NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Efficient Federated-Learning Model Debugging

Anran Li

Research Fellow

Nanyang Technological University

2022.12.12

Outline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Outline

Background

Problem Description

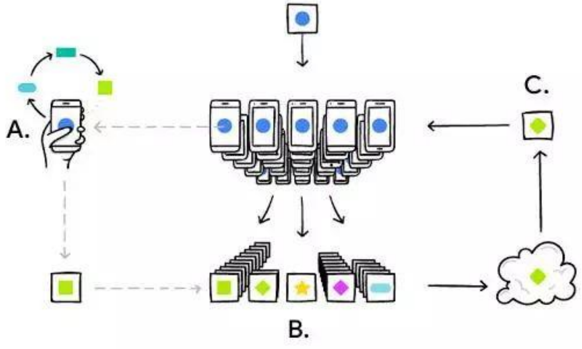
System Design

Evaluation

Conclusion

Federated Learning

- Federated learning (FL)
  - Local data, multiple clients cooperation
  - Privacy preserving, bandwidth saving



Overview of FL

- Data sources in FL

Crowdsourcing

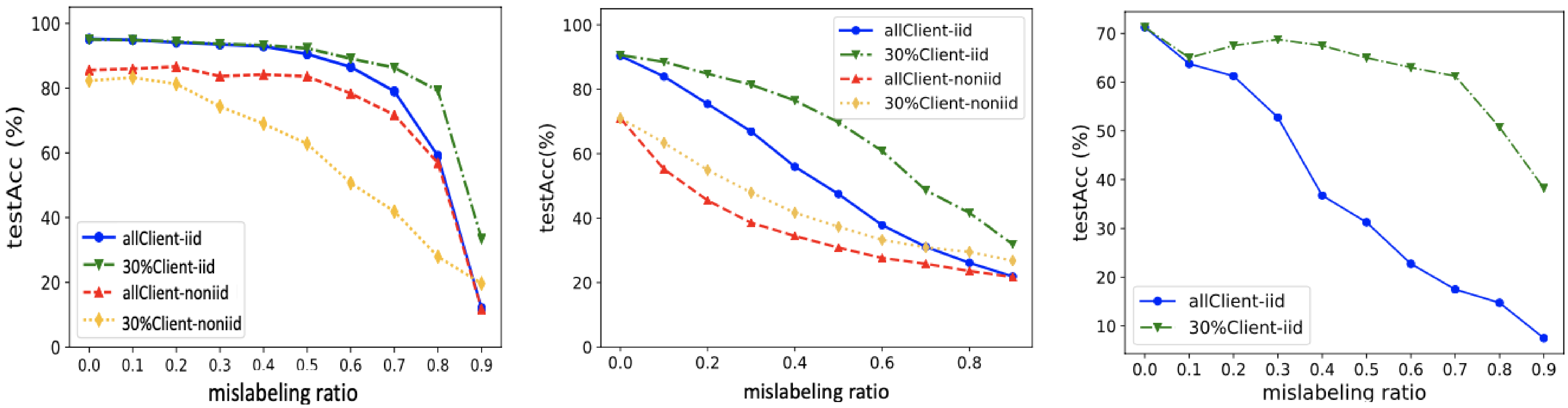
Web Crawling

Low-quality data leads to serious consequences

Low-quality Data in FL

Local erroneous data results low-performance global FL models

Test Accuracy v.s Mislabeling Ratio



(a) FedAVG-MNIST (b) FedAVG-CIFAR (c) FedAVG-ESC

Our Goal

Towards erroneous data in FL clients.

An Efficient Data-based Model Debugging or Interpretation for Federated Learning.

Model debugging: giving explanations of model prediction results (ICML 2017).

Existing Work

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Model Debugging	Sample-level	Client-level
Centralized Learning	Through perturbing a subset of the data samples [CVPR16, S&P16]. Analyzing the influence of data samples on the model's predictions [ICML19, NIPS19]	--
Federated Learning	--	Client contribution to FL models [AAAI21, BigData19]

Direct access to data

High computation/ communication cost

Privacy concerns

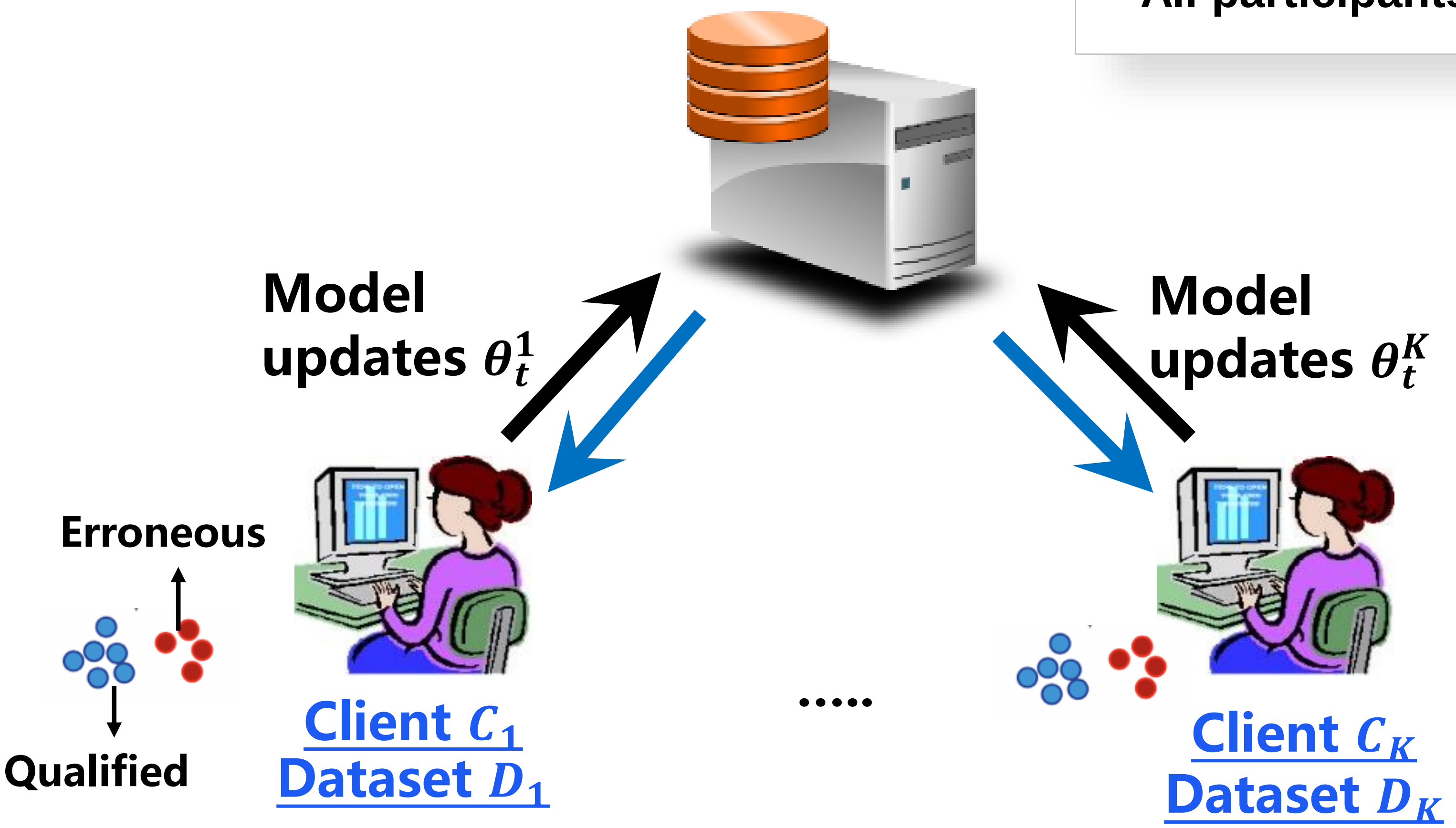
Lack of efficient sample-based model debugging or interpretation methods for FL models.

Outline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Problem Description

All participants are semi-honest



Efficient Federated-learning Model Debugging from the perspective of data

Influence Function for FL

- Influence functions were proposed to avoid retraining the model by providing a first-order approximation to the actual effect
- The parameter change after removing sample $z_{k,i}$ from client C_k.

$$I_f(w_k) \approx \nabla_\theta^\top f(\hat{\theta}(\mathbf{1}))(\frac{1}{K}\sum_{k=1}^{K} H_k + \lambda I)^{-1} g_{\hat{\theta},f}(w_k),$$

$$g_{\hat{\theta},f}(w_k) = \sum_{k=1}^{K} \sum_{i\in\mathcal{P}_k} w_{k,i} \nabla_\theta L(z_{k,i};\hat{\theta}) \qquad H_k = \frac{1}{n_k}\sum_{i\in\mathcal{P}_k} \nabla_\theta^2 L(z_{k,i};\hat{\theta})$$

Opportunities and Insights

- Insight 1: The influence function for FL has an additive property when measuring the change in test predictions
  - If $w_k = w_{k,1} + w_{k,2}$
  - Then $I_f(w_k) = I_f(w_{k,1}) + I_f(w_{k,2})$

