Regular Paper

How to Handle Excessively Anonymized Datasets

Ryo Nojima[1,a) Hidenobu Oguri[2] Hiroaki Kikuchi[3] Hiroshi Nakagawa[4] Koki Hamada[5]
Takao Murakami[6] Yuji Yamaoka[2] Chiemi Watanabe[7]

Received: September 19, 2017, Accepted: March 6, 2018

Abstract: Many companies and organizations have been collecting personal data with the aim of sharing it with partners. To prevent re-identification, the data should be anonymized before being shared. Although many anonymization methods have been proposed thus far, choosing one from them is not trivial since there is no widely accepted criteria. To overcome this situation, we have been conducting a data anonymization and re-identification competition, called PWS CUP, in Japan. In this paper, we introduce a problem appeared at the competition, named an *excessive* anonymization, and show how to formally handle it.

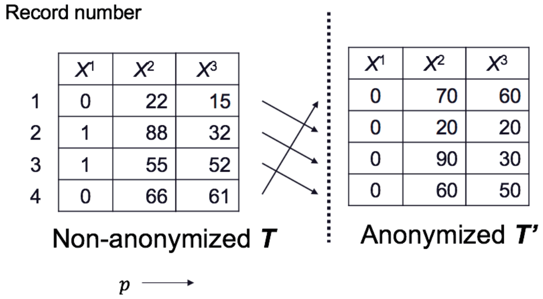
Keywords: PWS CUP, excessively anonymized datasets, distance function, bounds

1. Introduction

Background: Many companies and organizations have become aware of the potential competitive advantage they can obtain from the use of big data. This situation has motivated them to share their data with their partners. Because a possibility exists that such data will contain private and sensitive information, it is often recommended that an anonymization method be applied to the data before it is shared.

Although many anonymization methods have been proposed thus far [5], [6], owing to a lack of widely accepted criteria for anonymization methods, selecting one has been a difficult task. To establish such criteria, among other purposes, we have held an anonymization and re-identification competition in Japan, called PWS CUP, since 2015.

There are two phases to PWS CUP, namely, *anonymization* and *re-identification*. During the anonymization phase, the participant is first given an original dataset, say $\boldsymbol{T}$, and their task is then to generate an anonymized dataset and a permutation, say $\boldsymbol{T}'$ and p, respectively. For ease of explanation, let us assume that datasets $\boldsymbol{T}$ and $\boldsymbol{T}'$ are represented by a matrix (or simply a table), where each row corresponds to a record of the customer (or user), and each column corresponds to an attribute. The key issue of the evaluation framework employed in the competition is the existence of permutation p. Intuitively, this permutation p indicates the relationship between the positions of each customer in $\boldsymbol{T}$ and that in $\boldsymbol{T}'$. For example, in **Fig. 1**, X^1, X^2, and X^3 are the attributes, and the permutation p is defined through each arrow. That is, the arrow at the top indicates that the first column $(0, 22, 15)$ of $\boldsymbol{T}$ is anonymized simply by rounding and then mapped to the second column $(0, 20, 20)$ of $\boldsymbol{T}'$.



	X^1	X^2	X^3
1	0	22	15
2	1	88	32
3	1	55	52
4	0	66	61

X^1	X^2	X^3
0	70	60
0	20	20
0	90	30
0	60	50

Fig. 1 An example of the anonymization.

The anonymized dataset $\boldsymbol{T}'$ is evaluated along with $\boldsymbol{T}$ and p from the viewpoints of security and utility. Intuitively, this is performed as follows:

- If $\boldsymbol{T}$ and $\boldsymbol{T}'$ are a similar (or near), then the triplet $(\boldsymbol{T}, \boldsymbol{T}', p)$ will be given a high score as the utility evaluation, and
- If predicting p given $\boldsymbol{T}$ and $\boldsymbol{T}'$ is hard, then the triplet will be given a high score as the security evaluation

Evaluation in this framework: Let us assume that $\boldsymbol{T}$ and $\boldsymbol{T}'$ are $n \times m$ matrices, where n is the number of customers, and m is the number of attributes. We denote the i-th column of $\boldsymbol{T}$ by $\boldsymbol{T}_i$.

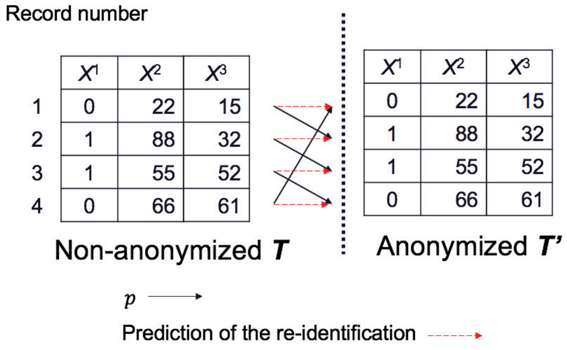
The evaluation based on the triplet $(\boldsymbol{T}, \boldsymbol{T}', p)$ has some serious difficulty, which we call an excessive swap or more generally *excessive* anonymization. That is, there is a possibility that the participant will anonymize the dataset as

$$\boldsymbol{T} = \boldsymbol{T}', p(x) = (x \bmod n) + 1, \text{where, } x \neq 0,$$

which results in high scores. This occurs for the following reasons:

- (Utility) Because $\boldsymbol{T} = \boldsymbol{T}'$, the utility evaluation will become a high score, and
- (Security) For each i, because $\boldsymbol{T}_i = \boldsymbol{T}'_i$, the natural re-identification algorithms will output the identity function

[1] National Institute of Information and Communications Technology, NICT, Koganei, Tokyo 184–8795, Japan
[2] FUJITSU LABORATORIES LTD., Kawasaki, Kanagawa 211–8588, Japan
[3] Meiji University, Nakano, Tokyo 164–8525, Japan
[4] The University of Tokyo, Bunkyo, Tokyo 113–8658, Japan
[5] NTT Secure Platform Laboratories, Musashino, Tokyo 180–8585, Japan
[6] National Institute of Advanced Industrial Science and Technology, AIST, Koto, Tokyo 135–0064, Japan
[7] University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
a) ryo-no@nict.go.jp



Record number	X^1	X^2	X^3
1	0	22	15
2	1	88	32
3	1	55	52
4	0	66	61

X^1	X^2	X^3
0	22	15
1	88	32
1	55	52
0	66	61

Fig. 2 An example of excessive anonymization.

