Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning*

Jun Luo[1], Shandong Wu[1,2,3,4]
jul117@pitt.edu; wus3@upmc.edu

[1]Intelligent Systems Program, [2]Department of Radiology, [3]Department of Biomedical Informatics, [4]Department of Bioengineering
University of Pittsburgh

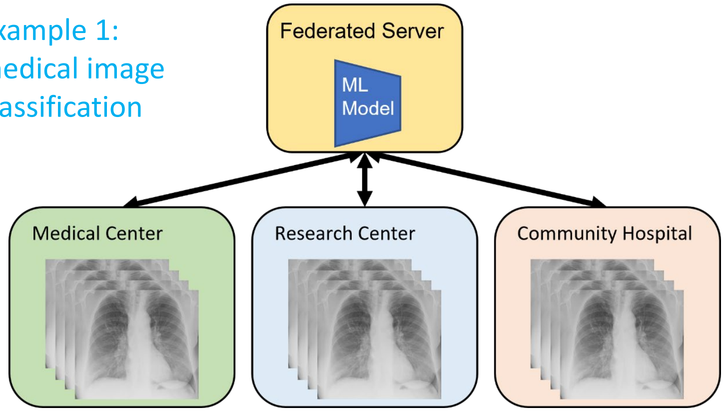
*An extended version of the paper (with the Appendix included) can be found at https://arxiv.org/abs/2110.08394. The code is publicly available at https://github.com/ljaiverson/pFL-APPLE.

Background

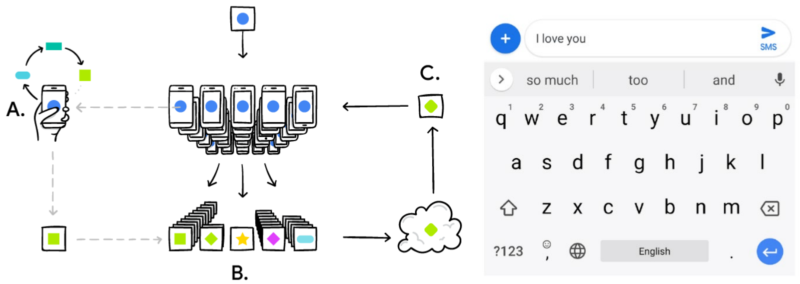
- Federated learning (FL) – privacy preserving machine learning
  - Pushes model to the clients that own privacy-sensitive data
  - Only model weights are shared while keeping the data decentralized

- Federated learning poses data heterogeneity challenge
  - Data heterogeneity – non-IID
  - Potential influence
    - slower convergence
    - inferior performance
    - Loss of clients' incentives to participate in the federation

Example 1: medical image classification



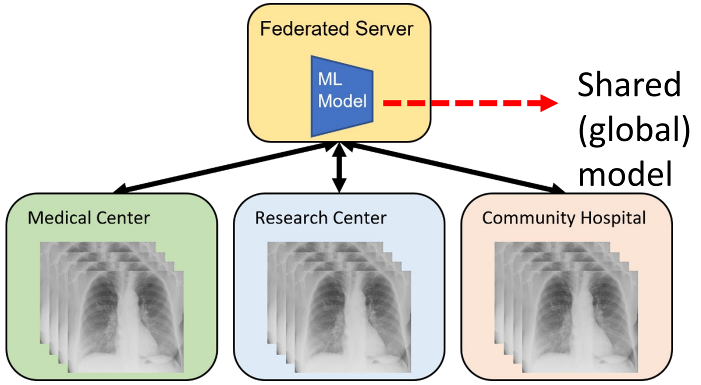
Example 2: smart phone keyboard next-word prediction

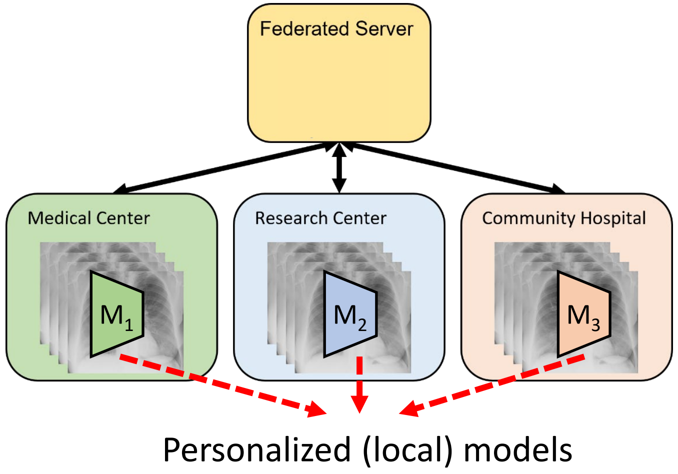


[Hard et al., 2019]

Background & Related Work

- FL algorithms that address the data heterogeneity fall into two categories
  - Generic FL algorithms ($\min_{w} f_G(w) = \min_{w} \sum_{i=1}^{N} p_i F_i(w)$)
    - Train a consensus global model that shared among all clients
      - FedAvg [McMahan et al. 2017]
      - FedProx [Li et al., 2020]
      - FedDyn [Acar et al., 2020]
      - Etc.
  - Personalized FL algorithms ($\min_{W} f_P(W) = \min_{w_i, i\in[N]} f_P(w_1, \dots, w_N) = \min_{w_i, i\in[N]} \sum_{i=1}^{N} p_i F_i(w_i)$)
    - Train multiple models (e.g. one model for each client)
      - Combined with multi-task learning / meta-learning [Smith et al., 2017, Fallah et al., 2020]
      - APFL [Deng et al., 2020]
      - FedFOMO [Zhang et al., 2021]
      - FedAMP [Huang et al., 2021]
      - Etc.

