IEICE TRANS. INF. & SYST., VOL.E106–D, NO.5 MAY 2023

PAPER Special Section on Data Engineering and Information Management

Privacy-Preserving Correlation Coefficient

Tomoaki MIMOTO[†a)], ***Nonmember***, **Hiroyuki YOKOYAMA**[†], **Toru NAKAMURA**[††], **Takamasa ISOHARA**[††], ***Members***, **Masayuki HASHIMOTO**[††], **Ryosuke KOJIMA**[†††], **Aki HASEGAWA**[†††], ***and*** **Yasushi OKUNO**[†††], ***Nonmembers***

SUMMARY Differential privacy is a confidentiality metric and quantitatively guarantees the confidentiality of individuals. A noise criterion, called sensitivity, must be calculated when constructing a probabilistic disturbance mechanism that satisfies differential privacy. Depending on the statistical process, the sensitivity may be very large or even impossible to compute. As a result, the usefulness of the constructed mechanism may be significantly low; it might even be impossible to directly construct it. In this paper, we first discuss situations in which sensitivity is difficult to calculate, and then propose a differential privacy with additional dummy data as a countermeasure. When the sensitivity in the conventional differential privacy is calculable, a mechanism that satisfies the proposed metric satisfies the conventional differential privacy at the same time, and it is possible to evaluate the relationship between the respective privacy parameters. Next, we derive sensitivity by focusing on correlation coefficients as a case study of a statistical process for which sensitivity is difficult to calculate, and propose a probabilistic disturbing mechanism that satisfies the proposed metric. Finally, we experimentally evaluate the effect of noise on the sensitivity of the proposed and direct methods. Experiments show that privacy-preserving correlation coefficients can be derived with less noise compared to using direct methods.

key words: *differential privacy, dummy data, correlation coefficient*

1. Introduction

Differential privacy, which is a confidentiality metric proposed in 2006 by Dwork [1], assumes an interactive query response with a database. Differential privacy is satisfied using a simple probabilistic mechanism that adds specific noise to the correct output result during a query process. However, from the perspectives of both privacy and utility, the mechanism must be tuned depending on the query. For example, the magnitude of noise added by the probabilistic disturbing mechanism is determined based on the effect that a record can have on the output value of the query, i.e., sensitivity. Therefore, it is necessary to reduce the noise and to maintain privacy by, for example, converting the data structure to reduce the sensitivity.

In this paper, we assume a query whose sensitivity is difficult to compute. Such difficulty is due to the differential privacy-specific definition. That is, because adjacent data for any given data set must be considered, datasets must be excluded that might be undefined or divergent depending on the query. This may seem too powerful an assumption at first glance, but differential privacy quantifies the minimum privacy strength that can be guaranteed for the entire data space without exception by considering any adjacent data for any given data set. New metrics that exclude certain datasets or situations could be proposed, but new metrics that exclude certain datasets or situations could be proposed, but exceptional situations should be considered on a individual query basis. In addition, because the data space changes, it is not possible to evaluate privacy strength strictly in combination with other differential privacy mechanisms. Therefore, we claim that the privacy strength of any queries should be accounted for by a uniform definition, rather than by exception handling, such as changing the data space for different queries. We solve this uncomputable sensitivity problem without using exception handling by proposing a confidentiality metric for datasets to which dummy data are added. Mechanisms that satisfy the proposed metric also satisfy differential privacy, and for queries for which the sensitivity is computable, the relationship between the proposed metric and differential privacy can be approximated. Furthermore, we focus on the correlation coefficient as a case study and propose and implement a probabilistic disturbing mechanism that satisfies the proposed metric. It is impossible to directly calculate the sensitivity of correlation coefficients, and no efficient mechanism has been proposed for deriving correlation coefficients that satisfy differential privacy. Therefore, we compare those that satisfy the differential privacy obtained using the sequential theorem and the local differential privacy mechanism.

2. Related Work

There are two well-known privacy metrics based on information theory: differential privacy, which assumes interactive query responses with databases [1], and local differential privacy, which assumes the provisions of the data themselves [2]. In particular, local differential privacy is a model that is suitable for such use cases as the collection of healthcare data by smartwatches and other devices, and companies such as Google [3] and Apple [4] are already putting them to practical use by having users provide data to which the local differential privacy mechanism is applied when using

Manuscript received June 27, 2022.
Manuscript revised November 16, 2022.
Manuscript publicized February 8, 2023.

[†]The authors are with Advanced Telecommunications Research Institute International (ATR), Kyoto-fu, 619–0237 Japan.
[††]The authors are with KDDI Research, Inc., Fujimino-shi, 356–8502 Japan.
[†††]The authors are with Kyoto University, Kyoto-shi, 606–8303 Japan.
a) E-mail: to-mimoto@atr.jp

DOI: 10.1587/transinf.2022DAP0014

their services. Differential privacy mechanisms are mainly used to obtain statistical information on large datasets under privacy protection, and several mechanisms have been proposed, such as histogram outputs [5], t-tests [6], chi-square tests [7], and so on. In addition, differential privacy mechanisms are expected to be applied to a wide range of use cases beyond simple statistical processing, such as machine learning applications [8]–[10]. One main goal of previous research is constructing a mechanism that allows for indentical privacy strength with less noise. For example, a privacy-enhanced machine learning model [10] suppresses noise by determining the upper bound of the impact of a single record through clipping. A more general mechanism was proposed [11] to reduce noise and improve utility by reducing the dimensionality of multidimensional data. Related to our study, [12] presents a systematic taxonomy of transformations and extensions of differential privacy depending on scenarios and adversary models. In addition, Zhang et al. [13] studies correlated differential privacy that aims to solve the issue that a data correlation may lead to leak privacy.

3. Preliminary

3.1 Differential Privacy

Differential privacy is a privacy protection metric used when statistical query responses to a database are assumed to be interactive. In many cases, query q is presented to the database, which satisfies differential privacy by applying mechanism M that adds specific noise to the correct query result. We define adjacent datasets to discuss differential privacy.

Definition 3.1 (Adjacent dataset)**.** Let the distance between datasets D and $D'(|D| = |D'| = n)$ be the number of different records $H(D, D') = |\{i : d_i \neq d'_i\}|$. Here $d_i \in D, d'_i \in D'$. Dataset D', for which $H(D, D') = 1$, is defined as an adjacent dataset to D.

