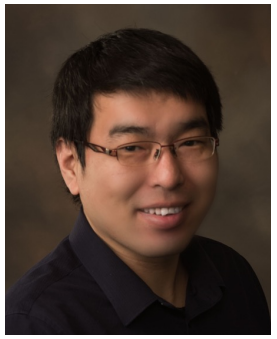
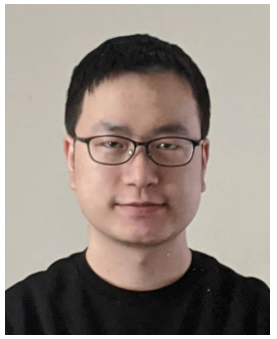
Optimizing the Collaboration Structure in Cross-Silo Federated Learning





Wenxuan Bao **Haohan Wang** **Jun Wu** **Jingrui He**

University of Illinois Urbana-Champaign

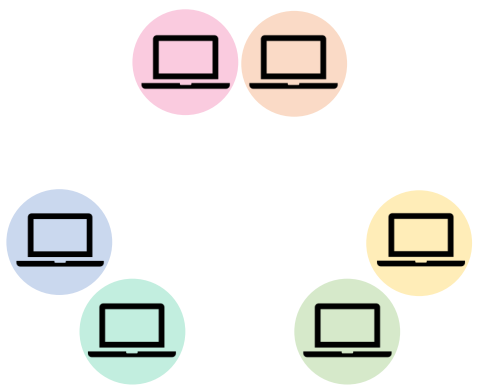
{wbao4,haohanw,junwu3,jingrui}@illinois.edu

baowenxuan.github.io, haohanwang.github.io,

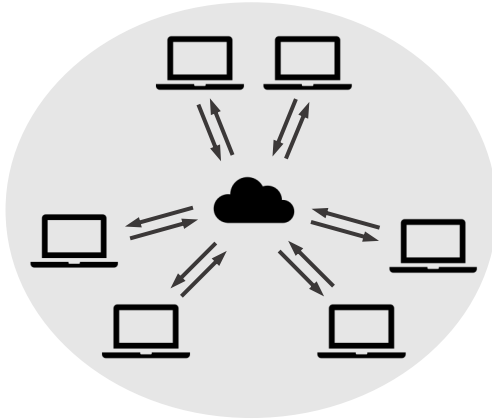
publish.illinois.edu/junwu3, hejingrui.org

Negative Transfer in Federated Learning

- *Federated Learning* (FL): Multiple clients collaborate to train machine learning models without sharing their raw data.
- *Global FL*: All clients share one global model.
  - Example: FedAvg, FedProx, etc.
- **Negative Transfer**: When clients have non-IID data, $\epsilon_i(h_G) > \epsilon_i(h_i)$ for some client i,
  - the global model h_G can be even worse than the local model h_i !

