BIAS PROPAGATION IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that participating in federated learning can be detrimental to group fairness. In fact, the bias of a few biased parties against under-represented groups (identified by sensitive attributes such as gender or race) propagates through the network to all parties. On naturally partitioned real-world datasets, we analyze and explain bias propagation in federated learning. Our analysis reveals that biased parties unintentionally yet stealthily encode their bias in a small number of model parameters, and throughout the training, they steadily increase the dependence of the global model on sensitive attributes. What is important to highlight is that the experienced bias in federated learning is higher than what parties would otherwise encounter in centralized training with a model trained on the union of all their data. This indicates that the bias is due to the algorithm. Our work calls for auditing group fairness in federated learning, and designing learning algorithms that are robust to bias propagation.

1 INTRODUCTION

Machine learning models can be biased against demographic groups. There has been plenty of work studying how machine learning algorithms reflect and even amplify the bias in the training data in the centralized setting, where the data is held by a single party (Hardt et al., 2016; Dwork et al., 2012; Calders et al., 2009; Hashimoto et al., 2018; Zhang et al., 2020; Blum and Stangl, 2020; Lakkaraju et al., 2017). However,, in practice, data is commonly owned by multiple parties and cannot be shared due to privacy concerns. Federated learning (FL) enables parties to collaboratively learn a global model without sharing the data. In each round, parties share the local model update computed on the dataset with a global server, and the server aggregates local updates to update the global model. Although FL is now widely deployed in various applications, such as healthcare, recruitment, and loan evaluation (Rieke et al., 2020; Yang et al., 2019), it is not fully understood whether and how FL algorithms could magnify the bias in the training datasets.

A few recent works study this problem focusing on measuring and mitigating the bias in FL with respect to a single global distribution (Chu et al., 2021; Zeng et al., 2021; Hu et al., 2022; Du et al., 2021; Abay et al., 2020; Papadaki et al., 2021; 2022; Hu et al., 2022). However, parties usually have *heterogeneous* data distributions. Evaluating the model bias with respect to the global distribution does not reflect how fair the FL model is with respect to parties' heterogeneous local data distributions (which are of relevance to users). This is exactly the critical problem that we address in this paper. Specifically, we ask: compared to standalone training (train model without any collaboration), can parties benefit from joining FL to reduce bias and build fairer models? Does FL provide parties with an equivalent fairness benefit of centralized training (on the union of their data)? Can parties with biased datasets deteriorate the local fairness of other parties? How and why does bias in a small number of parties affect the network? To the best of our knowledge, we provide the first comprehensive analysis of how the FL algorithm impacts local fairness.

We provide an empirical analysis based on real-world datasets. We show that the FL might not sustain the benefits of collaboration in terms of fairness, as compared to its accuracy benefit. Specifically, we find that the model trained in a centralized setting can be, on average, fairer on local data distributions than the standalone models. However, in those cases, the FL model, trained on the same dataset, can be more biased. This suggests that the FL algorithm itself can introduce new bias in the final model, deteriorating fairness for parties. Furthermore, we demonstrate that FL impacts local fairness for parties in different ways. Specifically, we find a strong correlation between parties' fairness gap in the standalone setting and the fairness benefit they obtain from joining FL: parties with greater bias in

the standalone setting (caused by their local data) would receive a fairer model from FL. In contrast, FL has negative impact on parties with less significant bias in the standalone setting, resulting in a more biased model in FL. We further demonstrate that this is due to the fact that **FL propagates bias**: bias from a few parties can now influence all parties in the network, hence aggravating the fairness problem among all parties.

Finally, we offer potential explanations for how bias is propagated in FL. In particular, we show that *local updates* from biased parties increase the model's prediction dependence on the sensitive attribute. Such increase is achieved by the norm increase of a few parameters (e.g., 104 parameters out of 1,792 parameters). This increase then impacts the global model through aggregation, subsequently impacting all other parties. In addition, we also show that the fairness gap of the final model can be governed by adjusting the value of those parameters. Surprisingly, by scaling those parameters, we can reduce the bias of the model to a small value of 0.05 (the measurement for bias takes a value from 0 to 1) with only 0.6% accuracy drop or increase the bias to a large value of 0.96 with 11% accuracy drop.

2 BACKGROUND

Federated Learning. We follow the conventional FL setting (McMahan et al., 2017). There is a network consisting of K parties. Each party $k \in [K]$ holds a dataset $\tilde{D}_k$ of size n_k sampled from local distribution $\mathcal{D}_k$. The goal of each party is to learn a model that minimizes the loss on $\mathcal{D}_k$. Federated learning achieves this goal by training a global model to minimize the average loss across parties, i.e., $\min_\theta \frac{1}{N}\sum_{k=1}^{K} n_k L(\theta, \tilde{D}_k)$, where N is the sum of all local training dataset size. At each communication round t, a global server sends the current global model to all parties. All parties train the global model locally on their local dataset and send the updated local model to the server. The server then aggregates those local models to obtain the new global model. Note that, we consider the case where all parties participate in the training in each round. It eliminates the potential bias caused by the sampling of participating parties.

Group Fairness. Fairness has a wide variety of meanings in literature. Group fairness entails, in particular, that the model should perform comparably across groups defined by sensitive attributes (e.g., sex). It is now common practice to evaluate discrimination in a model (or system) based on quantitative measurements of group fairness. In light of this, we focus on two widely-used group fairness notions, equalized odds (Hardt et al., 2016) and demographic parity (Dwork et al., 2012). To formally define those fairness notions, we assume each data point is associated with a sensitive attribute $a \in \mathcal{A}$, and we use Y to denote the true label.

Definition 1 (Equalized Odds (Hardt et al., 2016))**.** *We say a model f_θ is fair with respect to equalized odds if the prediction is conditionally independent of the sensitive attribute A given the true label Y, i.e., $\forall a, a' \in \mathcal{A}, y \in \mathcal{Y}, \Pr(f_\theta(X) = y | A = a, Y = y) = \Pr(f_\theta(X) = y | A = a', Y = y)$. To measure fairness, we use the fairness gap with respect to Equalized Odds Difference, defined as*

$$\Delta^{EO}(\theta, \mathcal{D}) := \max_{a,a' \in \mathcal{A}, y \in \mathcal{Y}} |\Pr_{\mathcal{D}}(f_\theta(X) = +|A = a, Y = y) - \Pr_{\mathcal{D}}(f_\theta(X) = +|A = a', Y = y)|$$

with an ideal value equal to zero.

Furthermore, in many applications, there exists a favorable prediction from the model (e.g., grant the loan). We assume the positive prediction (+) as the favorable outcome. *Demographic parity* (Dwork et al., 2012) group fairness notion asks the model to give a favorable label to groups with equal rates.

