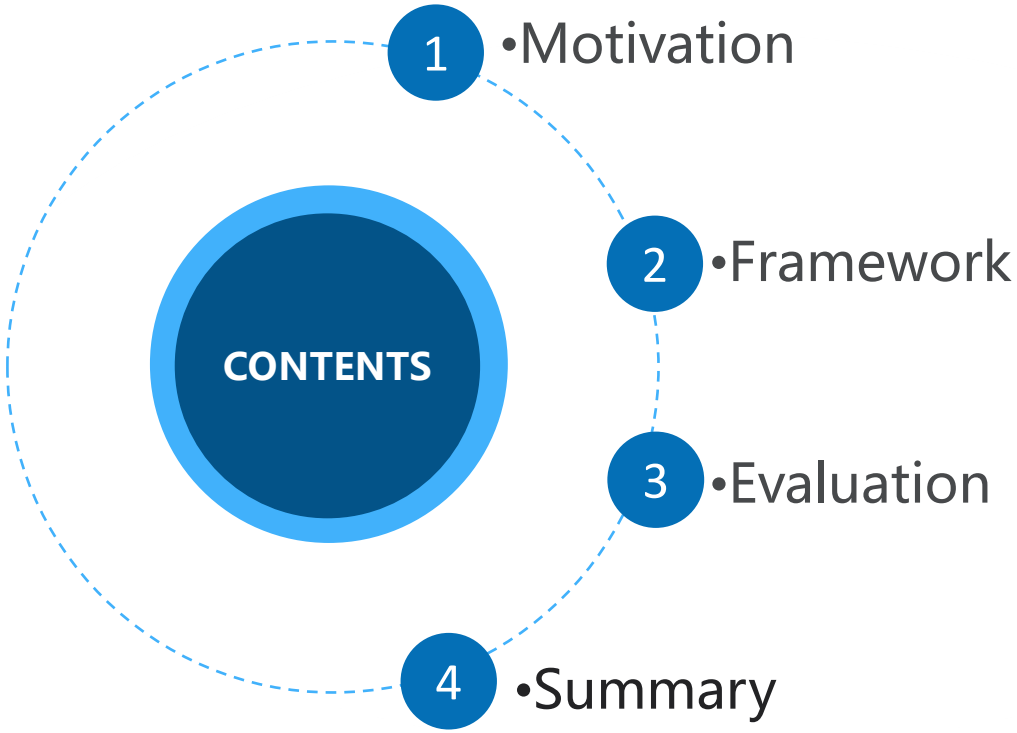
Private data quality and contribution degree assessment in decentralized network learning

