Institute for Communications Engineering
Department of Electrical and Computer Engineering
Technical University of Munich

LeadFL: Client Self-Defense against Model Poisoning in Federated Learning[1]

Ramazan Tan, Claus Guthmann

Introduction

Problem: Federated Learning is susceptible to **bursty poisoning attacks**, where sudden spikes in malicious clients bypass standard server-side defenses.

Lingering Impact: Once the model is poisoned, the attack effect persists for many subsequent rounds even without further attacks, as servers cannot eliminate this lasting damage.

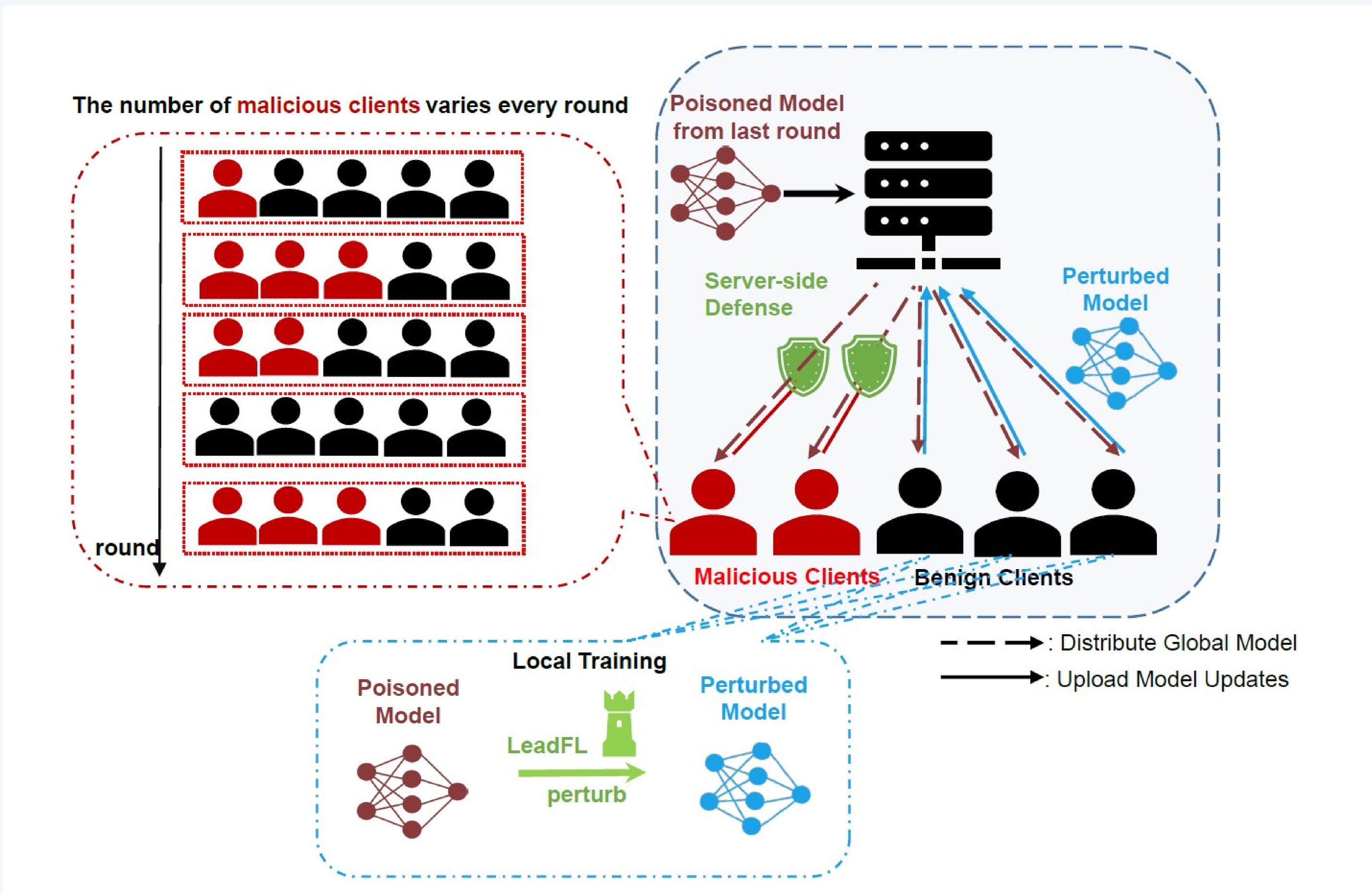

Fig 1: Bursty Adversarial Patterns

Why Existing Defenses Fail: Server-side defenses are designed assuming constant and low number of malicious clients.

