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

Zhangchen Xu
University of Washington
Seattle, USA
zxu9@uw.edu

Fengqing Jiang
University of Washington
Seattle, USA
fqjiang@uw.edu

Luyao Niu
University of Washington
Seattle, USA
luyaoniu@uw.edu

Jinyuan Jia
The Pennsylvania State University
State College, USA
jinyuan@psu.edu

Radha Poovendran
University of Washington
Seattle, USA
rp3@uw.edu

ABSTRACT

Federated learning (FL) enables multiple participants to train a global machine learning model without sharing their private training data. Peer-to-peer (P2P) FL advances existing centralized FL paradigms by eliminating the server that aggregates local models from participants and then updates the global model. However, P2P FL is vulnerable to (i) *honest-but-curious* participants whose objective is to infer private training data of other participants, and (ii) *Byzantine* participants who can transmit arbitrarily manipulated local models to corrupt the learning process. P2P FL schemes that simultaneously guarantee Byzantine resilience and preserve privacy have been less studied. In this paper, we develop Brave, a protocol that ensures Byzantine Resilience And priVacy-prEserving property for P2P FL in the presence of both types of adversaries. We show that Brave preserves privacy by establishing that any honest-but-curious adversary cannot infer other participants' private data by observing their models. We further prove that Brave is Byzantine-resilient, which guarantees that all benign participants converge to an identical model that deviates from a global model trained without Byzantine adversaries by a bounded distance. We evaluate Brave against three state-of-the-art adversaries on a P2P FL for image classification tasks on benchmark datasets CIFAR10 and MNIST. Our results show that global models learned with Brave in the presence of adversaries achieve comparable classification accuracy to global models trained in the absence of any adversary.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**;
• **Security and privacy**;

ACM Reference Format:
Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, and Radha Poovendran. 2024. POSTER: Brave: Byzantine-Resilient and Privacy-Preserving Peer-to-Peer Federated Learning. In *ACM Asia Conference on Computer and Communications Security (ASIA CCS '24), July 1–5, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3634737.3659428

ASIA CCS '24, July 1–5, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0482-6/24/07
https://doi.org/10.1145/3634737.3659428

1 INTRODUCTION

Peer-to-Peer federated learning (P2P FL) [5] allows multiple participants to collaboratively train a global model while avoiding sharing their private data in a distributed manner. However, P2P FL introduces vulnerabilities due to the lack of central coordination, which can allow unauthorized participants to access sensitive data and/or disrupt the training process. An *honest-but-curious* participant may be able to infer private training data by launching a membership inference attack (MIA) [9], highlighting the urgent need of privacy-preserving training techniques in P2P FL. Moreover, Byzantine participants can send inconsistent models to different participants in P2P networks, inject biased models, or even withhold their updates [2, 10], ultimately degrading the global model or causing the FL to fail. This underscores the need for Byzantine-resilient training techniques. However, techniques that simultaneously preserve privacy and are resilient to Byzantine adversaries have not yet been thoroughly investigated.

In this work in progress, we design a P2P FL scheme that achieves Byzantine resilience while preserving privacy. We consider the presence of both honest-but-curious and Byzantine participants, and define three properties: *information-theoretic privacy*, *ϵ-convergence*, and *agreement*. Information-theoretic privacy ensures that no information about the participants' local models is leaked during the training process, ϵ-convergence implies that the distance between global models learned with and without Byzantine participants is at most ϵ, and agreement indicates the global model of all benign participants are identical. We guarantee information-theoretic privacy by first letting each participant make a commitment of its local model, which will be 'locked' and thus not editable in the future. We then utilize multiparty computation (MPC) [12] to compare and sort the participants' local models without disclosing their true values. Each participant then invokes a trimming scheme to exclude the largest and smallest f local models when updating the global model, where f is the maximum number of Byzantine participants. The main contributions of our paper are summarized as follows.

- We propose Brave, a Byzantine Resilience And priVacy-prEserving protocol for P2P FL. We prove that Brave ensures the local model of each participant to be information-theoretically private during the learning process.

- We design a privacy-preserving trimming scheme to ensure ϵ-convergence in the presence of Byzantine adversaries. We leverage distributed consensus to ensure agreement. We theoretically prove Brave is resilient to Byzantine participants given $N > 3f+2$, where N is the total number of participants.
- We evaluate Brave against three state-of-the-art adversaries on two image classification tasks. Our results show that Brave guarantees ϵ-convergence if $N > 3f + 2$ holds. Furthermore, the global model trained using P2P FL that implements Brave achieves comparable classification accuracy to a global model learned in the absence of any adversary.