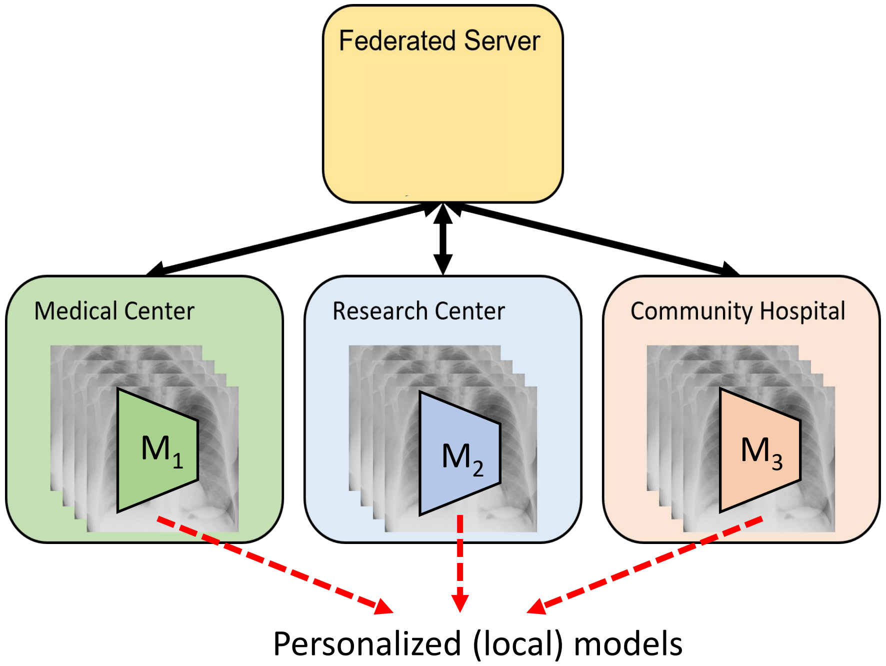




Motivation

- Investigate a personalized FL framework that adaptively learns how much each client can benefit from other clients' models.

- Flexibly control the focus of training between global and local objectives.



Method

- ***A**da**p**tive **P**ersonalized Cross-Si**l**o F**e**derated Learning* (**APPLE**)
- The model of a client
  - Personalized model $w_i^{(p)}$: used to do inference on client i
  - Core model $w_i^{(c)}$: a constructing part of personalized model on client i
  - $w_i^{(p)} = \sum_{j=1}^{N} p_{i,j} w_j^{(c)}$
  - Directed relationship (DR) vector p_i: learnable weights (coefficients for core models) on client i, always kept locally

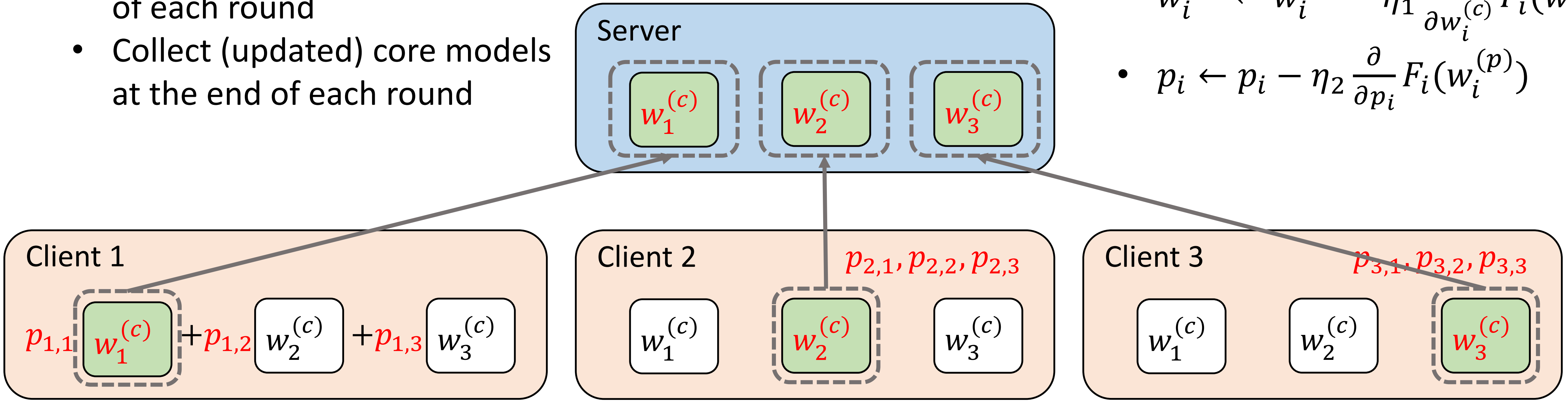
Client 1

$$w_1^{(p)} = p_{1,1} \boxed{w_1^{(c)}} + p_{1,2} \boxed{w_2^{(c)}} + p_{1,3} \boxed{w_3^{(c)}}$$

Method

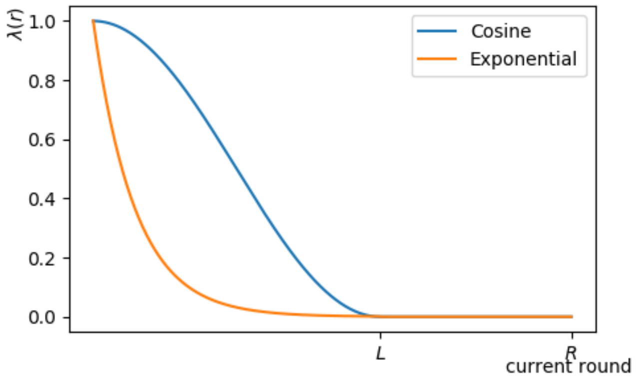
- Server
  - Broadcast core models to each client at the beginning of each round
  - Collect (updated) core models at the end of each round

- Local training
  - Clients' own core models and DR vectors are updated
  - $w_i^{(c)} \leftarrow w_i^{(c)} - \eta_1 \frac{\partial}{\partial w_i^{(c)}} F_i(w_i^{(p)})$
  - $p_i \leftarrow p_i - \eta_2 \frac{\partial}{\partial p_i} F_i(w_i^{(p)})$