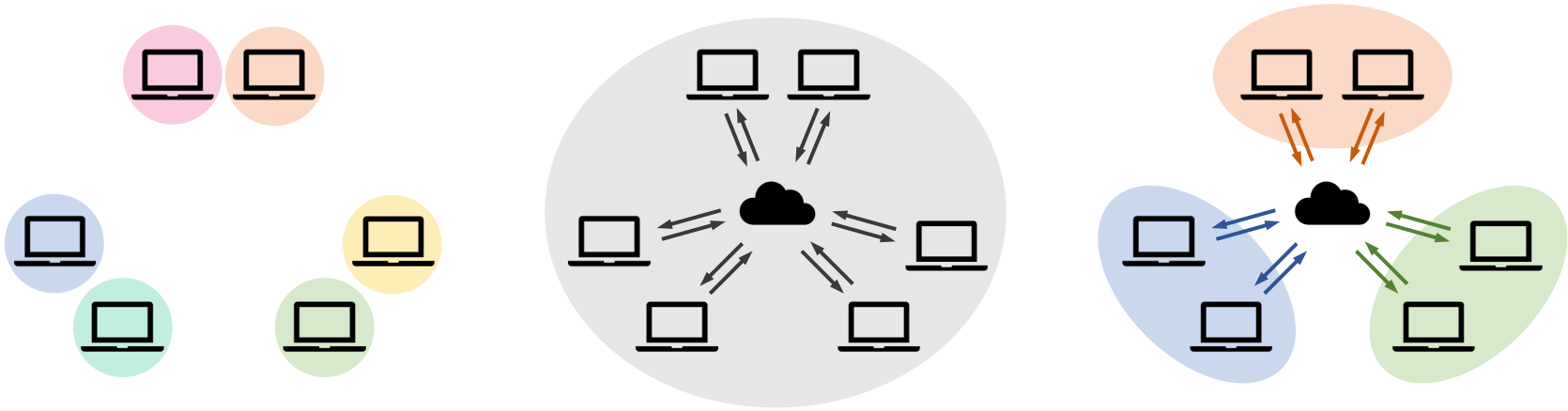


Local Training



Global FL

Clustered FL and Collaboration Structure



- *Clustered FL* groups clients into coalitions based on distributions; each client only shares model with clients in the same coalition.
- **Question**: What is the *optimal collaboration structure,* i.e., which clients should train shared model?

Our contributions

- **Theory**: We analyze how clustered FL performance is affected by two key factors: *distribution distance* and *data quantity*.
- **Algorithm**: We propose FedCollab to solve for the best collaboration structure.
- **Extensive experiments**: We test FedCollab under label shift, feature shift and concept shift with various models / datasets.

Theory: Error Bound for Clustered FL

Theorem 3.3. (informal) Let $\hat{h}_{\boldsymbol{\alpha}_i} = \arg\min_{h\in\mathcal{H}} \sum_{j=1}^{N} \alpha_{ij}\hat{\epsilon}_j(h)$ where $\sum_{j=1}^{N} \alpha_{ij} = 1$. With probability at least $1-2\delta$,

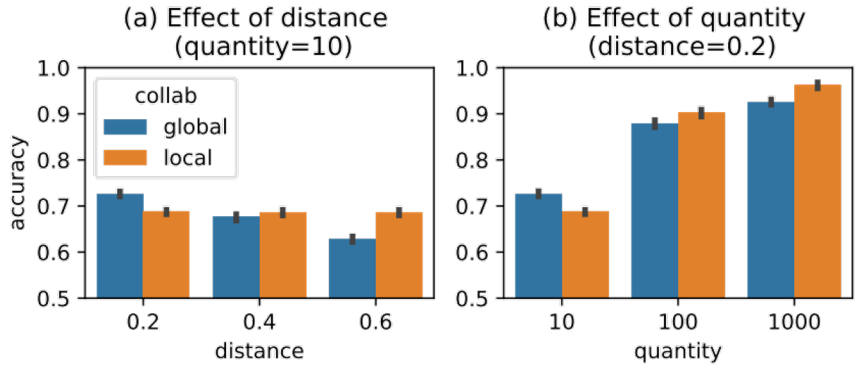
$$\epsilon_i(\hat{h}_{\boldsymbol{\alpha}_i}) \leq \epsilon_i(h_i^*) + 2\phi_{|\mathcal{H}|}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, m, \delta) + 2\sum_{j\neq i} \alpha_{ij} D(\mathcal{D}_i, \mathcal{D}_j)$$

where $\phi_{|\mathcal{H}|}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, m, \delta) = \sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_{ij}^2}{\beta_j}\right)\left(\frac{2d\log(2m+2)+\log(4/\delta)}{m}\right)}$, and $D(\mathcal{D}_i, \mathcal{D}_j) = \max_{h\in\mathcal{H}} |\epsilon_i(h) - \epsilon_j(h)|$

- The error upper bound of client i is controlled by
  - Collaboration structure α_{ij}
  - Pairwise distribution distances $D(\mathcal{D}_i, \mathcal{D}_j)$
  - Data quantities β_j

Theory: Optimal Collaboration Structure

- *The optimal collaboration structure (that minimizes the error bound) depends on distribution distances and data quantities!*

