DICE: Data Influence Cascade in Decentralized Learning

Speaker: Tongtian Zhu

Cascade

Collaborators

Many thanks to collaborators!

Tongtian Zhu
Zhejiang University

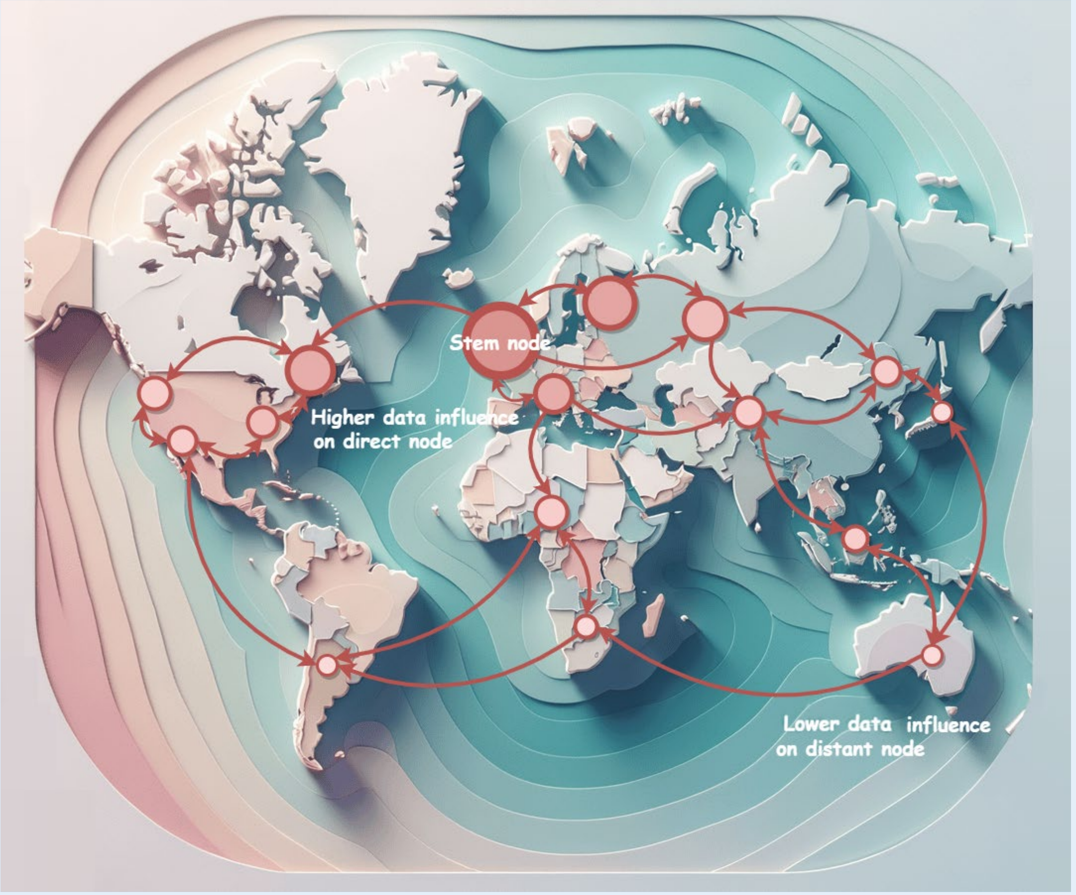
Wenhao Li
Zhejiang University

Can Wang
Zhejiang University

Fengxiang He
The University of Edinburgh

Main Results

Data Influence Cascade in Decentralized Learning

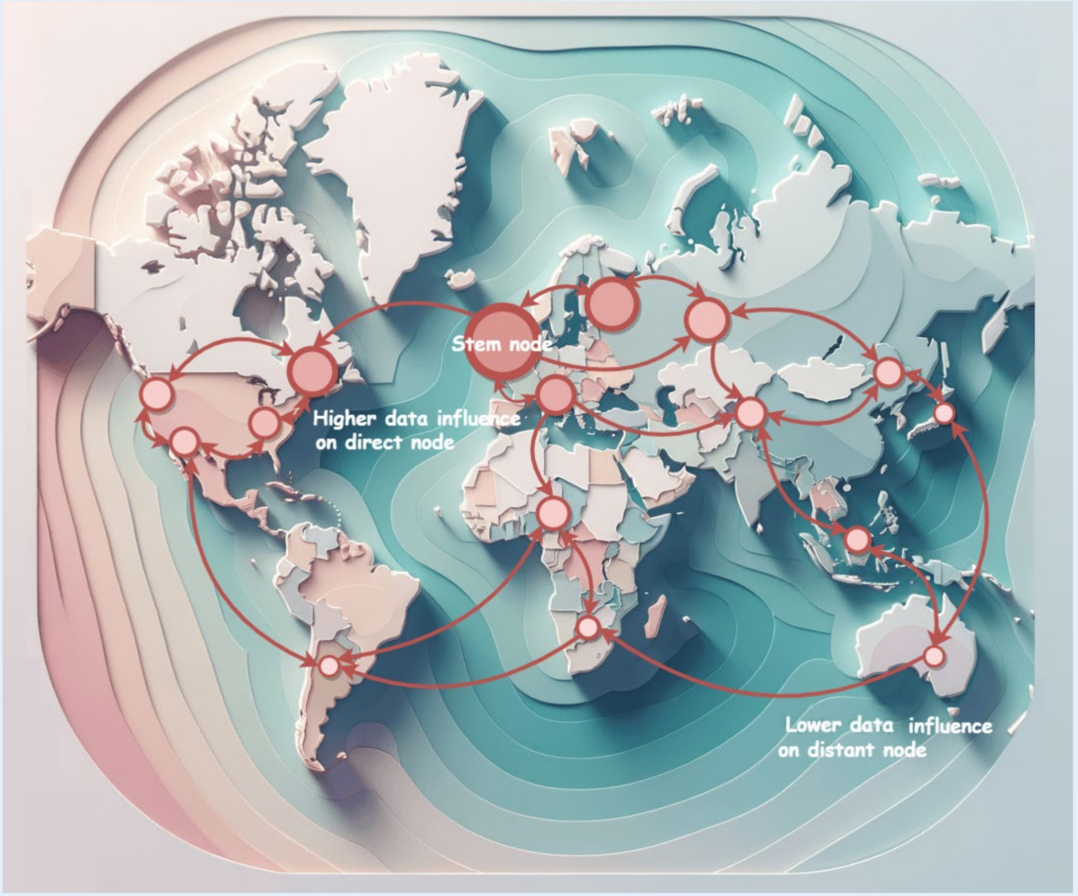


What scientific problem does this paper study?

Q: In decentralized financial systems, proof of work (PoW) ensures security and consensus through computational effort.
How can PoW be formally defined and quantified in the context of decentralized distributed machine learning?

Main Results

Data Influence Cascade in Decentralized Learning



What scientific problem does this paper study?

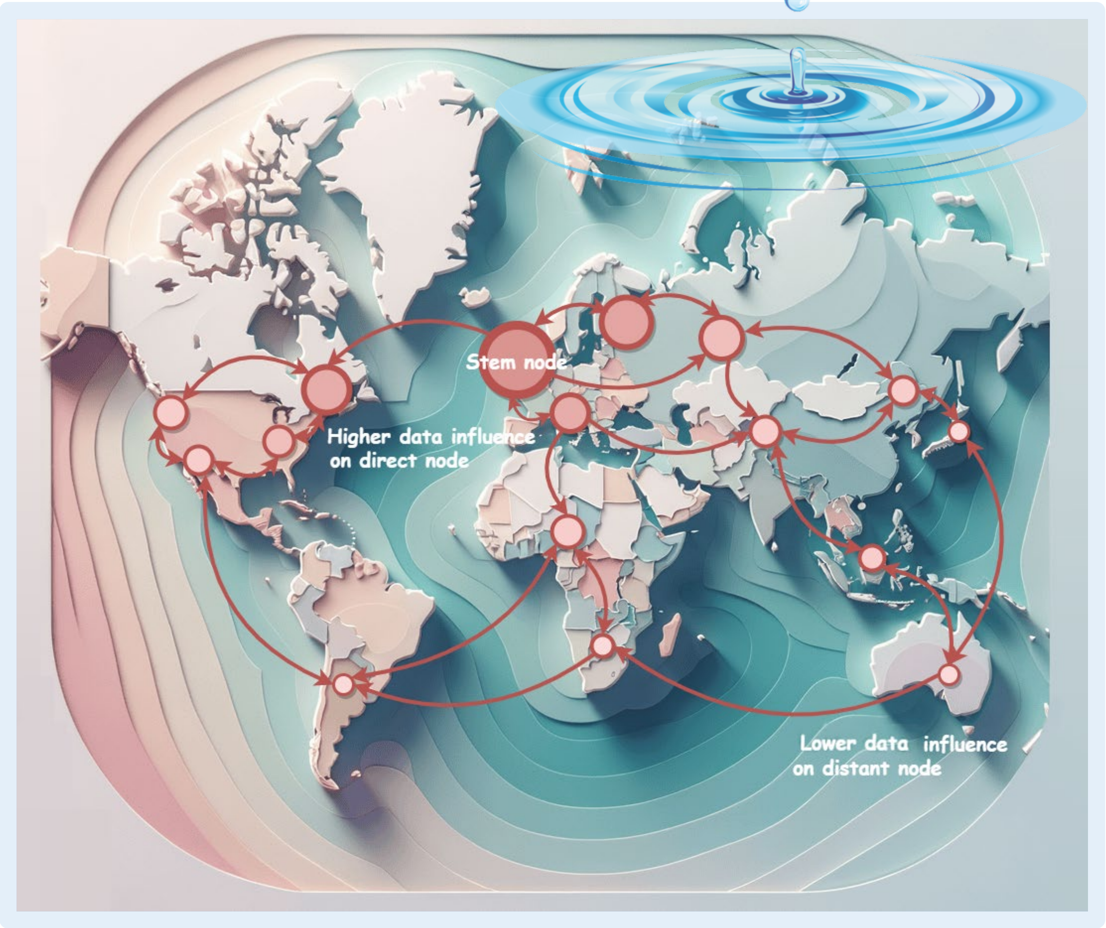
Q: In decentralized financial systems, proof of work (PoW) ensures security and consensus through computational effort. How can PoW be formally defined and quantified in the context of decentralized distributed machine learning?

