Computers & Security 122 (2022) 102907

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Privacy, accuracy, and model fairness trade-offs in federated learning

Xiuting Gu [a], Zhu Tianqing [a,*], Jie Li [a], Tao Zhang [b], Wei Ren [a,c], Kim-Kwang Raymond Choo [d]

[a] School of Computer Science, China University of Geosciences, Wuhan, PR China
[b] School of Computer Science, University of Technology Sydney, Sydney, Australia
[c] Guangxi Key Laboratory of Cryptography and Information Security, Guilin 541004, PR China
[d] Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

ARTICLE INFO

Article history:
Received 21 March 2022
Revised 12 August 2022
Accepted 31 August 2022
Available online 5 September 2022

Keywords:
Federated learning
Differential privacy
Discrimination
Fairness
Privacy preservation
Machine learning

ABSTRACT

As applications of machine learning become increasingly widespread, the need to ensure model accuracy and fairness while protecting the privacy of user data becomes more pronounced. On this note, this paper introduces a federated learning training model, which allow clients to simultaneously learn their model and update the associated parameters on a centralized server. In our approach, we seek to achieve an acceptable trade-off between privacy, accuracy, and model fairness by using differential privacy (DP), which also helps to minimize privacy risks by protecting the presence of a specific sample in the training data. Machine learning models can, however, exhibit unintended behaviors, such as unfairness, which result in groups with certain sensitive characteristics (e.g., gender) receiving different patterns of outcomes. Hence, we discuss the fairness and privacy effect of local DP and global DP when applied to federated learning by designing a fair and privacy quantification mechanism. In doing so, we can achieve an acceptable trade-off between accuracy, privacy, and model fairness. We quantify the level of fairness based on the constraints of three definitions of fairness, including demographic parity, equal odds, and equality of opportunity. Finally, findings from our extensive experiments conducted on three real-world datasets with class imbalance demonstrate the positive effect of local and global DP on fairness. Our study also shows that privacy can come at the cost of fairness, as stricter privacy can intensify discrimination. Hence, we posit that careful parameter selection can potentially help achieve a more effective trade-off between utility, bias, and privacy.

1. Introduction

Federated learning (FL) demonstrated in Brendan McMahan et al., 2016 implements a distributed model that selects independent clients with decentralized datasets as participants for training the client model, such that only the parameters of the client models are ended as response and aggregated by a centralized server. In the next round, the server replies the server model to the participants to improve the quality of the client models. FL has broad applications, including in on applications involving Internet of Things (IoT) systems. However, optimizing only the average loss in a cross-device distributed model may have a negative impact on model performance, accuracy, fairness, and/or privacy. For example, studies such as that conducted by Zhang et al. (2020b) have revealed potential bias in the prediction process; for example, dark-skinned defendants have a higher probability of being deemed guilty than defendants with light skin. From the privacy perspective, moreover, information regarding specific examples can be leaked through machine learning training on personal data. For example, Cummings et al. (2019) find that a word prediction model can reveal social security numbers and credit card numbers, as they are learned using machine learning on a dataset that contains that information. It is further important to note that, recently, new laws and regulations to protect the privacy of personal data are in place and legislated. Desiato (2018), the General Data Protection Regulation (GDPR) introduced by the European Community contains measures designed to prevent invasion of privacy. In the federated learning setting, data privacy exposure risks as introduced in this study cannot be avoided. These findings reinforce the importance of ensuring model accuracy; however, this conflicts with the need to protect the privacy of clients, along with the need to reduce discrimination against certain sensitive groups (genders, races, etc.) in FL applications. For example, Bagdasaryan et al., 2019 suggested that as differentially private technology becomes more widely used, discrimination between groups with different sizes of examples will occur, unintentionally resulting in unfair-

* Corresponding author.
E-mail addresses: tianqing.zhu@IEEE.org (Z. Tianqing), raymond.choo@fulbrightmail.org (K.-K.R. Choo).

https://doi.org/10.1016/j.cose.2022.102907
0167-4048/

ness. The trade-off of privacy, accuracy, and fairness must therefore be carefully considered.

Maximizing the utility of the model under differential privacy settings also has been explored. There are two primary settings: local model (DP applied in each client) and trusted curator model (DP applied in the server). However, the influence of differential privacy on group fairness has not been considered. If we apply differential privacy, will the error of sensitive groups be affected by the equilibrium? Is it feasible to improve the group fairness by applying differential privacy in the federated learning context? Can an acceptable compromise between accuracy, privacy, and fairness be obtained? These questions are important and must be answered, as interest in applying FL has greatly increased, especially for multiple organizations who wish to collaborate on training a model while keeping data decentralized.

In this work, we look beyond the trade-off between group fairness, privacy, and accuracy. The contributions of our work include the following. First, we apply two settings of differential privacy in a federated learning setting by training FL neural networks separately, using local DP and global DP, to evaluate the influence on fairness. In the models, the influence of gradient clipping and noise added from the privacy mechanism is explored. The two methods are the main factors that cause fluctuations in results. We demonstrate the impacts by extensive experiments using three real-world datasets with class imbalance.

We demonstrate that the cost of model accuracy can be ignored if we choose the proper noise scale and gradient clipping norm. The present study quantitatively estimates the scope to achieve a better trade-off between accuracy and fairness. The trade-off between fairness and privacy can be achieved using the differential privacy stochastic gradient descent algorithm (DP-SGD) proposed by Abadi et al., 2016 with an acceptable cost of accuracy. The experiments show that DP improves group fairness, although there is a fairness cost associated with achieving stricter privacy.

The remainder of this article is organized as follows. Section 2 reviews the existing fairness protection methods and related literature on fairness and privacy collaborative learning in the FL context. The key technologies and the fairness quantitative mechanism that underpin the study are introduced in Section 3. Details of the method for applying differential privacy algorithm in federated learning are presented in Section 4. The discussion and a summary of the experiments performed to evaluate the performances of DP are provided in Section 5. Finally, Section conclusions concludes this study.

2. Related work

In this section, we introduce the relevant literature on fairness protection, privacy preservation, and fairness risk metrics.

2.1. Fairness in federated learning

Extensive literature presents available methods for use in training a centralized and fair machine learning model. The methods include constrained and distributional robust optimization, along with post-shifting, first proposed by Bilal Zafar et al. (2015).

Previous literature such as Eckhouse et al., 2019 proposed that an imbalance in training sets will bring about bias in models. In the neural network context, Bagdasaryan et al., 2019 demonstrate that DP-SGD produces an unfair model when the training data is unbalanced. Here, “unfair” denotes a difference in accuracy between small and large classes of DP models; the DP model will tend to discriminate against the underrepresented classes. Kairouz et al. (2019) and Zhang et al. (2021) discovered that imbalance in datasets used for training can introduce unfairness into the prediction models. In a federated learning setting, an imbalance in the size of examples used by different clients may produce bias. A participant client will be assigned a larger weight when it generates more data. In this case, bias is produced locally on the client training side.