$p'(i) = i$ for all i, where p' is the prediction of p. That is, the security evaluation will become a high score as well.

An example of an excessive anonymization is shown in **Fig. 2**. From this example, it is easy to see that, for every i, because $\boldsymbol{T}_i = \boldsymbol{T}'_i$, the natural re-identification algorithm tends to predict p as $p(i) = i$ for every i.

At first glance this dataset seems to be "non-anonymized" and hence worthless for an anonymization. However, another interpretation is possible. That is, we can regard this anonymization dataset as a result of swapping of two records $\boldsymbol{T}_i$, $\boldsymbol{T}_{(i \bmod 4)+1}$ for every i. Since a swap is one of the widely accepted anonymization operations, it is difficult to strongly assert that this dataset has not been anonymized. Thus, two choices may exist for an interpretation of excessive anonymizations:

(1) the dataset is secure but not useful, or

(2) the dataset is insecure but useful.

In PWS CUP, we regard the dataset has order, that is, $\boldsymbol{T} = (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n)$. This makes two records

$$(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_i, \ldots, \boldsymbol{T}_j, \ldots, \boldsymbol{T}_n) \neq (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_j, \ldots, \boldsymbol{T}_i, \ldots, \boldsymbol{T}_n)$$

if $\boldsymbol{T}_i \neq \boldsymbol{T}_j$, and the interpretation (1) becomes possible. On the other hand, if we regard that the dataset is a *set* $\boldsymbol{T} = \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n\}$, then

$$\{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_i, \ldots, \boldsymbol{T}_j, \ldots, \boldsymbol{T}_n\} = \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_j, \ldots, \boldsymbol{T}_i, \ldots, \boldsymbol{T}_n\}.$$

The interpretation (2) is well suited in this case. In PWS CUP, we assume that the dataset has order and hence our choice is (1) naturally.

Contributions: In this paper, we show how to handle excessive anonymization to avoid giving the triplet high scores. Our approach is as follows: (1) we show how to design the distance function d, which measures the distance between two datasets $\boldsymbol{T}$ and $\boldsymbol{T}'$, and (2) show how to choose threshold t, which becomes the key factor in determining whether the given anonymized dataset is excessive anonymization. More formally, if

$$d_p(\boldsymbol{T}, \boldsymbol{T}') = d(\boldsymbol{T}, \boldsymbol{T}'; p) \geq t,$$

then we regard the triplet as excessive anonymization [*1]. In the competition, the triplet satisfying the above inequality is not accepted as the proper anonymization and is rejected by the system. To push this approach forward, we must determine the following:

(1) a method to construct the distance function d, and

(2) a method to choose the threshold t.

It is easy to see that choosing a correct t is not easy because if t is extremely small, no secure anonymization methods exist. Our contributions include a method to handle this difficulty.

2. Approach for Handling Excessive Anonymization

2.1 Overview of PWS CUP 2016 and the Problem

2.1.1 Overview

In this section, the competition of PWS CUP 2016 is introduced briefly. For further information, please see Ref. [4].

There are two phases, *anonymization* and *re-identification*, in the competition. During the anonymization phase, the participant's task is to anonymize the dataset $\boldsymbol{T}$. Let us denote the anonymized dataset as $\boldsymbol{T}'$. In addition to $\boldsymbol{T}'$, the participants have to generate the permutation p to determine the relation between $\boldsymbol{T}$ and $\boldsymbol{T}'$. Let us assume that $\boldsymbol{T}$ and $\boldsymbol{T}'$ are $n \times m$ matrices, where n is the number of customers, and m is the number of attributes. If $\boldsymbol{T}_i$ is anonymized and becomes $\boldsymbol{T}'_j$, then the permutation p satisfies $p(i) = j$. For Fig. 1 as an example, X^1, X^2, and X^3 are the attributes, and the permutation p is defined through each arrow. In this case, $\boldsymbol{T}_1$ becomes $\boldsymbol{T}'_2$, $\boldsymbol{T}_2$ becomes $\boldsymbol{T}'_3$, $\boldsymbol{T}_3$ becomes $\boldsymbol{T}'_4$, and $\boldsymbol{T}_4$ becomes $\boldsymbol{T}'_1$ and hence $p(i) = (i \bmod 4) + 1$.

In the re-identification phase, each participant tries to guess other participants' p's. The security is essentially evaluated as

$$\frac{|\{p(i) = q(i) \mid 1 \leq i \leq n\}|}{n},$$

where q is the guess submitted by other participants. The utility evaluation can be defined by the distance between $\boldsymbol{T}$ and $\boldsymbol{T}'$. For example, the difference between RFM (Recency, Frequency, Monetary) analysis of $\boldsymbol{T}$ and $\boldsymbol{T}'$ was used in the competition [4].

As we have explained in Section 1, this framework has a problem to overcome. Let us consider a triplet $(\boldsymbol{T}, \boldsymbol{T}', p)$ such that $\boldsymbol{T} = \boldsymbol{T}'$ and p being a random permutation. Then, it is hard for other participants to guess p since p is random. Hence, the dataset $\boldsymbol{T}'$ is evaluated as secure. Moreover, the utility evaluation becomes a high score because $\boldsymbol{T} = \boldsymbol{T}'$. We call this type of anonymizations as *excessive anonymizations* and our motivation in this paper is defeating these problematic anonymizations.

2.1.2 Typical Excessive Anonymizations

Let us consider a triplet $(\boldsymbol{T}, \boldsymbol{T}', p)$ such that $\boldsymbol{T} = \boldsymbol{T}'$. In PWS CUP, the following powerful excessive anonymizations cause the problem:

- (Shift-type excessive anonymization) $p = p_{\text{shift}}(i) = (i \bmod n) + 1$
- (Random-type excessive anonymization) $p = p_{\text{random}}$ is the random permutation on $\{1, \ldots, n\}$.

There are many possible extensions to the above types of excessive anonymization. For the case that every $\boldsymbol{T}_{ij}$ is real, one of the simplest extensions is that for every i, j

$$\boldsymbol{T}'_{ij} = \boldsymbol{T}_{ij} + \epsilon_{ij} \tag{1}$$

[*1] Since we concentrate on the difference between $\boldsymbol{T}_i$ and $\boldsymbol{T}'_{p(i)}$, we also use p.