Definition 2 (Demographic parity (Dwork et al., 2012))**.** *We say a model f_θ is fair with respect to demographic parity if the positive prediction is independent of the sensitive attribute A, i.e., $\forall a, a' \in \mathcal{A}, \Pr(f_\theta(X) = +|A = a) = \Pr(f_\theta(X) = +|A = a')$. To measure fairness, we use the fairness gap with respect to Demographic Parity Difference, defined as*

$$\Delta^{DP}(\theta, \mathcal{D}) := \max_{a,a' \in \mathcal{A}} |\Pr_{\mathcal{D}}(f_\theta(X) = +|A = a) - \Pr_{\mathcal{D}}(f_\theta(X) = +|A = a')|$$

with an ideal value equal to zero.

In the rest of the paper, we use the fairness gap Δ (including Δ^{EO} and Δ^{DP}) to measure the bias (fairness) of a model. The more significant fairness gap means the model is more biased (less fair).

Measuring impact of FL. Federated learning strives to improve model performance compared to standalone training, and the ultimate goal is to achieve comparable performance to centralized training. Thus, in order to quantify the impact of FL on local fairness, we use centralized training and standalone training as baselines. In standalone training, each party trains a model θ_k independently to minimize the loss on its training data $\tilde{D}_k$. Note that the standalone model's fairness gap for a party is contributed by herself. Thus, we use the fairness gap of a party's standalone model to represent the party's bias. On the other hand, in centralized training, all local training datasets from all parties are combined, and a centralized model θ_c is trained using this global training dataset.

We define the benefit of FL and collaboration (i.e., centralized training) in terms of fairness and accuracy using these baselines. The benefit of FL for each party k is defined by the difference between the standalone model θ_k and the FL model θ_g. It reveals the extent to which a party benefits from participating in FL. More formally, for a party k, we define the **accuracy benefit of FL** as $Acc(\theta_g, \mathcal{D}_k) - Acc(\theta_k, \mathcal{D}_k)$ and **fairness benefit of FL** as $\Delta(\theta_k, \mathcal{D}_k) - \Delta(\theta_g, \mathcal{D}_k)$. Similarly, the benefit of collaboration is based on the difference between the standalone model and the centralized model. It indicates how much an individual party could gain from collaborating with others by sharing the entire training dataset. The average benefit of FL or collaboration is the average benefit across all parties in the network. The positive benefit indicates that FL (or collaboration) helps to improve accuracy or fairness, while the negative benefit implies that FL or collaboration hurts.

Measuring bias aside fairness gap. Measuring the bias of a model based on the fairness gap reveals the model's performance disparity across groups, but does not explains why the model is biased. Specifically, is the prediction of the model attributed to the sensitive attribute or any other insensitive attributes whose distributions vary between groups? The former is typically referred to as disparate treatment (direct discrimination) as the sensitive attribute directly influences the model prediction (Grabowicz et al., 2022; Zafar et al., 2017). Towards measuring bias in this aspect, we employ the feature attribution method to explain the model's prediction on each data point. More precisely, given a model f_θ, and the input features $x = (x_1, x_2, ..., x_d)$, the attribution of the prediction at input x relative to a baseline input x' is a vector $(a_1, ..., a_d)$ where a_i is the contribution of x_i to the prediction of $f_\theta(x)$. We use Integrated Gradient (Sundararajan et al., 2017) to compute the feature attribution, which only requires a few calls to the gradient operation. Following the suggestions in the paper, we use the average feature values computed on the whole test dataset as the baseline for non-sensitive and non-binary sensitive features. For the binary sensitive attribute, we use the opposite feature value as the baseline. For instance, if the test input is (Sex: female, Age:28), the baseline input would be (Sex: male, Age: average age over the data points). By employing this feature attribution, we are able to discover the bias associated with disparate treatment. For more details about the Integrated Gradient, please refer to the original paper (Sundararajan et al., 2017).

3 Empirical Analysis

Our objective is to comprehend how FL impacts fairness for parties. Toward this end, we start by the comparison of average performance between FL and baselines (i.e., standalone training and centralized training). In particular, we ask: ***does FL improve fairness for clients compared to standalone training, and fairness-wise, does FL provide the same benefit as centralized training?*** We provide answers in Section 3.2, where we demonstrate that FL can exacerbate the fairness issue for parties and does not retain the fairness benefit of collaboration. After that, we further investigate how FL impacts performance for each party. Specifically, we aim to answer the following two questions: ***does FL provide the same benefit in terms of fairness for parties? What causes the disparate impact of FL on parties' fairness?*** We provide analysis in Section 3.3, where we show FL can propagate the bias among parties. As a consequence, FL improves fairness for parties with greater bias while worsening fairness issues for parties with less bias. Finally, we investigate ***how the bias is propagated in FL***. The analysis in Section 3.4 shows that biased parties encode the bias in a few parameters through local update, which is propagated to the aggregated model and ultimately to other parties via parameter aggregation.

3.1 Setup

We use the two datasets for our empirical analysis: ***US Census Data*** and ***CelebA (Liu et al., 2015)***. The US Census dataset consists of census data from different places in the US. We naturally partition

Table 1: **Average benefit of collaboration and FL**. The lowest accuracy and highest fairness gap are bold. The standard deviation across five runs is indicated between parentheses. FL improves the accuracy but worsens the fairness issue for clients.

	Accuracy			Δ^{EO}			Δ^{DP}		
Dataset	Standalone	Centralized	FedAvg	Standalone	Centralized	FedAvg	Standalone	Centralized	FedAvg
Income- Race	**.738** (.001)	.787 (.003) ↑	.785 (.002) ↑	.506 (.026)	.478 (.013) ↓	**.514** (.021) ↑	.321 (.011)	.307 (.012) ↓	**.322** (.004) ↑
Income- Sex	**.738** (.001)	.787 (.003) ↑	.785 (.002) ↑	.146 (.007)	.177 (.030) ↑	**.198** (.007) ↑	.167 (.005)	.203 (.008) ↑	**.217** (.005) ↑
Health- Race	**.683** (.002)	.714 (.002) ↑	.709 (.001) ↑	**.226** (.013)	.17 (.021) ↓	.161 (.021) ↓	**.108** (.004)	.079 (.004) ↓	.072 (.008) ↓
Health- Sex	**.683** (.002)	.714 (.002) ↑	.709 (.001) ↑	**.042** (.003)	.034 (.002) ↓	.029 (.005) ↓	**.019** (.002)	.015 (.002) ↓	.013 (.004) ↓
Employment- Race	**.759** (.001)	.824 (.002) ↑	.823 (.001) ↑	**.374** (.023)	.243 (.010) ↓	.237 (.003) ↓	**.287** (.009)	.240 (.007) ↓	.241 (.010) ↓
Employment- Sex	**.759** (.001)	.824 (.002) ↑	.823 (.001) ↑	**.092** (.004)	.059 (.005) ↓	.059 (.003) ↓	**.054** (.003)	.040 (.007) ↓	.038 (.006) ↓
CelebA- Sex	**.728** (.007)	.759 (.021) ↑	.793 (.002) ↑	.281 (.009)	.267 (.039) ↓	**.315** (.008) ↑	.398 (.011)	.416 (.049) ↑	**.485** (.009) ↑

the dataset based on the source of the data. Thus, we have 51 parties in total, representing 50 states in the US and Puerto Rico. We consider three tasks defined in the folktables (Ding et al., 2021): Income (ACSIncome) [1], Health (ACSPublicCoverage) and Employment (ACSEmployment). We train two-layer neural network models for all the tasks. We consider the binary notion of sex (i.e., male and female) and multi-value race as the sensitive attribute for every task. Each party has 1,000 training points and 2,000 test points. Note that we use a large test dataset to ensure an accurate estimation of the fairness gap with respect to parities' test distribution. CelebA (Liu et al., 2015) is an image dataset consisted of 200,000 celebrity images. We consider the task of predicting whether the person in a given image is attractive, which may be critical for job applications in particular fields (e.g., model, actors and actresses) (Dion et al., 1972; Eagly et al., 1991). We consider the binary attribute "Male" (male and non-male) as the sensitive attribute, which we will refer to as the Sex attribute. We partition the data uniformly at random among all parties and train CNN models.

