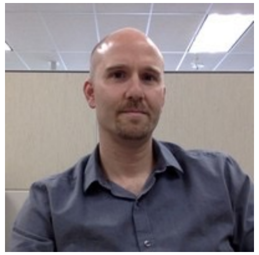
NUS National University of Singapore

intel

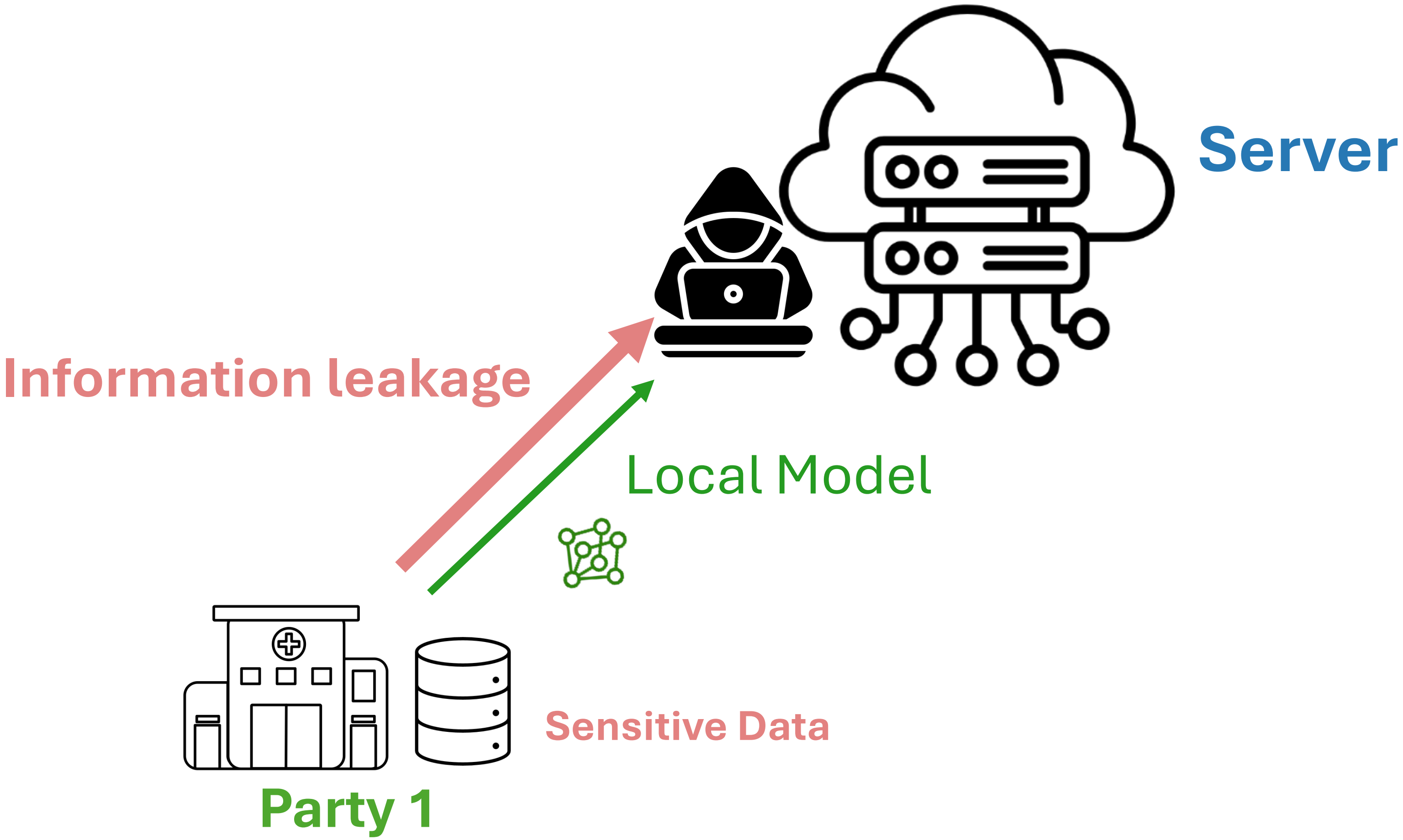
usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Efficient Privacy Auditing in Federated Learning

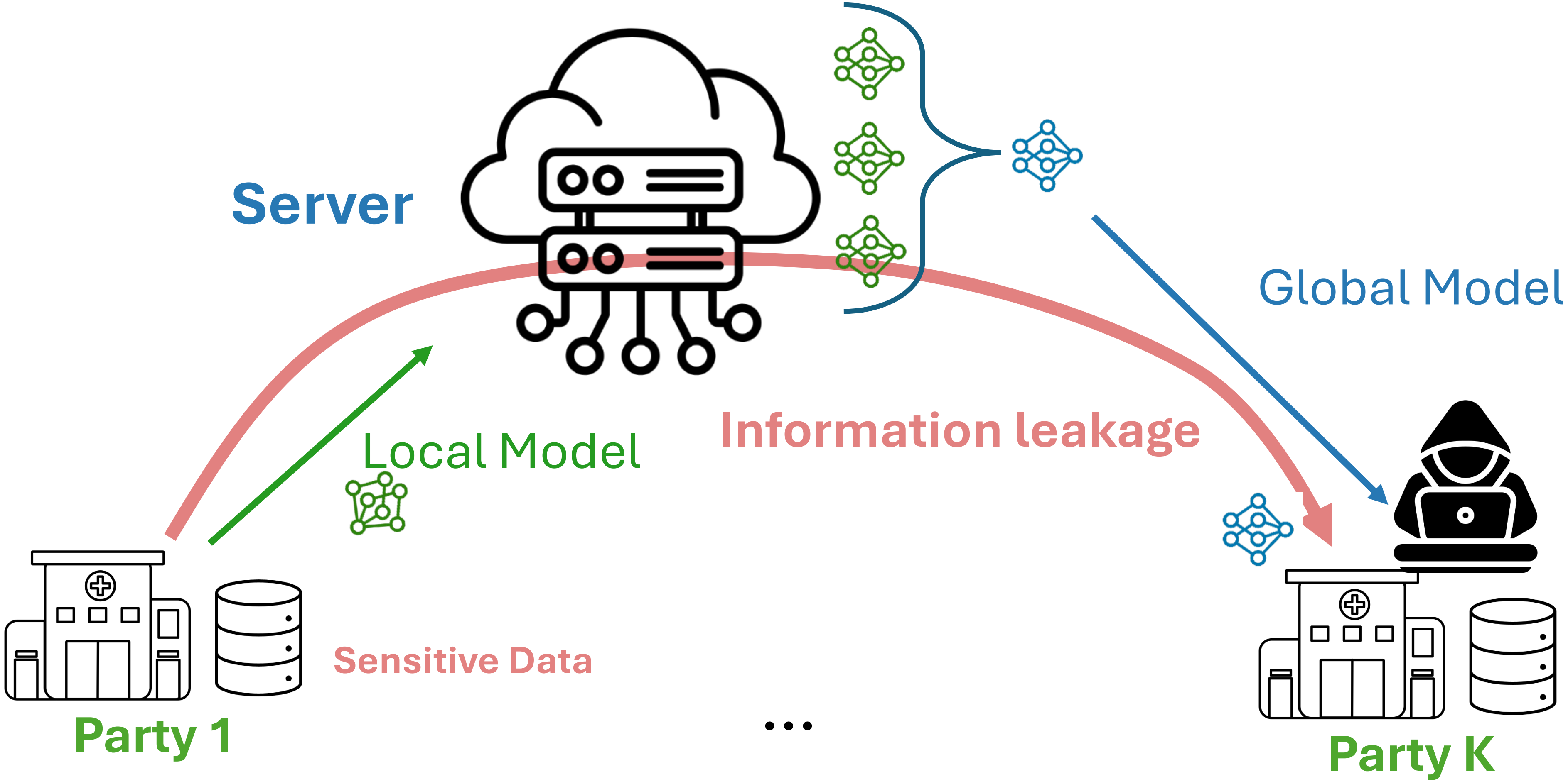
Hongyan Chang, Brandon Edwards, Anindya S. Paul, Reza Shokri



