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

What if the attackers are inside?

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

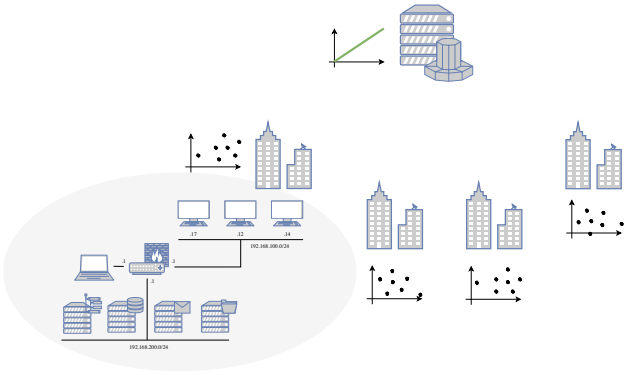
Léo Lavaur[1], Yann Busnel[2], and Fabien Autrel[1]
[1] IMT Atlantique, [2] IMT Nord Europe

4th International Workshop on Behavioral Analysis for System Security
co-located with the ARES conference, August 2, 2024

Figure: Typical NIDS workflow.

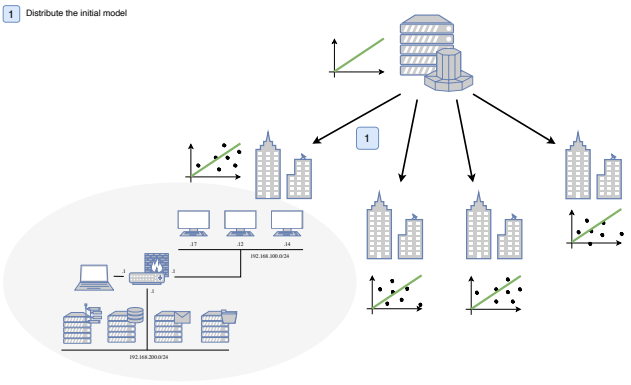
- Great performance with Deep Learning (DL) (on public datasets at least)
- **Limitations**: lack of labelled data, risk of local bias or skewed data distribution, inefficient against new attacks.

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm [1].
- Participants train a global model without sharing local data.



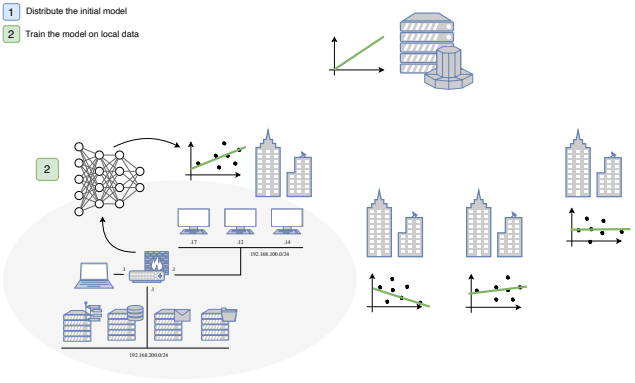
Scaling NIDS with FL

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm [1].
- Participants train a global model without sharing local data.



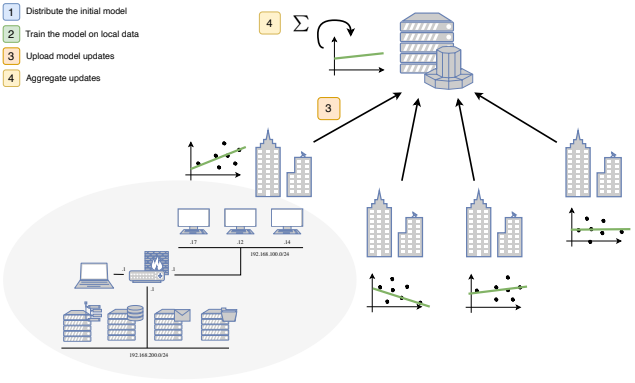
Scaling NIDS with FL

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm [1].
- Participants train a global model without sharing local data.