The pre-processing method is utilized to decrease bias from unbalanced training data. Some previous works such as Chawla et al. (2002) proposed oversampling, while Madras et al. (2018) proposed adversarial training, and cost-sensitive learning. Madras et al. (2018) found that models that generate artificial data points resembling minority group data can also effectively reduce data imbalance. Cummings et al. (2019) train models with post-processing and in-processing method.

In terms of fairness criteria in federated learning, Gupta and Raskar (2018) and Li et al. (2019) calculate the fairness level using the variance in accuracy across clients. The way to achieve fairness is to reduce the variance without sacrificing average accuracy. Donini et al. (2018) present an algorithm that employs an empirical risk minimization method. The method injects a fairness constraint that meets a definition of fairness into the objective function; however, this type of method is considered NP-hard. In our work, we apply a de-identification technique in order to reduce bias, which has a low computational cost and is good at preventing the specific samples from being identified. As for fairness criteria, we use variants of three definitions of fairness as the metrics to quantitatively evaluate fairness. Zhang et al. (2020a) developed a mechanism to balance the distributions of training datasets with semi-supervised learning to improve accuracy and decrease discrimination. Xu et al. (2020) proposed the conditional fairness method, which adds a variable representing fairness to restrict fairness level. Vasudevan and Kenthapadi (2020) defined a framework to satisfy the need to evaluate the discrimination level of scale machine learning systems.

In federated learning, the server balances aggregation by averaging the model loss across the sampled devices. Zhu et al. (2022) showed that this may lead to differing model performance on different devices. In this case, bias is produced during the aggregation process.

2.2. Fairness and privacy in federated learning

Differential private learning tends to hide specific sample characteristics to achieve privacy protection. Fair definitions tend to achieve balanced predictions on sensitive attributes. However, this requires knowing information about individuals in sensitive groups. Our work is focused on the trade-off between privacy, accuracy, and model fairness in federated learning. Cummings et al. (2019) show that the trade-off between privacy and discrimination when differential privacy is employed has been surveyed in the non-federated model. The study shows that the work yields an efficient algorithm for classification that maintains utility while satisfying both privacy and approximate fairness requirements with high probability.

Jagielski et al. (2018) and Bagdasaryan et al., 2019 suggest that the model fairness may be destroyed by privacy protection, which encourages the model to perform well on under-represented classes. Kearns et al. (2017) focuses on subgroup fairness, selecting a group fairness constraint and asking whether most of the subgroups satisfy this constraint.

For privacy protection, Geyer et al., 2017a adopt a client-level perspective. Their work sets up a central server that aggregates models trained by independent clients using local differential privacy, which aims to hide the contributions made by clients during training. In our work, we add a global-level perspective by applying global differential privacy settings in federated learning. We also explore fairness protection based on privacy protection.

2.3. Fairness risk measures

As noted by Chouldechova et al. (2018), there is widespread concern regarding the need to reduce discrimination in models related to protected features. Williamson and Menon, 2019 proposed a new definition of fairness, which allows for generic sensitive features and generates a new convex objective using constraints on the fairness definition.

To design a fairness measure, it is necessary to develop definitions of "fair" to assess group fairness. As demonstrated by Dwork et al. (2012), demographic parity requires the distribution of predictions to be identical for all values of feature S. Hardt et al., 2016 found that equalized odds are necessary to satisfy demographic parity when using the true label Y if the accuracy is to be balanced. Zafar et al. (2017) found that in the absence of disparate (mis)treatment, subgroup errors tend to be the same. Dwork et al. (2012) and Kusner et al. (2017) further developed approaches as individual fairness, according to which the circumscription of perfect fairness is the only ideal condition of the classifier.

When using practical machine learning, it is often necessary to apply constraints based on approximate fairness or precise fairness. The work of Menon and Williamson (2018) aimed to balance fairness and accuracy loss. Calmon et al. (2017) explored the maximal discrimination of accuracy between different groups. The method of Agarwal et al. (2018) is a variant that addresses a related issue. Williamson and Menon, 2019 proposed an alternative in terms of their, which holds that specific random variables are independent by Kullback–Leibler (KL) divergence. Komiyama and Shimao (2017) replace the KL divergence with f-divergence, while Pérez-Suay et al. (2017) replace the KL divergence with the Hilbert–Schmidt criterion. These measures can use to deal with multi-class as well as multi-label and continuous S. Cummings et al. (2019) shows that exact fairness can not cooperate with differential privacy, then propose a measure of approximate fairness to adapt fairness and privacy restrictions using A-discrimination; this requires that the discrepancies of true positive rates between different groups are smaller than A.

3. Preliminaries

In this section, we introduce the key technologies that underpin the present study, including differential privacy (DP), DP in federated learning, and fairness metrics.

3.1. Differential privacy

Review the standard definitions presented in Dwork (2011). A randomized mechanism M : $D \leftarrow R$ with a domain D and range R satisfies (ϵ, δ)-differential privacy if for any two adjacent datasets d, d′∈ D, and for any subset of outputs S ⊆ R, Pr[M(d)∈ S]≤ e^{ϵ} Pr[M(d′) ∈ S]+δ.

ϵ represents how strict the privacy is, while δ is the broken probability of the DP. A low value of ϵ and δ will lead to a higher level of privacy. Every ϵ-DP computation depletes the budget by δ; once the budget has been expended no further learning is available.

3.2. Differential privacy in federated learning

Local differential privacy(DP) technology is applied to the training data of an independent client then uploaded and grouped by the server. For this reason, this study injects DP-SGD into model training with an independent client dataset and protects specific sample information of local data from being leaked. Local dp adds perturbation to the clients' training process, after which the parameters of the local models are uploaded to the central server. This study applies the "moments accountant" approach developed by Abadi et al., 2016 to quantify the privacy cost. showed that this method of calculation results in much tighter bounds being imposed.

For global differential privacy settings, differentially private algorithms are operated in a trusted server, which is considered credible. Clients train their own models independently, which are aggregated by the server, after which the server applies algorithms to ensure the data is private. In federated learning, participants jointly train a model, after which the server aggregate clients' updates by the FedAvg algorithm developed by Brendan McMahan et al., 2016. This study adds perturbation to server aggregation, applying the DP-FedAvg method proposed by McMahan et al., 2017b. The method tends to limit the influence of any participant by clipping and adding Gaussian noise $N(0, \sigma^2)$. The structure of differential privacy settings in federated learning is illustrated in Fig. 1.