Information Leakage in FL

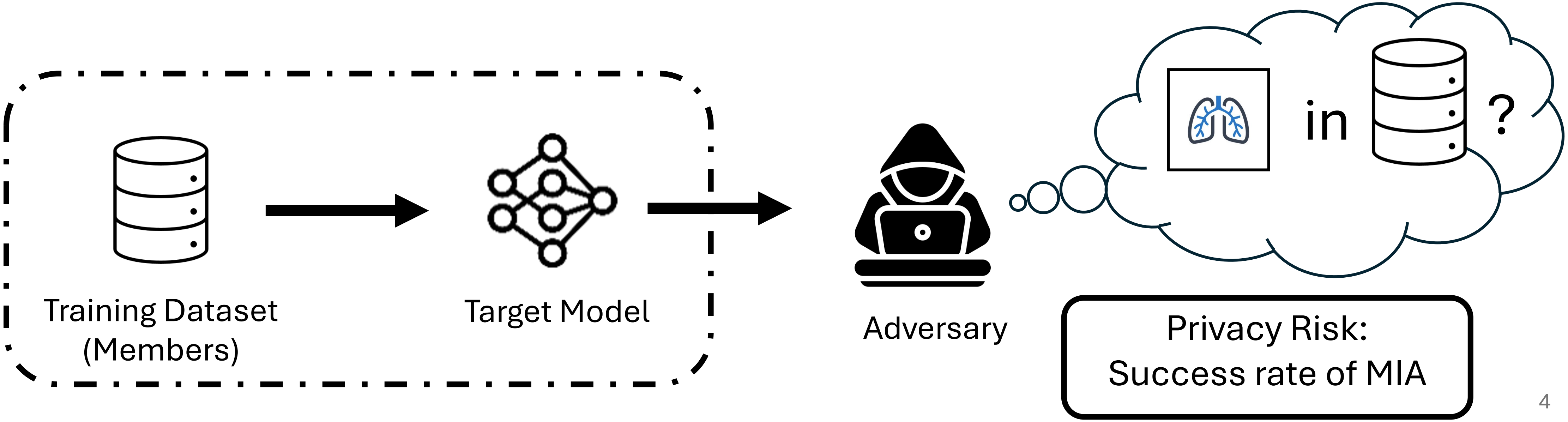


Information Leakage in FL

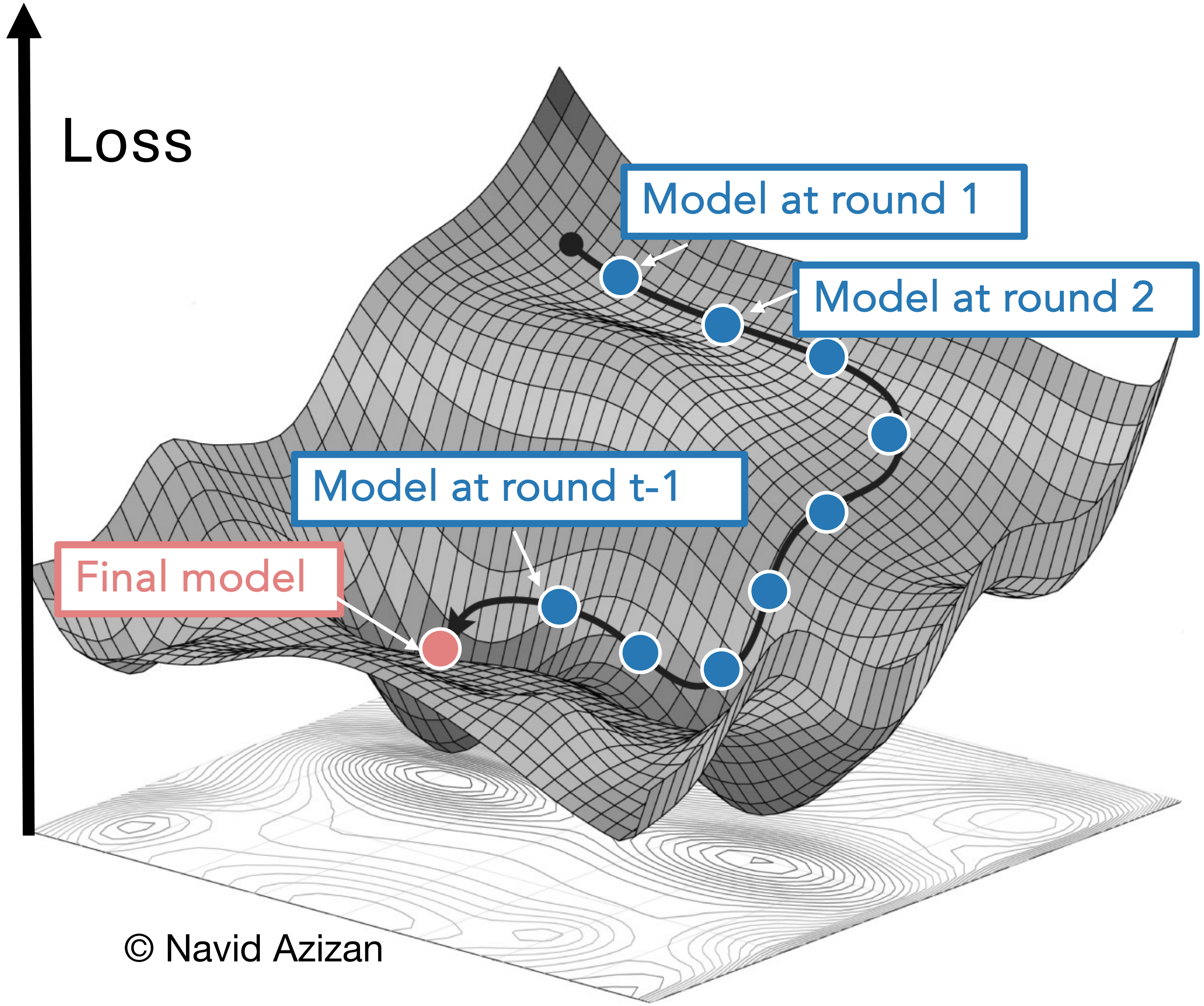


Measuring Information Leakage

- **Approach:** membership inference attacks (MIA)
- **Goal**: Infer whether a data point is in the training dataset



Information Leakage in FL



The adversary observes **multiple model snapshots**—the ***whole training dynamic***

Existing Solutions

- **MIA in Federated Learning**: leverage **all** snapshots
- ***Train*** inference models on
  - Computationally expensive signals (e.g., per-sample gradient)
  - Concatenation on all model snapshots

Nasr, Milad, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." SP, 2019

Existing Solutions

- **MIA in Federated Learning**: leverage **all** snapshots
- ***Train*** inference models on
  - Computationally expensive signals (e.g., per-sample gradient)
  - Concatenation on all model snapshots
- Efficiency
  - **Computing signal alone**
    - **~380 times** long than local training
    - 3 GPU hours -> 46 GPU days!
  - Not feasible for parties with limited resources