Opportunities and Insights

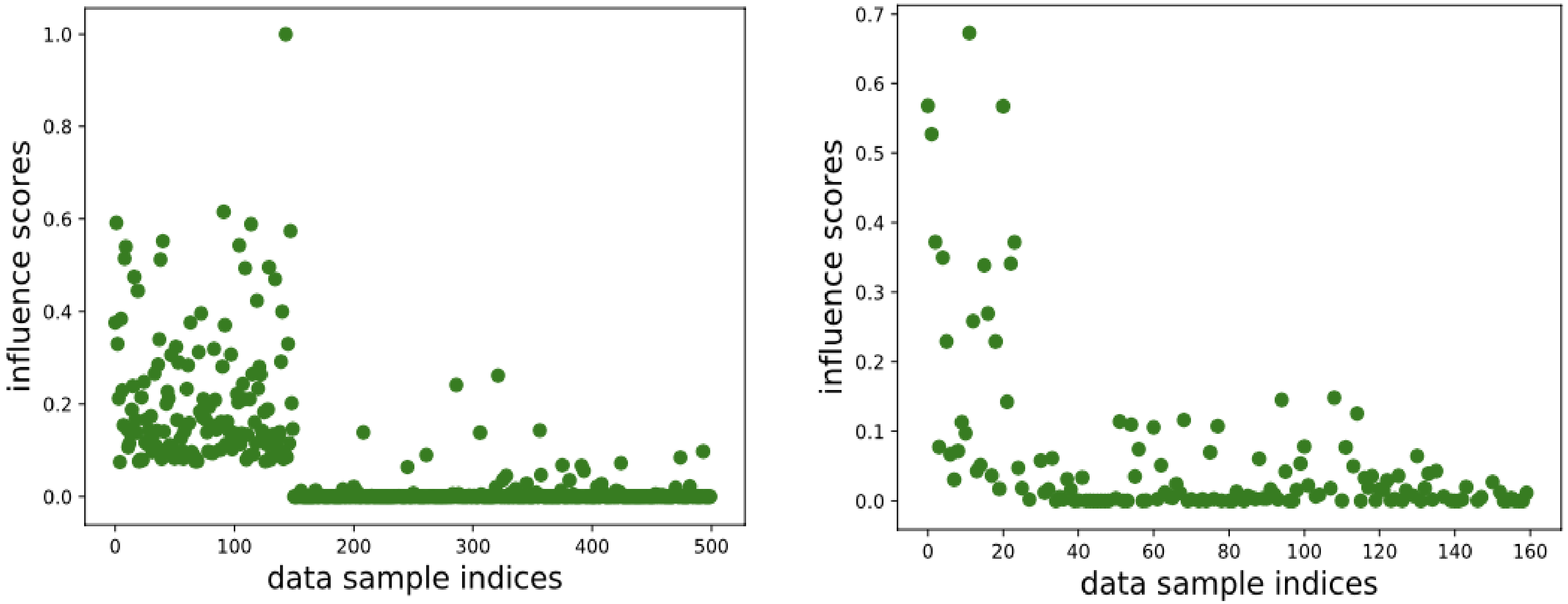
- Insight 1: The influence function for FL has an additive property when measuring the change in test predictions
  - If $w_k = w_{k,1} + w_{k,2}$

Enlighten hierarchical influence analysis: identifies influential clients first to save large cost for sample-level influence analysis.

Opportunities and Insights

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

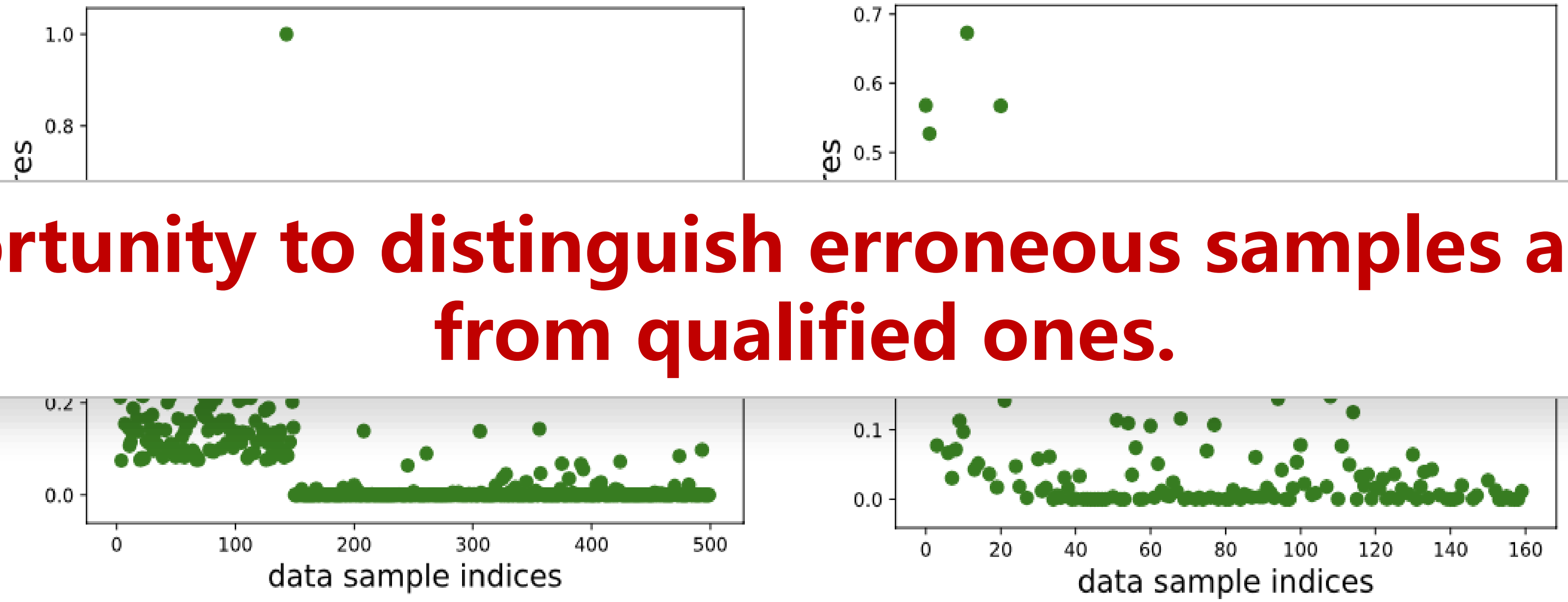
- Insight 2: When there are more qualified training samples than erroneous training samples, erroneous samples have obviously larger absolute influence values than qualified ones



Influence values for erroneous and qualified samples

Opportunities and Insights

- Insight 2: When there are more qualified training samples than erroneous training samples, erroneous samples have obviously larger absolute influence values than qualified ones



An opportunity to distinguish erroneous samples and clients from qualified ones.

