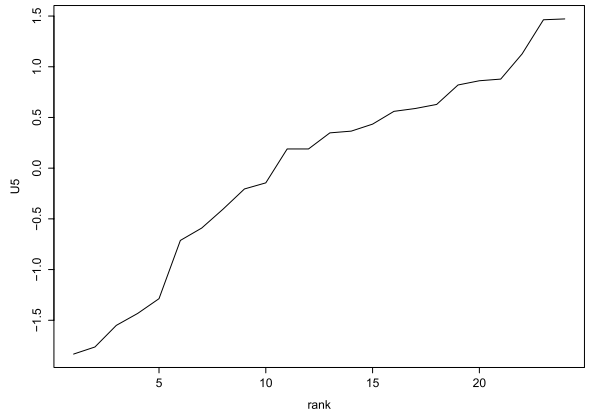


Fig. 8 Distribution of utility U_5.

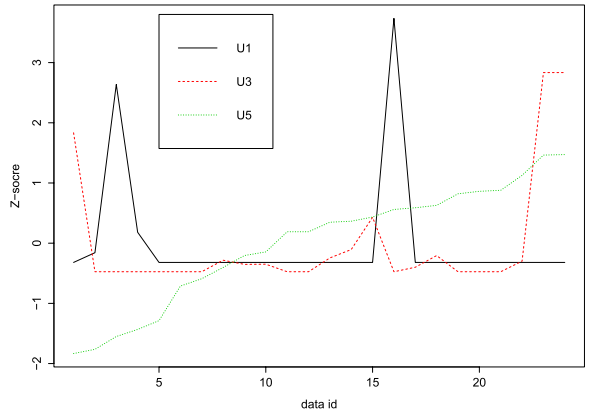


Fig. 9 Distribution of utilities U_1, U_3, U_5.

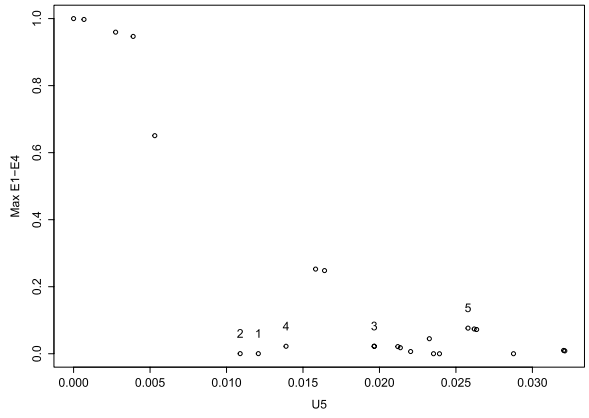


Fig. 10 Relationship between utility and security.

number of success re-identified records indicated horizontally. The re-identification technique quite varies and hence it is very hard to assume a fixed level of adversary.

We also note that the average re-identification ratio for the teams was 20.9%. Even though the data were carefully altered in any smart algorithm, it is almost impossible to prevent all data from being re-identified. There is no perfect algorithm for data anonymization. The competition result implied the limitation of anonymization technique.

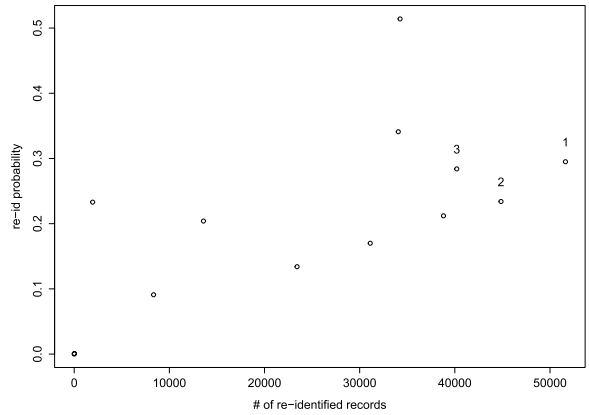


Fig. 11 Re-Id Probability with respect to total number of re-identified records.