Further details of the algorithms will provided in a later section.

3.3. Fairness metrics

In fair classification, we have some data $X \in R^n$, labels $y \in \{0, 1\}$, and sensitive attributes $A \in \{0, 1\}$. The predictor outputs a prediction $Y' \in \{0, 1\}$. The model seeks to learn to predict outcomes that are independent with respect to A but accurate concerning Y; that is, the predictor is fair for groups and accurate.

The literature quantitatively delimit fairness among groups and applies as fairness metrics to facilitate quantitative evaluation of a model's discrimination level.

Definition 1 ((Equalized odds) Proposed by Hardt et al., 2016)**.** If a predictor Y ′ meets the limit of equalized odds about protected attribute A and label Y, Y ′ and A are independent conditional on Y.

$$Pr\{Y' = 1 \mid A = 0, Y = y\} - Pr\{Y' = 1 \mid A = 1, Y = y\} = 0, y \in \{0, 1\} \quad (1)$$

Equalized odds (EOD) requires equal false positive and false negative rates between groups.

Definition 2 ((Equal opportunity) Proposed by Hardt et al., 2016)**.** If a binary predictor Y′ meets the limit of equal opportunity between A (sensitive attribute) and Y (label), we have the following:

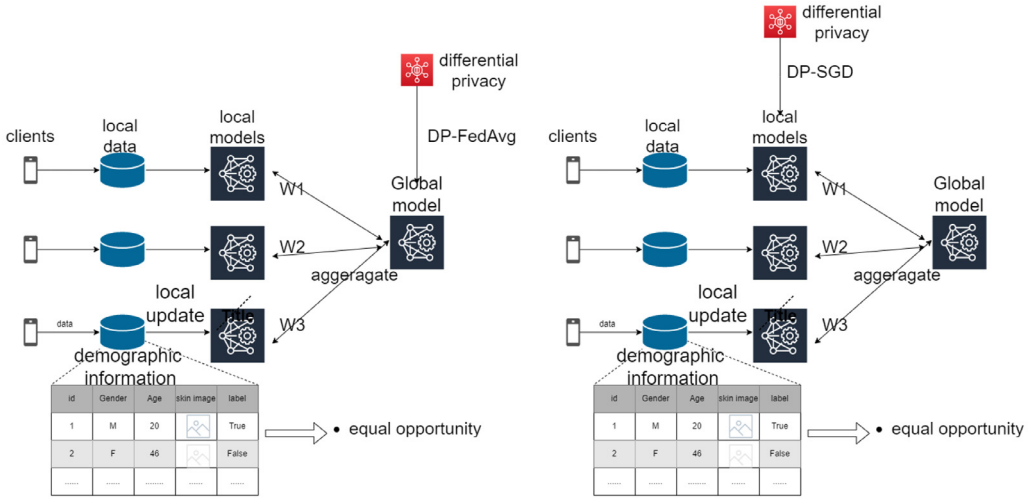
$$\Pr\{Y' = 1 \mid A = 0, Y = 1\} - Pr\{Y' = 1 \mid A = 1, Y = 1\} = 0 \quad (2)$$

The goal of Equal Opportunity (EOP) is to obtain a prediction that is equal to the true label Y based on supervised training data while limiting the accuracy of the model on the prediction sample to ensure fairness with respect to a sensitive attribute A. In supervised learning, it is considered that the training has access to the labels of the training data.

Definition 3 (Demographic parity)**.** The definition assumes that the same positive prediction rate of two groups concerning sensitive attribute A. It requires that the decision be unrelated to the protected attribute.

$$Pr\{Y' = 1 \mid A = 0\} - Pr\{Y' = 1 \mid A = 1\} = 0 \quad (3)$$

Demographic parity (DOP) ensures that positive outcomes are given to the two groups at the same rate. However, the usefulness



(a) global-dp setting (b) local-dp setting

Fig. 1. The structure of differential privacy (DP) settings in a federated learning context.

of demographic parity can be limited if the base rates of the two groups differ.

Definition 4 (Discrimination level)**.** The quantitative representation of the discrimination level between groups.

$$\alpha = \Pr\{Y' = 1 \mid A = 0, Y = 1\} - Pr\{Y' = 1 \mid A = 1, Y = 1\} \tag{4}$$

$$\beta = Pr\{Y' = 1 \mid A = 0, Y = y\} - Pr\{Y' = 1 \mid A = 1, Y = y\}, y \in \{0, 1\} \tag{5}$$

$$\theta = Pr\{Y' = 1 \mid A = 0\} - Pr\{Y' = 1 \mid A = 1\} \tag{6}$$

α is the output of EOD when the predictor does not meet the limit of EOD. α can be used to quantitatively evaluate the bias between groups. β is the output of EOP when the predictor does not meet the limit of EOP. β can be used to quantitatively evaluate the bias between groups. θ is the output of DOP when the predictor is not meet the limit of Demographic parity(DOP). θ can be used quantitatively evaluate the bias between groups.

The study applies the A-discrimination proposed by Cummings et al. (2019) to the quantitative evaluation fairness metric. A is the output parameter used to evaluate the level of fairness; a smaller value indicates that the model is fairer to groups. The study gets three results (α, β, θ) that satisfy three fairness definition constraints to express the fairness level. When (α, β, θ) are smaller, the level of fairness is better. For example, when the positive prediction rates of two groups concerning sensitive attributes are different, the limit of demographic parity is not reached. According to the definition of discrimination level-θ, the difference in the positive prediction rate of the two groups will be calculated as θ.

4. Method

There are two primary settings of DP in the federated learning context, namely: the local DP (LDP) and global DP (GDP) mechanisms.

4.1. Local DP in federated learning

Implement LDP Local differential privacy (LDP) applies DP-SGD to clients, which involves adding a differential privacy mechanism to stochastic gradient descent (SGD), then conducting a comparison with the baseline (non-differential privacy model) to see whether adding local differential privacy brings about an increase in fairness.

Local differential privacy training requires adjusting the hyperparameters. LR denotes the local learning rate of training on client models. The clipping bound is used for updating the gradients of every training round on the client. Noise is the Gaussian distribution parameter. The clipping bound S and noise are the parameters of DP before gradient update, making them the most important factors.

4.2. Global DP in federated learning

Implementing GDP To implement global differential privacy we use the DP-FedAvg algorithm on the server. For each round t, each participant i$\in d_C$ locally trains a model on their private data, producing a new model. The global server then aggregates these models and updates the global model as $G_{t+1} = G_t + \frac{\eta_g}{n} \sum_{i \in d_C} (L^i_{t+1} - G_t)$ using the global learning rate η_g. DP-FedAvg clips the norms to S for each update vector $\pi_s(L^i_{t+1} - G_t)$ and adds Gaussian noise N(0, σ^2) to the sum: $G_{t+1} = G_t + \frac{\eta_g}{n} \sum_{i \in d_C} \pi_S(L^i_{t+1} - G_t) + N(0, \sigma^2 I)$.