Motivation: A client-side intervention is critical to suppress attack propagation locally, regardless of server-side defenses.

Understanding the Attack Effect

Attack Effect on model Parameter[2]: We quantify the impact using the Attack Effect on Parameter (AEP), denoted as δ_t.

$$\delta_t \triangleq \underbrace{\boldsymbol{\theta_t}}_{\text{Benign global model}} - \underbrace{\boldsymbol{\theta_t^M}}_{\text{Poisoned global model}} \tag{1}$$

Benign global model **Poisoned global model**

Propagation of AEP[2]: When malicious clients attack in rounds τ_1 and τ_2, we estimate the attack effect for the intermediate rounds ($\tau_1 < t < \tau_2$) as follows:

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k \in S_t} p^k \prod_{i=0}^{I-1}(I - \eta_t \boldsymbol{H_{t,i}^k})\right]\hat{\delta}_{t-1} \tag{2}$$

Where **Hessian Matrix** is: $H_{t,i}^k \triangleq \nabla^2 L(\theta_{t,i}^k)$

Key Insight: If $\hat{\delta}_{\tau_1}$ resides in the kernel (null space) of $H_{t,i}^k$, then $\hat{\delta}_t = \hat{\delta}_{\tau_1}$, causing the attack effect to persist unchanged.

If $H_{t,i}^k$ is highly sparse $\Longrightarrow \delta_t \approx \delta_{t-1}$

Why server-side defenses fail: The propagation of attack effects is determined by $H_{t,i}^k$ during local client training, which is inaccessible to the central server.

Previous Client-Side Defense (FL-WBC)[2]: Adds random noise to reduce Hessian sparsity.

- Problem: Uncalibrated noise degrades model accuracy

Core Idea of LeadFL: Perturb the Hessian matrix to minimize the coefficient $(I - \eta_t H_{t,i}^k)$, reducing the lingering attack effect.

- Problem: Computing Hessian Matrix is expensive

Hessian Matrix Approximaton[3]

- **Diagonalization:** We approximate H using only its diagonal elements to reduce complexity:

$$H \approx \text{diag}(H)$$

- **Finite Difference:** The diagonal is estimated via the change in gradients between iterations:

$$H \approx \text{diag}\left(\nabla L(\theta_{t,i+1}^k) - \nabla L(\theta_{t,i}^k)\right)$$

- **Parameter Estimation:** To avoid extra backpropagation, we approximate gradient changes using model weight differences:

$$\tilde{H}_{t,i}^k \approx \frac{\text{diag}(\tilde{\theta}_{t,i+1}^k - \theta_{t,i}^k - \Delta\theta_{t,i}^k)}{\eta_t} \tag{3}$$

LeadFL Solution

LeadFL Client-Side Defense: A secondary backpropagation process is deployed utilizing a regularization term to minimize the coefficient involved in the propagation of the AEP.

Step 1: $$\underbrace{\tilde{\theta}_{t,i+1}^k \leftarrow \theta_{t,i}^k - \eta_t \nabla L(\theta_{t,i}^k)}_{\text{Local Training (Standard SGD)}}$$

Step 2: $$\theta_{t,i+1}^k \leftarrow \tilde{\theta}_{t,i+1}^k - \underbrace{\eta_t \alpha \,\text{clip}\left[\nabla(I - \eta_t \tilde{H}_{t,i}^k), q\right]}_{\text{Regularization Term}}$$

Where $\tilde{H}_{t,i}^k$ denotes the estimated diagonal of the Hessian Matrix, η_t is the learning rate, α represents the regularization rate controlling perturbation magnitude, and clip$(\cdot, q)$ is an element-wise clipping function with threshold q ensuring theoretical convergence.