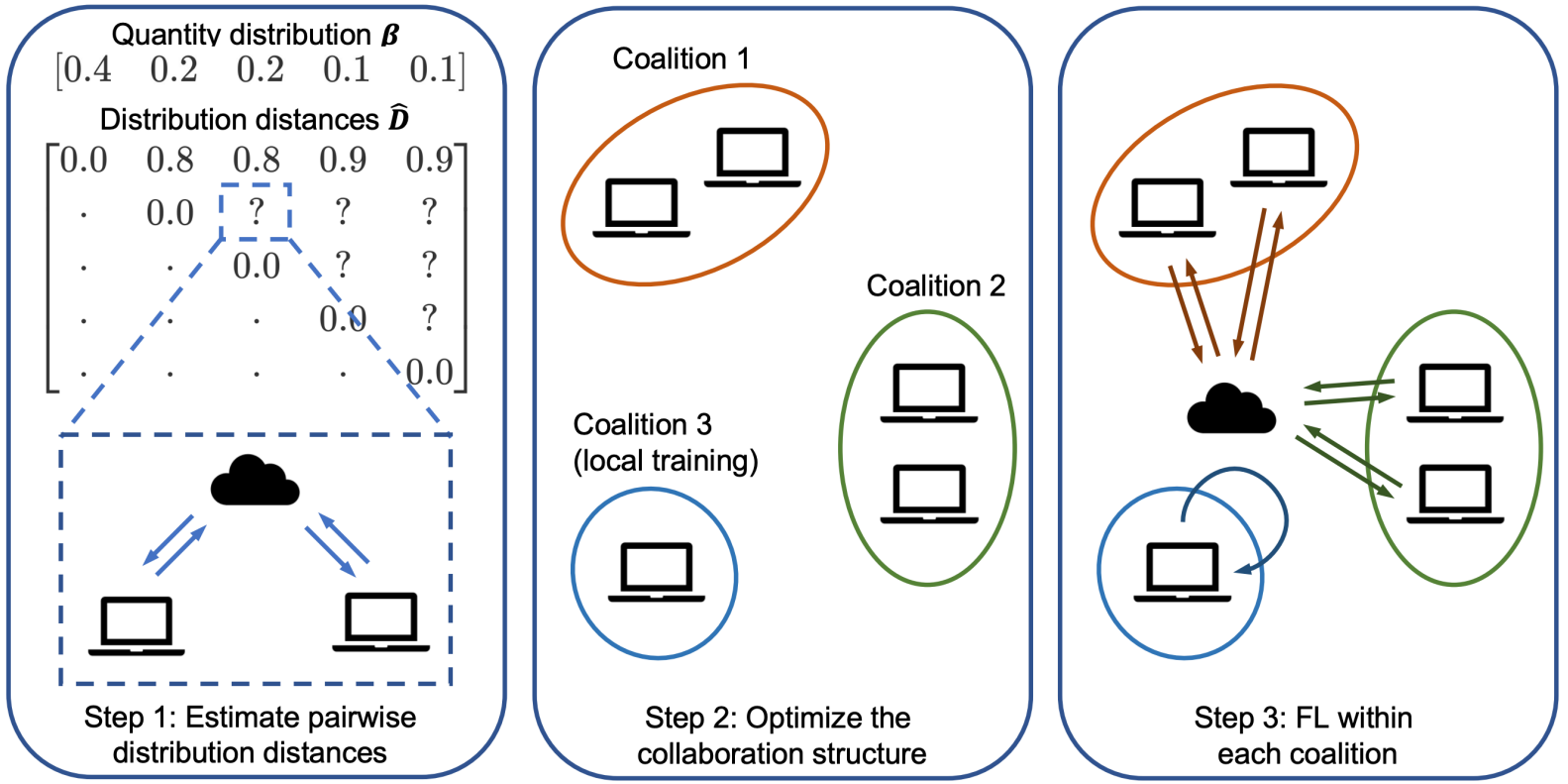


- Clients prefer collaborators with smaller distribution distances.
  - Collaboration is only beneficial when distribution distance is small enough.
- Clients with more data are *pickier* in the choice of collaborators.
  - Collaboration is only beneficial when quantity is small.

