PPAI-24

Brave: Byzantine-Resilient and Privacy-Preserving Peer-to-Peer Federated Learning

Zhangchen Xu* (UW), Fengqing Jiang* (UW), Luyao Niu (UW), Jinyuan Jia (PSU), Radha Poovendran (UW)

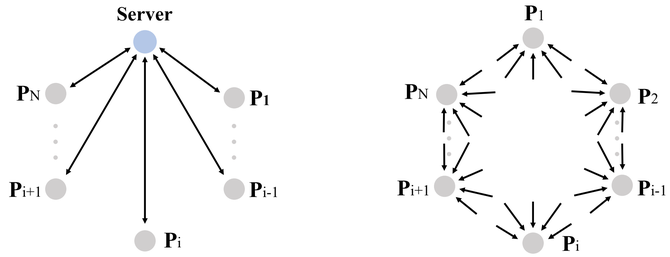
(* Equal Contribution)

TL;DR

1. We propose Brave, a **B**yzantine **R**esilience **A**nd Pri**V**acy PresE**r**ving Protocol for P2P FL.
2. Brave is information-theoretic **private** and **Byzantine resilient**, and **preserves the utility** in attack-free scenarios

Background

Federated Learning (FL) [1] is a collaborative learning paradigm, where multiple clients collaboratively learn a global model without sharing their private data.



Centralized FL | Peer-to-peer FL

System Model & Problem Formulation

System Model for P2P FL

- In communication round t, each participant P_i updates its local model $w_i(t)$ using gradient descent as: $w_i(t+1) = w(t) - \eta g_i(t)$
- Client P_i receives local models $w_j(t+1)$ from other participants, and update the global model as:

$$w(t+1) = \frac{1}{N}\sum_{i=1}^{N} w_i(t+1)$$

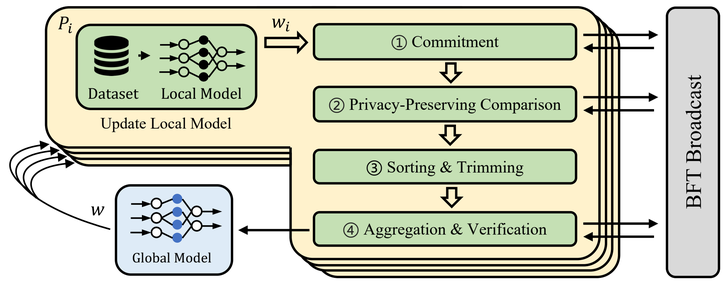
Threat Model

- **Passive Adversaries**: Follow P2P FL but aim to obtain the local model from other participants, there by extrapolating private training data
- **Byzantine Adversaries**: Aim at compromising the learning performance of P2P FL by biasing the local models of other participants. Include (1) creating compromised local models, (2) sending different model to different participants.

Design Goal

- **Information-Theoretic Privacy**: A benign participants P_i's local model cannot be revealed by a passive adversary
- **Byzantine Resilience**: **(1) ε-convergence**: the distance between the global models with and without Byzantine participants is at most ε, **(2) Agreement**: the global model of all benign participants is identical

Brave: Design



① Commitment

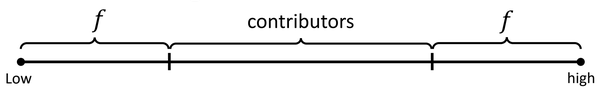
- Generate commitment of its local model w_i using Pedersen Commitment

② Privacy-Preserving Comparison

- Pairwise compare local models of all participants
- Aggregate pairwise comparison results to a sorted sequence

③ Sorting & Trimming

- In each coordinate k, trim the lowest f and the highest f local models, and denote the remaining as contributors.



④ Aggregation & Verification

- Contributors generate a cloaked local model and send it to other clients
- Sum all clocked local models to get the global model
- Verify the summation using commitments generated in Step ①

Brave: Theoretical Guarantees

Theorem 1 Information-theoretic Privacy. Consider a P2P FL in the presence of passive adversaries who are not colluding. Brave, guarantees information-theoretic privacy of the participants' local models.