Influence values for erroneous and qualified samples

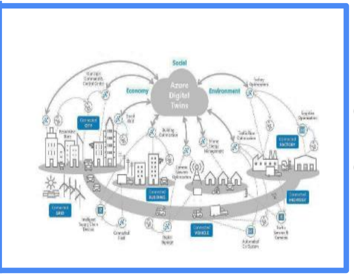
Challenging Issues and Solutions

Data privacy

Privacy-preserving

Hierarchical influence analysis

Restricted resources



Efficient

Training log based

Heterogeneous system

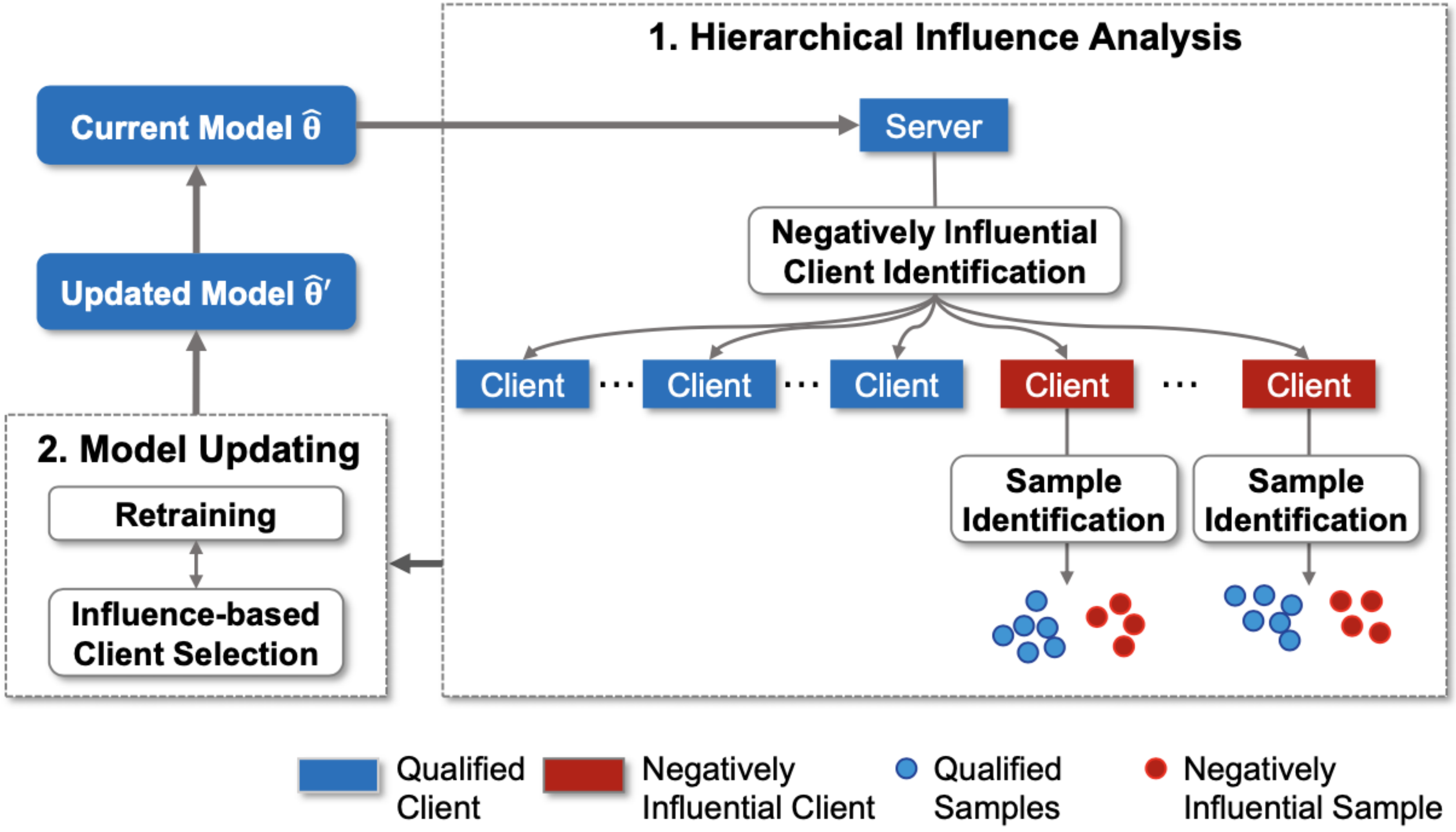
Adaptive

Two identification methods

Outline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

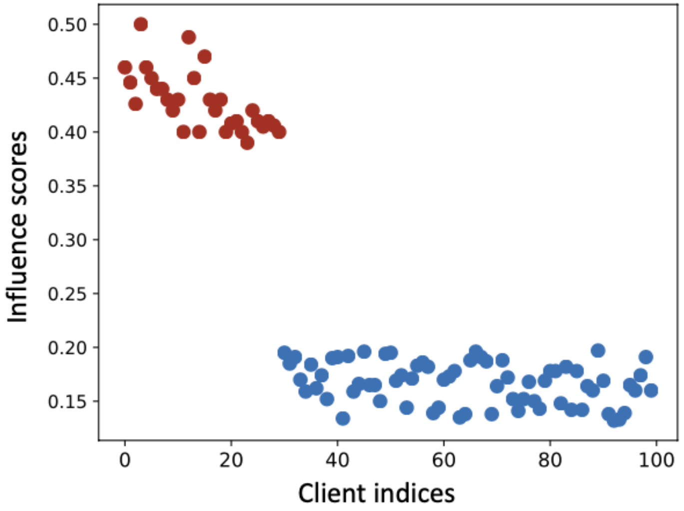
System Design



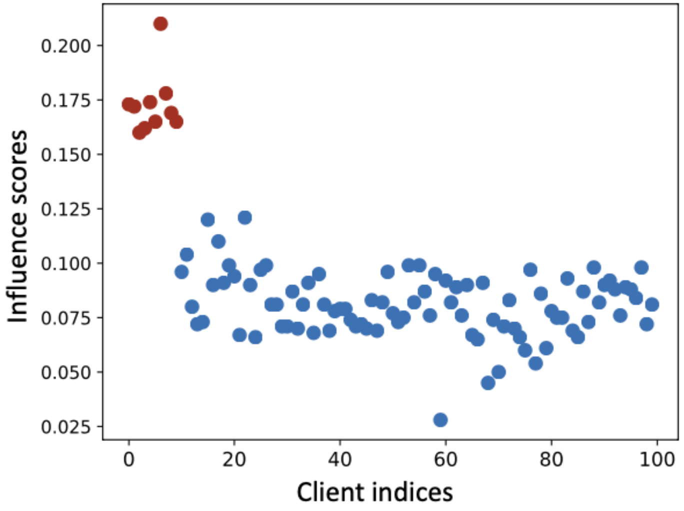
System Overview

Basic method of client identification

- Influential clients have obviously larger absolute influence values than qualified ones
  - ➢ Sum up influence values of all its samples



(a) The first 30 clients possess mislabeled data.



(b) The first 10 clients possess noisy data).

Large Computation cost

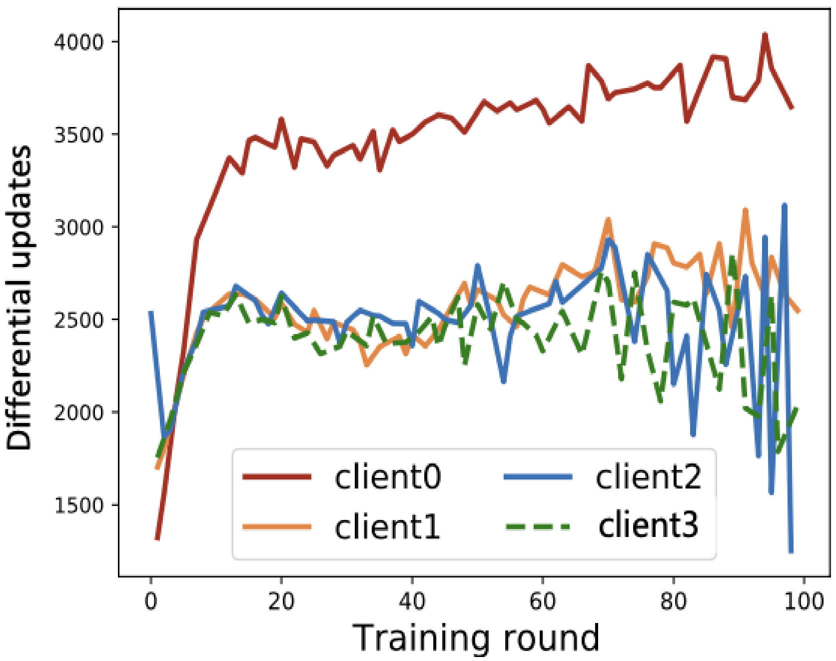
Training log based client identification

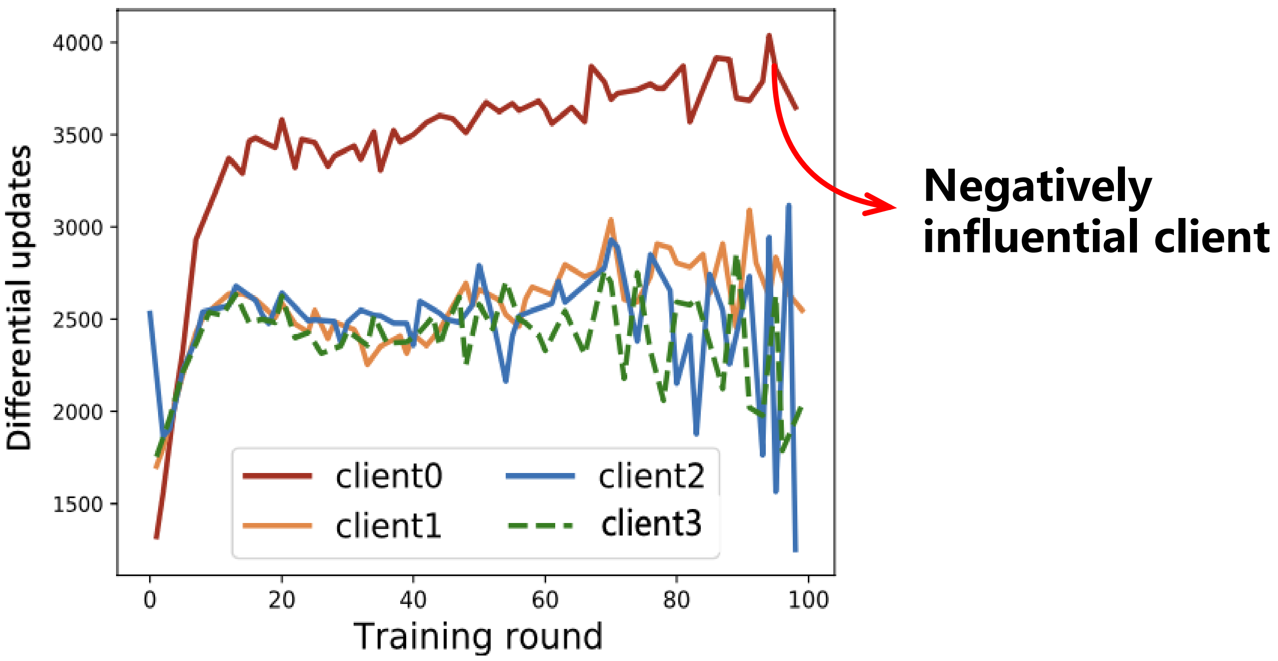
- A client is negatively influential if his distance is significantly greater than the median distances of all clients

$$\frac{D_k}{median\{D_l | l \in [K]\}} > \delta_T \qquad D_k = \frac{1}{N(k)} \sum_{t=\frac{T}{2}}^{T} s_t^k ||\theta_t^k - \theta_t||$$

Training log based client identification

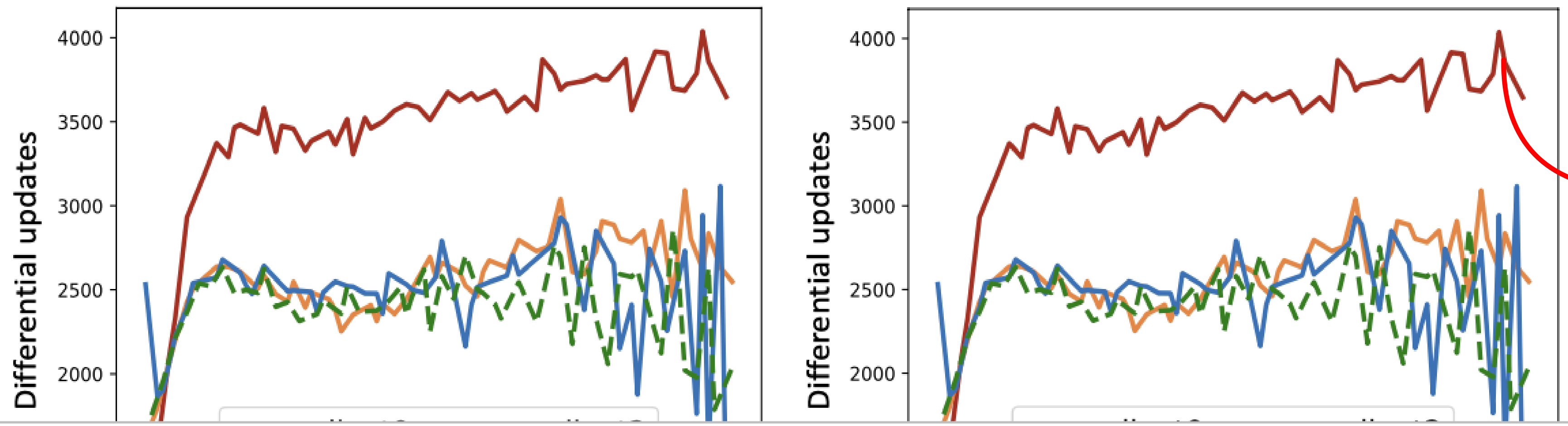
- A client is negatively influential if his distance is significantly greater than the median distances of all clients





Training log based client identification

- A client is negatively influential if his distance is significantly greater than the median distances of all clients



Negatively influential client

Dramatically saves both computation and communication cost by orders of magnitude

