FLAME: Taming Backdoors in Federated Learning

Thien Duc Nguyen[1], Phillip Rieger[1], Huili Chen[2], Hossein Yalame[1], Helen Mollering[1], Hossein Fereidooni[1], Samuel Marchal[3], Markus Miettinen[1], Azalia Mirhoseini[4], Shaza Zeitouni[1], Farinaz Koushanfar[2], Ahmad-Reza Sadeghi[1], and Thomas Schneider[1]

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

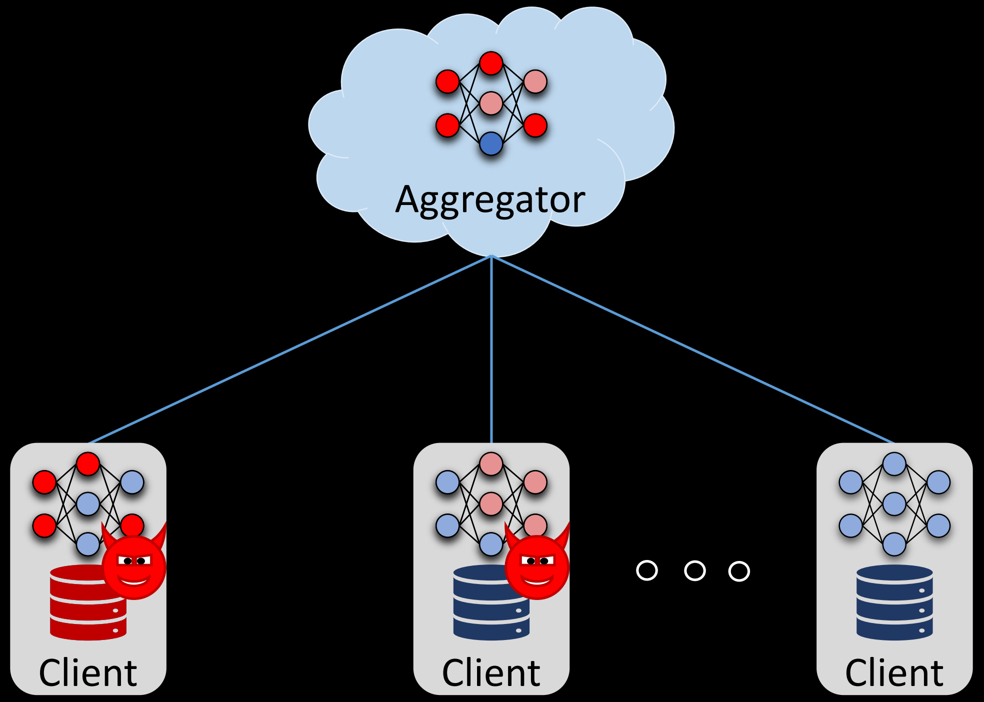
The 31st USENIX Security Symposium, 2022