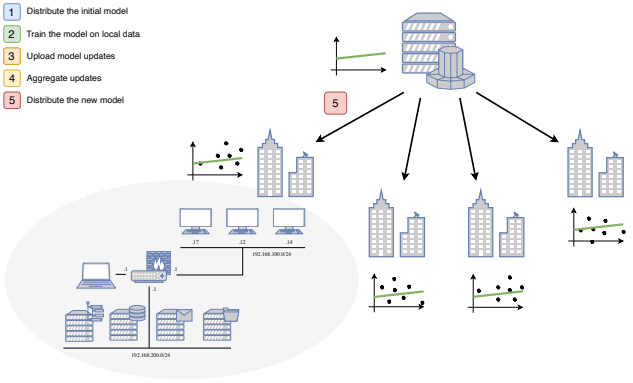
Scaling NIDS with FL

- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm [1].
- Participants train a global model without sharing local data.



Scaling NIDS with FL

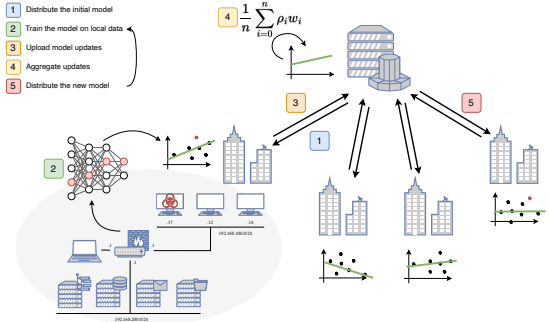
- Federated Learning (FL) is a distributed Machine Learning (ML) paradigm [1].
- Participants train a global model without sharing local data.



Federated Intrusion Detection System (FIDS)

FL can be used in Collaborative Intrusion Detection System (CIDS) [2]:

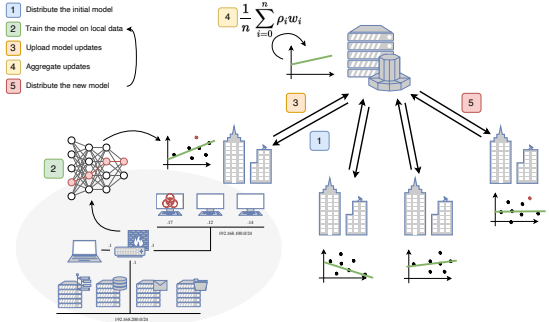
- Extend the training data with Horizontal Federated Learning (HFL)
  - → Reduce the risk of local bias



Federated Intrusion Detection System (FIDS)

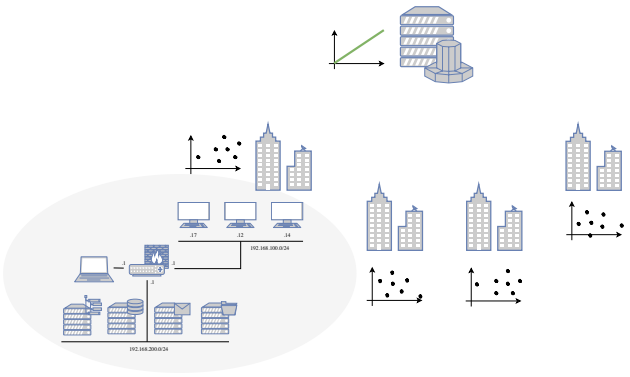
FL can be used in Collaborative Intrusion Detection System (CIDS) [2]:

- Extend the training data with Horizontal Federated Learning (HFL)
  - → Reduce the risk of local bias
- Effectively share knowledge (*e.g.*, on specific classes, instances) between participants
  - → Share the knowledge about a new attack [3];
  - → Improve the characterization of specific devices; ...