Differential privacy is defined as follows.

Definition 3.2 (Differential privacy)**.** Mechanism M satisfies (ϵ, δ)-difference privacy (DP) if for any adjacent dataset, assuming that Range(M) is every possible output that M can take, then $\mathcal{M} \subseteq$ Range(M):

$$Pr[M(D) \in \mathcal{M}] \leq e^{\epsilon} \cdot Pr[M(D') \in \mathcal{M}] + \delta. \quad (1)$$

In dealing with differential privacy mechanisms, we introduce some important properties. Our proposal and experiments exploit these properties [14].

Proposition 3.3 (Sequential Theorem)**.** Given dataset D, N probabilistic mechanisms q_i that satisfy (ϵ_i, δ_i)-DP, and any function g, then $Q(D) = g(q_1(D), \cdots, q_N(D))$, which combines q_i for D, satisfies $(\sum_{i=1}^{N} \epsilon_i, \sum_{i=1}^{N} \delta_i)$-DP.

Proposition 3.4 (Parallel Theorem)**.** Given dataset $D = \bigcup_i^N D_i$, N probabilistic mechanisms q_i that satisfies (ϵ_i, δ_i)-DP, and any function g, then $Q(D) = g(q_1(D_1), \cdots, q_N(D_N))$ satisfies $(\max \epsilon_i, \max \delta_i)$-DP.

3.2 Sensitivity

Sensitivity is a metric that shows the maximum impact on any given record by a mechanism. Sensitivity is defined as follows:

3.2.1 Global Sensitivity

Definition 3.5 (Global sensitivity)**.** Define global sensitivity GS_q for query $q : \mathcal{D} \to \mathbb{R}^k$ as follows. Here, $\|\cdot\|_p$ is the L_p norm function:

$$GS_q = \max_{\forall D, D' : H(D,D')=1} \|q(D) - q(D')\|_p. \quad (2)$$

With global sensitivity, the probabilistic mechanisms given below satisfy differential privacy.

Proposition 3.6 (Gaussian mechanism [15])**.** Let $q : \mathcal{D} \to \mathbb{R}^k$ be a query, let $N(\mu, v)$ be Gaussian noise with mean μ and variance v, and then the following mechanism M_q satisfies (ϵ, δ)-DP:

$$M_q(D) = q(D) + N\left(0, \frac{GS_q^2 \cdot 2\log(2/\delta)}{\epsilon^2}\right). \quad (3)$$

Proposition 3.7 (Laplacian mechanism [16])**.** Let $q : \mathcal{D} \to \mathbb{R}^k$ be a query, let $L(\mu, v)$ be Laplacian noise with mean μ and variance v, and then the following mechanism M_q satisfies $(\epsilon, 0)$-DP.

$$M_q(D) = q(D) + L\left(0, \frac{GS_q}{\epsilon}\right). \quad (4)$$

3.2.2 Smooth Sensitivity Framework

In global sensitivity, since sensitivity to arbitrary datasets is addressed, the sensitivity may be very large for some queries. Therefore, a previous work [17] devised sensitivity for dataset D instead of arbitrary datasets.

Definition 3.8 (Local sensitivity)**.** Given dataset D, we define local sensitivity LS_q for query $q : \mathcal{D} \to \mathbb{R}^k$ as follows. Here $\|\cdot\|_p$ is the L_p norm function:

$$LS_q(D) = \max_{\forall D' : H(D,D')=1} \|q(D) - q(D')\|_p. \quad (5)$$

Furthermore, smooth sensitivity is defined for the situations where the local sensitivity fails to satisfy the definition of differential privacy. Smooth sensitivity takes the local sensitivity for arbitrary datasets into account.

Definition 3.9 (Smooth sensitivity)**.** Suppose $\beta > 0$, and dataset D is given. We define smooth sensitivity $S_{q,\beta}$ for query $q : \mathcal{D} \to \mathbb{R}^k$ as follows:

$$S_{q,\beta}(D) = \max_{anyD'}(LS_q(D') \cdot e^{-\beta H(D,D')}). \quad (6)$$

Using smooth sensitivity, the probabilistic mechanism given below satisfies differential privacy.

Proposition 3.10. Let $q : \mathcal{D} \to \mathbb{R}^k$ be a query, let $\alpha = \epsilon/\sqrt{\ln(1/\delta)}$ and $\beta = \Omega(\epsilon/\sqrt{k\ln(1/\delta)})$, and then the following mechanism M_q satisfies (ϵ, δ)-DP:

$$M_q(D) = q(D) + \frac{S_{q,\beta}(D')}{\alpha} \cdot N(0,1). \quad (7)$$

3.3 Correlation Coefficient

The correlation coefficient is a measure of the relationship between two attributes, A and B. Correlation coefficient C is given by the following equation:

$$C = \frac{s_{AB}}{s_A s_B} = \frac{\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_A)(b_i - \mu_B)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_A)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - \mu_B)^2}}, \quad (8)$$

where s_{AB} denotes the covariance between A and B, s_A, s_B denote the standard deviation of A, B, n denotes the number of bivariate data (a_i, b_i), and μ_A, μ_B respectively denote the mean values of A, B. The correlation coefficient takes value $-1 \le C \le 1$, where A, B have a stronger negative correlation when C is close to -1 and a stronger positive correlation when C is close to 1. Since the correlation coefficient is a linear measure of the relationship, even when the correlation coefficient is close to 0, the relationship might be quadratic, or outliers might significantly affect the correlation coefficient. Although the correlation coefficient alone cannot identify all the relationships among the attributes, it is an indispensable measure as a basis for data analysis.