Nasr, Milad, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." SP, 2019

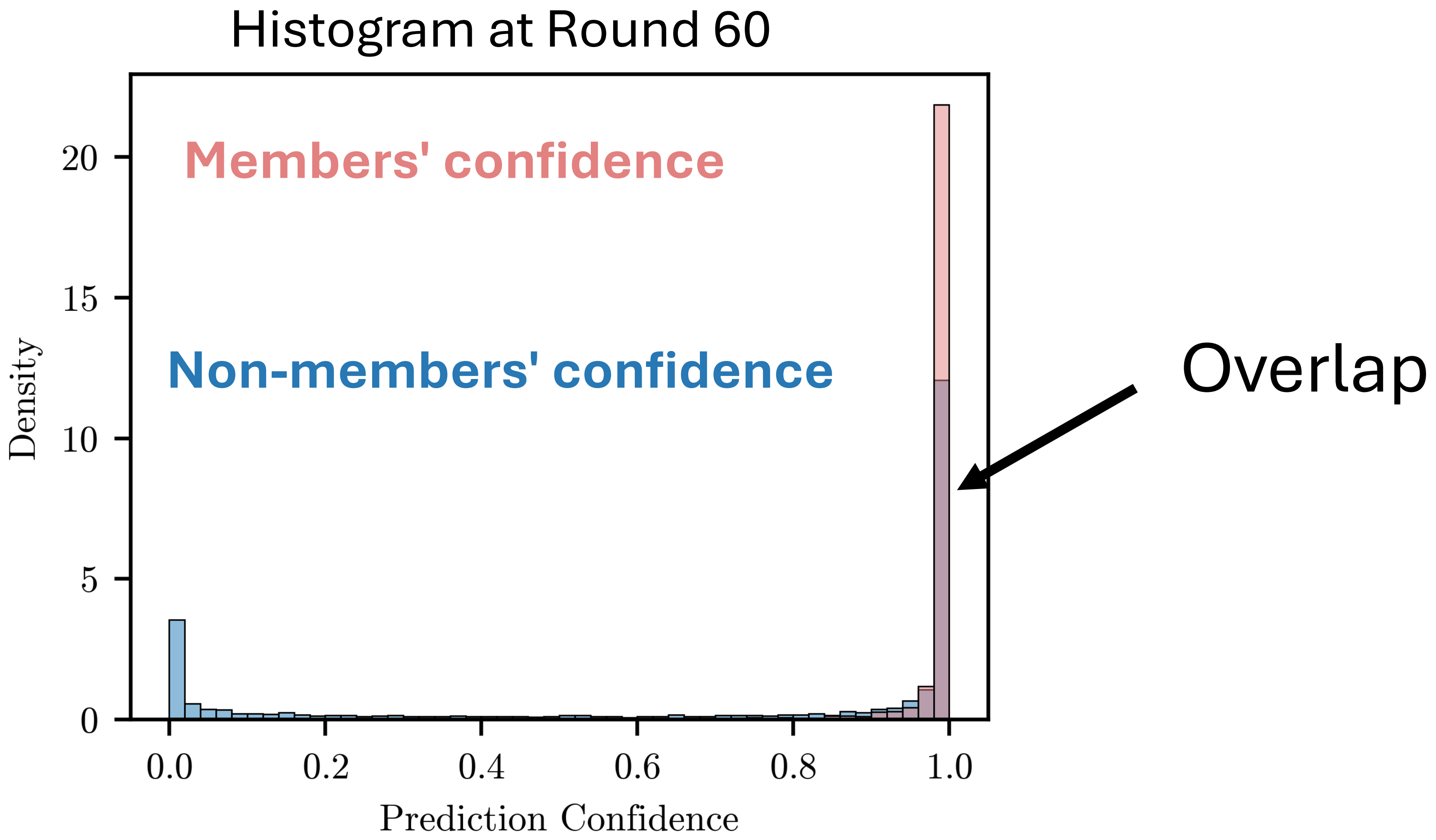
Existing Solutions

- **MIA in Centralized Learning**: leverage **one** snapshot
- ***Train*** a set of reference models
  - Simulate the model behavior: trained with/without the target point
- Efficiency:
  - Need to train lots of reference models (>16 models)
- Effectiveness:
  - Ignore multiple model snapshots

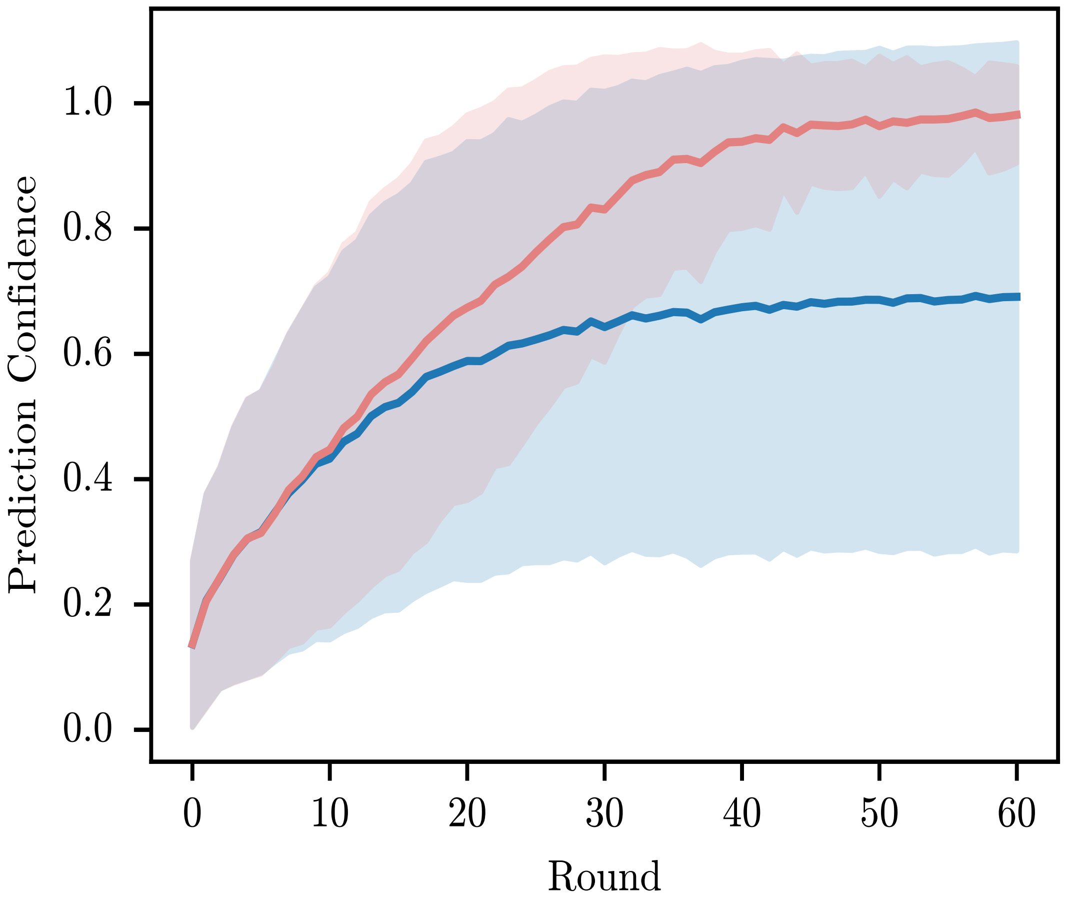
Ye, Jiayuan, et al. "Enhanced membership inference attacks against machine learning models." CCS, 2022.
Carlini, Nicholas, et al. "Membership inference attacks from first principles." SP, 2022.

How can parties effectively audit privacy risks without training additional models?

At a single round



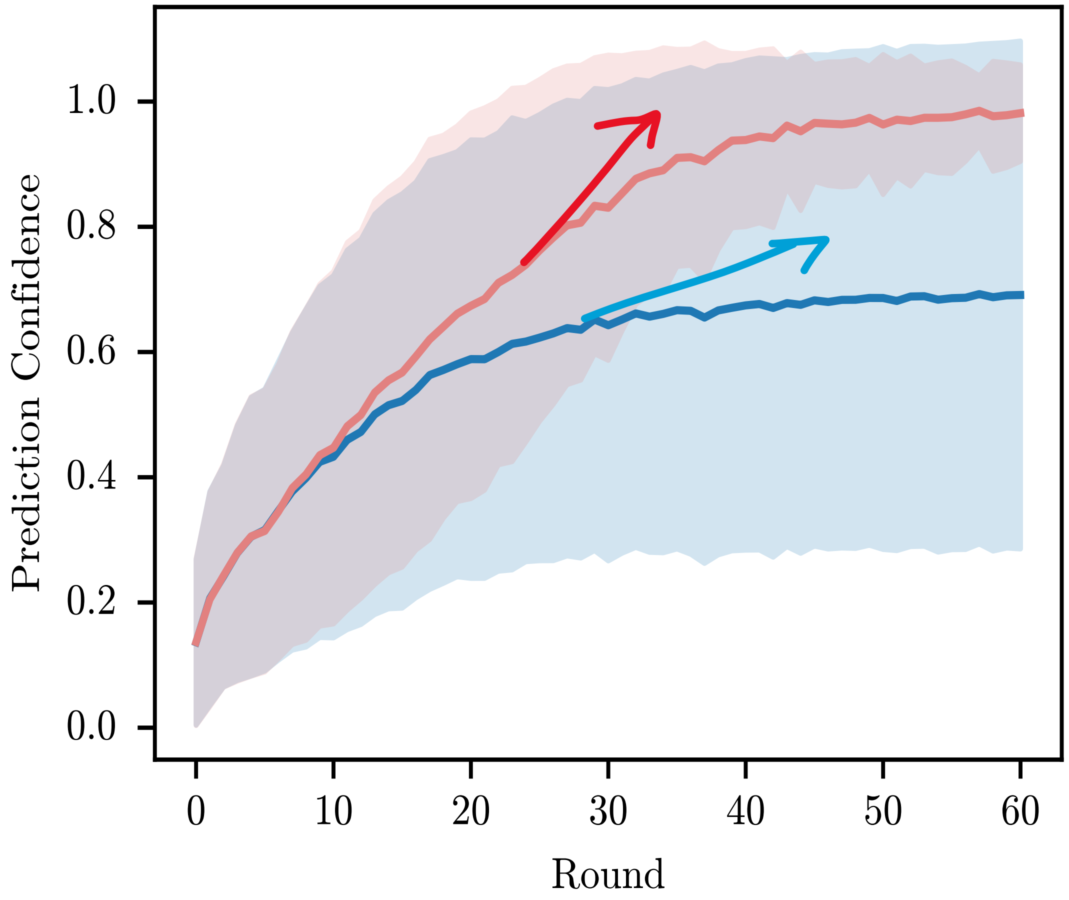
At any single round



Members' confidence

Non-members' confidence

Our solution: whole training dynamic



Members' confidence grows ***faster***

Non-members' confidence grow slower

- **Slope:** rate of change in model performance
- **Computation:** fit a linear function
  - $\hat{c}_t = bt + a$