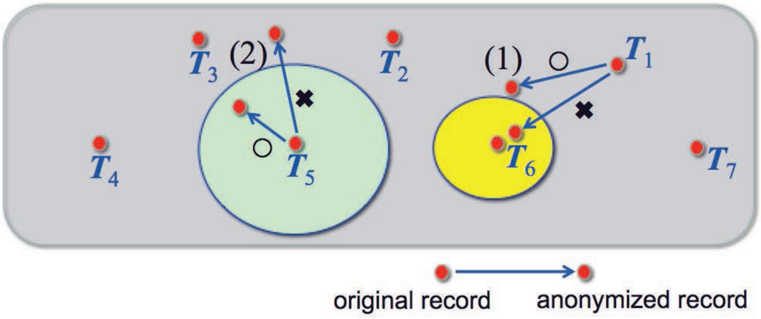


Fig. 3 Two Approaches.

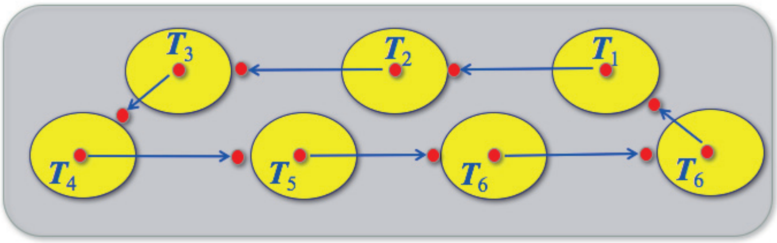


Fig. 4 Variant of the shift-type anonymization.

with the above p's, where each ϵ_{ij} is a small real number. Intuitively, we say that $(\boldsymbol{T}, \boldsymbol{T}', p)$ is an excessive anonymization if

$$(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n) \text{ and } (\boldsymbol{T}'_{p(1)}, \ldots, \boldsymbol{T}'_{p(n)})$$

are "too far" from each other. In this paper, we discuss how to formally determine the meaning of the term "too far."

2.2 Two Approaches

Let $\boldsymbol{T}'_{p(i)}$ be an anonymized record of $\boldsymbol{T}_i$, $d(\boldsymbol{T}_i, \boldsymbol{T}'_{p(i)})$ be their distance, and let t', t be any non-negative real numbers thresholds, where $t = n \times t'$.

There seem to exist at least two approaches to defeat excessive anonymization. For both approaches, the key ingredient is viewing each $\boldsymbol{T}_i$ as a point in a particular space. In **Fig. 3**, each point $\boldsymbol{T}_i$ is mapped to the space. With this figure, we can view the shift-type excessive anonymization as the swapping of $\boldsymbol{T}_1$ to $\boldsymbol{T}_2$, $\boldsymbol{T}_2$ to $\boldsymbol{T}_3$, and so on. To reject the excessive anonymizations, we have at least the following two approaches:

- (Approach I) Determine $\boldsymbol{T}'_{p(i)}$ as an excessive anonymized *record* if i and j exist such that $i \neq j$ and

  $$d(\boldsymbol{T}'_{p(i)}, \boldsymbol{T}_j) \leq t. \tag{2}$$

  The intuition of this approach is shown in (1) of Fig. 3.
- (Approach II) Determine $\boldsymbol{T}'_{p(i)}$ as an excessive anonymized record if

  $$d(\boldsymbol{T}'_{p(i)}, \boldsymbol{T}_i) > t.$$

  The intuition of this approach is shown in (2) of Fig. 3.

At first glance, Approach I may seem better than Approach II. However, this approach seems to have potential difficulty; for example, its naïve implementation of this approach results in accepting the variant of the shift-type excessive anonymization shown in **Fig. 4**. That is,

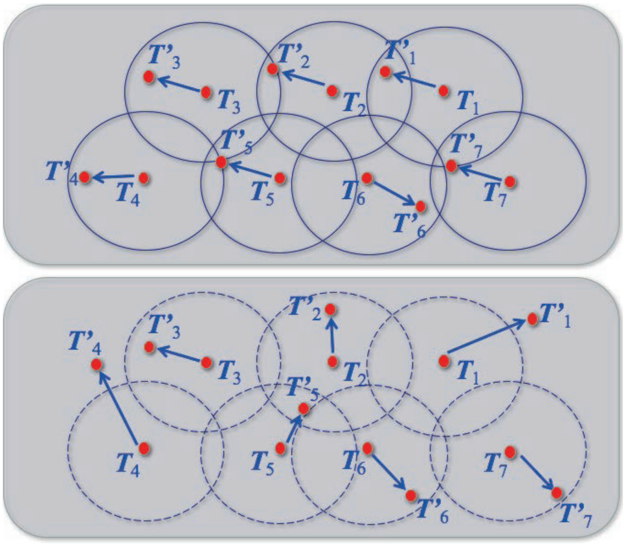


Fig. 5 The difference between strict and average.

- (step 1) apply the shift-type excessive anonymization to the dataset, and then
- (step 2) increase the distance so as not to satisfy Eq. (2).

The resulting anonymized dataset seems to obtain a high score for the security and utility evaluations. On the other hand, if the method based on Approach II is implemented in our system, the system can then reject all of these anonymized datasets. Intuitively, this is because the space for the anonymization is limited compared to Approach I. Hence, our choice is Approach II.

Strict or Average: If Approach II is chosen, then there are two choices for the consideration:

- (Strict) If there exists $\boldsymbol{T}_i$ such that

  $$d(\boldsymbol{T}_i, \boldsymbol{T}'_{p(i)}) > t',$$

  then determine this anonymized dataset as excessive anonymization (see the upper half of **Fig. 5**).

- (Average) If the average is bigger than t', that is
$$\sum_{1\le i\le n} d(\boldsymbol{T}_i, \boldsymbol{T}'_{p(i)}) > t = n \times t', \tag{3}$$
then determine this anonymized dataset as excessive anonymization (see the lower half of Fig. 5).

Thus, if "strict" is chosen, no anonymized record goes beyond threshold t', as shown in the upper half of Fig. 5. On the other hand, if "average" is chosen, then although some of the records can go beyond the threshold, their average does not, as shown in the lower half of Fig. 5. In other words, if we choose "strict", then our system can reject all excessive swapping. However, because the swapping is one of the anonymization operations, we chose the "average" in this paper.

Summary of this section: Our approach is based on Approach II, and determines whether the given dataset is an excessive anonymization based on the "average". Based on this approach, we consider how to design the distance function d and determine the threshold t'.