Method

- Proximal Directed Relationships
  - Since downloaded core models are not trained from local empirical risk, training might be drawn to resembling individual learning (DR matrix drawn to identity matrix)
  - Penalize DR vector by a proximal term
  - $F_i\left(w_i^{(p)}\right) = \frac{1}{n_i}\sum_{\xi\in D_i^{tr}} \mathcal{L}\left(w_i^{(p)};\xi\right) + \lambda(r)\frac{\mu}{2}\|p_i - p_0\|_2^2$
  - Prox-center $p_0 = [\frac{n_1}{n}, \dots, \frac{n_N}{n}]$
  - Loss scheduler $\lambda(r) \in [0,1]$: a decreasing function w.r.t. current round, controls the focus of training; μ: the peak value of the proximal term coefficient
  - Proximal term coefficient: ∞ → FedAvg; large → facilitate learning global high-level feature; small → concentrate on local empirical risk, learning the personalization



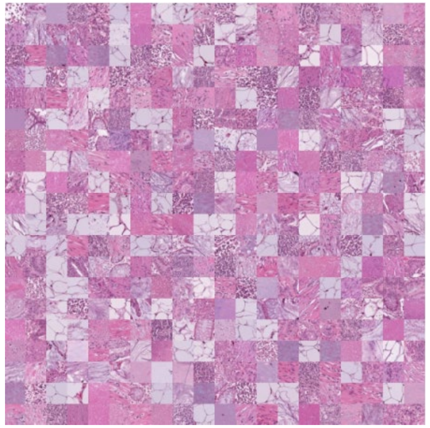
Experiments

- Datasets

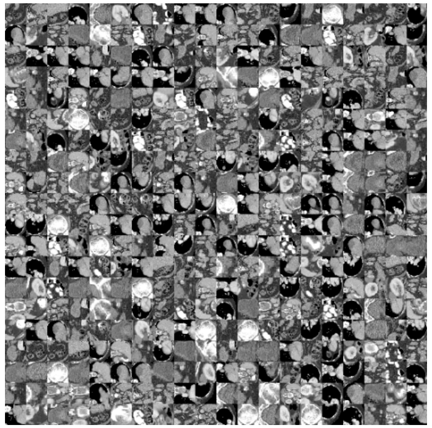
  - Two benchmark datasets
    - MNIST
    - CIFAR10

  - Two medical imaging datasets from MedMNIST collection [Yang et al., 2021]
    - OrganMNIST (axial) (11-class liver tumor images)
    - PathMNIST (9-class colorectal cancer images)

PathMNIST



OraganMNIST (axial)



Experiments

- Two non-IID settings

  - Pathological non-IID
    - Randomly select 2 classes for each client
    - In each class, assign a random number of images

  - Practical non-IID
    - Randomly partition each class of the dataset into 12 shards (10 x 1%, 1 x 10%, 1 x 80%)
    - Randomly assign one shard from each class to each client
    - Allows each client to have images from all classes, with more images from some classes while less from others
    - A simulation that is closer to real-world medical applications

Experiments

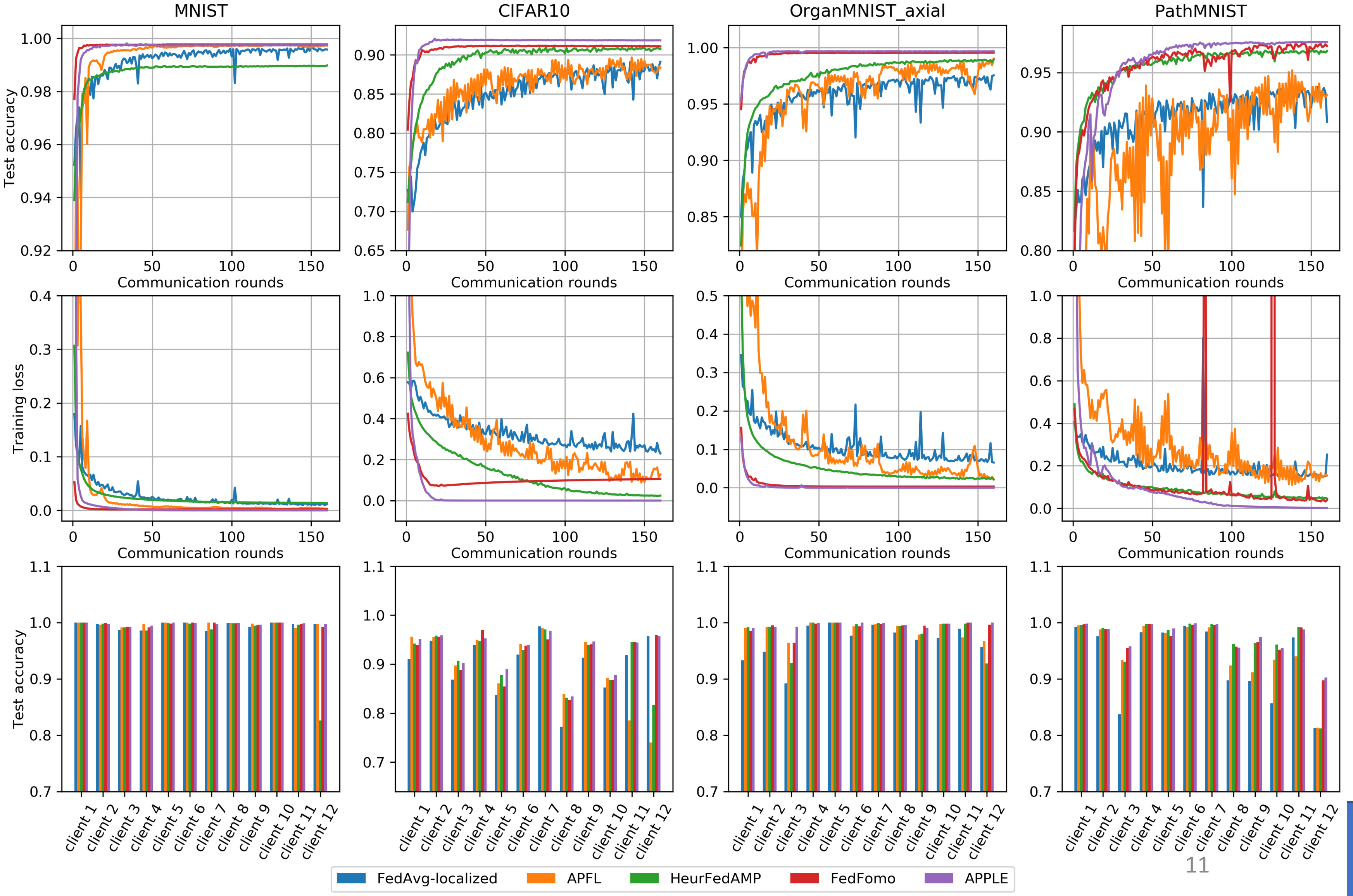
- Evaluation metrics
  - Numerical metrics: two types of test accuracies
    - ***B**est **M**ean **C**lient **T**est **A**ccuracy (**BMCTA**)*
      - Mean over all clients
      - Best over all rounds
  - Plots
    - Training loss curve
    - Test accuracy curve
    - Client wise test accuracies bar chart