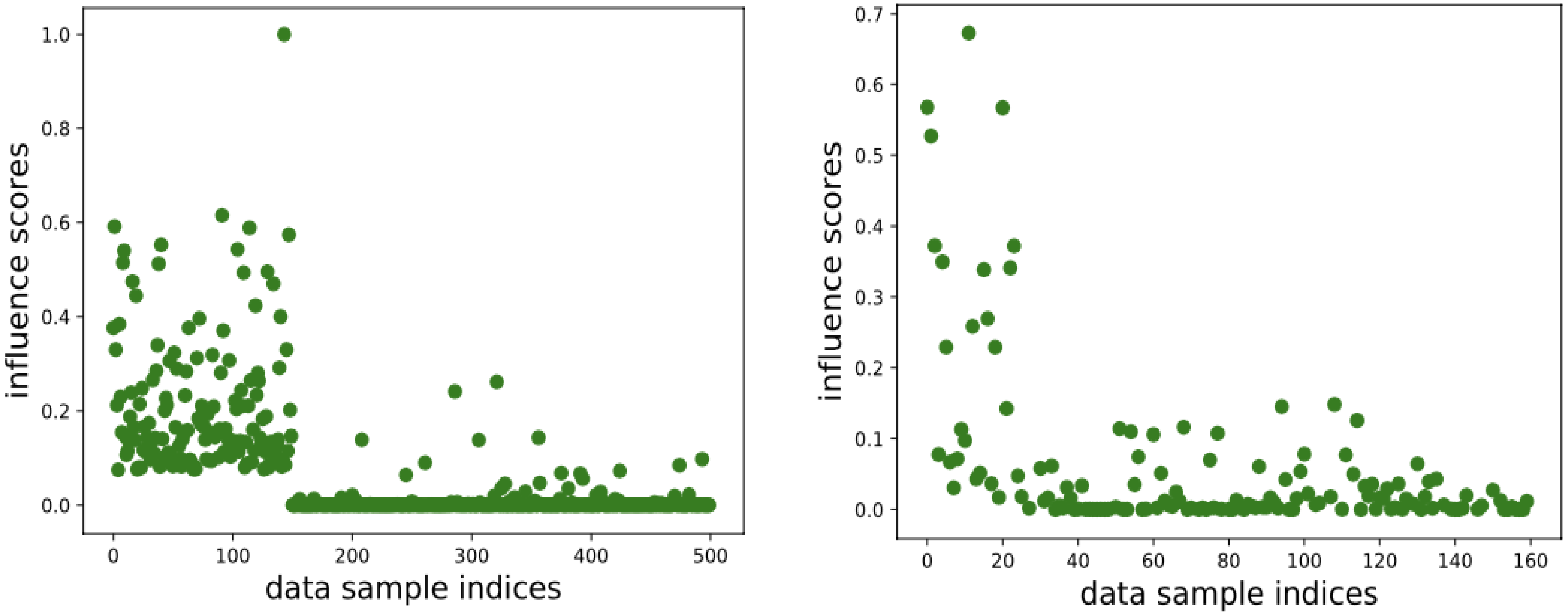
Basic method sample identification

- A sample is negatively influential if its influence value is significantly greater than the median influence values

$$\frac{I_f(z_{k,i})}{median\{I_f(z_{k,j})|z_{k,j} \in \bigcup_{C_k \in C_N} D_k} > \delta_S$$

Basic method sample identification

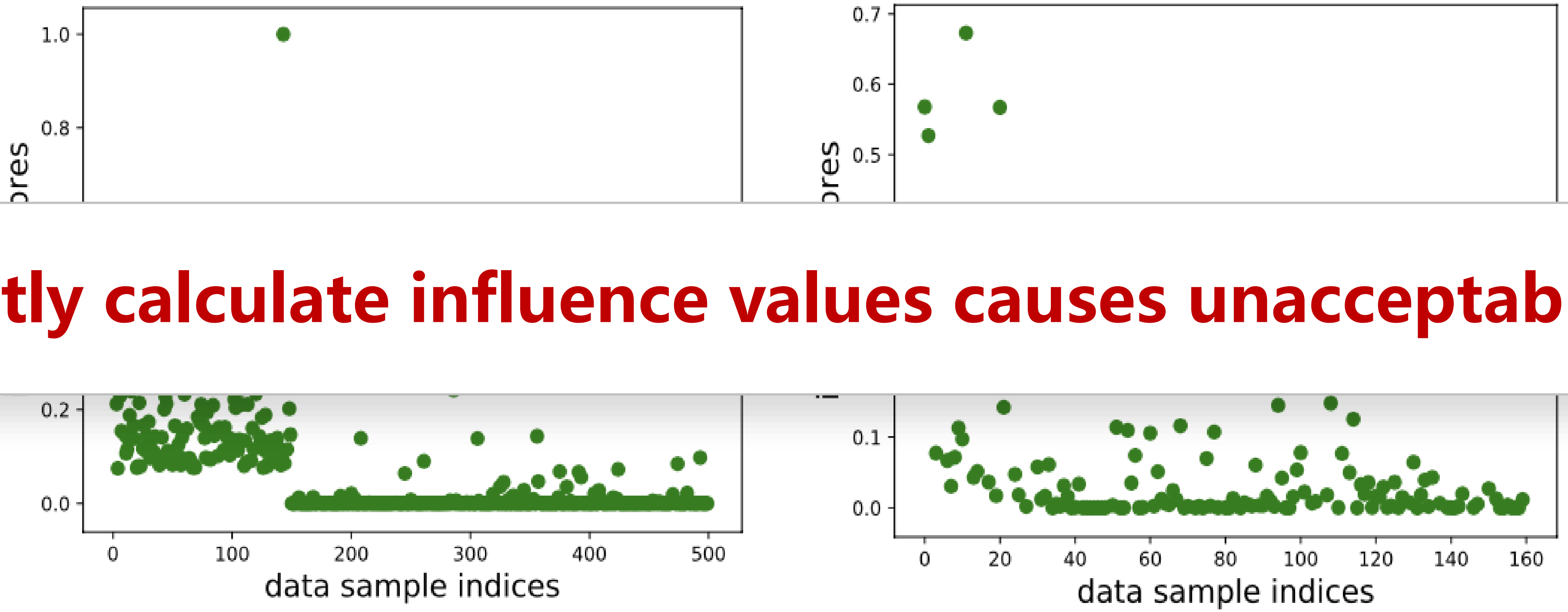
- A sample is negatively influential if its influence value is significantly greater than the median influence values



Influence values for erroneous and qualified samples

Basic method sample identification

- A sample is negatively influential if its influence value is significantly greater than the median influence values



Directly calculate influence values causes unacceptable cost

Influence values for erroneous and qualified samples

Computation-efficient sample identification

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- Use Hessian-vector products (HVP) to approximate $\left(\frac{1}{K}\sum_{k=1}^{K} H_k + \lambda I\right)^{-1} \nabla_\theta L(z_{test};\hat{\theta})$

$\boldsymbol{O(p^3)}$

$$H_j^{-1} = \sum_{i=0}^{j} (I-H)^i \xrightarrow{①} H_j^{-1} = I + (I-H)H_{j-1}^{-1}$$

$\boldsymbol{O(p^2)}$

$$\Big\downarrow ②$$

$$\begin{bmatrix} \frac{\partial^2 L}{\partial^2 \theta_1}, \dots, & \frac{\partial^2 L}{\partial \theta_p \partial \theta_1} \\ \dots & \dots \\ \frac{\partial^2 L}{\partial \theta_1 \theta_p}, \dots, & \frac{\partial^2 L}{\partial^2 \theta_p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix}$$

$\boldsymbol{O(p^2)}$

$$\xleftarrow{③}$$

$$\frac{\partial\left(\frac{\partial L}{\partial \theta_1} a_1, \dots, \frac{\partial L}{\partial \theta_p} a_p\right)}{\partial \theta}$$

$\boldsymbol{O(p)}$

Computation-efficient sample identification

- Use batch to decrease communication cost

for *each round* $j = 1, 2, \ldots r$ **do**

  The server uniformly selects a client $C_i, i \in [K]$ and sends $x_{j-1}, \hat{\theta}$ to client C_i

  Client C_i randomly selects $\lceil \xi n_i \rceil$ samples from $\mathcal{D}_i$; computes $x_j = v + \sum_{s=1}^{\lceil \xi n_i \rceil} (I - \nabla_\theta^2 L(z_{i,s}, \hat{\theta})) x_{j-1} / \lceil \xi n_i \rceil$; sends x_j to the server

Complexity Analysis

Model Debugging	Computation cost	Communication cost
Strawman method	$\boldsymbol{O(np^2 + p^3)}$	$\boldsymbol{O(Kp^2)}$
Computation-efficient method	$\boldsymbol{O(np)}$	$\boldsymbol{O(rp)}$

Complexity greatly reduced!

Communication-saving sample identification

- View the calculation of $H^{-1}v$ as the optimization problem $min_x||Hx - v||^2$

$$x_j = x_{j-1} + \frac{v_l - h_l x_{j-1}}{||h_l||^2} h_l$$

  - Uniformly sample one client at each step, use batch to estimate h_l

for *each round* $j = 1, 2, \ldots r_1$ **do**

  The server randomly selects l from the set $\{1, 2, \cdots, p\}$; uniformly selects a client $C_i, i \in [K]$; sends l to client C_i;

  Client C_i calculates h_l using all his/her samples; sends h_l to the server

  The server computes x_{j+1} using Eq. (10)

Complexity Analysis

Model Debugging	Computation cost	Communication cost
Strawman method	$O(np^2 + p^3)$	$O(Kp^2)$
Communication-saving method	$O(np)$	$O(Kp)$

Complexity greatly reduced!