For all the experiments, we report the average results over five runs with different random seeds. We use FedML (He et al., 2020), a library for FL based on PyTorch Paszke et al. (2019), to train models in FL. We also utilize Captum (Kokhlikyan et al., 2020) for computing the feature attribution. More details about the datasets and models can be found in Appendix A.

3.2 Collaboration via FL can worsen fairness issue

Table 1 shows the average accuracy and fairness gaps of the centralized model, FL model, and standalone models on local datasets. We observe that centralized training can improve accuracy and fairness at the same time for parties. For instance, on the Income dataset, centralized training improves the accuracy by 6% and reduces the fairness gap across racial groups by 5.5% with respect to equalized odds and by 5.5% with respect to demographic parity. However, we notice that FL may not achieve the same benefit as the centralized training in terms of fairness but can even worsen the fairness issue for parties. For instance, on the Income dataset, the EO fairness gap of the FL model for the sex groups increases by 35%, and the DP fairness gap increases by 29.9%, compared to standalone models. Note that the centralized and FL models are trained on the same dataset. The difference between the FL model and the centralized model in terms of fairness suggests FL algorithm can introduce more bias compared to standard training. Consequently, the explanation of how bias is introduced to the model during the standard training in the centralized setting may not adequately explain how FL introduces bias to the model. In the following, we further explore how FL impacts fairness from the party level.

3.3 FL propagates bias among parties

Disparate impact of FL on group fairness over parties. Figure 1 shows the benefit of FL on fairness and accuracy for parties. FL improves accuracy for almost all parties, and the improvement variance across parties is small. On the other hand, the fairness benefit of FL is negative for most parties. It implies that most parties obtain a more biased model in FL compared to standalone training. Moreover, we notice that the variance of the fairness benefit across parties is large. It shows that even though all parties have the same global model in FL, they are not benefiting from the model equally (recall that each party tests the model on her local test dataset).

[1]We use the Adult Reconstruction dataset (Sarah et al., 2020).

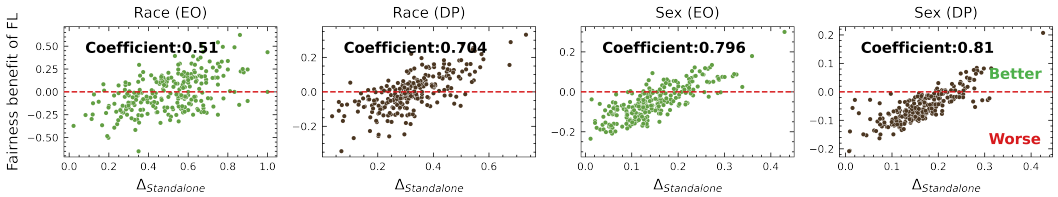

Figure 2: **Correlation between the fairness gap of the standalone model and the benefit obtained from FL - Income.** The x-axis shows the fairness gap of the standalone model, and the y-axis shows the fairness benefit of FL, defined in Section 2. The Pearson correlation coefficients between the fairness gap of the parties' standalone models and the fairness benefit they obtain from FL are presented in the Figure. FL improves fairness for more biased parties (i.e., who obtain a more biased model in the standalone setting) while exacerbating fairness issues for less biased parties.

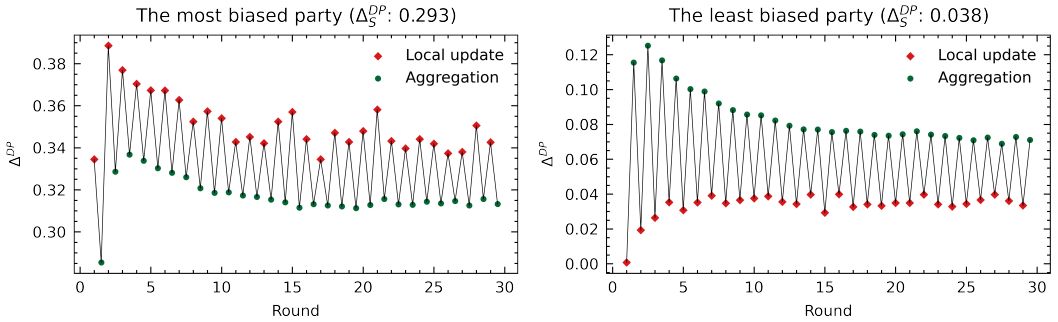

Figure 3: **Dynamic of Δ^{DP} during the training - Income (sex).** Figure shows the fairness gap of the aggregated model and local updated model from the most biased party and least biased party in the first 30 rounds (Figure 14 in Appendix B shows the results for all training rounds.). The most (least) biased party is the party with the highest (lowest) fairness gap in the standalone setting. The fairness gap in the standalone setting for those parties is shown in the title. Aggregation and local updates have contradictory impacts on model fairness for those parties.

Figure 2 shows strong correlations between the fairness benefit a party acquires in FL and the fairness gap of the standalone model for the party (i.e., bias level of the party). It further demonstrates the disparate impact of FL on fairness: FL can improve fairness for more biased parties but at the cost of worsening the issue for less biased parties.

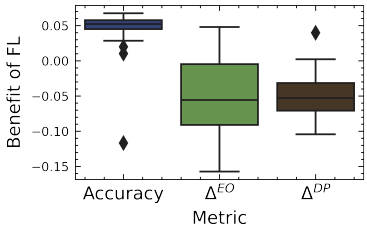

Figure 1: **Accuracy and Fairness Benefit of FL - Income (Sex).** FL has disparate impact on fairness across parties.

Aggregation contradicts with local update. Figure 3 shows the dynamic of the fairness gap before and after the aggregation. We find that aggregation and local update contradicts each other during the training. More specifically, the local update for the least biased party (whose standalone model has the lowest fairness gap) reduces the fairness gap of the model during the local update. However, this reduction is eliminated by aggregation. On the other hand, the most biased parties increase the models' bias through the local update, which will be reduced via aggregation. It indicates that the aggregation contributes to the disparate impact of FL on fairness, improving fairness for more biased parties while worsening the fairness issue for less biased parties compared to the standalone setting.