3. Design of Distance Functions

Let $\mathcal{T}$ be the set of all possible non-anonymized and anonymized *records*, i.e., $\boldsymbol{T}_i, \boldsymbol{T}'_i \in \mathcal{T}$ for all i. Let $d : \mathcal{T}\times\mathcal{T} \to \mathbb{R}^+$ be a distance function, where $\mathbb{R}^+$ is the set of non-negative real numbers. Although the proposal in this paper does not depend on the specific distance functions, the cardinalities may be as follows:

Euclid distance: If each row in $\boldsymbol{T}$ is the customer's location, then we can view the attributes X^1 and X^2 as the longitude and latitude, respectively. When $\boldsymbol{T}_i = (x_1, x_2)$, $\boldsymbol{T}'_i = (x'_1, x'_2)$, then the Euclidian distance
$$d_E(\boldsymbol{T}_i, \boldsymbol{T}'_i) = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2}$$
can be considered.

Jaccard distance: Let A, B be multi-sets. Then the Jaccard index J for the multi-sets is defined by
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4}$$
Further, the Jaccard distance is defined by
$$d_J(A, B) = 1 - J(A, B).$$
We can employ this distance function when each record of $\boldsymbol{T}$ is the multi-set. For example, $\boldsymbol{T}_i$ is a set of gifts bought by customer i.

The distance between the non-anonymized and anonymized records of customer i is defined by $d(\boldsymbol{T}_i, \boldsymbol{T}'_{p(i)})$. Further, the distance between a non-anonymized $\boldsymbol{T}$, and anonymized $\boldsymbol{T}'$ is defined by
$$d(\boldsymbol{T}, \boldsymbol{T}'; p) = \sum_{1\le i\le n} d(\boldsymbol{T}_i, \boldsymbol{T}'_{p(i)}). \tag{5}$$

4. How to Determine the Threshold

4.1 Basic Properties of the Distance Function

To determine the threshold, we focus on powerful excessive anonymization such as those introduced in Section 2.1. To do so, we estimate the amount of $t = \min(d(\boldsymbol{T}, \boldsymbol{T}'; p))$ for the case of $\boldsymbol{T} = \boldsymbol{T}'$. However, for "all" possible p's, $\min(d(\boldsymbol{T}, \boldsymbol{T}; p)) = 0$ because p may be an identity function. Hence to estimate t, the number of fixed points
$$Re(p) = |\{p(i) = i\}|$$
is considered and we then concentrate on estimating
$$s_{\text{exact}}(r) = \min\nolimits_{p \text{ s.t.} Re(p)\le r}(d(\boldsymbol{T}, \boldsymbol{T}; p)) \tag{6}$$
according to each integer $r \ge 0$.

Remark 1. *The readers should note that the number of fixed points of the permutation p_{shift} used in the shift-type excessive anonymization is zero because $p_{\text{shift}}(i) \ne i$ for all i. Thus, if we know $s_{\text{exact}}(0)$ and assign the threshold t as $t < s_{\text{exact}}(0)$ then we can detect any shift-type excessive anonymization. Moreover, we can "expect" that inequality (3) works with $(\boldsymbol{T}, \boldsymbol{T}', p_{\text{shift}})$ even if $\boldsymbol{T}$ and $\boldsymbol{T}'$ are close in the sense of Eq. (1) with approximately the same t.*

For the property of $s_{\text{exact}}(r)$, we have a simple but useful lemma:

Lemma 1. *For any positive integer r, $s_{\text{exact}}(r-1) \ge s_{\text{exact}}(r)$.*

Proof. Recall that from the definition of Eq. (6),
$$s_{\text{exact}}(r-1) = \min\nolimits_{p \text{ s.t.} Re(p)\le r-1}(d(\boldsymbol{T}, \boldsymbol{T}; p)),$$
$$s_{\text{exact}}(r) = \min\nolimits_{p \text{ s.t.} Re(p)\le r}(d(\boldsymbol{T}, \boldsymbol{T}; p)).$$
Since
$$\{p \mid Re(p) \le r-1\} \subseteq \{p \mid Re(p) \le r\},$$
we have $s_{\text{exact}}(r-1) \ge s_{\text{exact}}(r)$. □

Our contribution comes from the following simple to prove theorem:

Theorem 1. *For any t, if $Re(p) \le r$ and $t < s_{\text{exact}}(r)$, then*
$$d(\boldsymbol{T}, \boldsymbol{T}; p) > t. \tag{7}$$

Proof. From the definition of $s_{\text{exact}}(r)$ in Eq. (6),
$$d(\boldsymbol{T}, \boldsymbol{T}; p) \ge s_{\text{exact}}(r).$$
Further, from the condition of the theorem,
$$s_{\text{exact}}(r) > t.$$
Therefore, $d(\boldsymbol{T}, \boldsymbol{T}; p) > t$. □

Thus, if we determine a triplet $(\boldsymbol{T}, \boldsymbol{T}, p)$ as excessive anonymization with t satisfying $t < s_{\text{exact}}(r)$, then since
$$t < s_{\text{exact}}(r) \le s_{\text{exact}}(r-1) \le \ldots \le s_{\text{exact}}(0),$$
the triplet using the permutation p' such that $Re(p') \le r$ is also judged as an excessive anonymization.

Remark 2. *With an excessive anonymization, when the permutation p with $Re(p) = r$ is used, then a "natural" re-identification algorithm can re-identify r customers. Therefore, if the number of fixed points $Re(p)$ becomes bigger, it becomes insecure. For example, if p with $Re(p) = 40$ is employed, then based on the re-identification algorithm, 40 or more customers will be re-identified.*

Table 1 The relation between $s_{\text{lower}}(\mathtt{TC400}, r)/400$ and the number of fixed points r.

num. of fixed points r	$s_{\text{lower}}(\mathtt{TC400}, r)/400$	Remarks
99	0.64352876022	
47	0.763481675361	The expectation of the number of fixed points with multi-random-type excessive anonymization
19	0.829107320761	
1	0.872342498246	The expectation of the number of fixed points with random-type excessive anonymization
0	0.87483931124	$s_{\text{exact}}(0) \approx 0.89 \times 400$ (See Section 4.3.1), The number of fixed points with shift-type excessive anonymization

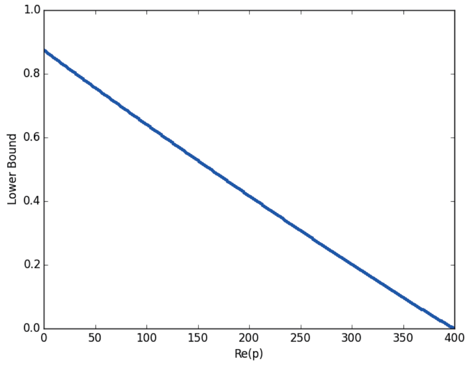


Fig. 6 The relation between $Re(p) = r$ and $s_{\text{lower}}(\mathtt{TC400}, r)/400$.