System Security Lab

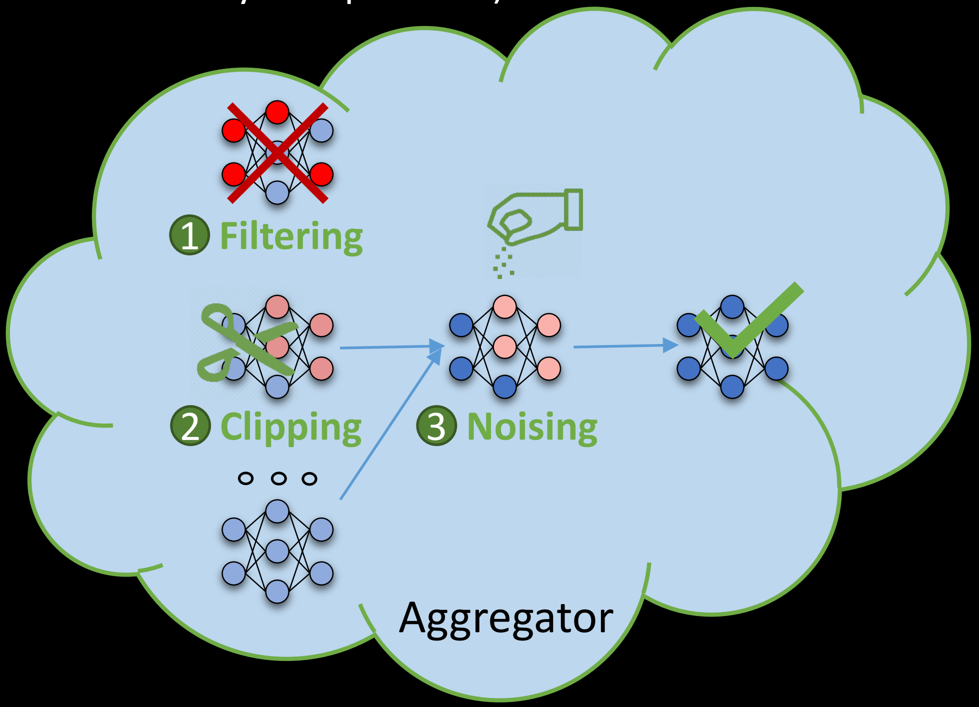
TECHNISCHE UNIVERSITÄT DARMSTADT

Big Picture

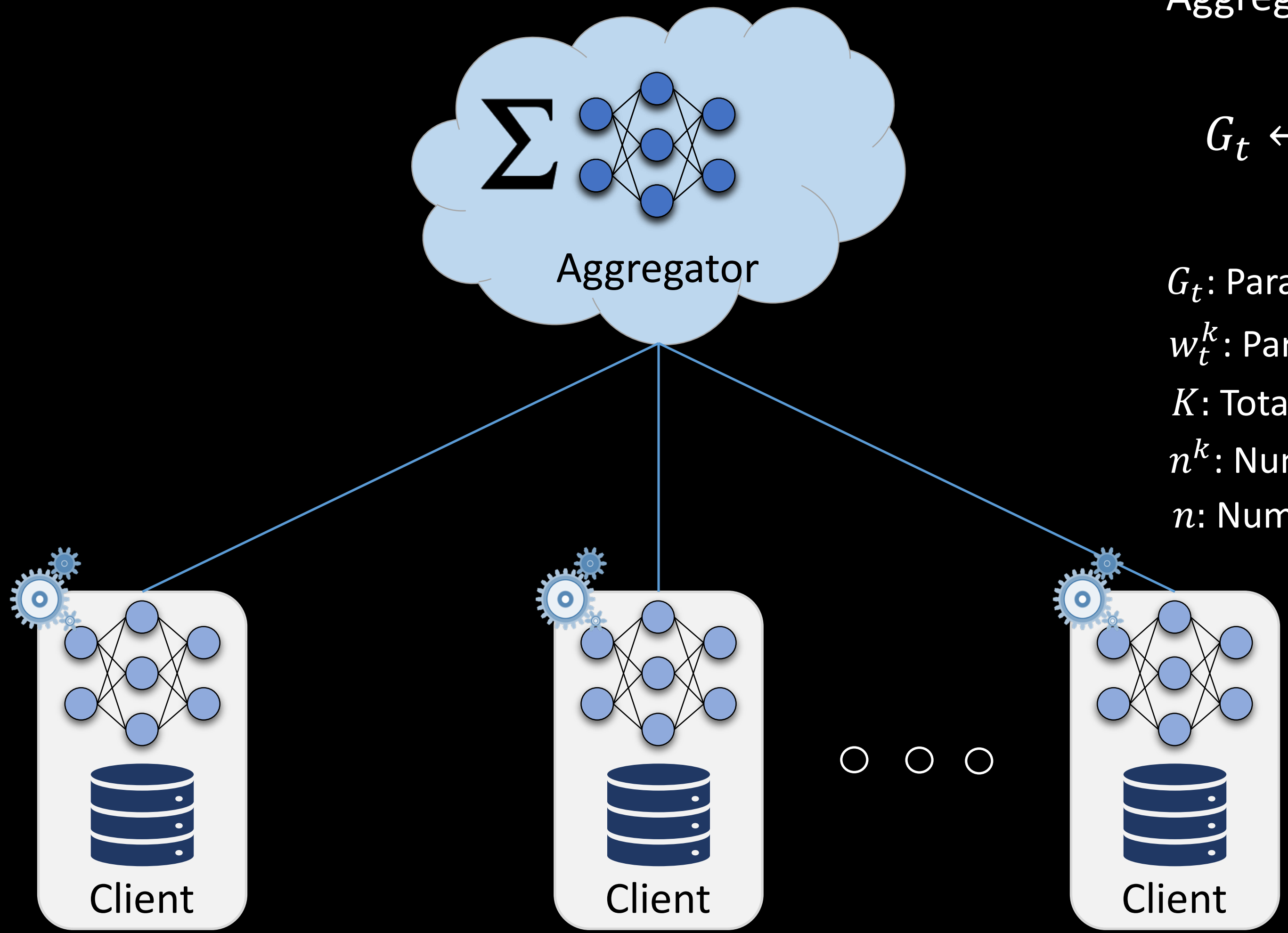
- Poisoning attacks on Federated Learning
  - Deteriorate model performance or **inject backdoors**
  - Existing defenses are not effective



- 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:

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

G_t: Parameters of aggregated model

w_t^k: Parameters of client's model

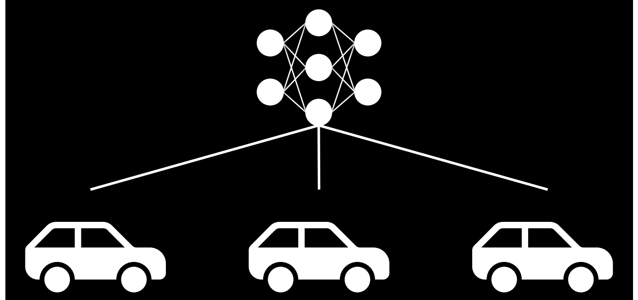
K: Total number of clients

n^k: Number of samples for client k

n: Number of samples for all clients

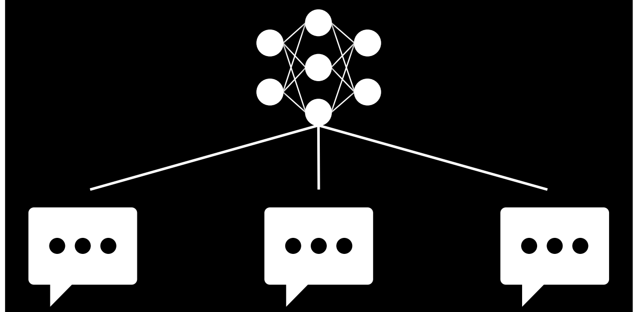
Federated Learning Applications

Autonomous Driving



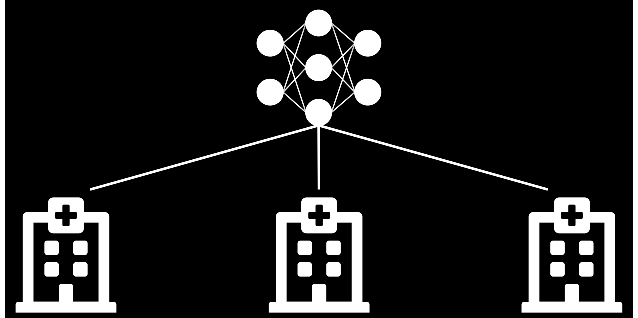
[Jallepalli et al. BigDataService 2021]

Nature Language Processing



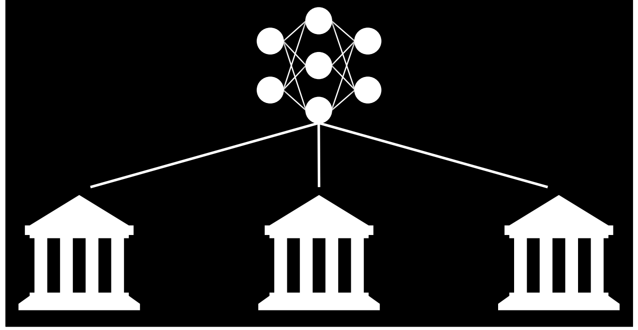
[McMahan et al. Google AI 2017]