Biased parties negatively influence other parties via aggregation throughout the training. Why does aggregation have disparate impact on local fairness for parties? Specifically, we ask which party's local update causes the increase (or decrease) of the fairness gap for other parties via aggregation. To answer this question, we measure the influence of a party on other parties' local fairness by the influence of the party's local update on the fairness of the aggregated model during the training. More

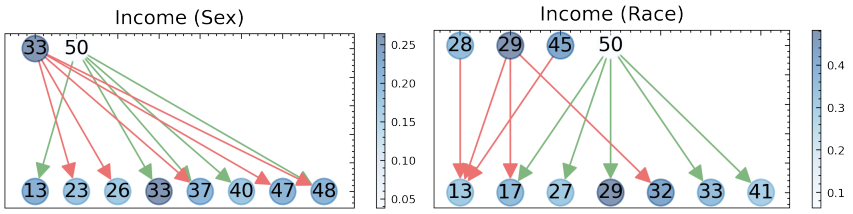

Figure 4: **Influence graphs - Income (Δ^{DP}).** We show the most influential pairs of parties. The nodes on the top represent parties who influence other parties, and the nodes on the bottom represent parties who are influenced by other parties. The number inside the node represents the party index. The color of the node represents the fairness gap in the standalone setting. The green edge (resp. red edge) connects the pair of parties when the party on the top positively (resp. negatively) influences the party on the bottom. We show the top 5 pairs with maximal positive influence and the top 5 pairs with maximal negative influence. The most influential parties are extremely biased parties and extremely less biased parties, and the biased parties have negative influences on other parties, while the less biased party has positive influences on other parties.

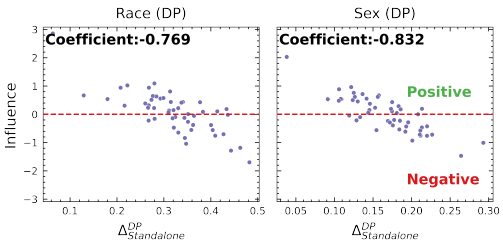

Figure 5: **Correlation between the influence and standalone bias - Income.** The y-axis represents the average influence of each party. The Pearson correlation between the fairness gap a party obtains in the standalone setting and the impact she has on local fairness for parties in FL are shown. The greater the party's bias, the greater its adverse effect on others.

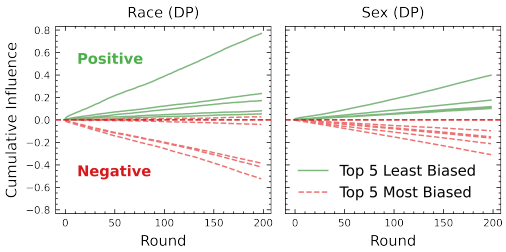

Figure 6: **Cumulative influence - Income.** At each round, the cumulative influence of each party up until the current round is shown. We show the results for the top five most and least biased parties. Less biased parties have a positive impact on the fairness of other parties, whereas biased parties continually have a negative impact on the fairness of other parties.

precisely, we compute the influence of party i on party j as the fairness gap increase when party i's local update is removed from the aggregation in each round t and sum over all the training rounds, which is formally defined as follows: $I_{i,j} = \sum_{t=1}^{T} \Delta(\theta_{t,-i}, D_j) - \Delta(\theta_t, D_j)$, where θ_t is the global model (i.e., the aggregated model overall local updated models) in round t and $\theta_{t,-i}$ is the aggregated model over local updated models from parties excluding party i. If party i improves fairness for party j, the influence is positive and vice versa. We compute the influence for all pairs of parties, and the most influential pairs are shown in Figure 4. We observe that a less biased party has positive influence on fairness for other parties while a more biased party has negative influence on other parties.

Furthermore, we show the correlation between the bias a party has (i.e., the bias of the party's standalone model) and its average influence on all parties' fairness (i.e., $\sum_{k=1}^{K} I_{i,k}/K$) in Figure 5. The strong correlation between those two further supports that the more biased parties have a stronger negative influence, while the less biased parties have a stronger positive influence on other parties. Lastly, Figure 6 shows the dynamic of the influence of the top 5 most biased parties (i.e., parties who have the largest fairness gap in the standalone setting) and the top 5 least biased parties. We notice that the influence is monotonically increasing for the most biased parties and decreasing for the least biased parties. In other words, biased (less biased) parties consistently have a negative (positive) influence on others' fairness gaps. Our results imply that FL can propagate bias among parties: the bias from biased parties negatively influence the fairness of other parties via aggregation throughout the training. Next, we will explore how the bias is propagated in FL.

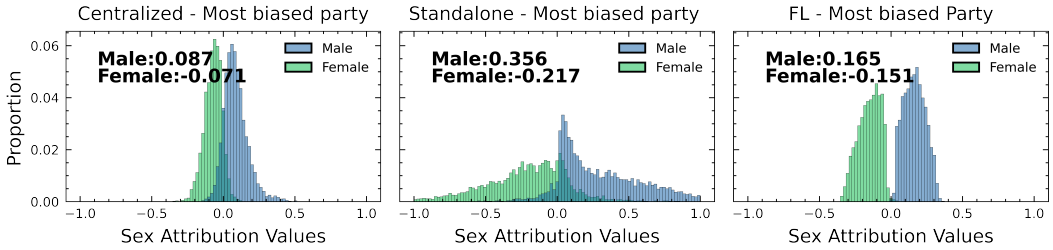

Figure 7: **Histogram of feature attribution value for "Sex" - Income.** The average attribution value for the female group and male group is shown in the figure. The results are computed on all test data points over five different runs. We show the results for the most biased party in the standalone setting. FL model's prediction has a larger dependency on the sensitive attribute.

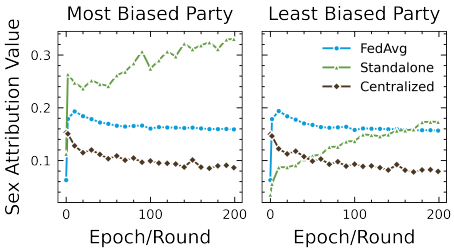

Figure 8: **Dynamic of absolute attribution value for sex - Income.** We show results for the parties who have the highest or lowest fairness gap in the standalone setting. The standalone training and FL gradually increase the model dependence on the sensitive attribute, while the centralized training reduces the dependence during the training.

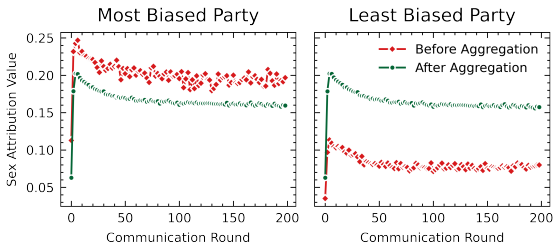

Figure 9: **Dynamic of absolute attribution value for sex before and after aggregation in FL - Income.** The local update (before aggregation) from the least biased party reduces the model dependence on the sensitive attribute while the most biased party increases the model dependence on the sensitive attribute throughout the training, which will affect the FL model via aggregation.