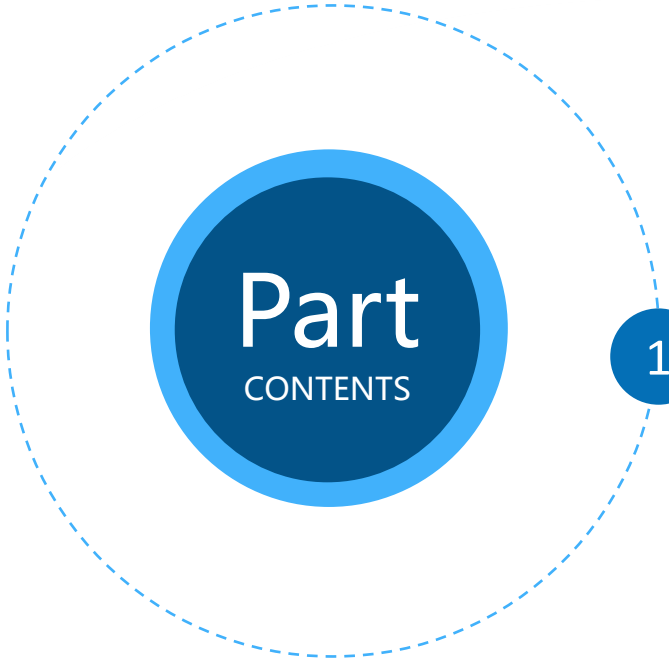
Course project presentation

BA21221024 兰牧沆
SA20216006 郭敏
SA21011249 姚志伟

Contents



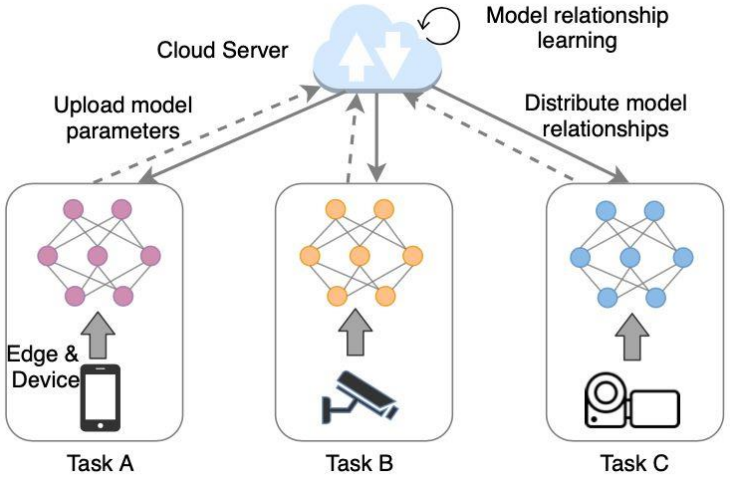
Contents

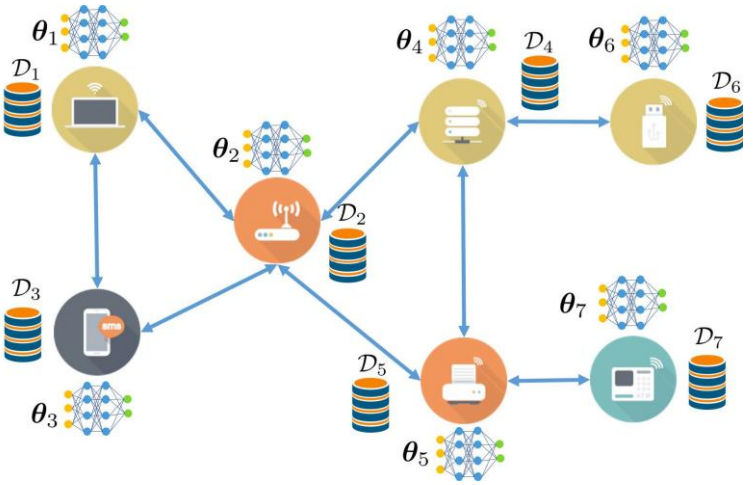


Motivation-Background

1. Decentralized Learning

- Privacy risk: share model, not local data
- Computation: from server to clients
- Robustness: avoid reliance on parameter servers

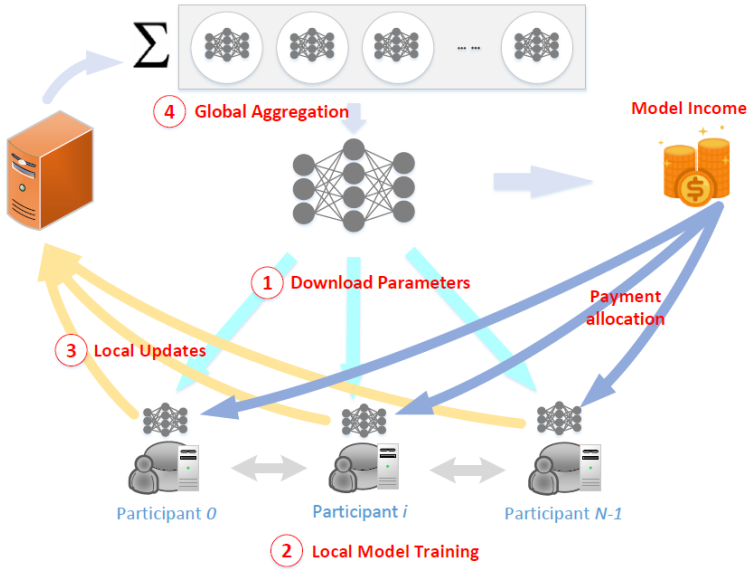




Motivation-Background

2. Incentive mechanism

- The performance will be deteriorated without sufficient training data and other resources in the learning process.
- Crucial to inspire more participants to contribute their valuable resources with some payments for federated learning



R. Zeng, C. Zeng, X. Wang, B. Li, and X. Chu, "A Comprehensive Survey of Incentive Mechanism for Federated Learning," *arXiv:2106.15406 [cs]*, Jun. 2021. [Online]. Available: http://arxiv.org/abs/2106.15406

Motivation-Background

3. Privacy matters

"Chinese consumers are often willing to authorize the use of certain personal data in exchange for more convenient services."

- On March 26, at the annual session of China Development Forum (CDF) 2018, Robin Li, Baidu

Is privacy leakage an evaluation indicator of contribution in decentralized learning?