Medical Image Processing



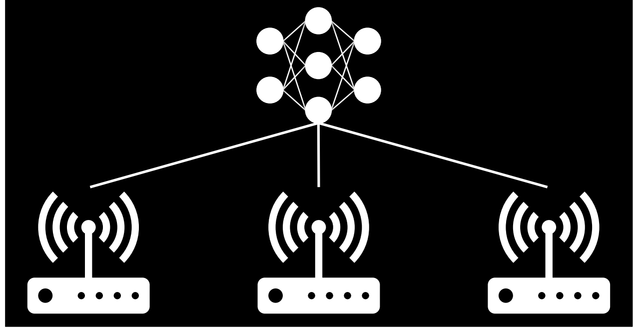
[Sheller et al. Intel AI 2018]

Financial Crime Detection



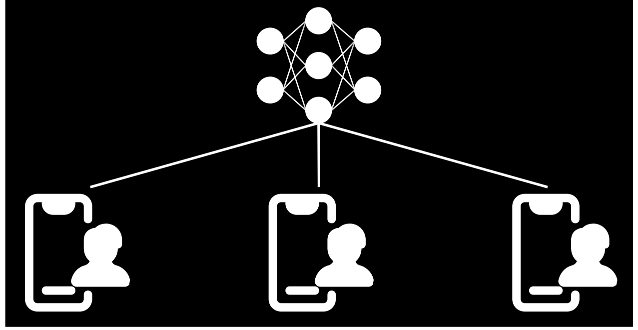
[Yang et al. BIGDATA 2019]

Network Intrusion Detection



[Nguyen et. al ICDCS 2019]

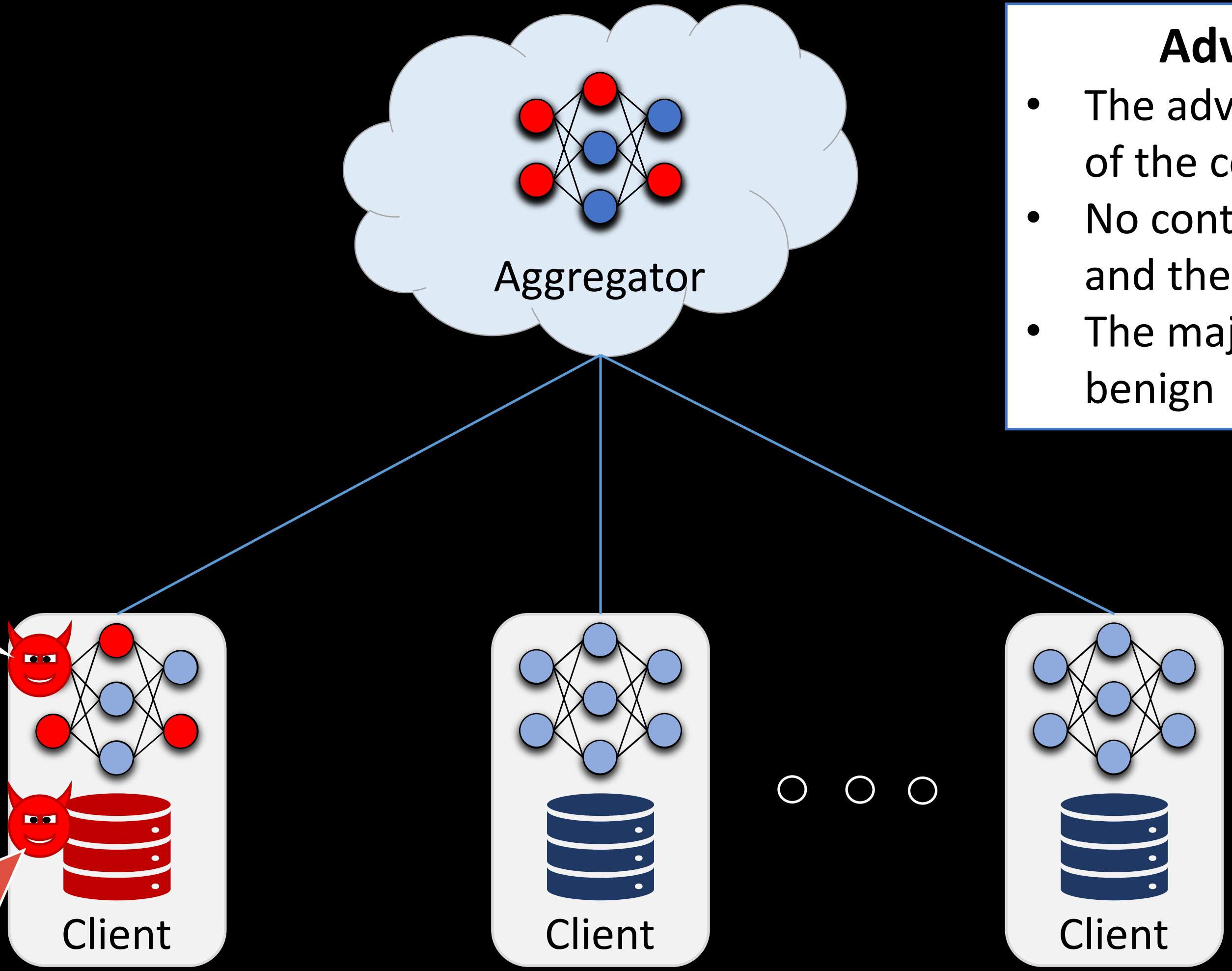
Cyber-Risk Intelligence



[Fereidooni et. al NDSS 2022]

Security & Privacy of Federated Learning

Poisoning Attacks on Federated Learning



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

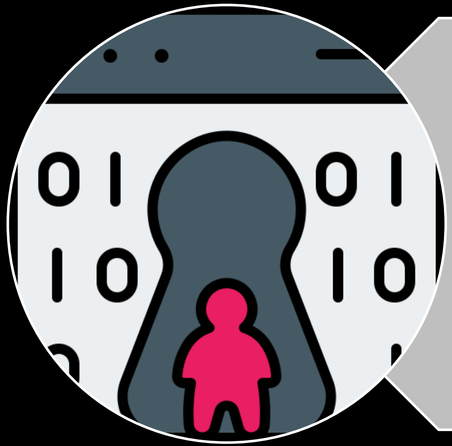
Model poisoning
Manipulate training process or craft model updates

