Privacy-Preserving Collaborative Data Collection and Analysis With Many Missing Values

Yuichi Sei, *Member, IEEE*, J. Andrew Onesimu, Hiroshi Okumura, and Akihiko Ohsuga, *Member, IEEE*

Abstract—Privacy-preserving data mining techniques are useful for analyzing various information, such as Internet of Things data and COVID-19-related patient data. However, collecting a large amount of sensitive personal information is a challenging task. In addition, this information may have missing values, which are not considered in the existing methods for collecting personal information while ensuring data privacy. Failure to account for missing values reduces the accuracy of the data analysis. In this article, we propose a method for privacy-preserving data collection that considers many missing values. The patient data are anonymized and sent to a data collection server. The data collection server creates a generative model and a contingency table suitable for multi-attribute analysis based on expectation–maximization and Gaussian copula methods. Using differential privacy (the de facto standard) as a privacy metric, we conduct experiments on synthetic and real data, including COVID-19-related data. The results are 50–80% more accurate than those of existing methods that do not consider missing values.

Index Terms—COVID-19, differential privacy, missing values, multi-dimensional analysis, privacy-preserving data collection

1 Introduction

To control a pandemic such as the coronavirus disease 2019 (COVID-19), we require the age, gender, family structure, and medical history of the infected individuals [1], [2]. Although such data may be provided to medical institutions by the patients themselves, the information is highly sensitive. If this information is anonymized, it can be shared among researchers worldwide without identifying the patients, which would help to elucidate the state of the pandemic and predict its course with greater accuracy.

Even when anonymized, a large amount of sensitive personal information is difficult to acquire. Moreover, this information may have missing values, as individuals who are willing to provide all confidential information are fewer than those who are willing to provide incomplete information. Researchers have proposed several methods that collect personal information while ensuring data privacy [3], [4], [5], [6]. In most of these methods, the privacy model is the ϵ-differential privacy [7], the de facto standard of privacy assurance [8]. Although these methods achieve differentially private data collection, they do not consider missing values. Consequently, the accuracy of the data analysis is significantly reduced, especially in multi-attribute analysis involving many missing values.

In this paper, we propose a method for privacy-preserving data collection that considers many missing values. The patient data are anonymized on the patient's device and/or computer in authorized hospitals, and are sent to a data collection server. Each patient can select which data to share or not share. The data collection server creates a generative model and contingency table suitable for multi-attribute analysis based on expectation–maximization and Gaussian copula methods.

We considered that if the value distribution of one or two attributes can be restored, the error in each attribute can be limited even when there are several missing values. Copula enables data generation when certain information (such as correlation and mutual information) is available for each pair of attributes. We thus combined the features of copula with those of data recovery using differential privacy. To our knowledge, this idea is novel to privacy-preserving data collection.

Applying a copula model to differentially private data collection with many missing values is our first contribution. The main technical contribution is as follows. To generate a copula model, a value distribution of each attribute and mutual information of all attributes are required. However, the server cannot collect original data, rather it collects noised data. Therefore, if the server generates value distributions and mutual information from the collected data, the generated copula model will be collapsed. Therefore, a technique to mitigate differentially private noise is necessary (described in Sections 4.2.1 and 4.2.2.)

- *Yuichi Sei is with the University of Electro-Communications, Tokyo 182-8585, Japan, and also with JST, PRESTO, Kawaguchi, Saitama 332-0012, Japan. E-mail: seiuny@uec.ac.jp.*
- *J. Andrew Onesimu is with the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India. E-mail: onesimu@gmail.com.*
- *Hiroshi Okumura is with Mitsubishi Research Institute, Tokyo 100-8141, Japan. E-mail: okumurah@mri.co.jp.*
- *Akihiko Ohsuga is with University of Electro-Communications, Tokyo 182-8585, Japan. E-mail: ohsuga@uec.ac.jp.*

Manuscript received 3 November 2021; revised 21 April 2022; accepted 9 May 2022. Date of publication 13 May 2022; date of current version 13 May 2023. This work was supported in part by JSPS KAKENHI under Grants JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19K12107, JP19H04113, and in part by JST, PRESTO under Grant JPMJPR1934.
(Corresponding author: Yuichi Sei.)
Digital Object Identifier no. 10.1109/TDSC.2022.3174887

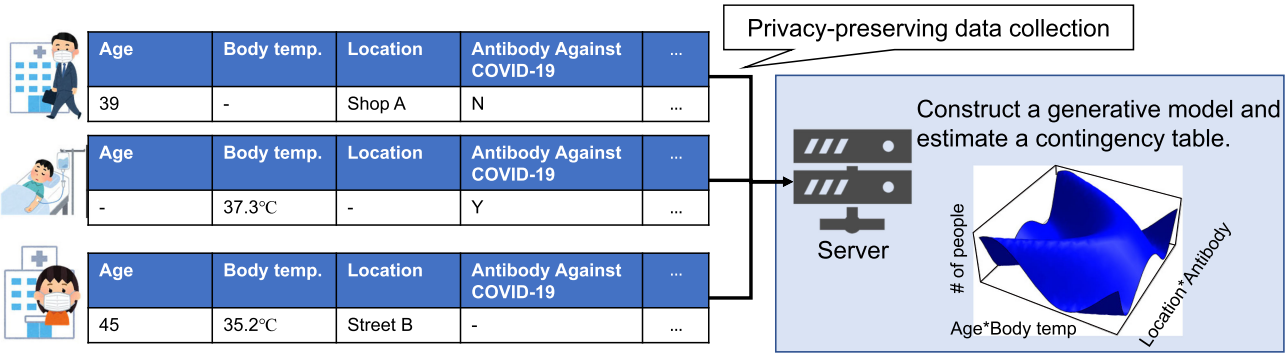


Fig. 1. Example application of the proposed privacy-preserving data collection method.

The remainder of this paper is organized as follows. Section 2 presents a motivating example and the assumptions of this study. Section 3 describes related work and Section 4 discusses the proposed method in detail. In Sections 5 and 6, we evaluate the results and discuss several practical considerations, respectively. Conclusions are presented in Section 7.

2 Background

2.1 Motivating Example

Fig. 1 presents a typical application of the proposed privacy-preserving data collection method. The data of each patient, such as age, gender, and medical history, are provided to authorized entities such as hospitals. The patients decide the information provided to the data collection server, which will be shared with researchers worldwide. Patients can also provide information not provided to hospitals, such as family structure and salary.

If the patient data are sufficiently detailed, researchers can identify patients from the information provided even when all identifiers are removed. However, if the information is insufficiently detailed, the effectiveness of the data analysis is significantly reduced. To solve this problem, we propose a differential privacy model.

Hospitals also apply a differential privacy mechanism to the information provided by each patient. The differential privacy mechanism can process any additional information provided by the patient. The differentially private information, including the additional differentially private information, is sent to the data collection server, which collects the differentially private information from several hospitals. The server then constructs a generative model that should be similar to a generative model created using true information, which is unknown to the server. Applying the generative model, researchers can construct a contingency table, mine the association rules, and perform machine learning with a suitable model such as a deep neural network.

2.2 Assumptions

In the COVID-19 scenario, we assume that all patients provide their information to the data collection server through authorized entities such as hospitals. The same assumption is made in COVID-19 contact tracing applications [9]. For example, the Ministry of Health, Labor, and Welfare in Japan launched a smartphone application called COVID-19 Contact-Confirming Application (COCOA).[1] When a COCOA user is confirmed to be infected with COVID-19, an authorized health center issues a code that the user can enter into COCOA. Only users with valid codes can register their infection.

Our proposed method can be used in other scenarios, such as crowd-sensing applications. In these applications, participants provide information such as their location and accelerometer data. Because smartphones can be used for health monitoring and cognitive function assessment [10], they can provide a medical-information portal to the data collection server. When the involvement of authorized entities is difficult, incentive and trustworthiness mechanisms such as those proposed in [11], [12] can be used.

We also assume many missing values in the collected data. As reported in the literature, many individuals hesitate to provide all their information [13], [14]. The rate of missing values ranges between 25% and 55% and may even be higher [13]. We also assume that the data collection server is honest-but-curious. That is, the server honestly follows the proposed scheme but attempts to reveal as much personal data as possible. Furthermore, we assume that the data collection server constructs a generative model and a contingency table. For this purpose, the server requires categorical attribute values. If the original value is a numerical value, it is classified into a predefined category in advance.

Several privacy-protection studies assume that users want to receive services from a service provider based on their attribute values. In such cases, the service provider requires the precise information on each user's attribute values [15]. However, in our scenario, all individuals voluntarily provide anonymized values to the data collection server and do not expect services from the data collection server based on their attribute values, although the server may provide various incentives such as financial rewards. The data collection server aims to create a dataset that can be statistically analyzed without precise information about each individual's attribute values.

3 Related Work

3.1 Differential Privacy

Differential privacy models [7] have been actively studied in the data mining field [16], [17]. Differential privacy ensures

1. https://play.google.com/store/apps/details?id=jp.go.mhlw.covid19radar&hl=en_US (accessed June 26, 2020)

that the output of the anonymization algorithm does not heavily rely on the data of a particular individual. Each individual can make a well-informed decision about whether to provide the data, and the risk of information leakage is controlled by the privacy budget ϵ.

A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if and only if for all pairs of individual values s_1 and s_2 and for all $\mathfrak{R} \subset Range(\mathcal{A})$, the following equation holds:

$$P(\mathcal{A}(s_1) \in \mathfrak{R}) \leq e^{\epsilon} P(\mathcal{A}(s_2) \in \mathfrak{R}). \quad (1)$$

Our research targets privacy-preserving data collection from each person. In this scenario, each datum from person can be considered as a database with one record. In this case, the privacy model is also known as ϵ-local differential privacy. Our research targets ϵ-local differential privacy. In this paper, "ϵ-differential privacy" refers to "ϵ-local differential privacy."

3.2 Anonymized Data Analysis With Differential Privacy

Mobile crowd-sensing is one application of anonymized data collection, and privacy-preserving techniques can incentivize participants [18]. Erlingsson *et al.* [19] proposed a privacy-preserving technique called Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR). Subsequently, Kairouz *et al.* [3] analyzed two algorithms, k-ary RR and RAPPOR, and extended them as O-RR and O-RAPPOR, respectively. O-RAPPOR outperformed k-ary RR, RAPPOR, and O-RR over the usual range of ϵ values. O-RAPPOR uses space-efficient probabilistic data structures called Bloom filters [20], which are changed at random to ensure differential privacy. However, as these methods do not consider missing values, all records with missing values must be removed before applying the methods.

Sei *et al.* [4] proposed S2Mb, which enhances the randomized response scheme [21], and proposed a method for estimating true counts from values with several errors [5]. They assumed a single attribute without missing values; if there were multiple attributes without missing values, they were converted into one attribute in advance. Several other studies on privacy-preserving data collection have been published [22], [23], [24], [25], [26]. However, all methods for differentially private anonymized data collection are heavily influenced by the number of records in the database. Thus, when the number of records is small, the accuracy of data analysis by these methods is significantly reduced.

Wang *et al.* proposed a differentially private deep neural network platform for sensitive crowd-sourced data [27]. This platform develops a deep neural network model using the sensitive data and publishes the trained model. However, attackers can use model inversion attacks [28], [29] and membership inference attacks [30], [31] to infer the sensitive raw data from the trained model. To protect these data, the platform adds noise in the training phase. Nevertheless, Wang *et al.* assumed that the platform is a trusted entity that can collect true information about sensitive data.

3.3 Missing Value Imputation

Wei *et al.* analyzed and compared the imputation accuracies of eight imputation methods [32]. The best-performing models were random forest and quantile regression imputation of left-censored data. In another study, Deb and Liew proposed an imputation method applicable to traffic accident data [33]. Their approach identifies a set of correlated records using a decision tree. The missing values are imputed from the correlation between the missing and non-missing attributes. Their method also samples several potential imputed values with high similarity.

Many imputation methods use fuzzy clustering algorithms. For example, Rahman *et al.* proposed a missing value imputation framework based on fuzzy expectation–maximization and fuzzy clustering [34]. This method searches and uses records with highest similarity to the record with missing values. The search is performed by a general fuzzy c-means clustering algorithm. Based on the membership degrees of all clusters, the missing values are then imputed by a fuzzy expectation–maximization algorithm [35], which is a modification of the regular expectation–maximization algorithm. Meanwhile, Sefidian and Daneshpour proposed the Gray-based fuzzy c-means and mutual information feature selection imputation method [36]. While executing the clustering algorithm, the distance between records is calculated using the gray relational grade and the highly related attributes (in terms of mutual information) are selected. Raja and Thangavel proposed a rough k-means centroid-based imputation method [37] that can handle inconsistencies and uncertainties in datasets. They reported that their proposed method outperforms the simple k-means and fuzzy c-means clustering methods.

All of the aforementioned methods assume that the obtained values represent the true values. However, the present study assumes that the server obtains disguised values because the true values are changed by differential privacy techniques. The structure of the disguised values depends on the applied differential privacy techniques. The structure can be a Bloom filter [3] and a set of dummy values [4]. Therefore, the existing missing value imputation methods are inapplicable to differential privacy scenarios.

3.4 Differentially Private Synthetic Datasets Generation

The literature includes several studies on differentially private data synthesis, such as [38], [39], [40], [41], [42]. These studies attempt to generate differentially private synthetic datasets from original (non-privatized) datasets. An example scenario is as follows. Assume a company holds an original (non-privatized) dataset that it wants to share with another organization. Because the original dataset contains sensitive personal information, the company should privatize the dataset. In this scenario, the company can use a differentially private data synthesis method before sharing the dataset.

Zhang *et al.* [38] proposed PrivSyn, an automatic synthetic data generation method, that calculates correlations of non-privatized attribute values and calculates multiple differentially private marginals to capture the characteristics of the non-privatized dataset. From these marginals, PrivSyn generates a differentially private synthetic dataset.

Vietri *et al.* [39] proposed two algorithms, FEM and sep-FEM. Their goal is to create a differentially private synthetic dataset from a non-privatized dataset that largely maintains

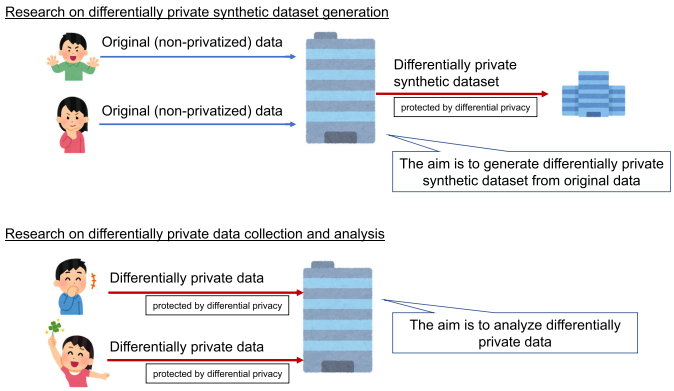


Fig. 2. The difference between researches on differentially private synthetic dataset, and differentially private data collection and analysis.

the answers to a large number of statistical queries. Although these algorithms follow the same dynamics as DualQuery [43] and MWEM [44], they could significantly improve accuracy. Similarly, Harder *et al.* [41] and Cai *et al.* [42] also assume a trusted data server has non-private personal information. In the framework proposed by Li *et al.* [40], the server can be honest-but-curious, but it needs true information that aggregates personal attribute values. These methods can release differentially private synthetic datasets with high accuracy; however, their assumptions and objectives differ from our study (see Fig. 2). We assume that a server does not have any original personal data, rather we assume it has differentially private data. The assumption and the objective of our research are very common in differentially private data collection research [3], [4], [5], [19], [45]. However, it is important to note that previous studies do not consider missing values.

The techniques of differentially private synthetic dataset generation are clearly more accurate than the techniques of differentially private data collection if the data collection server has original values; however, in this study, this assumption is not valid, and hence, a different method is needed in the present case.

4 Proposed Method

Based on differential privacy, we anonymize the patient personal data at the client side. The server collects the anonymized data and reconstructs the distributions of each attribute and all combinations of two attributes. From the two-attribute distributions, the mutual information of all pairs of attributes is calculated. Next, the generative model of the patient personal data is calculated from the mutual information using a Gaussian copula [46], [47]. Because our proposed method requires only the information about the combination of every attribute pair, it is robust to missing values. Finally, to visualize the generative model, we construct a contingency table from the generative model and the distribution of each attribute. The notations used in this study are listed in Table 1.

In the proposed method, to analyze collected differentially private data, the server constructs a copula model that mitigates the noise added by the differentially private technique. Constructing a copula model requires a value distribution of each attribute and mutual information about all attributes, as described in Section 4.2.3. Therefore, the proposed method first estimates single-attribute distribution (Section 4.2.1) and then estimates attribute-pair distribution (Section 4.2.2). Generation of the copula model is described generated in Section 4.2.3. The copula model can generate an arbitrary number of data samples without missing values. From these data samples, a contingency table is constructed (Sections 4.2.4 and 4.2.5.)

TABLE 1
Notations

ϵ	Privacy budget for differential privacy
n	Number of participants
g	Number of attributes for data collection
A_j	jth attribute
V_j	Domain of A_j
f_j	Size of V_j
V_{jk}	kth value of V_j
s_{ij}	True attribute value of A_j of Person i
R_{ij}	Disguised attribute value of A_j of Person i
c	Number of targeted attributes for analysis (used in experiments only)
m	Missing value rate (used in experiments only)

4.1 Anonymization at the Client Side

Let s_{ij} represent the value of attribute A_j of patient i. The number of attributes is g; that is, patient i has attribute values $s_{i1}, \ldots, s_{ig}$. Some values of s_{ij} may be missing. Let f_j be the number of categories of A_j.

We anonymize each non-missing value s_{ij}. Let V_j represent the domain of A_j and V_{jk} represent the kth value of V_j. For example, assume that A_1 represents the attribute of a disease {COVID-19, flu, cancer}. In this case, $f_1 = 3$ and V_{11}, V_{12}, and V_{13} are COVID-19, flu, and cancer, respectively.

Based on a previous method [4], we create a value set R_{ij} for each attribute A_j as follows:

$$R_{ij} = \begin{cases} \{s_{ij}\} \cup Ran(V_j \backslash \{s_{ij}\}, h_j - 1) & \text{with prob. } p_j \\ Ran(V_j \backslash \{s_{ij}\}, h_j) & \text{otherwise,} \end{cases} \tag{2}$$

where $Ran(S, h)$ represents a function that randomly selects h elements without duplication from set S. For example, assume that $S = \{A, B, C\}$, and h=2. In this case, $Ran(S, h)$ outputs $\{A, B\}$, $\{B, C\}$, or $\{A, C\}$. To satisfy ϵ-differential privacy, the parameters h_j and p_j are respectively determined as

$$h_j = \max\left(\left\lfloor \frac{f_j}{1 + e^{\epsilon}} \right\rfloor, 1\right) \quad \text{and}$$
$$p_j = \frac{e^{\epsilon} h_j}{f_j - h_j + e^{\epsilon} h_j}, \tag{3}$$

following [4]. As there are g attributes in our scenario, each R_{ij} should satisfy ϵ/g-differential privacy [48].

Algorithm 1 is the anonymization algorithm from the client side.

The privacy budget allocated to each attribute is ϵ/g. Even if all attributes are the same, i.e., the correlations among attributes are 1, we can satisfy ϵ-differential privacy due to the composition property of differential privacy [48].

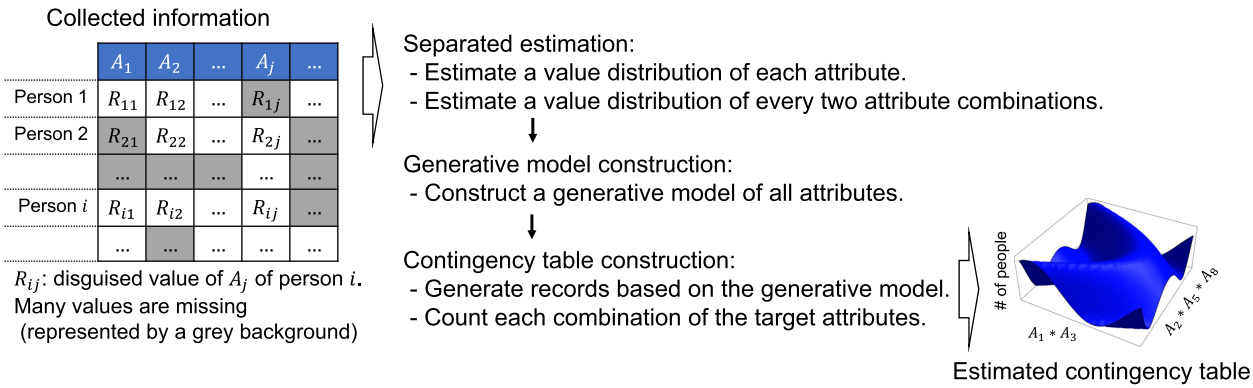


Fig. 3. Overview of the proposed estimation scheme.

Algorithm 1. Anonymization Algorithm for Patient i

Input: Privacy parameter ϵ, original data $\{s_{i1}, \ldots, s_{ig}\}$, each domain V_j
Output: Anonymized version of $\{s_{i1}, \ldots, s_{ig}\}$
1: **for** $j = 1, \ldots, g$ **do**
2: $f_j \leftarrow |V_j|$
3: Based on (3), determine p_j and h_j by substituting ϵ/g into ϵ
4: Based on (2), obtain R_{ij} from s_{ij} and V_j
5: **end for**
6: **return** $\boldsymbol{R_i} = \{R_{i1}, \ldots, R_{ig}\}$

4.2 Estimation at the Server Side

The data collection server first estimates the value distribution of each attribute as described in Section 4.2.1. It then estimates the value distribution of each attribute pair as described in Section 4.2.2. Using these estimated value distributions, the server creates a generative model (a Gaussian copula; see Section 4.2.3). Finally, it generates n complete data records and creates a contingency table of target attributes, which is specified by a data analyzer (Sections 4.2.4 and 4.2.5). Fig. 3 presents the overall structure of the proposed estimation scheme.

4.2.1 *Separated Estimation: Estimation of a Value Distribution of Each Attribute*

Each client sends its true value and $(h_j - 1)$ randomly selected values other than the true value with probability p_j and sends h_j randomly selected values other than the true value with probability $(1 - p_j)$ for attribute j, as represented in Algorithm 1. As a result, the probability that the true value is sent is p_j, and the probability that another value is sent is

$$q_j = \frac{p_j(h_j - 1)}{f_j - 1} + \frac{(1 - p_j)h_j}{f_j - 1} = \frac{h_j - p_j}{f_j - 1}, \tag{4}$$

as for attribute j. Here, because a total of h_j values are sent, $p_j + (f_j - 1)q_j = h_j$.

Let w_{jk} represent the number of occurrences of V_{jk} in $\{\boldsymbol{R_1}, \ldots, \boldsymbol{R_n}\}$, and let u_{jk} represent the true number of occurrences of V_{jk}. Thus, we have the following equation:

$$\begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jf_j} \end{pmatrix} = M \begin{pmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jf_j} \end{pmatrix}, \tag{5}$$

where M is the matrix where the diagonal elements are p_j and other elements are q_j. The symbol z_{jk} represents the estimated number of occurrences of V_{jk}. We can easily estimate these values by calculating the following equation:

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ \vdots \\ z_{jf_j} \end{pmatrix} = M^{-1} \begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jf_j} \end{pmatrix}, \tag{6}$$

where M^{-1} represents the inverse matrix of M. However, the estimation accuracy is very low [3]. Moreover, calculating the inverse function requires significant computation time, particularly for a large matrix. To overcome these limitations, we selected the expectation-maximization (EM)-based algorithm. If we know the values of u_{jk}, we can calculate each expected value of w_{jk}. In our problem setting, we know the actual values of w_{jk}; however, we do not know u_{jk}. Therefore, considering u_{jk} as unobserved latent variables, the EM-based algorithm can provide maximum a posteriori estimation. It can find the unobserved latent variables that best explain the observed values. Moreover, the EM-based algorithm can ensure the likelihood increase with each iteration [49], [50].

The symbol $\widetilde{n_j}$ represents the number of records in which a value exists for attribute A_j

$$\widetilde{n_j} = \sum_{k=1}^{f_j} w_{jk}. \tag{7}$$

Let z_{jk} represent the estimated number of occurrences of V_{jk} in A_j. From the expectation–maximization-based algorithm [4], we obtain z_{jk} by repeating the following substitution:

$$z_{jk} \leftarrow z_{jk}(p_j \mathcal{D}_k + q_j(\mathcal{E} - \mathcal{D}_k)), \tag{8}$$

where

$$q_j = \frac{h_j - p_j}{f_j - 1}, \tag{9}$$

$$\mathcal{D}_k = \frac{w_{jk}}{p_j z_{jk} + q_j(h_j \widetilde{n_j} - z_{jk})}, \tag{10}$$

TABLE 2
Example Table Created by the Privacy-Preserving Data Collection

Record ID	Age (A_1) (years)	Body temp. (A_2) (°C)	Location (A_3)
1	{39, 40, 58}	{35.2, 35.5}	-
2	{12, 22, 30}	-	{Shop A, Hospital D}
3	{25, 40, 61}	-	{Street B, Hospital D}
4	{33, 34, 88}	{37.5, 37.6}	{School C, Shop E}

and

$$\mathcal{E} = \sum_{k=1}^{f_j} \mathcal{D}_k. \tag{11}$$

4.2.2 Separated Estimation: Estimation of a Value Distribution of Every Two Attribute Combinations

Let $V_{jj'}$ be the combinations of the elements of attributes A_j and $A_{j'}$

$$V_{jj'} = V_j \times V_{j'}. \tag{12}$$

Let $w_{jj'kk'}$ represent the number of simultaneous occurrences of V_{jk} and $V_{j'k'}$ in each record of $\{\boldsymbol{R_1}, \ldots, \boldsymbol{R_n}\}$. The symbol $\widetilde{n_{jj'}}$ represents the number of records in which a value exists for both attributes A_j and $A_{j'}$

$$\widetilde{n_{jj'}} = \sum_{k=1}^{f_j} \sum_{k'=1}^{f_{j'}} w_{jj'kk'}. \tag{13}$$

As an example, assume that Table 2 was created by the privacy-preserving data collection. $\widetilde{n_1}$, $\widetilde{n_2}$, and $\widetilde{n_3}$ are 4, 2, and 3, respectively because attribute A_1 has four values, attribute A_2 has two values, and attribute A_3 has three values. $\widetilde{n_{1,2}}$ is 2 because two records (the first and fourth records) contain values in both A_1 and A_2 (the values are [39, 40, 58, 35.2, 35.5] and [33, 34, 88, 37.5, 37.6]). Similarly, $\widetilde{n_{1,3}}$ and $\widetilde{n_{2,3}}$ are 3 and 1, respectively.

As in Section 4.2.1, we estimate the occurrence of each combination V_{jk} and $V_{j'k'}$ of attributes A_j and $A_{j'}$ for n patients. By calculating these values for all combinations A_j and $A_{j'}$, we can estimate all value distributions of all attribute pairs.

After estimating attribute-pair distribution, the method of calculating mutual information is as follows. Mutual information of attributes j and j' is calculated as follows:

$$\sum_{k \in V_j} \sum_{k' \in V_{j'}} p(k, k') \log \frac{p(k, k')}{p(k)p(k')}, \tag{14}$$

where $p(k, k')$ represents the joint probability that V_{jk} and $V_{j'k'}$ occur, and $p(k)$ represents the probability that V_{jk} occurs.

4.2.3 Generative Model Construction: Constructing a Generative Model as the Gaussian Copula

Let $X_1, \ldots, X_g$ be random variables and let $F(x_1, \ldots, x_g)$ represent the joint probability distribution function of $X_1, \ldots, X_g$. The marginal distribution functions $F_1, \ldots, F_g$ and the joint probability distribution function have the following relationship.

Theorem 4.1 (Sklar's Theorem [51]). *A function C uniquely satisfies the following expression:*

$$\begin{aligned} F(x_1, \ldots, x_g) &= Pr(X_1 \leq x_1, \ldots, X_g \leq x_g) \\ &= C(F_1(x_1), \ldots, F_g(x_g)). \end{aligned} \tag{15}$$

From Sklar's Theorem, we have

$$C(u_1, \ldots, u_g) = F(F_1^{-1}(u_1), \ldots, F_g^{-1}(u_g)), \tag{16}$$

for arbitrary $\boldsymbol{u} = (u_1, \ldots, u_g)$ ($u_i \in [0, 1]$). Based on Sklar's Theorem, we have

$$\begin{aligned} \Phi_g(x_1, \ldots, x_g; \Sigma) &= Pr(X_1 \leq x_1, \ldots, X_g \leq x_g) \\ &= C(\Phi(x_1), \ldots, \Phi(x_g)), \end{aligned} \tag{17}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of a standard Gaussian distribution, and $\Phi_g(x_1, \ldots, x_g; \Sigma)$ represents the cumulative distribution function of a g-dimensional Gaussian distribution with random variables $X_1, \ldots, X_g$, and a covariance matrix Σ.

From (17), the cumulative distribution of the Gaussian copula can be expressed as

$$C(u_1, \ldots, u_g) = \Phi_g(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_g); \Sigma). \tag{18}$$

The Gaussian copula C represents the cumulative distribution function of each marginal distribution, which is a uniform distribution in the range [0,1]. The probability density function of the Gaussian copula $c(u_1, \ldots, u_g; \Sigma)$ satisfies the following relationship:

$$\phi(x_1, \ldots, x_g) = c(\Phi(x_1), \ldots, \Phi(x_g)) \prod_{j=1}^{g} \phi(x_j), \tag{19}$$

where $\phi(\cdot)$ represents the probability density function of a standard Gaussian distribution, i.e.,

$$\phi(x_1, \ldots, x_g) = \frac{1}{\sqrt{(2\pi)^g |\Sigma|}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right). \tag{20}$$

Therefore, we have

$$c(u_1, \ldots, u_g) = \frac{1}{\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \boldsymbol{\omega}^T (\Sigma^{-1} - \boldsymbol{I}) \boldsymbol{\omega}\right), \tag{21}$$

where $\boldsymbol{\omega} = \Phi^{-1}(\boldsymbol{u})$.

Σ must be estimated from the collected data. Let $\boldsymbol{u}^i$ and $\boldsymbol{\omega}^i$ represent the ith $\boldsymbol{u}$ and ith $\boldsymbol{\omega}$, respectively. From (21), the log-likelihood function of the Gaussian copula is given by

$$l(\Sigma) = -\frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\omega}^{i^T} (\Sigma^{-1} - \boldsymbol{I}) \boldsymbol{\omega}^i, \tag{22}$$

where $\boldsymbol{\omega}^j = \Phi^{-1}(\boldsymbol{u}^j)$. Differentiating (22) with respect to Σ^{-1}, we obtain [52]

$$\frac{\partial l(\Sigma)}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\omega}^i \boldsymbol{\omega}^{i^T}. \tag{23}$$

Therefore, the maximum likelihood estimator $\widehat{\Sigma}$ is

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\omega}^i \boldsymbol{\omega}^{i^T}. \tag{24}$$

To alleviate the high computational cost of (24), we estimate Σ using a suboptimal approach [47]. First, we calculate the mutual information of every pair of attributes using the reconstructed data in Section 4.2.2. We then determine each suboptimal element of Σ that minimizes the distance between the mutual information of the estimated joint distribution and that calculated from the reconstructed data (see Section 4.2.2).

Although a copula assumes that it can obtain original (non-privatized) data, it can construct a generative model from small data. There are several types of copulas, such as Gaussian and Student-t copulas. Garcia-Jorcano and Benito demonstrated experimentally that the Gaussian and Student-t copulas were best among Gaussian, Student-t, Clayton, Gumbel, and Frank copulas [53]. Lasmar and Berthoumieu preferred the Gaussian copula to the Student-t copula because the Gaussian copula can achieve high accuracy and its parameters can be easily estimated [54]. Based on these previous studies, the Gaussian copula was selected in this study. However, several other studies pointed out the shortcomings of the Gaussian copula [55]. We intend to consider using other types of copulas in future work.

4.2.4 Contingency Table Construction: Generation of Records Based on the Generative Model

We generated n complete data from the Gaussian copula C and the reconstructed data in Section 4.2.1. The n values of each attribute A_j were determined based on the estimated attribute distribution in Section 4.2.1. We also generated random values $\bar{x_1}, \ldots, \bar{x_g}$ based on an g-dimensional Gaussian distribution with covariance matrix $\widehat{\Sigma}$. We then obtained $u_i = \Phi(x_j)$ for all $i = 1, \ldots, g$. From the reconstructed data in Section 4.2.1, we finally obtained $F_j^{-1}(u_j)$ for each attribute value, where F_j represents the marginal distribution of attribute A_j.

4.2.5 Contingency Table Construction: Counting Each Combination of the Target Attributes

After the above process, we obtained n complete data records with g attributes. If a contingency table is used for many attributes, it loses its primary value [56], [57]. Therefore, data analyzers generally select several attributes. The target contingency table is then constructed by simply counting the occurrences of each combination of attribute values from the n generated complete data records.

Algorithm 2 presents the server algorithm.

4.3 Security Analysis

The proposed algorithm satisfies ϵ-differential privacy, as proven below. Because the estimation algorithm at the server side uses only the anonymized data generated at the client side, we must prove the safety of the anonymization algorithm at the client side.

Algorithm 2. Algorithm for Generating the Gaussian Copula and Contingency Table

Input: Privacy parameter ϵ, anonymization parameters p_j and h_j, anonymized dataset $\{\boldsymbol{R_1}, \ldots, \boldsymbol{R_n}\}$, set of target attributes for contingency table
Output: Contingency table of target attributes

for $j = 1, \ldots, g$ **do**
  $Z_j \leftarrow$ estimated value distribution of A_j calculated by Equation (8) based on $R_{ij} (i = 1, \ldots, n)$.
end for
for $j = 1, \ldots, g$ **do**
  for $j' = 1, \ldots, g$ **do**
    if $j \neq j'$ **then**
      $Z_{jj'} \leftarrow$ estimated value distribution of the combination of A_j and $A_{j'}$ calculated by Equations (12), (13), and (8) based on R_{ij} and $R_{ij'} (i = 1, \ldots, n)$.
    end if
  end for
end for
for $j = 1, \ldots, g$ **do**
  Construct cumulative distribution function F_j based on Z_j.
end for
for $j = 1, \ldots, g$ **do**
  for $j' = 1, \ldots, g$ **do**
    if $j \neq j'$ **then**
      for until the change of Σ converges **do**
        Create a temporary Copula with F_j, F_j' and a temporal Σ.
        jth row and j'th column of Σ are calculated based on the mutual information of $Z_{jj'}$.
      end for
    end if
  end for
end for
$O \leftarrow \varnothing$
for $i = 1, \ldots, n$ **do**
  $\{x_1, \ldots, x_g\} \leftarrow$ generated based on g-dimensional Gaussian distribution with covariance matrix Σ.
  for $j = 1, \ldots, g$ **do**
    $u_j \leftarrow \Phi(x_j)$
    $z_j \leftarrow F_j^{-1}(u_j)$
    $O \leftarrow O \cup \{\{z_1, \ldots, z_g\}\}$
  end for
end for
Count each occurrence of the combination of attribute values in O.

Theorem 4.2. *The anonymization algorithm at the client side satisfies ϵ-differential privacy.*

Proof. First, we prove that the anonymization for each attribute satisfies ϵ/g-differential privacy. The probability that R_{ij} contains s_{ij} and $h_j - 1$ specified elements is given by

$$\mathcal{P} = \frac{p_j}{{}_{f_j - 1}C_{h_j - 1}}, \tag{25}$$

and the probability that R_{ij} does not contain s_{ij} but contains h_j specified elements is

TABLE 3
Analysis of the Time Complexity

Step	Time Complexity
Estimate a value distribution of each attribute. (lines 1-3 in Algorithm 2)	$O(n+g)$
Estimate a value distribution of every two attribute combinations. (lines 4-10 in Algorithm 2)	$O(n+g^2)$
Construct a generative model of all attributes. (lines 11-23 in Algorithm 2)	$O(g^2)$
Generate records based on the generative model. (lines 24-32 in Algorithm 2)	$O(ng)$
Count each combination of the target attributes. (line 33 in Algorithm 2)	$O(\prod_{j=1}^{c} f'_j)$ where f'_j represents the domain size of jth targeted attribute for analysis

$$\mathcal{Q} = \frac{1-p_j}{{}_{f_j-1}C_{h_j}}. \tag{26}$$

By Equation (1), the constraints on the values of p_j and h_j based on ϵ/g-differential privacy are given by

$$\begin{aligned} e^{\epsilon/g} &\geq \max\left(\frac{\mathcal{P}}{\mathcal{Q}}, \frac{\mathcal{Q}}{\mathcal{P}}\right) \\ &= \max\left(\frac{p_j(f_j-h_j)}{(1-p_j)h_j}, \frac{(1-p_j)h_j}{p_j(f_j-h_j)}\right). \end{aligned} \tag{27}$$

Substituting Equation (3) into the right side of Expression (27), we obtain

$$\max\left(e^{\epsilon/g}, e^{-\epsilon/g}\right). \tag{28}$$

Because e is greater than 1 and ϵ and g are greater than 0, Expression (27) is satisfied.

As mentioned above, each attribute value is protected by ϵ/g-differential privacy. Because there are g attributes satisfying ϵ/g-differential privacy, the final output of the anonymization algorithm satisfies ϵ-differential privacy [48]. □

4.4 Time Complexity Analysis

The proposed system involves five steps as shown in Fig. 3. The relationship between each step and time complexity is described in Table 3.

The total time complexity is $O(ng + g^2 + \prod_{j=1}^{c} f'_j)$. Please note that the last item $\prod_{j=1}^{c} f'_j$ is common for all the existing methods because the purpose is to generate histograms with the number of bins as $\prod_{j=1}^{c} f'_j$.

Because the task "Construct a generative model of all attributes" is more complex than the task "Generate records based on the generative model," the impact of $O(g^2)$ is larger than the impact of $O(ng)$. Therefore, if too many attributes are found, the calculation cost will increase greatly. However, even when the number of attributes is 500 and the number of people is 10,000, the calculation time was 310 min in our experiments. All experiments were conducted on an Intel Xeon CPU W-2295 PC with 64 GB RAM.

5 Evaluation

5.1 Evaluation Setting

We compared the performances of the proposed method and four state-of-the-art methods: O-RAPPOR [3], S2Mb [4], MDN [23], and PDE/ETE [5].

Please note that as described in Section 3.4, evaluating the techniques of differentially private synthetic dataset generation is impossible in this study because the data collection server does not access the original values of people in the experiments although the techniques of differentially private synthetic dataset generation require the original values.

The experimental results of the simple combination of the differentially private technique at the client side and the copula technique at the server side are also shown. This method is referred to as DF+Copula.

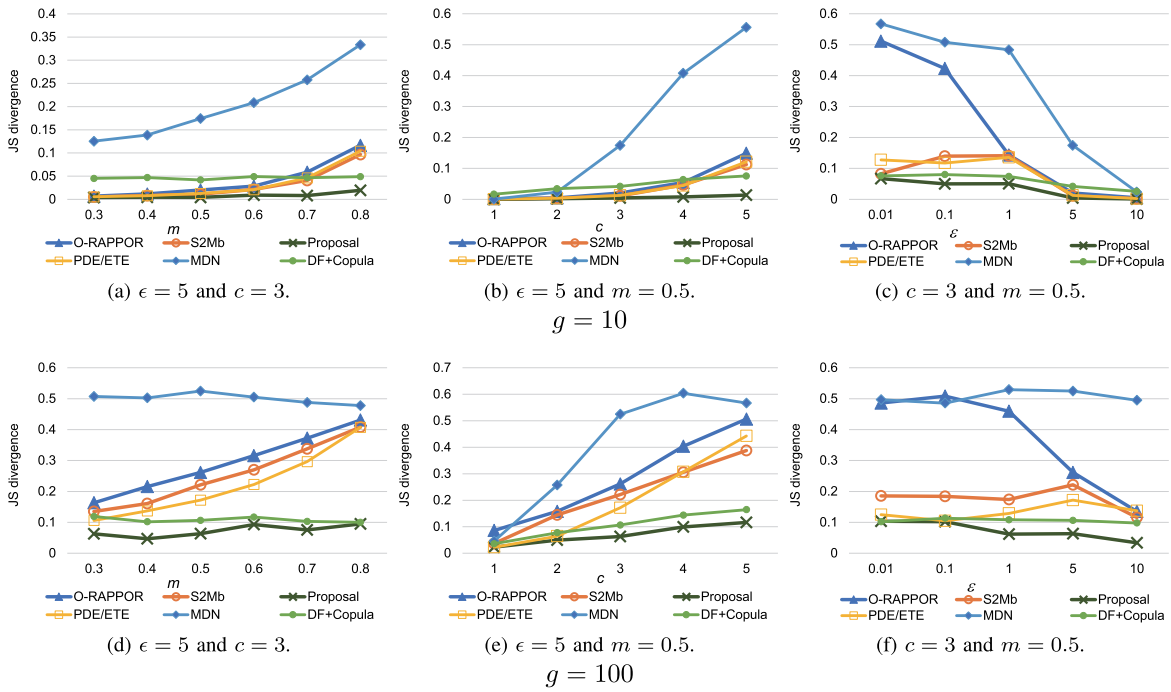
If the estimated contingency table generated by each method was similar to that generated from the valid data, which was unknown to the data collection server, the estimated contingency table was considered to be well generated by the model.

In this study, a contingency table is considered as a probability distribution of attribute values. To measure the difference between the probability distributions, we applied the Jensen–Shannon (JS) divergence rather than the usual Kullback–Leibler (KL) divergence, because the KL divergence assumes all non-zero probabilities. If any probabilities are zero, the KL divergence fails due to a division-by-zero error. The JS divergence is based on the KL divergence but does not impose the non-zero constraint.

In the Apple implementation, ϵ equals 1 or 2 per datum [58]. In evaluations by the Apple differential privacy team, ϵ was set to 2, 4, and 8 [59]. Microsoft described their differentially private framework and, in their paper, they set ϵ from 0.1 to 10 [60]. In the paper that proposed RAPPOR [19], which was developed by Google, $\epsilon = \log(3)$ is used as the main setting. Hsu showed that, in the literature, ϵ ranges from 0.01 to 10 [61]. Based on these settings, we set the value of ϵ from 0.01 to 10.

We varied the missing value rate m from 0.3 to 0.8, and the number of attributes c in the analysis from 1 to 5. The reported results are the averages of 100 experiments for each parameter setting. As the default parameters, we set $m = 0.5$, $c = 3$, and $\epsilon = 5$.

Note that the missing value rate m is used only for the experiments, and the proposed algorithm does not require



(a) $\epsilon = 5$ and $c = 3$. (b) $\epsilon = 5$ and $m = 0.5$. (c) $c = 3$ and $m = 0.5$.

$g = 10$

(d) $\epsilon = 5$ and $c = 3$. (e) $\epsilon = 5$ and $m = 0.5$. (f) $c = 3$ and $m = 0.5$.

$g = 100$

Fig. 4. Results of synthetic data ($f_j = 2$).

this information. The number of targeted attributes for analysis c can be freely determined by the data analyst according to the purpose of the analysis.

5.2 Experiments on Synthetic Data

We first evaluated the JS divergence on synthetic datasets, changing the number of attributes g from 10 to 100. The attribute values of each individual were determined randomly, the number of patients n was set to 10,000, and $f_j = 2$ for all j.

Several existing studies target a binary (i.e., $f_j = 2$) scenario. For example, Kairouz *et al.* [3] proposed a technique targeting $f_j = 2$ first. Then, they extended the technique and proposed O-RAPPOR. We conducted this experiment to verify the effectiveness of the proposed method in such a basic setting. To evaluate the practical application of the proposed method, we conducted experiments with four real datasets.

Figs. 4a, 4b, 4c and 4d, 4e, 4f present the results for $g = 10$ and $g = 100$, respectively. Under almost all parameter settings, the JS divergence was lower in the proposed method than in the established methods. As the number of attributes g increased, the results in Figs. 4d, 4e, and 4f were higher than those in Figs. 4a, 4b, and 4c. Meanwhile, increasing the privacy budget ϵ lowered the privacy-protection level. Therefore, when ϵ was large, the JS divergence decreased in all methods.

As the missing value rate m and several target attributes c increased, the JS divergence increased in the O-RAPPOR, S2Mb, MDN, and PDE/ETE methods, but remained low in the proposed method and DF+Copula. PDE/ETE does not transform the numerical values into categorical values.

The results demonstrate that in terms of accuracy, the proposed method outperforms the simple combination of a differentially private technique and a copula technique (DF +Copula.)

This observation confirms the robustness of the proposed method to missing values.

Then, we varied the number of attributes (g) from 10 to 200. Fig. 5 shows the results. The fewer the attributes, the fewer are the bins in the histogram, and accordingly, the JS divergence tends to be smaller. Conversely, if many attributes are found, most of the values in the histogram will be zero. Therefore, by predicting the value of most bins to be 0, we can expect a small JS divergence in this case as well. Because of these two characteristics, for many methods, we can see from the figure that the JS divergence increases as the number of attributes increases to some extent, and then, it decreases as the number of attributes increases further. For all settings, the proposed method exhibited the smallest JS divergence among all the tested methods.

Although several previous studies used a dataset with more than 200 attributes [41], many studies use datasets with less than 100 attributes [38], [39], [40], [42]. Because our proposed method calculates attribute-pair distribution, the processing time for datasets with many attributes is relatively long. In future, we will address ways to reduce the simulation time and will conduct additional experiments with the proposed method on datasets with a large number of attributes.

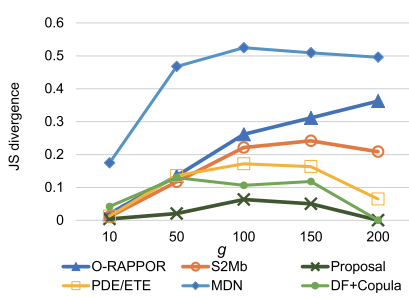


Fig. 5. Results of synthetic data ($\epsilon = 5, c = 3, m = 0.5$).

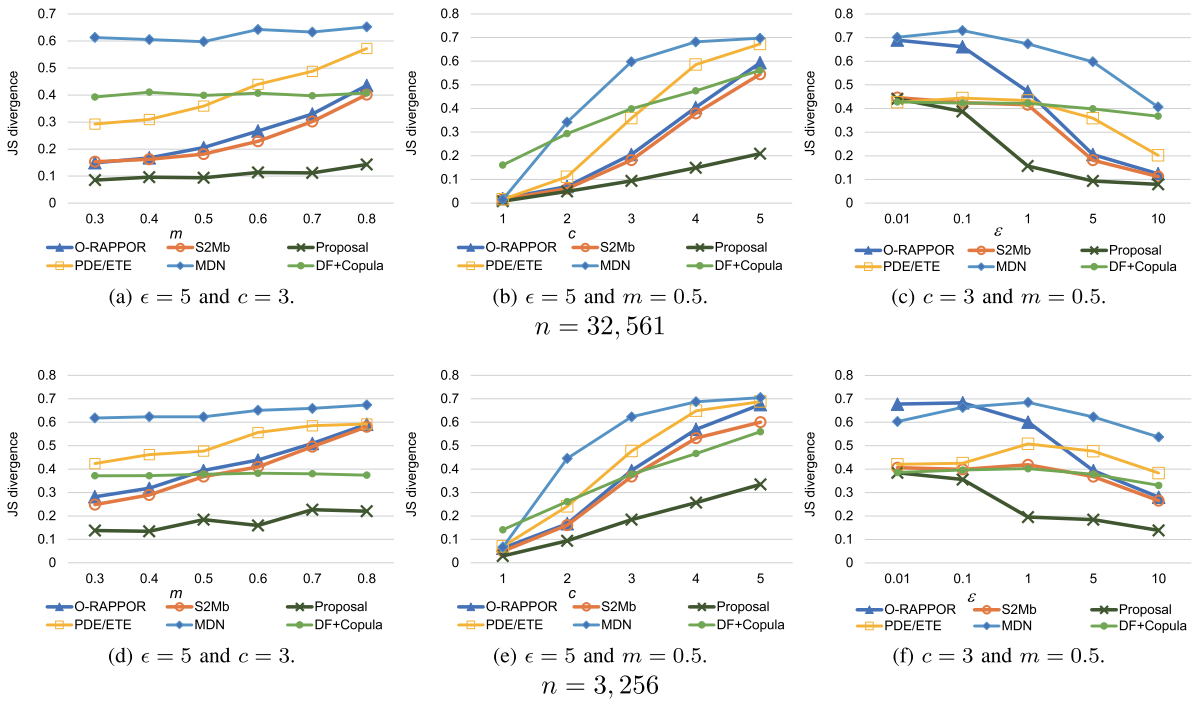


Fig. 6. Results of the Adult dataset.

5.3 Experiments on Real Data

In the real-data experiments, we first investigated the Adult dataset [62], which is widely used in evaluations of privacy-preserving data mining techniques (for example, see [63], [64], [65]). The Adult dataset consists of 15 attributes (e.g., age, income) in 32,561 records. The number of categories in these experiments was set from 2 to 9 per attribute.

Figs. 6a, 6b, and 6c present the experimental results.

When the missing value rate was small or ϵ was large, the JS divergence of the proposed method was similar to those of S2Mb, PDE/ETE, and O-RAPPOR. Similarly, when ϵ was small, the JS divergence of the proposed method was similar to the JS divergence of S2Mb, PDE/ETE, and DF+Copula.

However, at high missing value rates, the proposed method outperformed the other methods, achieving a high level of privacy protection.

To determine whether the proposed method is applicable to small datasets, we randomly sampled 10% of the 32,561 records in the Adult dataset and measured their JS divergence. Figs. 6d, 6e, and 6f present the results. Owing to the data sparsity, the estimation task was more difficult than in the other experiments and the JS divergence in all methods was higher for the 3,256 records than for the 32,561 records. However, the proposed method was robust to the small dataset. On a larger dataset with an insignificant missing value rate, the JS divergence was higher in the proposed method than in the existing methods. Therefore, regardless of missing value rate, the proposed method outperformed the other methods on smaller datasets.

We then used the Communities and Crime Unnormalized dataset [66] (hereafter referred to as the Community dataset). This dataset contains 124 predictive attributes such as the percentage of individuals aged 25 and over with a bachelor's or higher degree, which may be considered as private information in some communities.

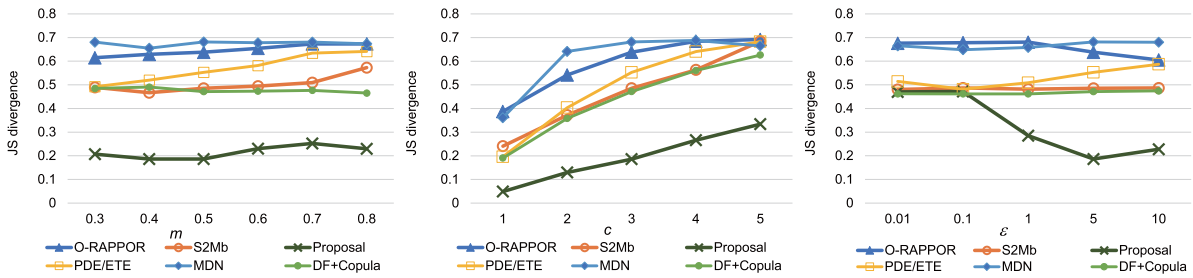
After removing the 22 attributes with more than 80% missing values, we obtained 102 attributes for analysis.

Fig. 7 presents the experimental results of the Community dataset. The results are similar to those of the synthetic Adult dataset. For almost all parameter settings, the proposed method outperformed the other methods. As the number of participants n was smaller than in the previous experiments, increasing the missing value rate increased the JS divergence of the proposed method. However, the increase in JS divergence is not considerable.

We next used a default dataset containing 21,985 records with the following attributes: sex, job, income, number of loans from other companies, number of delayed payments, and a default flag (0 or 1). Here, the word default means that a debtor failed to pay off a loan. The results of this dataset, which was generated from authentic default data, are plotted in Fig. 8. As shown in Fig. 8a, the proposed method accurately reconstructed the contingency tables even when the missing value ratio (m) increased to 0.8. On the contrary, the accuracies of the existing methods greatly decreased as the missing value ratio increased. Increasing the number of attributes used for generating contingency tables (c) also increased the reconstructed error (Fig. 8b). However, the proposed method was more resistant to increasing c than the other methods. Fig. 8c shows the effect of ϵ on the reconstruction error in the five methods. When ϵ was sufficiently large, the accuracies of all methods were very similar, but when ϵ was small, the reconstructed error was clearly lowest in the proposed method.

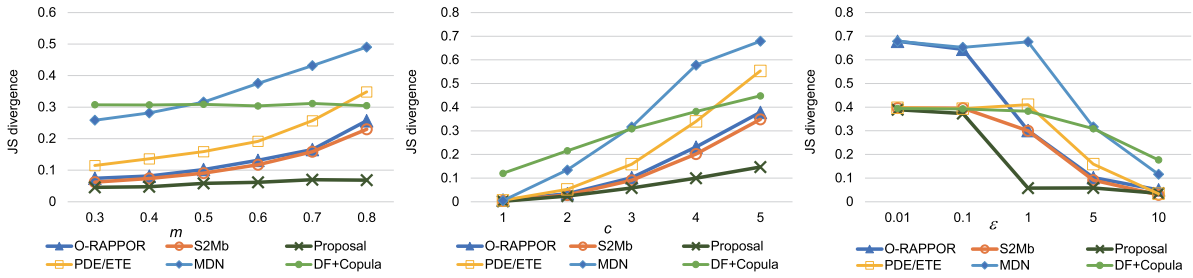
Finally, we applied a dataset related to the 2019 coronavirus disease (COVID-19) called Patient Medical Data for Novel Coronavirus COVID-19.[2] Hereafter, we refer to this dataset as the COVID-19 dataset. This dataset contains 427,036 records

2. https://datarepository.wolframcloud.com/resources/Patient-Medical-Data-for-Novel-Coronavirus-COVID-19/ (accessed June 20, 2020)



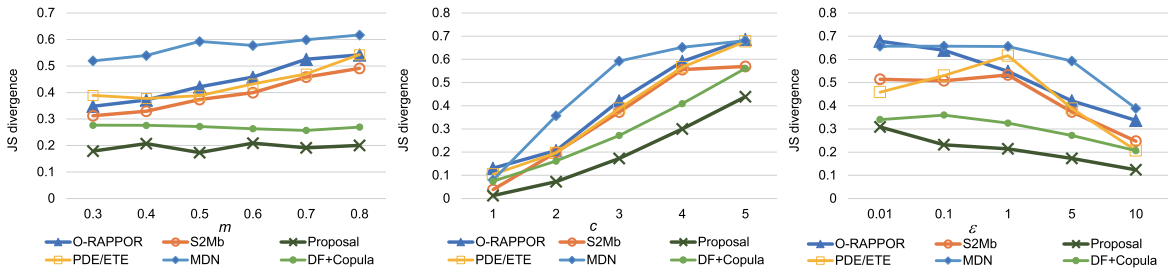
(a) $\epsilon = 5$ and $c = 3$. (b) $\epsilon = 5$ and $m = 0.5$. (c) $c = 3$ and $m = 0.5$.

Fig. 7. Results of the Community and Crime Unnormalized datasets.



(a) $\epsilon = 5$ and $c = 3$. (b) $\epsilon = 5$ and $m = 0.5$. (c) $c = 3$ and $m = 0.5$.

Fig. 8. Results of the default dataset.



(a) $\epsilon = 5$ and $c = 3$. (b) $\epsilon = 5$ and $m = 0.5$. (c) $c = 3$ and $m = 0.5$.

Fig. 9. Results of the Patient Medical Data for Novel Coronavirus COVID-19 dataset.

with 23 attributes. More than 90% of the values are missing for 12 attributes, and approximately 27% are missing even for basic attributes such as age and sex. From the COVID-19 dataset, we extracted the Japanese medical data and analyzed the attributes with few missing values (namely, age, sex, administrative division, date of confirmation, and chronic disease status). The date of confirmation was categorized by month and the number of categories in each attribute ranged from 2 to 29.

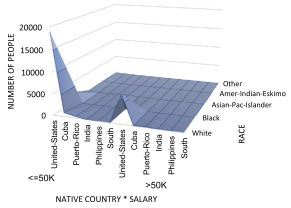
Fig. 9 presents the results of the COVID-19 dataset. Under all parameter settings, the JS divergence was lower in the proposed method than in the other methods. As the rate of missing values in the original COVID-19 dataset was 68.7%, we concluded that the proposed method effectively handles real datasets with missing values.

As examples, Figs. 10 and 11 respectively display histograms generated by the compared methods for combinations of the race, native country, and salary attributes in the Adult dataset and for combinations of the age, gender, and month of confirmation attributes in the COVID-19 dataset. Shown are the true histograms and those generated by O-RAPPOR, S2Mb, MDN, PDE/ETE and DF+Copula. On the Adult dataset, the generated histograms of O-RAPPOR and S2Mb did not differ greatly from the original histogram, but several values differed considerably from the true values. In contrast, the proposed method determined the true distribution almost perfectly, although several values contained errors. The proposed method also reconstructed the histogram of the COVID-19 dataset with high accuracy, whereas the histograms reconstructed by the existing methods noticeably differed from the true histogram.

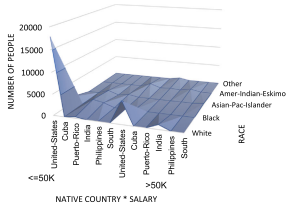
The execution time of the proposed system was 61 seconds on the Adult dataset (with 15 attributes and 32,561 records) and 1,399 seconds for an artificial dataset with 100 attributes and 10,000 records. As the proposed system calculates the occurrence frequencies of the attribute values for all pairs of attributes, its runtime obviously increased with number of attributes; nevertheless, the execution time on the artificial dataset was short enough for practical use.

5.4 Discussion of the Results of Experiments

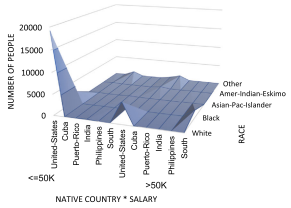
Existing private data collection methods (O-RAPPOR, S2Mb, PDE/ETE, MDN) do not consider missing values. For example, if the server wants to analyze the relationship between attributes A_1, A_2, and A_3, such methods only use data samples that have all these attribute values. For example, if the missing value rate is 0.5 among all attributes, the probability that three attributes have values is $(1 - 0.5)^3$=0.125. In other words, only 12.5% of data samples can be used to analyze these



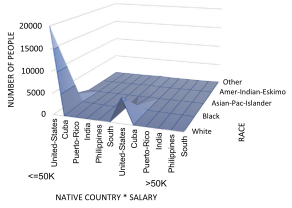
(a) Original histogram.



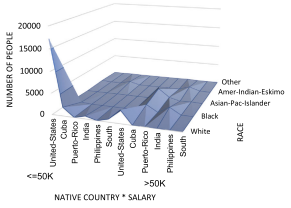
(b) Histogram for O-RAPPOR.



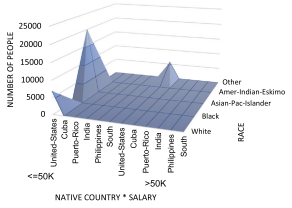
(c) Histogram for S2Mb.



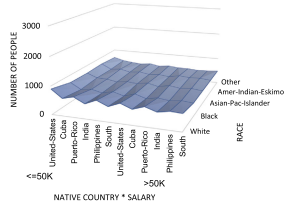
(d) Histogram for proposed method.



(e) Histogram for PDE/ETE.

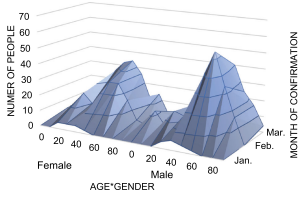


(f) Histogram for MDN.

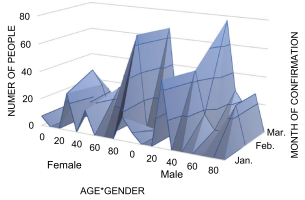


(g) Histogram for DF+Copula.

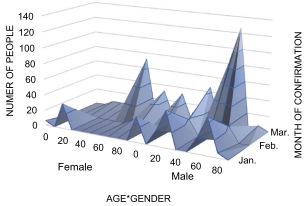
Fig. 10. Example of generated histograms of the race, native country, and salary attributes in the Adult dataset ($\epsilon = 5$, $m = 0.5$, $c = 3$).



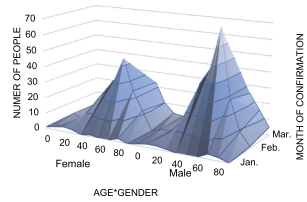
(a) Original histogram.



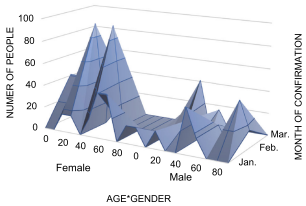
(b) Histogram for O-RAPPOR.



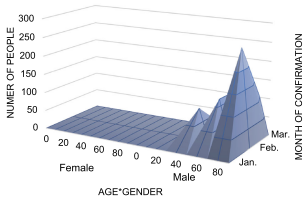
(c) Histogram for S2Mb.



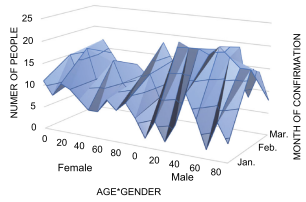
(d) Histogram for proposed method.



(e) Histogram for PDE/ETE.



(f) Histogram for MDN.



(g) Histogram for DF+Copula.

Fig. 11. Example of generated histograms of the age, gender, and confirmation month attributes in the COVID-19 dataset ($\epsilon = 5$, $m = 0.5$, $c = 3$).

attributes. The larger the value of m or the larger the value of c (which represents the number of targeted attributes for analysis), the greater the decrease in the probability that all target attributes have values (Figs. 4a–9a and Figs. 4b–9b.)

Differing from existing methods, the proposed method can reduce the effect of missing values because it estimates the attribute value distribution while mitigating the noise added by the differential privacy. The proposed method estimates single-attribute distribution and pair-attribute distribution, i.e., the relationship among three or more attributes does not need to be calculated to construct a copula model. Nevertheless, the constructed copula model can generate data samples that represent the true relationship between all attributes because the model is built to reproduce the characteristics of all pair-attribute distributions.

DF+Copula uses differentially private data as it is and constructs a copula model based on the differentially private data. Although each data sample has a significant amount of noise, true information is not completely lost. Therefore, DF+Copula could realize relatively high accuracy, particularly when m is large. However, DF+Copula does not mitigate the noise, and the accuracy is less than that of the proposed method.

In this experiment, we set the value of ϵ to between 0.01 and 10. An epsilon value of 0.01 is a very strict setting, while a value of 10 is a less privacy-preserving setting. Therefore, depending on the datasets, the accuracy of the proposed method and several other methods are similar when ϵ is 0.01 or 10. However, for all ϵ values, the accuracy of the proposed method is better or comparable to other methods. Regardless of the value of ϵ, it is evident that the proposed noise reduction technique has a positive impact.

6 Discussion

6.1 Kinds of Missing Values

In general, missing data can be categorized into the following three types:

- Missing completely at random (MCAR): MCAR signifies that the probability of an attribute value being observed or missed does not depend on any attributes.

- Missing at random (MAR): MAR signifies that the probability of an attribute value being observed or missed depends on other attributes but not on the missing attribute value itself.
- Missing not at random (MNAR): MNAR signifies that the probability of an attribute value being observed or missed depends on the missing attribute value itself.

In our experiments, we generated MCAR values. To our knowledge, we introduce the first privacy-preserving method for anonymized data collection with many missing values. The proposed method creates a Gaussian copula using the estimated value distributions of each attribute and of each pair of attributes. By combining the estimation with existing methods for missing value imputation of MAR and MNAR data (e.g., [67], [68]), our proposed method could be extended to MAR and MNAR data.

If all attributes are independent of each other, a multi-dimensional analysis is not required. Therefore, we assumed that at least several attributes depend on each other. However, the proposed algorithm can also be used when all attributes are independent. In this case, the proposed algorithm will perform similarly to the existing methods.

6.2 Treating Continuous Attributes

Continuous attributes in our method can be handled by two approaches. In the first approach, the continuous attribute values are discretized into several categories and the proposed algorithm is applied to the categorized values. For example, suppose that the domain of an attribute is [0, 10). When the domain is discretized into [0,0.1), [0.1,0.2),..., [9.9, 10), the attribute is divided into 100 categories. In the second approach, a continuous attribute is not discretized, but differential privacy is achieved by adding Laplace noise to each continuous attribute value [69]. The value distributions of each attribute and each pair of attributes can be reconstructed from a set of noise-added values generated with a certain probability distribution [5], [70]. These techniques can be used for determining the reconstructed value distribution in our method, which is needed for applying the Gaussian copula. Although the proposed method is applicable to continuous attributes, the present study considers discrete attributes to emphasize our approach.

6.3 Treating Massive Attributes

In massive-attribute cases, the privacy budget will be limited. This challenge is faced not only by the proposed method, but by all methods based on differential privacy. In our research, we allocated the same privacy budget to all attributes for technical convenience. However, reasonable privacy-budget allocation techniques such as [71], which can be used in the existing methods, can also be implemented in the proposed method.

6.4 Treating Data Streams

We here apply our method to data streams. Let t be the number of times that a differentially private value is reported by a user to the data collection server. In this case, the privacy budget of each report can be computed as ϵ/t. This naive solution decreases the utility at the data collection server. More sophisticated approaches such as those in [72], [73] could be applied to data streaming in our method.

The following sampling technique can be used. Even if the number of rounds is t, each person can send their differentially private value only $t'(< t)$ times because our proposed method is robust to missing values, as shown in the experimental results. In this case, the privacy budget is ϵ/t', and this value is larger than ϵ/t.

When a set of new data comes in, we can reapply the proposed method based only on the new data. We acknowledge that it may be possible to reduce the time required to generate a new copula model by reusing a copula model that has already been generated. We intend to investigate this possibility in future work.

Such techniques can be used for reconstructing the value distribution at each time stamp in our method, which is needed for applying the Gaussian copula. In this way, our proposed method can be applied to data streaming. Of course, these techniques can only partially prevent the usefulness degradation of the data.

Dwork *et al.* [74] recommended that the system clarifies the value of epsilon being used per datum, the lifetime of the data, cumulative privacy loss incurred before the data are retired, etc. In our research, these values can be freely determined under the agreement between the data collector and a person. Let ϵ_0 represent the value of epsilon being used per datum. Each person sends g attribute values to the data collector; hence, the total privacy loss is $g\epsilon_0$ for each report. If each person sends attribute values under $(g\epsilon_0)$-differential privacy h times during the lifetime of the data, the cumulative privacy loss incurred before the data are retired is $hg\epsilon_0$. In our experiments, we assumed h=1, and we set $\epsilon = g\epsilon_0$. For example, when the value of ϵ was set to 0.01 and g was 10, the value of ϵ_0 was set to 0.01/10. Further, we can obtain the result for $h > 1$. When h is set to 100, the value of ϵ_0 is reduced by a factor of 100. Therefore, the results obtained for $g = 10$, $h = 1$, and ϵ=0.01 can be considered equivalent to those obtained for $g = 10$, $h = 100$, and ϵ=1.0 because ϵ_0 is set to 0.001 in both cases.

If the person wants to keep the total privacy loss during the lifetime of the data below ϵ_{target} and sends their attribute values h times during the lifetime, each datum of their attribute value should be protected under (ϵ_{target}/hg)-differential privacy. In other words, if each person will send their differentially private value $t'(< t)$ times under ϵ_0-differential privacy for each attribute value, the value of t' should be set to $(\epsilon_{target}/\epsilon_0\ g)$ where $\epsilon = \epsilon_0\ g$.

6.5 Extensions of ϵ-Differential Privacy

In this study, we focused on ϵ-differential privacy and ϵ-local differential privacy. Several relaxations of ϵ-differential privacy (and ϵ-local differential privacy) have been proposed. Examples are the Gaussian differential privacy [75], concentrated differential privacy [76], Bayesian differential privacy [77], and Renyi differential privacy [78]. Nevertheless, many studies, including those based on differential private-federated learning (which generates a machine-learning model based on distributed data), still target ϵ- or (ϵ, δ)-differential privacy [79], [80], [81], [82], [83], [84]. In fact, ϵ-differential privacy is the fundamental concept underlying various differential privacy definitions and ensures stronger privacy than these relaxations of differential privacy [75],

[76], [77], [78]; accordingly, ϵ-differential privacy has been employed in many of the latest studies [82], [83], [84], [85].

Therefore, we focus on ϵ-differential privacy (and ϵ-local differential privacy) in the present study. In future work, (ϵ, δ)–differential privacy and other extensions of differential privacy will be considered.

6.6 Several Issues on Differential Privacy

In this study, it is assumed an individual provides data based on the protocol of the proposed method. However, there exists the threat of a manipulation attack in which an individual alters the proposed protocol with the goal of forcing the data collection server to draw false conclusions. Non-interactive local protocols, including our proposed protocol, are not robust to manipulation attacks, and their effectiveness will increase as privacy guarantees are strengthened. Cheu *et al.* stated that multiparty computation or shuffling might solve this problem [86]. Multiparty computation is a technology that requires long computation time but can process data confidentially [87]. Shuffling assumes a trusted shuffler exists, and the shuffler anonymizes the origins of individuals' messages [88], [89]. Using techniques for finding malicious input data in machine-learning models such as [90] is another option. Verification of whether these methods against manipulation attacks will also work for our proposed method is a future issue.

Determining the level of privacy protection in relation to business requirements is not an easy task. Dandekar *et al.* proposed an algorithm that takes the privacy–utility trade-off and minimizes the compensation budget [91].

Ding *et al.* stated that several differential privacy algorithms have bugs and they do not actually ensure differential privacy [92]. Although the method proposed in this paper was proven to ensure differential privacy, it is important to recognize that such problems exist.

7 Conclusion

Patient information is required for monitoring the status of patients' infections such as COVID-19. For this purpose, it is often collected and shared with researchers. Although privacy protection is a significant concern and necessitates privacy-protection techniques, excessive emphasis on privacy-protection processing stifles the data analysis. In addition, many patients elect not to provide personal information and some patients provide only some of their personal attributes values due to privacy concerns. In such cases, the collected privacy-protected data have many missing values. Several methods have been proposed for privacy-protection data mining, but these methods do not consider missing values. Consequently, the accuracy of data analysis is significantly reduced when the missing values are numerous.

In this paper, we inferred the value distributions of single attributes and combinations of two attributes, and generated a Gaussian copula. Because it uses the information about combinations of two attributes, the proposed method is robust to missing values in data. The generated Gaussian copula utilizes the information from all combinations of two attributes, which enhances its data reproducibility. On the real COVID-19 dataset, we demonstrated that the proposed method significantly reduces the JS divergence from those of the existing methods. In this study, we evaluated the proposed method on public data, but in future work, we expect to collect more sensitive attribute values using the proposed method.

Acknowledgments

We would like to thank the anonymous reviewers for their various insightful comments and suggestions.

References

[1] C. Mazza *et al.*, "A nationwide survey of psychological distress among italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors," *Int. J. Environ. Res. Public Health*, vol. 17, no. 9, 2020, Art. no. 3165.

[2] N. W. Chew *et al.*, "A multinational, multicentre study on the psychological outcomes and associated physical symptoms amongst healthcare workers during COVID-19 outbreak," *Brain, Behavior, Immun.*, vol. 88, pp. 559–565, 2020.

[3] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2436–2444.

[4] Y. Sei and A. Ohsuga, "Differential private data collection and analysis based on randomized multiple dummies for untrusted mobile crowdsensing," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 4, pp. 926–939, Apr. 2017.

[5] Y. Sei and A. Ohsuga, "Differentially private mobile crowd sensing considering sensing errors," *Sensors*, vol. 20, no. 10, pp. 2785:1–2785:25, 2020.

[6] J. Xu, A. Wang, J. Wu, C. Wang, R. Wang, and F. Zhou, "SPCSS: Social network based privacy-preserving criminal suspects sensing," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 1, pp. 261–274, Feb. 2020.

[7] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata Lang. Program.*, 2006, pp. 1–12.

[8] A. Roy Chowdhury, C. Wang, X. He, A. MacHanavajjhala, and S. Jha, "Cryptϵ: Crypto-assisted differential privacy on untrusted servers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2020, pp. 603–619.

[9] J. Abeler, M. Bäcker, U. Buermeyer, and H. Zillessen, "COVID-19 contact tracing and data protection can go together," *JMIR mHealth uHealth*, vol. 8, no. 4, 2020, Art. no. e19359.

[10] E. L. Brown, N. Ruggiano, J. Li, P. J. Clarke, E. S. Kay, and V. Hristidis, "Smartphone-based health technologies for dementia care: Opportunities, challenges, and current practices," *J. Appl. Gerontol.*, vol. 38, no. 1, pp. 73–91, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28774215/

[11] M. Pouryazdan, B. Kantarci, T. Soyata, L. Foschini, and H. Song, "Quantifying user reputation scores, data trustworthiness, and user incentives in mobile crowd-sensing," *IEEE Access*, vol. 5, pp. 1382–1397, 2017.

[12] A. Suliman, H. Otrok, R. Mizouni, S. Singh, and A. Ouali, "A greedy-proof incentive-compatible mechanism for group recruitment in mobile crowd sensing," *Future Gener. Comput. Syst.*, vol. 101, pp. 1158–1167, 2019.

[13] H. Kurasawa *et al.*, "Missing sensor value estimation method for participatory sensing environment," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 103–111.

[14] L. Cheng *et al.*, "Compressive sensing based data quality improvement for crowd-sensing applications," *J. Netw. Comput. Appl.*, vol. 77, pp. 123–134, 2017.

[15] Z. Wu *et al.*, "A location privacy-preserving system based on query range cover-up or location-based services," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5244–5254, May 2020.

[16] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: A survey," *IEEE Commun. Surv. Tuts.*, vol. 22, no. 1, pp. 746–789, Jan.–Mar. 2020.

[17] P. Zhao, G. Zhang, S. Wan, G. Liu, and T. Umer, "A survey of local differential privacy for securing internet of vehicles," *J. Supercomputing*, vol. 76, pp. 8391–8412, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s11227–019-03104-0

[18] Y. Wang, Z. Cai, Z. H. Zhan, B. Zhao, X. Tong, and L. Qi, "Walrasian equilibrium-based multiobjective optimization for task allocation in mobile crowdsourcing," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 1033–1046, Aug. 2020.

[19] 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.

[20] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[21] S. Agrawal and J. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proc. 21st Int. Conf. Data Eng.*, 2005, pp. 193–204.

[22] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2012, pp. 144–152.

[23] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Application and analysis of multidimensional negative surveys in participatory sensing applications," *Pervasive Mobile Comput.*, vol. 9, no. 9, pp. 372–391, 2013.

[24] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 251–262.

[25] R. Chaytor and K. Wang, "Small domain randomization: Same privacy, more utility," *Proc. VLDB Endow.*, vol. 3, no. 1/2, pp. 608–618, 2010.

[26] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2003, pp. 211–222.