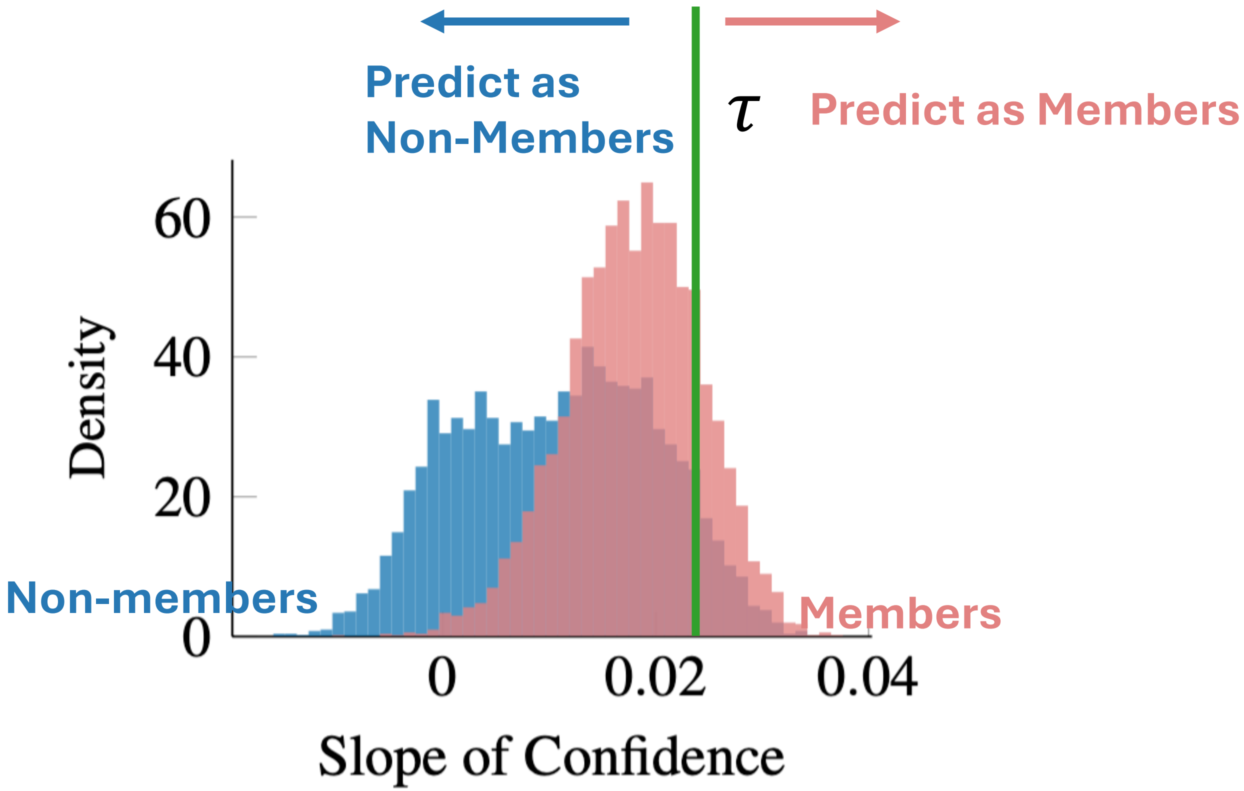
Slope

Efficiency

- Computed on the confidence, loss, and logits of the model.
  - **Already computed** in FL (no addition cost)
- Slope signal is a weighted sum
  - Also fast
- Real-time auditing

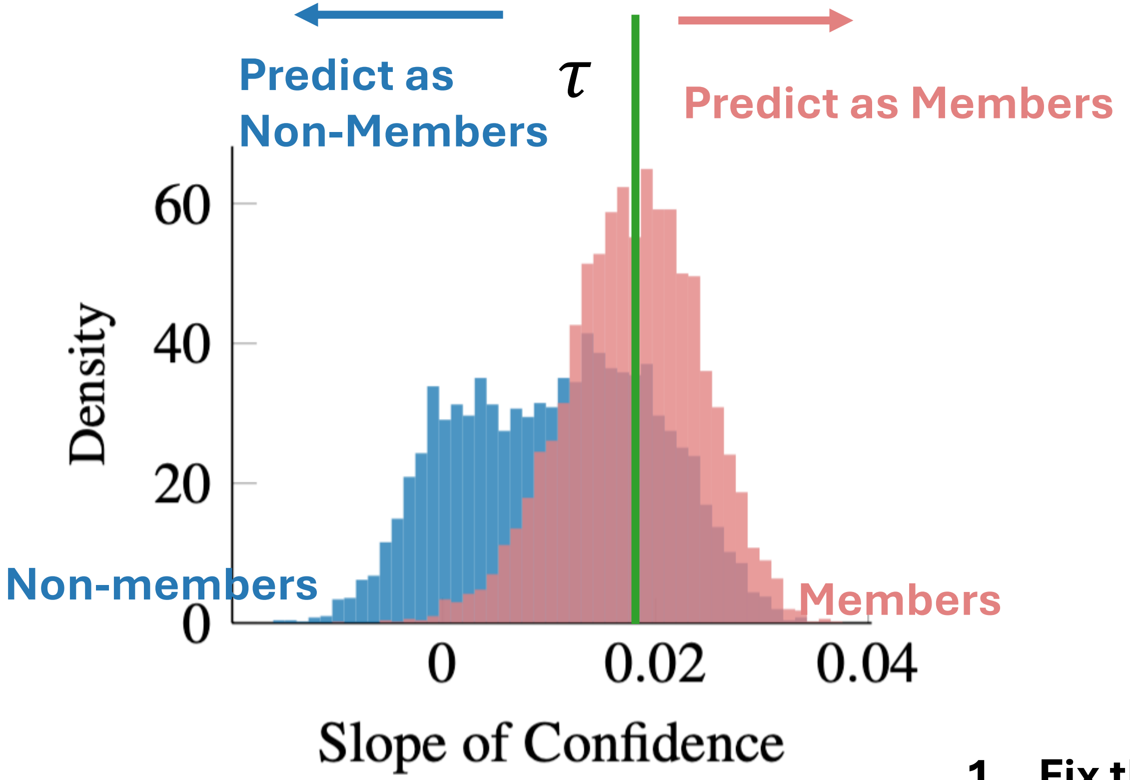
Algorithm

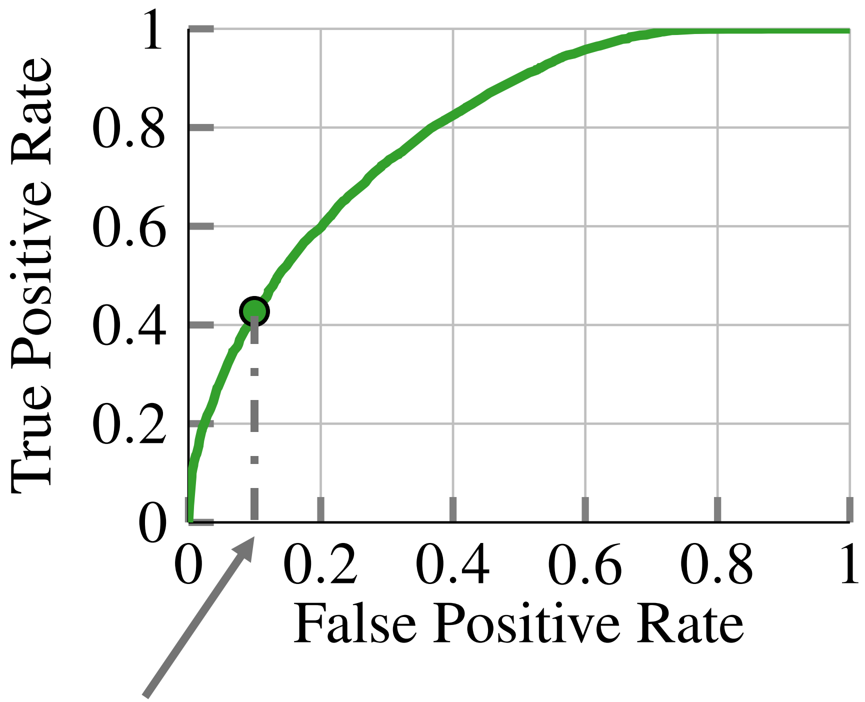
- Compare the computed slope with threshold τ