Motivation-Related work

Representative solutions

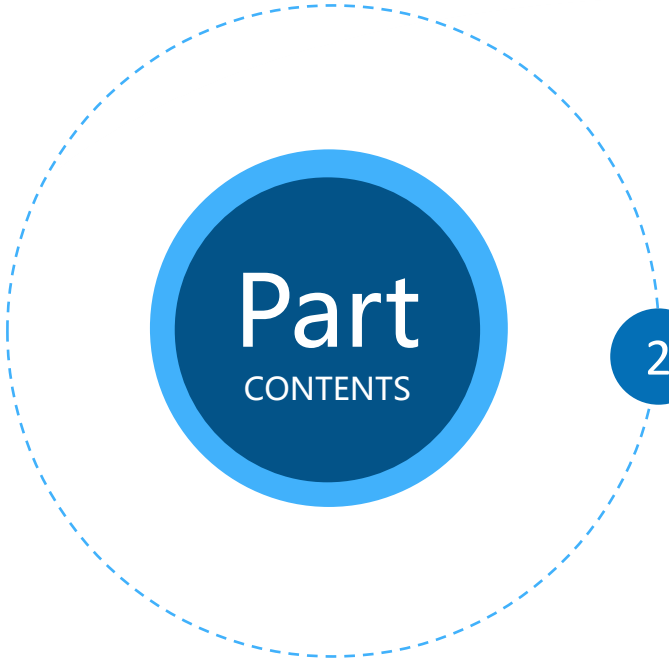
- Google[1]: once a model is trained, it is evaluated. If bias in device participation or other issues lead to an inferior model, it will be detected.
- Sharply value[2]: accumulated gradient combination among clients
- Game theory[3]: a monetary gain by utilizing the aggregated model. Then, the agents deduct the communication cost c from what they earn from MAE.

$$u(S) = f\left(\frac{1}{\mathbb{E}_{\boldsymbol{x}}[\mathcal{L}(\theta^S)]}\right) - c(S)$$

None of the existing solutions considers privacy issue!

[1] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *Proceedings of Machine Learning and Systems*, A. Talwalkar, V. Smith, and M. Zaharia, Eds., vol. 1, 2019, pp. 374–388.
[2] T. Song, Y. Tong, and S. Wei, "Profit allocation for federated learning." in *BigData*. IEEE, 2019, pp. 2577–2586.
[3] C. Hasan, "Incentive mechanism design for federated learning: Hedonic game approach," *CoRR*, vol. abs/2101.09673, 2021. [Online]. Available: https://arxiv.org/abs/2101.09673

Contents



Framework

Data exchange

- Randomly select local data and send to neighbors
- Privacy leakage
- Reduce parameter drift

Decentralized learning

- Robust learning scheme
- Consensus propagation without central server
- Performance enhancement

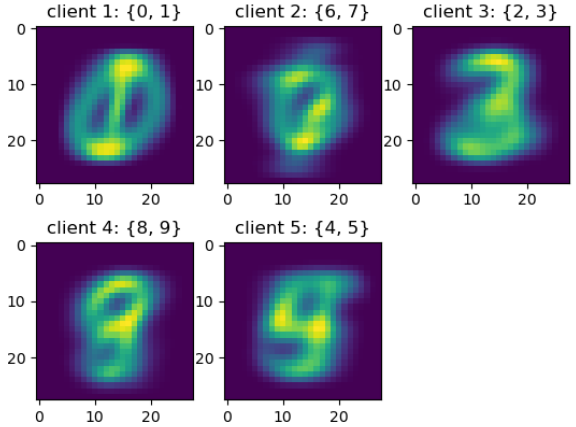
Data exchange

Data exchange

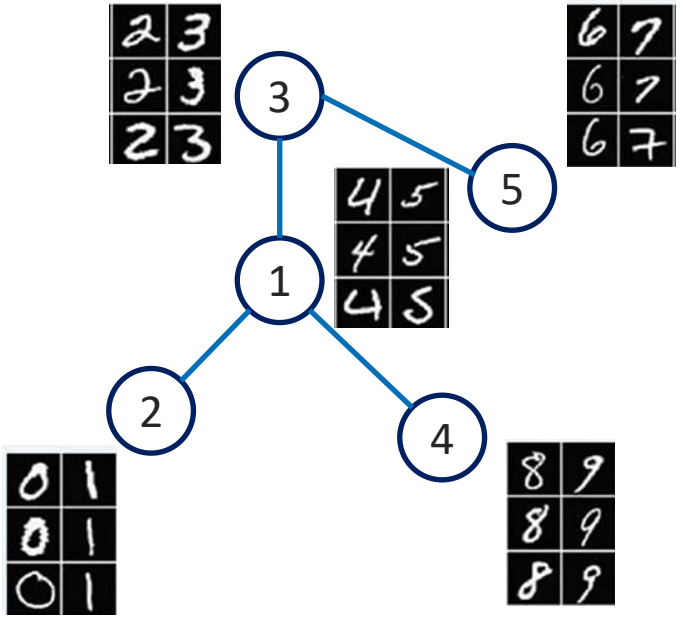
- Non i.i.d. data distribution
- Randomly select data and transmit to neighbors

Privacy leakage calculation

- Intensity distribution



- KL divergence between the distribution of transmitted images and neighbor's images



$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Decentralized learning

Target:

- Collaboratively solve a consensus optimization problem

$$\underset{w\in\mathbb{R}^d}{\text{minimize}} \quad f(w)=\sum_{i=1}^{L} f(w,\mathcal{D}_i)=\sum_{i=1}^{L} f_i(w)$$

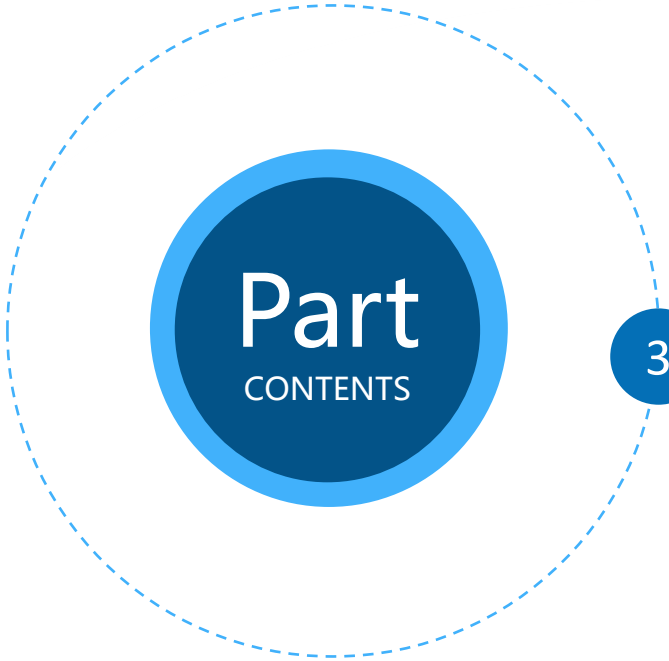
Methods

- Compute gradient via local dataset
- Update local parameters
- Compute neighborhood weighted average

$$w_{(i),t+1}=\sum_{j=1}^{L} p_{ij}\left(w_{(j),t}-\alpha\nabla f_j\left(w_{(j),t}\right)\right), \quad i=1,2,\ldots,n$$

p for mixing matrix $P=\begin{bmatrix} 1-2\tau & \tau & \tau \\ \tau & \tau & 1-2\tau \\ \tau & 1-2\tau & \tau \end{bmatrix}$

Contents



Evaluation

Metric

MNIST accuracy

Privacy leakage

Settings

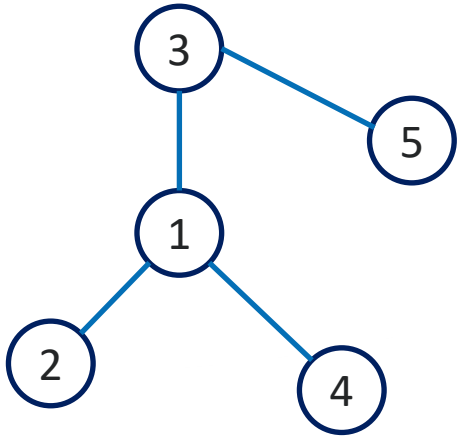
Local epoch: 3
Local batch size: 64
Momentum: 0.5
Learning rate: 0.01

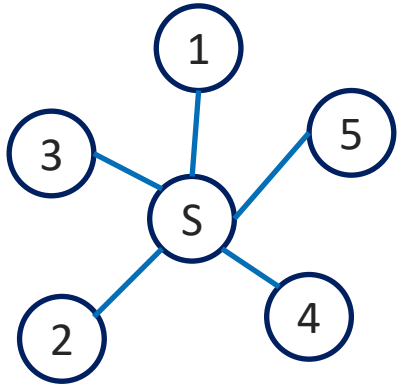
Training rounds: 10
GPU: GeForce GTX 1080 Ti

Experimental setup

DL models & distributed structure:

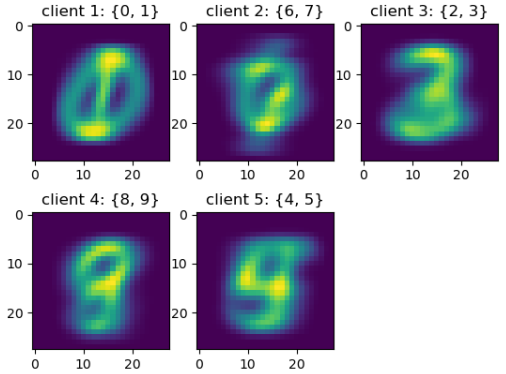
- Decentralized & Parameter server
- MLP & CNN (Non-IID MNIST)