- Compared baselines
  - Separate training
  - FedAvg (McMahan et al., 2017)
  - FedAvg-local
  - FedAvg-FT, FedProx-FT (Wang et al., 2019)
  - APFL (Deng et al., 2020)
  - HeurFedAMP (Huang et al., 2021)
  - FedFomo (Zhang et al., 2021)

Results

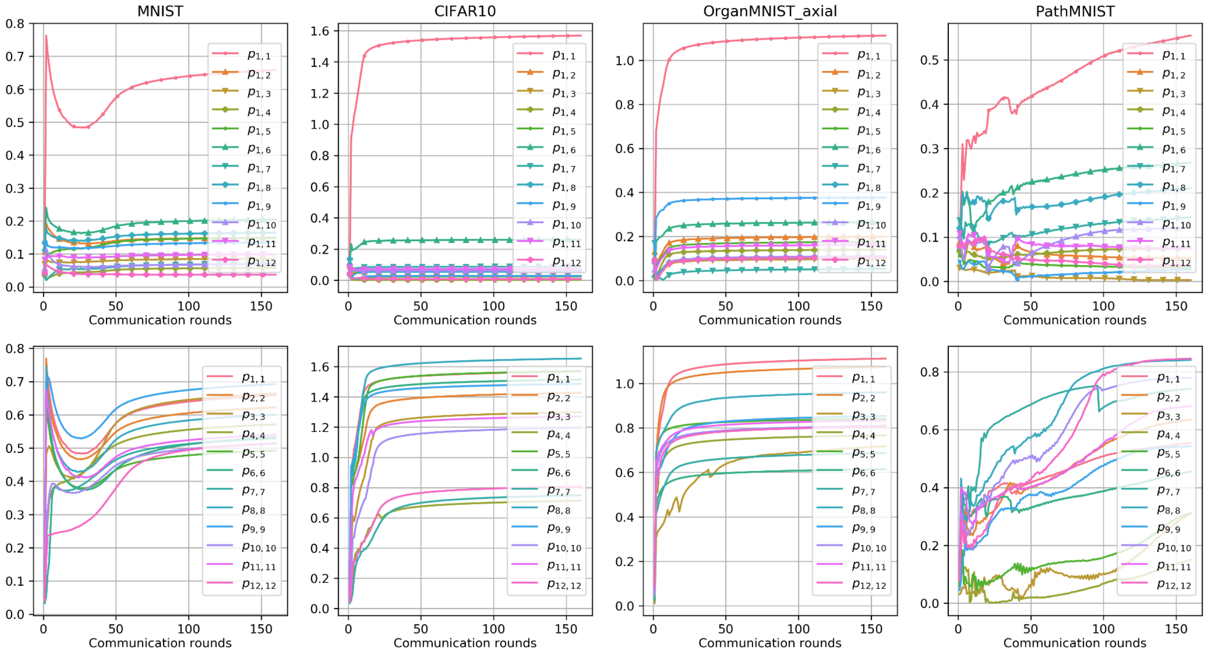
- Pathological non-IID

	Pathological non-IID			
	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
Separate	97.34	74.96	93.14	87.09
FedAvg	95.71	51.44	59.43	56.61
FedAvg-local	99.52	90.10	96.76	93.21
FedAvg-FT	99.43	90.49	97.03	92.31
FedProx-FT	99.43	90.49	97.03	92.38
APFL	99.75	89.30	98.72	94.98
HeurFedAMP	98.13	91.10	98.39	96.55
FedFomo	99.71	91.96	99.31	97.24
APPLE, $\mu = 0$	99.73	92.22	**99.66**	96.78
APPLE, $\mu \neq 0$	**99.77**	**92.68**	99.61	**97.51**

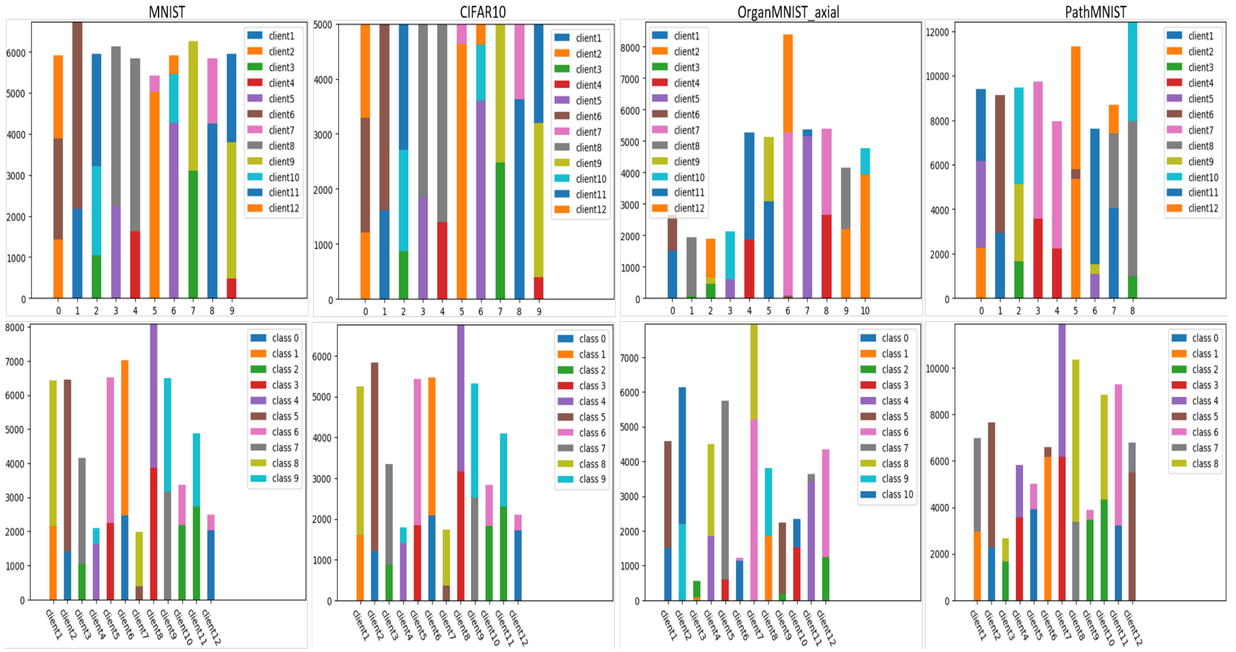


Results

- Visualization of Directed Relationships (Pathological non-IID)