Implementing LDP + GDP For the model Both-DP, this study applies both local and global differential privacy mechanisms with two parameters, namely local and global noise σ), creating a combination of two differential privacy models to improve the training.

4.3. Optimization perturbation method

The clipping bound is the main reason for discrimination using DP-SGD, as demonstrated by Zhang et al. (2021). A class containing a larger number of samples will typically have a greater impact than one with a smaller number, regardless of gradient clipping. The present study sets the clipping bound to be fixed for different groups in order to limit the effect of the gradient on bias.

The accuracy loss is limited by the clipping gradients and noise added to the gradient. For each group, this study sets the same clipping bound to prevent the average gradient from being too dissimilar to that of small-sized groups. In the case of uneven gradient distribution, this study adds optimization perturbation in the form of Gaussian noise to gradients with local differential privacy

to weaken the severity of the unevenness, thereby reducing the bias.

Training with imbalanced datasets is another main cause of discrimination in models produced by clients. High accuracy is often obtained when building models from unbalanced samples. In each round, model training with a more imbalanced data set gets more updates. The DP algorithm averages these updates to the gradient provided by clients to get the shared model. This reduces the fairness problems caused by data imbalance among clients with extremely imbalanced data. The present study adds output perturbation with global differential privacy. The purpose of DP-FedAvg is to avoid data imbalance, which has a significant impact on the global models obtained by each round of training, and promote a steady increase in fairness and accuracy.

To verify the relationship between privacy assurance and fairness, this study limits the broken probability of the privacy mechanism. To do so, it adjusts the privacy level by changing the scale of noise to weaken the effect of gradient clipping for groups, thereby bringing about acceptable privacy protection and reduced discrimination.

5. Experimental evaluation

The experiments with the benchmark LEAF conducted by Caldas et al., 2018 apply differential privacy federated learning models. The discrimination level is used to get a digital representation of unfairness between groups with respect to the sensitive attributes. When the discrimination level is lower, the model is fairer.

This work applies the DP-SGD algorithm developed by Abadi et al., 2016 on clients as local DP, or DP-FedAvg on a server as global DP, to achieve fairness protection and avoid discrimination, while simultaneously protecting fairness with lower privacy cost and a smaller accuracy decrease.

The study restricts privacy loss using the moments accountant and fixs δ as 10^{-5} to get ϵ as the privacy loss of different noise levels. Next, the noise scale is changed to implement fairness protection and fix privacy loss in order to obtain a trade-off between privacy loss, fairness and accuracy.

5.1. Classification of the adult dataset

5.1.1. Dataset

The classification experiment is based on the Adult dataset.[1] It includes over 40,000 rows of data and is obtained from the 1994 US Census. Our task is to predict each person's income. The output is the result of a comparison with the figure of 50 K/year (i.e., greater than or less than). The attribute "gender" is set to be the sensitive characteristics.

5.1.2. Model

Our experiment implements a linear regression model using Mini-batch Gradient Descent on every client with the learning rate = 0.01 and batch size = 10.

5.1.3. Results

Fig. 2 (a) shows that after carefully adding random noise, the fairness metrics (α, β, θ) are reduced in the local-DP model compared to the baseline model, which means that the discrimination between different groups is decreased. For local DP, when the noise equals 0.01 or 0.05, the three fairness metrics (α, β, θ) are both lower than non-DP. Fig. 2(b) shows that as the global-DP noise (σ) is larger for stricter privacy, the fairness metrics (α, β, θ) are bigger, thus discrimination is increased. For global DP, when noise is between 0.05 to 1.5, the three fairness metrics (α, β, θ) lower than non-DP. When adding noise on client and server to get both-DP model, the discrimination level falls instead.

5.1.4. Trade-off between privacy, fairness, and accuracy

From Fig. 3(a), when the noise at some points (e.g., noise between 0.81 and 3.0) the fairness result can get a smaller value than non-DP (noise is set as 0), which indicates that training with differential privacy decreases discrimination between groups. With regard to accuracy, from Fig. 3(b), the maximum reduction after adding local DP is 0.07, a superior result compared with non-DP. For privacy, the noise bigger epsilon smaller. The smallest epsilon is 0.18 when noise is 3.0. The experiment achieves 77% test set accuracy for (0.18, 10^{-5})-differential privacy with less damage to accuracy and fairness. On this dataset, the proposed method achieves an acceptable trade-off between fairness, privacy, and accuracy.

5.2. Classification of bank data

5.2.1. Dataset

The UCI Bank Marketing dataset[2] is produced from the phone marketing records of a Portuguese banking institution. The output indicates whether the client agrees to make a deposit. There are over 45,000 rows of clients' information. The sensitive attribute is "personal loan".

5.2.2. Model

Our experiment implements a Neural Networks model with two hidden layers for every client, and is trained based on the parameters: learning rate = 0.0001, batch size = 10.

5.2.3. Results

From Fig. 4, for local DP, when the noise = 0.001, DEP and EOD fairness metrics (β, θ) are lower than for non-DP. For global DP, when noise = 0.01, the DEP fairness metric (θ) is lower than for non-DP. As the global-DP noise (σ) is made larger for stricter privacy, the values of the fairness metrics (α, β, θ) are bigger, indicating higher levels of discrimination. For the both-DP model, the discrimination level falls instead.

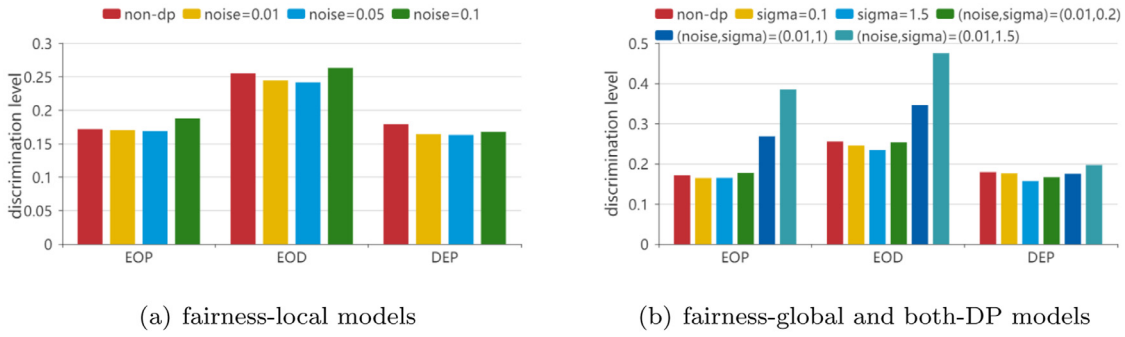
5.2.4. The trade-off between privacy, accuracy, and fairness

From Fig. 5, with regard to fairness, it can be seen that when the noise is at a certain level (e.g., noise between 0.1 and 0.35), the value of the discrimination level tends to be smaller than for non-DP (noise is set as 0). This indicates that training with differential privacy decreases discrimination between groups. From Fig. 5(b), a larger noise value is associated with a smaller epsilon value; when the noise value exceeds 0.45, the value of epsilon is acceptable (11). The noise must be a smaller value than 0.36, otherwise there will be a certain degree of damage to fairness. Thus, for this dataset, the proposed approach cannot achieve a suitable trade-off between fairness and privacy. There is a group fairness cost for achieving privacy.

