FLAME: Taming Backdoors in Federated Learning

Thien Duc Nguyen1, Phillip Rieger1, Huili Chen2, Hossein Yalame1, Helen Mollering1, Hossein Fareidooni1, Samuel Marchal3, Markus Miettinen1, Azalia Mirhoseini4, Shaza Zeitouni1, Farinaz Koushanfar2, Ahmad-Reza Sadeghi1, and Thomas Schneider1

1Technical University of Darmstadt, Germany; 2University of California San Diego, USA; 3Aalto University and F-Secure, Finland; 4Google, USA

The 31st USENIX Security Symposium, 2022
Big Picture

• Poisoning attacks on Federated Learning
  • Deteriorate model performance or **inject backdoors**
  • Existing defenses are not effective

• Our solution: **FLAME**
  • Eliminates poisoned updates effectively
  • Maintains model performance
  • Preserves privacy of clients’ data (based on Secure Two-Party Computation)
Federated Learning: Basics

Aggregation at round $t$:

$$G_t \leftarrow \sum_{k=1}^{K} \frac{n^k}{n} w^k_t$$

- $G_t$: Parameters of aggregated model
- $w^k_t$: Parameters of client’s model
- $K$: Total number of clients
- $n^k$: Number of samples for client $k$
- $n$: Number of samples for all clients
Federated Learning Applications

Autonomous Driving

Nature Language Processing

Medical Image Processing

Financial Crime Detection

Network Intrusion Detection

Cyber-Risk Intelligence

[Jallepalli et al. BigDataService 2021]

[McMahan et al. Google AI 2017]

[Sheller et al. Intel AI 2018]

[Yang et al. BIGDATA 2019]

[Nguyen et al. ICDCS 2019]

[Fereidooni et al. NDSS 2022]
Security & Privacy of Federated Learning
Poisoning Attacks on Federated Learning

Model poisoning
Manipulate training process or craft model updates

Data poisoning
Manipulate the local data used for training

Adversary Model
• The adversary has full control of the compromised clients
• No control of benign clients and the aggregator
• The majority of clients are benign
Poisoning Attacks on Federated Learning

Untargeted Poisoning
Renders the ML model useless (Denial-of-Service)

Targeted Poisoning (Backdoor)
Injecting malicious functionality using predefined (triggered) inputs
Examples of Backdoor Attacks: Adversary Chosen Label

**Word prediction**
Select end words, e.g.,
• "buy a CPU from AMD"

[Bagdasaryan et al. AISTATS 2020]

**Image classification**
Change labels, e.g.,
• Speed limit signs from 30kph to 80kph

[Shen et al. ACSAC 2016]

**IoT malware detection**
Inject malicious traffic, e.g., use compromised IoT devices

[Nguyen et al. DISS@NDSS 2020]
Existing Defenses Against Backdoor Attacks

Update Smoothing

- Differential Privacy
  - Bagdasaryan et al. AISTATS 2020
  - Naseri et al. NDSS 2022

- Gradient Pruning (GP)
  - Mondal et al. PPAI 2022

Distribution-based Filtering

- Outlier Detection
  - Shen et al. 2016 CCS
  - Munoz et al. arXiv 2019

- Sybil Detection
  - Fung et al. RAID 2020
  - Awan et al. ESORICS 2021

It is challenging to determine clipping bound and noise-level
  - Bagdasaryan et al. AISTATS’20

Make specific assumptions about attack strategy and data
  - Shen et al. CCS’19, Fung et al. RAID’20
FLAME Overview

Dynamic Model Filtering → Adaptive Model Clipping → Adaptive Noising

Aggregator

Client

Client

Client

Client

Client
Backdoor Characteristics

\( \vec{W}_1' \) (Large angular deviation)

\( \vec{W}_2' \) (Large magnitude)

\( \vec{W}_3' \) (Stealthy)

Global mode from training round t-1
Benign models at round t
Malicious models at round t

\( S_t \): Clipping bound, \( \sigma_t \): Noise level
FLAME: Dynamic Model Filtering

\[ W_i \leftarrow \text{Client Update}(G_{t-1}) \]
\[ (c_{11}, ..., c_{nn}) \leftarrow \text{Cosine Distance}(W_1, ..., W_n) \]
\[ (b_1, ..., b_L) \leftarrow \text{Clustering}(c_{11}, ..., c_{nn}) \]

- Global mode from training round \( t-1 \)
- Benign models at round \( t \)
- Malicious models at round \( t \)

HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
FLAME: Adaptive Model Clipping

\[ S_t \] : Clipping bound

\[ G_{t-1} \rightarrow W, \quad (\text{Large magnitude}) \]

\[ S_t \leftarrow \text{Median}(e_1, ..., e_n) \]

\[ w_j \leftarrow G_{t-1} + (W_j - G_{t-1}) \times \text{Min}\left(1, \frac{S_t}{e_j}\right) \forall \in \{b_1, ..., b_L\} \]

\[ (e_1, ..., e_n) \leftarrow \text{Euclidean\_Distance}(G_{t-1}, (W_1, ..., W_n)) \]

\[ S_t \leftarrow \text{Median}(e_1, ..., e_n) \]

\[ w_j \leftarrow G_{t-1} + (W_j - G_{t-1}) \times \text{Min}\left(1, \frac{S_t}{e_j}\right) \forall \in \{b_1, ..., b_L\} \]

\[ L_2\text{-norms} \]

\[ L_2\text{-norms (Euclidean distances) of model updates depending on the training rounds and datasets} \]
FLAME: Adaptive Noising - Theoretical Background

- Differential Privacy negative impact of individual (backdoor) samples e.g., [Du et al. ICLR 2020]:
  \[ Pr[M(D_1) \in B] \leq e^\varepsilon \cdot Pr[M(D_2) \in B] + \delta \]

- We prove that backdoor resilience from centralized learning can be transformed to federated learning

- Determine \( \sigma_t \) dynamically based on \( S_t \)

- Clipping and filtering reduce necessary noise, i.e., minimize the effect on the model performance

\[
G_t \leftarrow \sum_{j \in \{b_1, \ldots, b_L\}} \frac{W_j}{L} \\
G_t \leftarrow G_t + N(O, \sigma_t^2) \text{ where } \sigma_t^2 \leftarrow \sqrt{\frac{S_t \cdot 2 \ln \left( \frac{1.25}{\delta} \right)}{\varepsilon}}
\]
## Evaluation

<table>
<thead>
<tr>
<th>ATTACK</th>
<th>Dataset</th>
<th>No Defense</th>
<th>FLAME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BA</td>
<td>MA</td>
</tr>
<tr>
<td>Constrain-and-Scale</td>
<td>Reddit</td>
<td>100.0</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>CIFAR-10</td>
<td>81.9</td>
<td>89.8</td>
</tr>
<tr>
<td></td>
<td>IoT-Traffic</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Distributed Backdoor Attack</td>
<td>CIFAR-10</td>
<td>93.8</td>
<td>57.4</td>
</tr>
<tr>
<td>[Xie et al. ICLR 2020]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge-Case</td>
<td>CIFAR-10</td>
<td>42.8</td>
<td>84.3</td>
</tr>
<tr>
<td>[Wang et al. NeurIPS 2020]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projected Gradient Decent</td>
<td>CIFAR-10</td>
<td>56.1</td>
<td>68.8</td>
</tr>
<tr>
<td>[Wang et al. NeurIPS 2020]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untargeted Poisoning</td>
<td>CIFAR-10</td>
<td>-</td>
<td>46.7</td>
</tr>
<tr>
<td>[Fang et al. USENIXSec 2020]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BA: Backdoor Accuracy, MA: Main Task Accuracy
## FLAME vs. Existing Defenses