2 THREAT MODEL

We consider a P2P FL framework where both *passive* and *Byzantine* adversaries exist and potentially overlap, with the remaining participants identified as *benign*.

Adversary Goals and Actions. The passive adversaries [9] follow the procedure of P2P FL but aim to obtain the local models from the other participants, thereby extrapolating private training data by launching MIA [9] on these local models. The Byzantine adversaries aim at compromising the learning performance of P2P FL by biasing the local models of other participants. In pursuit of their objectives, the Byzantine adversaries can create compromised local models, and send different local models to different participants or just remain silent in the communication process.

Adversary Capabilities. The adversaries are assumed to have full access to the messages they receive but are incapable of eavesdropping or intercepting the communications of others. Furthermore, these adversaries have limited computational capability to solve the discrete logarithms problem [7]. This assumption is widely adopted for security of public key systems and protocols [3].

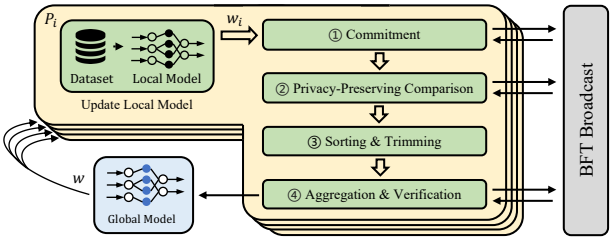


Figure 1: This figure shows the overall workflow of Brave.

3 DESIGN OF BRAVE

We design Brave, a protocol in P2P FL that comprises four stages: *Commitment*, *Privacy-preserving Comparison*, *Sorting & Trimming*, and *Aggregation & Verification*. The proposed protocol is proven to achieve information-theoretic privacy and Byzantine resilience if $N > 3f + 2$. To defend against Byzantine adversaries who might input differing models at various stages, Brave starts with a commitment stage (①) after participants update their local models in each iteration. Following this, Brave enters a privacy-preserving comparison stage (②), which enables the sorting of local models without revealing any information about their true values. Brave then incorporate trimmed mean [13] in the sorting & trimming stage (③) to remove the outliers possibly introduced by Byzantine adversaries. In the last aggregation & verification stage (④), MPC [1] is performed to aggregate models while preserving privacy. This stage also verifies the consistency of the aggregated model with the commitment.

4 EXPERIMENTS

In this section, we evaluate Brave against three state-of-the-art adversaries using two image classification tasks.

4.1 Experimental Setup

In what follows, we describe the experimental setup.

Brave Setup. We implement Brave on two P2P FL with different settings. In the first P2P FL, the participants train a 2-hidden-layer model (2NN) using samples from the MNIST dataset [6]. In the second P2P FL, the participants learn a Convolutional Neural Network (CNN) model using the CIFAR10 dataset [4]. In both P2P FL, the training images from the CIFAR10 or MNIST dataset are independently and identically distributed (i.i.d.) to the participants so that each participant has $|\mathcal{D}_i| = 2000$ images within its private dataset. The participants update their local models $w_i(t)$ using stochastic gradient descent (SGD) algorithm with learning rate $\eta = 0.01$ [8].

Baseline Setup. We present the effectiveness of Brave by comparing with a baseline, P2P FL-naïve. The baseline implements the classic P2P FL as in [5].

Threat Models. We evaluate Brave against Byzantine adversaries who adopt distinct strategies: No Attack, Label Flip Attack [10], Sign Flip Attack [11] and Gaussian Attack [2].

Evaluation Metric. We use classification accuracy over the testing dataset as the evaluation metric.

4.2 Experimental Results

In what follows, we demonstrate the effectiveness of Brave.

Byzantine Resilience of Brave. We evaluate the classification accuracy of the learned global models obtained by P2P FL with $N = 10$ participants and $f = 2$ Byzantine adversaries. In Table 1, we present the classification accuracy of the learned 2NN and CNN when the Byzantine adversaries adopt different attack strategies. We observe that if the Byzantine adversaries does not initiate any attack, then Brave retains comparable accuracy compared with classic P2P FL that does not implement Brave. Furthermore, Brave guarantees significantly higher accuracy of the learned model once the Byzantine adversaries send compromised local models to the other participants, and thereby is Byzantine-resilient.