Data poisoning
Manipulate the local data used for training

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]

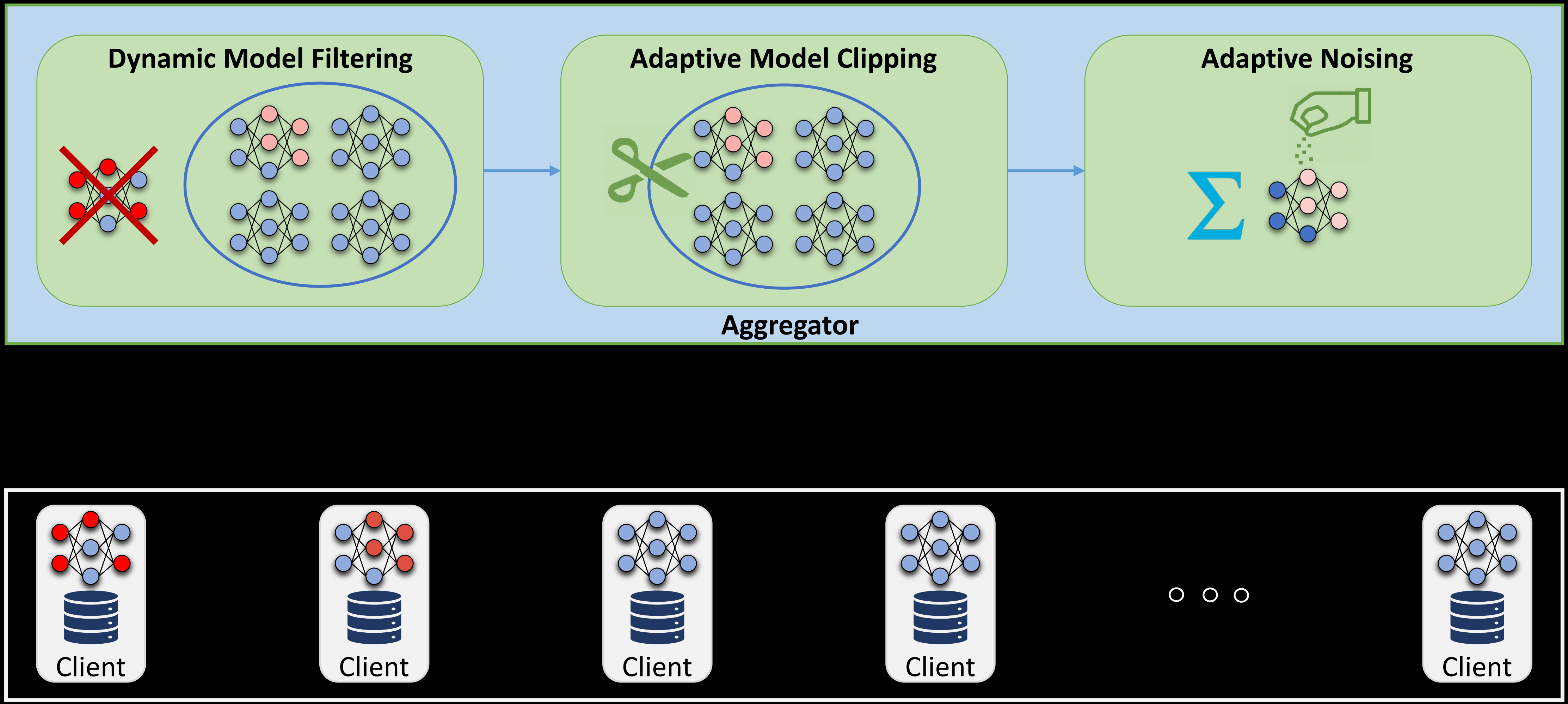
It is challenging to determine clipping bound and noise-level
[Bagdasaryan et al. AISTATS'20]

Distribution-based Filtering

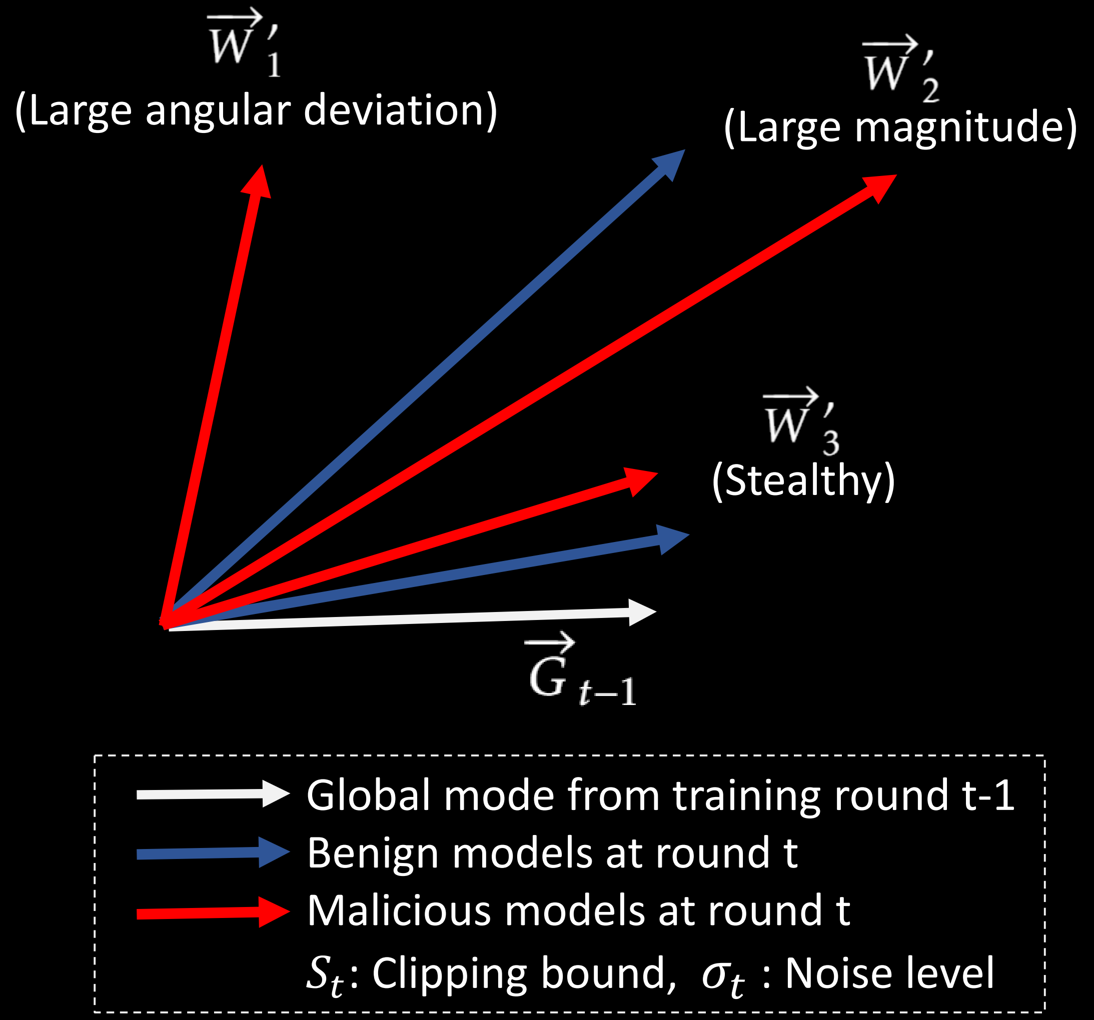
- **Outlier Detection**
  [Shen et al. 2016 CCS]
  [Munoz et al. arXiv 2019]
- **Sybil Detection**
  [Fung et al. RAID 2020]
  [Awan et al. ESORICS 2021]

Make specific assumptions about attack strategy and data
[Shen et al. CCS'19],[Fung et al. RAID'20]

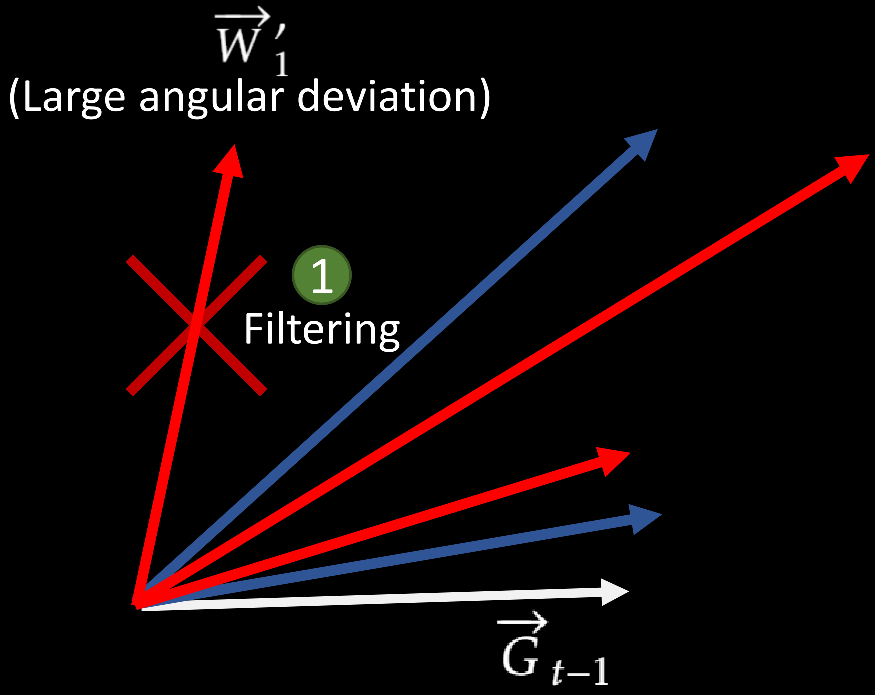
FLAME Overview



Backdoor Characteristics

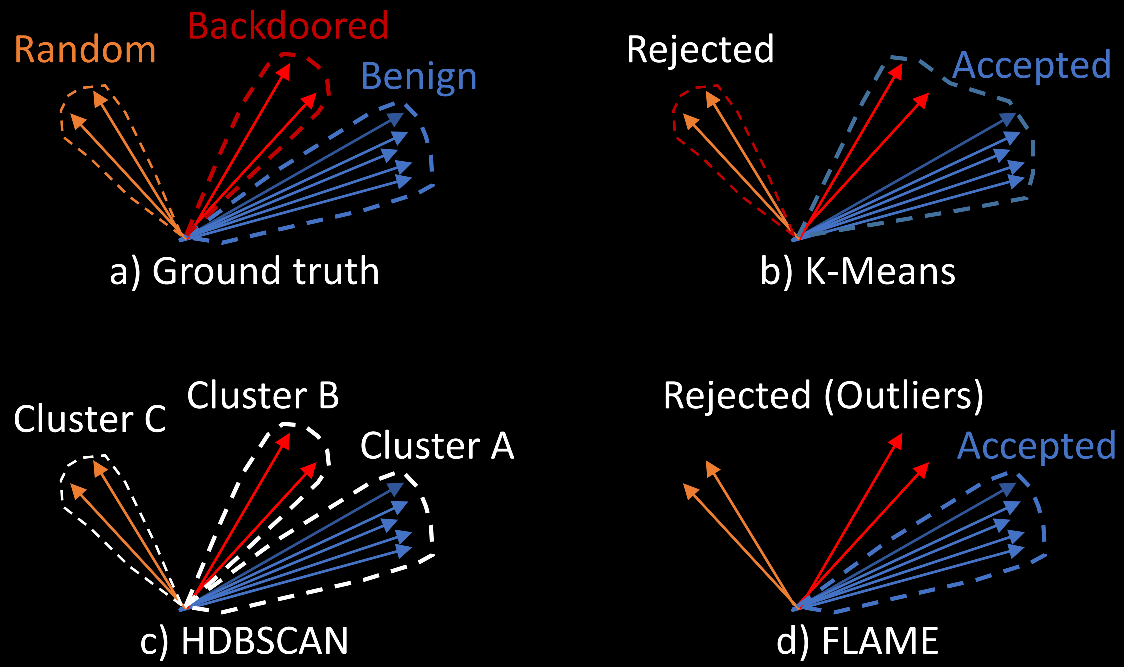


FLAME: Dynamic Model Filtering



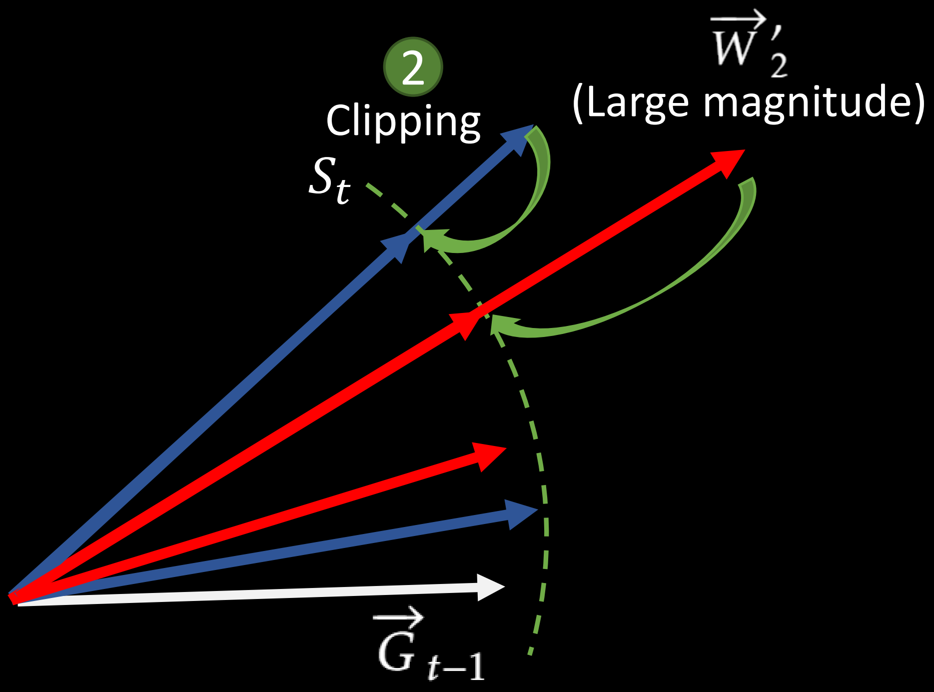
- Global mode from training round t-1
- Benign models at round t
- Malicious models at round t

$$W_i \leftarrow Client\_Update(G_{t-1})$$
$$(c_{11}, \dots, c_{nn}) \leftarrow Cosine\_Distance(W_1, \dots, W_n)$$
$$(b_1, \dots, b_L) \leftarrow Clustering(c_{11}, \dots, c_{nn})$$



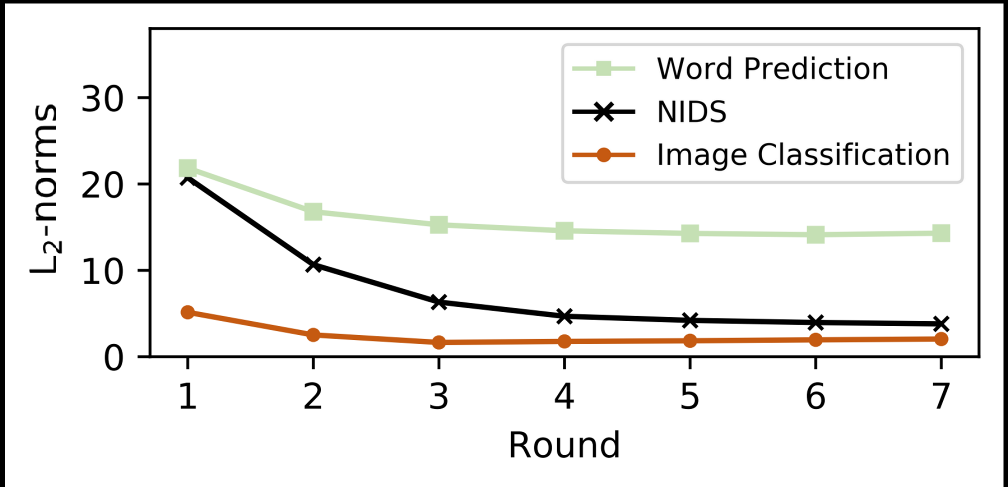
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise

FLAME: Adaptive Model Clipping



Global mode from training round t-1
Benign models at round t
Malicious models at round t
S_t: Clipping bound

$$(e_1, \dots, e_n) \leftarrow Euclidean\_Distance(G_{t-1,} (W_1, \dots, W_n))$$
$$S_t \leftarrow Median(e_1, \dots, e_n)$$
$$w_j \leftarrow G_{t-1} + (W_j - G_{t-1}) * Min\left(1, \frac{S_t}{e_j}\right) \forall \in \{b_1, \dots, b_L\}$$

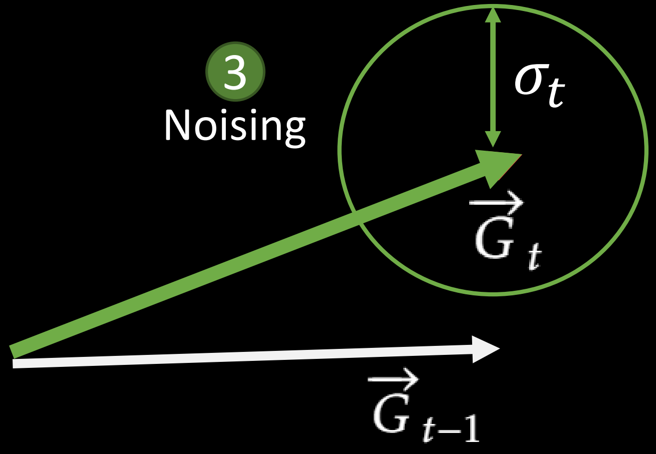


L_2-norms (Euclidean distances) of model updates depending on the training rounds and datasets

FLAME: Adaptive Noising - Theoretical Background

$$G_t \leftarrow \sum_{j \in \{b_1, \ldots, b_L\}} \frac{W_j}{L}$$

$$G_t \leftarrow G_t + N(0, \sigma_t^2) \ \ where \ \ \sigma_t^2 \leftarrow \frac{S_t \cdot \sqrt{2 \ln(\frac{1.25}{\delta})}}{\varepsilon}$$



Global mode from training round t-1
Benign models at round t
Malicious models at round t
S_t: Clipping bound, σ_t : Noise level

- Differential Privacy negative impact of individual (backdoor) samples e.g., [Du et at. ICLR 2020]:

$$Pr[M(D_1) \in \mathcal{O}] \leq e^{\epsilon} \cdot Pr[M(D_2) \in \mathcal{O}] + \delta$$

- We prove that backdoor resilience **from centralized learning** can be transformed to federated learning
- Determine σ_t dynamically based on S_t
- Clipping and filtering reduce necessary noise, i.e., minimize the effect on the model performance

Evaluation

ATTACK	Dataset	No Defense		FLAME	
		BA	MA	BA	MA
Constrain-and-Scale [Bagdasaryan et al. AISTATS 2020]	Reddit CIFAR-10 IoT-Traffic	100.0 81.9 100.0	22.6 89.8 100.0	**0.0** **0.0** **0.0**	22.3 91.9 **99.8**
Distributed Backdoor Attack [Xie et al. ICLR 2020]	CIFAR-10	93.8	57.4	**3.2**	76.2
Edge-Case [Wang et al. NeurIPS 2020]	CIFAR-10	42.8	84.3	**4.0**	79.3
Projected Gradient Decent [Wang et al. NeurIPS 2020]	CIFAR-10	56.1	68.8	**0.5**	65.1
Untargeted Poisoning [Fang et al. USENIXSec 2020]	CIFAR-10	-	**46.7**	-	**91.3**

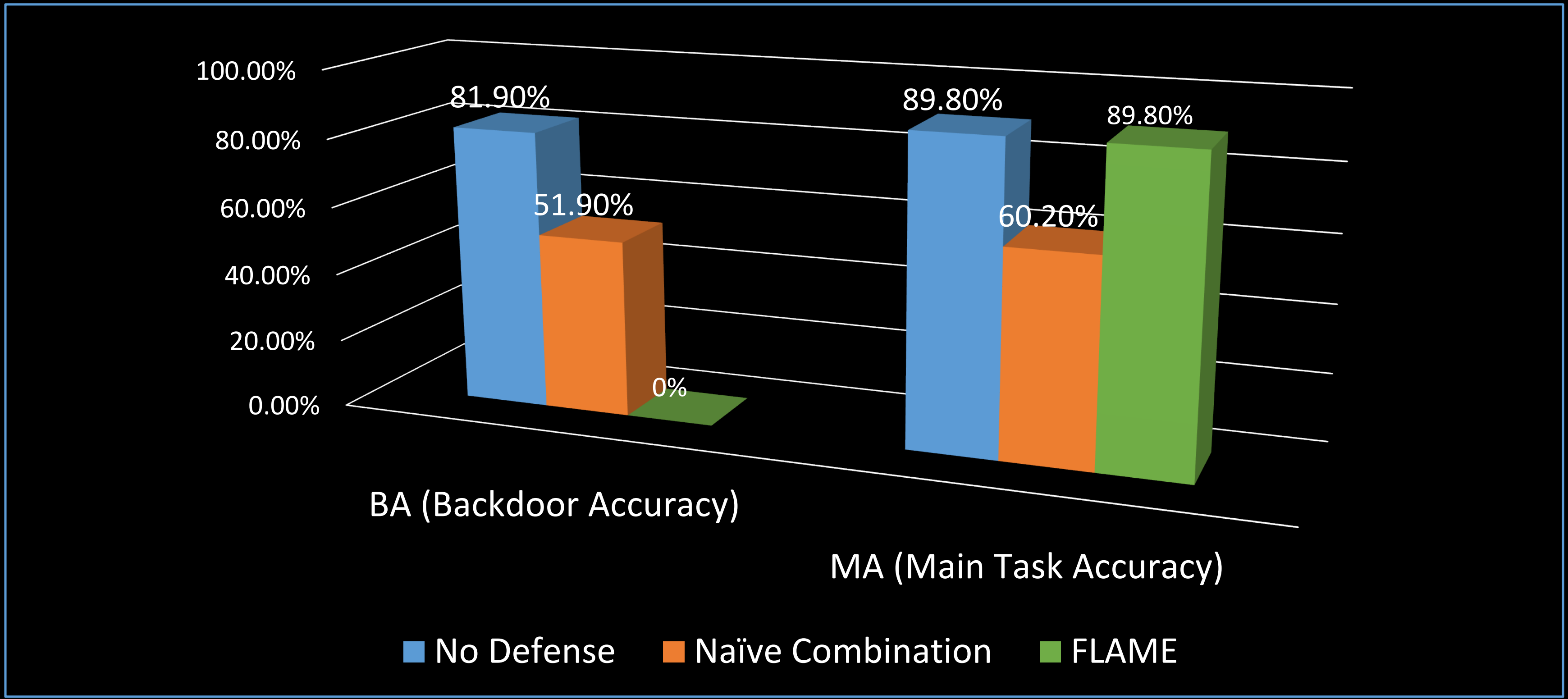
BA: Backdoor Accuracy, MA: Main Task Accuracy

FLAME vs. Existing Defenses

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

BA: Backdoor Accuracy, MA: Main Task Accuracy

FLAME vs. Naïve Combination



Comparison between FLAME and a combination of existing defenses against constrain-and-scale attack [Bagdasaryan et al. AISTATS 2020] on the CIFAR-10 dataset