<table>
<thead>
<tr>
<th>Defenses</th>
<th>Reddit BA</th>
<th>Reddit MA</th>
<th>CIFAR-10 BA</th>
<th>CIFAR-10 MA</th>
<th>IoT-Traffic BA</th>
<th>IoT-Traffic MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign Setting</td>
<td>100.0</td>
<td>22.6</td>
<td>-</td>
<td>92.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>No defense</td>
<td>-</td>
<td>22.7</td>
<td>81.9</td>
<td>89.8</td>
<td>-</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Krum</strong> [Blanchard et al. NIPS 2017]</td>
<td>100.0</td>
<td>9.6</td>
<td>100.0</td>
<td>56.7</td>
<td>100.0</td>
<td>84.0</td>
</tr>
<tr>
<td><strong>FoolsGold</strong> [Fung et al. RAID 2020]</td>
<td>0.0</td>
<td>22.5</td>
<td>100.0</td>
<td>52.3</td>
<td>100.0</td>
<td>99.2</td>
</tr>
<tr>
<td><strong>Auror</strong> [Shen et al. ACSAC 2016]</td>
<td>100.0</td>
<td>22.5</td>
<td>100.0</td>
<td>26.1</td>
<td>100.0</td>
<td>96.6</td>
</tr>
<tr>
<td><strong>AFA</strong> [Muñoz-González et al. arXiv 2019]</td>
<td>100.0</td>
<td>22.4</td>
<td>0.0</td>
<td>91.7</td>
<td>100.0</td>
<td>87.4</td>
</tr>
<tr>
<td><strong>DP</strong> [Sun et al. NeurIPS 2019]</td>
<td>14.0</td>
<td>18.9</td>
<td>0.0</td>
<td>78.9</td>
<td>14.8</td>
<td>82.3</td>
</tr>
<tr>
<td><strong>Median</strong> [Yin et al ICML 2018]</td>
<td>0.0</td>
<td>22.0</td>
<td>0.0</td>
<td>50.1</td>
<td>0.0</td>
<td>87.7</td>
</tr>
<tr>
<td><strong>FLAME</strong></td>
<td>0.0</td>
<td>22.3</td>
<td>0.0</td>
<td>91.9</td>
<td>0.0</td>
<td>99.8</td>
</tr>
</tbody>
</table>

BA: Backdoor Accuracy, MA: Main Task Accuracy
Comparison between FLAME and a combination of existing defenses against constrain-and-scale attack [Bagdasaryan et al. AISTATS 2020] on the CIFAR-10 dataset.
Private FLAME: Privacy Preserving Aggregation
Using secure Multi-party Computation
Private FLAME: Motivation

• Privacy attack: A curious aggregator can learn information from the training data by inference attacks
  • E.g., [Pyrgelis et al. NDSS 2018, Shokri et al. S&P 2017]

• Existing defenses prohibit access to the model updates to investigate backdoors
  • E.g., [Bonawitz et al. CCS 2017, Kairouz et al. PMLR 2021]

• Our goal: Introduce private FLAME such that FLAME algorithms are computed under encryption
Private FALME: Solution

Utilizing Secure Two-Party Computation (STPC)

1. Client i splits the parameters of $W_i$ into two secret shares $\langle X \rangle_i^A$ and $\langle X \rangle_i^B$
   - Each aggregator only receives an encrypted part of $W_i$

2. Two aggregators use STPC to run FLAME algorithms (clustering and clipping) on the secret shares
   - The aggregators learn nothing about $W_i$
Private FLAME: Evaluation

- The runtime of Private FLAME is significantly higher than standard FLAME.
- However, such runtime overhead would be acceptable to maintain privacy.
- Private FLAME provides similar results w.r.t accuracy in comparison to standard FLAME.

<table>
<thead>
<tr>
<th>#client</th>
<th>Reddit</th>
<th>CIFAR-10</th>
<th>IoT-Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>P</td>
<td>S</td>
</tr>
<tr>
<td>10</td>
<td>2.35</td>
<td>519.92</td>
<td>0.32</td>
</tr>
<tr>
<td>50</td>
<td>62.55</td>
<td>5895.70</td>
<td>2.62</td>
</tr>
<tr>
<td>100</td>
<td>252.13</td>
<td><strong>22081.65</strong></td>
<td>8.56</td>
</tr>
</tbody>
</table>

BA: Backdoor Accuracy, MA: Main Task Accuracy, TPR: True Positive Rate, TNR: True Negative Rate

Runtime in sec. of standard FLAME (S) compared to private FLAME (P) using secure two-party computation.

### Effectiveness

<table>
<thead>
<tr>
<th></th>
<th>Reddit</th>
<th>CIFAR-10</th>
<th>IoT-Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>P</td>
<td>S</td>
</tr>
<tr>
<td>BA</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>MA</td>
<td>22.3</td>
<td>22.2</td>
<td>91.9</td>
</tr>
<tr>
<td>TPR</td>
<td>22.2</td>
<td>20.4</td>
<td>23.8</td>
</tr>
<tr>
<td>TNR</td>
<td>100.0</td>
<td>100.0</td>
<td>86.2</td>
</tr>
</tbody>
</table>

Effectiveness of standard FLAME (S) in comparison to private FLAME (P) using secure two-party computation.
Conclusion & Future Work

• FLAME, a novel backdoor defense for FL:
  • Mitigates state-of-the-art backdoor attacks effectively
  • Negligible impact on the benign performance of the models
  • Preserves privacy of clients’ data

• Working on privacy-preserving poisoning defenses
  • Improving computational efficiency