4. Proposal

4.1 Differential Privacy with Dummy Data

We next consider queries for which sensitivity is difficult to derive. As stated in Definitions 3.5 and 3.8, sensitivity is the maximum difference between the output values for the query in adjacent datasets. This may seem like an excessive value, but it is necessary to ensure a lower bound of privacy for the entire data space. Here recalling the correlation coefficient, its denominator is $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_A)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - \mu_B)^2}$. If all the data in A or B have the same value, i.e., $a_1 = \cdots = a_n$ or $b_1 = \cdots = b_n$, the denominator is zero, and the correlation coefficient cannot be calculated. Even in the case of smooth sensitivity, where a dataset is given as input, the calculation of $S_{q,\beta}$ is impossible as long as there exists D' whose denominator is zero. Thus, when considering arbitrary datasets, computing sensitivity is difficult for undefined or divergent queries. The definition of differential privacy makes it difficult to calculate sensitivity of queries when there are non-computable dataset on the data space. One policy to combat this is to restrict special circumstances. For example, in the case of correlation coefficients, it is acceptable to exclude datasets in which all records coincide from the data space. However, exception handling should be considered each time the query or system design changes. In addition, exception handling may deviate from the strictly original definition of differential privacy, such as a change in data space, and may be difficult to compare directly with the privacy strength of other mechanisms. Therefore, we propose differential privacy with dummy data as an approach that avoids considering exceptions.

Definition 4.1 (Differential privacy with dummy data)**.** Mechanism M satisfies (ϵ', δ')-differential privacy with dummy data (DPwD) if for any adjacent datasets D, D', and $R = \{R_1, \cdots, R_w\}$, the following is satisfied:

$$Pr[M(D \cup R) \in \mathcal{M}] \le e^{\epsilon'} \cdot Pr[M(D' \cup R) \in \mathcal{M}] + \delta'. \quad (9)$$

Similarly, we define global sensitivity with dummy data.

Definition 4.2 (Global sensitivity with dummy data)**.** Global sensitivity with dummy data GS_q^R to query $q : \mathcal{D} \to \mathbb{R}^k$ is defined as follows:

$$GS_q^R = \max_{\forall D,D':H(D,D')=1} \|q(D \cup R) - q(D' \cup R)\|_p. \quad (10)$$

Definitions 4.1 and 4.2 differ from previous definitions 3.2 and 3.5 in that dummy dataset $R = \{R_1, \cdots, R_w\}$ is added to D. In Definitions 4.1 and 4.2, dummy dataset R is fixed, and any adjacent datasets (except R) are considered. Conventional differential privacy mechanisms also satisfy differential privacy with dummy data. As a typical example, we consider the Laplacian mechanism.

Theorem 4.3. Let query $q : \mathcal{D} \to \mathbb{R}^k$, let $L(\mu, v)$ be Laplacian noise with mean μ and variance v, and then the following mechanism M_q^R satisfies $(\epsilon', 0)$-DPwD:

$$M_q^R(D \cup R) = q(D \cup R) + L\left(0, \frac{GS_q^R}{\epsilon'}\right). \quad (11)$$

Proof. Consider $D \cup R, D' \cup R$ for adjacent D and D'. Let the query be $q : \mathcal{D} \to \mathbb{R}^k$, and let p_x be the probability density function of $M_q^R(x)$. At this time, for any $z \in \mathbb{R}^k$, the following holds:

$$\frac{p_{(D\cup R)}(z)}{p_{(D'\cup R)}(z)} = \prod_{i=1}^{k} \frac{\exp(-\frac{\epsilon'|q(D\cup R)_i - z_i|}{GS_q^R})}{\exp(-\frac{\epsilon'|q(D'\cup R)_i - z_i|}{GS_q^R})} = \prod_{i=1}^{k} \exp\left(\frac{\epsilon'(|q(D' \cup R)_i - z_i| - |q(D \cup R)_i - z_i|)}{GS_q^R}\right) \le \prod_{i=1}^{k} \exp\left(\frac{\epsilon'|q(D' \cup R)_i - q(D \cup R)_i|}{GS_q^R}\right)$$

$$= \exp\left(\frac{\epsilon' \|q(D' \cup R) - q(D \cup R)\|_1}{GS_q^R}\right)$$
$$\leq \exp(\epsilon'). \quad (12)$$

Note that $\frac{p_{(D\cup R)}}{p_{(D'\cup R)}} \geq \exp(-\epsilon')$ is symmetric. □

Definitions 4.1 and 4.2 differ from previous definitions 3.2 and 3.5 only in that they cover any dataset $D \cup R$ except R. Since this difference does not affect the proof that the conventional Laplace mechanism satisfies differential privacy, Theorem 4.3 can also be easily proved. Note that, although Definition 4.1 does not guarantee the privacy of R, nor does it affect the privacy of D because R is dummy data and generated internally by a data processor. Furthermore, we can approximate the relationship with conventional differential privacy as follows when GS_q and GS_q^R are non-zero real numbers:

$$\begin{aligned} M_q^R(D \cup R) &= q(D \cup R) + L\left(0, \frac{GS_q^R}{\epsilon'}\right) \\ &= q(D \cup R) + L\left(0, \frac{GS_q}{GS_q \cdot \epsilon' / GS_q^R}\right) \\ &= q(D) + L\left(0, \frac{GS_q}{\epsilon}\right) + \omega, \end{aligned} \quad (13)$$

where ω is $q(D \cup R) - q(D)$, but R is generated internally by a data processor and contains no information. For example, consider a query for the average. Since the sensitivities of each data space are $GS_q = \frac{m}{n}$ and $GS_q^R = \frac{m}{n+w}$, $\epsilon' = \frac{n}{n+w} \cdot \epsilon$ is obtained. As for ω, $\omega = q(D \cup R) - q(D) = \frac{n\sum^w R_i - w\sum^n d_i}{n(n+w)}$ follows from $q(D \cup R) = \frac{\sum^n d_i + \sum^w R_i}{n+w}$ and $q(D) = \frac{\sum^n d_i}{n}$. Mechanisms that satisfy the proposed definition are realized by superposition of noise based on sensitivity and error introduced by dummy data, where the former noise can be expressed in terms of the parameters of differential privacy. Smooth sensitivity with dummy data can be defined similarly as Definitions 4.4 and 4.5.