4.2 How to Estimate the Minimum of $s_{\text{exact}}(r)$ According to r

Our purpose is to estimate $s_{\text{exact}}(r)$. However, as we will mention in Remark 3, computing $s_{\text{exact}}(r)$ seems difficult. Hence in this section, we show how to estimate the *lower bound*, $s_{\text{lower}}(r)$, of $s_{\text{exact}}(r)$.

To estimate the lower bound, $s_{\text{exact}}(r)$ is first expanded as follows:

$$\begin{aligned} s_{\text{exact}}(r) &= \min_{p \text{ s.t.} Re(p) \leq r}(d(\boldsymbol{T}, \boldsymbol{T}; p)) \\ &= \min_{p \text{ s.t. } Re(p) \leq r}(d(\boldsymbol{T}_1, \boldsymbol{T}_{p(1)}) + d(\boldsymbol{T}_2, \boldsymbol{T}_{p(2)}) \\ &\quad + \cdots + d(\boldsymbol{T}_n, \boldsymbol{T}_{p(n)})). \end{aligned}$$

For every j, mdis_j is defined as

$$\text{mdis}_j = \min_{1 \leq k \leq n, j \neq k} d(\boldsymbol{T}_j, \boldsymbol{T}_k).$$

That is, for every $\boldsymbol{T}_j$, the nearest $\boldsymbol{T}_k$ is taken. Further, $\{\text{mdis}_j\}_{1 \leq j \leq n}$ are arranged as:

$$\text{mdis}_{j_1} \leq \text{mdis}_{j_2} \leq \ldots \leq \text{mdis}_{j_n}.$$

Then, as the lower bound of $s_{\text{exact}}(r)$, we have

$$s_{\text{exact}}(r) \geq s_{\text{lower}}(r) = \sum_{k=1}^{n-r} \text{mdis}_{j_k} \tag{8}$$

for all $0 \leq r \leq n-2$ [*2].

Application to the dataset of PWS CUP 2016: For PWS CUP 2016, $d = d_J$, and the dataset $\boldsymbol{T}$ was $\boldsymbol{T} = \mathtt{TC400}$, which is in fact a subset of the dataset used in Ref. [1]. In $\mathtt{TC400}$, there are 400 customers, whereas there are about 5,000 customers in the original dataset. Within this setting, the lower bound (8) is computed as shown in **Table 1** and **Fig. 6**.

As we explain later in Section 4.3.1,

$$s_{\text{exact}}(\mathtt{TC400}, 0) \approx 0.891 \times 400.$$

On the other hand, the lower bound derived from Eq. (8) becomes

$$s_{\text{lower}}(\mathtt{TC400}, 0) = 0.8748 \times 400.$$

Hence, within this dataset, the lower bound seems relatively tight.

4.3 Choice of Threshold t

In this section, we estimate the threshold based on the number of fixed points used in a typical excessive anonymization, such as the ones introduced in Section 2.1.

4.3.1 Shift-type Excessive Anonymization

As we noted previously, because the permutation used in a shift-type excessive anonymization has a property that

$$\forall i : p(i) \neq i,$$

we have

$$Re(p) = 0.$$

Hence, if threshold t is chosen such that $t < s_{\text{lower}}(0) \leq s_{\text{exact}}(0)$ and p satisfies $Re(p) = 0$, then the given triplet $(\boldsymbol{T}, \boldsymbol{T}, p)$ is judged as excessive anonymization.

Although computing $s_{\text{exact}}(r)$ seems difficult *in general*, we know how to compute $s_{\text{exact}}(r)$ when $r = 0$. To compute $s_{\text{exact}}(0)$, we introduce the symmetric matrix D whose elements are defined as follows:

$$[D_{jk}] = \begin{cases} d(\boldsymbol{T}_j, \boldsymbol{T}_k) & \text{if } j \neq k \\ \infty & \text{otherwise.} \end{cases}$$

For example, D becomes

[*2] Precisely, since $s_{\text{exact}}(n-1) = s_{\text{exact}}(n) = 0$, we should remove the case when $r = n - 1$.

$$D = \begin{pmatrix} \infty & \underline{4} & 7 & 6 \\ 4 & \infty & \underline{3} & 2 \\ 7 & 3 & \infty & \underline{5} \\ \underline{6} & 2 & 5 & \infty \end{pmatrix}.$$

Thus, $p(1) = 2$, and $D_{1,2} = d(\boldsymbol{T}_1, \boldsymbol{T}_2) = 4$ is the distance between $\boldsymbol{T}_1$ and $\boldsymbol{T}_{p(1)}$. Hence, when

$$\begin{aligned} &p(1) = 2, p(2) = 3, p(3) = 4, p(4) = 1, \\ &d(\boldsymbol{T}, \boldsymbol{T}, p) = d(\boldsymbol{T}_1, \boldsymbol{T}_{p(1)}) + \cdots + d(\boldsymbol{T}_4, \boldsymbol{T}_{p(4)}) \\ &\qquad\qquad = 4 + 3 + 5 + 6 = 18. \end{aligned}$$

The minimum is $s_{\text{exact}}(0)$, which is the instance of the assignment problem that can be solved using the Hungarian algorithm with a time complexity of $O(n^3)$.

Remark 3. *Although in Section 4.2 we showed how to compute $s_{\text{lower}}(r)$, we were unable to find an efficient algorithm which computes $s_{\text{exact}}(r)$ unless $r = 0$. We believe this belongs to an NP-hard problem, but do not know how to prove it.*

Application to the dataset of PWS CUP 2016: By setting $\boldsymbol{T} = \texttt{TC400}$, and $d = d_J$ as the input, the Hungarian algorithm outputs

$$\begin{aligned} s_{\text{exact}}(\texttt{TC400}, 0) &= \min\nolimits_{p \text{ s.t.} Re(p) \leq 0}(d(\texttt{TC400}, \texttt{TC400}, p)) \\ &= 356.395595858 \\ &> 0.89 \times 400. \end{aligned}$$

Hence, our system can reject the shift-type excessive anonymization by setting $t = 0.89 \times 400 \geq s_{\text{lower}}(\texttt{TC400}, 0) = 0.8748 \times 400$, as an example.