Influence-based client selection and model updating

- Dynamically selects clients to participate according to their influences

$$P_{t+1}^{k} = \frac{n_k||\theta_t^k - \theta_t||}{\sum_{C_k \in S_t} n_k||\theta_t^k - \theta_t||} \times \sum_{C_k \in S_t} P_t^k$$

Assign clients with larger influence on the current global model higher probabilities

Privacy Analysis during Model Debugging

No local training data transmitted during training, debugging process and updating process

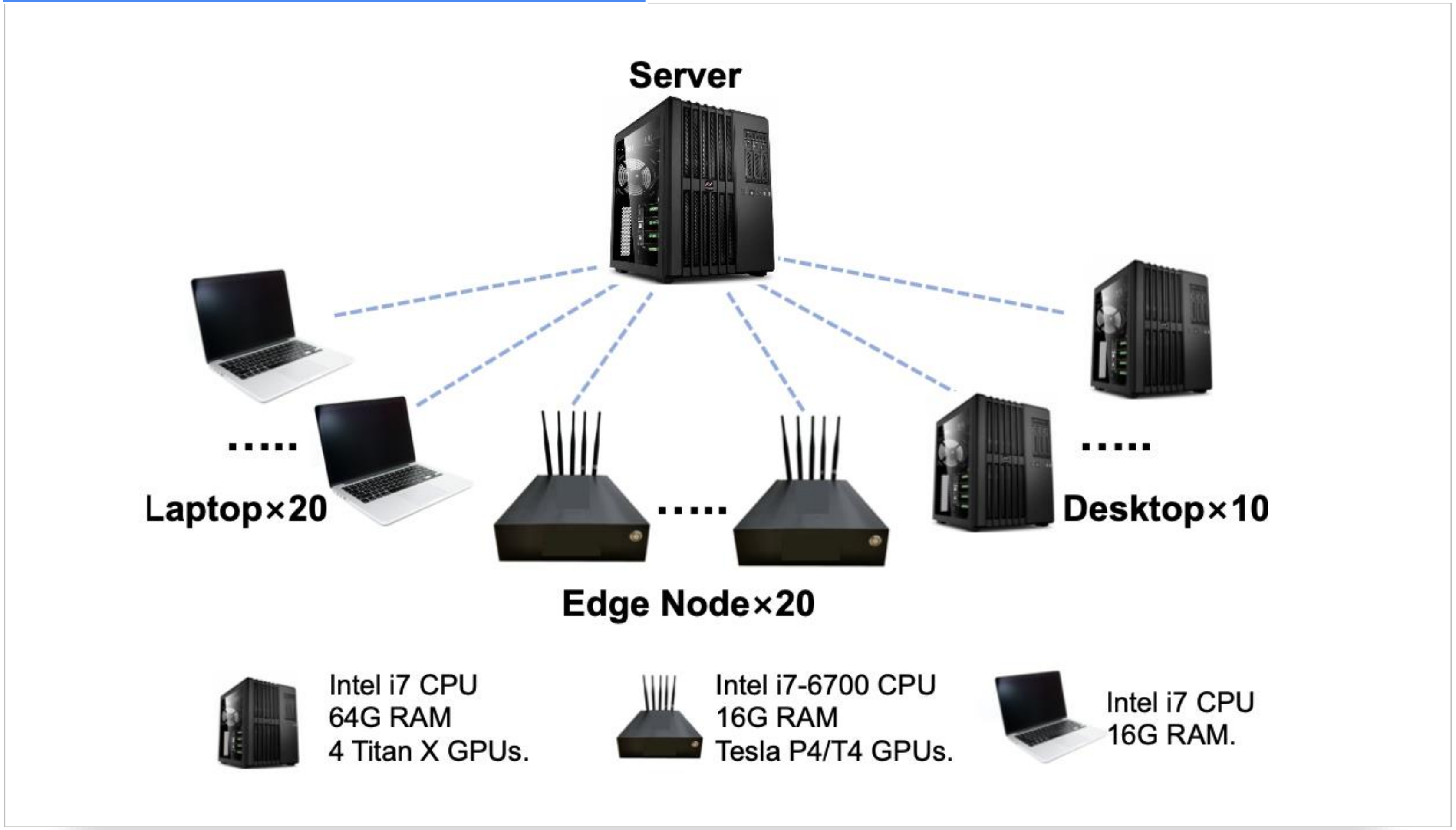
The transmitted second order information cannot be used to infer the local training data

Outline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Experiment Configuration

System Deployment



Experiment Configuration

Datasets

Modality	Notation	Size	Description
Image	$\mathcal{D}_M$	60,000	original training data of MNIST
	$\mathcal{D}_M^T$	10,000	original test data of MNIST
	$\mathcal{D}_M^m$	60,000	$\mathcal{D}_M$ with 9%-40% mislabeled samples
	$\mathcal{D}_C^n$	50,000	$\mathcal{D}_M$ with 9%-40% noisy samples
	$\mathcal{D}_C$	50,000	original training data of CIFAR10
	$\mathcal{D}_C^T$	10,000	original test data of CIFAR10
	$\mathcal{D}_C^m$	50,000	$\mathcal{D}_C$ with 9%-40% mislabeled samples
	$\mathcal{D}_C^n$	50,000	$\mathcal{D}_C$ with 9%-40% noisy samples
	$\mathcal{D}_R^m$	100,000	REAL dataset with 9% mislabeled samples
	$\mathcal{D}_R^T$	10,000	clean test dataset of REAL
	$\mathcal{D}_O^m$	10,000	MOTOR dataset with 9% noisy samples
	$\mathcal{D}_O^T$	1,000	clean test dataset of MOTOR
Audio	$\mathcal{D}_E$	320	original training data of ESC10
	$\mathcal{D}_E^T$	80	original test data of ESC10
	$\mathcal{D}_E^m$	320	$\mathcal{D}_E$ with 9%-40% mislabeled samples

Experiment Configuration

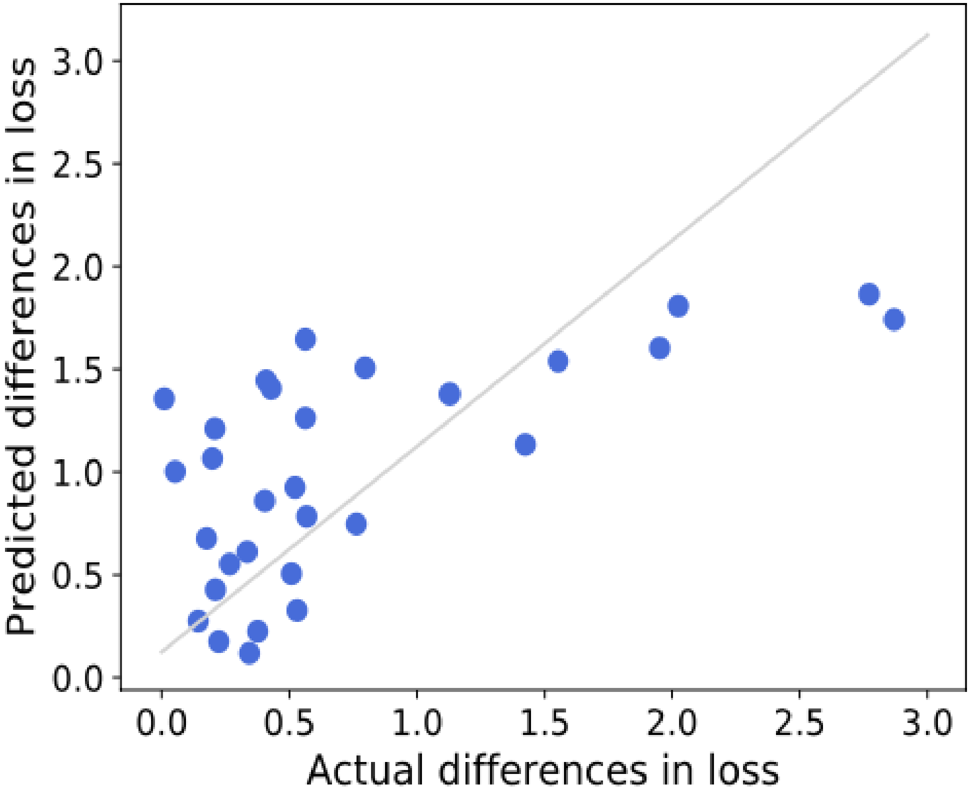
FL Models

Model	# of para	Task
FedAVG-MNIST [25]	$1,663,370$	digit number recognition
FedAVG-CIFAR [22]	$11,173,962$	image recognition
FedAVG-REAL [22]	$11,419,722$	image recognition
FedAVG-MOTOR [22]	$11,219,010$	image recognition
FedAVG-ESC [26]	$22,017,322$	environment classification

Settings of FL Models

Model	Dataset	r_m or r_n	# Clients	# NI-clients
FedAVG-MNIST	$\mathcal{D}_M^m$	$r_m = 9\%$	50	15
FedAVG-MNIST	$\mathcal{D}_M^n$	$r_n = 10\%$	50	15
FedAVG-CIFAR	$\mathcal{D}_C^m$	$r_m = 9\%$	15/50	3/15
FedAVG-CIFAR	$\mathcal{D}_C^n$	$r_n = 10\%$	15	3
FedAVG-REAL	$\mathcal{D}_R^m$	$r_m = 9\%$	10	3
FedAVG-MOTOR	$\mathcal{D}_O^m$	$r_m = 9\%$	10	3
FedAVG-ESC	$\mathcal{D}_E^m$	$r_m = 9\%$	4	1

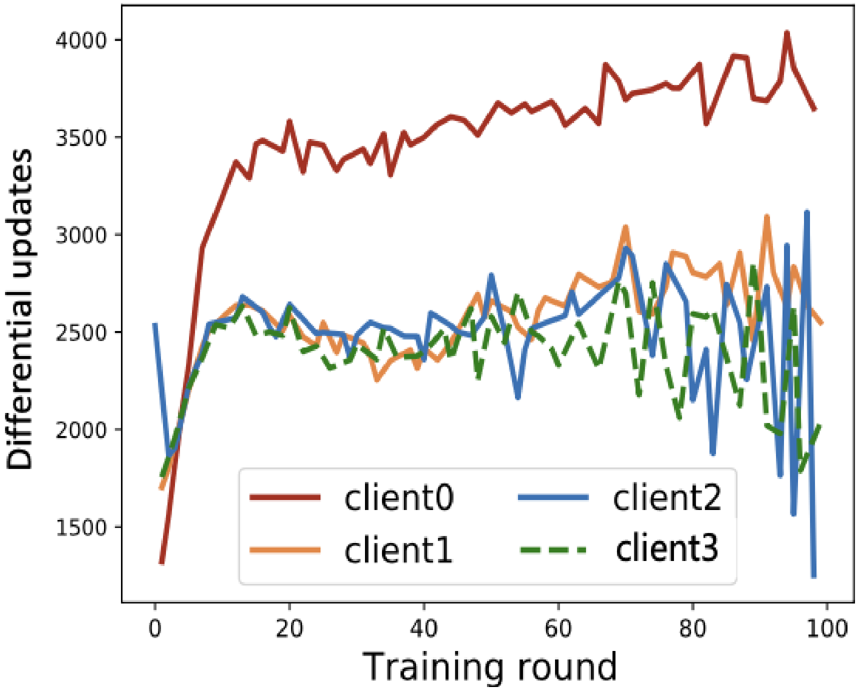
Influence function for FL vs. Leave-some-out retraining