Table 7 Ranking of Re-identification qualities.

rank	# of identified records	# of trials	# of tests	re-id ratio
1	51628	21	174993	29.5
2	44852	23	191659	23.4
3	40204	17	141661	28.4
4	38811	22	183326	21.2
5	34247	8	66664	51.4
6	34059	12	99996	34.1
7	31110	22	183326	17.0
8	23420	21	174993	13.4
9	13584	8	66664	20.4
10	8344	11	91663	9.1
11	1943	1	8333	23.3
12	18	2	16666	0.1
13	5	5	41665	0.0

3.6 Effect of k-Anonymity

We found that some data processed so that k-anonymity were satisfied for some $k > 1$. However, the k-anonymized data did not always improve the security against re-identification.

To see the effect of k-anonymity, we show the maximum re-identification ratio of anonymized data with respect to the average measures of k (S_2) in Fig. 12. Most anonymized data with $k = 1$ (nothing performed for k-anonymity) have the maximum re-identification ratio distributed from 0 to 1.0, shown at the left edge in the figure. In the figure, the highest k is at $S_2 = 107$ and its re-identification is almost zero. Generally, higher S_2 data are more secure against re-identification than the data without k-anonymity. However, there are some exception around S_2 of 30.

Figure 13 illustrates the bar-plot of re-identification ratio for each of minimum k (S_1). The mean re-identification for $k > 1$ is 0.013, which is smaller than that of $k = 1$. Note the mean of data for $k = 3$ is worse than that of $k = 1$. Hence, a naive processing for k-anonymity is not necessarily significant for security.

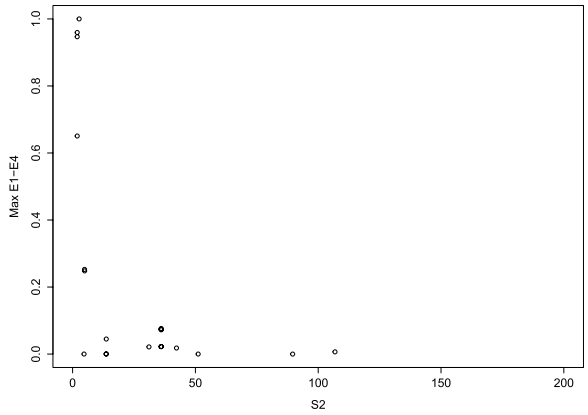


Fig. 12 Re-identification ratio with respect to mean k-anonimity (S_2).

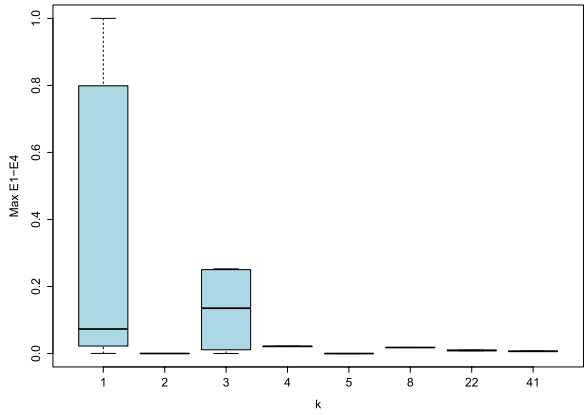


Fig. 13 Bar-plot of re-identification ratio with respect to k.

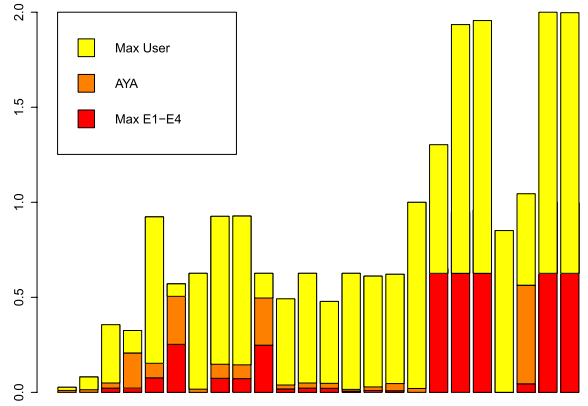


Fig. 14 Ratio of re-identification methods.

3.7 Effect of AYA Re-Identification

Finally, we evaluate the re-identification algorithm as an countermeasure of the cheating permuted data.