4.3.2 Random-type Excessive Anonymization

With excessive anonymization, choosing p randomly is useful for making the re-identification difficult:

```
Algorithm 1: Random-type excessive anonymization
Input:   $T$
Step 1.  Choose random permutation $p$
Step 2.  Output $(T, T, p)$ as $(T, T', p)$
```

In this subsection, we estimate the number of fixed points when p is chosen randomly. More precisely, we estimate

$$q^*(l) = \Pr_{p \in_R \mathsf{Perm}}[Re(p) \leq l],$$

where Perm is a set of all the permutations on $\{1, \ldots, n\}$. Here, because the number of fixed points is equal to or less than l with probability $q^*(l)$, if threshold t is chosen such that $s_{\text{exact}}(l) > t$, then with a probability at least $q^*(l)$, the output generated by the above algorithm will be rejected by the system.

For the estimation of $q^*(l)$, the corollary below is useful:

Corollary 1. *If the permutation p on $\{1, \ldots, n\}$ is chosen randomly, then the probability that the number of fixed points will be exactly k is*

$$\Pr_{p \in_R \mathsf{Perm}}[Re(p) = k] = \frac{1}{k!} \sum_{i=0}^{n-k} \frac{-1^i}{i!}. \tag{9}$$

Proof. The proof depends on the following theorem:

Theorem 2 (Extracted from Proposition 3.5 in Ref. [2]). *The number of permutations p on $\{1, \ldots, \hat{n}\}$ such that $Re(p) = 0$ is*

$$\sum_{i=0}^{\hat{n}} (-1)^i \binom{\hat{n}}{i} (\hat{n} - i)! = \hat{n}! \sum_{i=0}^{\hat{n}} \frac{(-1)^i}{i!}.$$

Hence, the number of permutations p such that $Re(p) = k$ is

$$\binom{n}{n-k} (n-k)! \sum_{i=0}^{n-k} \frac{(-1)^i}{i!} = \frac{n!}{k!} \sum_{i=0}^{n-k} \frac{(-1)^i}{i!}. \tag{10}$$

The corollary follows because the number of permutations is $n!$. □

With this corollary, $q^*(l)$ can be computed as

$$q^*(l) = \Pr_{p \in_R \mathsf{Perm}}[Re(p) \leq l] = \sum_{k=0}^{l} \Pr_{p \in_R \mathsf{Perm}}[Re(p) = k], \tag{11}$$

where each $\Pr_{p \in_R \mathsf{Perm}}[Re(p) = k]$ can be computed using Eq. (9).

Application to the dataset of PWS CUP 2016: By setting $\boldsymbol{T} = \texttt{TC400}$ ($n = 400$), $d = d_J$, and $l = 19$, Eq. (11) is computed concretely as [*3]

$$\begin{aligned} q^*(19) &= \Pr_{p \in_R \mathsf{Perm}}[Re(p) \leq 19] \\ &= \sum_{k=0}^{19} \Pr_{p \in_R \mathsf{Perm}}[Re(p) = k] \\ &\approx 1 - \frac{1}{2^{62}}, \end{aligned}$$

where $l = 19$ is chosen such that $q^*(19)$ is close enough to 1. In our case, we choose $> 1 - 2^{60}$ or $> 1 - 2^{50}$. Therefore, if we set the threshold t such that $t < s_{\text{lower}}(\texttt{TC400}, 19)$, then our system can reject the random-type excessive anonymization with a probability of at least $q^*(19)$.

4.3.3 Multi Random-type Excessive Anonymization

In PWS CUP 2016, as the utility evaluation, there is `ut-cmae2`. (Please see Ref. [4] for the detail.) To obtain a high score, employing excessive anonymization that takes this utility evaluation into account is one of the best strategies if there is no mechanism to reject it. In this section, we estimate the threshold t to defeat this excessive anonymization.

First let us consider the subsets $S_1, \ldots, S_c \subset \{1, \ldots, n\}$ such that

$$\begin{aligned} &\{1, \ldots, n\} = \cup_{1 \leq k \leq c} S_k, \\ &\forall k, j, \text{s.t. } k \neq j, S_k \cap S_j = \emptyset. \end{aligned}$$

In addition, let us denote the procedure for choosing the permutation on S_j randomly as $p_j \in_R \texttt{Perm}(S_j)$. Then, our next target is estimating the threshold for rejecting the output of the following algorithm.

```
Algorithm 2: Multi-random-type excessive anonymization
Input:   $T, S_1, \ldots .S_c$
         (For PWS CUP 2016, each $S_i$ is the subset of $\{1, \ldots, n\}$
         based on gender and nationality, which results in $c = 47$.)
Step 1.  For every $1 \leq j \leq c$, choose random permutation $p_j \in_R \texttt{Perm}(S_j)$
Step 2.  Output $(T, T, p)$ such that $p(x) = p_j(x)$ if $x \in S_j$
```

[*3] This is the result of Python2.7 using Decimal, NumPy packages.

Let us denote the probability that the output p generated by Algorithm 2 satisfies $Re(p) \leq l$ by

$$q^*(l) = \Pr_{(p_1,\dots,p_c)\in_R(\mathsf{Perm}(S_1),\dots,\mathsf{Perm}(S_c))}[Re(p) \leq l],$$

where $p(x) = p_j(x)$ if $x \in S_j$. If t is chosen such that $s_{\mathrm{exact}}(l) > t$, then with a probability at least $q^*(l)$, the triplet generated by the above algorithm will be rejected.

