PrivFairFL: Privacy-Preserving Fairness in Federated Learning

Sikha Pentyala[1], Nicola Neophytou[2], Anderson Nascimento[1], Martine De Cock[1,3], Golnoosh Farnadi[2,4]
{sikha, andclay, mdecock}@uw.edu, {nicola.neophytou, farnadig}@mila.quebec
[1]University of Washington Tacoma [2]Mila, Quebec AI Institute [3]Ghent University [4]HEC Montréal

Motivation

Group Fairness aims to ensure bias-free outputs of a machine learning (ML) model, in order to mitigate social biases amongst groups.

Federated learning (FL) protects raw user data, as each user trains a local model and shares only the learned parameters with a central server. However, achieving group fairness to mitigate biases in ML requires a central aggregator to collect sensitive attribute data on all clients. This contradicts the privacy goals of FL.

PrivFairFL: We address the above conflict of privacy and fairness, with a method to train group-fair models with complete privacy guarantees in a cross-device FL setup.

Approach

We use an effective combination of privacy-preserving techniques (PETs):

- **Federated Learning (FL)** to protect the raw data of the users. Our FL models also use DP-SGD during training.
- **Secure Multiparty Computation (MPC)** to protect sensitive information of the users. MPC enables two or more servers to jointly compute the output of the fairness algorithm on users' private attributes in a distributed way, without revealing anything other than the final computed result with each other.
- **Differential Privacy (DP)** to protect the output of the fairness algorithms computed by MPC.

Contributions

- **First-of-its-kind** method for training group-fair ML models in FL under complete privacy guarantees.
- A privacy-preserving pre-processing technique using training sample reweighing to mitigate bias, **PrivFairFL-Pre**.
- A privacy-preserving post-processing technique that identifies classification thresholds to achieve fairness between groups, **PrivFairFL-Post**.

Methodology

We propose pre- and post-processing techniques to achieve group fairness in FL. To preserve privacy,

1. We design MPC protocols to collect aggregated statistics of labels and sensitive attributes across the federation of clients, preserving privacy of the sensitive data.
2. These MPC protocols run algorithms for achieving group fairness: reweighing and threshold optimization.
3. To simulate global DP, we add the appropriate noise within the MPC protocols themselves, to make the outputs of the fairness algorithms private.

PrivFairFL-Pre

- Pre-processing technique.
- Debiases the input dataset using reweighing algorithm.
- Each data point is assigned a corresponding sample weight during training, $\frac{1}{C(s,y)}$, where $C(s,y)$ is the total number of examples with this combination of sensitive attribute s and label y.
- MPC protocols compute the weights.

PrivFairFL-Post

- Post-processing technique.
- Finds alternative classification thresholds for each group of sensitive attribute s.
- Generates ROC-curves based on label y and s.
- MPC protocols compute ROC-curves.

Extensions to PrivFairFL

PrivFairFL can be

- Easily extended to a **cross-silo** setup.
- Extended to **multi-class classification** or multi-valued sensitive attributes using a one-vs-rest approach.
- Used in **dynamic scenarios** with client dropout and changes in data distributions, by re-weighing after specified sets of FL rounds.

Results

Task : Binary classification; Sensitive attribute: Gender

ADS Dataset. With 109 clients

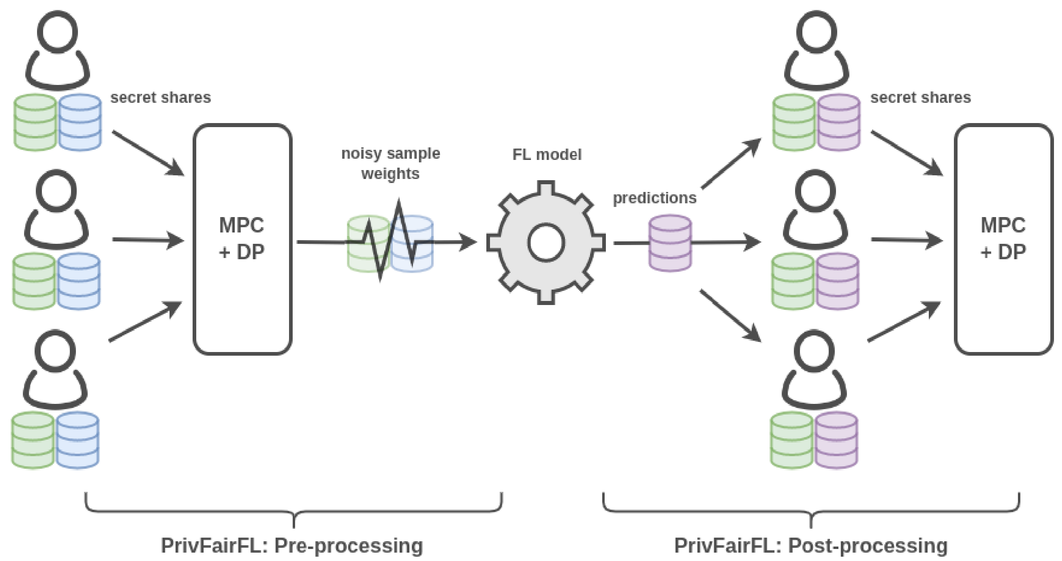
	Fairness	Privacy	Acc.	|1-DI|	ΔEOP	ΔEODD	ΔSP
CL	–	–	85.81%	1.214	0.070	0.037	0.004
FL	–	–	85.04%	1.654	0.089	0.050	0.018
FL-DP-SGD	–	DP-SGD	85.21%	6.606	0.070	0.043	0.023
FL-DP-SGD	Pre	Local DP	83.52%	0.260	0.036	0.033	0.032
PrivFairFL	Pre	MPC+DP	83.57%	0.122	0.018	0.027	0.036
PrivFairFL	Post	MPC+DP	85.17%	0.572	0.008	0.005	0.001

ML1M Dataset. With 75 clients

	Fairness	Privacy	Acc.	|1-DI|	ΔEOP	ΔEODD	ΔSP
CL	–	–	62.15%	0.138	0.091	0.141	0.094
FL	–	–	59.04%	0.096	0.081	0.096	0.091
FL-DP-SGD	–	DP-SGD	58.30%	0.027	0.026	0.027	0.052
FL-DP-SGD	Pre	Local DP	58.47%	0.045	0.042	0.052	0.061
PrivFairFL	Pre	MPC+DP	58.52%	0.045	0.042	0.051	0.063
PrivFairFL	Post	MPC+DP	58.46%	0.006	0.006	0.014	0.045

The results of accuracy and four fairness metrics for our techniques, evaluated against the centralized system (CL) and FL with DP-SGD. Better fairness is achieved when these metrics are closer to zero.

PrivFairFL-Pre benefits the highly imbalanced data in ADS, while **PrivFairFL-Post** successfully finds optimal thresholds for the almost balanced subset of data in ML-1M. Our techniques are independent of the model used and the statistical notions of fairness.



Acknowledgements Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Facebook Research Award for Privacy Enhancing Technologies, and the UW Azure Cloud Computing Credits for Research program.