3.4 How bias is propagated in FL?

Large fairness gap is caused by disparate treatment. We first investigate what the bias represents. More specifically, is the bias (fairness gap) increase in FL directly caused by the disparate treatment of the model among sensitive groups? To answer this question, we utilize Integrated Gradients (Sundararajan et al., 2017) to measure the attribution of each input feature to the predictions of models with respect to the positive class. Figure 7 shows the histogram of the attribution value for the sensitive attribute, "Sex", over individual test points from the female group and the male group. We notice that the sex attribute has a large attribution value for the standalone model's predictions and FL model's predictions. It means that the prediction of those models are heavily dependent on the sensitive attribute of the test data. Additionally, individuals' sensitive attribute affects the model predictions in a different way. The average attribution value for the "Sex" attribute is positive for the male group and negative for the female group. Moreover, we find that there is less difference in the attribution value for the female and male groups with respect to other attributes (see Figure 17 in Appendix B). It implies that the large fairness gap for models is not caused by the distinct distribution of other insensitive attributes over protected groups; rather, it is mainly caused by the disparate treatment of the model on protected groups.

FL model learns more biased patterns. Comparing the attribution value of "Sex" for different models in Figure 7, we find that, although the predictions of FL model depend less heavily on sensitive attribute compared to that of the standalone model, the dependence is still stronger than that of the centralized model. It suggests that the model trained in FL learns a more biased pattern compared to the model that could otherwise be learned in the centralized setting.

Figure 8 shows the dynamic of the average absolute feature attribution for "Sex" during the training. We find that, for the most biased party, standalone training increases the model's dependence on

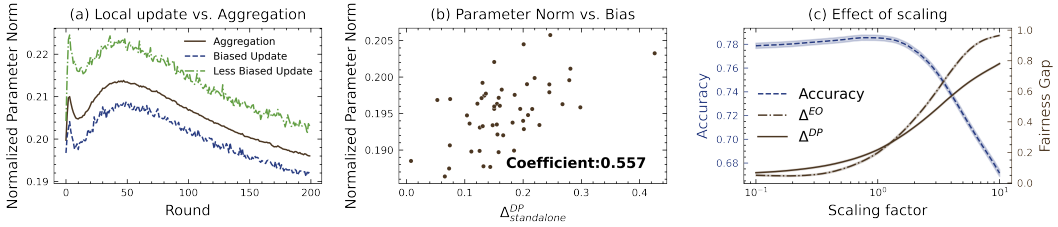

Figure 10: **Bias is encoded in the model parameters.** (a) depicts how the norm of parameters that are related to sex attribute changes during local update and aggregation. The most biased party raises the parameter norm, while the least biased party decreases it. (b) illustrates the relationship between the bias a party receives in the standalone setting and the parameter norm (associated with sex attribute) of its local updated model in the last round. (c) displays the effect of scaling the parameters associated with the sex attribute on fairness and accuracy.

the sensitive attribute throughout training, but centralized training decreases it. It suggests that collaboration through centralized training improves local fairness by guiding the model to learn less biased features. In FL, however, the absolute attribution value barely changes after 100 rounds and remains significantly greater than that in the centralized setting. It implies that FL algorithm may introduce the bias to the final model by inhibiting the model from learning less biased features.

Biased parties increase the model dependence on the sensitive attribute. Figure 9 shows the attribution value of the aggregated model and local updated model from the most biased party and least biased party. We observe that the biased party increases the global model dependence on sensitive attribute during the local update, and this increase is persistent over the training, while the least biased party, on the other hand, reduces the dependence on the sensitive attribute, which is aligned with the trend of fairness gap in Figure 3.

Bias is encoded in a few parameters. Biased parties increase the model dependence on the sensitive attribute. But how does this increase propagate to other parties in the network? Note that parties share the model parameters of the local updated model with the server; thus, the bias should be encoded in the parameters. Thus, we examine which parameters are related to the bias of the model. Intuitively, the parameters used to extract sensitive attribute information impact the attribution value of the sensitive attribute to the model prediction. If the absolute value (signal) of those parameters is substantial, the value of the sensitive attribute will have a significant effect on the model's prediction. In our evaluation, sensitive attribute is part of the input feature. Hence, the parameters directly applied to the "Sex" attribute in the first layer of the neural network should contribute to the bias of the model.

Figure 10(a) depicts the normalized norm for those parameters (the norm of parameters associated with the sensitive attribute divided by the parameter norm of the first layer) of aggregated models and local updated models for both the most and least biased parties. We find that the least biased party reduces the parameter norm, hence decreasing the model's sensitivity to the sensitive attribute. In contrast, the biased party raises norm for those parameters, hence amplifying the impact of the sensitive attribute on the model's prediction. Through aggregation, this amplification will be propagated to the global model. Figure 10(b) reveals a moderate correlation between the fairness gap party experiences in a standalone setting and the normalized parameter norm for the parameters used to extract sensitive attribute information. It also demonstrates that biased parties increase the parameter value associated with sensitive attributes during the local update, thereby boosting the model's susceptibility to the sensitive attribute.

Controlling fairness gap by scaling a few parameters. To further validate the effect of these parameters on model fairness, Figure 10(c) demonstrates that by scaling a few parameters associated with the sensitive attribute (i.e., 104 out of 1,792), the fairness gap can be greatly widened or narrowed at the expense of a moderate degree of accuracy. Specifically, by scaling the parameter value by 0.1 for the trained FL model, we can reduce the EO gap by 74.6% (from 0.198 to 0.05) and the DP gap by 69.1% (from 0.217 to 0.067) with just a 0.8% accuracy loss (from 0.785 to 0.779). In contrast, by raising the parameter value by a factor of 10, the EO gap increases to 0.96 (the maximal fairness gap is 1), which is almost five times large. The DP gap increases to 0.78, which is increased by 259%. At the same time, the accuracy decreases from 0.785 to 0.67. These findings explain how bias

propagates in FL: the biased parties magnify the impact of sensitive attributes on model predictions by increasing the model parameter used to extract sensitive attributes. This rise in parameters is subsequently propagated to the global model through aggregation, further aggravating the issue of fairness for other parties. Our results explain how bias is propagated in FL: **Biased parties encode bias in a few parameters through a local update, and this bias is consequently propagated to the entire network through parameter aggregation.**

4 Related work

Fairness has received considerable attention due to the growing deploying of machine learning in decision-making processes. Various definitions of fairness have been presented (Hardt et al., 2016; Dwork et al., 2012; Calders et al., 2009). Specifically, group fairness requires that the model behave similarly for groups defined by a sensitive attribute (e.g., race). Although how machine learning algorithms propagate data bias to the final model has been extensively investigated in a centralized setting (Blum and Stangl, 2020; Lakkaraju et al., 2017; Rambachan and Roth, 2020; Friedler et al., 2019; Dullerud et al., 2022), the effect of FL on model fairness remains unclear.

