IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

What if attackers are indeed inside?

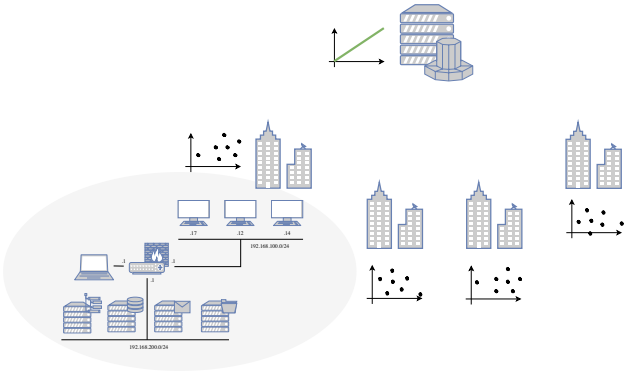
A Systematic Analysis of Label-flipping Attacks against Federated Learning for Intrusion Detection

Léo Lavaur[1], Yann Busnel[2], and Fabien Autrel[1]
[1] IMT Atlantique, [2] IMT Nord Europe

ARES (BASS) 2024, August 2, 2024

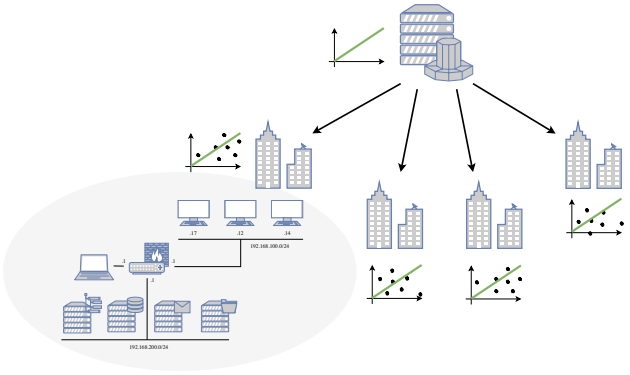
Federated Intrusion Detection System (FIDS)

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm.
- Can train a global model without sharing local data.
- Can be used in Collaborative Intrusion Detection System (CIDS):
  - Extend the training data
  - Effectively share knowledge (*e.g.*, on specific classes, instances) between participants



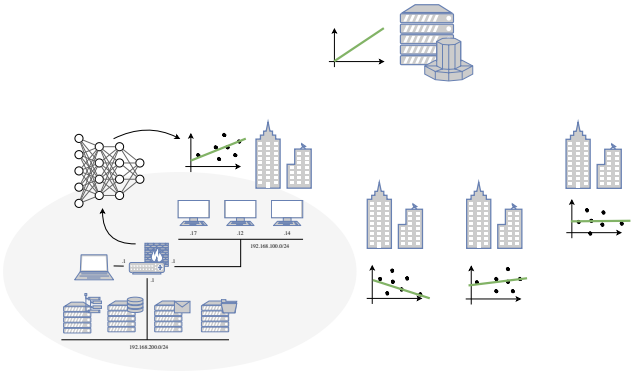
Federated Intrusion Detection System (FIDS)

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm.
- Can train a global model without sharing local data.
- Can be used in Collaborative Intrusion Detection System (CIDS):
  - Extend the training data
  - Effectively share knowledge (*e.g.*, on specific classes, instances) between participants



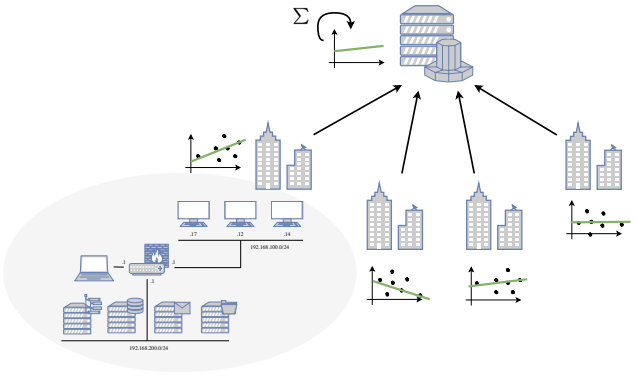
Federated Intrusion Detection System (FIDS)

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm.
- Can train a global model without sharing local data.
- Can be used in Collaborative Intrusion Detection System (CIDS):
  - Extend the training data
  - Effectively share knowledge (*e.g.*, on specific classes, instances) between participants