Figure 14 illustrates the ratio of method successfully performed to re-identify records, where the ratio are indicated by the length of bars. The most major method was made by users, labeled as "Max User", followed by "Max E1-E4" employed the sample re-identification. The some data, e.g., third one from the right, has the long bar for the label "AYA", which means that the data might be processed by the cheating permutation.

4. Conclusions

We have studied reasonable methods for evaluating the quality of data anonymization in the style of a competition. We have designed the measures for anonymized data in terms of data utility and security against the threat of re-identification. We have developed a competition platform that enables players to participate from remote sites.

As far as we know, this is the first data-anonymization competition in the world wide. We believe that it is a significant undertaking because the competition style is attractive to many engineers and the techniques are evaluated in a common environment. Therefore, methodologies for useful and secure data anonymization are sure to be improved via the competition. We now plan to analyze the results of our competition to identify the most significant elements in anonymization.

References

[1] 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), pp.1035–1042, Crans-Montana, 2016.

[2] Information Commissioner's Office (ICO), Anonymisation: managing data protection risk code of practice, 2012.

[3] K. El Emam and L. Arbuckle, Anonymizing Health Data Case Studies and Methods to Get You Started, O'Reilly, 2013.

[4] PWS CUP 2015 Website, http://www.iwsec.org/pws/2015/pwscup.html

[5] N.Z. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," Usenix 2016.

[6] R. Dewri, T. Ong, and R. Thurimella, "Linking health records for federated query processing," Proc. Privacy Enhancing Technologies (PETS 2016), vol.3, pp.4–23, 2016.

[7] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata," Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp.111–133, 2001.

[8] J. Domingo-Ferrer, S. Ricci, and J. Soria-Comas, "Disclosure risk assessment via record linkage by a maximum-knowledge attacker," 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), IEEE, 2015.

[9] G. Danezis, J. Domingo-Ferrer, M. Hansen, J-H. Hoepman, D.L. Metayer, R. Tirtea, and S. Schinffer, "Privacy and data protection by design – from policy to engineering," ENISA, 2014.

[10] H. Akiyama, K. Yamaguchi, S. Ito, N. Hoshino, and 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).

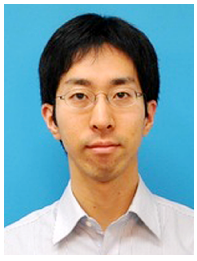
[11] C.C. Aggarwal and P.S. Yu, "A general survey of privacy-preserving data mining, models and algorithms," Privacy-Preserving Data Mining, vol.34, pp.11–52, Springer, 2008.

[12] FTC Report, "Protecting consumer privacy in an era of rapid change," 2012 (available at http://www.ftc.gov/sites/default/files/documents/reports/).

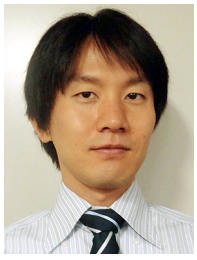
[13] C.J. Skinner and M.J. Elliot, "A measure of disclosure risk for microdata," J. Royal Statistical Society: Series B (Statistical Methodology), vol.64, no.4, pp.855–867, 2002.
[14] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," J. Am. Med. Inform. Assn., vol.17, no.2, pp.169–177, 2010.
[15] UCI Machine Learning Repository (http://archive.ics.uci.edu/ml).
[16] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," Trans. Data Privacy, vol.7, no.3, pp.337–370, 2014.
[17] M. Templ, A. Kowarik, and B. Meindl, "Statistical disclosure control methods for anonymization of microdata and risk estimation," sdcMicro package in R, 2015.
[18] P. Samarati, "Protecting respondents identities in microdata release," IEEE Trans. Knowl. Data Eng., vol.13, no.6, pp.1010–1027, 2001.
[19] L. Sweeny, "k-anonymity," Int. J. Uncertainty, Fuzziness & Knowledge-Based System, vol.10, pp.571–588, 2002.
[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, vol.1, no.1, Article 3, 2007.
[21] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," Proc. IEEE 23rd Int'l Conf. on Data Engineering (ICDE'07), IEEE, 2007.
[22] C. Dwork, "Differential privacy," Automata, Languages and Programming Lecture Notes in Computer Science, vol.4052, pp.1–12, 2006.
[23] East Japan Railway Company, "The numbers of passengers in 2012" (in Japanese) (available at http://www.jreast.co.jp/passenger/, referred in 2013).
[24] H. Yasunaga, H. Horiguchi, K. Kuwabara, S. Matsuda, K. Fushimi, and H. Hashimoto, "Outcomes after laparoscopic or open distal gastrectomy for early-stage gastric cancer: A propensity-matched analysis," Annals of Surgery, vol.257, no.4, pp.640–646, 2012.
[25] C. Yao, X.S. Wang, and S. Jajodia, "Checking for k-anonymity violation by views," Proc. 31st International Conference on Very Large Data Bases (VLDB'05), VLDB Endowment, pp.910–921, 2005.
[26] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," Proc. 32nd International Conference on Very Large Data Bases (VLDB'06), VLDB Endowment, pp.139–150, 2006.
[27] G. Poulis, S. Skiadopoulos, G. Loukides, A. Gkoulalas-Divanis, "Apriori-based algorithms for k^m-anonymizing trajectory data," Trans. Data Privacy, vol.7, no.2, pp.165–194, 2014.

