IEEE Open Journal of the
Computer Society

Received 19 April 2022; accepted 16 May 2022. Date of publication 20 May 2022;
date of current version 2 June 2022. The review of this paper was arranged by Associate Editor Zhi Liu.

Digital Object Identifier 10.1109/OJCS.2022.3175999

Re-Identification in Differentially Private Incomplete Datasets

YUICHI SEI[1,2] (Member, IEEE), HIROSHI OKUMURA[3], AND AKIHIKO OHSUGA[1] (Member, IEEE)

[1]University of Electro-Communications, Tokyo 182-8585, Japan
[2]JST, PRESTO, Kawaguchi, Saitama 332-0012, Japan
[3]Mitsubishi Research Institute, Tokyo 100-0014, Japan

CORRESPONDING AUTHOR: YUICHI SEI (e-mail: seiuny@uec.ac.jp)

This work was supported in part by JSPS KAKENHI under Grants JP18H03229, JP18H03340, JP18K19835, JP19K12107, JP19H04113, and JP21H03496, and in part by JST, PRESTO under Grant JPMJPR1934.
This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by The University of Electro-Communications under Application No. 19066, and performed in line with the Declaration of Helsinki.

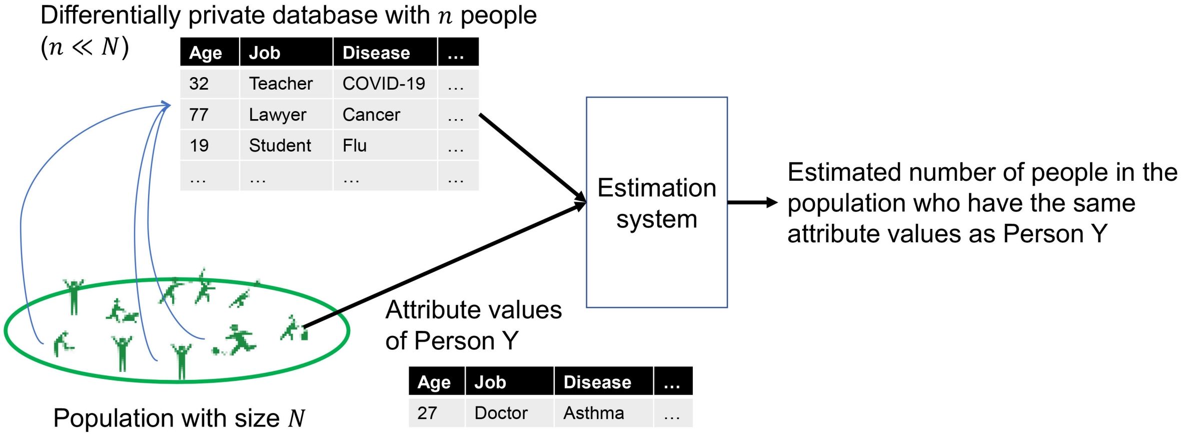
ABSTRACT Efforts to counter COVID-19 reaffirmed the importance of rich medical, behavioral, and sociological data. To make data available to many researchers who can conduct statistical analyses and machine learning, personally identifiable information must be excluded to protect individual privacy. It is essential to remove explicit identifiers, sample population data, and apply differential privacy, the de facto standard privacy metric. Despite the general belief that the risk of re-identification is insignificant when these techniques are applied, this study shows that even after applying these techniques, the risk of being re-identified is highly significant for some data. This study proposes in detail an algorithm for estimating the number of people in a population who have certain attribute values based on incomplete, differentially private databases. If the estimated number is one, the probability that only one person with that attribute value is present in the population is high, which means that there is a high probability of re-identification. Therefore, this study concludes that the re-identification risk must be evaluated even after applying state-of-the-art techniques to protect privacy.

INDEX TERMS Differential privacy, ethical and privacy framework, re-identification.

I. INTRODUCTION

Analyses of people's medical, behavioral, and sociological data are essential for understanding the pandemic situation and devising remedial measures [1]. Betsch *et al.*, for example, evaluated continuous data from approximately 7000 people to determine the impact of governmental policies concerning COVID-19 on people's compliance and mask-wearing behavior [2]. Grouping participants were analyzed according to age, sex, and beliefs about governmental policies. Swayamsiddha *et al.* surveyed multiple Internet-of-Things (IoT) healthcare services to address COVID-19. These healthcare services were designed to monitor each individual's health state, but they can also be used to plan long-term countermeasures against pandemics using statistical analyses and machine learning based on personal medical data [3]. Drefahl *et al.* analyzed data on COVID-19 deaths recorded in Sweden. They discovered that being male, having less disposable income, having a lower level of education, not being married, and being an immigrant from a low- or middle-income country were independently associated with a potentially higher risk of death from COVID-19 [4]. Ray *et al.* proposed a data-sharing marketplace to connect data owners and buyers. The marketplace data were exchanged based on the data value and the data buyer's reliability. Data owners can directly control their data and make data-sharing decisions based on risks and compensation [5]. These studies appropriately handled the personal data. However, there is always the risk of cases where personal data are overprotected, and the people to whom data is provided are relatively limited, or conversely, events in which personal data are not adequately protected.

To protect individual privacy, it is required to eliminate personally identifiable data to make the data available to many


Differentially private database with n people
($n \ll N$)
Estimation system
Estimated number of people in the population who have the same attribute values as Person Y
Attribute values of Person Y
Population with size N

Age	Job	Disease	…
32	Teacher	COVID-19	…
77	Lawyer	Cancer	…
19	Student	Flu	…
…	…	…	…

Age	Job	Disease	…
27	Doctor	Asthma	…

FIGURE 1. Target scenario of this research. There may be more than one database. Person Y's information may or may not be contained in the differentially private database.

researchers capable of conducting statistical analyses and machine learning. It is essential to remove explicit identifiers, sample population data, and apply differential privacy, which is the de facto standard privacy metric used by companies such as Google, Apple, and Microsoft [6]–[8].

However, there has been no verification of the success of the data re-identification after applying the processes. This study shows that a person may be correctly re-identified with 20%–60% accuracy, depending on the dataset and experimental setup.

For instance, consider the following scenario. A city has established an IoT environment to collect personal data for COVID-19 measures. Assuming that Alice's set of attribute values (the attributes include estimated age, height, weight, and COVID-19 infection status) is unique among the city citizens, the collection of the attribute value data of Alice uniquely identifies her. Therefore, if anonymity is required, it is desirable to discard Alice's data. In this scenario, we Image deleted. Please check. assume the city holds a dataset of several citizens. This study builds a system that can predict the number of citizens who have X as their attribute value with high accuracy, given a person's attribute value data, regardless of whether the person exists in the database (Fig. 1).