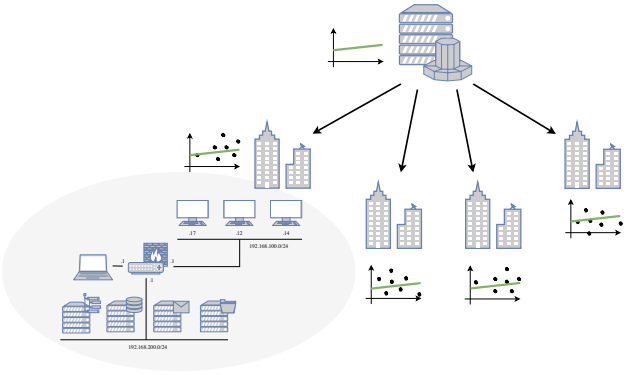
Federated Intrusion Detection System (FIDS)

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm.
- Can train a global model without sharing local data.
- Can be used in Collaborative Intrusion Detection System (CIDS):
  - Extend the training data
  - Effectively share knowledge (*e.g.*, on specific classes, instances) between participants



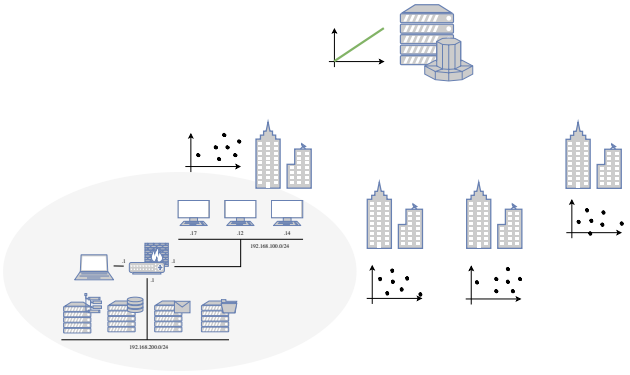
Federated Intrusion Detection System (FIDS)

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm.
- Can train a global model without sharing local data.
- Can be used in Collaborative Intrusion Detection System (CIDS):
  - Extend the training data
  - Effectively share knowledge (*e.g.*, on specific classes, instances) between participants



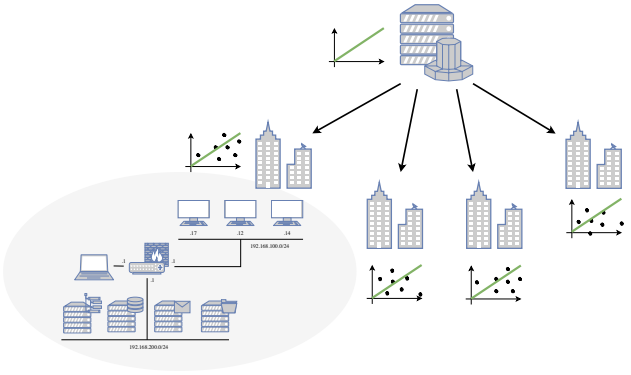
FL against malicious contributions

- FL is highly susceptible to poisoning.
  - Few studies on their impact in FIDS.



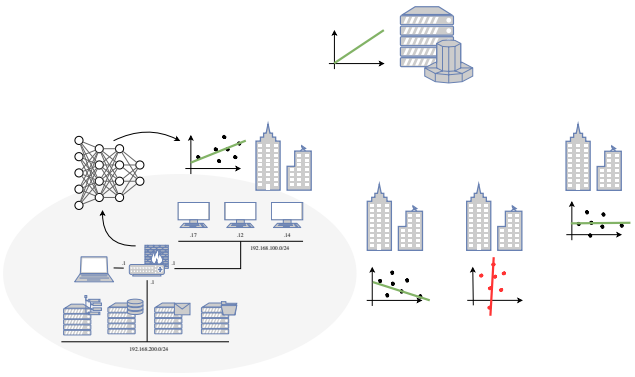
FL against malicious contributions

- FL is highly susceptible to poisoning.
  - Few studies on their impact in FIDS.



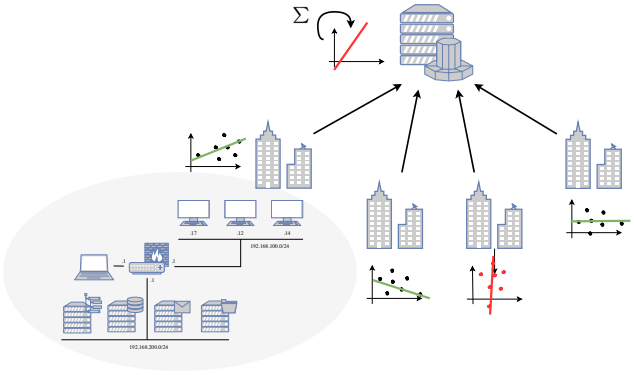
FL against malicious contributions

- FL is highly susceptible to poisoning.
  - Few studies on their impact in FIDS.



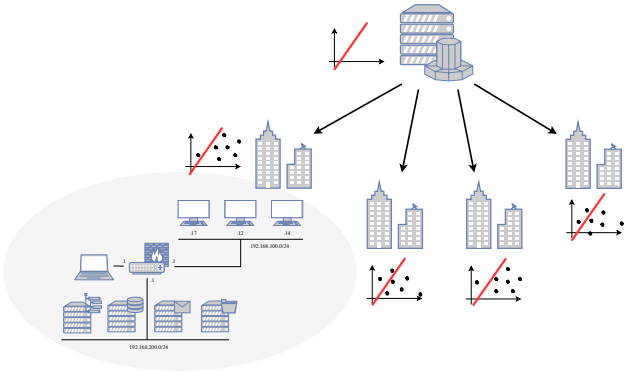
FL against malicious contributions

- FL is highly susceptible to poisoning.
  - Few studies on their impact in FIDS.



FL against malicious contributions

- FL is highly susceptible to poisoning.
  - Few studies on their impact in FIDS.



Types of poisoning attacks

- By component:
  - Data poisoning (*e.g.*, label-flipping, clean-label attacks, backdoors)
  - Model poisoning (*e.g.*, gradient boosting, noising)
- By target:
  - Untargeted: affect the model's global performance
  - Targeted: modify its behavior on specific classes or instances
- By frequency:
  - one-shot: attacks are performed once
  - iterative/continuous: at each round
  - adaptive: reacts to the model aggregation

Types of poisoning attacks

- By component:
  - Data poisoning (*e.g.*, **label-flipping**, clean-label attacks, backdoors)
  - Model poisoning (*e.g.*, gradient boosting, noising)
- By target:
  - **Untargeted**: affect the model's global performance
  - **Targeted**: modify its behavior on specific classes or instances
- By frequency:
  - one-shot: attacks are performed once
  - **iterative/continuous: at each round**
  - adaptive: reacts to the model aggregation

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

- Research questions (RQ):
  - RQ1. Is the behavior of poisoning attacks predictable?
  - RQ2. Are there beneficial or harmful combinations of hyperparameter under poisoning attacks?
  - RQ3. Can FL heal itself from poisoning attacks?
  - RQ4. Are IDS backdoors realistic using label-flipping attacks?
  - RQ5. Is there a critical threshold where label-flipping attacks begin to impact performance?

Experimental setup

- Used dataset: sampled NF-V2 version of CSE-CIC-IDS2018
  - ports and IP addresses are removed
- Same class distribution in the training and testing sets
  - 80% of the dataset is used for training
  - 20% of the dataset is used for testing

Experimental setup

- Used dataset: **sampled** NF-V2 version of CSE-CIC-IDS2018
  - ports and IP addresses are removed
- Same class distribution in the training and testing sets
  - 80% of the dataset is used for training
  - 20% of the dataset is used for testing

Experimental setup

- Used dataset: sampled NF-V2 version of CSE-CIC-IDS2018
  - ports and IP addresses are removed
- Same class distribution in the training and testing sets
  - 80% of the dataset is used for training
  - 20% of the dataset is used for testing
- Assessment of the representativity of the dataset sampling
  - Cross-projections of the malicious traffic from two datasets in two dimensions using PCA

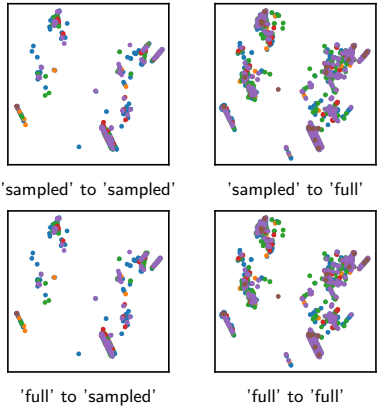


Figure: Cross-projection of the malicious traffic from two datasets in two dimensions using PCA.