Evaluation

- Dataset: FashionMNIST
- 100 clients (25% malicious), 10 selected per round
- Attacks: 9-pixel backdoor (Periodic burtsy attacks)

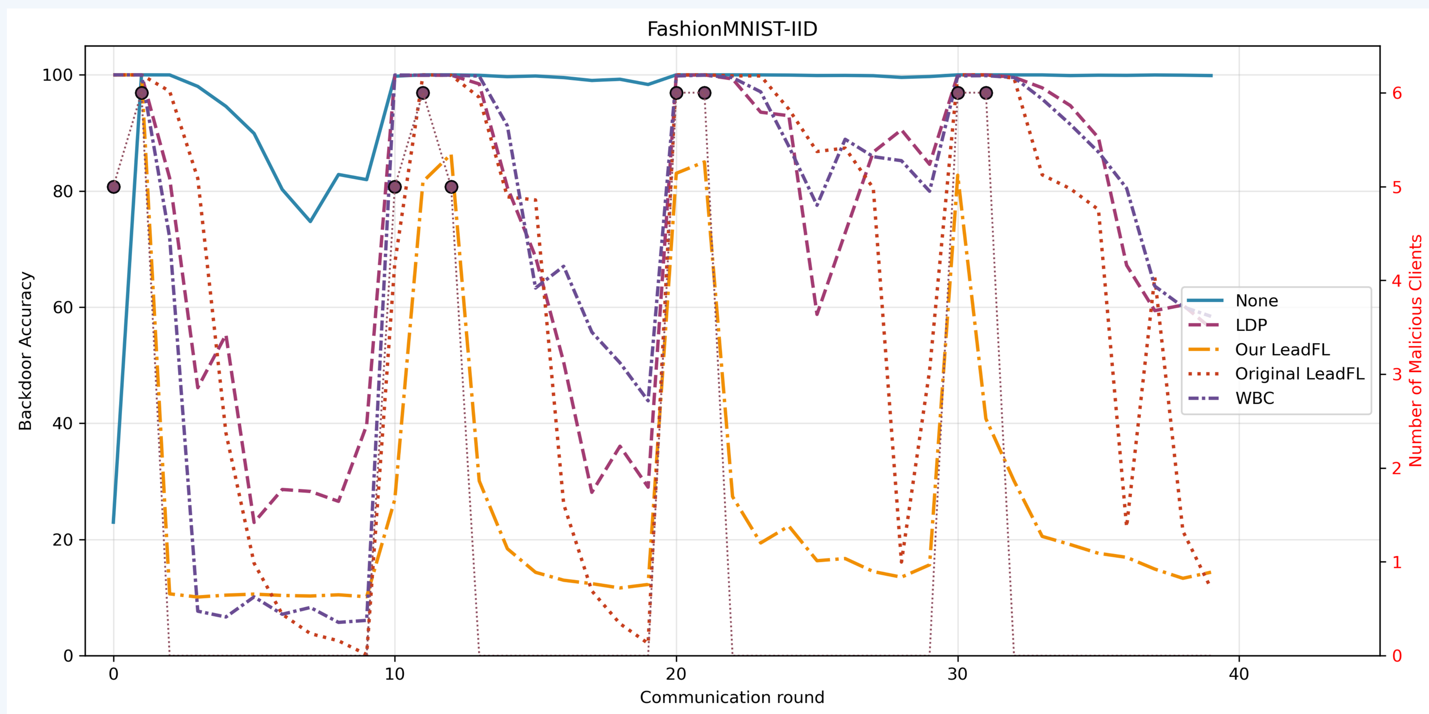

Fig 2: Backdoor Accuracy Comparison of Different Client-side Defenses

Client-Side Defense: Comparison of defenses under a 9-pixel pattern backdoor attack (periodic patterns) with Bulyan server-side defense on IID and non-IID FashionMNIST datasets in Fig 3 and Table 1.

	IID					Non-IID				
	None	LDP	Our LeadFL	Original LeadFL	WBC	None	LDP	Our LeadFL	Original LeadFL	WBC
MA	88.8	85.0	72.9	86.9	85.8	60.6	74.3	38.1	73.0	65.0
BA Avg	95.5	73.1	29.4	62.2	70.8	97.2	76.8	25.6	54.3	68.0
BA Final	99.9	56.4	14.4	11.6	58.4	99.2	67.6	2.7	18.8	45.6

Table 1

Regularization Analysis: Analyzed the effect of varying the regularization rate in Fig. 3.