What phenomena does this paper uncover?

The influence of data “cascades” through the communication graph, resembling “ripples in water”.

Main Results

Data Influence Cascade in Decentralized Learning



What scientific problem does this paper study?

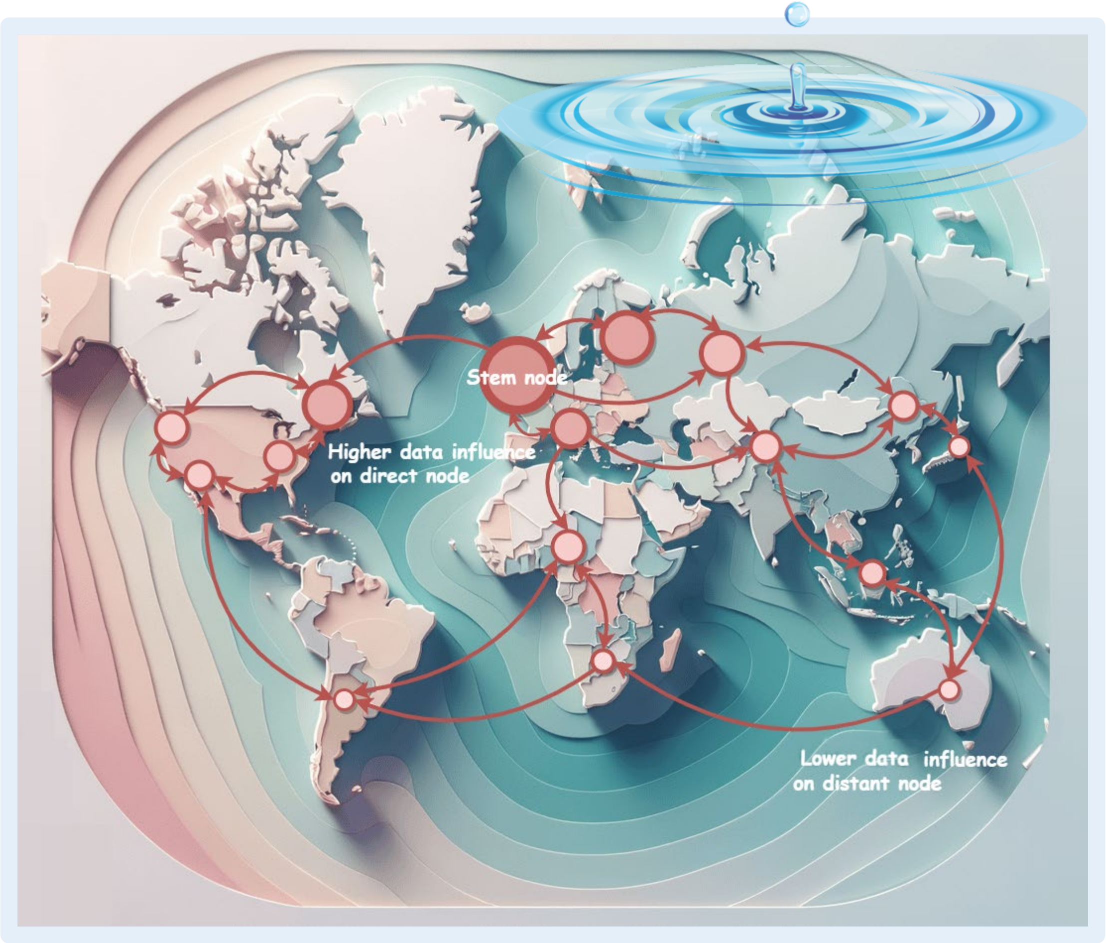
Q: In decentralized financial systems, proof of work (PoW) ensures security and consensus through computational effort. How can PoW be formally defined and quantified in the context of decentralized distributed machine learning?

What phenomena does this paper uncover?

The influence of data "cascades" through the communication graph, resembling "ripples in water".

Main Results

Data Influence Cascade in Decentralized Learning



What phenomena does this paper uncover?

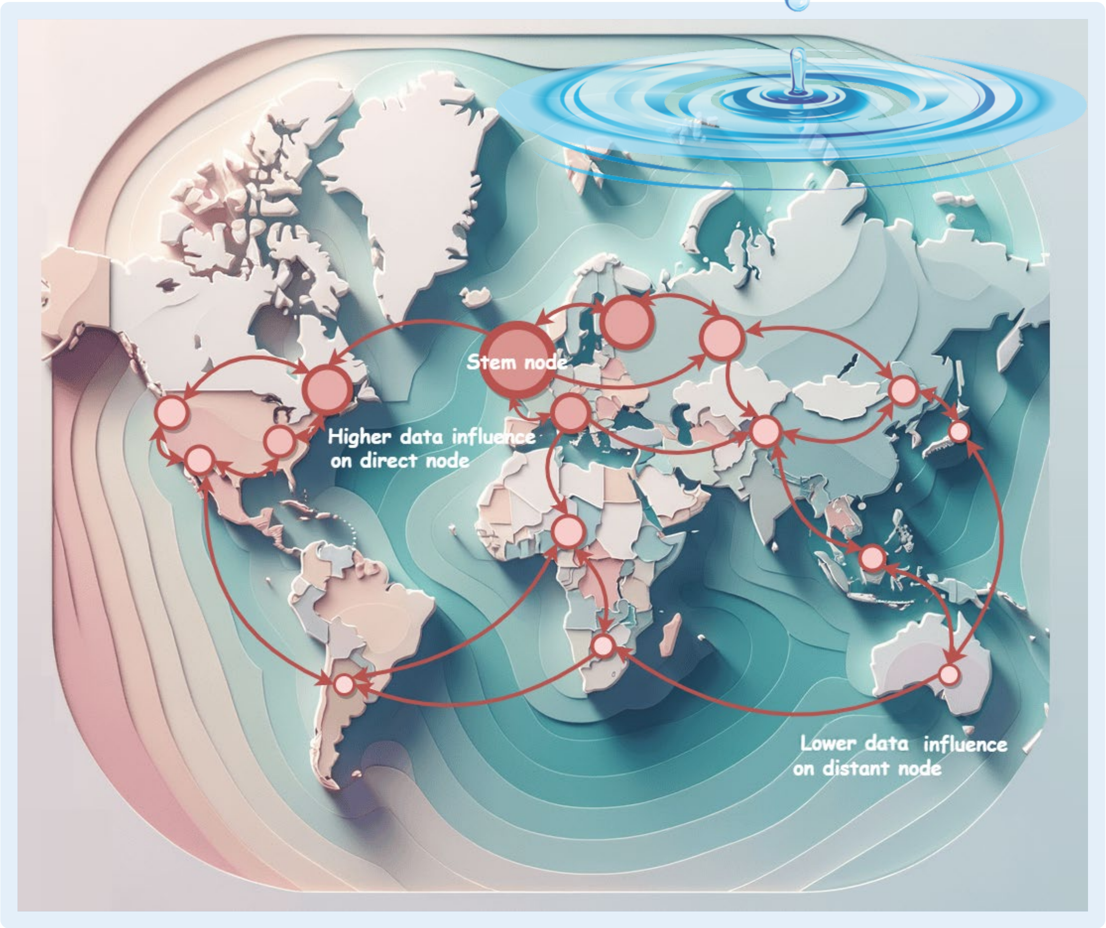
In decentralized learning, the influence of data “cascades” through the communication graph, resembling “ripples in water”.

This influence is determined by both the original data and the topological position of the data-holding node within the communication network.

Intuition?

Main Results

Data Influence Cascade in Decentralized Learning

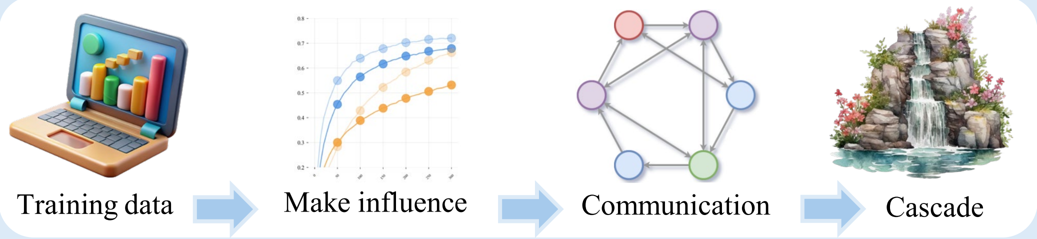


What phenomena does this paper uncover?