Definition 4.4 (Local sensitivity with dummy data)**.** Given dataset $D \cup R$, local sensitivity LS_q^R with dummy data to query $q : \mathcal{D}^{|\chi|} \to \mathbb{R}^k$ is defined as follows:

$$LS_q^R(D \cup R) = \max_{\forall D' : H(D,D')=1} \|q(D \cup R) - q(D' \cup R)\|_p. \quad (14)$$

Definition 4.5 (Smooth sensitivity with dummy data)**.** Suppose $\beta > 0$, and dataset D is given. Smooth sensitivity S_q^R with dummy data to query $q : \mathcal{D} \to \mathbb{R}^k$ is defined as follows:

$$S_{q,\beta}^R(D \cup R) = \max_{anyD'} (LS_q^R(D' \cup R) \cdot e^{-\beta H(D,D')}). \quad (15)$$

Theorem 4.6. Let $q : \mathcal{D} \to \mathbb{R}^k$ be a query. When $\alpha = \epsilon' / \sqrt{\ln(1/\delta)}$ and $\beta = \Omega(\epsilon' / \sqrt{k \ln(1/\delta')})$, the following mechanism satisfies (ϵ', δ')-DPwD.

$$M_q(D \cup R) = q(D \cup R) + \frac{S_{q,\beta}^R(D' \cup R)}{\alpha} \cdot N(0, 1) \quad (16)$$

As in a previous work [17], Proposition 3.10 requires the consideration of a dataset such that $H(D, D') = k$. Since $H(D, D') = H(D \cup R, D' \cup R)$, adding dummy data has no effect on the proof of the theorem; the only difference is that record R in $D \cup R$ is fixed, guaranteeing the privacy of any data D. Although the privacy of R is not guaranteed, there is no effect on each datum because R is dummy data.

4.2 Correlation Coefficient Satisfying Differential Privacy with Dummy Data

As we have described, mechanisms that satisfy our proposed definition satisfy conventional differential privacy and can be compared with other differential privacy mechanisms using privacy parameters ϵ. Next, returning to the purpose for which we extended the definition, we consider a mechanism for queries that cannot be directly constructed with traditional differential privacy. As an example, we use Definitions 4.4 and 4.5 to calculate the correlation coefficient that satisfies differential privacy with dummy data.

In the following, values corresponding to D in D' are denoted by using $()'$. For example, the set corresponding to set A in D', i.e., $a_1 = (a_1)', \cdots, a_n = (a_n)', a_i \neq (a_i)'$, is denoted by $(A)'$ and the average value of $(A)'$ is denoted by $(\mu_A)'$. Even if the i-th record (a_i, b_i) and $((a_i)', (b_i)')$ are the different record in adjacent datasets, we will not lose generality and they are denoted as $\Delta a_i = (a_i)' - a_i, \Delta b_i = (b_i)' - b_i$. The following is the correlation coefficient of D':

$$\begin{aligned} (C)' &= \frac{(s_{AB})'}{(s_A)'(s_B)'} \\ &= \frac{h(\Delta a_i, \Delta b_i)}{\sqrt{\frac{1}{n} f(\Delta a_i)} \sqrt{\frac{1}{n} g(\Delta b_i)}}, \end{aligned} \quad (17)$$

where

$$\begin{aligned} f(\Delta a_i) &= \frac{n-1}{n}(\Delta a_i)^2 + 2(a_i - \mu_A)\Delta a_i \\ &\quad + \sum^n (a_j - \mu_A)^2 \\ &= C_n(\Delta a_n)^2 + 2C_a \Delta a_n + C_{aa}, \\ g(\Delta b_i) &= \frac{n-1}{n}(\Delta b_i)^2 + 2(b_i - \mu_B)\Delta a_i \\ &\quad + \sum^n (b_j - \mu_B)^2 \\ &= C_n(\Delta b_i)^2 + 2C_b \Delta b_i + C_{bb}, \\ h(\Delta a_i, \Delta b_i) &= \frac{n-1}{n}\Delta a_i \Delta b_i + (b_i - \mu_B)\Delta a_i \\ &\quad + (a_i - \mu_A)\Delta b_i + \sum^n (a_j - \mu_A)(b_j - \mu_B) \\ &= C_n \Delta a_i \Delta b_i + C_b \Delta a_i + C_a \Delta b_i + C_{ab}. \end{aligned} \quad (18)$$

We then obtain the following theorem.

Theorem 4.7. Let C be the correlation coefficient of dataset

$D \in [0,m]^2$ with two attributes, A and B, let C' be the correlation coefficient of an adjacent dataset. Then we obtain $\arg\max_{r_i=(a_i,b_i)} |C'-C| \in \left\{(0,m),(m,m),(0,0),(\frac{D_3D_5-D_2D_6}{D_1D_5-D_2D_4},\frac{D_3D_4-D_1D_6}{D_2D_4-D_1D_5})\right\}$. Here, $D_1 = C_aC_b - C_nC_{ab}$, $D_2 = C_nC_{aa} - C_a^2$, $D_3 = C_aC_{ab} - C_{aa}C_b$, $D_4 = C_nC_{bb} - C_b^2$, $D_5 = C_aC_b - C_nC_{ab}$, and $D_6 = C_bC_{ab} - C_{bb}C_a$.

Proof. $f(\Delta a_i), g(\Delta b_i)$ are downward convex quadratic functions, and the denominator is minimized when $f(-\frac{n}{n-1}(a_i - \mu_A)) = \sum_{j=1}^{n}(a_j-\mu_A)^2 - \frac{n}{n-1}(a_i-\mu_A)^2$, $g(-\frac{n}{n-1}(b_i-\mu_B)) = \sum_{j=1}^{n}(b_j-\mu_B)^2 - \frac{n}{n-1}(b_i-\mu_B)^2$. Since C is not affected by $\Delta a_i, \Delta b_i$, to maximize $|C-C'|$, $\Delta a_i, \Delta b_i$ either maximizes or minimizes C'.

In the following, we consider a dataset such that both D, D' contain at least one different data set, i.e., the denominator of C, C' is non-zero. The following is the partial differentiation of C' by Δa_i:

$$\frac{d}{d\Delta a_i}C' = \frac{(C_aC_b - C_nC_{ab})\Delta a_i + (C_nC_{aa} - C_a^2)\Delta b_i + (C_{aa}C_b - C_aC_{ab})}{1/n \cdot g(\Delta b_i)^{1/2} \cdot f(\Delta a_i)^{3/2}}. \tag{19}$$

Similarly, a partial differentiation of C' by Δb_i yields

$$\frac{d}{d\Delta b_i}C' = \frac{(C_aC_b - C_nC_{ab})\Delta b_i + (C_nC_{bb} - C_b^2)\Delta a_i + (C_{bb}C_a - C_bC_{ab})}{1/n \cdot f(\Delta a_i)^{1/2} \cdot g(\Delta b_i)^{3/2}}. \tag{20}$$

From Eqs. (19) and (20), C' can take an extreme value at $\frac{d}{d\Delta a_i}C' = \frac{d}{d\Delta b_i}C' = 0$, where $(\Delta a_i, \Delta b_i) = (\frac{D_3D_5-D_2D_6}{D_1D_5-D_2D_4}, \frac{D_3D_4-D_1D_6}{D_2D_4-D_1D_5})$. Here, $D_1 = C_aC_b - C_nC_{ab}, D_2 = C_nC_{aa} - C_a^2, D_3 = C_aC_{ab} - C_{aa}C_b, D_4 = C_nC_{bb} - C_b^2, D_5 = C_aC_b - C_nC_{ab}$, and $D_6 = C_bC_{ab} - C_{bb}C_a$.