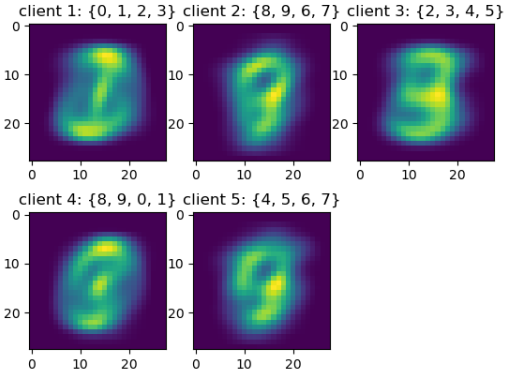




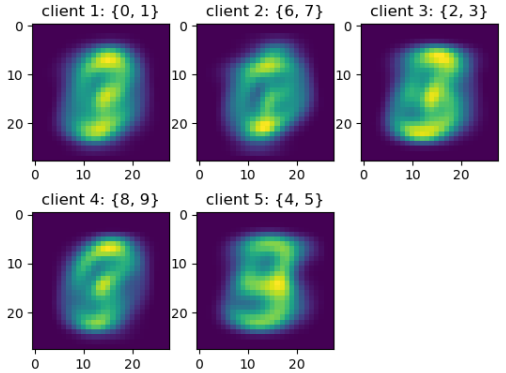
Data exchange

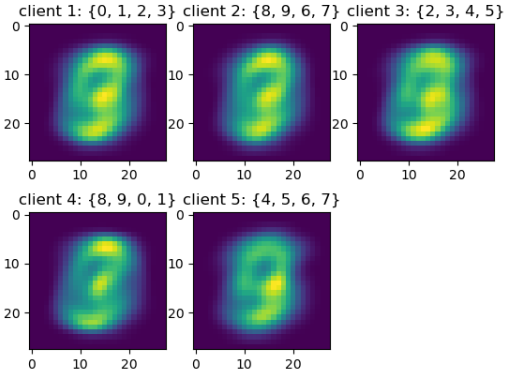
Original distribution





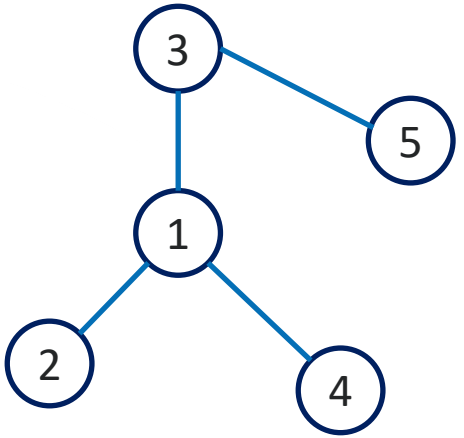
Exchanged distribution





Share ratio = 0.2

Share ratio = 0.8



Heterogeneity comes from

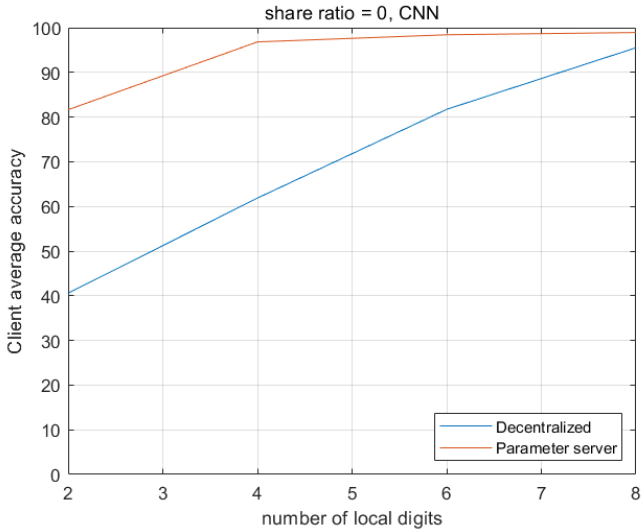
- Initial data distribution
- Volume of exchanged data
- Degree of nodes

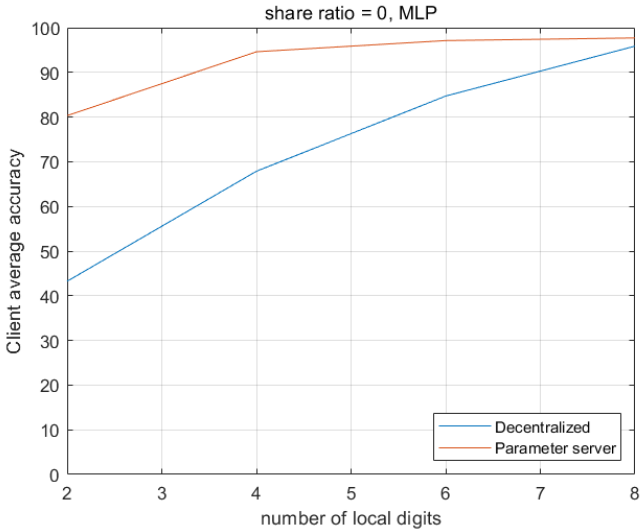
Learning Performance

中国科学技术大学 University of Science and Technology of China

Parameter server vs Decentralized learning

- MLP performs better in parameter server structure, CNN performs better in decentralized structure
- PS structure is more robust against data heterogeneity