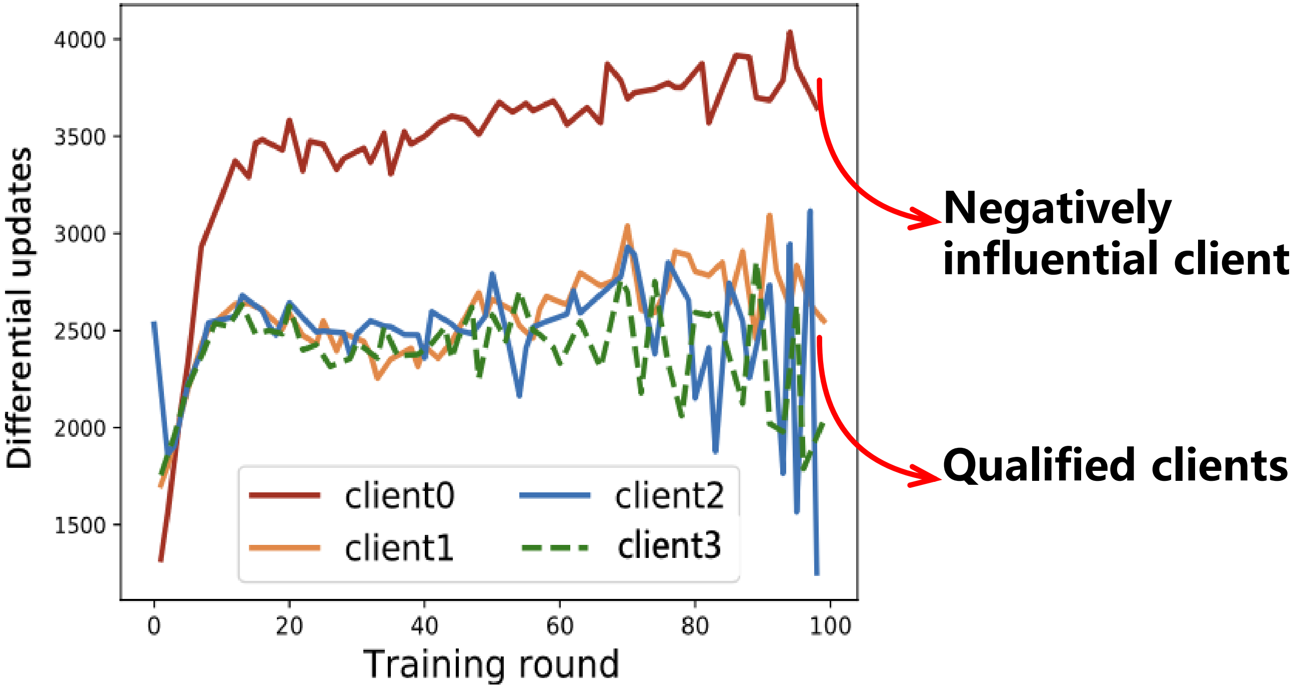


The predicted influences and actual changes in loss are correlated, with PCC=0.6

Identifying negatively influential clients

Accuracy of training log based identification

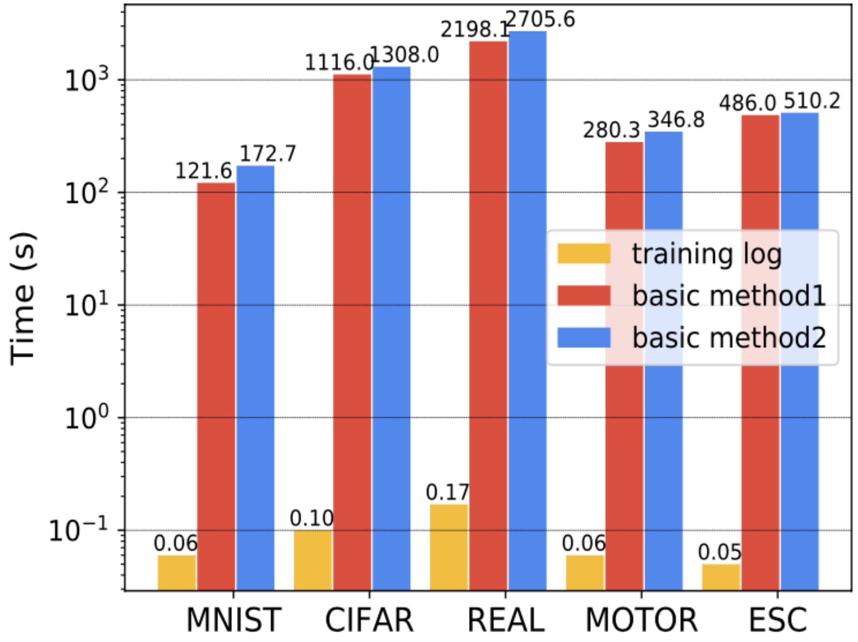




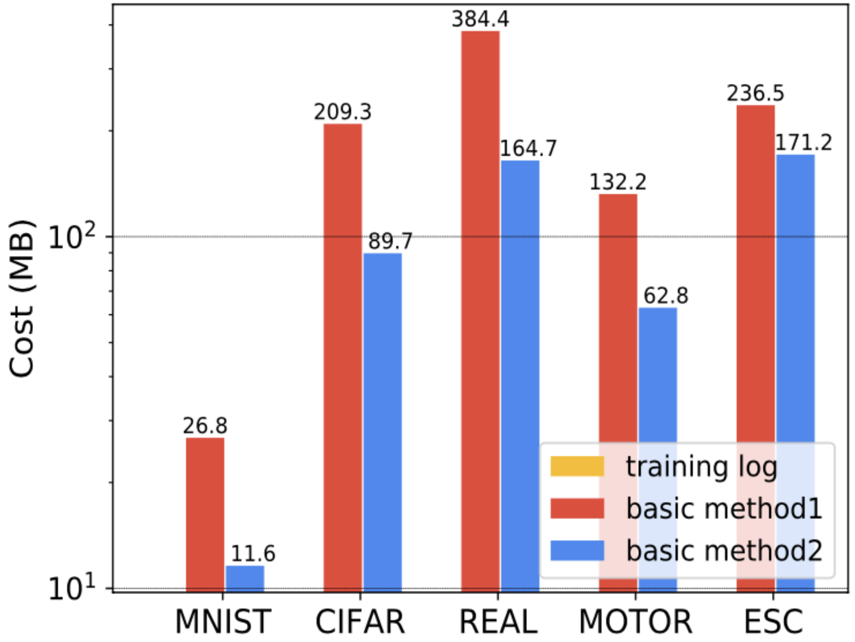
Accuracy, precision and recall are all 100% with threshold $\delta_T = 1.50$

Identifying negatively influential clients

Efficiency of training log based identification



(a) Computation cost.

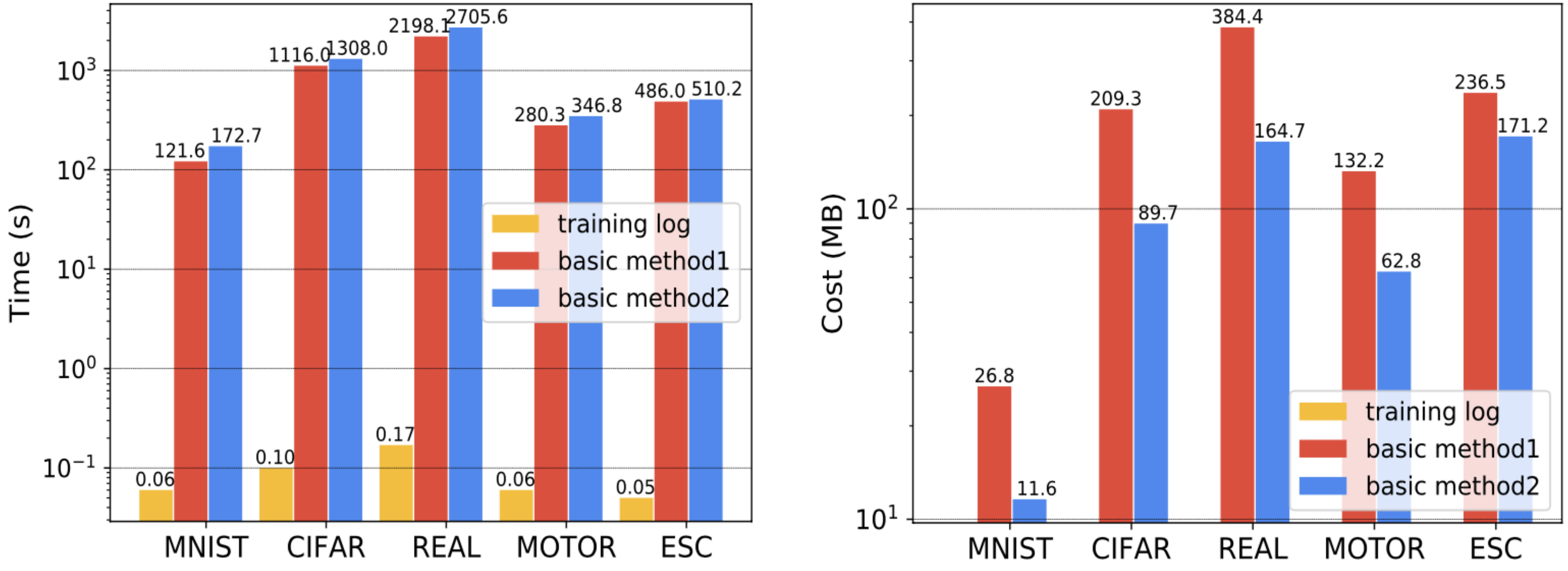


(b) Communication cost.

The runtime of the training log method is 0.1s, while 1116.0s and 1308.0s for two basic methods

Identifying negatively influential clients

Efficiency of training log based identification



(a) Computation cost.