We can compute $q^*(l)$ using the probability generating function. To do so, we firstly define q_{jk} as

$$q_{jk} = \Pr_{p_j\in_R\mathsf{Perm}(S_j)}[Re(p_j) = k], |S_j| = N_j,$$

and then focus on the following probability generating function:

$$G(x) = \prod_{1\leq j\leq c}\left(q_{j0} + q_{j1}x + q_{j2}x^2 + \cdots + q_{jN_j}x^{N_j}\right).$$

Then, let us consider the expansion of $G(x)$. The coefficient of the degree k term of the expanded $G(x)$ becomes

$$\Pr[Re(p) = k],$$

where each q_{jk} can be computed using Corollary 1. Thus, to compute $q^*(l)$, we firstly expand the equation as

$$q^*(l) = \sum_{k=0}^{l}\Pr_{(p_1,\dots,p_c)\in_R(\mathsf{Perm}(S_1),\dots,\mathsf{Perm}(S_c))}[Re(p) = k] \qquad (12)$$

and then compute each $\Pr[Re(p) = k]$ using the probability generating function.

Application to the dataset of PWS CUP 2016: By setting $\boldsymbol{T}$ = TC400 (n = 400), and l = 99, Eq. (12) is computed concretely as

$$\begin{aligned} q^*(99) &= \sum_{k=0}^{99}\Pr_{(p_1,\dots,p_c)\in_R(\mathsf{Perm}(S_1),\dots,\mathsf{Perm}(S_c))}[Re(p) = k] \\ &\approx 1 - \frac{1}{2^{51}}, \end{aligned}$$

where the concrete value of $N_1, \dots, N_c$ can be computed from Ref. [4]. Also $l = 99$ is chosen such that $q^*(99) > 1-2^{60}$ or $1-2^{50}$. Therefore, by setting the threshold t as $t < s_{\mathrm{lower}}(\mathtt{TC400}, 99)$, with probability at least $q^*(l)$, the multi-random-type excessive anonymization will be rejected.

Summary of Section 4.1, 4.2 and 4.3: To reject shift-type, random-type, and multi-random-type excessive anonymization using our system from PWS CUP 2016, it is sufficient to choose t satisfying $t < s_{\mathrm{lower}}(\mathtt{TC400}, 99)$. As a concrete value, we can find

$$t = 0.64 \times 400 < s_{\mathrm{lower}}(\mathtt{TC400}, 99) \qquad (13)$$

from Table 1.

4.4 Threshold Causing Anonymization Shortage

By setting a small threshold, we can reduce the chance of excessive anonymization. However, if the threshold is too small there might be no *secure* anonymization. In this section, we estimate such t.

To obtain the intuition regarding this, let us map each $\boldsymbol{T}_i$ to a high dimensional space as shown in **Fig. 7**. For example, if each record is anonymized so as to not go beyond the circle in Fig. 7, an adversary can re-identify each i using Algorithm 3.

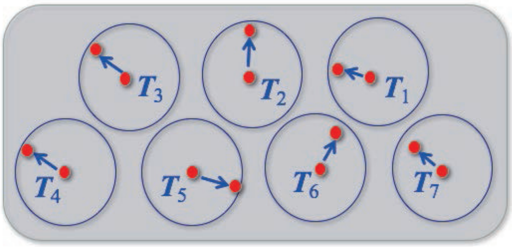


Fig. 7 Anonymization shortage due to the small threshold.

```
Algorithm 3
Input: $\boldsymbol{T}, \boldsymbol{T}'$
Step 1. For $1 \leq k \leq n$, find $k_j = \mathrm{argmin}_j d(\boldsymbol{T}_j, \boldsymbol{T}'_k)$
Step 2. Output $(k_1, \dots, k_n)$
```

This is because the nearest anonymized record of $\boldsymbol{T}_j$ is always $\boldsymbol{T}'_{p(i)}$.

More formally, let us denote the radius of circle as $t' = \frac{t}{n}$. To estimate t', we consider the *minimum distance*

$$\mathrm{Min} = \min_{j,k \text{ s.t. } j\neq k} d(\boldsymbol{T}_j, \boldsymbol{T}_k).$$

That is, if t' satisfies $\mathrm{Min} > 2t'$, then any two circles do not cross each other, as shown in Fig. 7, and thus every record $\boldsymbol{T}_i$ becomes re-identified through Algorithm 3. Therefore, t' must satisfy $\mathrm{Min} < 2t'$ at least.

Application to the dataset of PWS CUP 2016: The minimum distance of TC400 used in PWS CUP 2016 is

$$\mathrm{Min} = \min_{j,k \text{ s.t. } j\neq k} d_J(\mathtt{TC400}_j, \mathtt{TC400}_k) \approx 0.713362. \qquad (14)$$

Hence, t' must satisfy

$$t' > \mathrm{Min}/2 \approx 0.357.$$

Summarizing the results based on Eq. (13), the threshold t can be chosen from within the following range:

$$0.357 \times 400 < t \leq 0.64 \times 400 \approx s_{\mathrm{lower}}(\mathtt{TC400}, 100). \qquad (15)$$

5. The Use Case: PWS CUP 2016

The dataset employed in PWS CUP 2016 [1] has many attributes: invoice numbers, stock codes, customer IDs, invoice dates and etc. To employ our proposal in PWS CUP 2016, only customer IDs and stock codes are used, which means we regard $\boldsymbol{T}_i$ as a set of gifts (stock codes) bought by customer i. Moreover, the distance is measured by the Jaccard distance. As we have noted, to prevent the shift-type, the random-type, and the multi-random-type excessive anonymizations, it is sufficient to chose the threshold $0.64 * 400$. We now demonstrate that this is true by computer simulation. **Figures 8**, **9** and **10** show the frequency of $d(\boldsymbol{T}, \boldsymbol{T}, p)/400$ during 10,000 trials in the setting of PWS CUP 2016, which is larger than 0.64 of Eq. (15) as we have expected.

It is easy to see that our proposal in this paper successfully rejects all the powerful excessive swaps introduced in Section 2.1. Next, we are going to explain some of the proper anonymization

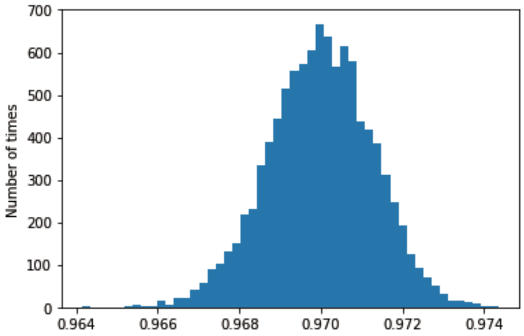


Fig. 8 Measuring $d(\boldsymbol{T}, \boldsymbol{T}, p)/400$ with the shift-type excessive swap, where p is chosen random such that $Re(p) = 0$.