Metric

- Effectiveness metric: TPR at low FPR





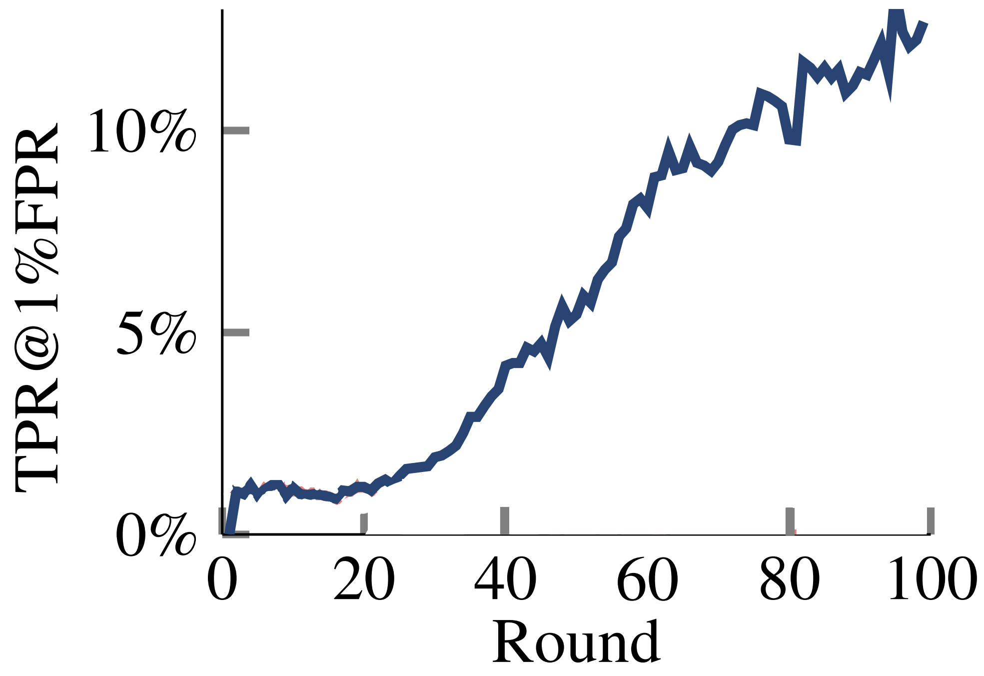
1. **Fix the FPR (e.g., 1%) we are interested in**
2. **Iterate through τ and pick the best one**

Auditing Pipeline

For each communication round:

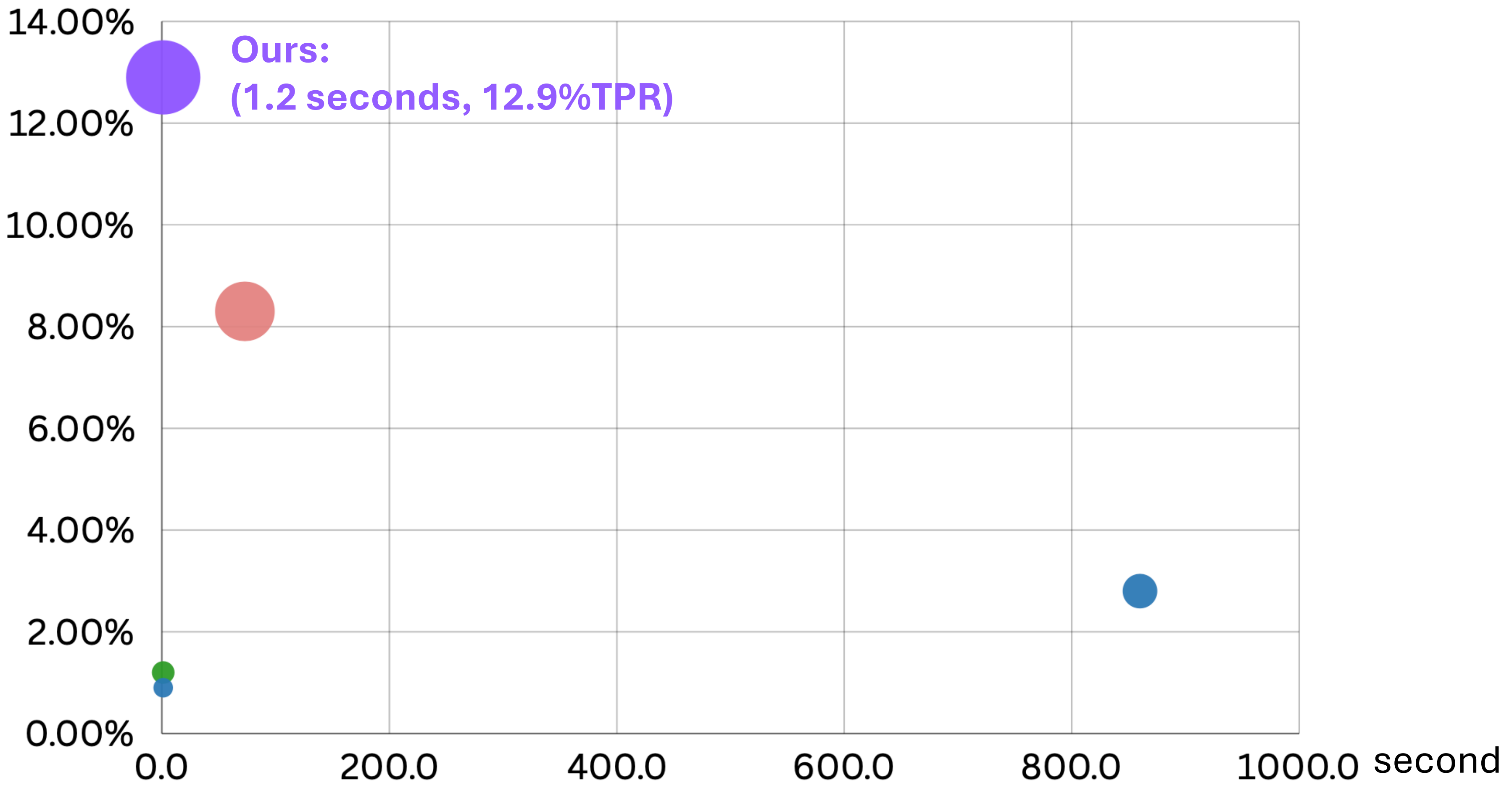
1. Update the global model using the local dataset to get the local model
2. **Evaluate the privacy risk**
3. Send ***local model*** to server

Auditing Results for a Party



Effectiveness VS Efficiency

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model



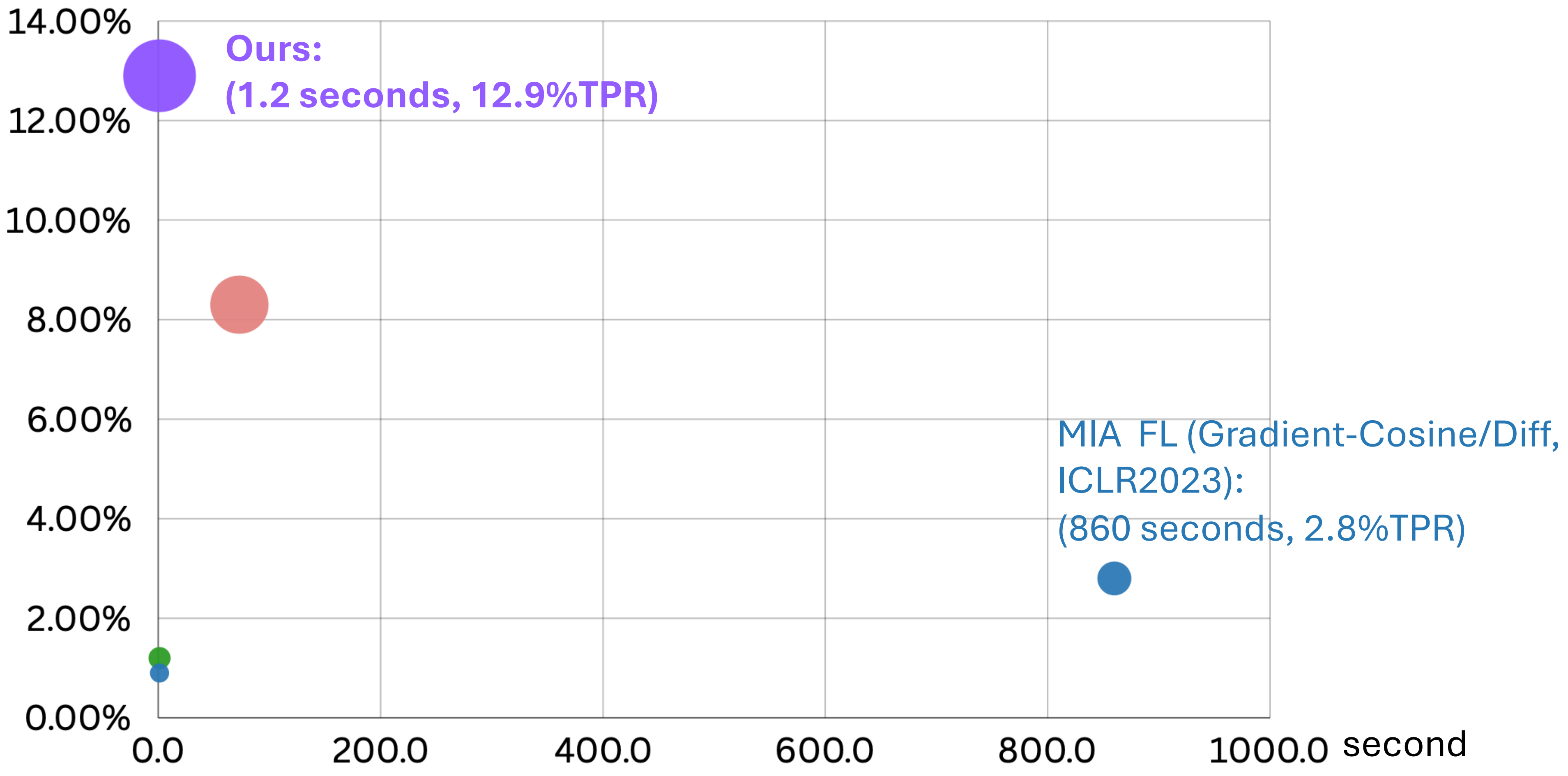
Effectiveness:
TPR at low FPR (1%)