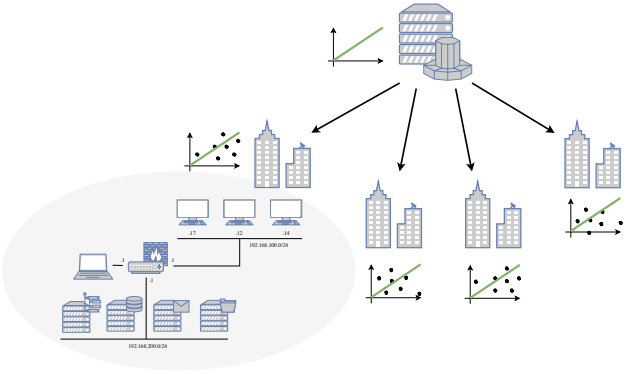


FL against malicious contributions

- FL is highly susceptible to poisoning [4].

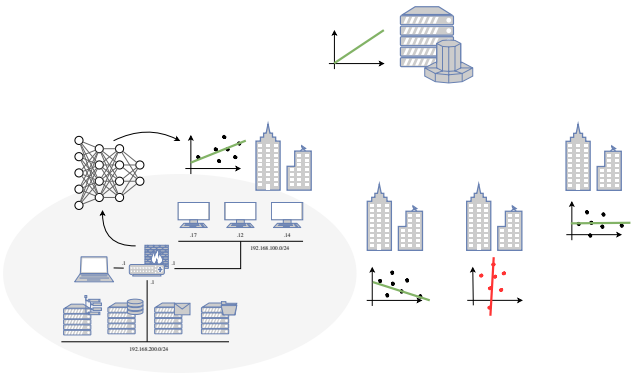


- FL is highly susceptible to poisoning [4].



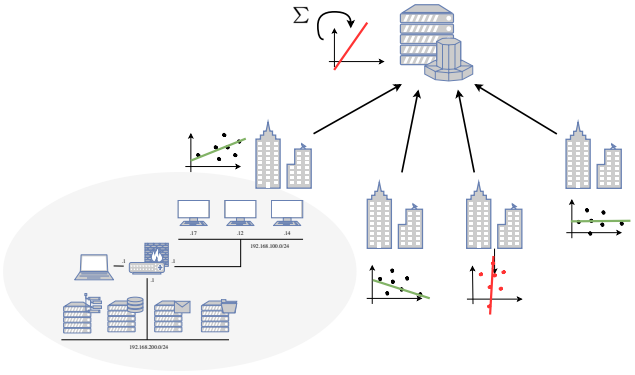
FL against malicious contributions

- FL is highly susceptible to poisoning [4].



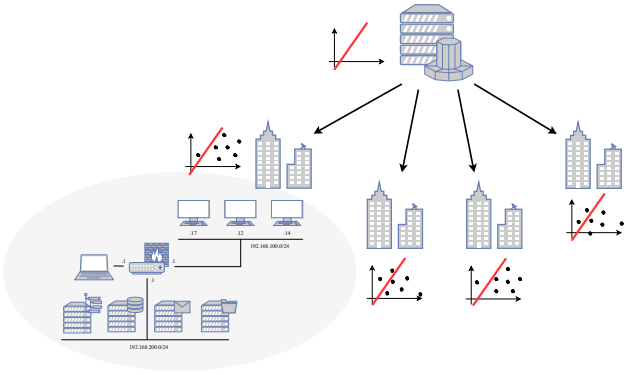
FL against malicious contributions

- FL is highly susceptible to poisoning [4].



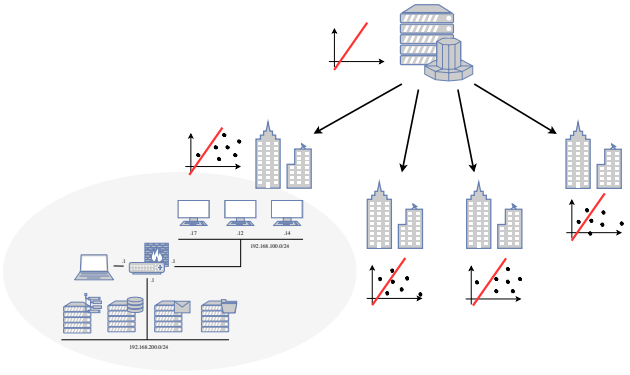
FL against malicious contributions

- FL is highly susceptible to poisoning [4].