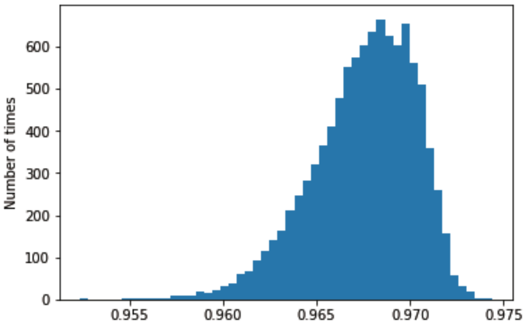


Fig. 9 Measuring $d(\boldsymbol{T}, \boldsymbol{T}, p)/400$ with the random-type excessive swap.

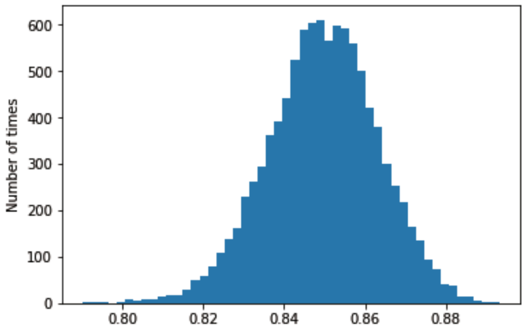


Fig. 10 Measuring $d(\boldsymbol{T}, \boldsymbol{T}, p)/400$ with the multi-random-type excessive swap.

methods still work. The candidacies of the anonymization methods for PWS CUP 2016 some of the authors of this paper considered were (i) the pseudonymization, and (ii) the perturbation on the attribute "date." Since we have only used the stock codes for the measure of the excessive swaps, the participants can employ (i) and (ii) without any penalty. Moreover, the participants can employ any anonymization method without a penalty if the stock codes are not modified too much.

6. Conclusion

In this paper, we explored an excessive anonymization. Our contributions to this topic are methods for constructing the distance function and choosing the threshold. As we explained in Section 2, Approach II does not reject all excessive anonymization because a certain number of swap operations are allowed. However, the number of swap operations can be controlled through this approach. For PWS CUP 2016, Approach II together with "average" was in fact employed. In the competition, the Jaccard distance is employed, threshold t becomes 0.7×400, which is slightly larger than the inequality in Eq. (15).

Finally, we emphasize that the proposed method is not only applied to the excessive anonymization problem in PWS CUP 2016 to derive the threshold but also applied to the excessive anonymization problem in other similar settings.

References

[1] Chen, D., Sain, S.L. and Guo, K.: Data Mining for the Online Retail Industry: A Case Study of RFM Model-Based Customer Segmentation Using Data Mining, *Journal of Database Marketing & Customer Strategy Management*, Vol.19, No.3, pp.197–208 (2012).

[2] Jukna, S.: *Extremal Combinatorics with Applications in Computer Science*, Springer-Verlag (2001).

[3] Kikuchi, H., Yamaguchi, T., Hamada, K., Yamaoka, Y., Oguri, H. and Sakuma, J.: Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization, *AINA*, pp.1035–1042 (2016).

[4] Kikuchi, H., Oguri, H., Nojima, R., Hamada, K., Murakami, T., Yamaoka, Y., Yamaguchi, T. and Watanabe, C.: PWSCUP Competition: De-identify Transaction Data Securely, *Computer Security Symposium*, 2A1-2 (2016).

[5] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity, *ACM Trans. Knowledge Discovery from Data*, Vol.1, No.1, Article 3 (2008).

[6] Sweeney, L.: k-anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, Vol.10, No.5, pp.557–570 (2002).

Ryo Nojima was born in 1976. He received his Ph.D. in 2005 from NAIST. In 2005, he became a postdoctoral fellow at The University of Tokyo and, then in 2006, he became a researcher at NICT. He is currently a research manager at NICT.

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 his Ph.D. degree from the Graduate University for Advanced Studies (SOKENDAI) in 2016. He is now working in FUJITSU LABORATORIES LTD. from 2017. His main research interests are anonymization techniques, privacy preserving data mining. He is a member of IPSJ.

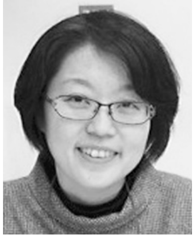
Hiroaki Kikuchi received his B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994. After he working in Fujitsu Laboratories Ltd. in 1990, he had worked in Tokai University from 1994 through 2013. He is currently a professor at Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University. He was a visiting researcher of the School of Computer Science, Carnegie Mellon University in 1997. His main research interests are network security, cyrptographical protocl, privacy-preserving data mining, and fuzzy logic. He received the Best Paper Award for Young Researcher of Japan Society for Fuzzy Theory and Intelligent Informatics in 1990, the Best Paper Award for Young Researcher of IPSJ National Convention in 1993, the Best Paper Award of Symposium on Cryptography and Information Security in 1996, the IPSJ Research and Development Award Award in 2003, the Journal of Information Processing (JIP) Outstanding paper Award in 2010, and the IEEE AINA Best Paper Award in 2013. He is a member of IEICE, IPSJ, the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM. He is a director of IPSJ since 2013. He receives IPSJ Fellow.

Hiroshi Nakagawa was born in 1953. He received his M.E. and Doctor of Engineering from The University of Tokyo in 1977 and 1980, respectively. He became an associate professor at Yokohama National University in 1981 and a professor at The University of Tokyo in 1999. He is also the group director of RIKEN AIP in 2017. His current research interests include artificial intelligence, machine learning and privacy protection technologies.

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.

Takao Murakami was born in 1981. He received his M.E. and Ph.D. from The University of Tokyo in 2006 and 2014, respectively. He is currently a researcher with the National Institute of Advanced Industrial Science and Technology (AIST). His research interest is biometrics and privacy. He received the IEEE TrustCom Best Paper Award in 2015. He is a member of IEEE, IEICE, and IPSJ.

Yuji Yamaoka received his Master of Information Science and Technology 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.



Chiemi Watanabe was born in 1975. She received her M.S. and Ph.D. from Ochanomizu University in 2000 and 2003, respectively. She became an assistant professor at Nara Women's University in 2003, and a lecturer at Ochanomizu University in 2005, and assistant processor at Tsukuba University in 2013. Her current research interest is privacy preserved database systems. She is a member of IPSJ, IEEE-CS, and ACM.