In addition, since $\Delta a_i, \Delta b_i \in [-m, m]$, and r' monotonically increases or decreases with respect to the other variable when either Δa_i or Δb_i is fixed, $|C'-C|$ may be maximum when Δa_i and Δb_i are maximum or minimum, respectively. Therefore, Theorem 4.7 holds. □

We now consider the actual algorithm for deriving privacy-enhancing correlation coefficients. The algorithm requires the local sensitivity of D_j. Let C_j and C'_j be the correlation coefficient for D_j and an adjacent dataset of D_j. Fixing the i-th record (a_i, b_i) of D_j, we can easily find the adjacent dataset, where $\|C'_j - C_j\|$ is maximum from Theorem 4.7. Specifically, assume an adjacency data set D_{j+1} with (a_i, b_i) converted to extreme values or values at the edge of the domain of definition. Since the local sensitivity is the maximum value of $\|C'_j - C_j\|$, we consider the adjacent dataset with the maximum $\|C'_j - C_j\|$ for each (a_i, b_i) and generate the adjacent dataset with the largest difference among them as D_{j+1}. This process can be performed recursively to obtain $S^r_{q,\beta}(D_0)$, and privacy can be assured by adding noise to the $q(D)$ result.

Algorithm 1 $CCDPwD(D_0, \epsilon', \delta', R_1, R_2(\neq R_1))$: Correlation coefficient achieving differential privacy with dummy data.

Input: Dataset $D = \{d_1, \cdots d_n\}$, privacy parameter ϵ', and random records R_1, R_2.

```
D_0 ← D_0 ∪ {R_1, R_2}
for j = 0; j < n; j++ do
    for i = 0; i < n; i++ do
        Calculate d'_i = {a'_i, b'_i} = arg max_{a'_i,b'_i} ||C'_j − C_j||, where C'_j and C_j are
        correlation coefficients obtained from an adjacent dataset D_{j+1}
        and D_j
    end for
    d' = max d'_i
    Generate D_{j+1} with different d' = {a', b'} from D_j
    Calculate LS^r_q(D_{j+1}) · e^{−βH(D_0,D_{j+1})}
end for
S^r_{q,β}(D_0) = max(LS^r_q(D_{j+1}) · e^{−βH(D_0,D_{j+1})})
return M^r_q(D_0) = q(D_0) + (S^r_{q,β}(D_0) / α) · N(0,1)
```

5. Experimentation

In this section, we compare the correlation coefficients derived by the proposed method with the privacy-protected correlation coefficients based on the direct method. The following experiments were conducted on a data set ($C \approx 1.0$ for ease of explanation) following a normal distribution with n data and an $m(a_i, b_i \in [0, m])$ data domain.

5.1 Correlation Coefficients Satisfying Differential Privacy Using Direct Methods

As described in Sect. 4.1, computing sensitivity is impossible using the correlation coefficient as a query. Thus, it is difficult to construct a probabilistic disturbing mechanism to find the correlation coefficient that directly satisfies differential privacy. On the other hand, with the sequential theorem, a probabilistic disturbing mechanism can be constructed to obtain the covariance and standard deviation that satisfy the differential privacy, and to calculate the correlation coefficient that satisfies the differential privacy from each result. The privacy parameter can be obtained by summing the privacy parameters. The covariance and standard deviation are obtained by $s_{AB} = h(\Delta a_i, \Delta b_i)$, $s_A = \sqrt{f(\Delta a_i)/n}$, $s_B = \sqrt{g(\Delta b_i)/n}$. Therefore, considering the application of the smooth sensitivity framework, local sensitivity is maximized when each is extreme or when $\Delta a_i, \Delta b_i$ is maximum or minimum. Now we can construct a mechanism that derives a correlation coefficient that satisfies differential privacy.

We set $(n, m) = (100, 100)$, fix privacy parameter $\delta' = 0.01$, vary ϵ', and evaluate the noise magnitude. The evaluation is expressed in quartiles of 100 runs. Privacy parameters ϵ', δ' are allocated equally; if $(\epsilon', \delta') = (1, 1/100)$, the

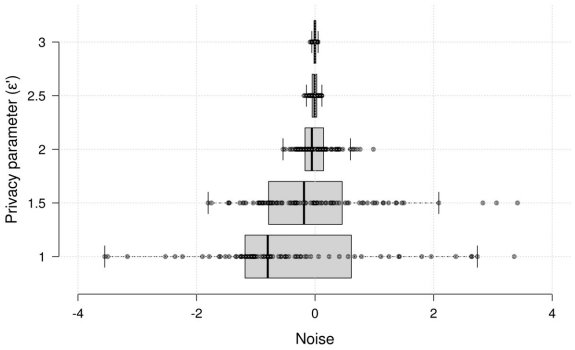


Fig. 1 Noise distribution with sequential theorem: $(m, n) = (100, 100)$

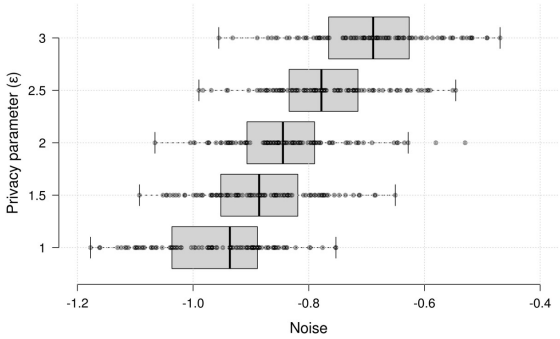


Fig. 2 Noise distribution with LDP: $(m, n) = (100, 100)$

privacy parameters for each s_{AB}, s_A, and s_B are $(\epsilon'_{s_{AB}}, \delta'_{s_{AB}}) = (\epsilon'_{s_a}, \delta'_{s_a}) = (\epsilon'_{s_b}, \delta'_{s_b}) = (1/3, 1/300)$ and the smooth sensitivity framework was applied. The experimental results are summarized in Fig. 1. The results show that when the correlation coefficient is calculated using the sequential theorem, noise that satisfies differential privacy is added to each of the three variables of standard deviation and covariance, amplifying their effects, and the overall noise also tends to increase. In the case of $(m, n) = (100, 100)$, $\epsilon' = 2.5$, the upper bound of noise is about 0.2, indicating that $\epsilon' \geq 3$ is required to obtain sufficient accuracy.anism.

