GAN-inspired Defense Against Backdoor Attacks on Federated Learning Systems

Faculty Research and Industry Networking Day

Prof. Feng Li | **Computer and Information Technology** | **SMART Lab Indy**

/Abstract

Federated Learning (FL) enables collaborative model training while preserving data privacy, but backdoor poisoning attacks—where malicious clients implant triggers to alter model behavior—pose a significant challenge. This study introduces a GAN-inspired approach using clients as Generators and the server as a Discriminator to detect backdoors. Using client-generated labelled backdoor and benign models, we train a binary classifier at the server to effectively distinguish backdoor-affected models from benign ones, enhancing detection of malicious activity in FL.

/Research Questions

- Can we build a binary classifier to differentiate between backdoor and benign model?
- How can we overcome the unavailability of labelled backdoor and benign model data?
- Can we leverage the help of clients to defend against backdoor attacks?

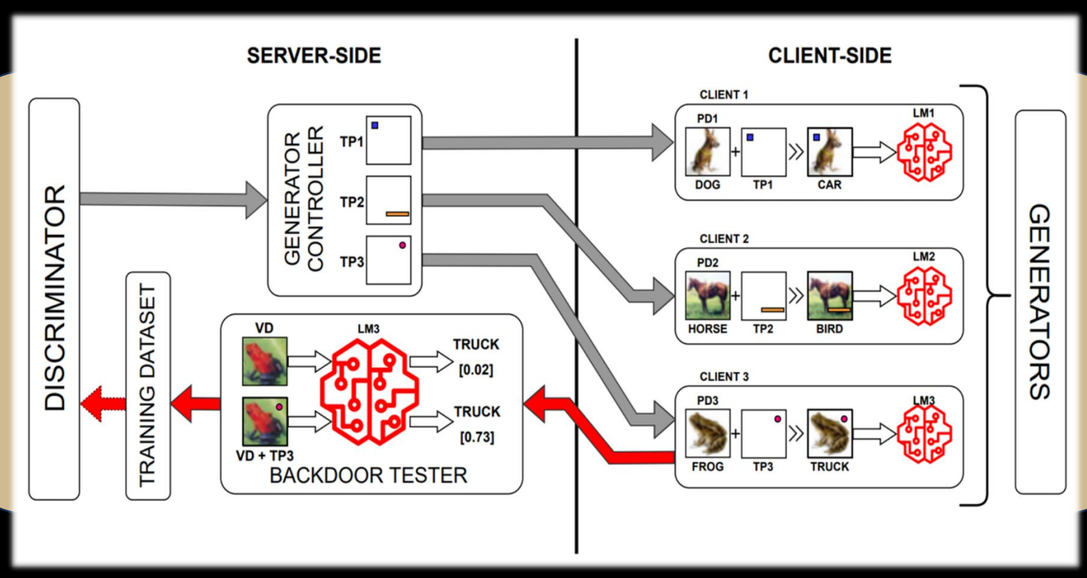
METHODOLOGY

/Client - Generators

The clients assist the server in building a binary classifier by providing the necessary training dataset. They are divided into several rotating groups contributing to different labels.

LABELED DATA GENERATION

- *BENIGN MODEL*: Models are labeled as benign after regular model updates from clients are clipped and rescaled to prevent unknown backdoor influence.
- *BACKDOORED MODEL*: The server shares the backdoor trigger and specifications with clients, which they use during model training. Such models are labeled as backdoored after undergoing backdoor testing.