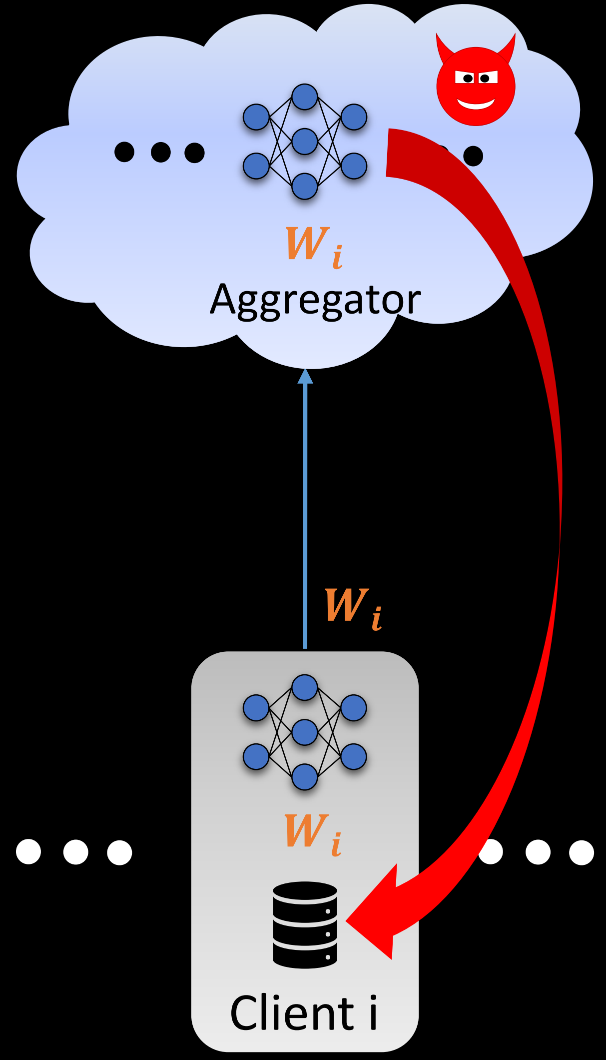
Private FLAME: Privacy Preserving Aggregation

Using secure Multi-party Computation

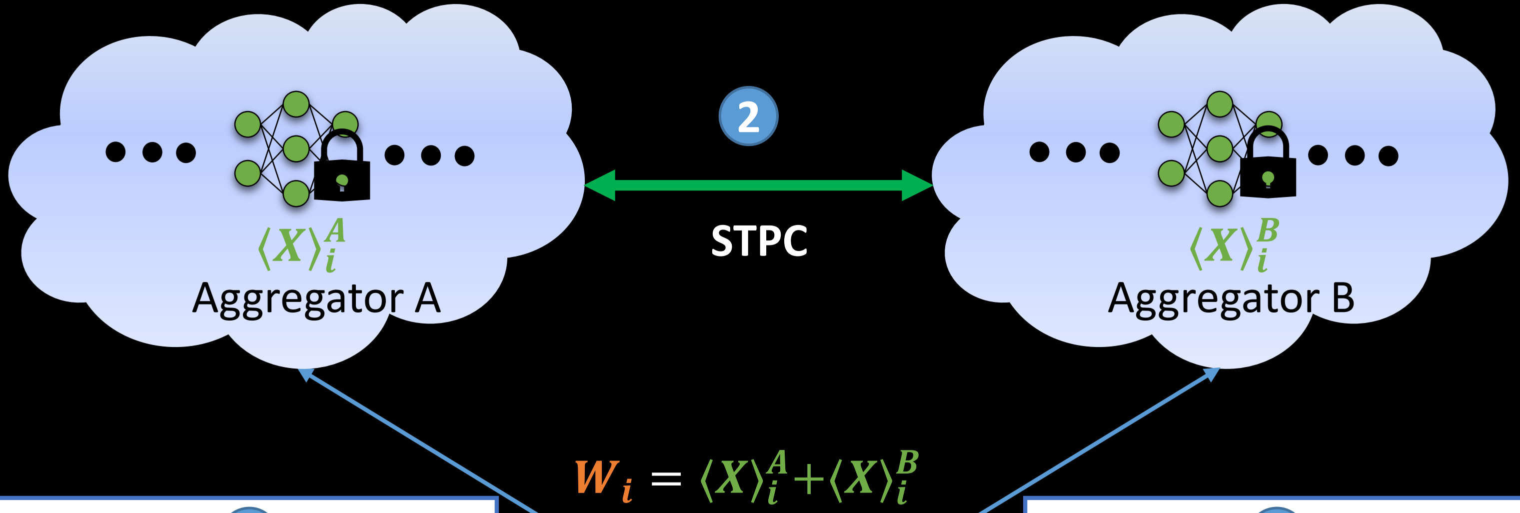
Private FLAME: Motivation

- 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



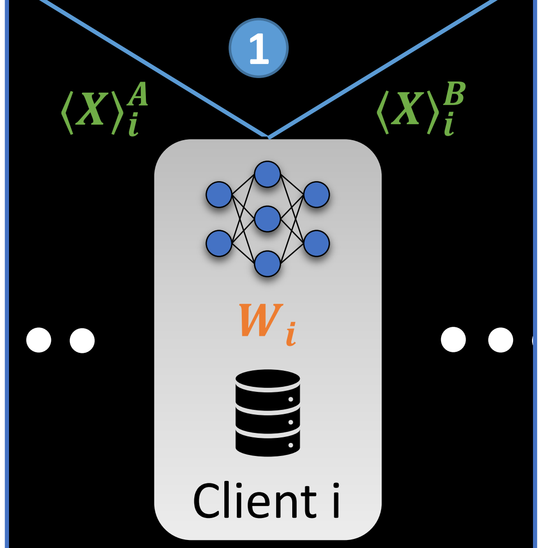
Utilizing Secure Two-Party Computation (**STPC**)



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 a 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 FALME 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

#client	Reddit		CIFAR-10		IoT-Traffic	
	S	P	S	P	S	P
10	2.35	519.92	0.32	134.93	0.07	108.16
50	62.55	5895.70	2.62	766.12	0.69	269.35
100	252.13	**22081.65**	8.56	**2568.23**	2.11	**876.96**

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

	Reddit		CIFAR-10		IoT-Traffic	
	S	P	S	P	S	P
BA	0.0	0.0	0.0	0.0	0.0	0.0
MA	22.3	22.2	91.9	91.7	99.8	99.7
TPR	22.2	20.4	23.8	40.8	59.5	51.0
TNR	100.0	100.0	86.2	100.0	100.0	100.0

Effectiveness of standard FALME (S) in comparison to
private FALME (P) using secure two-party computation

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

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