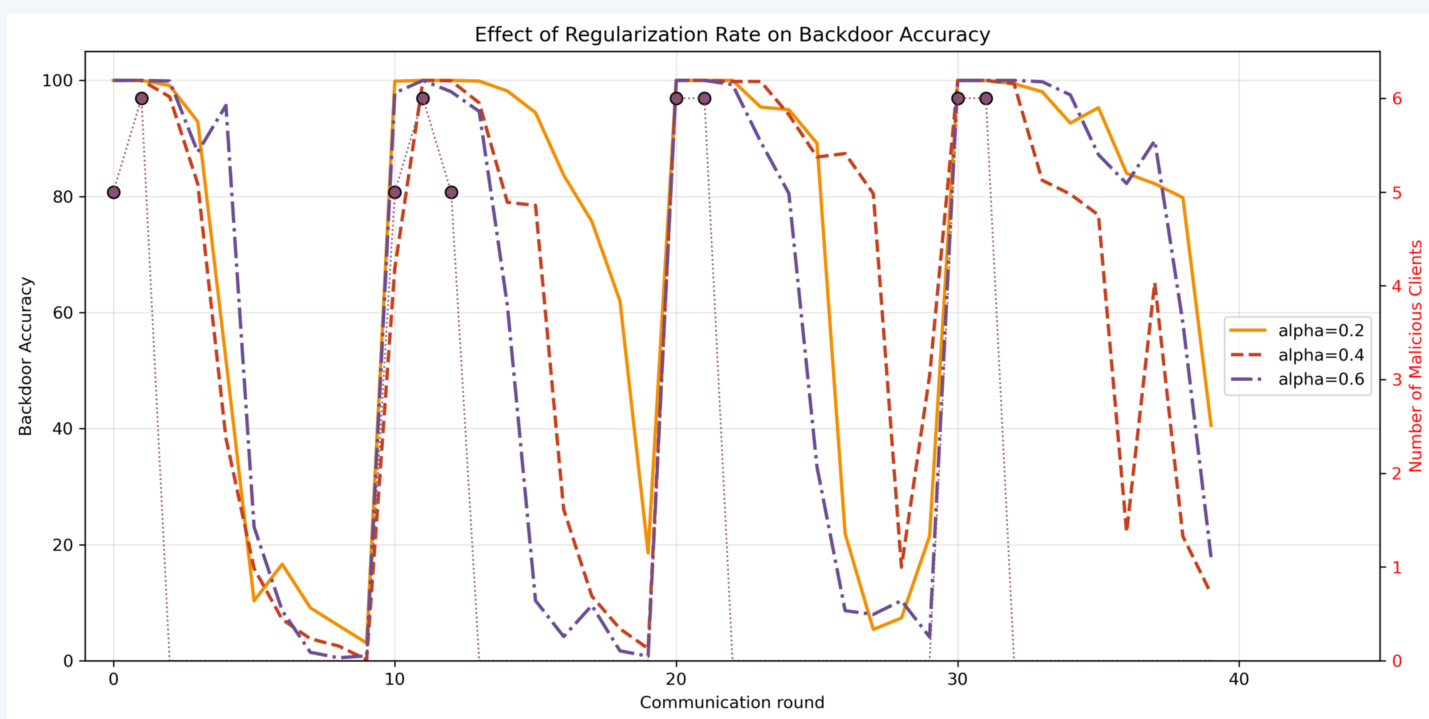

Fig 3: Backdoor Accuracy Comparison of Different Regularization Rates

Conclusion

Increase in attack recovery: LeadFL mitigates the lingering impact of bursty poisoning attacks by regularizing the client-side Hessian matrix.

Decline in main task accuracy: Theoretical and empirical evaluations prove that LeadFL defend against attacks with a low degradation of the main task accuracy.

References

[1] Zhu, C., Roos, S., & Chen, L. Y. (2023). LeadFL: Client Self-Defense against Model Poisoning in Federated Learning. *ICML 2023.*

[2] Sun, J., et al. (2021). FL-WBC: Enhancing Robustness Against Model Poisoning Attacks in Federated Learning from a Client Perspective. *NeurIPS 2021.*

[3] LeCun, Y., Denker, J. S., & Solla, S. A. (1990). Optimal Brain Damage. *Advances in Neural Information Processing Systems 2 (NIPS 1989)*, 598-605.