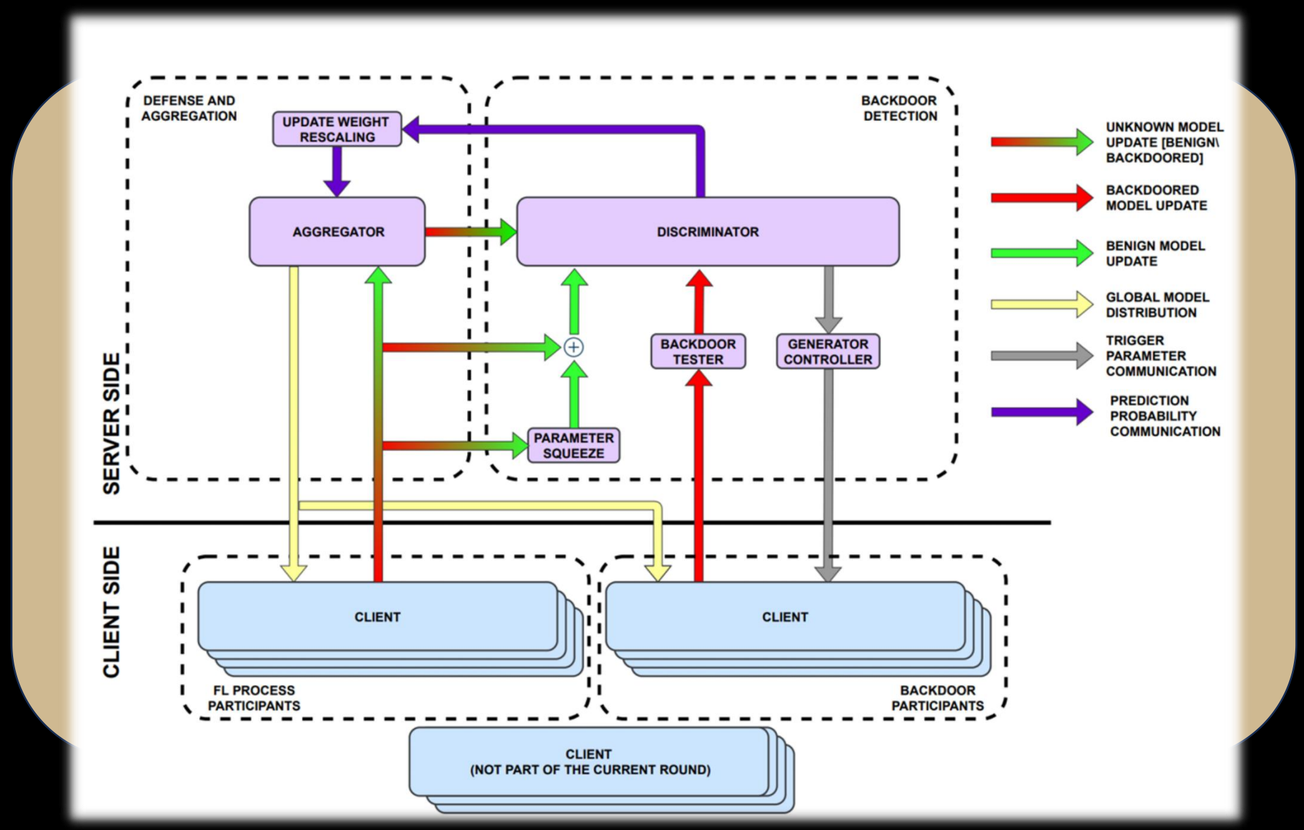


/Server - Discriminator

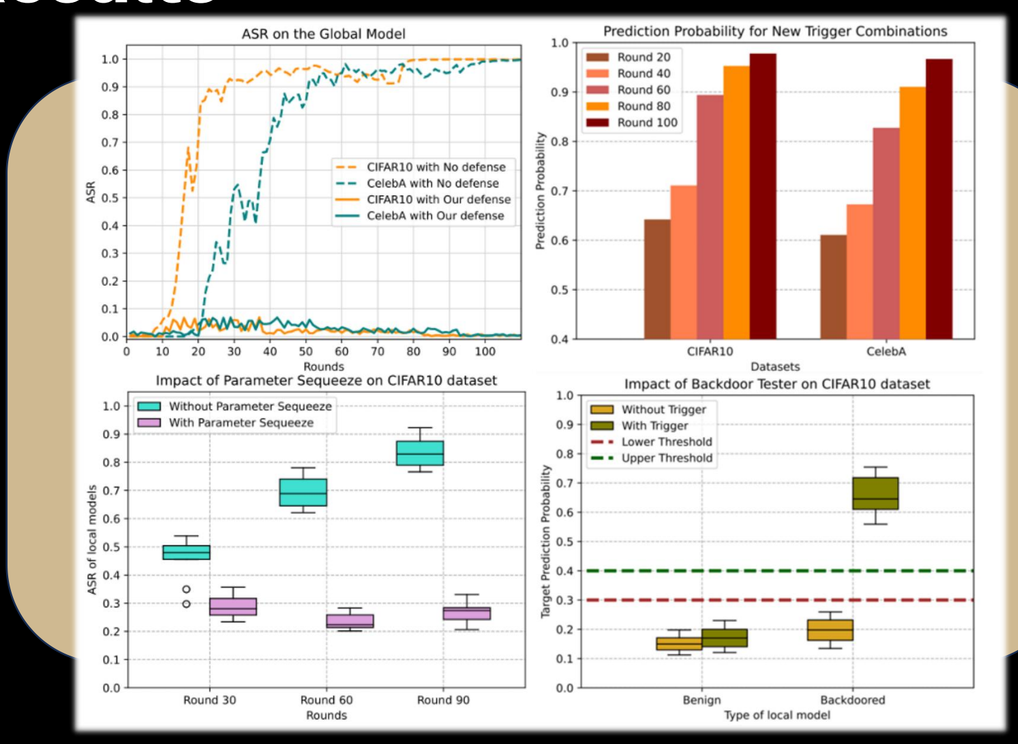
Along with performing the federated aggregation step, the server is also responsible for building the binary classifier as the Discriminator and countering potential poisoned labeled data provided by malicious clients.

ENSURING LABELED DATA QUALITY:

- *PARAMETER SQUEEZING*: Applies perturbation techniques to submitted benign data to remove any potential backdoor influence.
- *GENERATOR CONTROLLER*: Determines the specifications for backdoor data generation based on the discriminator's performance.
- *BACKDOOR TESTER*: Evaluates the received backdoor models to ensure they align with the specifications set by the generator controller.



/Results



PURDUE UNIVERSITY | Polytechnic Institute