Such a system is important in terms of privacy and determining city policies. To make policy decisions, it is necessary to know the number of people with certain attributes in the entire population. In some cases, such as census data, the data are collected from a large portion of the population. If it were possible to collect data from everyone, there would be no need for a "system to predict the overall number of people," as the system does. However, in non-mandatory surveys at the local government level or voluntary surveys conducted by companies, the data acquisition rate of the population may be less than a small percentage. In such instances, estimating the number of people with certain attributes in the total population is difficult. However, this is a critical question.

This study first examines the above problem regarding differential privacy and incomplete multiple databases. The experimental results show that the possibility of being randomly identified is high even with differential privacy. It also shows that the possibility of being identified increases when multiple databases are considered.

The main contributions of this article are as follows.

1) We proposed a new algorithm to re-identify individuals from incomplete datasets collected based on differential privacy, a de facto standard privacy metric.
2) Using six datasets, we showed that the proposed algorithm can identify individuals with higher accuracy than existing methods.
3) We conducted a smart house experiment on 20 subjects for 2 weeks to 2 months to collect IoT data containing private information and made some of these data publicly available.
4) We analyzed the information of the 20 subjects after protecting it with differential privacy. We showed that there is a 79% probability that one subject's information is the only one protected among 10 million people in their 20s and 30s in the Japan Kanto region. Furthermore, the results showed that subjects may be re-identified even if protected with differential privacy.

The main technical challenges are 1) generating an estimation model from differentially private categorical attributes, 2) estimating the existence probability using the generated model, ensuring it does not make the probability of the presence of at least one person too small, and 3) modifying the estimated probabilities using the information on the number of people present in each table.

The remainder of the article is structured as follows: Section II explains three application scenarios of the proposed system. Section III reviews some related work. Section IV presents the proposed system that estimates the number of people with the same attribute value existing in the population. Section V presents the evaluations based on real-world datasets and their results. Section VI addresses the techniques that were not employed in this study. Finally, Section VII presents the conclusion.

II. SCENARIOS

The three example scenarios covered by this study are as follows:

Scenario I: Using attributive values, City A attempted to publish Bob's documented COVID-19 infection experience. Bob is in his 30 s, lives in Town G in City A, is male, has lung disease, and is infected with COVID-19. City A does not know whether other people in City A have the same attribute values. If there is a high probability that more than ten other people have the same attribute values as Bob, City A will publish this article with Bob's attribute values.

Scenario II: Company B collected anonymized personal data from an IoT system. Suppose the anonymized data sample can be linked to a specific person with a high probability. In that case, Company B will delete the data in response to a legal request, such as the general data protection regulation.

Scenario III: Town C has personal data for all its residents. To use machine learning techniques to reveal the relationships between personal attributes and annual income, Town C will extract several data samples from the dataset, remove identifiers, and provide the data to a research institute it undergoes a private treatment using differential privacy. Town C wants to know if any sample still had the potential to identify an individual uniquely.

Therefore, it is essential to estimate the number of people with certain attribute values.

III. RELATED WORK

A. ESTIMATING RE-IDENTIFICATION SUCCESS

Rocher *et al.* [1] demonstrated that the risk of re-identification is high even after removing explicit identifiers and sampling population data. However, their study did not consider differential privacy, an essential technique in the privacy field. Currently, differential privacy is the most actively researched privacy-protection technology, and many companies have also been widely used, such as Apple and Google. However, it may be incorrect to assume that differential privacy technology can solve the problem of the re-identification of sample data. Their study also assumed a situation where there was only one database and did not consider the risk of multiple databases.

In this article, we focus on re-identification from personal attributes. However, several studies on re-identification from images have also been conducted. For example, Lin *et al.* proposed a method for re-identifying pedestrians from their videos [9]. They integrated the learning of attribute values by building an attribute–person recognition network to improve the re-identification accuracy. Large amounts of labeled data are needed to re-identify people from images. To address this problem, Wang *et al.* proposed a deep learning algorithm that can transfer labeled training data from the target region to a new region for the re-identification task without collecting new labeled data [10]. Zhou *et al.* proposed a system that can identify a given pedestrian from a network of surveillance cameras [11]. Their method is based on adaptive graph learning and can be used for unsupervised machine learning. Using four datasets, they demonstrated that their method outperformed previous studies. Although image-based re-identification is beyond the scope of this article, research on re-identification after applying differential privacy to images is considered a future issue.

Information on human behavior recognition obtained from cameras and sensing data can be used as input for the proposed method. For example, Luo *et al.* proposed a video semantic recognition system [12]. Their method can accurately perform semantic recognition of video images using semi-supervised feature analysis, even when labeled data are scarce. Chen *et al.* proposed a human activity recognition system [13]. Their method can perform human behavior recognition from sensing data with high accuracy, even in environments with unbalanced and scarce labeled data. Person behavior estimated in this way may be used for person re-identification. This article does not detail how to collect personal attributes but assumes that personal attributes can be obtained with a high degree of accuracy as with these methods. Differential privacy techniques are applied to the information thus obtained to protect privacy.

B. DIFFERENTIAL PRIVACY

Differential privacy has been widely studied in the recent decade for privacy-preserving data mining [14]. We assume a scenario where we collect attribute values from each person while protecting their privacy based on differential privacy. Differential privacy may also be referred to as local differential privacy in this scenario.

For simplicity, we assume that the number of attributes is one. However, the same process can be applied to multiple attributes. Additionally, we assumed that the target attribute was categorical. If the attribute is numerical, it is categorized, and this assumption is general [15].

Definition [ϵ-differential privacy]

Let V be the set of possible values of the attribute value, v and v' be elements of V, and ϵ be a positive real number. A randomized mechanism A satisfies ϵ-differential privacy if and only if the following equation holds for any output y

$$P\left(A\left(v\right) = y\right) \leq e^{\epsilon} P\left(A\left(v'\right) = y\right) \text{ for all } v,\ v'. \qquad (1)$$

Here, the definition assumes that the number of attributes is one. However, we assumed that the number of attributes is more than one in this study. When the number of attributes is more than one, we can consider the set of attribute values as one attribute value.

Let f be the number of categories, i.e., $f = |V|$. For example, consider that there are two attributes: gender and age. Consider that the possible values of gender are {male, female}, and the possible values of age are {0 s, 10 s, ..., 90 s}. In this case, $V = \{[\text{male}, 0\text{s}], [\text{male}, 10\text{s}], \ldots, [\text{female}, 80\text{s}], [\text{female}, 90\text{s}]\}$ and $f = 2 \cdot 10 = 20$.

Each person provides the correct attribute value to the data collector with a probability β. Moreover, with probability

$(1-\beta)$, each person provides the data collector with an attribute value randomly selected from all other attribute values. Then, ϵ-differential privacy can be guaranteed if β satisfies the following equation [16]:

$$\beta = \frac{e^{\epsilon}}{-1+e^{\epsilon}+f} \tag{2}$$

This is the simplest approach to realizing differential privacy. Section VI discusses what happens when we use more sophisticated techniques to achieve differential privacy. Generally, the more sophisticated and advanced the technology, the greater the risk of being personally identified.