Hiroaki Kikuchi was born in Japan. He received B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994, respectively. After he working in Fujitsu Laboratories Ltd. from 1990 through 1993, he had worked at Tokai University from 1994 through 2013. He is currently a Professor in the Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University. He was a visiting researcher at the School of Computer Science, Carnegie Mellon University in 1997. His main research interests are fuzzy logic, cryptographical protocols, and network security. He received the Best Paper Award for Young Researcher of Japan Society for Fuzzy Theory and Intelligent Informatics in 1990, the Best Paper Award of Symposium on Cryptography and Information Security in 1996, the IPSJ Research and Development Award in 2003, the Journal of Information Processing (JIP) Outstanding Paper Award in 2010 and 2017 and the IEEE AINA Best Paper Award in 2013. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM. He is a fellow of the Information Processing Society of Japan (IPSJ).

Takayasu Yamaguchi received B.E. and M.E. degrees from the University of Electro-Communications in 1999 and 2001. He is currently a Senior research engineer, Research Laboratories, NTT DOCOMO, Inc. His research interests include privacy-preserving data mining, statistical machine learning, and big data applications.



Koki Hamada received his B.E. and M.I. degrees from Kyoto University, Kyoto, Japan, in 2007 and 2009. In 2009, he joined NTT Corporation. He is currently a researcher at NTT Secure Platform Laboratories. He is presently engaged in research on cryptography and information security. He received the Best Paper Award of Symposium on Cryptography and Information Security in 2011, the Outstanding Paper Award of Computer Security Symposium in 2012 and 2014, and the IPSJ Yamashita SIG Research Award in 2014. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) and the Information Processing Society of Japan (IPSJ).



Yuji Yamaoka received his M.E. degree from the University of Tokyo in 2003 and has been engaged in Fujitsu Laboratories Ltd. since 2003. His research interests are information security and privacy protection technologies.

Hidenobu Oguri received Bachelor of Literature from Waseda University in 1997. After he worked in TAITO Corporation and other works since 1997, he was engaged in R&D work on data analysis and privacy protection technology at NIFTY Corporation since 2007. After that, he received Ph.D. degree from the Graduate University for Advanced Studies (SOKENDAI) in 2016. He is now working in FUJITSU CLOUD TECHNOLOGIES LIMITED from 2017. His main research interests are anonymization techniques, privacy preserving data mining. He is a member of the Information Processing Society of Japan (IPSJ),

Jun Sakuma received a Ph.D. degree in Engineering from the Tokyo Institute of Technology, Tokyo Japan in 2003. He has been a professor in the Department of Computer Science, School of System and Information Engineering, University of Tsukuba, Tsukuba, Japan, since 2016. He has also been a team leader of Artificial Intelligence Security and Privacy team in the Center for Advanced Intelligence Project, RIKEN, since 2016. Prior to that, he worked as an associate professor in the same department of University of Tsukuba (2009–2016). He worked as an assistant professor in the Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan (2004–2009). He worked as a researcher at Tokyo Research Laboratory, IBM, Tokyo Japan (2003–2004). His research interests include data mining, machine learning, data privacy, and security. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE).