Accuracy and Privacy Evaluations of Collaborative Data Analysis

Akira Imakura, Anna Bogdanova, Takaya Yamazoe, Kazumasa Omote, Tetsuya Sakurai.
1-1-1 Tennodai, Ibaraki, Tsukuba 305-8573, University of Tsukuba, Japan,
imakura@cs.tsukuba.ac.jp, bogdanova.anna.fw@u.tsukuba.ac.jp, s1920600@u.tsukuba.ac.jp, omote@risk.tsukuba.ac.jp,
sakurai@cs.tsukuba.ac.jp.

Abstract

Distributed data analysis without revealing the individual data has recently attracted significant attention in several applications. A collaborative data analysis through sharing dimensionality reduced representations of data has been proposed as a non-model sharing-type federated learning. This paper analyzes the accuracy and privacy evaluations of this novel framework. In the accuracy analysis, we provided sufficient conditions for the equivalence of the collaborative data analysis and the centralized analysis with dimensionality reduction. In the privacy analysis, we proved that collaborative users' private datasets are protected with a double privacy layer against insider and external attacking scenarios.

Introduction

Background

Recently, the problem of real-life data availability for machine learning and data analysis applications came to the forefront of actual research challenges. In particular, use-cases that pertain to sensitive personal information or corporate secrecy can benefit from the ability to process distributed data without revealing it to other parties.

Various methods have been proposed over recent years, involving sharing a machine learning model that is collectively trained among several parties. In the present work, we analyze an alternative method of distributed and privacy-preserving data analysis that does not require sharing the machine learning model. The non-model-sharing approach has certain advantages over model-sharing methods: (a) maintaining the secrecy of particular model architecture; (b) protection from model poisoning attacks; (c) avoiding iterative communications necessary for machine learning training; (d) option to outsource data analysis to a third party or a data analysis competition.

The collaborative data analysis considered in this paper had been previously proposed in (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020). However, the method's proper accuracy and privacy analysis were lacking. In present work, we fill in this gap by providing conditions for equivalence of the data analysis in centralized and distributed settings, as well as conducting a thorough privacy analysis and disclosure risk evaluation of the collaborative data analysis.

Main purposes and contributions

The accuracy and privacy analyses are essential in practical use of the collaborative data analysis. In this paper, we analyze the equivalence of the collaborative data analysis and the centralized analysis with dimensionality reduction. We also analyze the privacy of the private dataset against insider and external attacking scenarios.

The main contributions of this paper are summarized as follows:

- We provided the sufficient condition for equivalence of the collaborative data analysis and the centralized analysis with dimensionality reduction.
- We proved that, in the collaborative data analysis, the private dataset is protected based on a double privacy layer against insider and external attacking scenarios.
- We demonstrated numerical evaluations for the accuracy and privacy analyses.

Related Work

The problem of deriving insights from data while maintaining the privacy of individual data records was first addressed in the context of Data Mining and Knowledge Discovery in Databases (KDD) (Agrawal and Srikant 2000) and consequently formed a large body of literature known as Privacy-Preserving Data Mining (PPDM). This field studies data-sanitizing operations, which can offer quantifiable privacy guarantees, while maintaining data utility for a variety of downstream analytical tasks, including supervised and unsupervised machine learning (Mendes and Vilela 2017).

As there are multiple definitions of what constitutes privacy and how it should be measured, distinct privacy guarantees and methods of privacy production became known in the literature as privacy models. Most influential privacy models that emerged from PPDM are k-anonymity, proposed by Samarati and Sweeny (1998), and ε-differential privacy, introduced by Dwork (2008). K-anonymity protects users' data from linkage attacks, ensuring that released data has at least k identical records. ε-differential privacy, on the other hand, guarantees that the inclusion of any record in

the dataset will not change the output of data mining operations by more than ε, thus preventing membership inference attacks. Both privacy models are theoretically sound, deployed in practice, and legislatively recognized. However, there are significant shortcomings that call for the development of alternative notions of privacy. Thus, k-anonymity is proven to be NP-hard (Verykios et al. 2004) and not attainable on high dimensional and sparse datasets (Narayanan and Shmatikov 2006). Similarly, providing record-level ε-differential privacy is not suitable for modern deep learning applications (Zhao, Chen, and Zhang 2019), as the amount of perturbation required diminishes data utility.

With the advancement of highly parameterized machine learning models, the focus of data privacy research shifted towards designing model architectures and sanitizing model parameters to enable Privacy-Preserving Machine Learning (PPML). One particularly successful approach had been Federated Learning proposed by McMahan et al. (2017). It is a machine learning framework that allows distributed training of deep learning models through the averaging of gradient descent steps taken on private datasets. Analogous algorithms were introduced for numerous machine learning models and various distributed settings, forming what became known as Federated Learning Systems (Li, Wen, and He 2019). Although currently, PPML cannot provide a formal privacy guarantee, as PPDM does, it satisfies privacy requirements through the data minimization approach, by sharing only the information necessary to the particular analytical task (Kairouz et al. 2019). Additionally, PPML is often combined with encryption schemes to prevent inference from intermediary results of the computation.

The collaborative data analysis (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020), the method considered in this paper, is positioned in between the two approaches to privacy-preserving data analysis. It shares with PPDM the focus on data transformation and the release of sample-vise information, which can be further explored for hidden relations and patterns. At the same time, it employs the information-minimization approach of PPML through the dimensionality reduction operation on the original data. Moreover, the shared intermediate representations can be formed by hidden layers of deep neural networks, strongly relating our method to the PPML domain.

Privacy-preserving properties of dimensionality reduction were previously explored in several papers. Thus, Tai et el. (2018), demonstrated that dimensionality reduction on average increases the value of k in k-anonymity privacy model, although it does not perform it reliably. Similarly, Nguyen and colleagues (2020) developed a ε-DR privacy framework of measuring the information loss of dimensionality reduction operations analogous to ε-differential privacy. Formal privacy guaranteed were demonstrated for particular methods of dimensionality reduction, such as non-metric multidimensional scaling (MDS) (Alotaibi et al. 2012) and random projections (Liu, Kargupta, and Ryan 2005). Moreover, specialized methods of dimensionality reduction were developed to satisfy certain privacy models, for instance differential-private Principal Component Analysis, and differential-private Linear Discriminant Analysis (Jiang et al. 2013). Since Data Collaboration method assumes an arbitrary dimensionality reduction function applied at the user's side, in practical applications such methods can be chosen to satisfy necessary privacy standards.

To the best of our knowledge, Data Collaboration is the only method so far offering collaborative data analysis through sharing dimensionality reduced representations of data and integration of such representations in a unified subspace. In this work, for the first time we propose privacy guarantees as well as utility measures of the transformed collaborative representations of data.

Collaborative data analysis

Distributed data analysis

Let m and n denote the numbers of features and training data samples. In addition, let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times m}$ and $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times \ell}$ be the training dataset and the corresponding ground truth. The n data samples are partitioned into c parties as follows:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_c \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_c \end{bmatrix}. \tag{1}$$

Then, the i-th party has partial dataset and the corresponding ground truth,

$$X_i \in \mathbb{R}^{n_i\times m}, \quad Y_i \in \mathbb{R}^{n_i\times \ell}.$$

A motivating example could be found in distributed medical data analysis. An analysis only using the dataset in each medical institution, i.e., *individual analysis* may not be sufficient for generating a high-quality prediction result due to insufficiency and imbalance of the data samples. If we can centralize the datasets from multiple institutions and analyze them as one dataset, i.e., *centralized analysis*, then we expect to achieve a high-quality prediction. However, it is difficult to centralise the original medical data samples with those from other institutions due to confidentiality concerns. Such kind of distributed data analysis is also essential in other applications, e.g., financial and manufacturing data analysis.

Outline of the collaborative data analysis

The collaborative data analysis has been proposed in (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020) as a method of distributed data analysis. A practical operation strategy regarding privacy and confidentiality concerns is also introduced. Here, we briefly introduce the algorithm based on the practical operation strategy.

In the practical operation strategy, the collaborative data analysis is operated by two roles: *user* and *analyst*. The users have the private dataset X_i and the corresponding ground truth Y_i and want to analyze them without sharing X_i. Each user individually constructs a dimensionality reduced intermediate representation and centralize it to analyst. To allow each user to use individual function for generating the intermediate representation, analyst transforms again the centralized intermediate representations to an incorporable form

called *collaboration representations*. For constructing the incorporable collaboration representations, users generate a shareable *anchor dataset* and centralize its intermediate representation to analyst. Then, the collaborative representation is analyzed as one dataset.

Training phase First, all users generate the same anchor dataset $X^{\mathrm{anc}} \in \mathbb{R}^{r\times m}$, which is a shareable data consisting of public data or dummy data randomly constructed, and partition it by features. Then, each user constructs the intermediate representations,

$$\widetilde{X}_i = f_i(X_i) \in \mathbb{R}^{n_i\times \widetilde{m}_i}, \quad \widetilde{X}_i^{\mathrm{anc}} = f_i(X^{\mathrm{anc}}) \in \mathbb{R}^{r\times \widetilde{m}_i},$$

where f_i denotes a linear or nonlinear row-wise mapping function and centralize the intermediate representations to the analyst. A typical setting for f_i is a dimensionality reduction, with $\widetilde{m}_i < m$, including unsupervised methods (Pearson 1901; He and Niyogi 2004; Maaten and Hinton 2008) and supervised methods (Fisher 1936; Sugiyama 2007; Li et al. 2017; Imakura et al. 2019). For privacy and confidentiality concerns, the function f_i should be set as

- The private data X_i can be obtained only if anyone has both the corresponding intermediate representation $\widetilde{X}_i$ and the mapping function f_i or its approximation.
- The mapping function f_i can be inferred only if anyone has both the input and output of f_i.

At the analyst side, the mapping function g_i for the collaboration representation is constructed satisfying

$$\widehat{X}_i^{\mathrm{anc}} = g_i(\widetilde{X}_i^{\mathrm{anc}}) \in \mathbb{R}^{r\times \widehat{m}} \quad \text{s.t. } \widehat{X}_i^{\mathrm{anc}} \approx \widehat{X}_{i'}^{\mathrm{anc}} \quad (i \neq i'),$$

in some sense. For computing g_i, authors of (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020) introduced a practical method via a total least squares problem when g_i is linear and also indicated an idea when g_i is nonlinear.

Then, the obtained collaboration representations $\widehat{X}_i = g_i(\widetilde{X}_i)$ can be analyzed as one dataset,

$$\widehat{X} = [\widehat{X}_1^{\mathrm{T}}, \widehat{X}_2^{\mathrm{T}}, \ldots, \widehat{X}_c^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n\times \widehat{m}},$$

with the shared ground truth Y_i using some supervised machine learning and the deep learning methods. The functions g_i and h are returned to the i-th user.

Prediction Phase Let $X_i^{\mathrm{test}} \in \mathbb{R}^{s_i\times m}$ be a test dataset of the i-th party. Then, for prediction phase, the predictive result Y_i^{test} of X_i^{test} is obtained by

$$Y_i^{\mathrm{test}} = h(g_i(f_i(X_i^{\mathrm{test}})))$$

via the intermediate and collaboration representations.

Accuracy analysis

We analyze the accuracy of the collaborative data analysis for the simple case that the mapping functions f_i and g_i are linear, i.e.,

$$\widetilde{X}_i = f_i(X_i) = X_iF_i, \quad F_i \in \mathbb{R}^{m\times \widetilde{m}} \quad (\mathrm{rank}(F_i) = \widetilde{m}),$$
$$\widehat{X}_i = g_i(\widetilde{X}_i) = \widetilde{X}_iG_i, \quad G_i \in \mathbb{R}^{\widetilde{m}\times \widetilde{m}} \quad (\mathrm{rank}(G_i) = \widetilde{m}).$$

Here, for simplicity, we assume that the dimensionality of $\widetilde{X}_i$ does not depend on i. Also, the matrices G_i are computed as introduced in (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020), that is,

$$\min_{G_i\in\mathbb{R}^{\widetilde{m}_i\times\widehat{m}}} \sum_{i=1}^{c} \|Z - \widetilde{X}_i^{\mathrm{anc}}G_i\|_{\mathrm{F}}^2, \tag{2}$$

where $Z \in \mathbb{R}^{r\times\widetilde{m}}$ is set as a column orthogonal matrix whose columns are the left singular vectors corresponding to the $\widetilde{m}$ largest singular values of a matrix

$$[\widetilde{X}_1^{\mathrm{anc}}, \widetilde{X}_2^{\mathrm{anc}}, \ldots, \widetilde{X}_c^{\mathrm{anc}}] = X^{\mathrm{anc}}[F_1, F_2, \ldots, F_c].$$

Theoretical evaluation for accuracy

In this paper, we analyze the accuracy of the collaborative data analysis compared with the centralized analysis with dimensionality reduction $B \in \mathbb{R}^{m\times\widetilde{m}}$ based on the norm

$$\left\| XB - \begin{bmatrix} X_1F_1G_1 \\ X_2F_2G_2 \\ \vdots \\ X_cF_cG_c \end{bmatrix} \right\|_{\mathrm{F}}^2 / \|X\|_{\mathrm{F}}^2$$
$$= \sum_{i=1}^{c} \|X_iB - X_iF_iG_i\|_{\mathrm{F}}^2 / \|X\|_{\mathrm{F}}^2. \tag{3}$$

With the anchor dataset X^{anc} preserving statistics of X, we evaluate the accuracy (3) by

$$\begin{aligned}
&\sum_{i=1}^{c} \|X_iB - X_iF_iG_i\|_{\mathrm{F}}^2 / \|X\|_{\mathrm{F}}^2 \\
&\quad\approx \sum_{i=1}^{c} \|X^{\mathrm{anc}}B - X^{\mathrm{anc}}F_iG_i\|_{\mathrm{F}}^2 / (c\|X^{\mathrm{anc}}\|_{\mathrm{F}}^2) \\
&\quad\leq \frac{\sum_{i=1}^{c} \|X^{\mathrm{anc}}B - Z\|_{\mathrm{F}}^2 + \|Z - X^{\mathrm{anc}}F_iG_i\|_{\mathrm{F}}^2}{c\|X^{\mathrm{anc}}\|_{\mathrm{F}}^2}.
\end{aligned}$$

Under the assumption that f_i are linear, we have $\widetilde{X}_i = X_iF_i$, where $X_i \in \mathbb{R}^{n_i\times m}$ and $F_i \in \mathbb{R}^{m\times\widetilde{m}}$. Let $F = [F_1, F_2, \ldots, F_c]$ and

$$X^{\mathrm{anc}}F = U\Sigma V^{\mathrm{T}} = [U_1, U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^{\mathrm{T}} \\ V_2^{\mathrm{T}} \end{bmatrix},$$
$$F = U_F\Sigma_F V_F^{\mathrm{T}} = [U_{F1}, U_{F2}] \begin{bmatrix} \Sigma_{F1} & \\ & \Sigma_{F2} \end{bmatrix} \begin{bmatrix} V_{F1}^{\mathrm{T}} \\ V_{F2}^{\mathrm{T}} \end{bmatrix}$$

be singular value decompositions of matrices $X^{\mathrm{anc}}F$ and F. Here, $\Sigma_1, \Sigma_{F1} \in \mathbb{R}^{\widetilde{m}\times\widetilde{m}}$ are the diagonal matrices corresponding to $\widetilde{m}$ largest singular values. Note that $Z = U_1$. Then, we have

$$\begin{aligned}
\|\Sigma_2\|_{\mathrm{F}}^2 &= \min_{\mathrm{rank}(\widetilde{X})=\widetilde{m}} \|X^{\mathrm{anc}}F - \widetilde{X}\|_{\mathrm{F}}^2 \\
&\leq \|X^{\mathrm{anc}}F - X^{\mathrm{anc}}U_{F1}\Sigma_{F1}V_{F1}^{\mathrm{T}}\|_{\mathrm{F}}^2 \\
&\leq \|X^{\mathrm{anc}}\|_{\mathrm{F}}^2 \|F - U_{F1}\Sigma_{F1}V_{F1}^{\mathrm{T}}\|_{\mathrm{F}}^2 \\
&= \|X^{\mathrm{anc}}\|_{\mathrm{F}}^2 \|U_{F2}\Sigma_{F2}V_{F2}^{\mathrm{T}}\|_{\mathrm{F}}^2 \\
&= \|X^{\mathrm{anc}}\|_{\mathrm{F}}^2 \|\Sigma_{F2}\|_{\mathrm{F}}^2.
\end{aligned}$$

Algorithm 1 Collaborative data analysis

Input (for user side): $X_i \in \mathbb{R}^{n_i \times m}, Y_i \in \mathbb{R}^{n_i \times \ell}, X_i^{\mathrm{test}}$, individually
Output (for user side): Y_i^{test} $(i = 1, 2, \ldots, c)$.

	user side (i)		*analyst side*
	———- Training phase ———		
1:	Generate X_i^{anc} and share to all users		
2:	Set X^{anc}		
3:	Generate f_i		
4:	Compute $\widetilde{X}_i = f_i(X_i)$ and $\widetilde{X}_i^{\mathrm{anc}} = f_i(X^{\mathrm{anc}})$		
5:	Share $\widetilde{X}_i, \widetilde{X}_i^{\mathrm{anc}}$ and Y_i to analyst	$\rightarrow$	Get $\widetilde{X}_i, \widetilde{X}_i^{\mathrm{anc}}$ and Y_i for all i
6:			Construct g_i from $\widetilde{X}_i^{\mathrm{anc}}$ for all i
7:			Compute $\widehat{X}_i = g_i(\widetilde{X}_i)$ for all i
8:			Set $\widehat{X}$ and Y
9:			Analyze $\widehat{X}$ and get h as $Y \approx h(\widehat{X})$
10:	Get g_i and h	$\leftarrow$	Return g_i and h to user
	———- Prediction phase ———		
11:	Compute $Y_i^{\mathrm{test}} = h(g_i(f_i(X_i^{\mathrm{test}})))$		

Using this inequality, the norm (2) can be bounded by

$$\begin{aligned}
&\min_{G_i} \sum_{i=1}^{c} \|Z - X^{\mathrm{anc}} F_i G_i\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^{c} \min_{G_i} \|Z - X^{\mathrm{anc}} F_i G_i\|_{\mathrm{F}}^2 \\
&\le \sum_{i=1}^{c} \|G_i\|_{\mathrm{F}}^2 \min_{G_i^{-1}} \|Z G_i^{-1} - X^{\mathrm{anc}} F_i\|_{\mathrm{F}}^2 \\
&\le \left(\max_i \|G_i\|_{\mathrm{F}}^2\right) \min_{G_i^{-1}} \|Z[G_1^{-1}, G_2^{-1}, \ldots, G_c^{-1}] \\
&\qquad\qquad - X^{\mathrm{anc}}[F_1, F_2, \ldots, F_c]\|_{\mathrm{F}}^2 \\
&= \left(\max_i \|G_i\|_{\mathrm{F}}^2\right) \|\Sigma_2\|_{\mathrm{F}}^2 \\
&\le \left(\max_i \|G_i\|_{\mathrm{F}}^2\right) \|X^{\mathrm{anc}}\|_{\mathrm{F}}^2 \|\Sigma_{F2}\|_{\mathrm{F}}^2.
\end{aligned}$$

Therefore, the accuracy of the collaborative data analysis (3) can be evaluated by $\|\Sigma_2\|_{\mathrm{F}}^2$ and $\|\Sigma_{F2}\|_{\mathrm{F}}^2$. Note that the value $\|\Sigma_2\|_{\mathrm{F}}^2$ can be obtained at the analyst side.

This bound provides the following theorem.

Theorem 1. *If the mapping functions satisfy*

$$\mathcal{R}(F_1) = \mathcal{R}(F_2) = \cdots = \mathcal{R}(F_c), \quad \mathrm{rank}(X^{\mathrm{anc}} F_i) = \widetilde{m}, \tag{4}$$

the predictive results of the collaborative data analysis is mathematically equivalent to that of the centralized analysis with dimensionality reduction $F_1 G_1$.

Proof. The condition (4) provides $\Sigma_{F2} = O$, then we have

$$F_1 G_1 = F_2 G_2 = \cdots = F_c G_c,$$

that proves the theorem. □

Numerical evaluation for accuracy

Here, we provide numerical evaluation of the accuracy analysis. We used a 10-class classification of handwritten digits (MNIST) (LeCun 1998), where $m = 784$. We set the number of parties as $c = 4$ and the number of samples for each party as $n_i = 50$.

Let $B \in \mathbb{R}^{784 \times 25}$ be a mapping function generated by PCA using all the training dataset X. We then set

$$F_i = B E_i^{(1)} + \varepsilon \|B\|_{\mathrm{F}} E_i^{(2)}, \quad i = 1, 2, \ldots, c,$$

with $E_i^{(1)} \in \mathbb{R}^{25 \times 25}$ and $E_i^{(2)} \in \mathbb{R}^{784 \times 25}$ whose entries are normally distributed random numbers. We used a kernel version of ridge regression (K-RR) (Saunders, Gammerman, and Vovk 1998) with a Gaussian kernel for analyzing the collaboration representation. The bandwidth σ of the Gaussian kernel is set based on the local scaling (Zelnik-Manor and Perona 2005). We set the regularization parameter for K-RR to $\lambda = 0.1$. The anchor data X^{anc} is constructed as a random matrix and $r = 2,000$. Then, we evaluate the following four values,

$$\begin{aligned}
\tau_1 &= \|\Sigma_2\|_{\mathrm{F}} / \|\Sigma_1\|_{\mathrm{F}}, \\
\tau_2 &= \|\Sigma_{F2}\|_{\mathrm{F}} / \|\Sigma_F\|_{\mathrm{F}}, \\
\tau_3 &= \left\| X F_1 G_1 - \begin{bmatrix} X_1 F_1 G_1 \\ X_2 F_2 G_2 \\ \vdots \\ X_c F_c G_c \end{bmatrix} \right\|_{\mathrm{F}} / \|X F_1 G_1\|_{\mathrm{F}}, \\
\tau_4 &= 1 - \mathrm{NMI}(Y_{\mathrm{CDA}}^{\mathrm{test}}, Y_{\mathrm{CA}}^{\mathrm{test}}),
\end{aligned}$$

where, in τ_4, the value $\mathrm{NMI}(Y_{\mathrm{CDA}}^{\mathrm{test}}, Y_{\mathrm{CA}}^{\mathrm{test}}) \in [0, 1]$ denotes the normalized mutual information (NMI) between the prediction results of test dataset X^{test} of the collaborative data analysis and the centralized analysis with dimensionality reduction $F_1 G_1$.

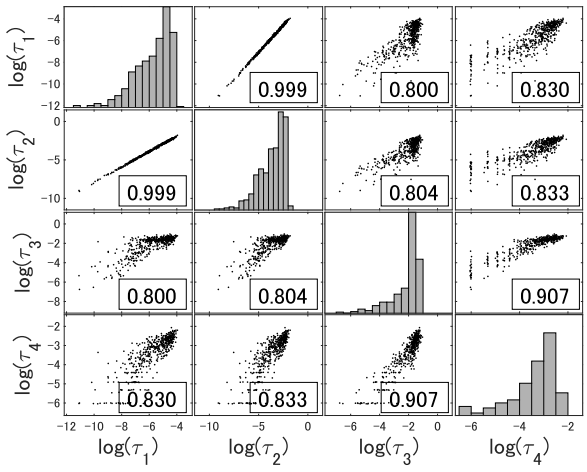

Figure 1: Scatter plot matrix and correlation coefficients for $\log(\tau_i)$.

All the numerical experiments were performed on Windows 10 Pro, Intel(R) Core(TM) i7-10710U CPU @ 1.10GHz, 16GB RAM using MATLAB2019b.

First, we perform the methods 10 times with $\varepsilon = 0$, that is $\mathcal{R}(F_i) = \mathcal{R}(F_{i'})$, but $F_i \neq F_{i'}$ for $i \neq i'$. The obtained average values are

$$\tau_1 = 8.42 \times 10^{-16}, \quad \tau_2 = 2.24 \times 10^{-16},$$
$$\tau_3 = 1.44 \times 10^{-13}, \quad \tau_4 = 0.00,$$

that mean the collaborative data analysis obtains the same result as the centralized analysis with the dimensionality reduction, i.e., $Y^{\text{test}}_{\text{CDA}} = Y^{\text{test}}_{\text{CA}}$. This result supports Theorem 1.

Next, we perform the methods 500 times with random $\varepsilon \in [10^{-2}, 10^{-6}]$ and evaluate correlation coefficients of $\log(\tau_i)$. Figure 1 shows the scatter plot matrix and correlation coefficients for $\log(\tau_i)$, which demonstrates that the accuracy of the collaborative data analysis regarding τ_4 has a strong correlation between τ_1, τ_2 and τ_3. Therefore, from this result, we observed that the accuracy of the collaborative data analysis τ_4 can be evaluated well by τ_1 in practice. Note that τ_1 can be obtained at the analyst side.

We can also observed from Figure 2 that τ_4 is roughly bounded by

$$\tau_4 \leq c\sqrt{\tau_1} \quad \left(\Leftrightarrow \log(\tau_4) \leq \frac{\log(\tau_1)}{2} + \log(c)\right)$$

with some constant c. Note that $c = \exp(0.5)$ in Figure 2.

Remarks on accuracy analysis

From the above analysis, we observed that

- If F_i satisfy the sufficient condition (4), the collaborative data analysis achieves the same result of the centralized analysis with dimensionality reduction.
- The accuracy of the collaborative data analysis compared with the centralized analysis with dimensionality reduction can be evaluated by $\|\Sigma_2\|_{\text{F}}$ in practice.

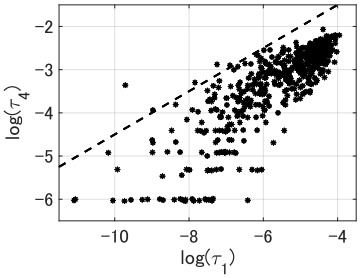

Figure 2: Scatter plot of $\log(\tau_4)$ v.s. $\log(\tau_1)$ and its rough bound $\log(\tau_1)/2 + 0.5$.

Note that, in order to obtain (approximately) the same predictive results as the centralized analysis with dimensionality reduction, we do not need to use the same mapping functions f_i, but use different functions with (approximately) the same range space.

Privacy analysis

For the analysis, this paper considers the privacy of the private data X_i of each user in the collaborative data analysis. Note that any information of the test data X_i^{test} does not have to be shared to others; see Algorithm 1.

Attacks for the data X_i can be classified into (i) attacks to infer the characteristics of the training data; (ii) attacks to infer the training data X_i itself; and (iii) attacks to infer whether a data sample is in the training dataset or not, so-called the membership inference attack.

This paper considers the privacy of the data X_i itself, rather than the characteristics of the data. We also shortly discuss the privacy against the membership inference attack.

Privacy definitions of dimensionality reduction: ε-DR privacy

Here, we introduce two dimensionality reduction (DR) privacy definitions: ε-DR privacy introduced in (Nguyen et al. 2019) and its variant, to evaluate the degree to which privacy is preserved through dimensionality reduction. Let $f : \boldsymbol{x} \in \mathbb{R}^m \to \widetilde{\boldsymbol{x}} \in \mathbb{R}^{\widetilde{m}}$ $(m > \widetilde{m})$ be a dimensionality reduced function and $f^\dagger$ be a reconstruction function of f. Then, we evaluate degree of privacy preservation using $\text{dist}(\boldsymbol{x}, \boldsymbol{x}')$ with a certain distance measure $\text{dist}(\cdot, \cdot)$, where $\boldsymbol{x}' = f^\dagger(f(\boldsymbol{x}))$.

The function f satisfies the ε-DR privacy regarding the expected value, if we have

$$E[\text{dist}(\boldsymbol{x}, \boldsymbol{x}')] \geq \varepsilon_1 \tag{5}$$

for each i.i.d. input sample $\boldsymbol{x}$. The value ε_1 depends on f and is always larger than 0 with $m > \widetilde{m}$.

Also, a function f satisfies the ε-DR privacy regarding a sample set $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, if we have

$$\min_{\boldsymbol{x} \in \mathcal{X}} \text{dist}(\boldsymbol{x}, \boldsymbol{x}') \geq \varepsilon_2. \tag{6}$$

The value ε_2 depends on f and a set of samples $\mathcal{X}$. Therefore, since ε depends on $\mathcal{X}$, ε_2 is possible to be 0 even if $m > \widetilde{m}$.

Attacking scenarios

In this paper, we consider the following two attacking scenarios: insider and external attacks.

- *Insider attacks*. Here, the users and analyst will strictly follow the strategy, but they try to infer the private data X_i.
- *External attacks*. Here, we consider a man-in-the-middle attack scenario where an attacker eavesdrops the information exchanged among users and analyst and infer the private data X_i.

Privacy against the honest-but-curious analyst

Theorem 2. *For the collaborative data analysis, an honest-but-curious analyst cannot infer the private dataset X_i of the users, unless analyst does not collude with user(s).*

Proof. For a privacy of X_i against the honest-but-curious analyst, each user shares the intermediate representations $\widetilde{X}_i$ and $\widetilde{X}_i^{\mathrm{anc}}$ to analyst. Here, we consider the possibility of recovering X_i from $\widetilde{X}_i$ and $\widetilde{X}_i^{\mathrm{anc}}$.

If analyst has a mapping function f_i, analyst can infer X_i by solving

$$\widetilde{X}_i = f_i(X_i).$$

However, the function f_i is private in the collaborative data analysis and also cannot be inferred by analyst, because analyst only has the output of f_i, that is the intermediate representations, but has no input data of f_i, if analyst does not collude with user(s). Note that the function f_i is constructed by some dimensionality reduction method with the private data X_i. The function f_i depends on X_i, so even if the dimensionality reduction method is identified, f_i itself cannot be inferred.

Thus, analyst cannot obtain the private data X_i from the intermediate representations, that proves the theorem. □

Privacy against the honest-but-curious users

Theorem 3. *For the collaborative data analysis, an honest-but-curious users cannot infer the private dataset X_i of a particular user against collusion of up to $c-2$ users.*

Proof. For a privacy of X_i against other users, each user shares the local anchor data X_i^{anc} to other users. Also, each user obtain the functions g_i and h from analyst that is constructed based on information of private data of other users.

First, we consider the possibility of recovering the private data from X_i^{anc}. The local anchor data does not contain the original X_i, but may preserve some useful information. Here, the local anchor data is constructed by users themselves using e.g., GAN and autoencoder with data augmentation. Users can control the containing information although it may have a trade-off relationship between the performance. Note that the collaborative data analysis works well even using random anchor data as demonstrated in (Imakura and Sakurai 2020; Imakura, Ye, and Sakurai 2020). Therefore, users cannot obtain the private information of X_i from X_i^{anc}.

Next, we consider the possibility of recovering the private data from g_i and h. When $c-1$ users $i \neq i'$, where the total number of users is c, are malicious and they collude together to retrieve information of the private dataset of the remaining (victim) user i', the colluding users have the function h and all X_i, f_i, g_i $(i \neq i')$. In this case, the function g_i $(i \neq i')$ and h are constructed by X_i, f_i $(i \neq i')$ of the colluding users and $X_{i'}$ of the victim user. Therefore, the private data $X_{i'}$ of the victim user will be inferred by solving an inverse problem. On the other hand, when the $c-2$ users collude, the obtained functions g_i and h of the colluding users are affected by remaining two users with equal importance. Therefore, users cannot infer each private dataset X_i of the victim users.

Thus, an honest-but-curious users cannot infer the private dataset of a particular user against collusion of up to $c-2$ users, that proves the theorem. □

Privacy against collision of analyst with users

Theorem 4. *If user(s) and analyst collude in the collaborative data analysis, the privacy of X_i is preserved regarding ε-DR privacy definitions* (5) *and* (6) *of each f_i.*

Proof. If user(s) and analyst collude, then they can obtain both the input X^{anc} and output $\widetilde{X}_i^{\mathrm{anc}}$ of f_i. In this case, they can infer f_i satisfying

$$\widetilde{X}_i^{\mathrm{anc}} = f_i(X^{\mathrm{anc}}).$$

Therefore, using the inferred f_i, they can infer X_i from $\widetilde{X}_i$. However, since f_i is a dimensionality reduced function, that is $m > \widetilde{m}_i$, the privacy is still preserved regarding ε-DR privacy definitions (5) and (6) of f_i. In other words, the exact data X_i cannot be recovered from $\widetilde{X}_i$, even using f_i. □

Privacy against the external attacks

Using secure data transmission protocols such as Transport Layer Security (TLS), in which the transferred information is encrypted using the private key of the involving parties, the collaborative data analysis also protects the private dataset X_i against the man-at-the-middle attackers. Note that, in this case, we do not use secure multi-party computations, but just use encrypted communication for non private data.

In the case that we do not use secure data transmission protocols, the situation is almost the same as the case that users and analyst collude. That is, man-at-the-middle attackers can infer f_i from X^{anc} and $\widetilde{X}_i^{\mathrm{anc}}$ and can infer X_i; however, the privacy is still preserved regarding ε-DR privacy definitions (5) and (6) of f_i.

Numerical evaluation for privacy analysis

Here, we provide numerical evaluation of the worst-case privacy analysis, i.e., the situation of Theorem 4. Let $B_i \in \mathbb{R}^{m \times m_i}$ be a matrix for dimensionality reduction for X_i as

$$\widetilde{X}_i = X_i B_i, \quad B_i \in \mathbb{R}^{m \times m_i}.$$

Table 1: Trade-off relationship between ε-DR privacy and prediction accuracy.

down-sampling parameter ε	min ε-DR (7)	Ave. # of samples	Ave. ACC
0.0	7.36×10^{-6}	100.00	92.8
0.0001	2.07×10^{-4}	99.97	92.8
0.001	1.01×10^{-3}	99.75	92.8
0.01	1.00×10^{-2}	97.50	92.8
0.1	1.00×10^{-1}	76.85	91.7
0.2	2.00×10^{-1}	56.43	90.3
0.3	3.00×10^{-1}	39.35	88.7
0.4	4.00×10^{-1}	25.84	86.1
0.5	5.00×10^{-1}	15.98	81.4
Centralized analysis		1000.00	93.6
Individual analysis		100.00	75.5

Let $X_i = [\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_{n_i}^{(i)}]^{\mathrm{T}}$. Then, if B_i and the center $\boldsymbol{\mu}_i \in \mathbb{R}^m$ of dataset X_i are stolen, X_i is inferred by

$$X_i' = [\boldsymbol{x}_1^{(i)'}, \boldsymbol{x}_2^{(i)'}, \ldots, \boldsymbol{x}_{n_i}^{(i)'}]^{\mathrm{T}} = \widetilde{X}_i B_i^{\dagger} + \mathbf{1}\boldsymbol{\mu}_i^{\mathrm{T}}(I - B_i B_i^{\dagger}),$$

where $B^{\dagger}$ is the pseudo-inverse of B and $\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}}$. As a ε-DR privacy regarding a sample set (6), we set

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathrm{dist}(\boldsymbol{x}, \boldsymbol{x}') = \min_{i,j} \frac{\|\boldsymbol{x}_j^{(i)} - \boldsymbol{x}_j^{(i)'}\|_2}{\|\boldsymbol{x}_j^{(i)}\|_2}. \tag{7}$$

Then, we use a down-sampling technique which removes training data samples satisfying (7) $< \varepsilon$ by changing ε and evaluate a trade-off relationship between (7) and a prediction accuracy of the collaborative data analysis.

We used MNIST again. The dimensionality reduction matrix B_i is constructed by PCA using each X_i. We set $c = 10, n_i = 100$ and $m_i = 25$ for parameters. Other settings of numerical evaluation are the same as used in the numerical evaluation for accuracy analysis.

Table 1 shows the trade-off relationship between a minimum ε-DR privacy (7), average number of samples after the down-sampling technique in each party, and average of prediction accuracy (ACC) of 10 trials. We also show averages of ACC for the centralized and individual analyses.

This result shows that, by the down-sampling technique, we can increase the value of ε-DR privacy (7) without loss of ACC; see the case of $\varepsilon = 10^{-2}$. Also, if the predictive accuracy is allowed to decrease slightly, it can take a larger value of ε-DR privacy (7); see the case of $\varepsilon = 0.2$. These results mean that a small number of samples significantly reduce the values of ε-DR privacy (7), while these samples do not significantly affect the predictive results. Additionally, even with a larger ε, e.g., $\varepsilon = 0.5$, the predictive accuracy (ACC) of the collaborative data analysis is still higher than that of the individual analysis.

Remarks on privacy analysis

The collaborative data analysis has the following double privacy layer for protection of the private data X_i.

- No one can have the private data X_i because f_i is private under the protocol (Theorems 2 and 3).
- Even if f_i is stolen, the private data X_i is still protected regarding ε-DR privacy definitions (5) and (6) (Theorem 4).

For satisfying ε-DR privacy definitions (5) and (6) with certain quantities $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, we need to pay attention to the construction of f_i. Dimensionality reduction method satisfying ε-DR privacy (5) with a given $\varepsilon_1 > 0$ has been proposed in (Nguyen et al. 2019). For satisfying ε-DR privacy (6) with a given $\varepsilon_2 > 0$, we can use a down-sampling technique, which removes data samples satisfying $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{x}') < \varepsilon_2$, or a constrained dimensionality reduction method, which adds (6) as a constrain in optimization.

We also observed from our numerical evaluation that the down-sampling technique can take a larger value of ε-DR privacy (7) with small decreasing of ACC.

Here, we also shortly discuss the privacy against the membership inference attack. The membership inference attacks involve constructing multiple reference datasets and observing the change in the output of the constructed model according to the presence or absence of the target data samples. Therefore, they are only feasible in scenarios when individual dimensionality reduction functions f_i are leaked in the collaborative data analysis. To secure data collaboration from the collusion of users in applications where membership inference is a concern, specialized dimensionality reduction algorithms providing Differential Privacy, such as in (Jiang et al. 2013), can be also considered.

Conclusions

In this paper, we analyzed the accuracy and privacy of a non-model sharing-type federated learning, so-called collaborative data analysis.

From the accuracy analysis, we provided the sufficient condition (4) for equivalence of the collaborative data analysis and the centralized analysis with dimensionality reduction. We also provided a criteria τ_1 for evaluating accuracy of the collaborative data analysis and numerically evaluated them.

From the privacy analysis, we proved that, in the collaborative data analysis, the privacy of the private dataset is preserved based on a double secureness against insider and external attacking scenarios. We also evaluated the trade-off relationship of privacy and accuracy and showed that the down-sampling technique can take a larger value of ε-DR privacy (7) with small decreasing of prediction accuracy.

In the future, we will further analyze the accuracy and privacy of the collaborative data analysis for more complicated situations, e.g, usage of nonlinear dimensionality reduction function and the case of vertical and horizontal data distribution, with numerical evaluation in real-world problems.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported in part by the New Energy and Industrial Technology Development Organization (NEDO).

References

Agrawal, R.; and Srikant, R. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 439–450.

Alotaibi, K.; Rayward-Smith, V. J.; Wang, W.; and de la Iglesia, B. 2012. Non-linear dimensionality reduction for privacy-preserving data classification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 694–701. IEEE.

Dwork, C. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, 1–19. Springer.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of human genetics* 7(2): 179–188.

He, X.; and Niyogi, P. 2004. Locality preserving projections. In *Advances in neural information processing systems*, 153–160.

Imakura, A.; Matsuda, M.; Ye, X.; and Sakurai, T. 2019. Complex Moment-Based Supervised Eigenmap for Dimensionality Reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3910–3918.

Imakura, A.; and Sakurai, T. 2020. Data Collaboration Analysis Framework Using Centralization of Individual Intermediate Representations for Distributed Data Sets. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 6: 04020018.

Imakura, A.; Ye, X.; and Sakurai, T. 2020. Collaborative Data Analysis: Non-Model Sharing-Type Machine Learning for Distributed Data. In *2020 Principle and Practice of Data and Knowledge Acquisition Workshop (PKAW2020), (accepted)*.

Jiang, X.; Ji, Z.; Wang, S.; Mohammed, N.; Cheng, S.; and Ohno-Machado, L. 2013. Differential-private data publishing through component analysis. *Transactions on data privacy* 6(1): 19.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* .

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* .

Li, Q.; Wen, Z.; and He, B. 2019. Federated learning systems: Vision, hype and reality for data privacy and protection. *arXiv preprint arXiv:1907.09693* .

Li, X.; Chen, M.; Nie, F.; and Wang, Q. 2017. Locality Adaptive Discriminant Analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2201–2207. AAAI Press.

Liu, K.; Kargupta, H.; and Ryan, J. 2005. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering* 18(1): 92–106.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9: 2579–2605.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282.

Mendes, R.; and Vilela, J. P. 2017. Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access* 5: 10562–10582.

Narayanan, A.; and Shmatikov, V. 2006. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105* .

Nguyen, H.; Zhuang, D.; Wu, P.-Y.; and Chang, M. 2019. AutoGAN-based Dimension Reduction for Privacy Preservation. *arXiv Preprint* arXiv:1902.10799.

Nguyen, H.; Zhuang, D.; Wu, P.-Y.; and Chang, M. 2020. AutoGAN-based dimension reduction for privacy preservation. *Neurocomputing* 384: 94–103.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11): 559–572.

Samarati, P.; and Sweeney, L. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression .

Saunders, C.; Gammerman, A.; and Vovk, V. 1998. Ridge regression learning algorithm in dual variables .

Sugiyama, M. 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of machine learning research* 8(May): 1027–1061.

Tai, B.-C.; Li, S.-C.; Huang, Y.; Suri, N.; and Wang, P.-C. 2018. Exploring the Relationship Between Dimensionality Reduction and Private Data Release. In *2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC)*, 25–33. IEEE.

Verykios, V. S.; Bertino, E.; Fovino, I. N.; Provenza, L. P.; Saygin, Y.; and Theodoridis, Y. 2004. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record* 33(1): 50–57.

Zelnik-Manor, L.; and Perona, P. 2005. Self-tuning spectral clustering. In *Advances in neural information processing systems*, 1601–1608.

Zhao, J.; Chen, Y.; and Zhang, W. 2019. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access* 7: 48901–48911.