In decentralized learning, the influence of data “cascades” through the communication graph, resembling “ripples in water”.

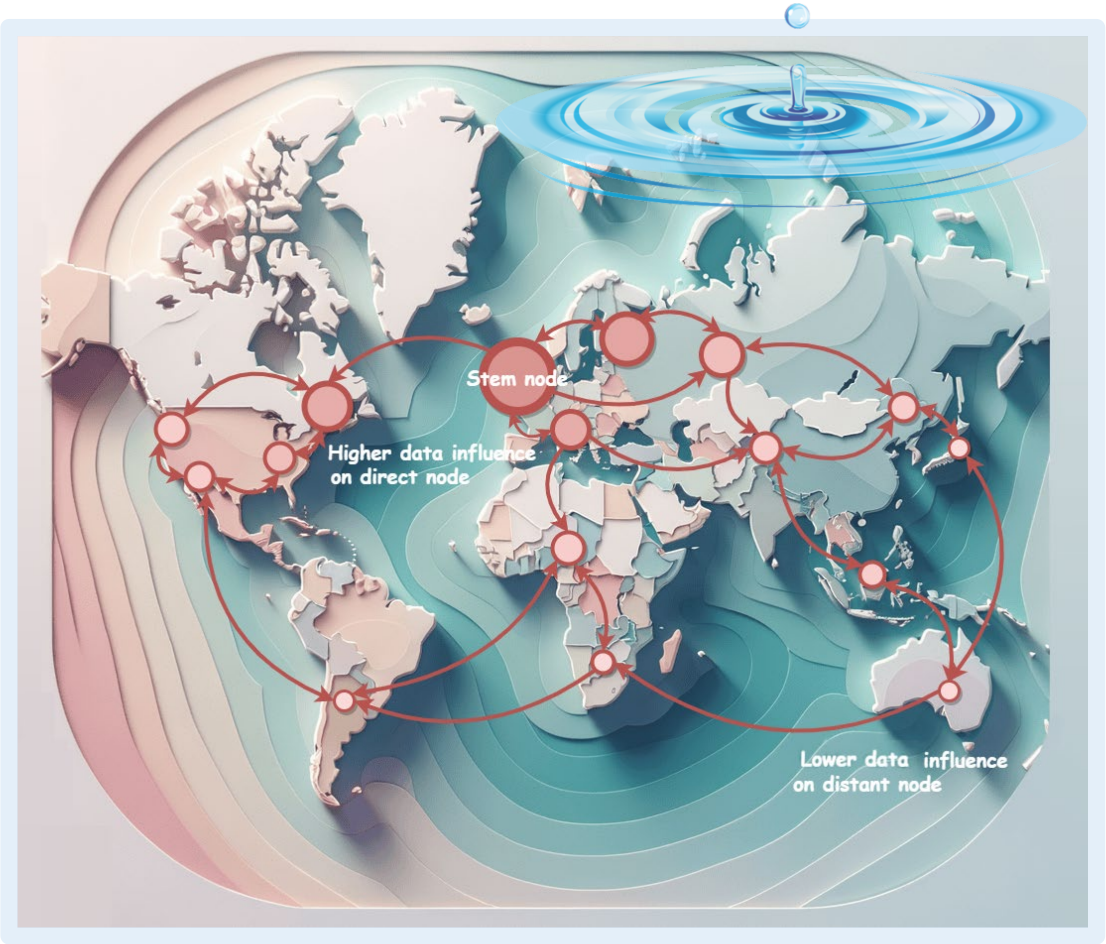
This influence is determined by both the original data and the topological position of the data-holding node within the communication network.

Intuition



Main Results

Data Influence Cascade in Decentralized Learning



What phenomena does this paper uncover?

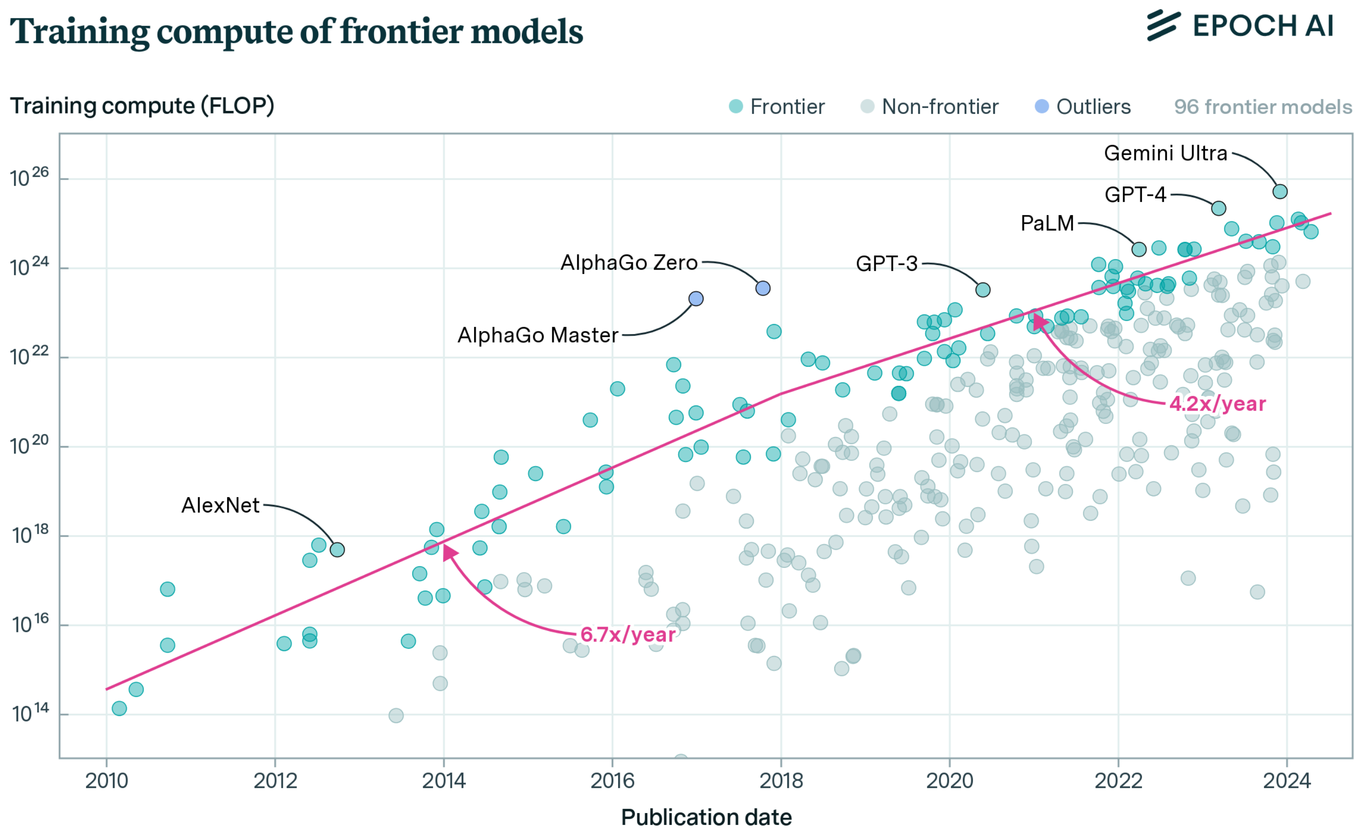
In decentralized learning, the influence of data “cascades” through the communication graph, resembling “ripples in water”.

This influence is determined by both the original data and the topological position of the data-holding node within the communication network.