Efficiency:
GPU time (second) for computing the signals for ***one round***

Effectiveness VS Efficiency

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

Effectiveness:
TPR at low FPR (1%)



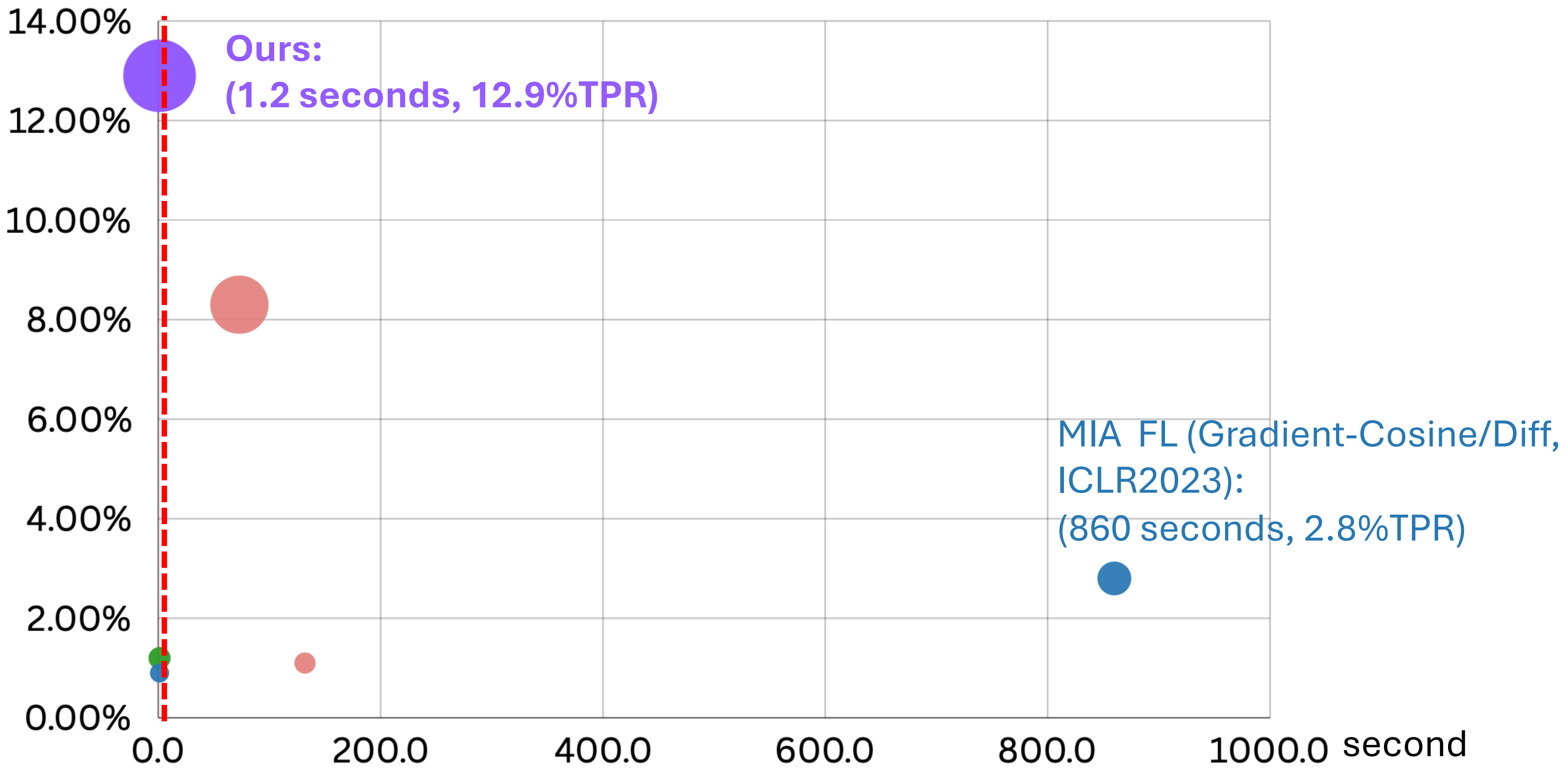
Efficiency:
GPU time (second) for computing the signals for ***one round***

Effectiveness VS Efficiency

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

Local update:
2.3 seconds

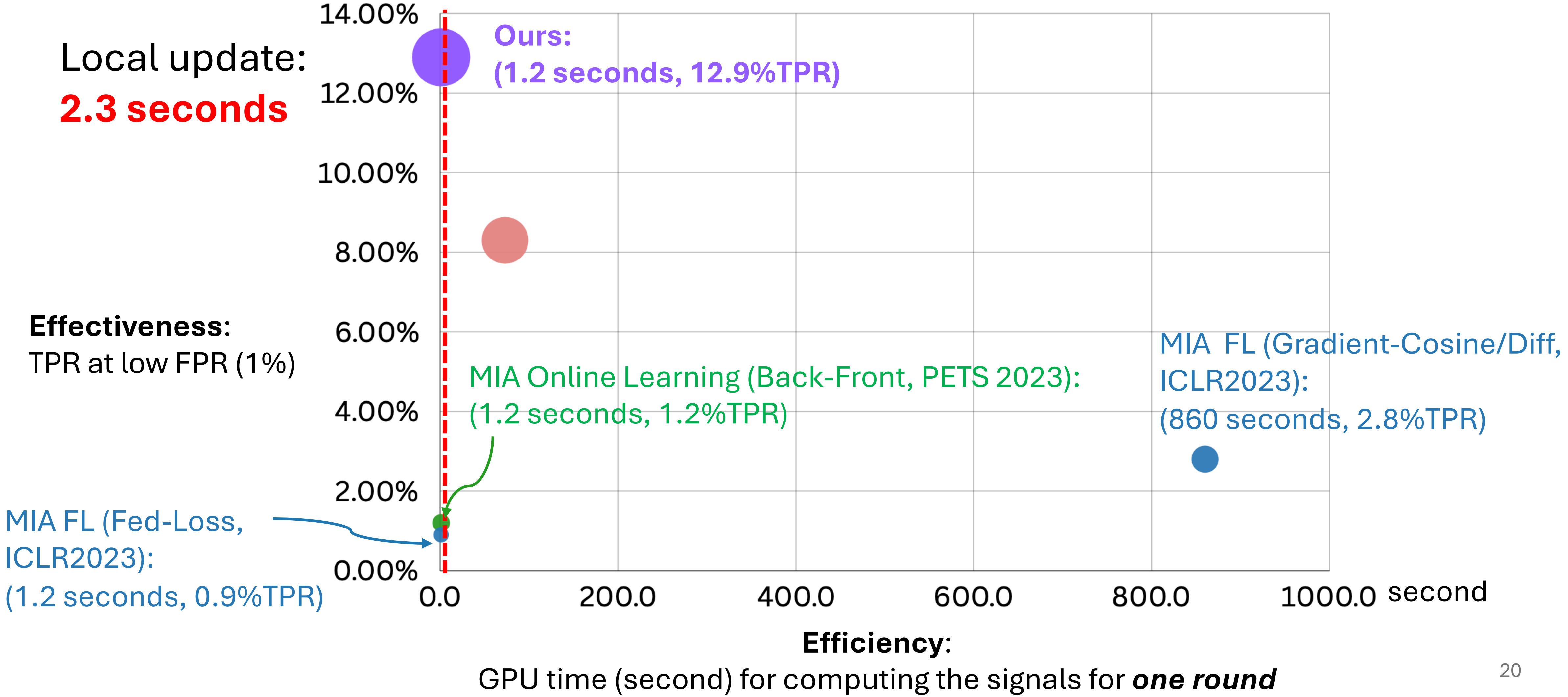
Effectiveness:
TPR at low FPR (1%)