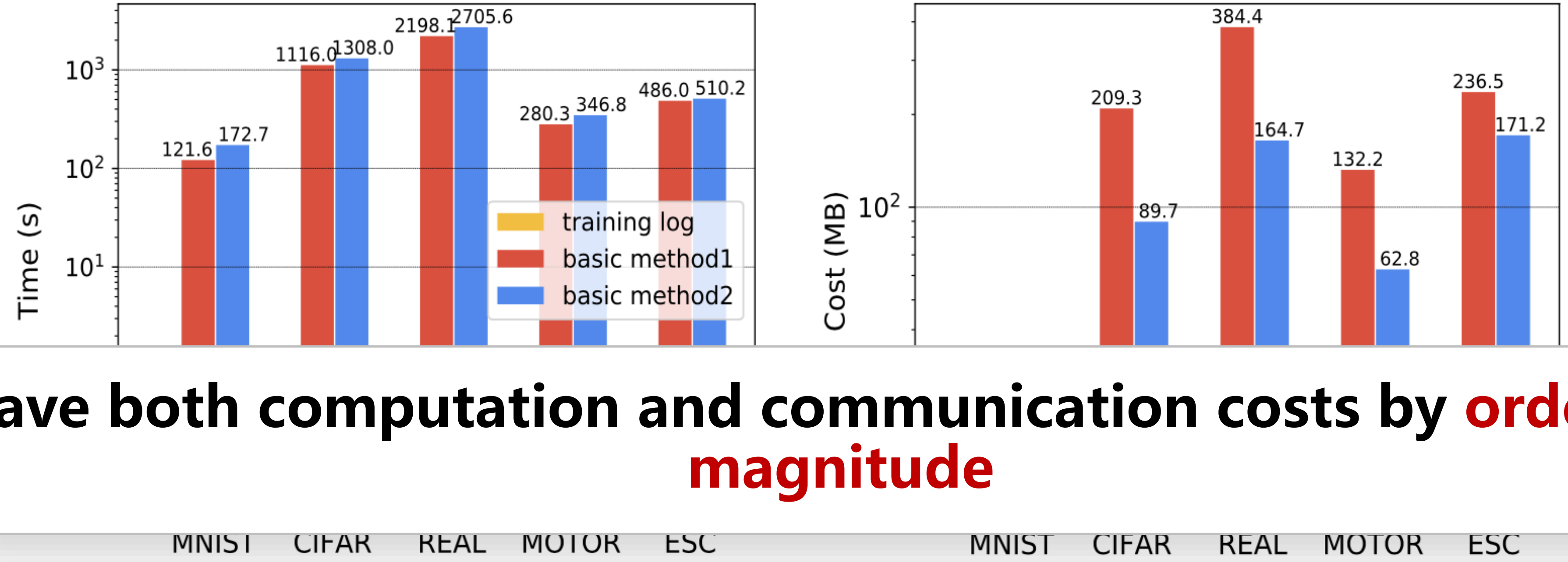
(b) Communication cost.

No communication cost of training log method, while 209.3MB and 89.7MB for two basic methods

Identifying negatively influential clients

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Efficiency of training log based identification



(a) Computation cost. (b) Communication cost.

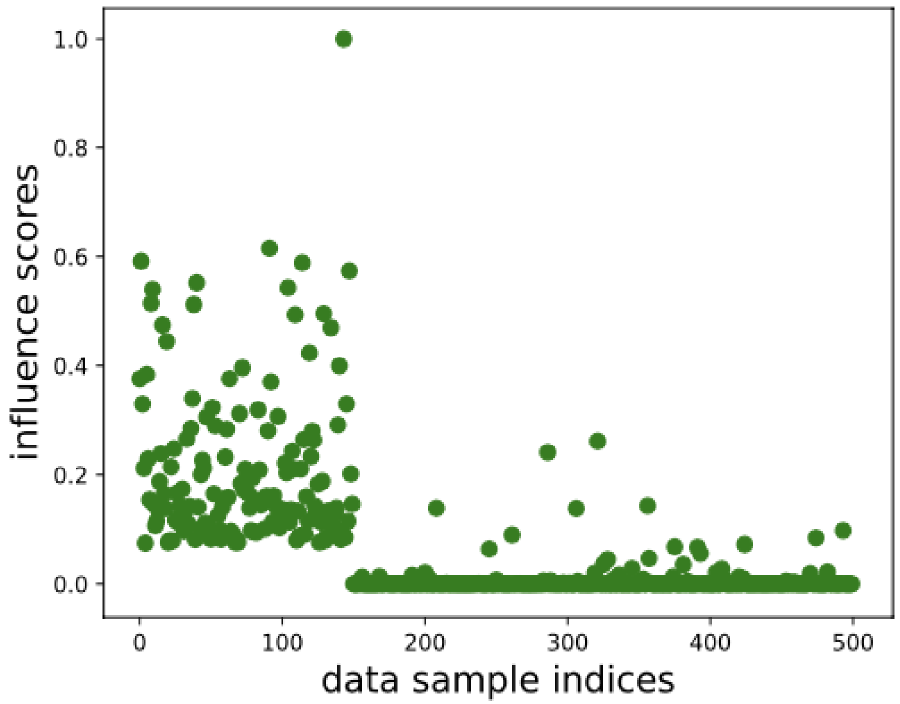
Save both computation and communication costs by orders of magnitude

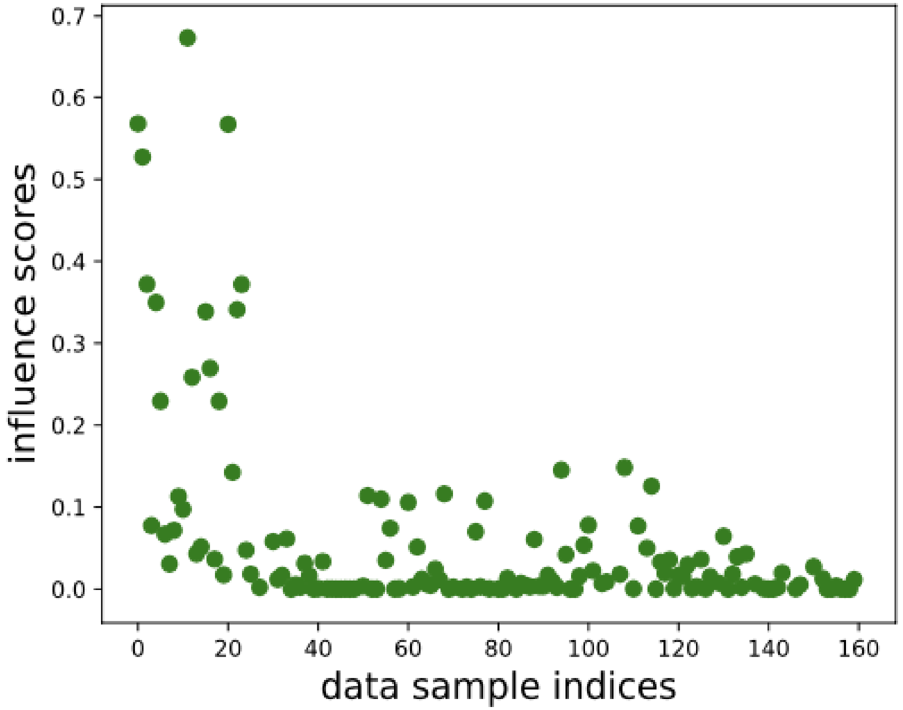
No communication cost of training log method, while 209.3MB and 89.7MB for two basic methods

Identifying negatively influential samples

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Accuracy





Influence values of data samples calculated using Algorithm 1

Identifying negatively influential samples

Accuracy

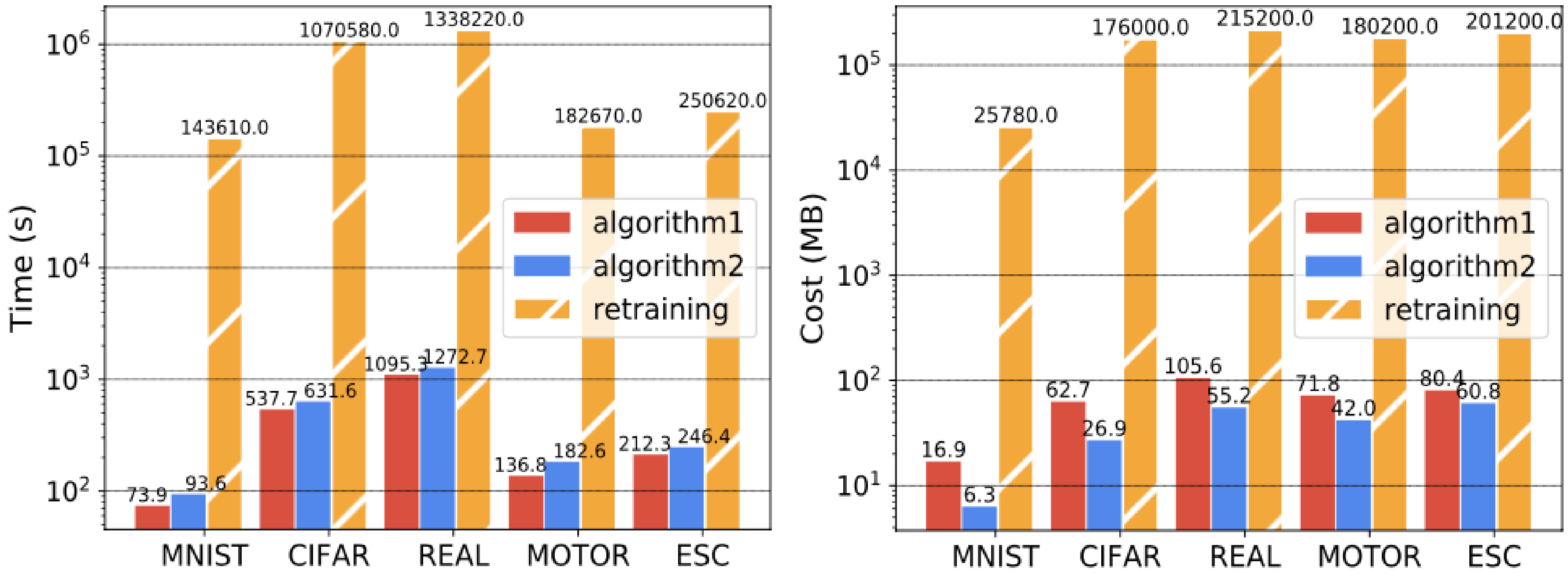
Dataset (# clients)	Algorithm	Accuracy	Precision	Recall
$\mathcal{D}_M^m(50)$	Algorithm 1	91.0%	90.5%	93.2%
	Algorithm 2	92.5%	94.0%	90.8%
$\mathcal{D}_M^n(50)$	Algorithm 1	94.2%	92.0%	91.0%
	Algorithm 2	92.1%	90.3%	92.3%
$\mathcal{D}_C^m(15)$	Algorithm 1	90.5%	75.3%	90.4%
	Algorithm 2	90.1%	77.0%	91.2%
$\mathcal{D}_C^m(50)$	Algorithm 1	82.1%	70.1%	82.0%
	Algorithm 2	83.0%	71.0%	81.4%
$\mathcal{D}_C^n(15)$	Algorithm 1	89.2%	78.5%	92.3%
	Algorithm 2	88.6%	76.4%	90.7%
$\mathcal{D}_R^m(10)$	Algorithm 1	84.0%	70.3%	80.0%
	Algorithm 2	85.1%	71.6%	81.2%
$\mathcal{D}_O^m(10)$	Algorithm 1	80.1%	62.0%	80.0%
	Algorithm 2	82.5%	64.1%	81.8%
$\mathcal{D}_E^m(4)$	Algorithm 1	81.3%	72.3%	93.0%
	Algorithm 2	72.0%	73.1%	92.6%