Formally,

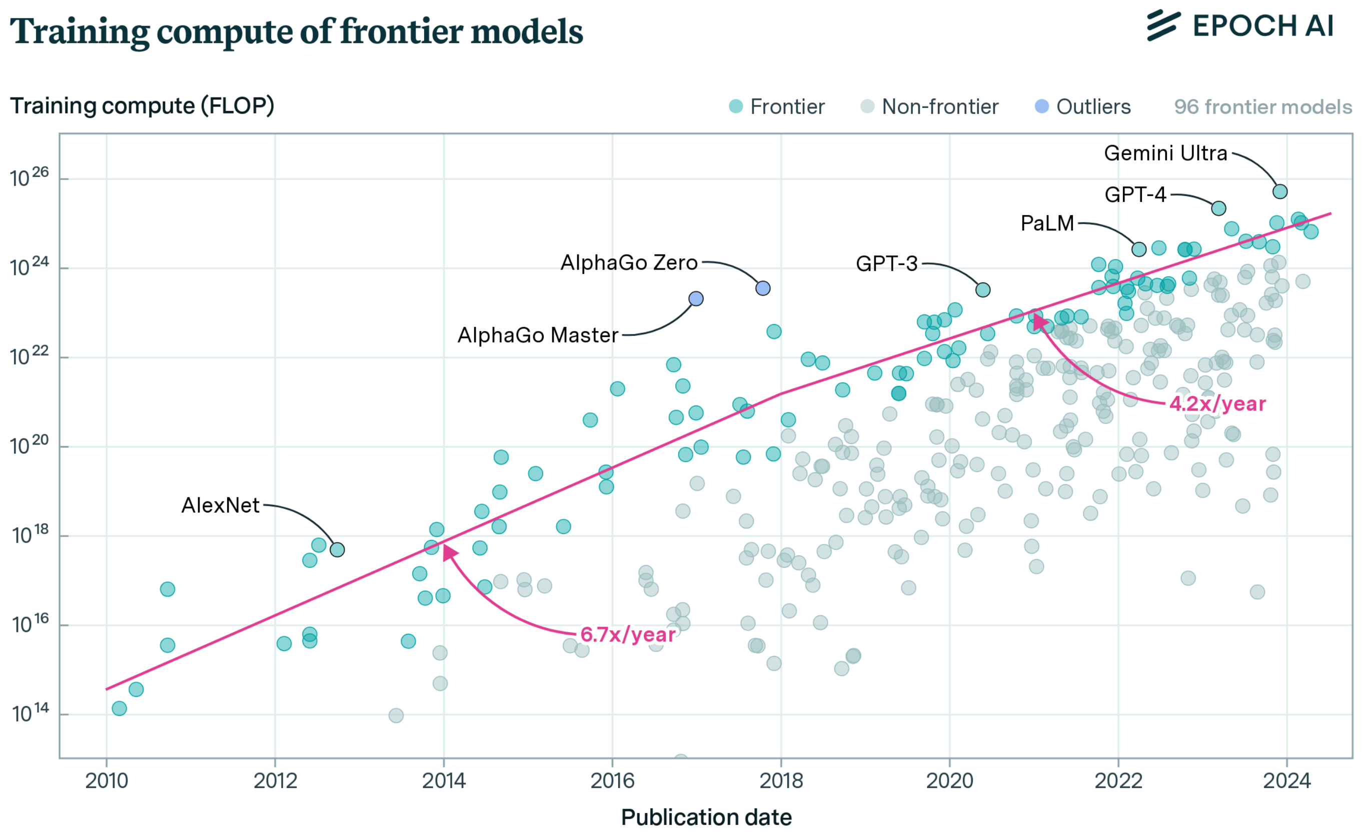
$$\mathcal{I}^{(r)}_{\text{DICE-E}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = -\sum_{\rho=0}^{r} \sum_{(k_1,\dots,k_\rho)\in P_j^{(\rho)}} \eta^t q_{k_\rho} \underbrace{\left(\prod_{s=1}^{\rho} \boldsymbol{W}^{t+s-1}_{k_s,k_{s-1}}\right)}_{\text{communication graph-related term}} \underbrace{\nabla L(\boldsymbol{\theta}^{t+\rho}_{k_\rho}; \boldsymbol{z}')^\top}_{\text{test gradient}} \times \underbrace{\left(\prod_{s=2}^{\rho}\left(\boldsymbol{I} - \eta^{t+s-1} \boldsymbol{H}(\boldsymbol{\theta}^{t+s-1}_{k_s}; \boldsymbol{z}^{t+s-1}_{k_s})\right)\right)}_{\text{curvature-related term}} \underbrace{\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t)}_{\text{optimization-related term}}.$$

Background



Training Compute of Frontier AI Models Grows by 4-5x per Year. Epoch AI, 2024.

Background

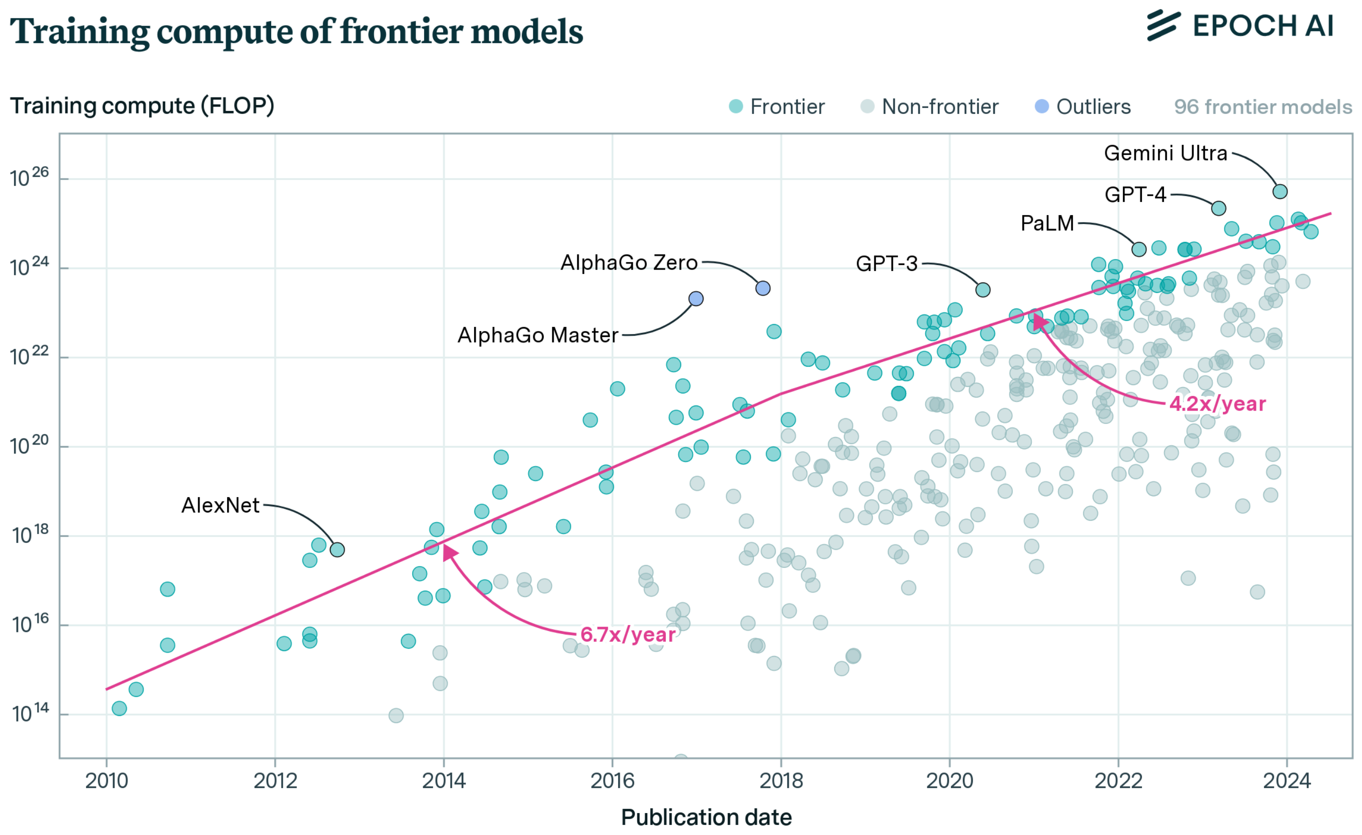


Estimated Compute Cost
GPT-4: $78 million
Gemini Ultra: $191 million

Training Compute of Frontier AI Models Grows by 4-5x per Year. Epoch AI, 2024.

The AI Index 2024 Annual Report. Institute for Human-Centered AI, Stanford University, 2024.

Background



Estimated Compute Cost
GPT-4: $78 million
Gemini Ultra: $191 million

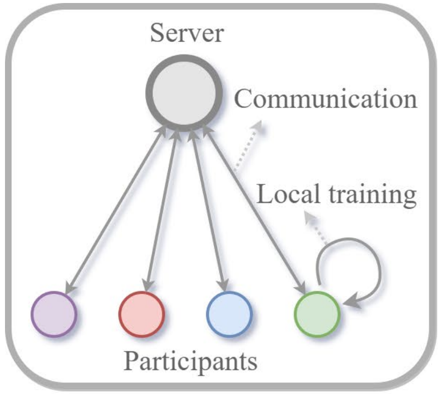
The exponentially growing compute demands imposes a financial burden far beyond the affordability of academia and individuals.

Training Compute of Frontier AI Models Grows by 4-5x per Year. Epoch AI, 2024.
The AI Index 2024 Annual Report. Institute for Human-Centered AI, Stanford University, 2024.

Background

The exponentially growing compute demands imposes a financial burden far beyond the affordability of academia and individuals.

Large-scale training are primarily performed in costly data centers.

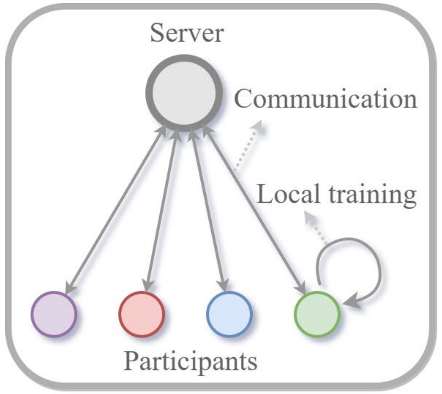


(a) Server-based Learning

Background

The exponentially growing compute demands imposes a financial burden far beyond the affordability of academia and individuals.

Large-scale training are primarily performed in costly data centers.