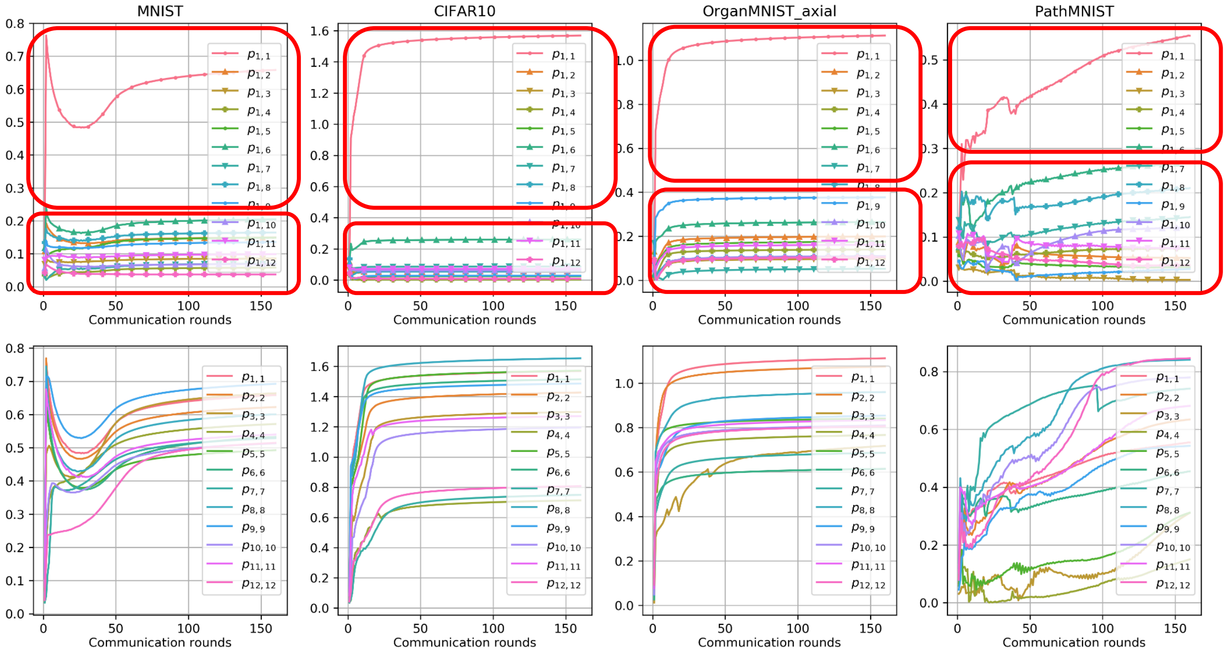
Visualization of DR



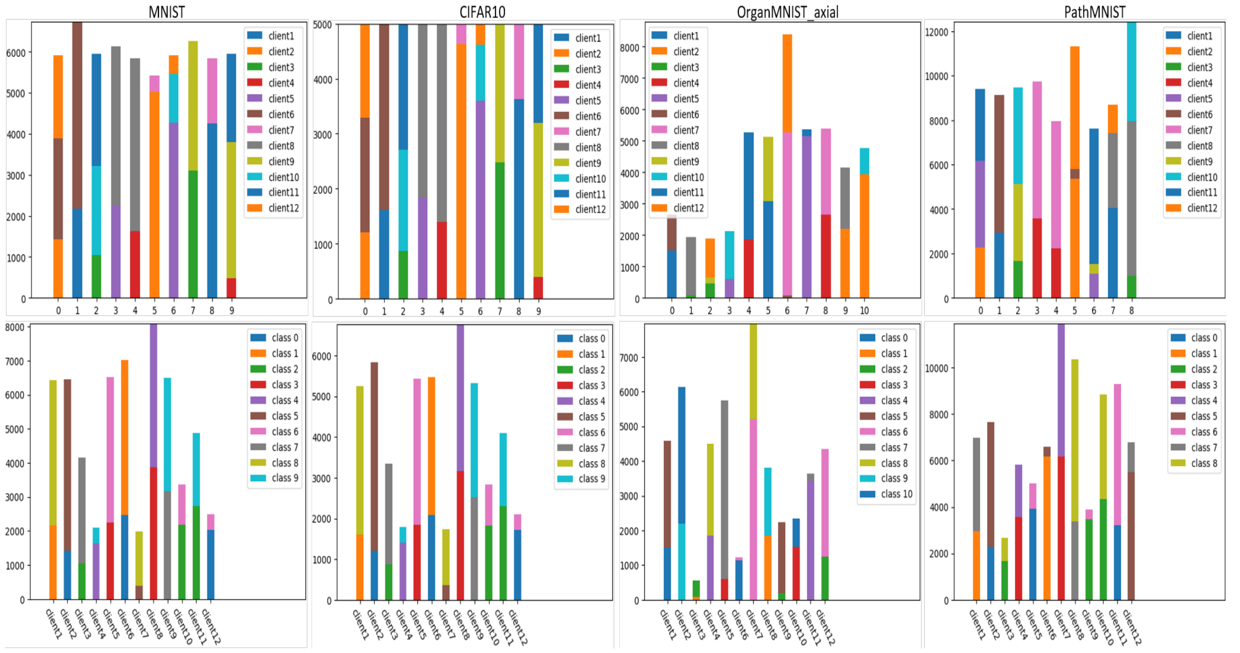
Data distribution

Results

- Visualization of Directed Relationships (Pathological non-IID)