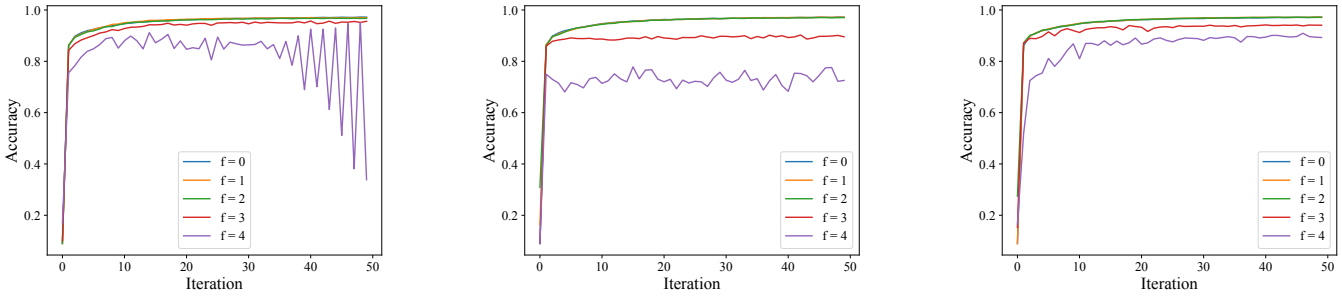
ϵ-convergence of Brave. We demonstrate ϵ-convergence of Brave in Fig. 2a-2c. We observe that when $N > 3f + 2$, then Brave guarantees that the global model learned in the presence of Byzantine adversaries remains ϵ-close to the global model learned when $f = 0$. We further notice that the value of ϵ increases with respect to the number of Byzantine adversaries. The presence of more Byzantine adversaries could introduce additional bias to the learned global model, or even cause FL to fail (as shown in Fig. 2a when $f = 4$).

5 CONCLUSION

In this paper, we considered P2P FL in the presence of both passive and Byzantine adversaries. The passive adversaries aimed at

Table 1: This table presents the classification accuracy of the learned 2NN and CNN using a P2P FL with $N = 10$ and $f = 2$. The second row of the table represents the scenario where the Byzantine adversaries send their true local models and do not launch any attack. Rows 3-6 correspond to different threat models. We observe that Brave ensures the P2P FL to learn a global model with near-optimal classification accuracy against all threat models, and hence is Byzantine-resilient.

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%**



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

Figure 2: This figure presents the accuracy of 2NN learned using P2P FL with $N = 10$ participants at each iteration t. When the number of Byzantine adversaries f satisfies $N > 3f + 2$, i.e., $f \in \{0, 1, 2\}$, Brave ensures ϵ-convergence property. When f violates $N > 3f + 2$, the Byzantine adversaries can corrupt the learned 2NN, and even prevent FL from converging (Fig. 2a, $f = 4$).

inferring the other participants' private information during the training process, whereas Byzantine adversaries could arbitrarily manipulate the information it sent to disrupt the learning algorithm. We developed a four-stage P2P FL protocol named Brave that information-theoretically preserves privacy and is resilient to malicious attacks caused by Byzantine adversaries. We evaluated Brave using two image classification tasks with CIFAR10 and MNIST datasets. Our results showed that Brave can effectively defend against the state-of-the-art adversaries.

ACKNOWLEDGEMENTS

This work is partially supported by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-20-1-0074 and National Science Foundation (NSF) under grant No. 2229876.

This work is supported in part by funds provided by the National Science Foundation, by the Department of Homeland Security, and by IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or its federal agency and industry partners.

REFERENCES

[1] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. ACM, 1175–1191.

[2] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. 2020. Local model poisoning attacks to Byzantine-robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 1605–1622.

[3] Antoine Joux, Andrew Odlyzko, and Cécile Pierrot. 2014. The past, evolving present, and future of the discrete logarithm. *Open Problems in Mathematics and Computational Science* (2014), 5–36.

[4] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images*. Ph. D. Dissertation. University of Toronto.

[5] Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar. 2019. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173* (2019).

[6] Yann LeCun, Corinna Cortes, and Chris Burges. 2010. MNIST handwritten digit database. Available: http://yann.lecun.com/exdb/mnist.

[7] Kevin S McCurley. 1990. The discrete logarithm problem. In *Proc. of Symp. in Applied Math*, Vol. 42. USA, AMS, 49–74.

[8] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54. PMLR, 1273–1282.

[9] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 739–753. https://doi.org/10.1109/SP.2019.00065

[10] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *Computer Security – ESORICS 2020*. Springer International Publishing, 480–501.

[11] Cong Xie, Sanmi Koyejo, and Indranil Gupta. 2019. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 6893–6901.

[12] Andrew C. Yao. 1982. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science*. IEEE, 160–164. https://doi.org/10.1109/SFCS.1982.38

[13] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 5650–5659.