Another way to obtain privacy-preserving correlation coefficients is to apply a local differential privacy mechanism. Although differential privacy assumes dataset D' adjacent to $D = \{d_1, \cdots, d_n\}$, local differential privacy assumes arbitrarily different data d'_i for each datum d_i. Given an ϵ-local differential privacy mechanism M and a dataset $D'' = \left\{d'_1, \cdots, d'_n\right\}$, $Pr[M(D)] \leq e^{\epsilon} \cdot Pr[M(D')] \leq e^{\epsilon} \cdot Pr[M(D'')]$ holds. Thus M satisfies ϵ-differential privacy. When using the local difference privacy mechanism, noise is added to each record, which adds more noise than necessary, and ways to maintain its utility must be devised. For example, a local difference privacy mechanism was proposed [6] for a t-test and maintains its accuracy by indirectly performing a t-test using data distribution characteristics after noise was added. Since no local differential privacy mechanism has been proposed that can efficiently derive correlation coefficients, we evaluated them using data to which an existing local differential privacy mechanism [11] was applied. Here, we varied privacy parameter ϵ and evaluated the magnitude of the noise. The evaluation is expressed in quartiles for 100 runs. The experimental results are summarized in Fig. 2. Our experimental results show that evaluating the correlation coefficient is very difficult using a local differential privacy mechanism by simply applying the existing mechanism. Since the dataset used in this study is $C \approx 1.0$, applying the local differential privacy mechanism almost eliminates the correlation. Although the noise can be reduced by increasing the privacy parameter, the results are impractical even at $\epsilon = 3.0$, indicating that special processing, as in a previous study [6], should be paid to the correlation coefficient when a local differential privacy mechanism is used.

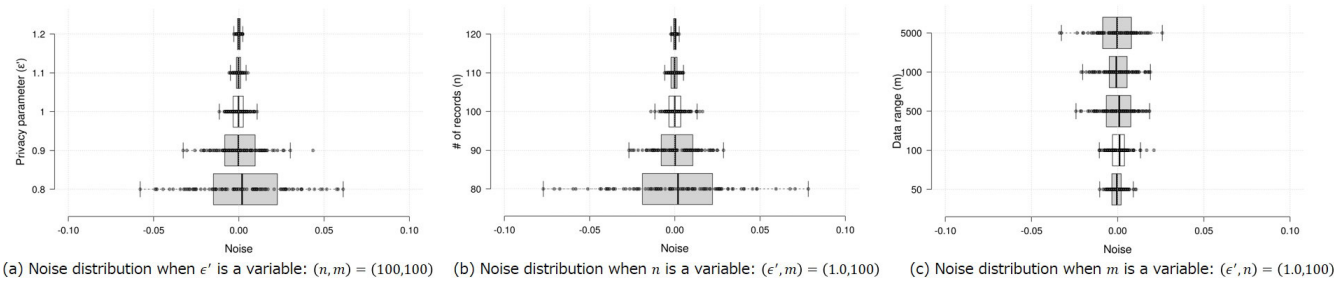
5.2 Proposal

We evaluate our proposed method. The following parameters might affect the output results: privacy parameter (ϵ', δ'), the number of data n, and the data definition domain m. In the experiment, each parameter is varied, and noise magnitude $\frac{S^{r}_{q,\beta}(D_0)}{\alpha} \cdot N(0, 1)$ is evaluated. The magnitude of the noise is expressed as a quartile for 100 runs. For comparison, the experimental results under the same conditions are represented by white bars.

Figure 3 (a) shows a case where the number of data is $n = 100$, the data definition range is $m = 100(a_i, b_i \in [0, m])$, privacy parameter $\delta' = 0.01$ is fixed, and privacy parameter ϵ' is variable. Naturally, the larger the privacy parameter is, the smaller the noise becomes. When $\epsilon' \geq 0.8$, the noise is at most ± 0.1, and correlation can be obtained relatively accurately. Figure 3 (b) shows the results when the privacy parameters $(\epsilon', \delta') = (1.0, 0.01)$, data domain $m = 100$ are fixed, and the number of data n is a variable. As the number of data increases, the impact of a single datum on the correlation coefficient decreases. Therefore, under identical privacy parameters and data domain, noise decreases as the number of data increases. Experimental results show that for $n \geq 80$, the noise is at most ± 0.1, and correlations can be obtained accurately even with relatively small samples.

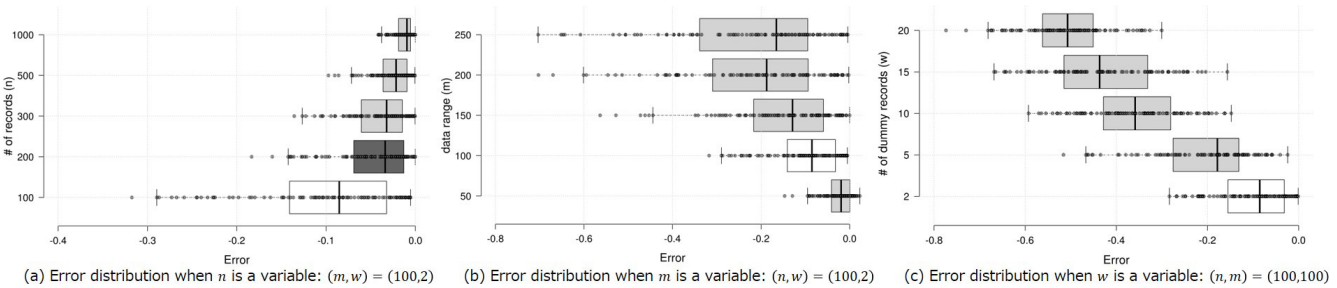
Finally, Fig. 3 (c) shows the results when privacy parameters $(\epsilon', \delta') = (1.0, 0.01)$, the number of data $n = 100$ are fixed, and data domain m is a variable. Data domain m affects the noise, although not as significantly as ϵ', n. This may be due to the fact that a'_i and b'_i can change to 0 and m, which are respectively the minimum and maximum values of the domain, during the construction of an adjacent dataset, and thus the data domain affects the magnitude of the noise. The results also show that many of the $r' = (a'_i, b'_i)$ replaced in constructing the adjacency datasets are outliers. Therefore, although this experiment was conducted on random data following a normal distribution, the magnitude of the noise is not easily affected by the bias or the characteristics of the dataset. This trend is similar for any dataset.