Algorithm: FedCollab

- FedCollab minimizes an empirical estimation of the error bound.

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{\beta}, m, \hat{\boldsymbol{D}}) = \sum_{i=1}^{N} \left(\frac{C}{\sqrt{m}} \sqrt{\sum_{j=1}^{N} \frac{\alpha_{ij}^2}{\beta_j}} + \sum_{j=1}^{N} \alpha_{ij} \hat{D}_{ij} \right)$$



Experiments: Alleviating Negative Transfer

- FedCollab alleviates negative transfer for both global FL and personalized FL.

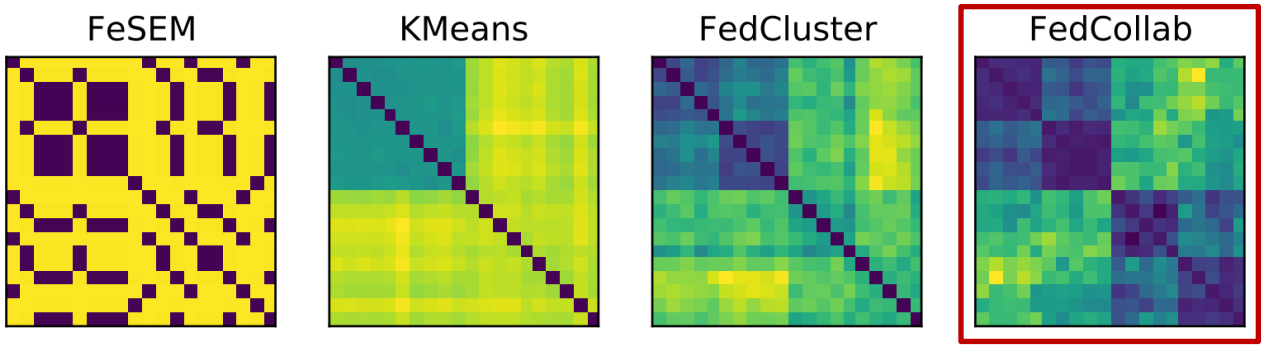
Method	Label Shift (FashionMNIST)			Feature Shift (CIFAR-10)			Concept Shift (CIFAR-100)		
	Acc ↑	IPR ↑	RSD ↓	Acc ↑	IPR ↑	RSD ↓	Acc ↑	IPR ↑	RSD ↓
Local Train	86.05 (0.28)	-	-	38.65 (0.44)	-	-	29.82 (0.56)	-	-
FedAvg	46.64 (0.12)	46.00 (2.24)	41.03 (0.24)	44.31 (0.98)	86.00 (4.18)	4.62 (0.58)	26.62 (0.12)	50.00 (0.00)	11.54 (0.45)
+FEDCOLLAB	92.45 (0.07)	100.00 (0.00)	5.99 (0.41)	52.61 (0.60)	100.00 (0.00)	3.30 (0.63)	40.94 (0.22)	100.00 (0.00)	2.78 (0.30)
FedProx	46.70 (0.08)	45.00 (5.00)	41.09 (0.29)	44.45 (0.58)	87.00 (4.47)	4.74 (0.56)	26.78 (0.14)	50.00 (0.00)	11.66 (0.36)
+FEDCOLLAB	92.39 (0.15)	100.00 (0.00)	6.02 (0.37)	52.73 (0.64)	100.00 (0.00)	3.16 (0.61)	40.99 (0.17)	100.00 (0.00)	2.79 (0.34)
FedNova	75.92 (1.14)	45.00 (3.54)	12.38 (1.25)	46.98 (0.57)	99.00 (2.24)	3.42 (0.22)	26.46 (0.13)	50.00 (0.00)	10.57 (0.32)
+FEDCOLLAB	92.47 (0.13)	100.00 (0.00)	5.97 (0.39)	52.72 (0.57)	100.00 (0.00)	3.18 (0.63)	40.92 (0.36)	100.00 (0.00)	2.75 (0.43)
Finetune	67.32 (3.17)	48.00 (2.74)	22.97 (2.82)	44.17 (0.99)	82.00 (2.74)	5.14 (0.32)	33.30 (4.79)	50.00 (0.00)	13.95 (0.57)
+FEDCOLLAB	92.57 (0.15)	99.00 (2.24)	6.07 (0.30)	51.53 (0.61)	100.00 (0.00)	2.92 (0.46)	40.94 (2.36)	100.00 (0.00)	2.54 (0.30)
Per-FedAvg	51.13 (4.10)	49.00 (2.24)	37.35 (4.15)	43.78 (0.69)	83.00 (9.08)	4.74 (0.65)	27.39 (0.24)	50.00 (0.00)	12.24 (0.46)
+FEDCOLLAB	92.16 (0.25)	97.00 (6.71)	6.00 (0.25)	52.64 (0.45)	100.00 (0.00)	3.03 (0.30)	41.04 (0.26)	100.00 (0.00)	2.85 (0.49)
pFedMe	55.31 (3.45)	47.00 (4.47)	33.71 (3.11)	39.74 (0.85)	60.00 (12.25)	4.81 (0.74)	27.04 (0.39)	48.00 (2.74)	10.39 (0.47)
+FEDCOLLAB	92.18 (0.43)	99.00 (2.24)	6.40 (0.81)	47.20 (1.29)	97.00 (2.74)	3.02 (0.30)	37.47 (0.31)	100.00 (0.00)	3.04 (0.23)
Ditto	68.73 (1.40)	48.00 (2.74)	20.29 (2.06)	47.04 (0.30)	97.00 (2.74)	3.85 (0.35)	32.50 (0.40)	50.00 (0.00)	12.22 (0.36)
+FEDCOLLAB	92.55 (0.08)	99.00 (2.24)	6.11 (0.30)	50.97 (0.75)	99.00 (2.24)	3.38 (1.55)	40.33 (0.33)	100.00 (0.00)	2.16 (0.30)