Experimental setup

- A simple multilayer perceptron (MLP) model with two hidden layers is used
- Baseline (centralized training)
  - F1-score: 0.966
  - Accuracy: 0.992
- FL setup
  - Cross-silo setting: all clients are available at each round
  - The dataset is partitioned into 10 Independent and Identically Distributed (IID) shards of 80,000 data points
  - Models are aggregated using `FedAvg`
- Attack model
  - Malicious participants can alter their local datasets before training
  - Data-poisoning: label-flipping attacks (targeted and untargeted)
  - Data Poisoning Rate (DPR): $(\alpha) \rightarrow$ proportion of flipped samples
  - Model Poisoning Rate (MPR): $(\tau) \rightarrow$ proportion of attackers

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

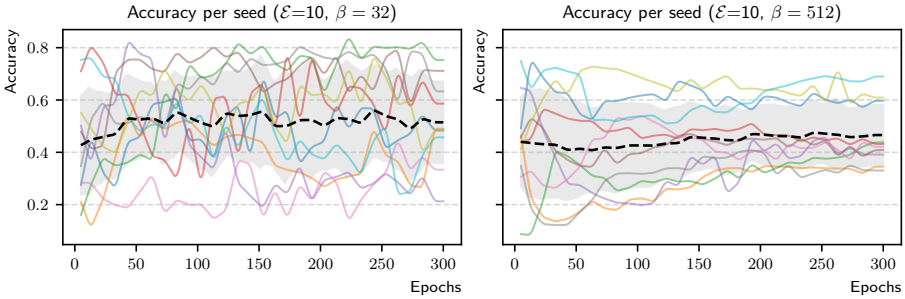
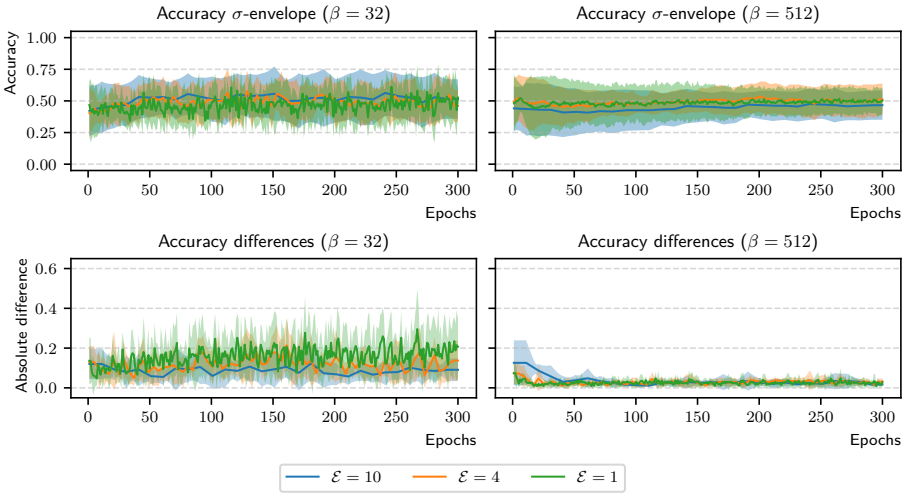


Figure: Accuracy of the poisoned model by seed.

RQ1: Is the behavior of poisoning attacks predictable?

Figure: Predictability depending on the hyperparameters.



RQ1: Is the behavior of poisoning attacks predictable?

Answer:

Answer: Nope.

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

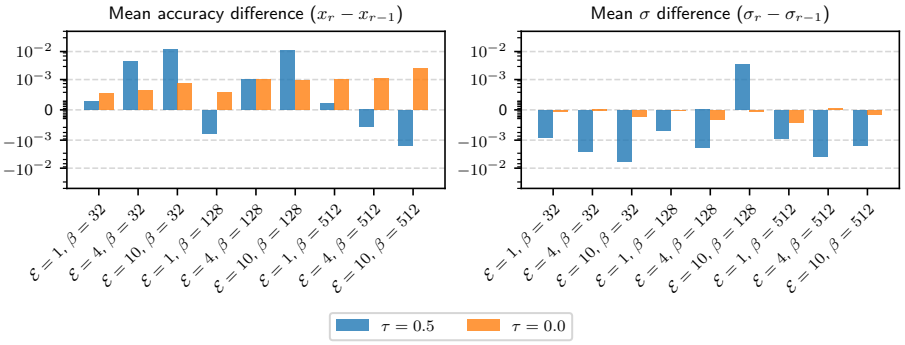


Figure: Effect of the hyperparameters on the accuracy of the poisoned model.