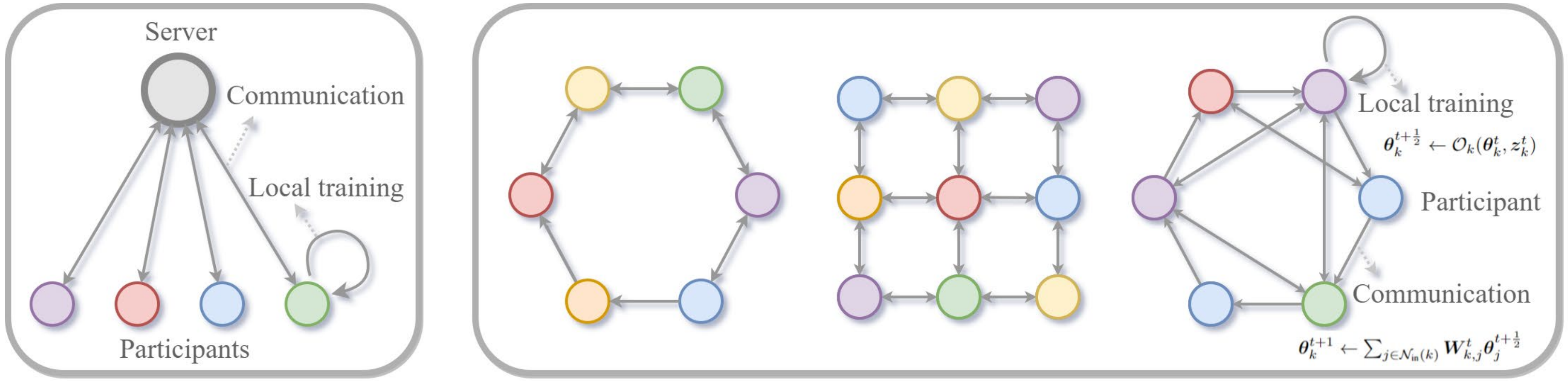


(a) Server-based Learning

Background

The exponentially growing compute demands imposes a financial burden far beyond the affordability of academia and individuals.

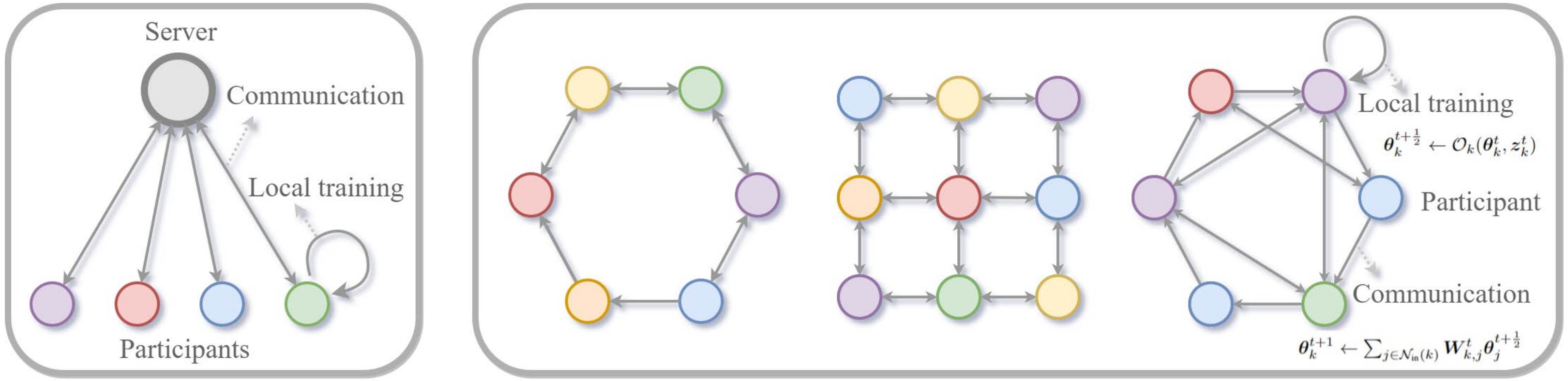
Large-scale training are primarily performed in costly data centers.



(a) Server-based Learning

(b) Decentralized Learning

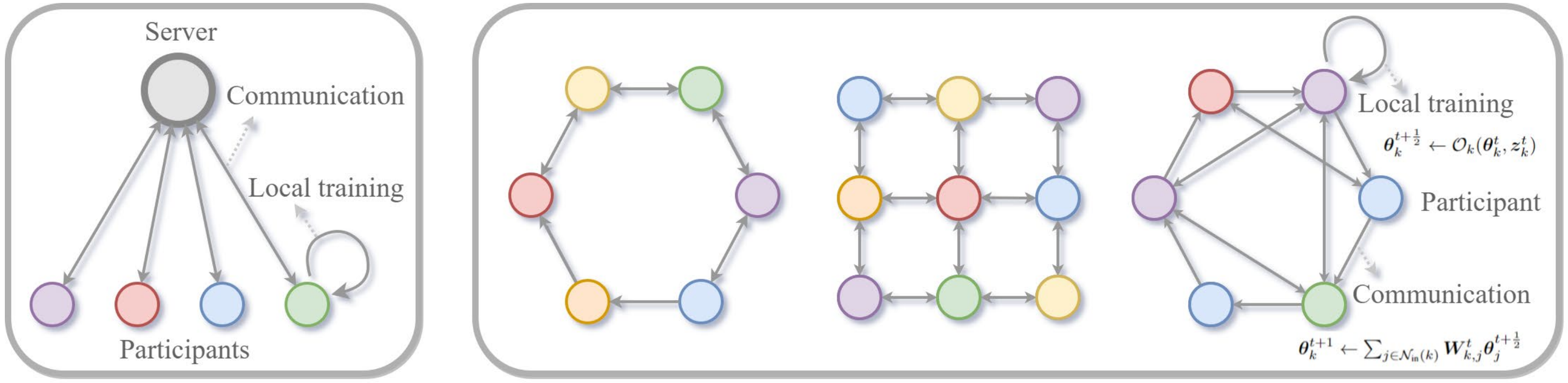
Motivating Question



(a) Server-based Learning

(b) Decentralized Learning

Motivating Question

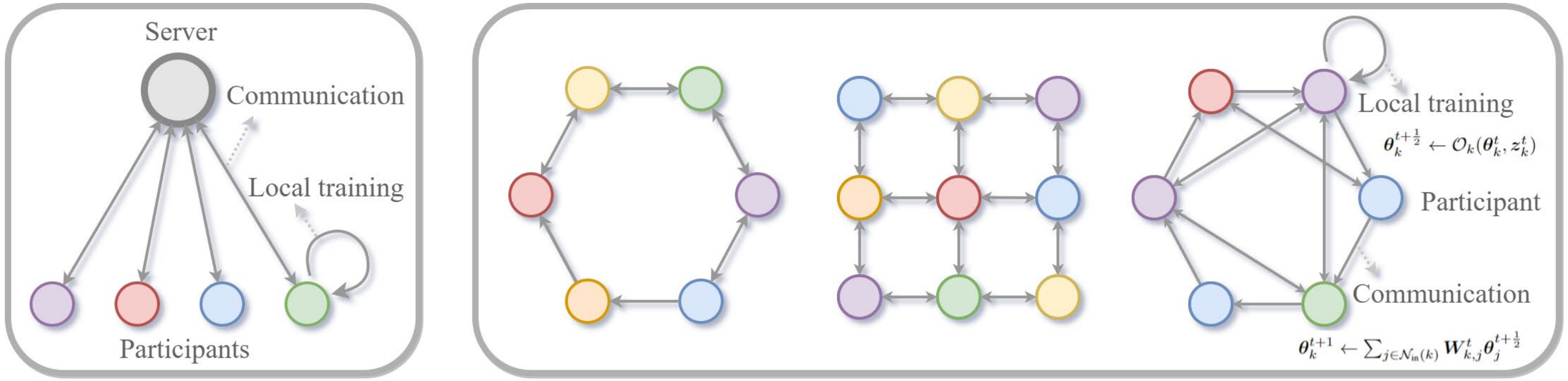


(a) Server-based Learning

(b) Decentralized Learning

Q: What motivates edge participants to engage in decentralized learning?

Motivating Question



(a) Server-based Learning

(b) Decentralized Learning

Q: What motivates edge participants to engage in decentralized learning?

Q: Can we quantify individual contributions in decentralized learning?
How can a proof-of-work mechanism be designed in this context?

How to Answer This Question?

Q: What motivates edge participants to engage in decentralized learning?

Q: Can we quantify individual contributions in decentralized learning?
How can a proof-of-work mechanism be designed in this context?

How to Answer This Question?

Q: What motivates edge participants to engage in decentralized learning?

Q: Can we quantify individual contributions in decentralized learning?
How can a proof-of-work mechanism be designed in this context?

A: Individual contributions can be quantified via data influence.

How to Answer This Question?

Q: What motivates edge participants to engage in decentralized learning?

Q: Can we quantify individual contributions in decentralized learning?
How can a proof-of-work mechanism be designed in this context?