In FL setting, a substantial body of literature on fairness focuses on the performance disparity of the FL model across parties rather than demographic groups (Li et al., 2021; Zhao and Joshi, 2022; Li et al., 2019; Mohri et al., 2019; Deng et al., 2020; Donahue and Kleinberg, 2021; Hao et al., 2021; Zhou et al., 2021; Yu et al., 2020; Lyu et al., 2020). We focus on group fairness, which concerns performance disparity among groups. Thus, our focus differs from those studies. In terms of group fairness, Abay et al. (2020) listed a few potential sources of bias in FL. Recently, considerable progress has been made in training group fair models in FL (Abay et al., 2020; Chu et al., 2021; Zeng et al., 2021; Hu et al., 2022; Du et al., 2021; Ezzeldin et al., 2021). However, the majority of these works suggest techniques for achieving fairness on a single test distribution. Instead, we focus on fairness issues for parties. Cui et al. (2021); Papadaki et al. (2022) proposed algorithms to improve local fairness for parties. Our purpose, instead, is to gain a comprehensive understanding of how FL influences local fairness on its own, which we believe is equally crucial as designing fair algorithms.

5 Future Work & Conclusion

Future Work. In this work, we have explained how bias is propagated in FL when the sensitive attribute is included as an input feature. In practice, however, sensitive attributes may be prohibited from being included in input features. In such situations, the model may still be heavily biased based on variables that are correlated with the (unobserved) sensitive attribute. For instance, a person's zip code may be highly correlated with their race, a phenomenon known as "redlining". To discern which model parameters contribute to the bias and to audit the bias propagation in this scenario is a promising future work direction. In addition, we also plan to extend our analysis on other types of FL algorithms (e.g., personalized FL) besides FedAvg.

Conclusion. Federated learning has been widely deployed in various applications, which have significant individual-level consequences. Thus, it is essential to anticipate the possible bias introduced by FL. Our paper makes the first step in this direction, offering a comprehensive analysis of the impact of FL (i.e., FedAvg) on local fairness for parties. We demonstrated that the FL algorithm could introduce bias on its own. Consequently, FL may exacerbate the issue of fairness for the involved parties. In addition, we demonstrated that this exacerbation is not equally distributed among parties. This is due to the fact that FL can propagate bias among parties. Finally, we provided potential explanations about how bias is propagated in FL. Specifically, biased parties encode their bias into the local updates by steadily increasing the signal of a few parameters throughout the training process. Such bias is then propagated to other parties via aggregation and ultimately, to the global model parameters.

6 ETHICS STATEMENT

Our analysis focuses on group fairness, which is typically used to audit the model or system for any bias or discrimination. Auditing the bias with respect to other definitions of fairness, such as individual fairness, may result in different conclusions. Moreover, our study focuses primarily on the binary notion of sex attributes and the multi-valued race attribute. We recognize that there are numerous protected groups outside those considered in the analysis, such as those defined by multiple sensitive attributes. The propagation of bias against fine-grained subgroups maybe even more substantial than we found in the paper.

7 REPRODUCIBILITY STATEMENT

We provide the code and script for downloading the data in the supplemental materials. More details can be found in the "REAME.md" file in the supplemental materials. In addition, we provide details about the model, datasets, and implementations in Appendix A.

REFERENCES

Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing*, 2020.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Fair and consistent federated learning. *arXiv e-prints*, pages arXiv–2108, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *arXiv preprint arXiv:2108.04884*, 2021.

Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285, 1972.

Kate Donahue and Jon M. Kleinberg. Models of fairness in federated learning. *CoRR*, abs/2112.00818, 2021. URL `https://arxiv.org/abs/2112.00818`.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, and Marzyeh Ghassemi. Is fairness only metric deep? evaluating and addressing subgroup gaps in deep metric learning. *arXiv preprint arXiv:2203.12748*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL `https://doi.org/10.1145/2090236.2090255`.

Alice H Eagly, Richard D Ashmore, Mona G Makhijani, and Laura C Longo. What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, 110(1):109, 1991.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021)*, 2021.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1905–1916, 2022.

Weituo Hao, Mostafa El-Khamy, Jungwon Lee, Jianyi Zhang, Kevin J Liang, Changyou Chen, and Lawrence Carin Duke. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3310–3319, 2021.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. Fedml: A research library and benchmark for federated machine learning. *NeurIPS 2020 FL Workshop Best Paper Award*, 2020.

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Provably fair federated learning via bounded group loss. *arXiv preprint arXiv:2203.10190*, 2022.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Federating for learning group fair models. *arXiv preprint arXiv:2110.01999*, 2021.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. *arXiv preprint arXiv:2201.08304*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ashesh Rambachan and Jonathan Roth. Bias in, bias out? evaluating the folk wisdom. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Flood Sarah, King Miriam, Rodger Renae, Ruggles Steven, and Warren J. Robert. Integrated public use microdata series, current population survey: Version 8.0 [dataset], 2020. URL https://www.ipums.org/projects/ipums-cps/d030.v8.0.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. Ffd: A federated learning based method for credit card fraud detection. In *International conference on big data*, pages 18–32. Springer, 2019.

Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 393–399, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

Zhiyuan Zhao and Gauri Joshi. A dynamic reweighting strategy for fair federated learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8772–8776. IEEE, 2022.

Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. Towards fair federated learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4100–4101, 2021.

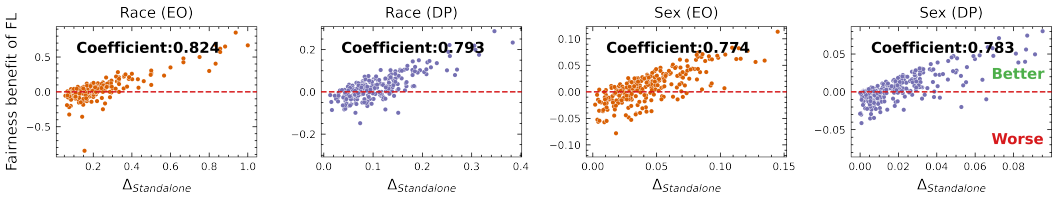

Figure 11: **Correlation between the fairness gap of the standalone model and the benefit obtained from FL - Health**

A Additional Details about Experiments

Datasets and Models We explain the datasets and models used in the paper.

Census Dataset We use the datasets provided from folktables. In particular, we consider the ACSIncome, ACSPublicCoverage, and ACSEmployment tasks defined in the forlktables. In the ACSIncome, the goal is to predict whether an individual's Income is above than $50,000. In ACSEmployment task, the goal is to predict whether an individual is employed. Similarly, in the Health (i.e., ACSPublicCoverage) task, the objective is to predict whether an individual is covered by public health insurance. We use the same pre-processing as in the folktables and train a fully connected neural network model with one hidden layer of 32 neurons for Income and 64 neurons for Employment and Health tasks. For all the tasks, we use the RELU activation function. We use an SGD optimizer with a learning rate of 0.001 for centralized training on Health and Employment datasets and 0.1 for other settings, and the batch size is 32. We train the NN models for 200 epochs. In FL, each client updates the global mdoel for 1 epoch and shares it with the server. We encode the categorical features based on the encoding template provided in folktables. After the encoding, the input feature size for Income is 54, 154 for Health, and 109 for Employment. We consider sex and race as sensitive attributes. Accordingly, there are two gender groups (male and female) and nine racial groups ("White alone," "Black or African American alone," "American Indian alone," "Alaska Native alone," and "American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other," "Asian alone," "Native Hawaiian and Other Pacific Islander alone," "Some Other Race alone," "Two or More Races").