Theorem 2 Agreement on Global Model. The global model w of all benign participants is identical given $N > 3f + 2$.

Theorem 3 ε-convergence (informal). The global model $w(t)$ obtained by applying Brave deviates from the optimal one by a bounded distance

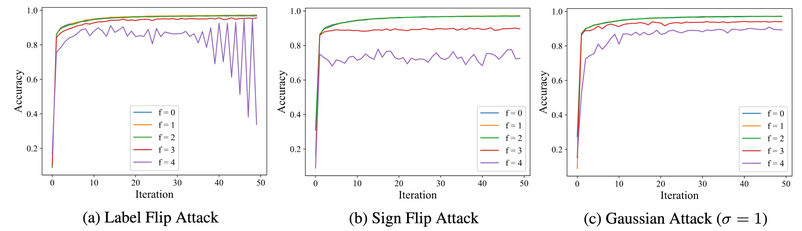
Experimental Results

- **Datasets**: CIFAR10 and MNIST **Model**: 2NN and CNN
- **Evaluation Metric**: Classification Accuracy
- **Byzantine Adversaries**: Label Flip [2], Sign Flip [3], Gaussian Attack [4]

Takeaway: Brave has near-optimal classification accuracy

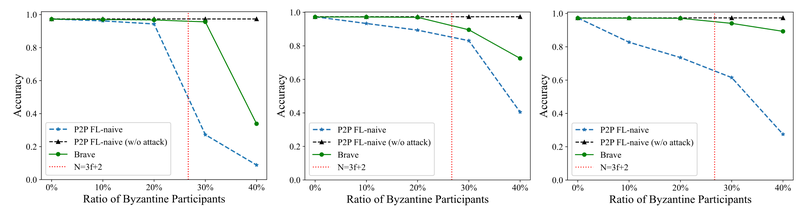
Adversary Strategy	w/o Brave		Brave	
	2NN+MNIST	CNN+CIFAR10	2NN+MNIST	CNN+CIFAR10
No Attack	**97.35%**	**63.94%**	97.21%	63.55%
Label Flip	89.91%	52.15%	**96.74%**	**60.91%**
Sign Flip	11.35%	48.68%	**97.02%**	**63.54%**
Gaussian ($\sigma = 0.1$)	92.02%	55.58%	**96.92%**	**63.08%**
Gaussian ($\sigma = 1$)	53.01%	10.01%	**97.12%**	**61.92%**

Takeaway: Brave ensures ε-convergence



(a) Label Flip Attack (b) Sign Flip Attack (c) Gaussian Attack ($\sigma = 1$)

Takeaway: Brave guarantees Byzantine Resilient if $N > 3f + 2$



(a) Label Flip Attack (b) Sign Flip Attack (c) Gaussian Attack ($\sigma = 1$)

Takeaway: Brave is Scalable

Adv. Strategy	$N = 10$	$N = 15$	$N = 20$
No Attack	97.21%	97.54%	**97.64%**
Label Flip	96.74%	97.18%	**97.50%**
Sign Flip	97.02%	97.34%	**97.51%**
Gaussian ($\sigma = 0.1$)	96.92%	97.39%	**97.42%**
Gaussian ($\sigma = 1$)	97.12%	97.27%	**97.59%**

References

[1] *McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics, 1273–1282. PMLR.*
[2] *Tolpegin, V.; Truex, S.; Gursoy, M. E.; and Liu, L. 2020. Data poisoning attacks against federated learning systems. In European Symposium on Research in Computer Security, 480–501. Springer.*
[3] *Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In International Conference on Machine Learning, 6893–6901. PMLR.*
[4] *Fang, M.; Cao, X.; Jia, J.; and Gong, N. 2020. Local model poisoning attacks to Byzantine-robust federated learning. In 29th USENIX Security Symposium (USENIX Security 20), 1605–1622.*