RQ2: Do hyperparameters influence the effect of poisoning attacks? 16

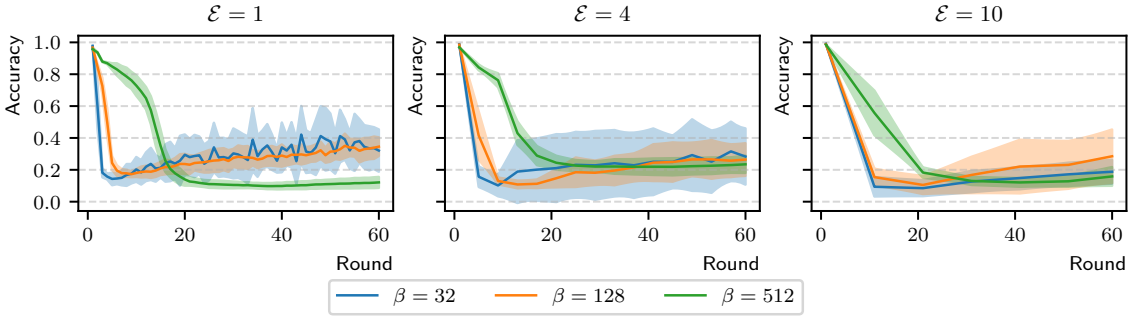


Figure: Effect of the hyperparameters on the accuracy of the poisoned model in the `late` scenario.

RQ2: Do hyperparameters influence the effect of poisoning attacks?

Answer:

- No impact on the average performance
- Significant impact on the variance of the results
- High batch size leads to more inertia
  - The impact is less instantaneous

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

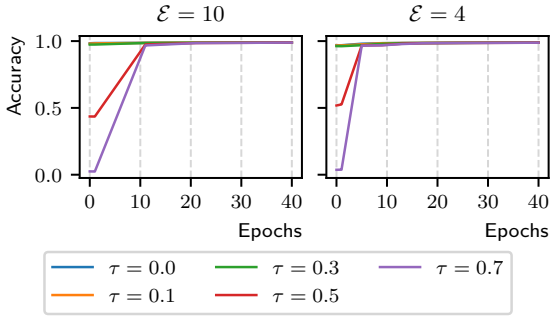


Figure: Recovery of the model after a poisoning attack.

RQ3: Can FL heal itself from poisoning attacks?

Answer:

- Yes.
- Attack stopping $\Longleftrightarrow$ randomize initial parameters
  - One round suffises to recover (in our settings).

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Absolute Attack Success Rate (AASR)

- Targeted attacks:
  - miss rate on a class: $\frac{\mathrm{FN}_{\mathrm{class}}}{\mathrm{TP}_{\mathrm{class}}+\mathrm{FN}_{\mathrm{class}}}$
- Untargeted attacks:
  - misclassification rate: $1 - \text{accuracy}$

Relative Attack Success Rate (RASR)

- $\frac{\max(\mathrm{AASR}_{\mathrm{benign}},\mathrm{AASR}_{\mathrm{attack}})-\mathrm{AASR}_{\mathrm{benign}}}{1-\mathrm{AASR}_{\mathrm{benign}}}$

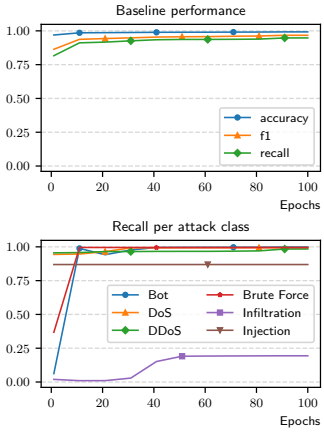


Figure: Baseline for benign runs.

RQ4: Are IDS backdoors realistic using label-flipping attacks?