CelebA Dataset We train CNN models on the dataset with one CNN layer whose output channel is 32, kernel size is 3, and stride is 1. We use 'same' padding for the CNN layer. Following this CNN layer, we have the Batch normalization layer and Max Pooling layer. After which, we have the connected layer. We train the model for 500 communication rounds or epochs with SGD optimizers. The learning rate is 0.1, and the batch size is 128. The train, test, and validation datasets ratio for each party is 6:2:2.

Implementation We run all experiments on Ubuntu with two NVIDIA TITAN RTX GPUs.

B Additional Experimental Results

B.1 Bias propagation effect

Census dataset We show the bias propagation effect of FL for Health and Employment task in Figure 11 and 12 respectively.

CelebA We show the bias propagation effect of FL on CelebA dataset. Besides partitioning the data in an IID manner (we refer to as setting (i)), we also evaluate two different settings, where we change the number of samples from the unattractive females. In this way, we aim to change the data bias in the local training datasets for clients. In particular, in setting (ii), half of the parties have more samples from the unattractive female than another half of the parties. The ratio is 8:2. In setting (iii), a single party has half of the data from the unattractive female, and the data is then iid partitioned among the other parties. In the figure, we find that the in the non-iid setting, the more biased parties have a larger and positive benefit while FL hurts the less biased parties. Figure 13 shows the correlation between the fairness gap of the standalone model and the benefit a party gets in the FL. We find that

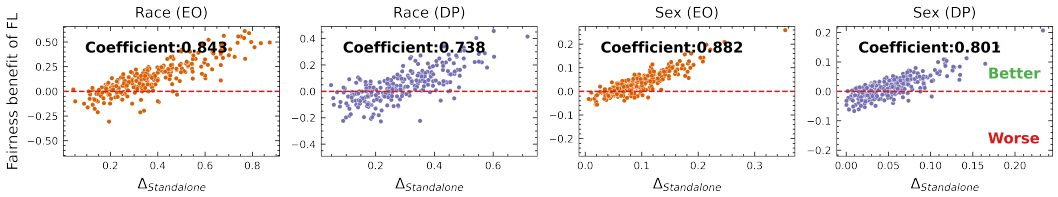

Figure 12: **Correlation between the fairness gap of the standalone model and the benefit obtained from FL - Employment**

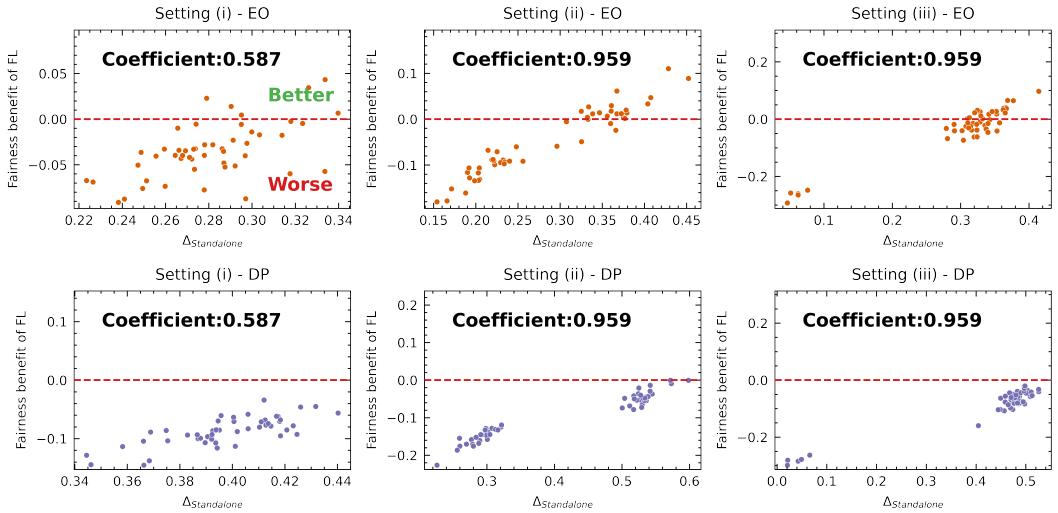

Figure 13: **Correlation between the fairness gap of the standalone model and the benefit obtained from FL - CelebA**

the in the non-iid setting, the more biased parties have a larger and positive benefit while FL hurts the less biased parties.

B.2 Aggregation contradicts with local update

Figure 14 shows the dynamic of the fairness gap for the aggregated model and local updated model from the most biased party and least biased party during the training.

B.3 Influence Subgraphs

We show the top 10 maximal positive and top 10 maximal native influence client pairs in Figure 15 and top 15 in Figure 16. We can see that the less biased client has a strong positive influence on other clients while the more biased client has a strong negative influence on other clients.

B.4 Attribution value for sensitive attribute

We show the attribution value for all input features for the most biased party and least biased party in Figure 17 and Figure 18, respectively. We observe that the attribution value for other features is similar for female and male groups. However, the sex attribute has the largest attribution value and affects the groups differently. It implies that the models are highly dependent on the sensitive attribute (i.e., "Sex") for making predictions.

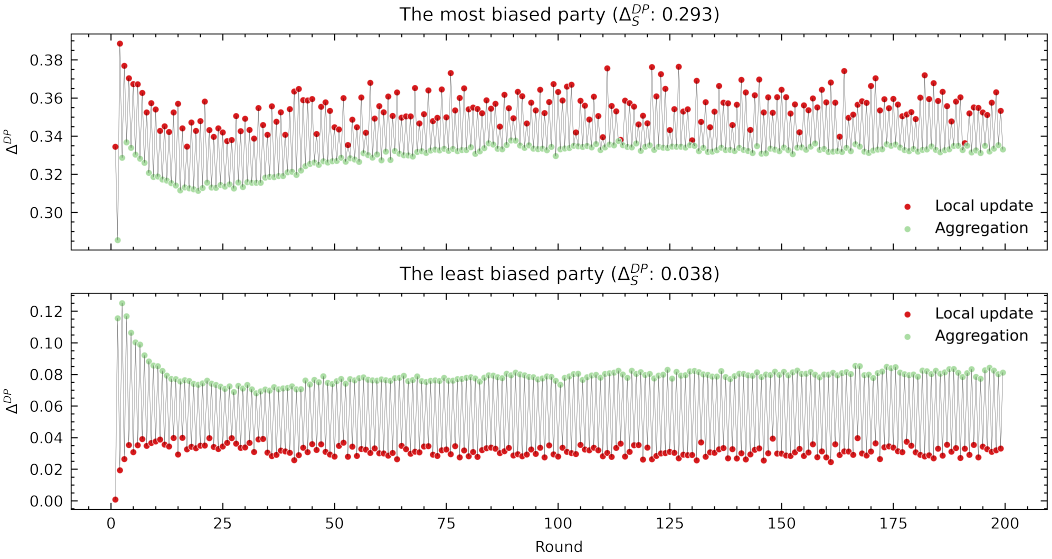

Figure 14: **Dynamic of Δ^{DP} during the training - Income (sex).** Figure shows the fairness gap for the aggregated model and local updated model from the most biased party and least biased party during the training. The most (least) biased party is the party with the highest (lowest) fairness gap in the standalone setting. The fairness gap in the standalone setting for those parties is shown in the title. Aggregation and local updates have contradictory impacts on model fairness for those parties.