With regard to accuracy, all values of noise keep accuracy above 90%, and the maximum reduction after adding local DP is 0.003, a superior result compared with non-DP with bias aggravated simultaneously (which can be ignored). The experiment achieves 90% test set accuracy for (11, 10^{-5})-differential privacy with damage to fairness.

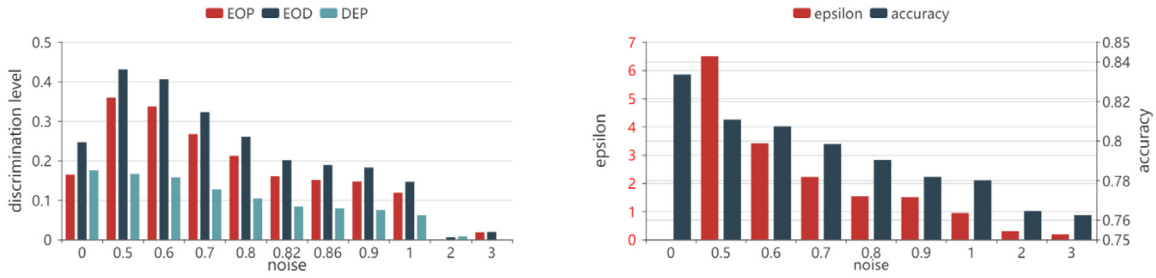
[1] https://archive.ics.uci.edu/ml/datasets/adult.

[2] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.



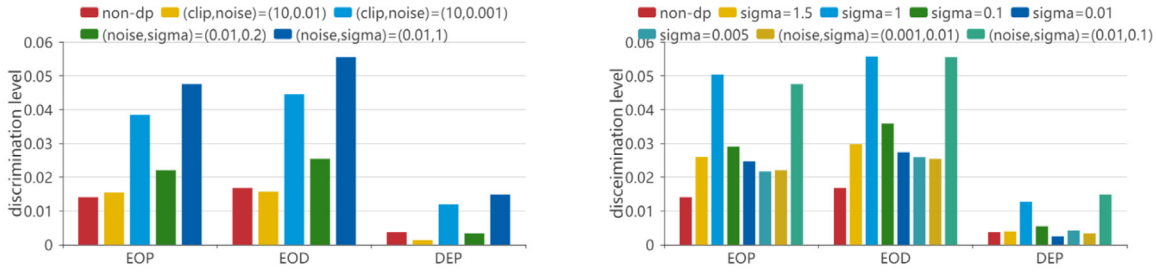
(a) fairness-local models (b) fairness-global and both-DP models

Fig. 2. Comparison of discrimination level and accuracy on the Adult dataset.



(a) fairness-noise (b) the trade-off between privacy cost and accuracy

Fig. 3. Results for the Adult dataset using different levels of noise. The maximum reduction of accuracy after adding local DP is 0.07, a superior result compared with non-DP.



(a) fairness-local models (b) fairness-global and both-DP models

Fig. 4. Comparison of fairness metrics and accuracy results on the bank dataset.

5.3. Classification of clients' default

5.3.1. Dataset

The UCI client credit card default dataset[3] records instances of customers' defaulting on payments in Taiwan and contrasts the predictive accuracy of the probability of default with six data mining methods. The output is whether or not the real probability of default has been accurately estimated and predicted. There are over 30,000 rows of data. The study sets "gender" as the sensitive characteristic.

5.3.2. Model

Our experiment implements a Neural Networks model with two hidden layers for every client, and is trained based on the following parameters: learning rate = 0.001, and batch size = 10.

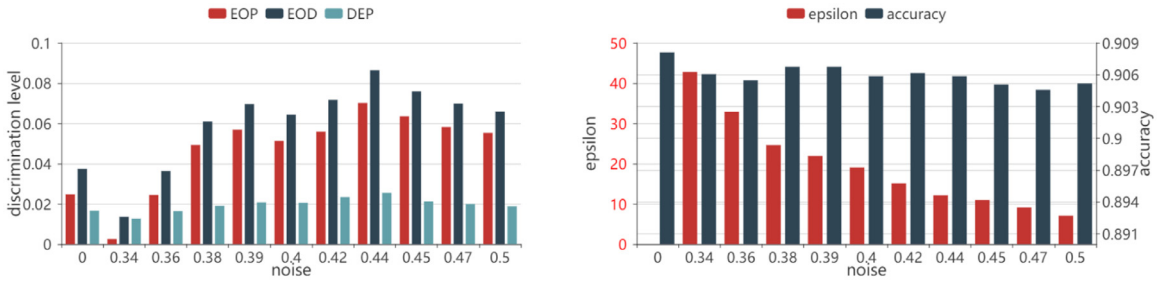
[3] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

5.3.3. Results

From Fig. 6, for local DP, when the noise is set as 0.001, the three fairness metrics (α, β, θ) are both lower than for non-DP. After carefully adding random noise, the local-DP model and global DP model can reduce the discrimination level compared with the non-DP model. For global DP, when the noise is between 0.01 to 0.05, the values of the three fairness metrics are lower than for non-DP. When the level of global-DP noise is higher, the level of privacy protection is also higher, but the overall fairness is lower. For the both-DP model, the discrimination is reduced.

