MPAF: Model Poisoning Attacks to Federated Learning based on Fake Clients

Xiaoyu Cao and Neil Zhenqiang Gong
Federated Learning
FL is Vulnerable to Poisoning Attacks
Limitations of Existing Attacks

• High cost
  – Based on a large fraction (e.g., 10%) of compromised genuine clients
  – Need to bypass malware detection systems

• Sub-optimal overall attack effect
  – Round-wise optimization problems
  – Attack effect may cancel out in multiple rounds
Fake-client-based Attacks

Key challenge: NO training data available

Low cost
No malware detection
Our Design

\[
\min_{\{g_i^t\}} \| w^T - w' \|
\]

- Fake model updates on \(i\)-th client in \(t\)-th round
- Base model with low accuracy
- Final model under attack
Our Design

\[ w^{t-1} \]
\[ w^t \]
\[ w^{t+1} \]
\[ w^* \]

fake local model updates

genuine local model updates
Evaluations

Purchase dataset, 1000 genuine clients, Trimmed-mean

Our attack can significantly decrease the test accuracy

- Use random noise as fake local model updates
- Reverse history global updates as fake local model updates
Takeaways

• Poisoning attacks based on fake clients can be as effective while incurring much lower cost

• Poisoning attacks based on fake clients and global optimization problems deserve more interest
Thank you!