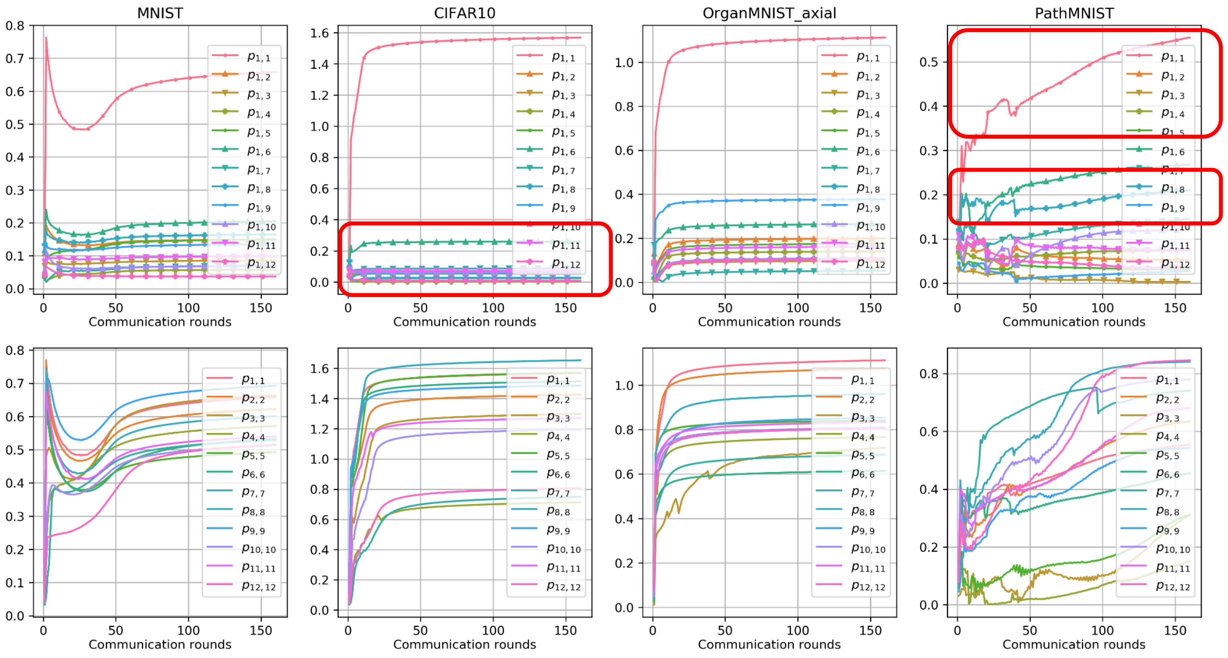
Visualization of DR



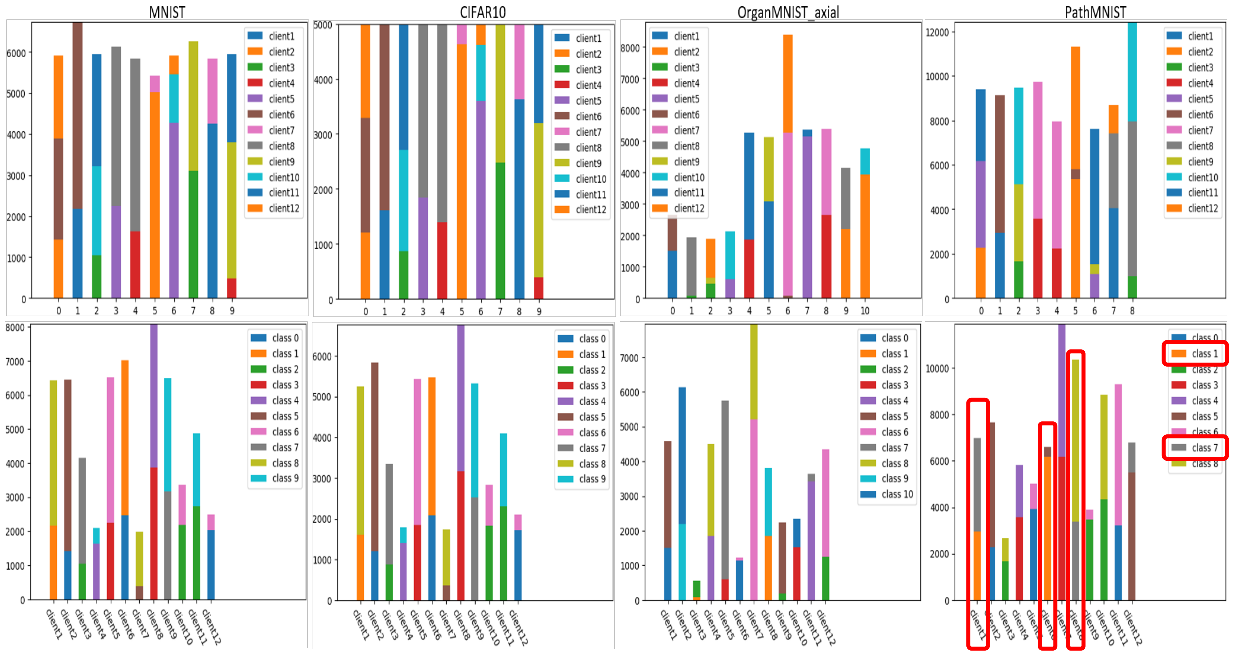
Data distribution

Results

- Visualization of Directed Relationships (Pathological non-IID)