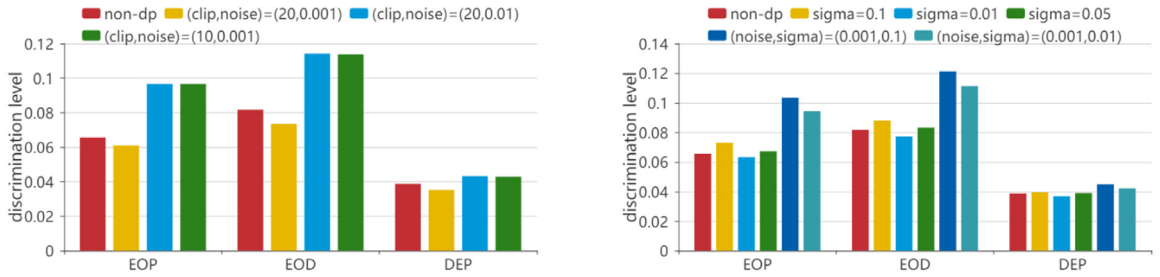
5.3.4. Trade-off between privacy, accuracy, and fairness

Each plot of Fig. 7 shows the result of varying the levels of noise for the corresponding ϵ, keeping δ fixed. To find a suitable level of noise to make epsilon smaller than 10, the value for fairness cannot simultaneously be smaller than for non-DP (noise is set as 0). As the strength of privacy protection increases, the degree of fairness decreases, which means that achieving a certain level of privacy protection undermines group fairness; in other words, there is a group fairness cost for achieving privacy.



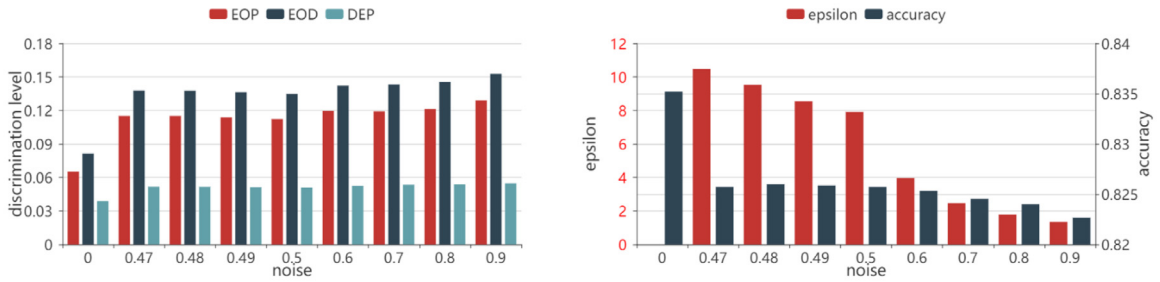
(a) fairness-noise (b) the trade-off between accuracy and privacy cost

Fig. 5. Results for different noise levels on the bank dataset. The maximum reduction of accuracy after adding local DP is 0.003, a superior result compared with non-DP.



(a) fairness-local models (b) fairness-global and both-DP models

Fig. 6. Comparison of fairness metrics and accuracy on the default dataset.



(a) fairness-noise (b) the trade-off between accuracy and privacy cost

Fig. 7. Results for different noise levels on the default dataset. The maximum reduction of accuracy after adding local DP is 0.01 lower than non-DP.

5.4. Impact of parameters

The goal is to find the parameters to get a suitable performance for local-DP and global-DP in certain cases. The present study evaluates the effects of hyperparameters using a real-world dataset based on a non-DP federated learning model. Experiments show that with a fixed clipping bound, noise can decrease discrimination; when noise is not at a proper level, however, the opposite results. Noise influences the direction of gradient migration. The impact is more significant in cases when the noise greatly alters the gradient.

5.4.1. Impact of training epochs

Fig. 8 shows that learning for a greater number of epochs may improve accuracy, but that the impact is reduced as the number of epochs increases. Fig. 8(a) shows that the fairness results can be better with a limited number of epochs. Moreover, Fig. 8(b) shows that more training epochs minimally impact the accuracy when the model is stable.

5.4.2. Impact of noise scale

In terms of noise behavior, the present study applies DP-SGD when training a client model. The algorithm adds Gaussian noise with clipping bound of 10. Fig. 9 shows, that increasing noise within a specific range on the gradient may produce limited higher fairness, but that this is not decisive.

5.5. Discussion and summary

5.5.1. Discussion

First, the study discusses how the noise scale and gradient clipping impact fairness, privacy, and accuracy.

How does differential privacy affect fairness? The accuracy loss is limited by clipping gradients and the addition of Gaussian noise to the gradient. This study sets the same clipping bound for each group to ensure that the average gradient is not too similar from that of small-sized groups. In the case of uneven gradient distribution, with local DP, the study adds optimized Gaussian noise perturbation to the gradient to weaken the severity of the unevenness,

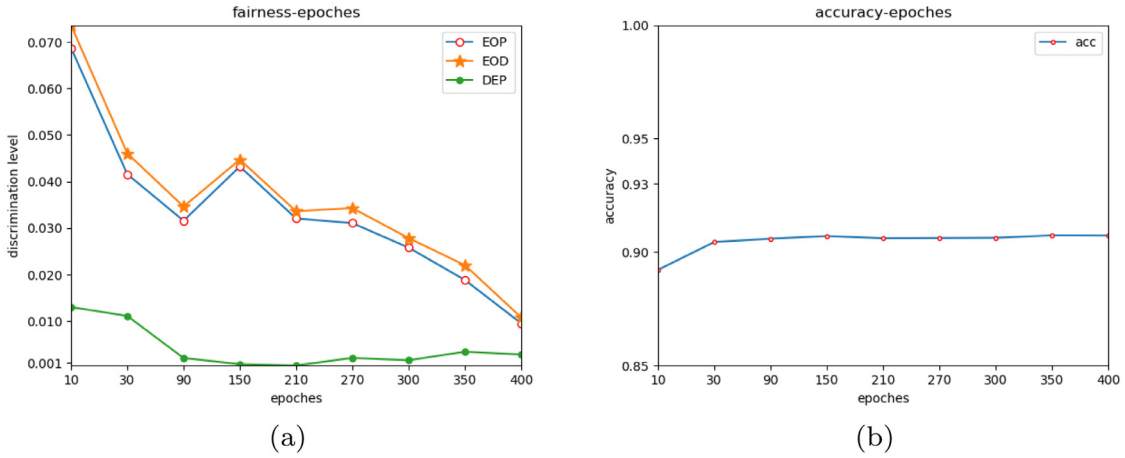


Fig. 8. Results for effect of epochs.

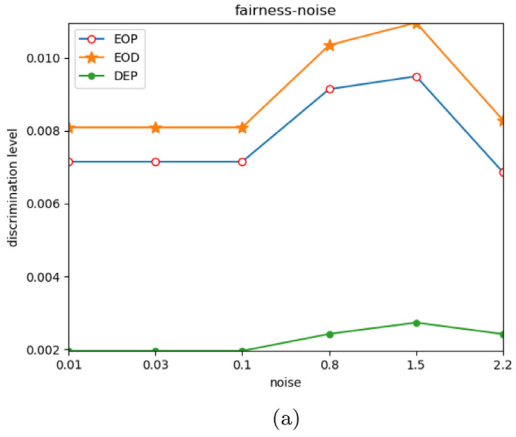


Fig. 9. Results of effect of noise.

