AUTHOR:
ATANAS DONEV
A.H.DONEV@STUDENT.TUDELFT.NL

BYZANTINE ATTACKS AND DEFENSES IN DECENTRALIZED LEARNING SYSTEMS THAT EXCHANGE CHUNKED MODELS

SUPERVISORS: JÉRÉMIE DECOUCHANT | BART COX

INTRODUCTION

Fully decentralized (peer-to-peer) learning removes the central server, so each node trains and exchanges model updates locally. One recent improvement is model chunking [2] (sending subsets of the model updates, rather than the whole ones). This may help boost privacy, but it appears that such network are still susceptible to byzantine model attacks (backdoor attacks, label flipping, model poisoning, etc.).

Most existing backdoor defenses—robust aggregators such as Krum/Bulyan, clustering-plus-noise filters, or validation sets—depend on a server with full visibility (federated learning), so they break down when models are exchanged only in small chunks and no peer ever sees the whole model.

Our work fills this gap by defining the chunk-level threat and adapting two defenses from DL to chunked DL: Norm Clipping (NC) and Sentinel (SL). We then alter NC to make the threshold adaptive.

RESEARCH QUESTION

How to defend against Byzantine attacks while exchanging chunked models in decentralized setting?

Sub-questions:
What is **decentralized learning**?
What are **chunked models**?
What are the **byzantine attacks** in literature?
What are the **proposed defenses** in literature?

BACKGROUND

Decentralized Learning (DL) enables collaborative machine learning without a central server, allowing direct peer-to-peer model exchanges. However, it remains vulnerable to Byzantine attacks. Backdoor attacks subtly poison models, embedding triggers causing specific misclassifications without affecting normal accuracy. Untargeted label-flipping randomly corrupts training labels, severely reducing overall accuracy. Defending DL, especially when models are exchanged in chunks, presents ongoing research challenges.

METHODOLOGY

We simulate fully decentralized learning on a 16-node static 3-regular graph using DecentralyzePy, where model parameters are split into chunks before sharing — ensuring no node sees the full model and privacy is preserved. We use the following datasets:

- MNIST (LeNet): 300 rounds (for rapid prototyping)
- CIFAR-10 (LeNet5): 300 rounds (final evaluation)
- Data is evenly split; 3 out of 16 nodes are malicious.

We employ **two** distinct **model attacks**, namely **backdoor attack [1]** and **untargeted label flipping [3]**. Against each of those attacks we test **three** different **defenses**: **Norm Clipping [4]**, **Adaptive Norm Clipping** and **Sentinel [3]**.

Each attack is tested against the 3 defenses and the baseline. This process is repeated 3 times. Key metrics:

- Clean Accuracy (model utility)
- Attack Success Rate (ASR) (malicious effectiveness)
- Convergence Rounds (training stability)

The data is set first as IID and then as non-IID to reflect realistic conditions.

Algorithm	Accuracy	ASR
ANC	57.93±0.02%	7.01±0.11%
ANC-BA	57.92±0.02%	6.70±0.51%
ANC-UF	55.68±0.21%	N/A
CDL	73.04±0.10%	7.98±0.04%
BA	72.71±0.23%	85.07±0.13%
NC	57.33±0.07%	7.01±0.48%
NC-BA	57.14±0.25%	20.83±2.06%
NC-UF	54.30±0.21%	N/A
SL	73.17±0.08%	N/A
SL-BA	72.14±0.28%	85.48±0.06%
SL-UF	9.94±0.00%	N/A
UF	9.91±0.06%	N/A

TABLE 1: IID RESULTS

Algorithm	Accuracy	ASR
ANC	44.12±0.08%	8.22±0.10%
ANC-BA	40.94±0.38%	17.80±0.09%
ANC-UF	41.42±0.40%	N/A
CDL	63.13±1.04%	7.94±0.18%
BA	62.66±0.03%	85.06±0.60%
NC	43.75±0.06%	7.04±0.65%
NC-BA	41.96±0.09%	23.12±1.35%
NC-UF	41.92±0.10%	N/A
SL	63.74±0.27%	N/A
SL-BA	61.94±0.84%	85.94±1.44%
SL-UF	9.91±0.06%	N/A
UF	9.87±0.06%	N/A

TABLE 2: NON-IID RESULTS

RESULTS

NC - Norm Clipping
ANC - Adaptive Norm Clipping
SL - Sentinel
BA - Backdoor attack
UF - Untargeted Label Flipping
CDL - vanilla chunked DL

ANC-BA – Adaptive NC under Backdoor attack
ANC-UF – Adaptive NC under UF
NC-BA – Norm Clipping under Backdoor attack
NC-UF – Norm Clipping under UF
SL-BA – Sentinel under Backdoor attack
SL-UF – Sentinel under UF

DISCUSSION

Although model chunking helps with privacy attacks, model attacks remain highly effective in this scenario. Backdoor attack achieves 85% ASR and untargeted label flipping manages to sabotage final test accuracy significantly. While defenses like Norm Clipping and Adaptive Norm Clipping manage to protect the network, they incur a modest degradation in performance. Sentinel does not degrade final performance, but does not succeed in defending at all.

CONCLUSION

Our evaluation confirms that decentralized learning is highly vulnerable to byzantine manipulation: the backdoor attack remained stealthy and achieved a high ASR, while untargeted label flip managed to degrade global accuracy significantly. The defenses tested produced mixed results. The static ones (Norm Clipping and Adaptive Norm Clipping) countered the attacks but degraded final accuracy. The robust aggregator Sentinel did not succeed at all in defending, possibly because the updates are chunks instead of whole models. The conclusion is that robust aggregators do not work in the scenario of model chunking and that static defenses need to be improved to not incur a cost on the final accuracy.

[1] E. BAGDASARYAN, A. VEIT, Y. HUA, D. ESTRIN, AND V. SHMATIKOV, "HOW TO BACKDOOR FEDERATED LEARNING," INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, PP. 2938–2948, JUL. 2018.
[2] S. BISWAS, M. EVEN, A.-M. KERMARREC, L. MASSOULIÉ, R. PIRES, R. SHARMA, AND M. DE VOS, "NOISELESS PRIVACY-PRESERVING DECENTRALIZED LEARNING," PROCEEDINGS ON PRIVACY ENHANCING TECHNOLOGIES, VOL. 2025, NO. 1, PP. 824–844, NOV. 2024.
[3] C. FENG, A. H. CELDRÁN, J. BALTENSPERGER, E. T. M. BELTRÁN, P. M. S. SÁNCHEZ, G. BOVET, AND B. STILLER, "SENTINEL: AN AGGREGATION FUNCTION TO SECURE DECENTRALIZED FEDERATED LEARNING," IN FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 2024.
[4] G. SYROS, G. YAR, S. BOBOILA, C. NITA-ROTARU, AND A. OPREA, "BACKDOOR ATTACKS IN PEER-TO-PEER FEDERATED LEARNING," ACM TRANSACTIONS ON PRIVACY AND SECURITY, OCT. 2024.