[27] Y. Wang, M. Gu, J. Ma, and Q. Jin, "DNN-DP: Differential privacy enabled deep neural network learning framework for sensitive crowdsourcing data," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 1, pp. 215–224, Feb. 2020.

[28] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1322–1333.

[29] C. Park, D. Hong, and C. Seo, "An attack-based evaluation method for differentially private learning against model inversion attack," *IEEE Access*, vol. 7, pp. 124 988–124 999, 2019.

[30] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Secur. Privacy*, 2017, pp. 3–18.

[31] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, and W. Cheng, "SocInf: Membership inference attacks on social media health data with machine learning," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 5, pp. 907–921, Oct. 2019.

[32] R. Wei *et al.*, "Missing value imputation approach for mass spectrometry-based metabolomics data," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018. [Online]. Available: https://www.nature.com/articles/s41598–017-19120-0

[33] R. Deb and A. W. C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data," *Inf. Sci.*, vol. 339, pp. 274–289, 2016.

[34] M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 389–422, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s10115–015-0822-y

[35] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[36] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," *Expert Syst. Appl.*, vol. 115, pp. 68–94, 2019.

[37] P. S. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," *Soft Comput.*, vol. 24, no. 6, pp. 4361–4392, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s00500–019-04199-6

[38] Z. Zhang *et al.*, "PrivSyn: Differentially private data synthesis," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 929–946.

[39] G. Vietri, G. Tian, M. Bun, T. Steinke, and Z. S. Wu, "New oracle-efficient algorithms for private synthetic data release," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9707–9716.

[40] J. Li *et al.*, "Efficient and secure outsourcing of differentially private data publishing with multiple evaluators," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 1, pp. 67–76, Jan./Feb. 2022.

[41] F. Harder, K. Adamczewski, and M. Park, "DP-MERF: Differentially private mean embeddings with RandomFeatures for practical privacy-preserving data generation," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, 2021, vol. 130, pp. 1819–1827.

[42] K. Cai, X. Lei, J. Wei, and X. Xiao, "Data synthesis via differentially private markov random fields," *Proc. VLDB Endow.*, vol. 14, no. 11, pp. 2190–2202, 2021.

[43] M. Gaboardi, E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu, "Dual query: Practical private query release for high dimensional data," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 1170–1178. [Online]. Available: https://proceedings.mlr.press/v32/gaboardi14.html

[44] M. Hardt, K. Ligett, and F. Mcsherry, "A simple and practical algorithm for differentially private data release," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1–9.

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

[46] C. Genest and J. MacKay, "The joy of copulas: Bivariate distributions with uniform marginals," *Amer. Statistician*, vol. 40, no. 4, pp. 280–283, 1986.

[47] 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.

[48] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?," *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2013.

[49] C. F. J. Wu, "On the convergence properties of the EM algorithm on JSTOR," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983. [Online]. Available: https://www.jstor.org/stable/2240463?seq=1#metadata_info_tab_contents

[50] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 885–896, Oct. 1999.

[51] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Pub. de l'Institut Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959.

[52] H. Tozaka and T. Yoshiba, "Specific applications of copulas in financial practice," *Financial Res.*, vol. 24, no. 2, pp. 115–162, 2005.

[53] L. Garcia-Jorcano and S. Benito, "Studying the properties of the bitcoin as a diversifying and hedging asset through a copula analysis: Constant and time-varying," *Res. Int. Bus. Finance*, vol. 54, Dec. 2020, Art. no. 101300. [Online]. Available: /pmc/articles/PMC7395826/ /pmc/articles/PMC7395826/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395826/

[54] N. E. Lasmar and Y. Berthoumieu, "Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2246–2261, May 2014.

[55] E. Kole, K. Koedijk, and M. Verbeek, "Selecting copulas for risk management," *J. Bank. Finance*, vol. 31, no. 8, pp. 2405–2423, Aug. 2007.

[56] M. J. Wood and J. Ross-Kerr, *Basic Steps in Planning Nursing Research: From Question to Proposal*. Burlington, MA, USA: Jones & Bartlett Publishers, 2010.

[57] R. Grover and M. Vriens, *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*, Thousand Oaks, CA, USA: SAGE Publications, Inc, 2006.

[58] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on MacOS 10.12," 2017, *arXiv:1709.02753*.

[59] Differential Privacy Team Apple, "Learning with privacy at scale," *Apple Mach. Learn. J.*, vol. 1, no. 8, pp. 1–25, 2017.

[60] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3574–3583.

[61] J. Hsu *et al.*, "Differential privacy: An economic method for choosing epsilon," in *Proc. IEEE Comput. Secur. Foundations Symp.*, 2014, pp. 398–410.

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

[63] J. Jia and W. Qiu, "Research on an ensemble classification algorithm based on differential privacy," *IEEE Access*, vol. 8, pp. 93 499–93 513, 2020.

[64] F. Song, T. Ma, Y. Tian, and M. Al-Rodhaan, "A new method of privacy protection: Random k-anonymous," *IEEE Access*, vol. 7, pp. 75 434–75 445, 2019.

[65] C. Eyupoglu, M. Aydin, A. Zaim, and A. Sertbas, "An efficient Big Data anonymization algorithm based on chaos and perturbation techniques," *Entropy*, vol. 20, no. 5, pp. 373:1–373:18, May 2018. [Online]. Available: http://www.mdpi.com/1099–4300/20/5/373

[66] Communities and crime unnormalized data set, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Communities_and_Crime_Unnormalized

[67] A. B. Pedersen *et al.*, "Missing data and multiple imputation in clinical epidemiological research," *Clin. Epidemiol.*, vol. 9, pp. 157–166, 2017. [Online]. Available: /pmc/articles/PMC5358992/?report= abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5358992/

[68] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies," *J. Proteome Res.*, vol. 15, no. 4, pp. 1116–1125, 2016. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.5b00981
[69] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr.*, 2006, pp. 265–284.
[70] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439–450.
[71] X. Feng and C. Zhang, "Local differential privacy for unbalanced multivariate nominal attributes," *Hum.-centric Comput. Inf. Sci.*, vol. 10, no. 25, pp. 1–21, 2020. [Online]. Available: https://doi.org/10.1186/s13673–020-00233-x
[72] X. Xiong, S. Liu, D. Li, Z. Cai, and X. Niu, "Real-time and private spatio-temporal data aggregation with local differential privacy," *J. Inf. Secur. Appl.*, vol. 55, no. 102633, pp. 1–10, 2020.
[73] X. Fang, Q. Zeng, and G. Yang, "Local differential privacy for data streams," in *Proc. Int. Conf. Secur. Privacy Digit. Economy*, 2020, pp. 143–160. [Online]. Available: http://dx.doi.org/10.1007/978–981-15-9129-7_11
[74] C. Dwork and G. J. Pappas, "Privacy in information-rich intelligent infrastructure," A Computing Community Consortium, Tech. Rep., 2017. [Online]. Available: https://arxiv.org/abs/1706.01985
[75] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," 2019, *arXiv:1905.02383*. [Online]. Available: http://arxiv.org/abs/1905.02383
[76] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptogr. Conf.*, 2016, pp. 635–658.
[77] A. Triastcyn and B. Faltings, "Bayesian differential privacy for machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9583–9592.
[78] I. Mironov, "Rényi differential privacy," in *Proc. IEEE Comput. Secur. Foundations Symp.*, 2017, pp. 263–275.
[79] X. Huang, Y. Ding, Z. L. Jiang, S. Qi, X. Wang, and Q. Liao, "DP-FL: A novel differentially private federated learning framework for the unbalanced data," *World Wide Web*, vol. 23, no. 4, pp. 2529–2545, 2020.
[80] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private AirComp federated learning with power adaptation harnessing receiver noise," in *Proc. IEEE Glob. Commun. Conf.*, 2020, pp. 1–6.
[81] 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. 10 639–10 651, Jul. 2021.
[82] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2134–2143, Mar. 2020.
[83] Y. Zhao *et al.*, "Local differential privacy-based federated learning for Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, Jun. 2021.
[84] L. Sun, J. Qian, and X. Chen, "LDP-FL: Practical private aggregation in federated learning with local differential privacy," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1571–1578.
[85] Y. Wang, J. Lam, and H. Lin, "Consensus of linear multivariable discrete-time multiagent systems: Differential privacy perspective," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2021.3135933.
[86] A. Cheu, A. Smith, and J. Ullman, "Manipulation attacks in local differential privacy," in *Proc. IEEE Symp. Secur. Privacy*, 2021, pp. 883–900.
[87] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptographic Techn.*, 2006, pp. 486–503. [Online]. Available: https://link.springer.com/chapter/10.1007/11761679_29
[88] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Proc. Annu. Int. Cryptol. Conf.*, 2019, pp. 638–667.
[89] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, 2019, pp. 2468–2479.
[90] T. Chiba, Y. Sei, Y. Tahara, and A. Ohsuga, "A countermeasure method using poisonous data against poisoning attacks on IoT machine learning," *Int. J. Semantic Comput.*, vol. 15, no. 2, pp. 215–240, Jul. 2021. [Online]. Available: www.worldscientific.com
[91] A. Dandekar, D. Basu, and S. Bressan, "Differential privacy at risk: Bridging randomness and privacy budget," in *Proc. Privacy Enhancing Technol.*, 2020, pp. 64–84. [Online]. Available: https://arxiv.org/abs/2003.00973v2
[92] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2018, pp. 475–489. [Online]. Available: https://doi.org/10.1145/3243734.3243818

Yuichi Sei (Member, IEEE) received the PhD degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2009. From 2009 to 2012, he was with the Mitsubishi Research Institute, Tokyo, Japan. He joined the University of Electro-communications, Tokyo, Japan, in 2013, 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.

J. Andrew Onesimu received the bachelor of engineering (BE) degree in CSE, in 2011, the master of engineering (ME) degree from Anna University, Chennai, India, in 2013, and the PhD degree from the Vellore Institute of Technology (VIT), Vellore, India, in 2021. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering (CSE), Manipal Institute of Technology, Manipal, India. He is an active researcher published scientific research articles in reputed journals and conferences. He also served as a speaker with prestigious conferences worldwide. He is having more than 8 years of teaching experience with undergraduate (UG) and postgraduate (PG) levels. His research interests include privacy preserving data, healthcare data analysis, deep learning, machine learning, computer vision, and blockchain technologies.

Hiroshi Okumura received the PhD degree in economics from the Kobe University, Hyogo, Japan, in 2012. He is working with Mitsubishi Research Institute, Tokyo, Japan. His research interests include statistics, econometrics, and statistical machine learning. He is a member of Japan Statistical Society and Information Processing Society of Japan (IPSJ).

Akihiko Ohsuga (Member, IEEE) received the PhD degree in computer science from Waseda University, in 1995. From 1981 to 2007 he was with Toshiba Corporation. He joined the University of Electro-Communications, in 2007. 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 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). He has been a fellow of IPSJ since 2017. He served as a 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 received IPSJ Best Paper Awards, in 1987 and 2017.