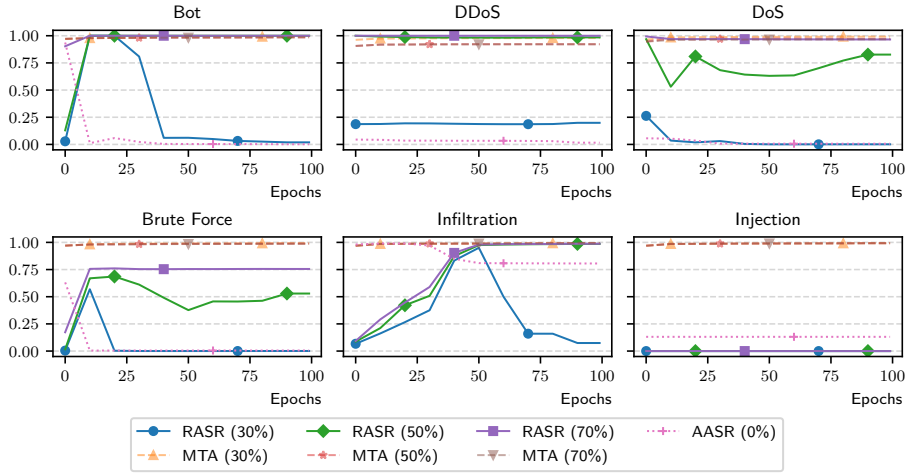


Figure: Backdoor success rate.

RQ4: Are IDS backdoors realistic using label-flipping attacks?

Answer:

- Yes, but...

RQ4: Are IDS backdoors realistic using label-flipping attacks?

Answer:

- Yes, but...
- The model's generalization capabilities can mitigate the impact
  - especially with class overlaps between characteristics
- The attack's effectiveness is highly dependent on the target
- We need a significant number of attackers.

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

RQ5: Is there a critical threshold for label-flipping to be effective?

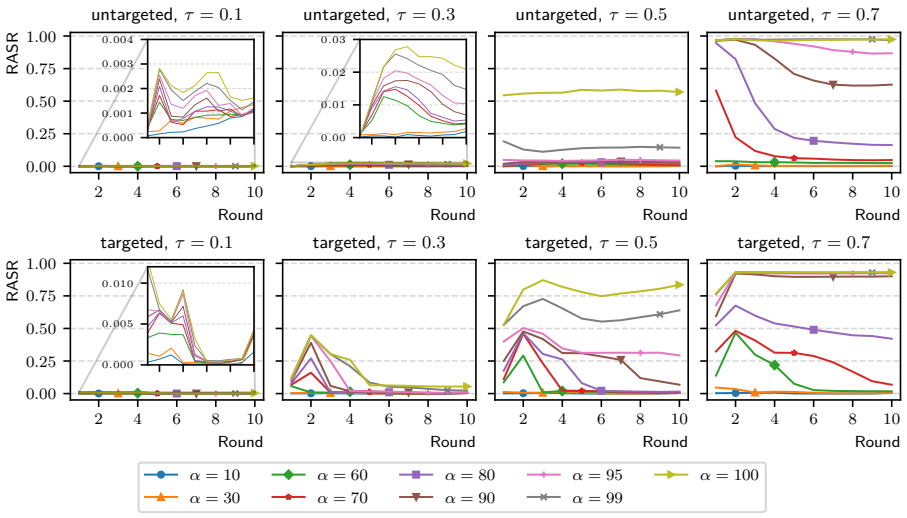


Figure: Impact of τ and α on the attack's effectiveness.

OUTLINE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Conclusion

We build a reproducible framework to study the impact of label-flipping attacks in FIDS using FL, here are our main findings (yet):

- Label-flipping attacks can have a significant impact on the performance of FL models, especially targeted ones
- The ASR is closely related to the number of flipped samples overall, which can be approximated in IID settings by $\alpha * \tau$
- Targeted label-flipping attacks strive on well-detected targets, but can be significantly mitigated by the model's generalization capabilities
- Mitigation strategies must be adapted to the use case specificities (e.g., constrained environments)

Future Work

- Extend the study to other datasets.
- Study the impact of the data distribution on the ability to detect attacks.
- Extend to other feature sets and poisoning attacks.
  - Our evaluation framework is generic enough (and open source!) to make extending the results easy.

Future Work

- Extend the study to other datasets. [DONE!]
- Study the impact of the data distribution on the ability to detect attacks. [DONE!]
- Extend to other feature sets and poisoning attacks.
  - Our evaluation framework is generic enough (and open source!) to make extending the results easy.

Want to know about the new results? I defend my thesis on ***October 7th, 2024*** *at* ***14:00*** *in* ***Rennes****. You are welcome to join!*