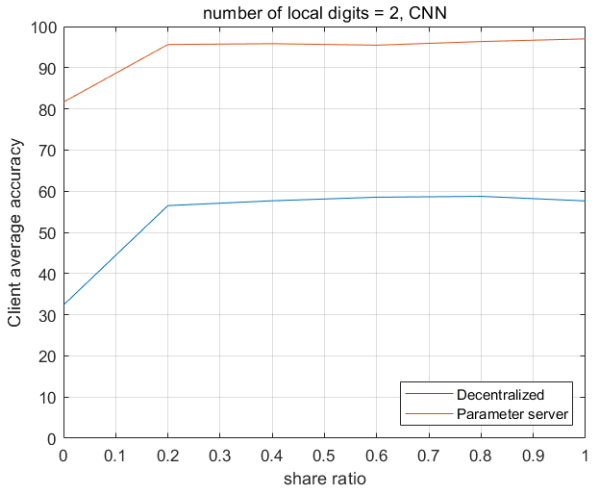


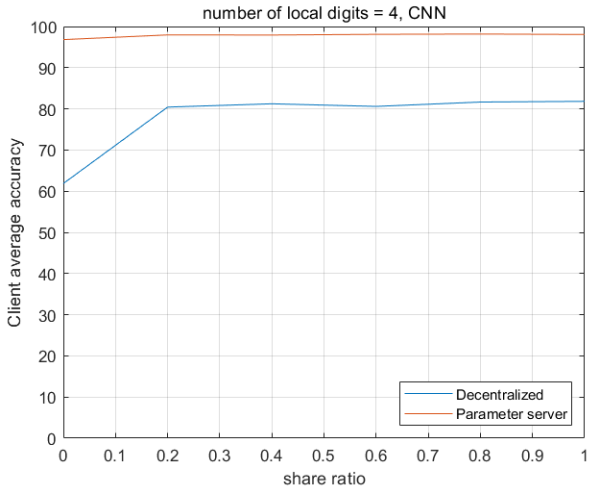
Learning Performance

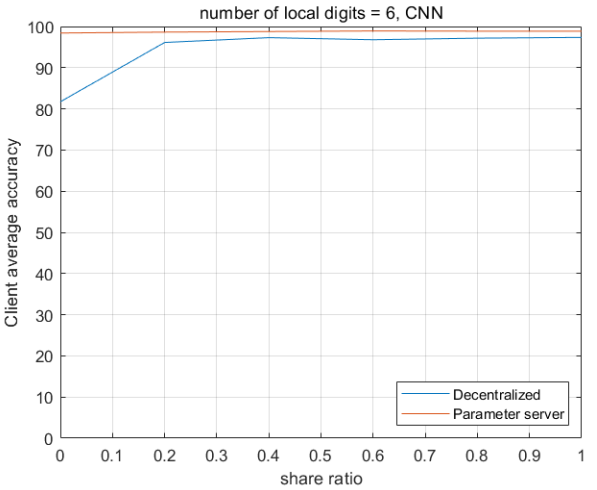
中国科学技术大学
University of Science and Technology of China

Parameter server vs Decentralized learning

- Data exchange helps improving performance, in both settings
- “The more data exchange, the better performance” is not true
- Initial distribution is crucial in decentralized learning