Efficiency:
GPU time (second) for computing the signals for ***one round***

Effectiveness VS Efficiency

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

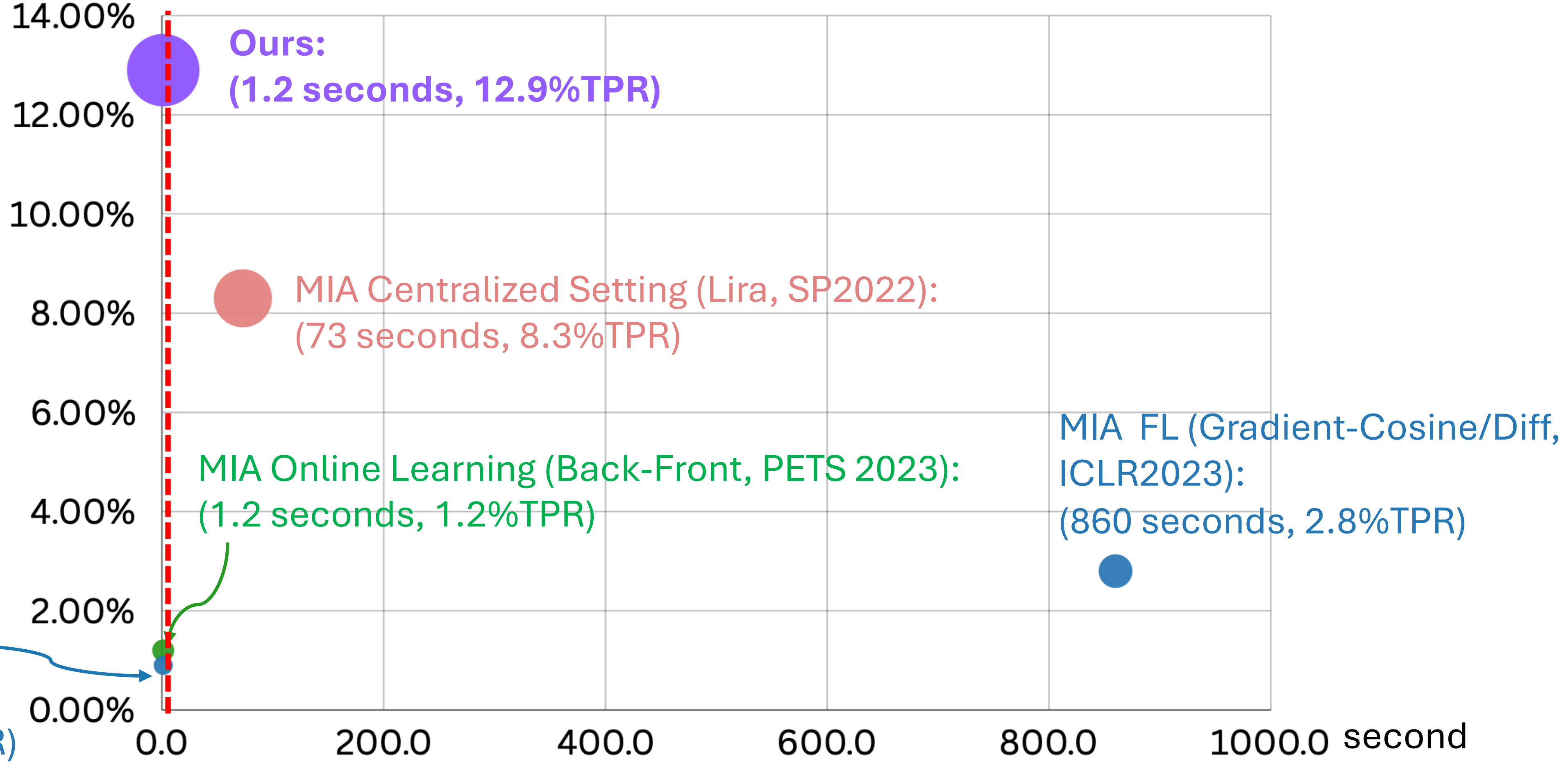


Effectiveness VS Efficiency

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

Local update:
2.3 seconds

Effectiveness:
TPR at low FPR (1%)



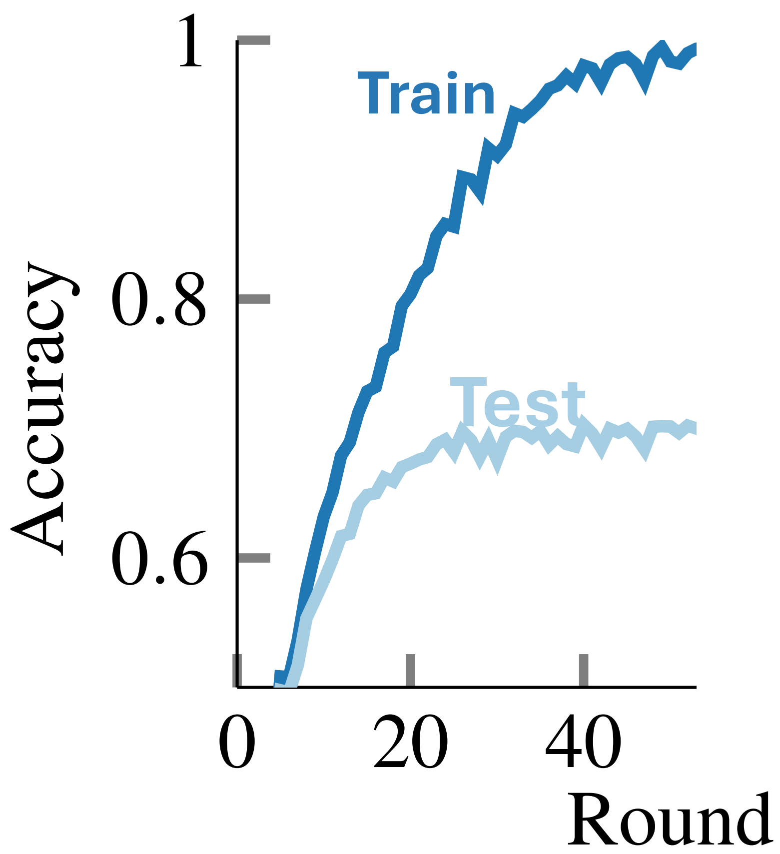
MIA FL (Fed-Loss, ICLR2023):
(1.2 seconds, 0.9%TPR)

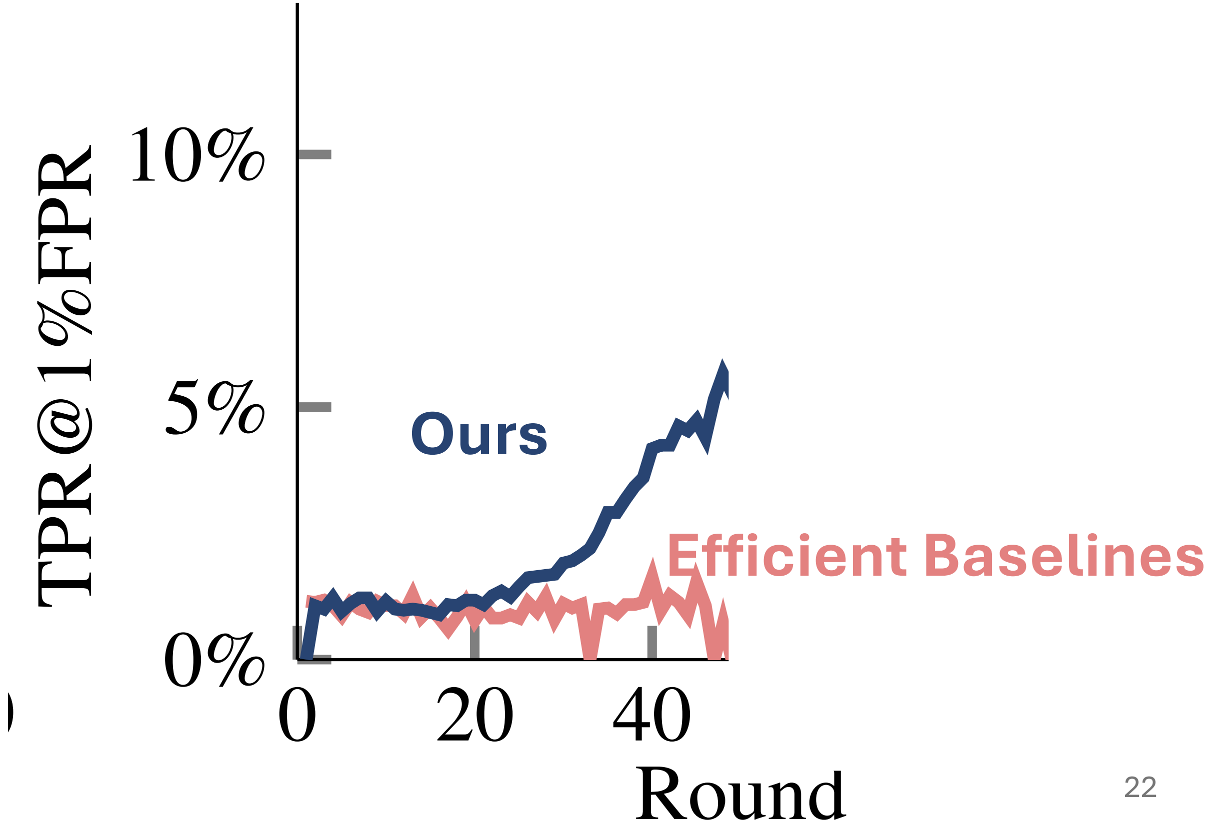
Efficiency:
GPU time (second) for computing the signals for ***one round***