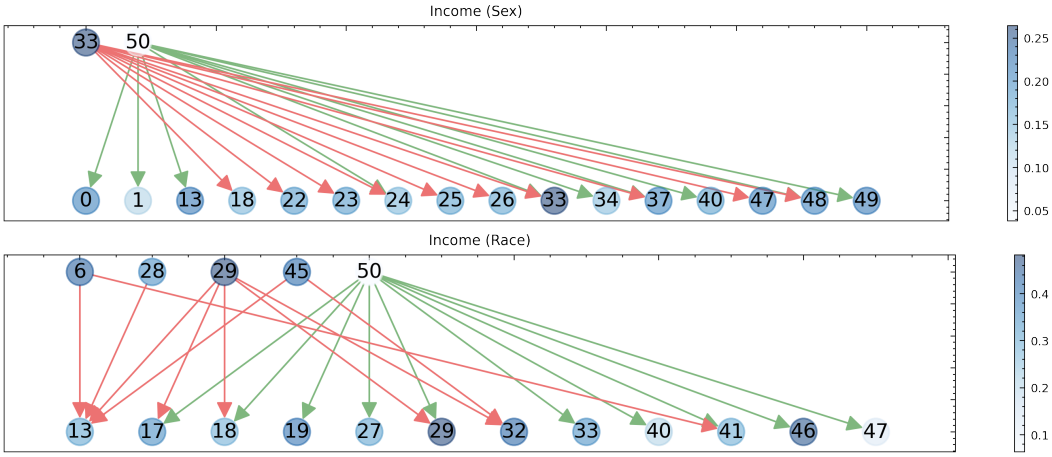

Figure 15: **Influence subgraph - Income (Δ^{DP})** Top 10 maximal positive influential pair and top 10 maximal negative influential pair. The green edge and red edge represent the positive and negative influence, respectively. The color of the node represents the fairness gap in the standalone setting.

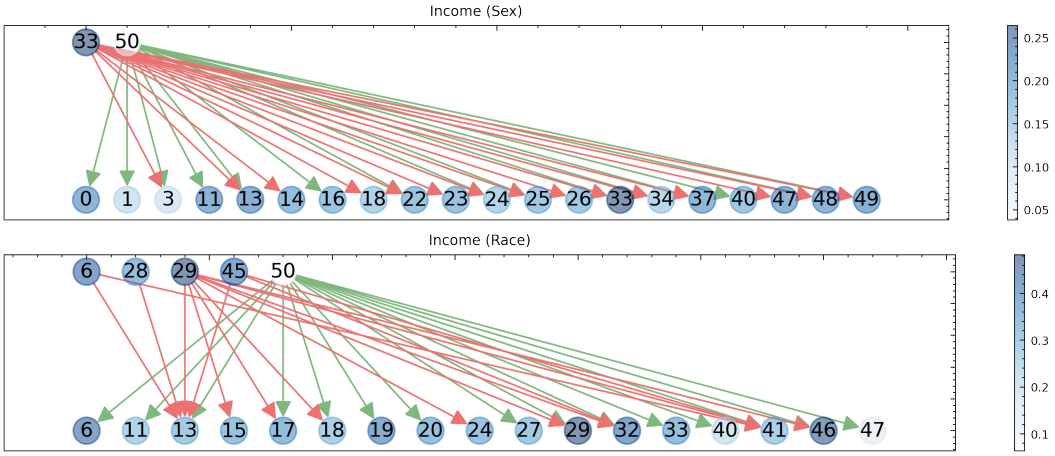

Figure 16: **Influence subgraph - Income** (Δ^{DP}) Top 15 maximal positive influential pair and top 15 maximal negative influential pair. The green edge and red edge represent the positive and negative influence, respectively. The color of the node represents the fairness gap in the standalone setting.

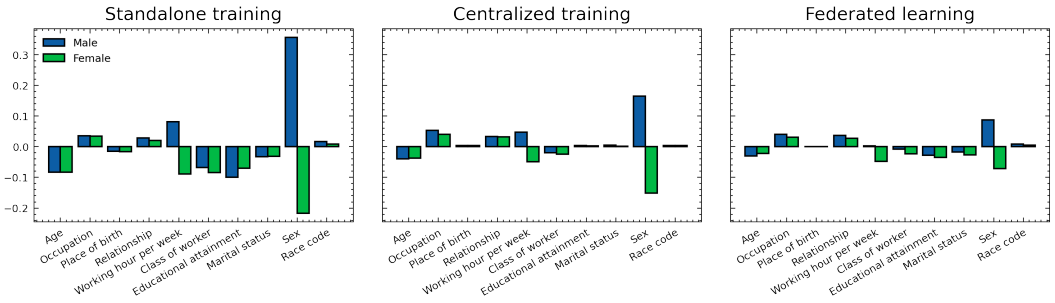

Figure 17: **Feature attribution value - Income (Most biased party)** We shows the feature attribution value for all the input features. The model trained in FL and standalone setting has a large dependency on the sensitive attribute "Sex".

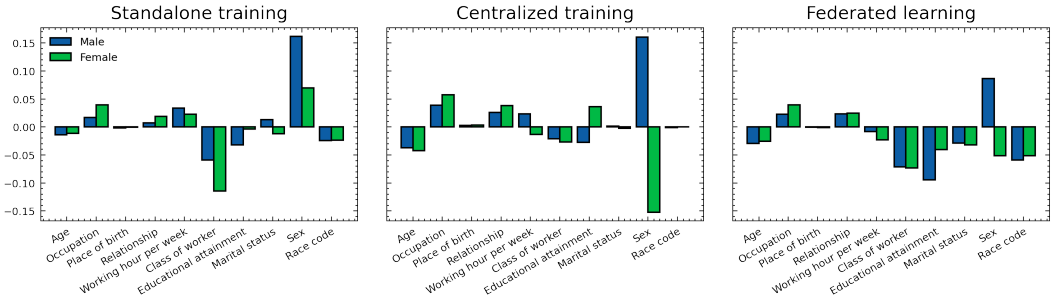

Figure 18: **Feature attribution value - Income (Least biased party)** We shows the feature attribution value for all the input features. The model trained in FL and standalone setting has a large dependency on the sensitive attribute "Sex".