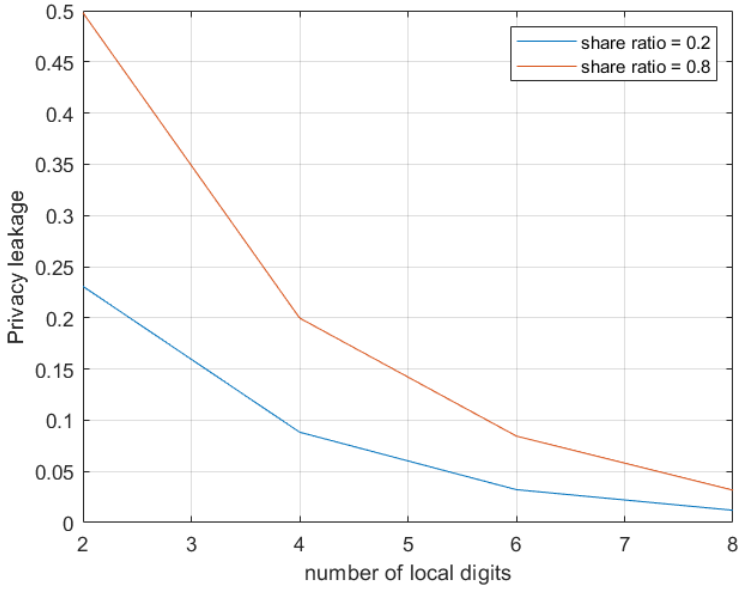


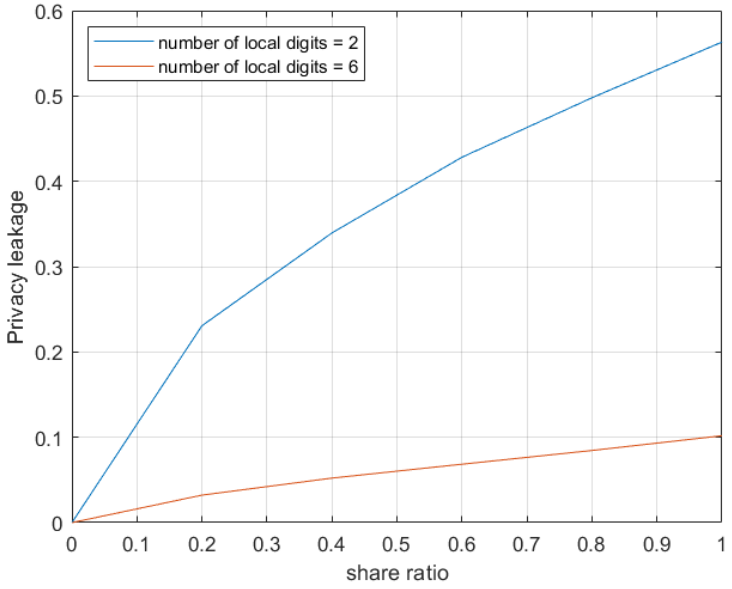


Privacy leakage

Parameter server vs Decentralized learning

- More similar data, less privacy leakage
- Less data exchange does not mean less privacy leakage



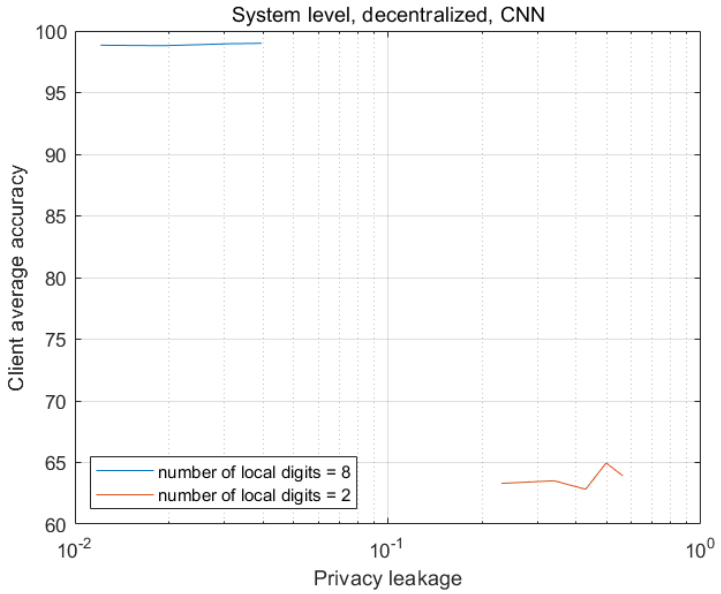


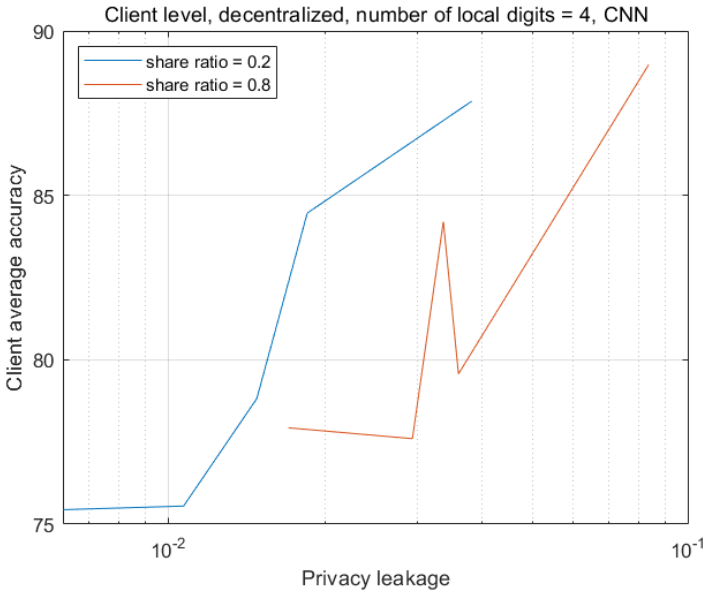
Should I share private data?