In our proposed algorithm, a noise ω is added by adding dummy data. Our proposal allows for the calculation of sensitivity by adding dummy data to an arbitrary data set,



(a) Noise distribution when ϵ' is a variable: $(n, m) = (100,100)$ (b) Noise distribution when n is a variable: $(\epsilon', m) = (1.0,100)$ (c) Noise distribution when m is a variable: $(\epsilon', n) = (1.0,100)$

Fig. 3 Noise distribution of the proposal algorithm



(a) Error distribution when n is a variable: $(m, w) = (100,2)$ (b) Error distribution when m is a variable: $(n, w) = (100,2)$ (c) Error distribution when w is a variable: $(n, m) = (100,100)$

Fig. 4 Error distribution of the proposal algorithm

which is accomplished by adding at least two dummy data with different values. The impact of dumy data on the correlation coefficient, as in the mean example, ω is independent of the privacy parameter ϵ' and becomes smaller as the number of reocrds n increases and larger as the data range m increases. Furthermore, the effect of the number of dummy data w depends on the distributions of the dummy data and the actual dataset, and the closer the distributions are, the less effect w has on ω. In this experiment, for ease of understanding the results, we deal with a dataset that follows a normal distribution with a correlation of almost 1. On the other hand, we did not have access to the dataset in the algorithm, and dummy data with a uniform distribution were given to the dataset. Figures 4 show the error due to dummy data on the correlation, which depends on the number of dummy data n, m, w. This error is caused by adding dummy data and does not depend on the privacy parameter. For the correlation coefficient, the number of dummy data required to obtain the sensitivity is $w = 2$, and considering the case $n = 100$, the impact of dummy data is at most 0.3. Therefore, the proposed method performs better than the existing methods even if the error due to dummy data and the noise that satisfies the differential privacy with dummy data are taken into account.

5.3 Improvement

The previous experiments showed that the proposed method has a greater impact of errors due to dummy data compared to noise to satisfy differential privacy. This is due to the fact that the data handled in the experiments have very high correlations, whereas the dummy data added by the proposed algorithm are uniformly distributed records, and their respective distributions are very different. In particular, since the dataset follows a normal distribution, values at the edge of the data range, i.e., close to m, are rarely available, and when m is large, uniformly distributed dummy data taking values close to m will result in a large error. Matching the distribution of the dummy data to the distribution of the dataset reduces the impact of the dummy data, but accessing the dataset is a use of the information. Therefore, to reduce the error of dummy data, Propositions 3.3 and 3.4 can be applied. Specifically, as a preprocessing step for the proposed algorithm, a simple differential privacy mechanism is applied to obtain the distribution of the dataset, and then the proposed algorithm is applied with dummy data following the obtained distribution. We compared the total noise of simply applying the proposed mechanism with that of using the improved mechanism. Figure 5 shows the results of the experiment with $n = 200$, $m = 100$, and $\epsilon' = \epsilon'_1 = \epsilon'_2 = 1.0$. ϵ'_1 and ϵ'_2 are privacy parameters for the preprocessing mechanism and the proposed mechanism. The preprocessing mechanism and the proposed mechanism

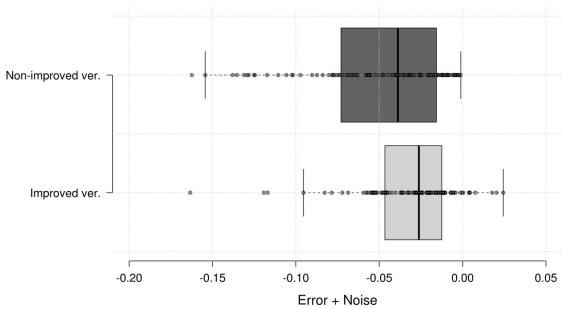


Fig. 5 Effect of assuming dummy data distribution

are performed on D_1 and D_2. Here, D_1 and D_2 are datasets where D is divided into 100 records each. The experimental result shows that the error is reduced by making the distribution of the dummy data closer to the distribution of the data set, and that the improved version has less total noise. Note that the experiment of the non-improve version is under the same condition as the black bar in Fig. 4 (a).

6. Conclusion

In this paper, we defined differential privacy with dummy data as an extension of differential privacy. Mechanisms that satisfy the proposed definition can be constructed simply by adding dummy data to the conventional differential privacy mechanism, and in addition, if the sensitivity in the conventional definition can be calculated, the relationship to the sensitivity in the extended definition can be expressed. We further took the correlation coefficient as an example of a mechanism that is difficult to construct according to the conventional definition, and described how to construct the mechanism. For queries for which it is difficult to directly determine the sensitivity, the privacy can be guaranteed by synthesizing differential privacy mechanisms, but the accuracy is significantly degraded by the superimposition of noise. On the other hand, our proposed method can guarantee privacy with less noise than the composition of differential privacy mechanisms, even after taking into account the additional error due to the addition of dummy data. Furthermore, we provided additional suggestions to lower the impact of errors and showed that privacy can be guaranteed more effectively when the number of data is sufficient.