Visualization of DR

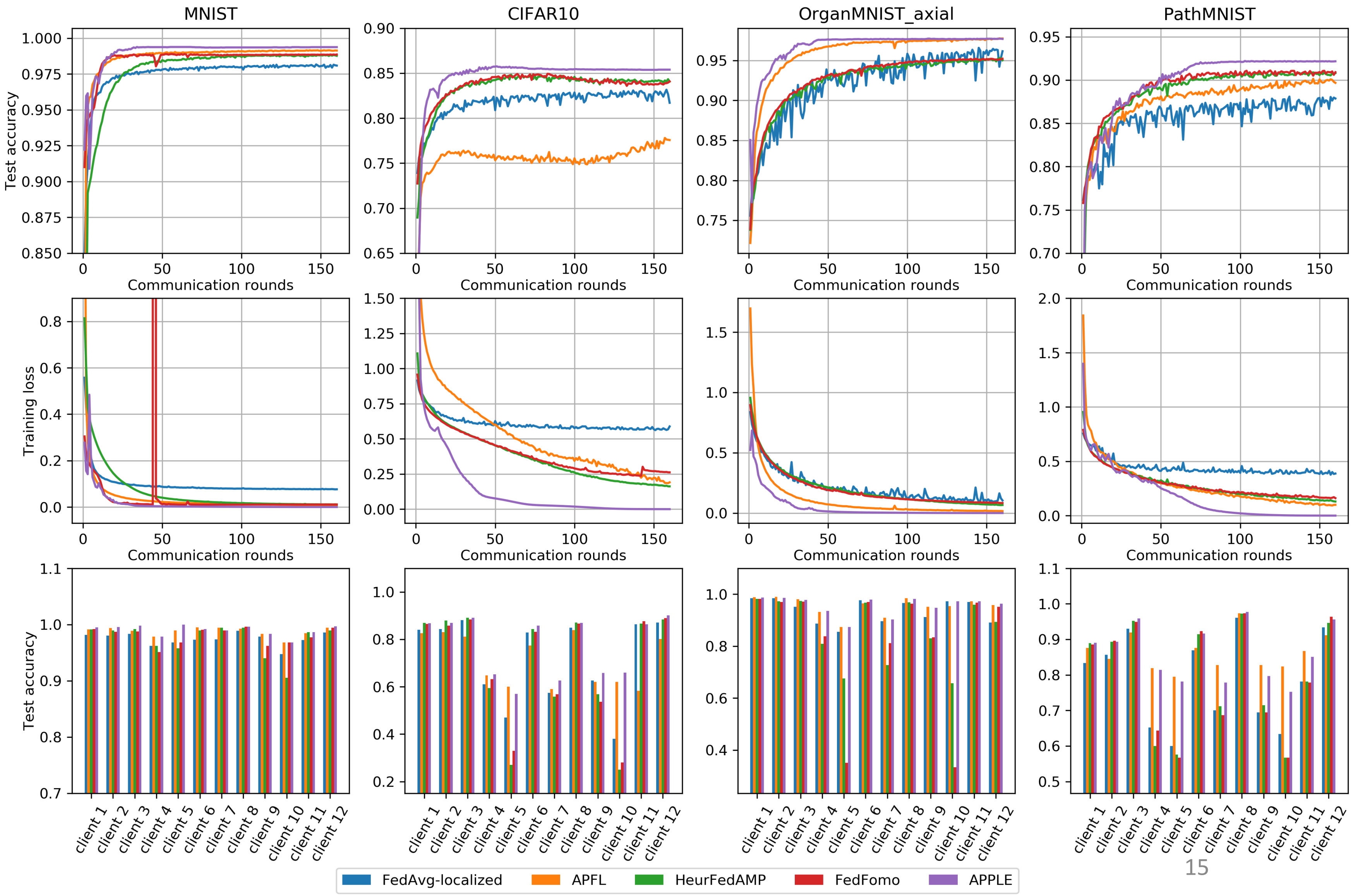


Data distribution

Results

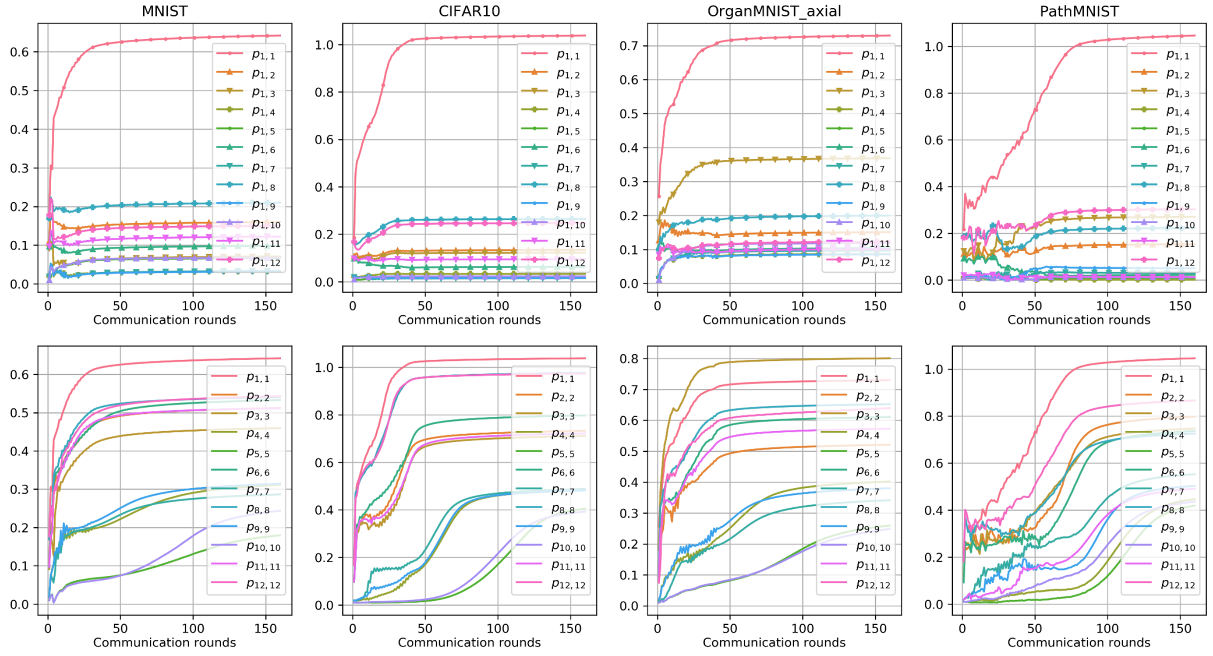
- Practical non-IID

	Practical non-IID			
	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
Separate	78.20	63.06	65.21	61.36
FedAvg	94.00	34.32	86.56	53.83
FedAvg-local	97.47	71.99	93.75	78.70
FedAvg-FT	97.66	72.08	94.13	78.69
FedProx-FT	97.66	72.08	94.13	78.69
APFL	98.80	71.19	95.53	86.35
HeurFedAMP	97.45	69.54	86.82	79.33
FedFomo	98.05	70.15	82.86	79.39
APPLE, $\mu = 0$	**99.00**	75.62	**95.70**	84.22
APPLE, $\mu \neq 0$	98.97	**77.41**	95.62	**86.39**