Achieve fairly high accuracy in all settings

Identifying negatively influential samples

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Efficiency

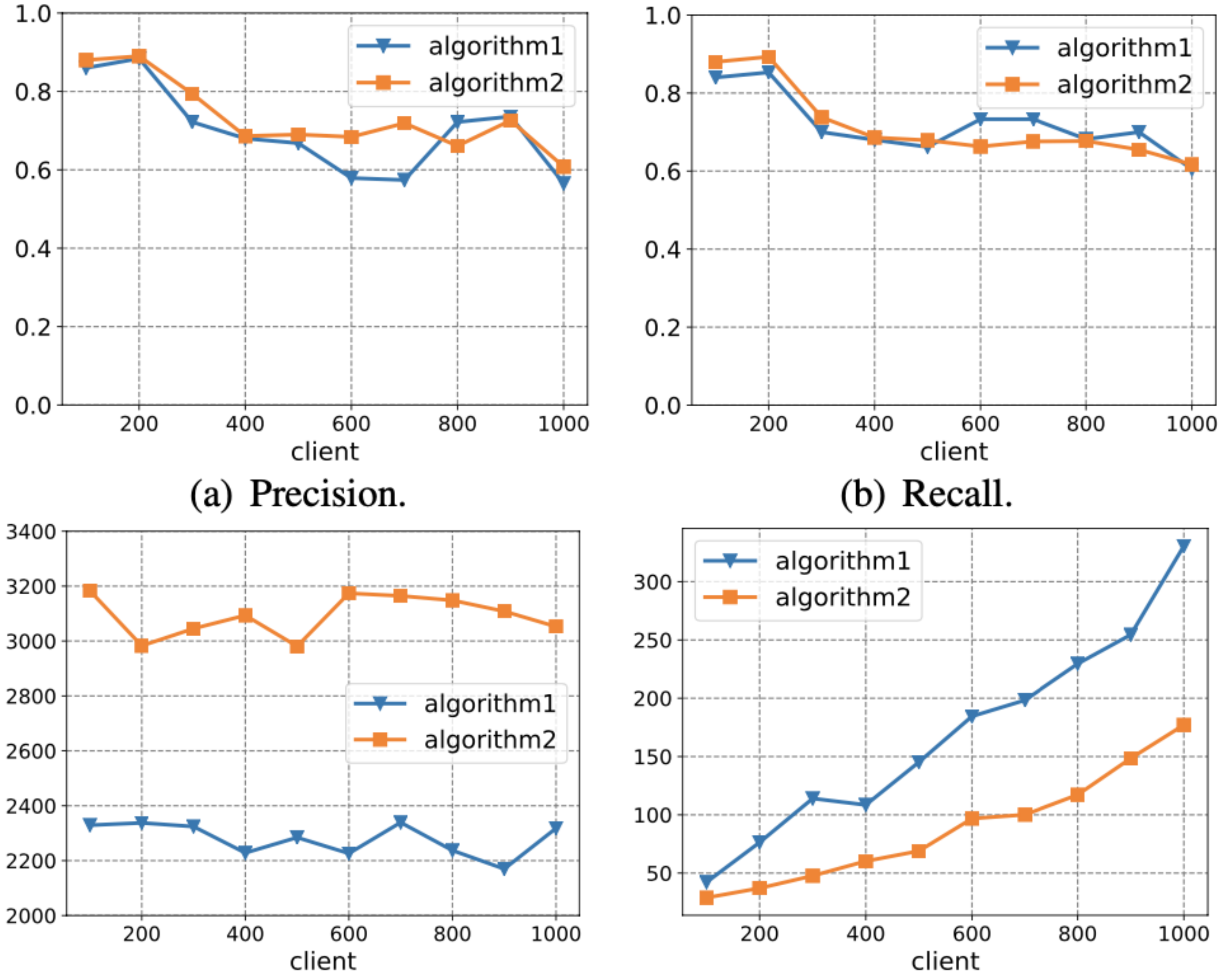


(a) Computation cost.

(b) Communication cost.

The costs are orders of magnitude lower, e.g., less than 0.051% computation cost, 0.060% communication cost

Large-scale simulations



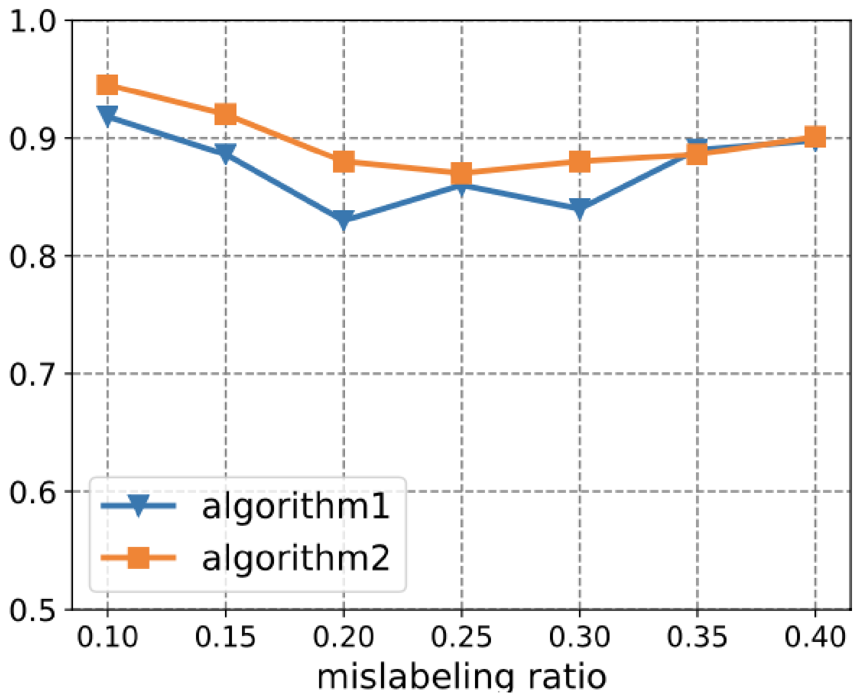
(a) Precision. (b) Recall.

(c) Computation time (s) (d) Communication cost (MB)

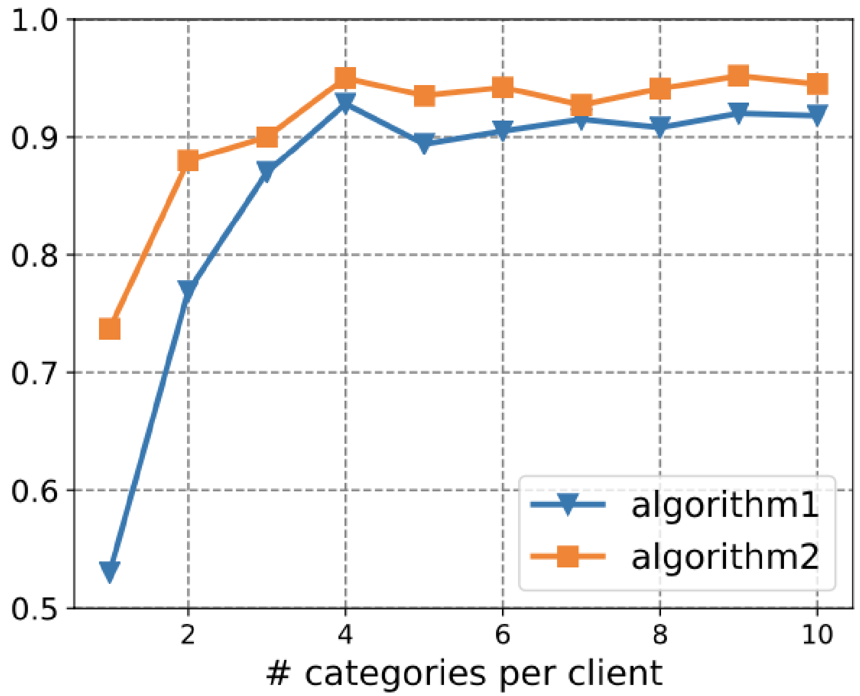
Scalable and robust in large environments

Large-scale simulations

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



Identification accuracy for different mislabeling ratios



Identification accuracy for different Non-IID scenarios

Scalable and robust in large environments

Outline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

1. Present the framework FLDebugger to accomplish both debugging and interpretability of FL models from the perspective of training data.

2. Design a hierarchical negatively influential clients and samples identification method with around 90% accuracy.

3. Design influence-based clients selection retraining method to facilitate the model training in terms of higher accuracy and faster convergence.

Thanks!

Email: anran.li@ntu.edu.sg