C. SECURE SHARING OF SENSITIVE DATA

Lian *et al.* proposed a blockchain platform to securely allow only legitimate users to access COVID-19 electronic medical records. This can prevent tampering by malicious users and keep the communication and storage overheads small [17]. Such technologies are designed to prevent malicious users from seeing sensitive data, whereas legitimate users are shown the data as they are. Thus, system users can identify individuals from the data. This technology is necessary for treating individual patients. However, when the data are to be used by researchers who want to perform statistical analyses or machine learning, we should restrict the transmission of personally identifiable information completely.

D. SECURE COMPUTATION OF SENSITIVE DATA

Secure multi-party computation (MPC) allows each organization holding personal data to perform statistical calculations on all data without disclosing each data [18], [19]. MPC had the disadvantage of being computationally time-consuming; however, its performance has been recently improved. MPC can be used for statistical analysis and machine learning. Knott *et al.* proposed a machine learning framework that can securely use each organization's sensitive data [20]. Tran *et al.* proposed an efficient framework for privacy-preserving deep neural networks. This framework is not only capable of training deep learning models at high speed but is also highly resistant to collusion attacks [21].

Federated learning is another approach for privacy-preserving distributed machine learning [22]–[24]. It does not directly access each organization's sensitive data but obtains only the information necessary to update the parameters of machine learning models. These approaches are useful for statistical analysis or generating machine learning models. However, we estimate how rare each individual's attribute values are in the population from incomplete databases protected by differential privacy.

IV. METHODS

Since the idea of privacy depends on sociocultural norms, such as tightness, it is necessary to set the flexible level of privacy protection [23]. Therefore, we must determine whether a person is uniquely re-identifiable and whether the many people with the same attribute value exist in the population. Table 1 presents the main notations used in this article.

TABLE 1. Notations

ϵ	Privacy budget of differential privacy
V	Set of possible values of the combination of the attributes
f	Number of possible combination values, i.e., $\lvert V\rvert$
β	Probability of reporting the true value
N	Number of people in the population
n	Number of people in a database
q	Estimated probability that a person with specified attribute values based on a created copula model

A. GENERATING A GAUSSIAN COPULA MODEL BASED ON DIFFERENTIALLY PRIVATE DATA

We can construct a Gaussian copula model if we know the cumulative distribution function of every attribute and the covariance between every two attributes. It is impossible to obtain the true values of the cumulative distribution function for every attribute and mutual information because we do not have true data but have differentially private data. This study assumes that each attribute value is categorical; therefore, covariance is not obtained.

Estimating a probability distribution function from differentially private data has been widely studied in the research on differential privacy [15], [25]. An iterative Bayes approach can be used. Let w_i be the number of times that attribute value i is reported, $\widehat{w_i}$ be the estimated number of people who have attribute value i , and n be the total number of people who have reported their attribute values. The value of $\widehat{w_i}$ can be calculated as follows:

$$\frac{\widehat{w_i}}{n'} \leftarrow \sum_{j=1}^{n'} \frac{\frac{w_j}{n'}\beta_{i,j}\widehat{w_i}}{\sum_{k=1}^{n'}\frac{\beta_{j,k}\widehat{w_k}}{n}}, \tag{3}$$

where

$$\beta_{i,j} = \begin{cases} \beta & (i=j) \\ (1-\beta)\,/\,(f-1) & (\text{otherwise,}) \end{cases} \tag{4}$$

and the initial value of $\widehat{w_i}$ is set to w_i.

Let X be a random variable and $H(X)$ be the entropy of X. $H(X)$ can be calculated using the estimated probability distribution function of X. The estimated mutual information for the two random variables, X_i and X_j, is calculated as follows:

$$M\left(X_i,\ X_j\right) = H\left(X_i\right) + H\left(X_j\right) - H\left(X_i, X_j\right) \quad , \tag{5}$$

where $H(X_i,\ X_j)$ represent the joint entropy of X_i and X_j.

From the values of $M(X_i,\ X_j)$, each covariance value between attributes should be estimated to generate a Gaussian copula model. Considering the value distribution of $X_i$as a Gaussian distribution with μ_i mean and $\sigma_i$standard deviation. The joint distribution of X_i and X_j follows the mutual Gaussian distribution with $(\mu_i,\ \mu_j)$ mean and $\sqrt{\sigma_{ij}}\ =\sqrt{\sigma_{ji}}$

standard deviation. Here, we have

$$H\left(X_i\right) = \frac{1 + log\, 2\pi\sigma_i^2}{2} \ , \tag{6}$$

$$H\left(X_j\right) = \frac{1 + log\, 2\pi\sigma_j^2}{2}, \tag{7}$$

$$H\left(X_i, X_j\right) = \frac{\log\det\left(2\pi e\Sigma_{ij}\right)}{2} \ , \tag{8}$$

where

$$\Sigma_{ij} = \begin{pmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ji} & \sigma_j^2 \end{pmatrix}.$$

From (5) – (8) we have

$$M\left(X_i,\ X_j\right) = \frac{1}{2}\log\frac{\sigma_i^2\sigma_j^2}{\det\Sigma_{ij}} = \frac{1}{2}\log\left(1 - \frac{\sigma_{ij}^2}{\sigma_i^2\sigma_j^2}\right).$$

B. ESTIMATING OCCURRENCE PROBABILITY FROM COPULA

A copula model can incorporate rich statistical information into the perturbed data while preserving the marginal distributions of the data. The probability of a person with specified attribute values can be estimated by generating several samples from the created copula model (e.g., larger than *N*).

Let q be the estimated probability. Since this is the estimated value, there are natural cases where the value is too small to be true.

This study estimates the number of people with specified attribute values. Since attribute values are based on the attribute values of a real person, there is at least one person with that attribute value in the population. Thus, if the value of q is too small, it should be corrected to a larger value. Let α be the probability that at least one person with specified attribute values exists in the population and $\hat{\alpha}$ be the target value determined by the system manager. The value of q is replaced by a larger value of q and a minimum value that satisfies $\alpha \geq \hat{\alpha}$. The probability distribution of the estimated number of people with specified attribute values is represented by a binomial distribution with parameters n and q. In this study, the value of N is sufficiently large, and the value of q is small. Therefore, the binomial distribution can be approximated by a Poisson distribution. Thus, we have the following equation:

$$\sum_{c=1}^{\infty} Poisson\left(c;\ N\,q\right) = \widehat{\alpha}, \tag{9}$$

where

$$Poisson\left(c; Nq\right) = \frac{(Nq)^{ce^{-Nq}}}{c!}. \tag{10}$$

By simplifying (7) and (8), we obtain the following equation that needs to be solved for q:

$$-e^{Nq} + \frac{\Gamma\left(1 + N,\ Nq\right)}{N!} = \alpha, \tag{11}$$

where $\Gamma(a, b)$ represents $\int_{\mathrm{b}}^{\infty} t^{a-1}e^{-t}dt$. This equation can be solved using the Newton–Raphson method [26] or other approximation methods; however, it can be computationally expensive. Here, the value of q is significantly small; therefore, $Poisson(c; N\,q)$ has the largest value when $c = 1$, and the values of $Poisson(c; N\,q)$ with $c \geq 2$ can be ignored. Thus, we have

$$\frac{Poisson\left(c; N\,q\right)}{Poisson\left(1; N\,q\right)} = \frac{(N\,q)^{c-\,1}}{c!}. \tag{12}$$

Hence, (9) can be replaced by the following equation:

$$Poisson\left(1; n\,q\right) = \alpha. \tag{13}$$

Solving this equation, we have

$$q \ = \ -\frac{W\left(-\alpha\right)}{n}, \tag{14}$$

where W represents the Lambert W function [27], i.e., it provides the principal solution for w in $z = we^{w}$, where z is an arbitrary complex number. This equation can be solved using *scipy.special.lambertw*, a Python library.