A: Individual contributions can be quantified via data influence.

parameter-level contribution → data-level contribution

Setup

We consider a general *personalized distributed optimization problem* over a graph $G = (V, E)$

$$\min_{\theta=\{\theta_k \in R^d\}_{k\in V}} [L(\theta) \triangleq \sum_{k\in V} q_k L_k(\theta_k)].$$

Here each local objective $L_k(\theta_k) = \mathrm{E}_{z_k \sim D_k}[L(\theta_k; z_k)]$, where D_k denotes the local data distribution. Empirical risk minimization involves optimizing the sample average approximation:

$$\hat{L}(\theta) = \sum_{k\in V} q_k \hat{L}_k(\theta_k) \text{ where } \hat{L}_k(\theta_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} L\,(\theta_k; z_{k_i}).$$

Setup

We consider a general *personalized distributed optimization problem* over a graph $G = (V, E)$

$$\min_{\theta=\{\theta_k \in R^d\}_{k\in V}} \left[L(\theta) \triangleq \sum_{k\in V} q_k L_k(\theta_k)\right].$$

model parameters

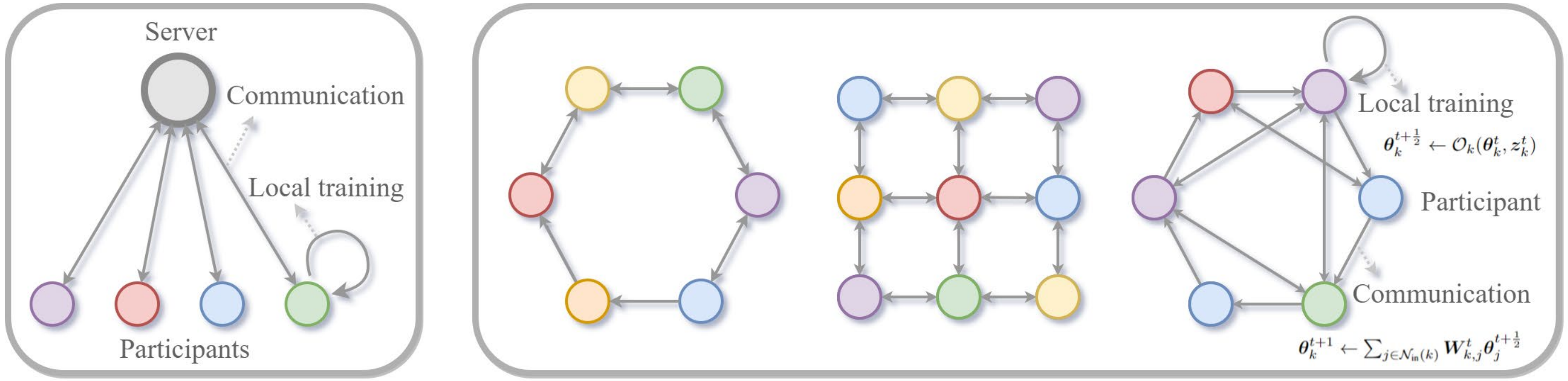
set of all participants

loss on local model and data

Here each local objective $L_k(\theta_k) = \mathrm{E}_{z_k \sim D_k}[L(\theta_k; z_k)]$, where D_k denotes the local data distribution. Empirical risk minimization involves optimizing the sample average approximation:

$$\hat{L}(\theta) = \sum_{k\in V} q_k \hat{L}_k(\theta_k) \text{ where } \hat{L}_k(\theta_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} L\left(\theta_k; z_{k_i}\right).$$

Setup



(a) Server-based Learning

(b) Decentralized Learning

Algorithm 1 Decentralized Learning with Flexible Gossip and Optimization

Require: $G = (\mathcal{V}, \mathcal{E})$, $\{\theta_k^0\}_{k\in\mathcal{V}}$, optimizer $\mathcal{O}_k$, number of communication rounds T, and mixing matrix distributions $\mathcal{W}^t$ $(\forall t \in [T])$

for $t = 1$ to T **do in parallel for all** participants $k \in \mathcal{V}$

Local Update:

Sample $z_k^t \sim \mathcal{D}_k$, update parameters with optimizer $\mathcal{O}_k$: $\theta_k^{t+\frac{1}{2}} \leftarrow \mathcal{O}_k(\theta_k^t, z_k^t)$

Gossip Averaging:

Send $\theta_k^{t+\frac{1}{2}}$ to $\{l \mid W_{l,k} > 0\}$ and receive $\theta_j^{t+\frac{1}{2}}$ from $\{j \mid W_{k,j} > 0\}$.

Sample $W^t \sim \mathcal{W}^t$, perform gossip averaging: $\theta_k^{t+1} \leftarrow \sum_{j\in\mathcal{N}_{in}(k)} W_{k,j}^t \theta_j^{t+\frac{1}{2}}$

End for

Background of Data Influence

Definition 1 (Leave-one-out Influence).

$$\mathcal{I}_{\mathrm{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus z}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Background of Data Influence

Definition 1 (Leave-one-out Influence)**.**

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus z}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Understanding Black-box Predictions via Influence Functions

Pang Wei Koh [1] **Percy Liang** [1]

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') \approx -\nabla_{\boldsymbol{\theta}} L\left(\boldsymbol{z}', \boldsymbol{\theta}^*\right)^\top H_{\boldsymbol{\theta}^*}^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{z}, \boldsymbol{\theta}^*)$$

Main Results

Definition 1 (Leave-one-out Influence).

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Question: what makes decentralized learning different?

1. The presence of multiple local models trained on Non-IID data, which may lead to diverse local optima.
2. The concept of "neighbors" plays a crucial role, as model parameters are exchanged only among neighboring nodes, allowing for the indirect propagation of data influence.

Main Results

Definition 1 (Leave-one-out Influence)**.**

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus z}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Key observations: *In decentralized learning,*
1) neighbors who serves as customers hold the rights to determine data influence;
2) data influence is not static but spreads across participants through gossips during training.

Main Results

Definition 1 (Leave-one-out Influence)**.**

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus z}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Key observations: *In decentralized learning,*
1) neighbors who serves as customers hold the rights to determine data influence;
2) data influence is not static but spreads across participants through gossips during training.

Unfortunately, the original formulation of data influence **cannot** account for these two key characteristics of decentralized learning.

Main Results

Definition 1 (Leave-one-out Influence).

$$\mathcal{I}_{\text{LOO}}(\boldsymbol{z}, \boldsymbol{z}') = L(\boldsymbol{\theta}^*; \boldsymbol{z}') - L(\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}; \boldsymbol{z}'),$$

where $\boldsymbol{z}$ denotes the training data instance under influence assessment, $\boldsymbol{z}'$ is the loss-evaluating instance, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\setminus \boldsymbol{z}}$ are the models trained on the entire dataset $\mathcal{S}$ and $\mathcal{S} \setminus \{z\}$, respectively.

Definition 2 (One-hop Ground-truth Influence). The one-hop DICE-GT value quantifies the influence of a data instance $\boldsymbol{z}_j^t$ from participant j on a loss-evaluating instance $\boldsymbol{z}'$ within itself and its immediate neighbors. Formally, for a given participant $j \in \mathcal{V}$:

$$\mathcal{I}^{(1)}_{\text{DICE-GT}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = \underbrace{q_j \left(L(\boldsymbol{\theta}_j^{t+\frac{1}{2}}; \boldsymbol{z}') - L(\boldsymbol{\theta}_j^t; \boldsymbol{z}') \right)}_{\text{direct marginal contribution of } \boldsymbol{z}_j^t \text{ to } j} + \underbrace{\sum_{k \in \mathcal{N}^{(1)}_{\text{out}}(j)} q_k \left(L(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{z}') - L(\boldsymbol{\theta}^{t+1}_{k \setminus \boldsymbol{z}_j^t}; \boldsymbol{z}') \right)}_{\text{indirect marginal contribution of } \boldsymbol{z}_j^t \text{ to one-hop neighbors}}.$$

Local Update:

Sample $z_k^t \sim \mathcal{D}_k$, update parameters with optimizer $\mathcal{O}_k$: $\boldsymbol{\theta}_k^{t+\frac{1}{2}} \leftarrow \mathcal{O}_k(\boldsymbol{\theta}_k^t, \boldsymbol{z}_k^t)$

Main Results

Definition 2 (One-hop Ground-truth Influence). The one-hop DICE-GT value quantifies the influence of a data instance $\boldsymbol{z}_j^t$ from participant j on a loss-evaluating instance $\boldsymbol{z}'$ within itself and its immediate neighbors. Formally, for a given participant $j \in \mathcal{V}$:

$$\mathcal{I}_{\text{DICE-GT}}^{(1)}(\boldsymbol{z}_j^t, \boldsymbol{z}') = \underbrace{q_j \left(L(\boldsymbol{\theta}_j^{t+\frac{1}{2}}; \boldsymbol{z}') - L(\boldsymbol{\theta}_j^t; \boldsymbol{z}') \right)}_{\text{direct marginal contribution of } \boldsymbol{z}_j^t \text{ to } j} + \underbrace{\sum_{k \in \mathcal{N}_{\text{out}}^{(1)}(j)} q_k \left(L(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{z}') - L(\boldsymbol{\theta}_{k \setminus \boldsymbol{z}_j^t}^{t+1}; \boldsymbol{z}') \right)}_{\text{indirect marginal contribution of } \boldsymbol{z}_j^t \text{ to one-hop neighbors}}.$$

Main Results

Definition 2 (One-hop Ground-truth Influence). The one-hop DICE-GT value quantifies the influence of a data instance $\boldsymbol{z}_j^t$ from participant j on a loss-evaluating instance $\boldsymbol{z}'$ within itself and its immediate neighbors. Formally, for a given participant $j \in \mathcal{V}$:

$$\mathcal{I}_{\text{DICE-GT}}^{(1)}(\boldsymbol{z}_j^t, \boldsymbol{z}') = \underbrace{q_j \left(L(\boldsymbol{\theta}_j^{t+\frac{1}{2}}; \boldsymbol{z}') - L(\boldsymbol{\theta}_j^t; \boldsymbol{z}') \right)}_{\text{direct marginal contribution of } \boldsymbol{z}_j^t \text{ to } j} + \underbrace{\sum_{k \in \mathcal{N}_{\text{out}}^{(1)}(j)} q_k \left(L(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{z}') - L(\boldsymbol{\theta}_{k \backslash \boldsymbol{z}_j^t}^{t+1}; \boldsymbol{z}') \right)}_{\text{indirect marginal contribution of } \boldsymbol{z}_j^t \text{ to one-hop neighbors}}.$$

Proposition 1 (Approximation of One-hop DICE-GT). The one-hop DICE-GT value (see Definition 2) can be linearly approximated as follow:

$$\mathcal{I}_{\text{DICE-E}}^{(1)}(\boldsymbol{z}_j^t, \boldsymbol{z}') = -q_j \nabla L(\boldsymbol{\theta}_j^t; \boldsymbol{z}')^\top \Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) - \sum_{k \in \mathcal{N}_{\text{out}}^{(1)}(j)} q_k \boldsymbol{W}_{k,j}^t \nabla L(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{z}')^\top \Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t),$$

where $\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) = \mathcal{O}_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) - \boldsymbol{\theta}_j^t$.