Algorithm 1: Differential privacy SGD (DP-SGD).

Input: Dataset (x_1, y_1),...,(x_n, y_n) of size N, bitch size b, learning rate η, sampling probability q, loss function L(θ(x),y), K iterations, noise σ, clipping bound S, π_S(x)=x $*$ min(1, $\frac{S}{\|x\|_2}$)

Initialize: Model θ_0 ;
for $k \in [K]$ **do**
  randomly sample batch from dataset N with probability q ;
  foreach (x_i, y_i) *in batch* **do**
    $g_i \leftarrow \nabla$ L($\theta_k(x_i), y_i$)
  end
  $g_{batch} = \frac{1}{qN}(\sum_{i \in batch} \pi_S(g_i) + N(0, \sigma^2 I))$;
  $\theta_{k+1} \leftarrow \theta_t - \eta g_{batch}$
end
Output: Model θ_k and accumulated privacy cost (ϵ, δ)

thereby reducing the bias. On the other hand, training with imbalanced data set is another main cause of discrimination from client models. The global DP adds noise on an average gradient over random batches of clients to reduce bias from uneven client data.

For privacy, with a fixed value of δ, ϵ can get an acceptable result when choosing from a range of noise values. To protect privacy, DP-SGD clips the norm of the gradient to limit the sensitivity of each example and add noise to the gradient. This algorithm calculates explicit privacy bounds by using the moments accountant with the exact form of the noise (Gaussian). The range of noise can affect fairness and accuracy. The question of how to determine the range of noise to get an ignorable accuracy cost and simultaneously improve fairness is a difficult one, as it that is interrelated with the amount and composition of the dataset.

At the same time, the total noise when using local DP is bigger than for global DP. Thus, local DP may harm the model's accuracy. As Abadi et al., 2016 show, with δ fixed, altering ϵ may have a more substantial impact on accuracy. Our experiments show that training for more epochs can bring about improvements in accuracy and fairness. As the noise increases, the privacy cost is proportionally smaller. This may be a reason for selecting a larger number of epochs, as long as this does not exceed the cumulative privacy budget.

How might favorable balance between fairness and privacy be struck? The present study finds that there can exist variations in levels of fairness and privacy protection. This presents an opportunity when trying to build an optimization function for these two targets, namely that of setting the privacy restrictions. The study accordingly applies a naive approach to test out a minimally acceptable threshold for privacy and then determine the acceptable fairness level.

5.5.2. Summary

The present study focuses on a quantitative approach to assessing the effect of differential privacy on group discrimination. Experiments are conducted on three original unbalanced real-world datasets. Local differential privacy (LDP) is widely used in privacy protection. Accordingly, the present study explores the effect of LDP on model fairness, privacy, and accuracy and tries to achieve a trade-off. The study further explores the effect of different hyperparameters in its experiments to create a non-DP federated learning model on the bank dataset. For group fairness metrics, the experiment set the approximate function to satisfy three definitions of group fairness as the result of discrimination.

To obtain a trade-off between fairness and accuracy, in the local DP and global DP model, a proper value of noise brings the accuracy close to saturation, while discrimination between groups also drops significantly. DP can reduce group discrimination at the expense of model average accuracy. It is believed that group fairness produces an average accuracy loss in binary classification.

For privacy protection, the experiment sets the learning rate, gradient norm bound, batch size, and δ as fixed to get ϵ as the privacy loss when varying amounts of noise are added at the gradient. Our paper shows that varying the value of noise added at the gra-

dient has a large impact on fairness and privacy. When the noise is at a certain level, the privacy loss can be limited to between 1 and 10, which was the choice made in the work of Abadi et al., 2016. At the same time, fairness can be better than non-DP with a limited decrease in accuracy. The study finds that varying the noise added on gradient has a large impact on privacy loss.

The present study demonstrates the positive effects of local differential privacy and global differential privacy on fairness. As privacy protection is made stricter, group fairness will decrease. A trade-off between fairness, accuracy, and privacy is therefore achievable.

6. Conclusions

The study demonstrates that differential privacy can reduce discrimination in FL approaches. In LDP and GDP, under the condition of proper noise levels and a fixing clipping bound, fairness can be achieved. When the privacy level increases to some extent, the fairness of the model might decrease. Adding local differential privacy can bring about an acceptable trade-off between fairness, accuracy, and privacy. Simultaneously, stricter privacy protection may reduce group fairness, as discrimination tends to increase under stricter privacy settings; in fact, sufficiently tight privacy guarantees can degrade utility to the extent that the model becomes essentially random. Moreover, groups with a large number of samples no longer have a dominant influence on the model gradient under the proposed approach, making it fairer.

Several open problems remain around the question of how to improve fairness and privacy in the federated learning context. More importantly, it is still unclear how DP's hyperparameters (L2 norm bound and noise variance) can be generically chosen as a function of the model size/architecture and types of devices. Since there is a fairness cost associated with differential privacy, we hope to explore methods other than differential privacy with the goal of protecting fairness and privacy. We hope the results outlined in this paper can motivate further research on fine-grained parameter selection to establish the most practical mechanism in order to achieve the trade-off of fairness and privacy.

CRediT author statement

Xiuting Gu led the manuscript drafting, experiment implementation, and idea presentation. **Tianqing Zhu** participated in the solution design, manuscrupt drafting, and student supervision. **Jie Li** contributed to the experiment design, implementation, and evaluation. **Tao Zhang** contributed to the solution design and experiment evaluation design. **Wei Ren** participated in the solution design, manuscript drafting and editing, and student supervision. **Kim-Kwang Raymond Choo** contributed to the solution design, manuscript drafting and presentation, and experiment evaluation design.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research was financially supported by the National Natural Science Foundation of China (No. 61972366). The research of K.-K. R. Choo was supported only by the Cloud Technology Endowed Professorship.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cose.2022.102907