C. USING THE FACT OF EXISTENCE IN EACH TABLE

First, we assume that there is a single database. The probability that the number of people with the specified attribute values is x in population size N, and is represented by

$$p_1\left(x\right) = {}_NC_x\, q^x(1-q)^{N-\,x} \text{ for } x \geq 0. \tag{15}$$

When the number x is greater than or equal to 1, we have

$$p_1\left(x\right) = \frac{{}_NC_x q^x(1-q)^{N-\,x}}{1-(1-q)^N} \text{ for } x \geq 1. \tag{16}$$

After that, we assume that there are m incomplete databases. Let D_i be the ith database. The population size is N and each database D_i samples people with a sampling rate s_i from the population. The sampling rates were independent of each other. Let c_i be the number of people with specified attribute values in D_i. Here, $c_i \in \{0,\ 1,\ \ldots,\ n\}$. The probability that the number of people with the specified attribute values is x in the population size N is represented by

$$p_m\left(x\right) = \frac{1}{z}p_1\left(x\right)\prod_{i=1}^{m}{}_xC_{c_i}s_i^{c_i}(1-s_i)^{x-c_i}, \tag{17}$$

where Z is the normalization term represented by

$$Z \ = \sum_{x'=1}^{n} p_1\left(x'\right)\prod_{i=1}^{m}{}_{x'}C_{c_i}s_i^{c_i}(1-s_i)^{x'-c_i}. \tag{18}$$

Finally, the expected value is calculated such that

$$E \ = \ \sum_{x=1}^{n} x \cdot p_m\left(x\right). \tag{19}$$

TABLE 2. Statistics of the Datasets

		Adult	Default	COVID-19	Census	Census5_all	Census5_3000
Number of people		30,162	40,498	674	95,130	95,130	3000
Number of attributes		9	6	5	25	5	5
Number of same records	Min	1	1	1	1	1	1
	Max	548	1920	20	4741	23,059	705
	Mean	51.9	486.1	5.1	488.0	6042.0	180.1
	Mode	1	1920	2	1	23,059	705
Product of the number of categories		5.4×10^7	1.3×10^4	3.5×10^3	4.8×10^{22}	7.7×10^4	7.4×10^4

V. EVALUATION

In this experiment, we used each database as the population. Then, the records were sampled from each dataset at a specified sampling rate to create a specified number (maximum of ten) of databases.

Consequently, we extracted 1000 records from each dataset. We predicted that the number of records with the same attribute values as the extracted records would exist in the population using the created databases. This prediction was performed by classifying whether the number of records with the same attribute values exceeded a specified threshold number of records, and then the accuracy was measured. If the threshold is one, we can uniquely identify a person in the population.

The default values of the number of databases, sampling ratio, threshold number, and values of ϵ per attribute were set to 1, 0.1, 10, and 1.0, respectively.

Accuracy was measured for the copula, OptMean, OptMode, IDUE_RAPPOR, and IDUE_OUE methods, as well as the proposed method. A copula is a method that directly applies the copula model; OptMean outputs the mean of the population's number of people with the same attribute values. Meanwhile, OptMode outputs the median of the number of people with the same attribute values in the population. IDUE_RAPPOR and IDUE_OUE were developed by Gu *et al.* [25] based on the RAPPOR [28] and OUE [29] algorithms, respectively.

We prepared three main datasets: adult, default, and census. First, we used an adult dataset [30], which is widely used to evaluate privacy-preserving data mining techniques [31], [32]. The adult dataset consists of 15 attributes (e.g., age and income) with 32,561 records, and 9 of the 15 categorical attributes were used in the experiments.

Next, we used a default dataset containing 40498 records with the attributes of gender, occupation, income, number of loans from other companies, number of late payments, and default flag (0 or 1). Here, the default flag indicates whether the borrower has failed to repay the loan. A private company provided the dataset.

Finally, we used a census dataset [33], [31]. This dataset contains 199,523 records; however, there are many missing values. After omitting records with missing values, 95,130 records remained. The number of categories per attribute in this dataset varied from 1 to 50. We used 23 categorical attributes from the census dataset. IDUE_RAPPOR and IDUE_OUE must generate a bit vector whose length is the product of the number of categories of all attributes. For the census dataset, the length was 4.8×10^{22}; therefore, it was very difficult to apply IDUE_RAPPOR or IDUE_OUE. Thus, we generated two additional datasets: Census5_all, which uses only 5 categorical attributes, and Census5_3000, which uses 5 categorical attributes with 3000 records.

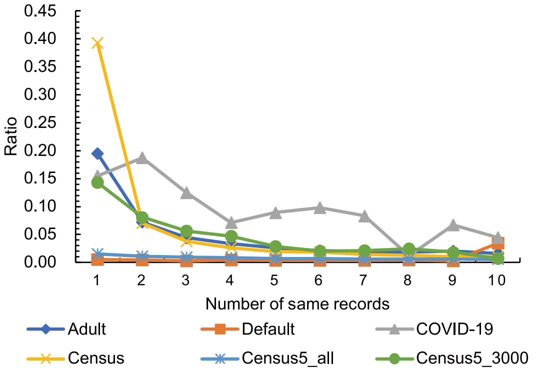
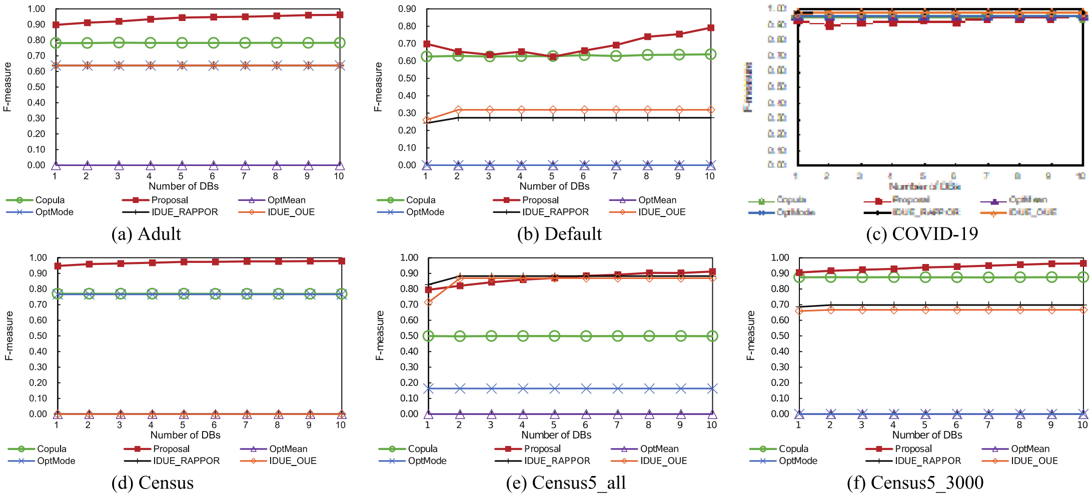