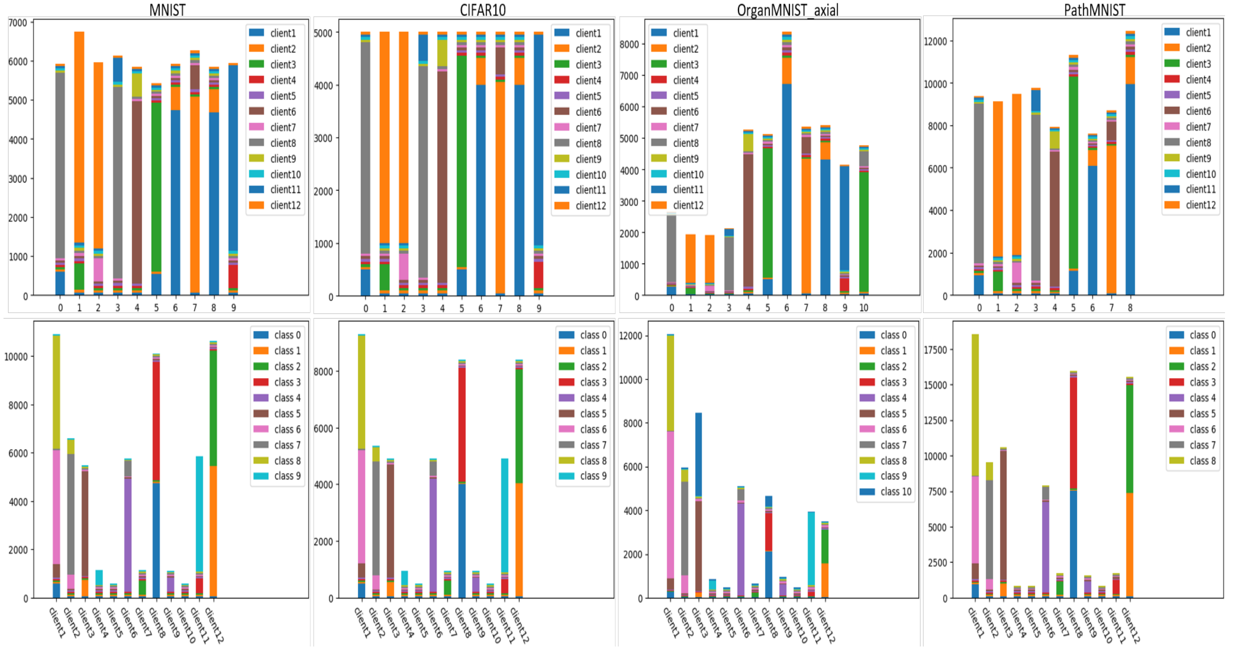


Results

- Visualization of Directed Relationships (Practical non-IID)



Visualization of DR



Data distribution

Results

- Under limited bandwidth
  - Restrict the number of models (M) a client can download per round with $1 \leq M \leq N - 1$
  - Client j's core model will be downloaded to client i with a probability positively correlated to $|p_{i,j}|$.
  - We limited $M = 1, 2, 5, 7, 11$ ($N = 12$), and compared our method against FedFomo.

		Pathological non-IID				Practical non-IID			
		MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
$M = 11$	FedFomo	99.71	91.96	99.31	97.24	98.05	70.15	82.86	79.39
	APPLE	99.73	92.22	99.66	96.78	99.00	75.62	95.70	84.22
$M = 7$	FedFomo	99.71	91.95	99.31	97.33	97.65	70.24	80.88	80.19
	APPLE	99.73	92.17	99.53	97.15	98.70	76.14	94.21	84.07
$M = 5$	FedFomo	99.71	91.94	99.31	97.40	97.47	70.44	82.83	79.62
	APPLE	99.72	92.28	99.48	97.17	98.45	75.63	94.49	85.46
$M = 2$	FedFomo	99.71	91.98	99.31	97.25	96.51	69.87	79.53	79.26
	APPLE	99.70	92.41	99.47	97.11	98.29	74.84	92.29	84.64
$M = 1$	FedFomo	99.71	91.95	99.31	97.15	91.54	69.93	78.37	75.17
	APPLE	99.66	92.31	99.59	96.29	98.52	73.03	93.55	83.35

Conclusion

- We proposed a personalized approach for cross-silo federated learning that
  - Allows clients to adaptively learn how much they can benefit from other clients' models
  - Flexibly controls the training focus between learning from global collaboration and local objective
- Our work does have some limitations, making it suitable only for a small federation (e.g. cross-silo FL)
  - Downloading the other clients' core models increases the communication overhead.
  - Training the DR vector – the coefficients for the core models – increases the local computing overhead.
- In the future, we will investigate personalized FL leveraging information or knowledge of datasets of the clients.

Acknowledgements:

Intelligent **C**omputing for **C**linical **I**maging (**ICCI**) Lab, University of Pittsburgh

- NIH/NCI #1R01CA218405, an NSF grant (CICI:SIVD:2115082), the grant 1R01EB032896 as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program, a Pitt Momentum Funds scaling award (Pittsburgh Center for AI Innovation in Medical Imaging), and an Amazon AWS Machine Learning Research Award.

- Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF grant number ACI-1548562, NSF award number ACI-1928147, the Bridges-2 system at the Pittsburgh Supercomputing Center (PSC), supported by NSF award #ACI-1928147.

Thank you!

- Check out the full version of the paper (with the Appendix included) at https://arxiv.org/abs/2110.08394.
- The code is publicly available at https://github.com/ljaiverson/pFL-APPLE.

Jun Luo
jul117@pitt.edu