中国科学技术大学
University of Science and Technology of China

Correlation between privacy leakage and learning performance

- Proper clustering is much more effective than encouraging sharing private data
- More privacy leakage of a client does improve its performance



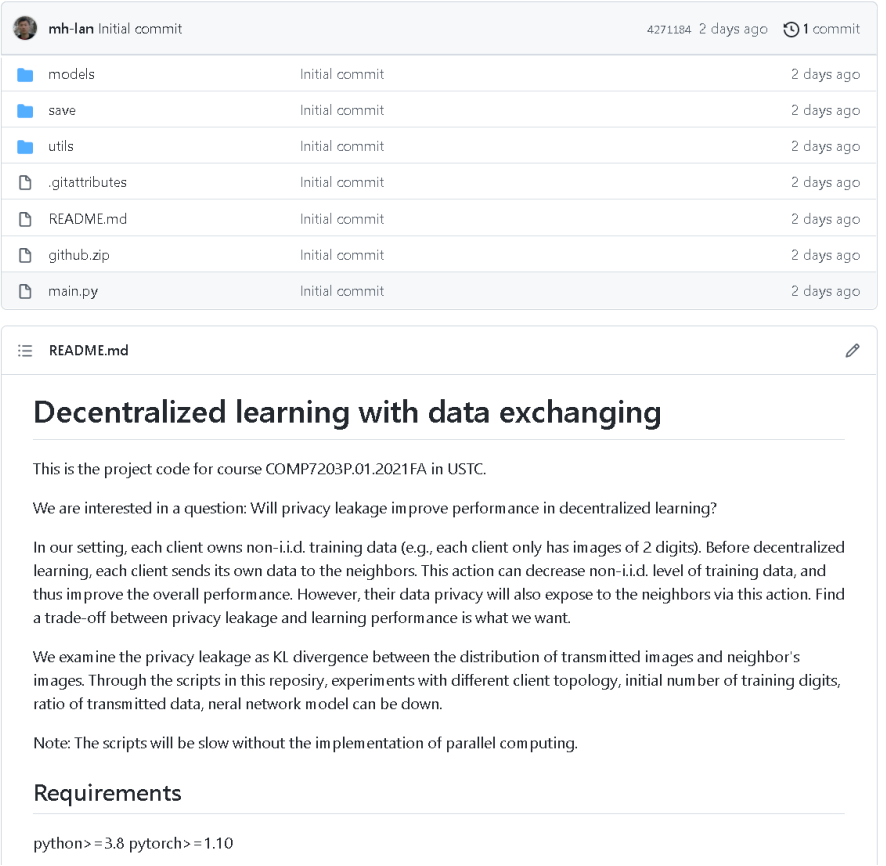


Code sharing

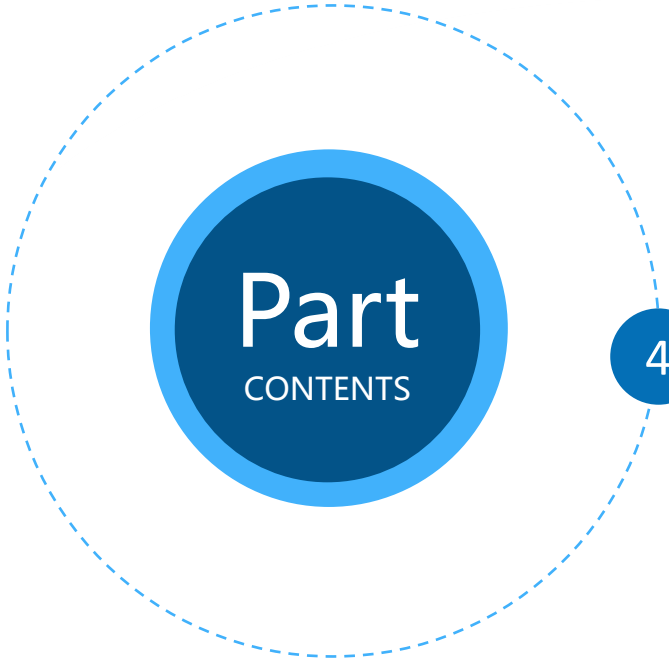
Find more in source code!

Repository:

https://github.com/mh-lan/privacy_performance_tradeoff



Contents



Summary

- Consider privacy in contribution assessment
- Extensive emulation with adjustable code framework
- **Proper clustering** is much more effective than encouraging sharing private data
- More privacy leakage of a client **does improve its performance**

Challenges & Future works

1. Mixing matrix/Topology

Key nodes in graph, graph regeneration

2. Real dataset

Multimodal learning, low quality dataset

3. Theoretical analysis with heterogeneity