FIGURE 2. Number of the same records and their ratio. A number of the same records of 1 indicates a unique sample.

Table 2 summarizes the statistics of the databases used in the experiment. Fig. 2 shows the relationship between the number of records and their ratio. For instance, the ratio of the number of the same records being 1 is 0.004 and 0.4 for the COVID-19 and census datasets, respectively. This means that 0.4% and 40% of the samples are unique for the COVID-19 and census datasets. Thus, the unique characteristics are very different for these datasets.

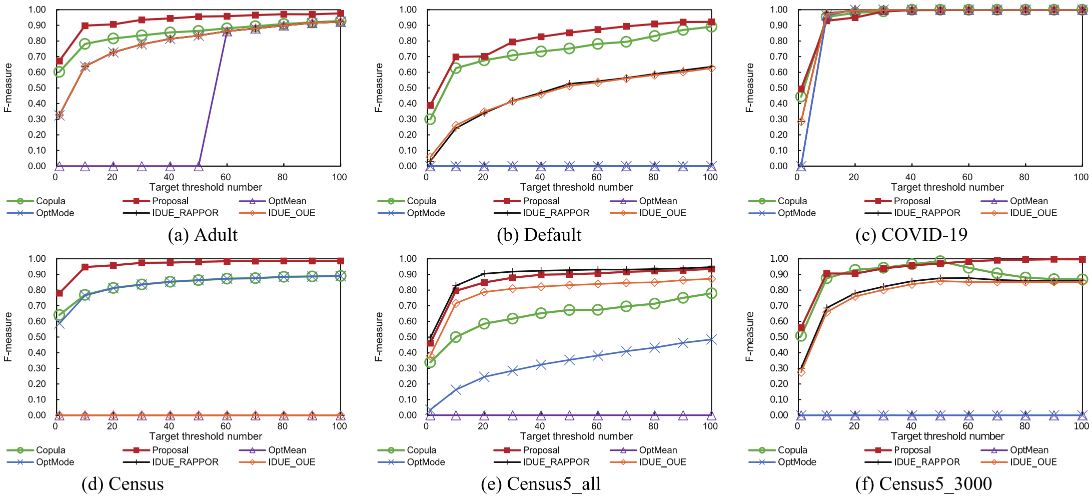
A. RESULTS FOR NON-PRIVATIZED DATASETS

Fig. 3 shows experimental results where the number of databases increased from 1 to 10. Each DB was created using independent random sampling from each dataset at a sampling rate of 0.1. Meanwhile, the accuracy hardly changed as the number of databases increased, except for the proposed method. Conversely, the proposed method slightly improved accuracy as the number of databases increased. The accuracy



(a) Adult (b) Default (c) COVID-19
(d) Census (e) Census5_all (f) Census5_3000

FIGURE 3. **Varying the number of databases with non-privatized datasets.**



(a) Adult (b) Default (c) COVID-19
(d) Census (e) Census5_all (f) Census5_3000

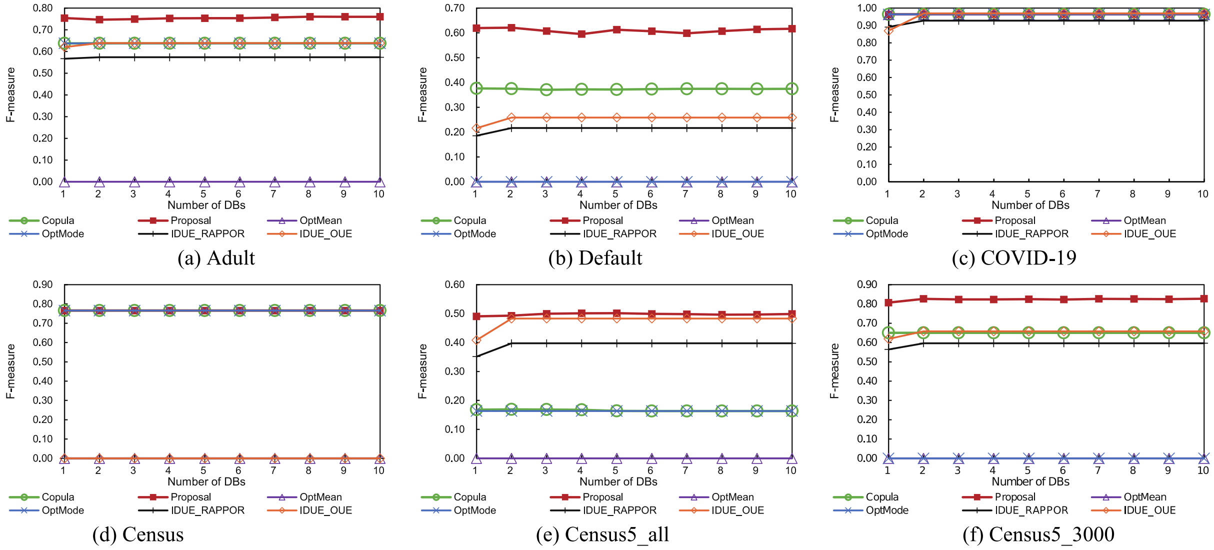
FIGURE 4. **Varying the target threshold number with non-privatized datasets.**

of OptMean and OptMode is not excellent, but high accuracy was attained only for the COVID-19 dataset. In the COVID-19 dataset, the variation in the number of records with the same value is small. Therefore, the OptMean and OptMode methods are not prone to large errors for such datasets. However, the accuracy of OptMean is always zero for all datasets except COVID-19. OptMode is more robust to variations in the number of records with the same value as OptMean; therefore, it is more accurate than OptMean. The accuracy of the copula is lower than that of the proposed method. It can be observed that the innovations in the proposed method are effective. IDUE_RAPPOR and IDUE_OUE methods are not used to determine the number of people who have the same attribute but to infer the frequency of occurrence of attribute values from a differentially private dataset. Although they can be used for this purpose, they do not achieve high accuracy because they are slightly different from the original purpose.