IPR: % of clients with accuracy gains, i.e., FL model is better than local model
RSD: standard deviation clients' accuracy gains

Experiments: Comparison

- FedCollab outperforms other clustered FL algorithms because it utilizes quantity information and provides high-quality estimation of distribution distances.

Method	Label Shift (FashionMNIST)			Feature Shift (CIFAR-10)			Concept Shift (CIFAR-100)		
	Acc ↑	IPR ↑	RSD ↓	Acc ↑	IPR ↑	RSD ↓	Acc ↑	IPR ↑	RSD ↓
IFCA	91.49 (0.61)	95.00 (5.00)	5.62 (0.54)	49.78 (1.01)	100.00 (0.00)	3.13 (0.52)	30.74 (4.46)	60.00 (22.36)	11.28 (5.04)
FedCluster	92.07 (0.47)	95.00 (7.07)	6.14 (0.49)	44.86 (1.90)	79.00 (17.10)	5.64 (1.81)	29.23 (2.18)	62.00 (12.55)	9.55 (0.69)
FeSEM	56.79 (6.71)	45.00 (11.18)	36.12 (2.08)	42.73 (0.37)	82.00 (5.70)	4.10 (0.62)	31.92 (3.12)	72.00 (12.55)	9.81 (1.77)
KMeans	69.30 (0.81)	72.00 (2.74)	35.87 (1.22)	48.61 (1.15)	96.00 (4.18)	4.54 (0.74)	34.24 (3.01)	85.00 (13.69)	6.47 (3.06)
FEDCOLLAB	92.45 (0.07)	100.00 (0.00)	5.99 (0.41)	52.61 (0.60)	100.00 (0.00)	3.30 (0.63)	40.94 (0.22)	100.00 (0.00)	2.78 (0.30)



Summary

- **Theory**: We analyze how clustered FL performance is affected by two key factors: *distribution distance* and *data quantity*.
- **Algorithm**: We propose FedCollab to solve for the best collaboration structure.
- **Extensive experiments**: We test FedCollab under label shift, feature shift and concept shift with various models / datasets.