References

[1] C. Dwork, "Differential privacy," Proc. ICALP 2006, LNCS, vol.4052, pp.1–12, 2006.
[2] J.C. Duchi, M.I. Jordan, and M.J. Wainwright, "Local privacy and statistical minimax rates," 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pp.429–438, IEEE, 2013.
[3] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," Proc. 2014 ACM SIGSAC conference on computer and communications security, pp.1054–1067, 2014.
[4] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," Proc. 2018 International Conference on Management of Data, pp.1655–1658, 2018.
[5] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," Proc. 2010 ACM SIGMOD International Conference on Management of data, pp.735–746, 2010.
[6] B. Ding, H. Nori, P. Li, and J. Allen, "Comparing population means under local differential privacy: with significance and power," Proc. AAAI Conference on Artificial Intelligence, vol.32, no.1, 2018.
[7] F. Yu, S.E. Fienberg, A.B. Slavković, and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies," Journal of biomedical informatics, vol.50, pp.133–141, 2014.
[8] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional mechanism: regression analysis under differential privacy," arXiv preprint arXiv:1208.0219, 2012.
[9] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong, "Differentially private naive bayes classification," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pp.571–576, IEEE, 2013.
[10] K. Wei, J. Li, M. Ding, C. Ma, H.H. Yang, F. Farokhi, S. Jin, T.Q.S. Quek, and H.V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," IEEE Trans. Inf. Forensics Security, vol.15, pp.3454–3469, 2020.
[11] N. Wang, X. Xiao, Y. Yang, J. Zhao, S.C. Hui, H. Shin, J. Shin, and G. Yu, "Collecting and analyzing multidimensional data with local differential privacy," 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp.638–649, IEEE, 2019.
[12] D. Desfontaines and B. Pejó, "Sok: Differential privacies," Proceedings on Privacy Enhancing Technologies, vol.2020, no.2, pp.288–313, 2020.
[13] T. Zhang, T. Zhu, R. Liu, and W. Zhou, "Correlated data in differential privacy: definition and analysis," Concurrency and Computation: Practice and Experience, vol.34, no.16, e6015, 2022.
[14] F.D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," Proc. 2009 ACM SIGMOD International Conference on Management of data, pp.19–30, 2009.
[15] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," Annual International Conference on the Theory and Applications of Cryptographic Techniques, vol.4004, pp.486–503, Springer, 2006.
[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," Theory of cryptography conference, vol.3876, pp.265–284, Springer, 2006.
[17] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," Proc. thirty-ninth annual ACM symposium on Theory of computing, pp.75–84, 2007.

Tomoaki Mimoto received B.E. and Ph.D degree from Osaka university, Japan, in 2012 and 2022, and received M.E. (Outstanding Performance Award) in information science from Japan Advanced Institute of Science and Technology in 2014. He joined KDDI in 2014, and was with KDDI research, Inc. from 2015 to 2020. He is currently an research engineer in Advanced Telecommunication Research Institute International (ATR).

Hiroyuki Yokoyama received B.E., M.E., and Ph.D. degrees in Electrical Engineering from Kyoto University, Kyoto, Japan, in 1992, 1994, and 2006, respectively. He joined the Research and Development Laboratories of Kokusai Denshin Denwa Company Ltd. (currently KDDI Corp.) in 1994 and has engaged in research on communications network planning, sensor data mining and its applications for mobile phones. He has been leading 5G related research projects as a director of the Adaptive Communication Research Laboratories in Advanced Telecommunications Research Institute International since 2018. He received the best paper awards of IPS in 2012 and IEEE CCWC in 2022.

Toru Nakamura is a research engineer in the Information Security Laboratory at KDDI Research, Inc. He received the B.E., M.E., and Ph.D degree from Kyushu University, in 2006, 2008, and 2011, respectively. In 2011, he joined KDDI and in the same year he moved to KDDI R&D Laboratories, Inc. (currently renamed KDDI Research, Inc.). In 2018, he moved to Advanced Telecommunications Research Institute International(ATR). Since 2020, he is a researcher in KDDI Research, Inc. again. He received CSS2016 SPT Best Paper Award. His current research interests include security, privacy, and trust, especially privacy enhanced technology and analysis of privacy attitudes. He is a member of IEICE and IPSJ.

Takamasa Isohara received his B.E. and M.E. degree in Information and Computer Science from Keio University, Japan, in 2005 and 2007, respectively. He joined KDDI and has been engaged in the research on network security, smartphone security, SIM-based IoT security, and Automotive cybersecurity. He is currently a senior manager at the Usable Trust Laboratory of KDDI Research, Inc. He received the IPSJ Kiyasu Special Industrial Achievement Award in 2011. He is a member of IEICE and IPSJ.

Masayuki Hashimoto received the B.E., M.E. and Ph. D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1995, 1997 and 2007, respectively. He received the M.B.A. from Graduate School of Management, GLOBIS University, Tokyo, Japan, in 2017. After he joined KDDI R&D Laboratories in 1997, he engaged research on digital image transmission, developed a mobile medical image transmission system and commercialized the system in medical fields in Japan. Since 2019, he has been a head of the Department of Advanced Security, Adaptive Communication Research Laboratories, Advanced Telecommunications Research Institute International. His research interests include data privacy, vulnerability-information extraction, group signature and supply chain security.

Ryosuke Kojima received a B.E. in computer science in 2012, an M.E. in information science and engineering in 2014, and a Ph.D. in information science and engineering in 2017, all from the Tokyo Institute of technology. He is currently a lecturer in Kyoto University. His research interests include AI, machine learning, and medical AI. He is a member of the JSAI, RSJ.

Aki Hasegawa recieved his B.E. and M.E. degrees in information engineering from the University of Electro-Communicstions in 1995 and 1997, respectively. From April 1997 to September 2000, He was a researcher in a project of JST ERATO. From October 2000 to March 2020, He worked at RIKEN. He is currently a program-specific researcher at the Graduate School of Medicine, Kyoto University. His current research interests include database development and chemoinformatics.

Yasushi Okuno 2021: Director, HPC- and AI-driven Drug Development Platform Division, Unit Leader, Biomedical Computational Intelligence Unit, Unit Leader, AI-driven Drug Discovery Collaborative Unit, RIKEN R-CCS (-present) Representative Director, Life Intelligence Consortium (-present). 2016: Professor, Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University (-present) 2014: Program-Specific Professor (Endowed Chair), Graduate School of Medicine, Kyoto University. 2008: Professor (Endowed Chair), 2006: Associate Professor, 2003: Program-Specific Assistant Professor, Graduate School of Pharmaceutical Sciences, Kyoto University. 2000: Ph.D. in Pharmaceutical Sciences, Kyoto University