Impact of communication rounds

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

- Privacy risk **keeps increasing** even though the accuracy barely changes

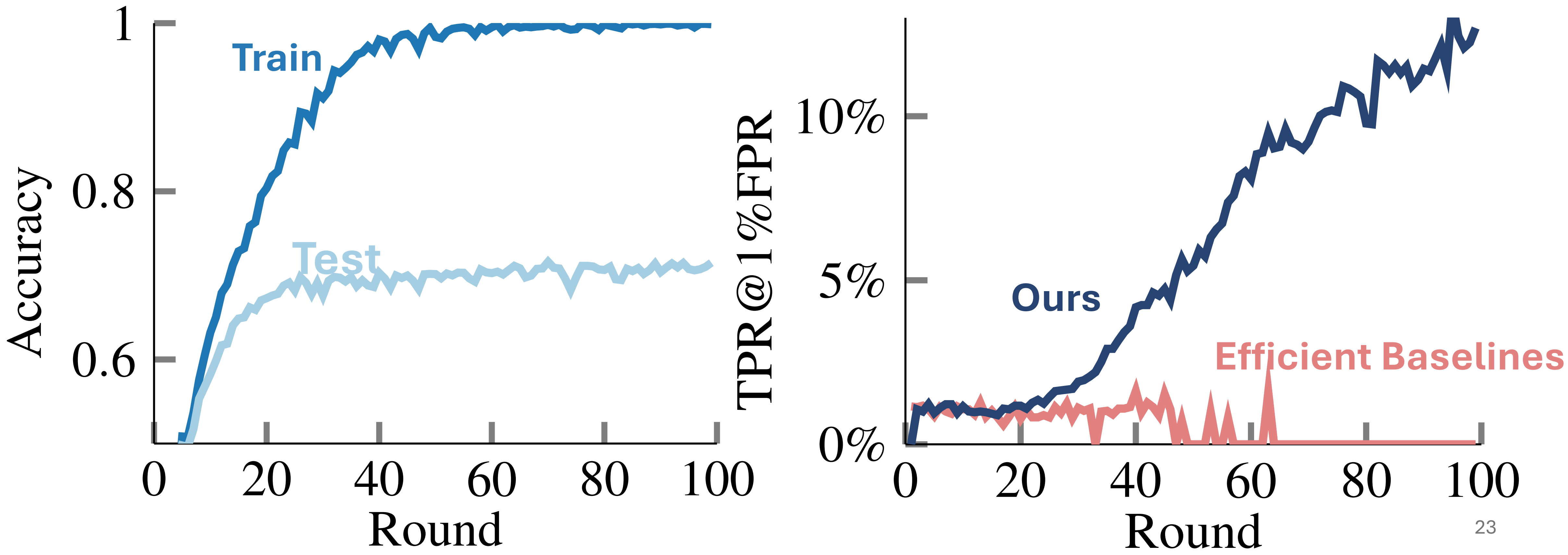




Impact of communication rounds

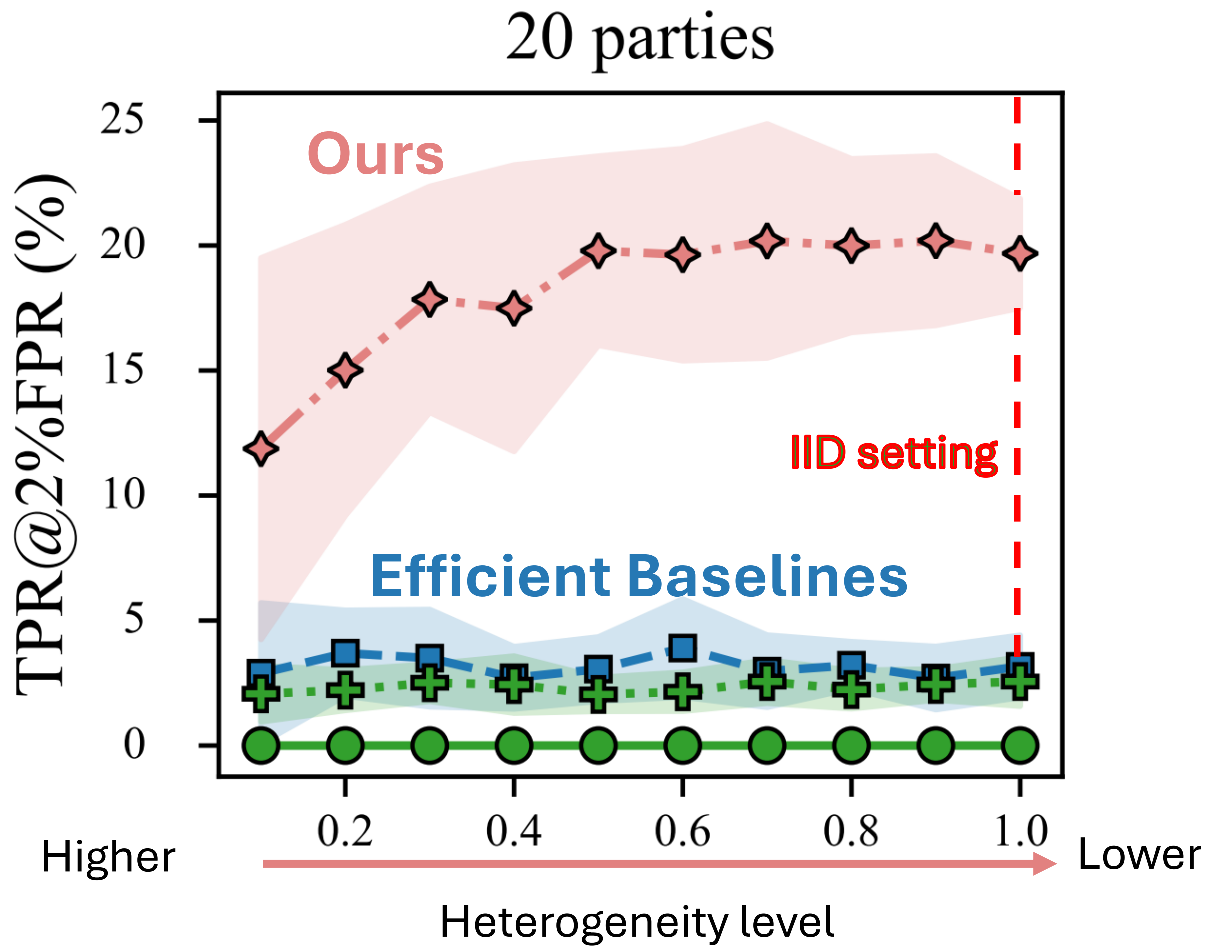
Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model

- Privacy risk **keeps increasing** even though the accuracy barely changes



Impact of data heterogeneity

Model: ResNet56
Data: CIFAR10
Number of Party: 4
Auditing global model



- Parties are experiencing **different levels of risk,** especially in the Non-IID setting.
- Average privacy risk **reduces** when **increasing** data heterogeneity.

Takeaway

- Privacy auditing framework
  - Slope: leverage whole training dynamic
- Effective and efficient
- Comprehensive evaluation
  - Check for more details