Fig. 4 shows the results of measuring the accuracy by varying the number of target thresholds from 1 to 100. The number of threshold values of one represents the task of determining whether there is only one person with a specified attribute value in the population. The proposed method achieves an accuracy of about 0.4–0.8, indicating that the risk of identifying an individual is high, even with sampling. This was consistent with the experimental results reported by Rocher *et al.*[1].

B. RESULTS FOR DIFFERENTIALLY PRIVATE DATASETS

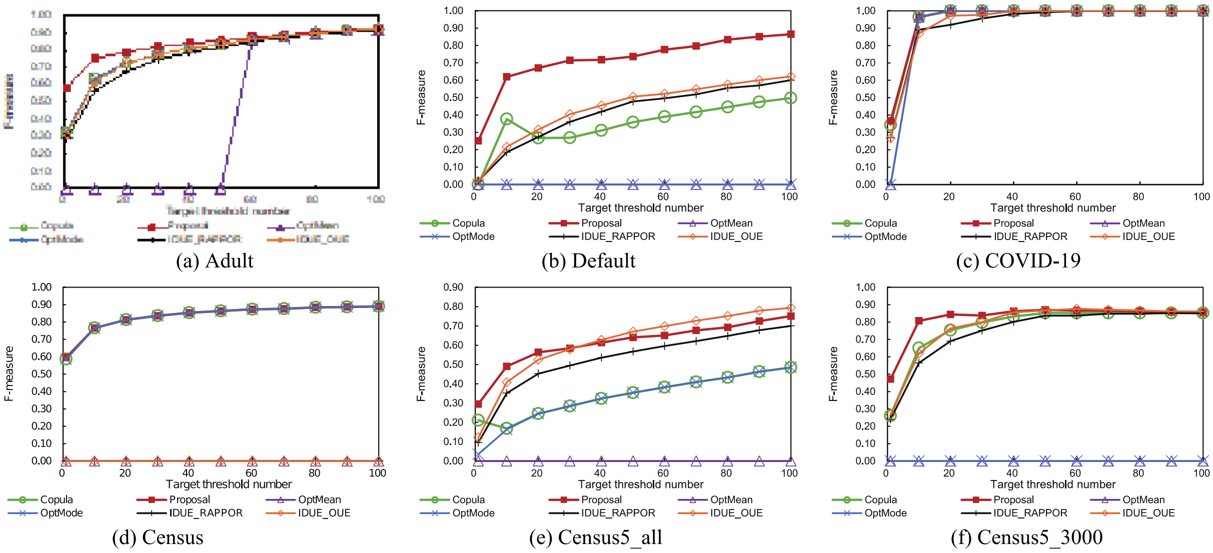
The same experiment was conducted on different private datasets. Fig. 5 shows the experimental results. The default value of ϵ was 1.0. A smaller value of ϵ corresponds to a more robust privacy level. A value of 1.0 indicates an adequate level of privacy protection, and many studies have used this value in their experiments [15], [25], [34]. The prediction accuracy for differentially privatized datasets is lower than that for



(a) Adult (b) Default (c) COVID-19

(d) Census (e) Census5_all (f) Census5_3000

FIGURE 5. **Varying the number of databases with differentially private datasets.**



(a) Adult (b) Default (c) COVID-19

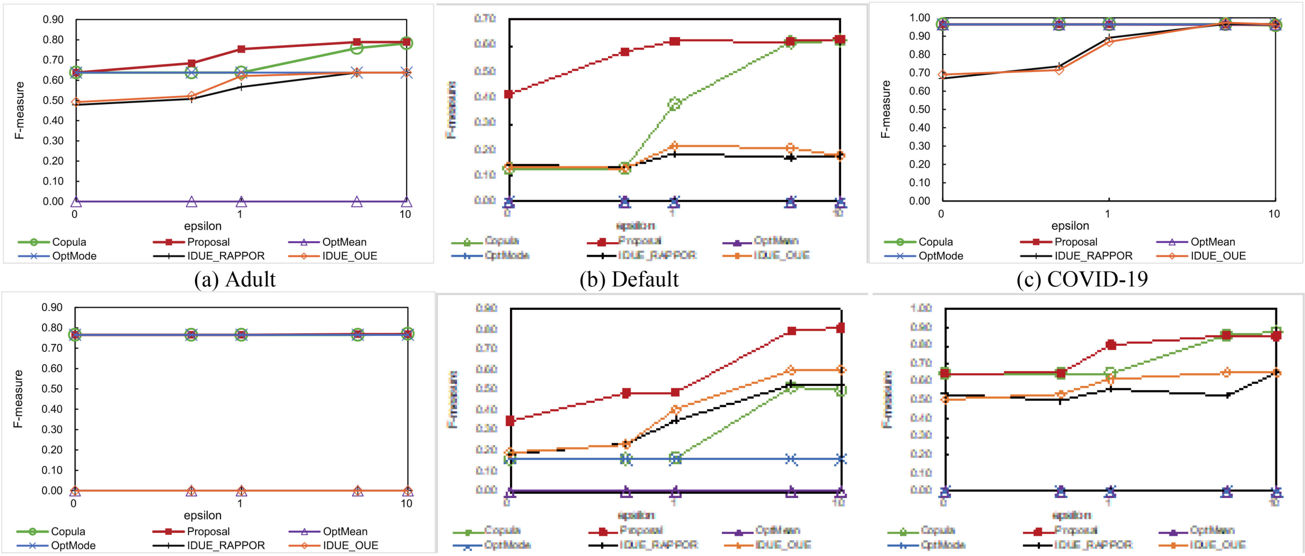
(d) Census (e) Census5_all (f) Census5_3000

FIGURE 6. **Varying the target threshold number with differentially private datasets.**

non-privatized data. However, the extent of this decrease was not statistically significant. As shown in Fig. 6, the prediction accuracy is still high when the number of target thresholds is one, ranging from 0.25 to 0.6.

Then, to analyze the effect of the value of ϵ, we conducted experiments with varying ϵ. We changed the value from 0.1 to 10.0. This range of ϵ covers the main values used in existing studies in scenarios in which individuals collect data [15], [25], [29], [35]. Fig. 7 shows the results.

For the COVID-19 and Census data sets, OptMode's accuracy is high because in these databases, the number of records with the same attribute value is usually less than 10 (see Fig. 2), and OptMode always outputs "2" on the COVID-19 data set and "1" on the Census data sets (see Table 2). However, this strategy could not generate accurate results for other data sets. IDUE_OUE outperformed IDUE_RAPPOR for almost all values of ϵ, but the accuracy of both methods was low for the Census data set. Because the product of the number of categories is considerably large, it was highly difficult for them to reconstruct the original distribution of the attribute values. Among all methods, the proposed method produced the best results for almost all values of ϵ. Therefore, the proposed method for building the copula model while mitigating the effect of differential privacy worked well regardless of the ϵ value.