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp. 308–318.
Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H., 2018. A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.
Bagdasaryan, E., Poursaeed, O., Shmatikov, V., 2019. Differential privacy has disparate impact on model accuracy. Advances in neural information process-ing systems 32.
Bilal Zafar, M., Valera, I., Gomez Rodriguez, M., Gummadi, K. P., 2015. Fairness constraints: mechanisms for fair classification. arXiv.
Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R., 2017. Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems, pp. 3992–4001.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chouldechova, A., Benavides-Prado, D., Fialko, O., Vaithianathan, R., 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Conference on Fairness, Accountability and Transparency, pp. 134–148.
Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J., 2019. On the compatibility of privacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 309–315.
Desiato, D., 2018. A methodology for GDPR compliant data processing. SEBD.
Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J.S., Pontil, M., 2018. Empirical risk minimization under fairness constraints. In: Advances in Neural Information Processing Systems, pp. 2791–2801.
Dwork, C., 2011. A firm foundation for private data analysis. Commun. ACM 54 (1), 86–95.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2012. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226.
Caldas, S., Duddu, S.M.K., Wu, P., Li, T., Kone˙cn'y, J., McMahan, H.B., Smith, V., Talwalkar, A., 2018. Leaf: A benchmark for federated settings. arXiv preprint arXiv:1812.01097.
Eckhouse, L., Lum, K., Conti-Cook, C., Ciccolini, J., 2019. Layers of bias: a unified approach for understanding problems with risk assessment. Crim. Justice Behav. 46, 185–209.
Gupta, O., Raskar, R., 2018. Distributed learning of deep neural network over multiple agents. J. Netw. Comput. Appl. 116, 1–8.
Geyer, R. C., Klein, T., Nabi, M., 2017a. Differentially private federated learning: a client level perspective. arXiv preprint arXiv:1712.07557 .
Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifimalvajerdi, S., Ullman, J., 2018. Differentially private fair learning. arXiv: Learning.
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: Learning.
Kearns, M., Neel, S., Roth, A., Wu, Z. S., 2017. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. arXiv preprint arXiv:1711.05144.
Komiyama, J., Shimao, H., 2017. Two-stage algorithm for fairness-aware machine learning. arXiv preprint arXiv:1710.04924.
Kusner, M., Russell, C., Loftus, J., Silva, R., 2017. Counterfactual fairness.
Li, T., Sanjabi, M., Smith, V., 2019. Fair resource allocation in federated learning. arXiv: Learning.
Madras, D., Creager, E., Pitassi, T., Zemel, R., 2018. Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309.
Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems 29.
Menon, A.K., Williamson, R.C., 2018. The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency, pp. 107–118.
Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., Camps-Valls, G., 2017. Fair kernel learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 339–355.
Vasudevan, S., Kenthapadi, K., 2020. LiFT: a scalable framework for measuring fairness in ML applications. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2773–2780.
McMahan, H.B., Ramage, D., Talwar, K., Zhang, L., 2017b. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.0696.
Brendan McMahan, H., Moore, E., Ramage, D., Hampson, S., Ag¨uera y Arcas, B., 2016. Communication-efficient learning of deep networks from decentralized data. arXiv e-prints , arXiv–1602.
Williamson, R., Menon, A., 2019. Fairness risk measures, in: International Conference on Machine Learning. PMLR 6786–6797.
Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., Cui, W., 2020. Algorithmic decision making with conditional fairness. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2125–2135.

Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P., 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide web, pp. 1171–1180.

Zhang, T., Li, J., Han, M., Zhou, W., Yu, P., et al., 2020. Fairness in semi-supervised learning: unlabeled data help to reduce discrimination. IEEE Trans. Knowl. Data Eng. 34, 1763–1774.

Zhang, T., Zhu, T., Gao, K., Zhou, W., Philip, S.Y., 2021. Balancing learning model privacy, fairness, and accuracy with early stopping criteria. IEEE Trans. Neural Netw. Learn. Syst. 1–13. doi:10.1109/TNNLS.2021.3129592.

Zhang, T., Zhu, T., Li, J., Han, M., Yu, P.S., 2020. Fairness in semi-supervised learning: unlabeled data help to reduce discrimination. IEEE Trans. Knowl. Data Eng. 34, 1763–1774. doi:10.1109/TKDE.2020.3002567.

Zhu, T., Ye, D., Wang, W., Zhou, W., Yu, P.S., 2022. More than privacy: applying differential privacy in key areas of artificial intelligence. IEEE Trans. Knowl. Data Eng. 34 (6), 2824–2843. doi:10.1109/TKDE.2020.3014246.

Gu Xiuting received her B.Eng. degree from Yanshan University, China, in 2013. She is currently a master student at China University of Geosciences, Wuhan, China. Her research interests include privacy preserving, AI security and privacy, and network security.

Zhu Tianqing received her B.Eng. degree and her M.Eng. degree from Wuhan University, China, in 2000 and 2004, respectively. She also holds a Ph.D. in computer science from Deakin University, Australia (2014). She is currently a professor at China University of Geosciences, Wuhan, China. Her research interests include privacy preserving, AI security and privacy, and network security.

Jie Li is currently an undergraduate student at China University of Geosciences, Wuhan, China. Her research interests include privacy preserving, AI security and privacy, and network security.

Tao Zhang received the B.Eng and M.Eng degrees from the Information Engineering School, Nanchang University, China, in 2015 and 2018, respectively. Currently, he is working towards his Ph.D. degree with the school of Computer Science in the University of Technology Sydney, Australia. His research interests include privacy preserving, AI fairness, and machine learning.

Wei Ren currently is a Professor at the School of Computer Science, China University of Geosciences (Wuhan), China. He was with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, USA in 2007 and 2008, the School of Computer Science, University of Nevada Las Vegas, USA in 2006 and 2007, and the Department of Computer Science, The Hong Kong University of Science and Technology, in 2004 and 2005. He obtained his Ph.D. degree in Computer Science from Huazhong University of Science and Technology, China. He has published more than 70 refereed papers, 1 monograph, and 4 textbooks. He has obtained 10 patents and 5 innovation awards. He is a senior member of the China Computer Federation and a member of IEEE.

Kim-Kwang Raymond Choo received his Ph.D. in Information Security in 2006 from Queensland University of Technology, Australia. He currently holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio. He is the founding co-Editor-in-Chief of ACM Distributed Ledger Technologies: Research & Practice, the founding Chair of IEEE Technology and Engineering Management Technical Committee on Blockchain and Distributed Ledger Technologies, an ACM Distinguished Speaker and IEEE Computer Society Distinguished Visitor (2021 - 2023), and a Web of Science's Highly Cited Researcher (Computer Science – 2021, Cross-Field – 2020). He is the recipient of the IEEE Systems, Man, and Cybernetics Technical Committee on Homeland Security Research and Innovation Award in 2022, and the 2019 IEEE Technical Committee on Scalable Computing Award for Excellence in Scalable Computing (Middle Career Researcher). He has also received best paper awards from IEEE Systems Journal in 2021, IEEE Computer Society's Bio-Inspired Computing Special Technical Committee Outstanding Paper Award for 2021, IEEE DSC 2021, IEEE Consumer Electronics Magazine for 2020, Journal of Network and Computer Applications for 2020, EURASIP Journal on Wireless Communications and Networking in 2019, IEEE TrustCom 2018, and ESORICS 2015.