FL against malicious contributions

- FL is highly susceptible to poisoning [4].
- The impact is difficult to estimate.
  - Few studies in the FIDS context, often partial.
  - Nothing on aggregated classes.



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

Our work

Continuous label-flipping attacks in collaborative Intrusion Detection System (IDS) context.

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?

Evaluation framework

Experiment orchestration using `Eiffel` [3].

- `Flower` simulation framework [5] for FL.
- `Hydra` [6] for experiment generation and configuration.
- Custom-made poisoning engine with different attack strategies.
- Nix [7] and Poetry to fix system and Python dependencies, enabling reproducibility.

Experiment orchestration using `Eiffel` [3].

- `Flower` simulation framework [5] for FL.
- `Hydra` [6] for experiment generation and configuration.
- Custom-made poisoning engine with different attack strategies.
- Nix [7] and Poetry to fix system and Python dependencies, enabling reproducibility.

Table: Experimental parameters.

Parameter	Values	Description
batch_size	32, 128, **512**	Batch size (β)
epochs	**100_10x10**, 100_4x25, 100_1x100, 300_10x30, 300_4x75, 300_1x300	Local epochs per round ($\mathcal{E}$)
distribution	10-0, 9-1, 7-3, **5-5**, 3-7	Proportion of attackers (τ)
scenario	continuous-{10,30,60,70,80,90,95,99}, **continuous-100**, late-3, redemption-3	Poisoning rate per round (α)
target	**untargeted**, bot, dos, ddos, bruteforce, infiltration, injection	Attack type and target
seed	1313, 1977, 327, 5555, 501, 421, 3263827, 2187, 1138, 6567	Seed for PRNG

494 experiments $\times$ 10 seeds $\rightarrow$ 685 hours of computation.

Absolute Attack Success Rate (AASR)

- Targeted attacks $\rightarrow$ miss rate on the targeted class:

$$\frac{\mathrm{FN}_{\mathrm{class}}}{\mathrm{TP}_{\mathrm{class}} + \mathrm{FN}_{\mathrm{class}}}$$

- Untargeted attacks $\rightarrow$ misclassification rate:

$$1 - \text{accuracy}$$

Absolute Attack Success Rate (AASR)

- Targeted attacks → miss rate on the targeted class:

$$\frac{\mathrm{FN}_{\mathrm{class}}}{\mathrm{TP}_{\mathrm{class}} + \mathrm{FN}_{\mathrm{class}}}$$

- Untargeted attacks → misclassification rate:

$$1 - \mathrm{accuracy}$$

Relative Attack Success Rate (RASR)

- ASR relative to the benign scenario:

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

OUTLINE

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

Experimental setup

- Dataset:
  - **sampled** NF-V2 version of CSE-CIC-IDS2018 [8], [9]
  - ports and IP addresses are removed
  - 80% for training, 20% for testing (evenly distributed)

Experimental setup

- Dataset:
  - **sampled** NF-V2 version of CSE-CIC-IDS2018 [8], [9]
  - ports and IP addresses are removed
  - 80% for training, 20% for testing (evenly distributed)
- Model:
  - Multilayer Perceptron (MLP) model with two hidden layers [10]
  - Baseline (centralized training): F1-score of 0.966, accuracy of 0.992

Experimental setup

- Dataset:
  - **sampled** NF-V2 version of CSE-CIC-IDS2018 [8], [9]
  - ports and IP addresses are removed
  - 80% for training, 20% for testing (evenly distributed)
- Model:
  - Multilayer Perceptron (MLP) model with two hidden layers [10]
  - Baseline (centralized training): F1-score of 0.966, accuracy of 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`