(a) Adult (b) Default (c) COVID-19

(d) Census (e) Census5_all (f) Census5_3000

FIGURE 7. Varying the privacy budget ϵ.

D. REAL SMART HOUSE EXPERIMENT

We experimented with testing the risks of collecting and sharing real people's data while protecting them from differential privacy. The ethics committee approved the experiment on human experimentation at The University of Electro-Communications (Management ID: 19066), and written consent from the subjects was obtained.

A smart house was prepared in Tokyo, Japan, where 20 subjects lived individually for 2 weeks to 2 months, and data on their use of home appliances and sleep status were collected. The experiment ran from March 2020 to December 2021. The total data size of the obtained data was about 3.5 GB.[1] The gender, height, weight, age, number of times the dryer was used per day, the average heart rate upon waking, average heart rate during sleep, and average number of hours of REM, shallow, and deep sleep per day were used for this evaluation. The significance of the experimental data is limited due to the small number of subjects. However, it serves to remind the risks of sharing the data even if data are collected using differential privacy.

The subjects originally lived in the Kanto region of Japan and were also between their 20s and 30s. The population of this age group in the Kanto region is approximately 10 million. Therefore, the population size was set to 10 million, and the probability that no one other than the subjects existed with the same attribute values was estimated. Here, two people were randomly selected, and the data for these two people were split in half, and each was treated as a separate person. In other words, the experiment was conducted by assuming that data were available for 22 people. For these two persons, the estimated probability is expected to be zero. Note that even if the data are for the same person, the data contents are different.

Consequently, the estimated probabilities of those two were almost zero, as expected. The highest estimated probability among the subjects was 79%. This means that the probability that no other person in a population of 10 million has the same attribute values as this person is 79%. Therefore, there is a risk in sharing this person's data. When we checked this person's data, we discovered that he slept very little: 0.6, 1.8, and 0.6 h of REM, shallow, and deep sleep, respectively, per day. Few, if any, other individuals with such a short average sleep time exist, if they exist at all. The estimated probability for one subject was 54% for the other subjects, and for all other persons, it was less than 0.1.

E. EXPERIMENTS ON CALCULATION TIME

The proposed algorithm mainly consists of two parts. The first is the generation of the estimation model based on differentially private data, and the second is an estimation of re-identification probability using the generated estimation model. Since the model only needs to be created once from the databases, even if the computation time for the first half of the part is relatively long, it is not a major practical problem. However, the second part needs to be done every time a person's attribute values are collected, so the computation time must be short.

To measure the calculation time, we conducted additional experiments on the Adult, Default, COVID-19, Census, Census5_all, and Census5_3000 datasets. All experiments were conducted on an Intel Xeon CPU W-2295 workstation with 128 GB of RAM. Table 3 shows the results. The calculation time of generating the estimation model on the Census dataset is relatively long. However, as described above, model building only needs to be done once. It is considered sufficiently

[1] https://github.com/ponyora/smarthouse/(For now, we will only provide data for 13 subjects for privacy-protection reasons.)

TABLE 3. Calculation Time

	Adult	Default	COVID-19	Census	Census5_all	Census5_3000
Generation of the estimation model based on differentially private data	2,205 [s]	543 [s]	108 [s]	11,463 [s]	168 [s]	144 [s]
Estimation of re-identification probability per person	0.027 [s]	0.097 [s]	0.012 [s]	0.099 [s]	0.075 [s]	0.046 [s]

short compared to the time required to collect personal data, so it is not considered a major problem in practical use.

VI. DISCUSSION

We used a very simple algorithm for the differential privacy technique. This section examines what happens when we use more advanced differential privacy techniques, such as [15], [36], [37], [38]. However, the techniques used in this article and more advanced differential privacy techniques can strictly adhere to a given privacy parameter, ϵ. Using more advanced techniques does not increase the degree to which privacy is protected. The degree of privacy protection depends on only the value of ϵ, and the more advanced the technology, the less extra protection it provides. However, advanced technology can provide more useful information for data analysis. This means that advanced technology does not reduce the risk of personal identification. Moreover, this risk can be expected to grow in many cases.

If other technologies, such as k-anonymity [39] or related technology, such as [40], [41], are used in addition to differential privacy technology, the risk of being personally identified may be reduced. This study only focuses on differential privacy, which is the most standard in the field of privacy-protected data analysis. However, research on other techniques will be left for future work.

Although the Copula-based method was used in this study, it may be possible to increase the accuracy of the proposed method by using a generative adversarial network (GAN)-based method, such as [42], and increasing the amount of data. Proposals for techniques to generate GAN models while taking differential privacy into account should be considered for future work.

VII. CONCLUSION

This study proposed an algorithm to quantify the likelihood of successful re-identification when the dataset contains only a small fraction of the population, and each record is protected by differential privacy.

We demonstrated that the success rate of re-identification is reduced compared with the case where the records are not protected by differential privacy; however, the accuracy is still high. Although differential privacy provides a high level of privacy protection, it is necessary to measure the likelihood of successful re-identification using an algorithm, such as the proposed algorithm. Data considered to have a high probability of successful re-identification must be deleted or otherwise handled appropriately.

REFERENCES

[1] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature Commun.*, vol. 10, no. 1, pp. 1–9, 2019.

[2] C. Betsch *et al.*, “Social and behavioral consequences of mask policies during the COVID-19 pandemic,” *Proc. Nat. Acad. Sci.*, vol. 117, no. 36, pp. 21851–21853, Sep. 2020. [Online]. Available: https://www.pnas.org/content/117/36/21851https://www.pnas.org/content/117/36/21851.abstract

[3] S. Swayamsiddha and C. Mohanty, “Application of cognitive Internet of Medical Things for COVID-19 pandemic,” *Diabetes Metabolic Syndrome: Clin. Res. Rev.*, vol. 14, no. 5, pp. 911–915, Sep. 2020.

[4] S. Drefahl *et al.*, “Socio-demographic risk factors of COVID-19 deaths in Sweden: A nationwide register study,” *Stockholm Res. Rep. Demography*, vol. 23, pp. 1–15, Sep. 2020. [Online]. Available: /articles/preprint/Socio-demographic_risk_factors_of_COVID-19_deaths_in_Sweden_A_nationwide_register_study/12420347/4

[5] S. Ray, T. Palanivel, N. Herman, and Y. Li, “Dynamics in data privacy and sharing economics,” *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 114–115, May 2021.

[6] A. Farzanehfar, F. Houssiau, and Y. A. de Montjoye, “The risk of re-identification remains high even in country-scale location datasets,” *Patterns*, vol. 2, no. 3, Mar. 2021, Art. no. 100204.

[7] Z. Li, T. Wang, M. Lopuhaä-Zwakenberg, N. Li, and B. Škoric, “Estimating numerical distributions under local differential privacy,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2020, pp. 621–635.