Local Update:

Sample $z_k^t \sim \mathcal{D}_k$, update parameters with optimizer $\mathcal{O}_k$: $\boldsymbol{\theta}_k^{t+\frac{1}{2}} \leftarrow \mathcal{O}_k(\boldsymbol{\theta}_k^t, \boldsymbol{z}_k^t)$

Main Results

Definition 2 (One-hop Ground-truth Influence)**.** The one-hop DICE-GT value quantifies the influence of a data instance $\boldsymbol{z}_j^t$ from participant j on a loss-evaluating instance $\boldsymbol{z}'$ within itself and its immediate neighbors. Formally, for a given participant $j \in \mathcal{V}$:

$$\mathcal{I}^{(1)}_{\text{DICE-GT}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = \underbrace{q_j \left(L(\boldsymbol{\theta}_j^{t+\frac{1}{2}}; \boldsymbol{z}') - L(\boldsymbol{\theta}_j^t; \boldsymbol{z}') \right)}_{\text{direct marginal contribution of } \boldsymbol{z}_j^t \text{ to } j} + \underbrace{\sum_{k \in \mathcal{N}^{(1)}_{\text{out}}(j)} q_k \left(L(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{z}') - L(\boldsymbol{\theta}_{k \setminus \boldsymbol{z}_j^t}^{t+1}; \boldsymbol{z}') \right)}_{\text{indirect marginal contribution of } \boldsymbol{z}_j^t \text{ to one-hop neighbors}}.$$

Definition 3 (Multi-hop Ground-truth Influence)**.** The multi-hop DICE-GT value quantifies the cumulative influence of a data instance $\boldsymbol{z}$ on a loss-evaluating instance $\boldsymbol{z}'$ across all nodes within r-hop neighborhoods of participant j. Formally, for a given participant $j \in \mathcal{V}$:

$$\mathcal{I}^{(r)}_{\text{DICE-GT}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = q_j \left(L(\boldsymbol{\theta}_j^{t+\frac{1}{2}}; \boldsymbol{z}') - L(\boldsymbol{\theta}_j^t; \boldsymbol{z}') \right) + \sum_{s=1}^{r} \sum_{k \in \mathcal{N}^{(s)}_{\text{out}}(j)} q_k \left(L(\boldsymbol{\theta}_k^{t+s}; \boldsymbol{z}') - L(\boldsymbol{\theta}_{k \setminus \boldsymbol{z}_j^t}^{t+s}; \boldsymbol{z}') \right).$$

Main Results

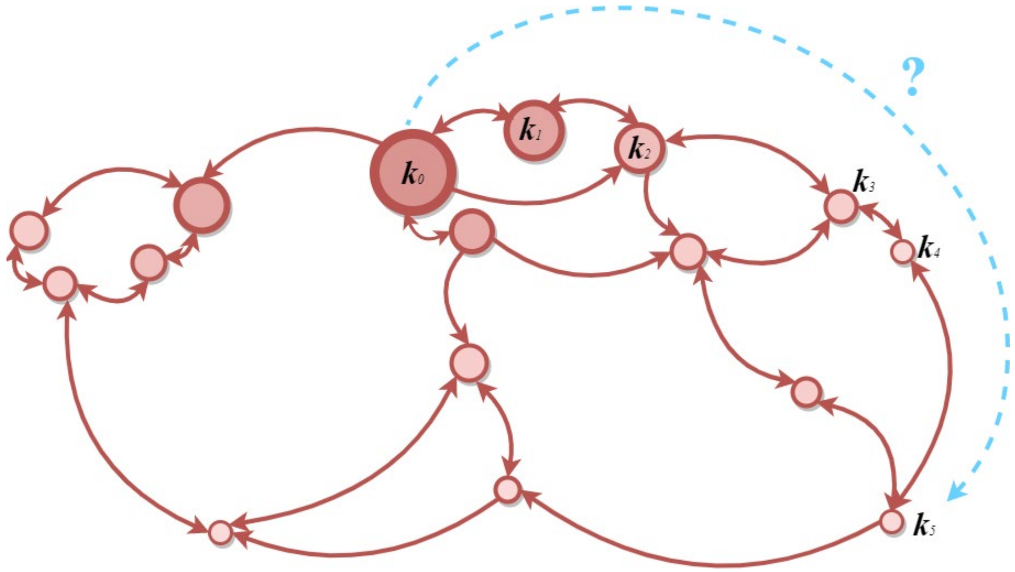
Theorem 2 (Approximation of r-hop DICE-GT). The r-hop DICE-GT influence $\mathcal{I}^{(r)}_{\text{DICE-GT}}(\boldsymbol{z}_j^t, \boldsymbol{z}')$ (see Definition 3) can be approximated as follows:

$$\mathcal{I}^{(r)}_{\text{DICE-E}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = -\sum_{\rho=0}^{r} \sum_{(k_1,\dots,k_\rho)\in P_j^{(\rho)}} \eta^t q_{k_\rho} \underbrace{\left(\prod_{s=1}^{\rho} \boldsymbol{W}_{k_s,k_{s-1}}^{t+s-1}\right)}_{\text{communication graph-related term}} \underbrace{\nabla L(\boldsymbol{\theta}_{k_\rho}^{t+\rho}; \boldsymbol{z}')^\top}_{\text{test gradient}}$$
$$\times \underbrace{\left(\prod_{s=2}^{\rho} \left(\boldsymbol{I} - \eta^{t+s-1} \boldsymbol{H}(\boldsymbol{\theta}_{k_s}^{t+s-1}; \boldsymbol{z}_{k_s}^{t+s-1})\right)\right)}_{\text{curvature-related term}} \underbrace{\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t)}_{\text{optimization-related term}} .$$

where $\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) = \mathcal{O}_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) - \boldsymbol{\theta}_j^t$, $k_0 = j$. Here $P_j^{(\rho)}$ denotes the set of all sequences $(k_1, \dots, k_\rho)$ such that $k_s \in \mathcal{N}_{\text{out}}^{(1)}(k_{s-1})$ for $s = 1, \dots, \rho$ (see Definition A.7) and $\boldsymbol{H}(\boldsymbol{\theta}_{k_s}^{t+s}; \boldsymbol{z}_{k_s}^{t+s})$ is the Hessian matrix of L with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_{k_s}^{t+s}$ and data $\boldsymbol{z}_{k_s}^{t+s}$. For the cases when $\rho = 0$ and $\rho = 1$, the relevant product expressions are defined as identity matrices, thereby ensuring that the r-hop DICE-E remains well-defined.

Main Results: Intuition

How can the influence of indirectly connected nodes—such as nodes k_0 to k_5—be quantified?



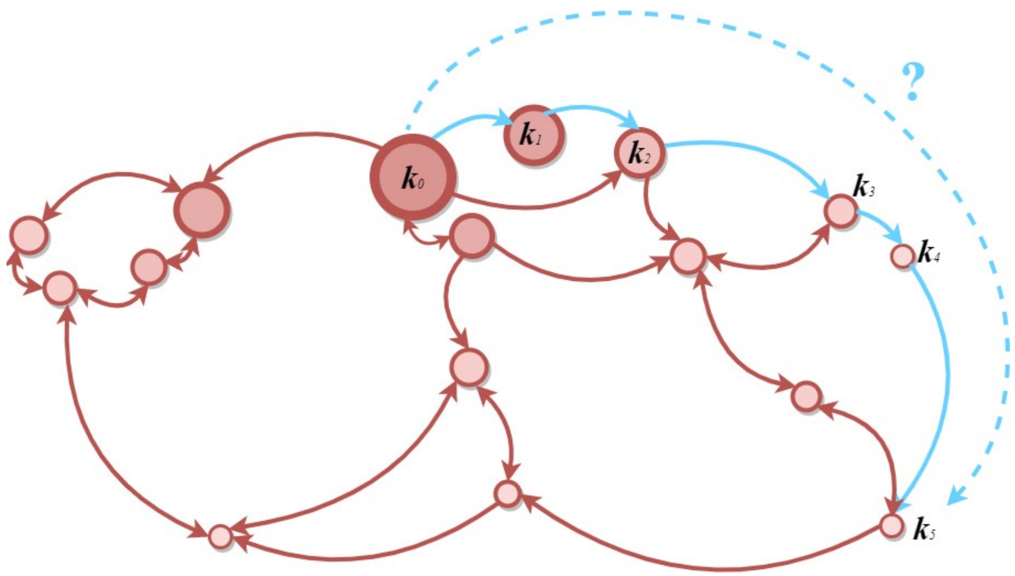
Theorem 2 (Approximation of r-hop DICE-GT). The r-hop DICE-GT influence $\mathcal{I}^{(r)}_{\text{DICE-GT}}(\boldsymbol{z}^t_j, \boldsymbol{z}')$ (see Definition 3) can be approximated as follows:

$$\mathcal{I}^{(r)}_{\text{DICE-E}}(\boldsymbol{z}^t_j, \boldsymbol{z}') = -\sum_{\rho=0}^{r} \sum_{(k_1,\ldots,k_\rho)\in P_j^{(\rho)}} \eta^t q_{k_\rho} \underbrace{\left(\prod_{s=1}^{\rho} \boldsymbol{W}^{t+s-1}_{k_s,k_{s-1}}\right)}_{\text{communication graph-related term}} \underbrace{\nabla L(\boldsymbol{\theta}^{t+\rho}_{k_\rho}; \boldsymbol{z}')^\top}_{\text{test gradient}} \times \underbrace{\left(\prod_{s=2}^{\rho} \left(\boldsymbol{I} - \eta^{t+s-1} \boldsymbol{H}(\boldsymbol{\theta}^{t+s-1}_{k_s}; \boldsymbol{z}^{t+s-1}_{k_s})\right)\right)}_{\text{curvature-related term}} \underbrace{\Delta_j(\boldsymbol{\theta}^t_j, \boldsymbol{z}^t_j)}_{\text{optimization-related term}} .$$

where $\Delta_j(\boldsymbol{\theta}^t_j, \boldsymbol{z}^t_j) = \mathcal{O}_j(\boldsymbol{\theta}^t_j, \boldsymbol{z}^t_j) - \boldsymbol{\theta}^t_j$, $k_0 = j$. Here $P_j^{(\rho)}$ denotes the set of all sequences $(k_1, \ldots, k_\rho)$ such that $k_s \in \mathcal{N}^{(1)}_{\text{out}}(k_{s-1})$ for $s = 1, \ldots, \rho$ (see Definition A.7) and $\boldsymbol{H}(\boldsymbol{\theta}^{t+s}_{k_s}; \boldsymbol{z}^{t+s}_{k_s})$ is the Hessian matrix of L with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{t+s}_{k_s}$ and data $\boldsymbol{z}^{t+s}_{k_s}$. For the cases when $\rho = 0$ and $\rho = 1$, the relevant product expressions are defined as identity matrices, thereby ensuring that the r-hop DICE-E remains well-defined. Full proof is deferred to Appendix C.3.

Main Results: Intuition

How can the influence of indirectly connected nodes—such as nodes k_0 to k_5—be quantified?



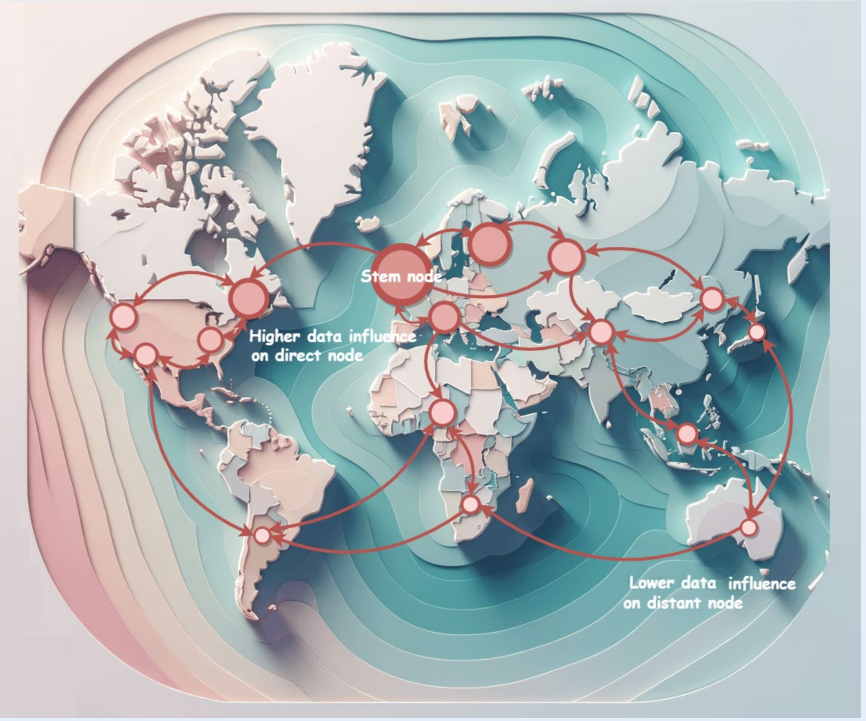
Theorem 2 (Approximation of r-hop DICE-GT). The r-hop DICE-GT influence $\mathcal{I}^{(r)}_{\text{DICE-GT}}(\boldsymbol{z}_j^t, \boldsymbol{z}')$ (see Definition 3) can be approximated as follows:

$$\mathcal{I}^{(r)}_{\text{DICE-E}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = -\sum_{\rho=0}^{r} \sum_{(k_1,\ldots,k_\rho)\in P_j^{(\rho)}} \eta^t q_{k_\rho} \underbrace{\left(\prod_{s=1}^{\rho} \boldsymbol{W}^{t+s-1}_{k_s,k_{s-1}}\right)}_{\text{communication graph-related term}} \underbrace{\nabla L(\boldsymbol{\theta}^{t+\rho}_{k_\rho}; \boldsymbol{z}')^\top}_{\text{test gradient}}$$

$$\times \underbrace{\left(\prod_{s=2}^{\rho} (\boldsymbol{I} - \eta^{t+s-1} \boldsymbol{H}(\boldsymbol{\theta}^{t+s-1}_{k_s}; \boldsymbol{z}^{t+s-1}_{k_s}))\right)}_{\text{curvature-related term}} \underbrace{\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t)}_{\text{optimization-related term}} .$$

where $\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) = \mathcal{O}_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t) - \boldsymbol{\theta}_j^t$, $k_0 = j$. Here $P_j^{(\rho)}$ denotes the set of all sequences $(k_1, \ldots, k_\rho)$ such that $k_s \in \mathcal{N}^{(1)}_{\text{out}}(k_{s-1})$ for $s = 1, \ldots, \rho$ (see Definition A.7) and $\boldsymbol{H}(\boldsymbol{\theta}^{t+s}_{k_s}; \boldsymbol{z}^{t+s}_{k_s})$ is the Hessian matrix of L with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{t+s}_{k_s}$ and data $\boldsymbol{z}^{t+s}_{k_s}$. For the cases when $\rho = 0$ and $\rho = 1$, the relevant product expressions are defined as identity matrices, thereby ensuring that the r-hop DICE-E remains well-defined. Full proof is deferred to Appendix C.3.

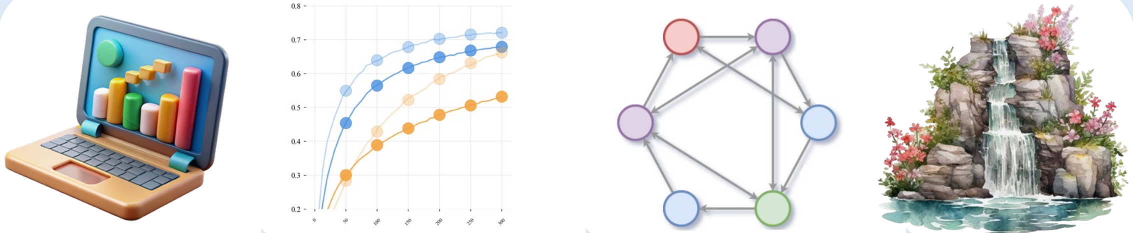
Takeaways



What phenomena does this paper uncover?

In decentralized learning, the influence of data "cascades" through the communication graph, resembling "ripples in water".

This influence is determined by both the original data and the topological position of the data-holding node.



$$\mathcal{I}^{(r)}_{\text{DICE-E}}(\boldsymbol{z}_j^t, \boldsymbol{z}') = -\sum_{\rho=0}^{r} \sum_{(k_1,\ldots,k_\rho)\in P_j^{(\rho)}} \eta^t q_{k_\rho} \underbrace{\left(\prod_{s=1}^{\rho} \boldsymbol{W}_{k_s,k_{s-1}}^{t+s-1}\right)}_{\text{communication graph-related term}} \underbrace{\nabla L(\boldsymbol{\theta}_{k_\rho}^{t+\rho}; \boldsymbol{z}')^\top}_{\text{test gradient}}$$

$$\times \underbrace{\left(\prod_{s=2}^{\rho} (\boldsymbol{I} - \eta^{t+s-1} \boldsymbol{H}(\boldsymbol{\theta}_{k_s}^{t+s-1}; \boldsymbol{z}_{k_s}^{t+s-1}))\right)}_{\text{curvature-related term}} \underbrace{\Delta_j(\boldsymbol{\theta}_j^t, \boldsymbol{z}_j^t)}_{\text{optimization-related term}} .$$

Collaborators

Many thanks to collaborators again!

Tongtian Zhu

Zhejiang University

Wenhao Li

Zhejiang University

Can Wang

Zhejiang University

Fengxiang He

The University of Edinburgh

Thank you!

DICE: Data Influence Cascade in Decentralized Learning
https://openreview.net/forum?id=2TIYkqieKw

Contact: raiden@zju.edu.cn (Tongtian Zhu)