[8] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, “Unique in the shopping mall: On the reidentifiability of credit card metadata,” *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[9] Y. Lin *et al.*, “Improving person re-identification by attribute and identity learning,” *Pattern Recognit.*, vol. 95, pp. 151–161, 2017.

[10] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2275–2284.

[11] R. Zhou, X. Chang, L. Shi, Y. D. Shen, Y. Yang, and F. Nie, “Person reidentification via multi-feature fusion with adaptive graph learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, May 2020.

[12] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, “An adaptive semisupervised feature analysis for video semantic recognition,” *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

[13] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, “A semisupervised recurrent convolutional attention model for human activity recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, May 2020.

[14] D. Jacobs, T. McDaniel, A. Varsani, R. U. Halden, S. Forrest, and H. Lee, “Wastewater monitoring raises privacy and ethical considerations,” *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 116–121, Sep. 2021.

[15] T. Murakami and Y. Kawamoto, “{Utility-optimized} Local differential privacy mechanisms for distribution estimation,” in *Proc. USENIX Secur. Symp.*, 2019, pp. 1877–1894.

[16] P. Kairouz, S. Oh, and P. Viswanath, “Extremal mechanisms for local differential privacy,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 492–542, 2016.

[17] L. Tan, K. Yu, N. Shi, C. Yang, W. Wei, and H. Lu, “Towards secure and privacy-preserving data sharing for COVID-19 medical records: A blockchain-empowered approach,” *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 271–281, Jan./Feb. 2022.

[18] S. Halevi, C. Hazay, A. Polychroniadou, and M. Venkitasubramaniam, “Round-optimal secure multi-party computation,” *J. Cryptology*, vol. 34, no. 3, pp. 1–63, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00145-021-09382-3

[19] M. Hastings, B. Hemenway, D. Noble, and S. Zdancewic, "SoK: General purpose compilers for secure multi-party computation," in *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 1220–1237.

[20] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten, "CrypTen: Secure multi-party computation meets machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, Art. no. 34.

[21] A. T. Tran, T. D. Luong, J. Karnjana, and V. N. Huynh, "An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation," *Neurocomputing*, vol. 422, pp. 245–262, Jan. 2021.

[22] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *J. Healthcare Inform. Res.*, vol. 5, no. 1, pp. 1–19, Mar. 2021.

[23] M. Wu, D. Ye, J. Ding, Y. Guo, R. Yu, and M. Pan, "Incentivizing differentially private federated learning: A multi-dimensional contract approach," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10639–10651, Jul. 2021.

[24] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, 2020. [Online]. Available: http://dx.doi.org/10.1038/s42256-020-0186-1

[25] X. Gu, M. Li, L. Xiong, and Y. Cao, "Providing input-discriminative protection for local differential privacy," in *Proc. IEEE 36th Int. Conf. Data Eng.*, 2020, pp. 505–516.

[26] T. J. Ypma, "Historical development of the Newton– Raphson method," *SIAM Rev.*, vol. 37, no. 4, pp. 531–551, 1995.

[27] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert*W* function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.

[28] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.

[29] T. Wang, J. Blocki, N. Li, T. Wang, J. Blocki, and N. Li, "Locally differentially private protocols for frequency estimation," in *Proc. USENIX Secur. Symp.*, 2017, pp. 729–745.

[30] D. Dua and C. Graff, "UCI machine learning repository," 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[31] J. Jia and W. Qiu, "Research on an ensemble classification algorithm based on differential privacy," *IEEE Access*, vol. 8, pp. 93499–93513, 2020.

[32] C. Ma, L. Yuan, L. Han, M. Ding, R. Bhaskar, and J. Li, "Data level privacy preserving: A stochastic perturbation approach based on differential privacy," *IEEE Trans. Knowl. Data Eng.*, early access, Dec. 21, 2021, doi: 10.1109/TKDE.2021.3137047.

[33] F. Harder, K. Adamczewski, and M. Park, "DP-MERF: Differentially private mean embeddings with randomfeatures for practical Privacy-preserving data generation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1819–1827.

[34] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," *Lecture Notes Comput. Sci.*, vol. 11693, no. 1565387, pp. 638–667, 2019.

[35] Y. Sei and A. Ohsuga, "Private true data mining: Differential privacy featuring errors to manage Internet-of-Things data," *IEEE Access*, vol. 10, pp. 8738–8757, 2022.

[36] D. Wang and J. Xu, "Differentially private high dimensional sparse covariance matrix estimation," *Theor. Comput. Sci.*, vol. 865, pp. 119–130, 2021.

[37] X. Ren *et al.*, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.

[38] H. Husain, B. Balle, Z. Cranko, and R. Nock, "Local differential privacy for sampling," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3404–3413.

[39] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[40] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," *IEEE Trans. Dependable Secure Comput.*, vol. 16, no. 4, pp. 580–593, Jul./Aug. 2019.

[41] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proc. 22nd Int. Conf. Data Eng.*, 2006, pp. 24–24.

[42] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 650–667.

YUICHI SEI (Member, IEEE) received the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2009. From 2009 to 2012, he was with Mitsubishi Research Institute, Tokyo, Japan. In 2013, he joined The University of Electro-Communications, Chofu, Japan, and is currently an Associate Professor with the Graduate School of Informatics and Engineering. He is also a Visiting Researcher with Mitsubishi Research Institute and an Adjunct Researcher with Waseda University, Tokyo, Japan. His current research interests include pervasive computing, privacy-preserving data mining, and software engineering. He was the recipient of the IPSJ Best Paper Award and JSCE Hydraulic Engineering Best Paper Award in 2017.

HIROSHI OKUMURA received the Ph.D. degree in economics from Kobe University, Hyogo, Japan, in 2012. He is currently with Mitsubishi Research Institute, Tokyo, Japan. His research interests include statistics, econometrics, and statistical machine learning. He is a Member of the Japan Statistical Society and Information Processing Society of Japan (IPSJ).

AKIHIKO OHSUGA (Member, IEEE) received the Ph.D. degree in computer science from Waseda University, Tokyo, Japan, in 1995. From 1981 to 2007, he was with Toshiba Corporation. In 2007, he joined The University of Electro-Communications, Chofu, Japan. He is currently a Professor with the Graduate School of Informatics and Engineering. He is also a Visiting Professor with the National Institute of Informatics. His research interests include agent technologies, web intelligence, and software engineering. He is a Member of the IEEE Computer Society (IEEE CS), Information Processing Society of Japan (IPSJ), Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society for Artificial Intelligence (JSAI), Japan Society for Software Science and Technology (JSSST), and Institute of Electrical Engineers of Japan (IEEJ). Since 2017, he has been a Fellow of IPSJ. He was the Chair of IEEE CS Japan Chapter, a Member of JSAI Board of Directors, a Member of JSSST Board of Directors, and a Member of JSSST Councilor. He was the recipient of the IPSJ Best Paper Awards in 1987 and 2017.