OUTLINE

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

Impact predictability

- Very high variance in the results
- The impact of the attack is highly dependent on the seed
  - → Initial parameters, data shuffling, ...
- Results tend to stabilize after a few rounds on different values

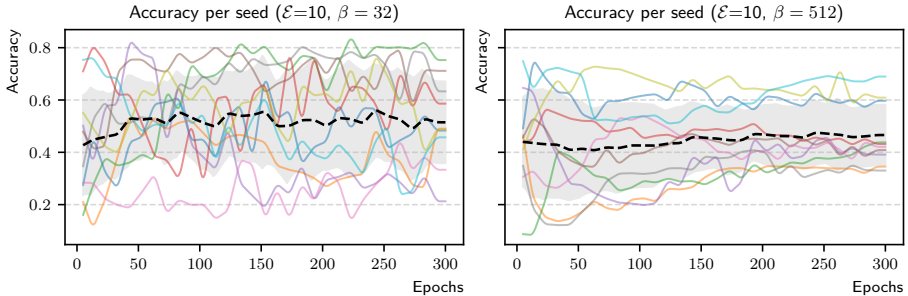


Figure: Accuracy of the poisoned model by seed (50% attackers).

Hyperparameter influence

- No impact on the average performance
- Significant impact on the variance of the results, but not really on convergence

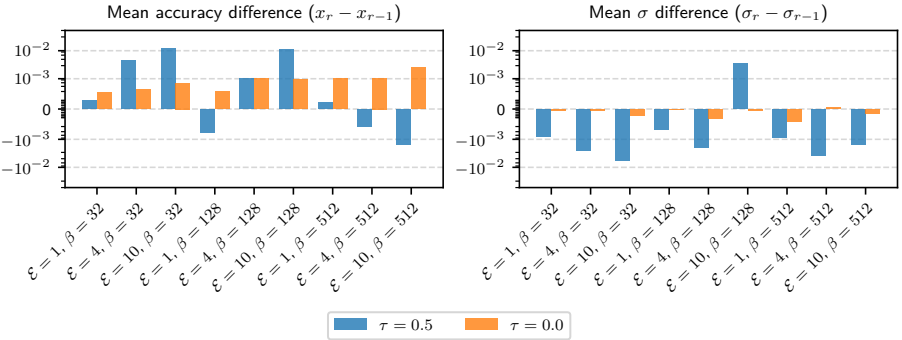


Figure: Effect of the hyperparameters on the accuracy of the poisoned model.

Hyperparameter influence

- `late-3` scenario: attackers start poisoning after 3 rounds
- High batch size leads to more inertia
  - The impact is less instantaneous → more impactful in constrained environments

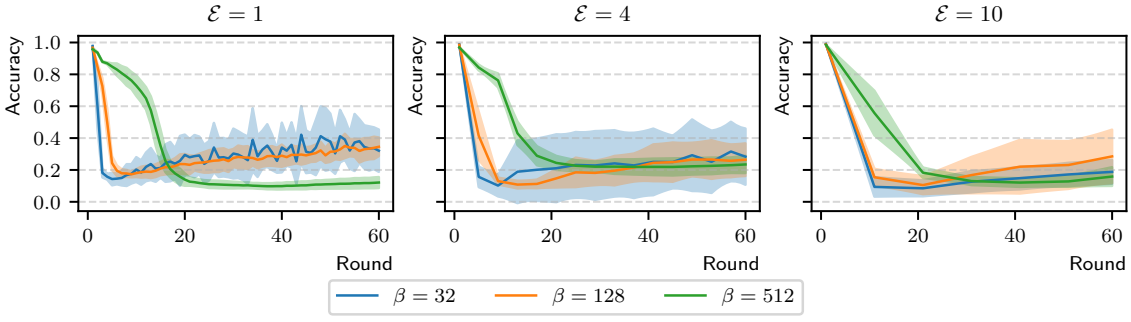


Figure: Effect of the hyperparameters on the accuracy of the poisoned model in the `late` scenario (50% attackers).

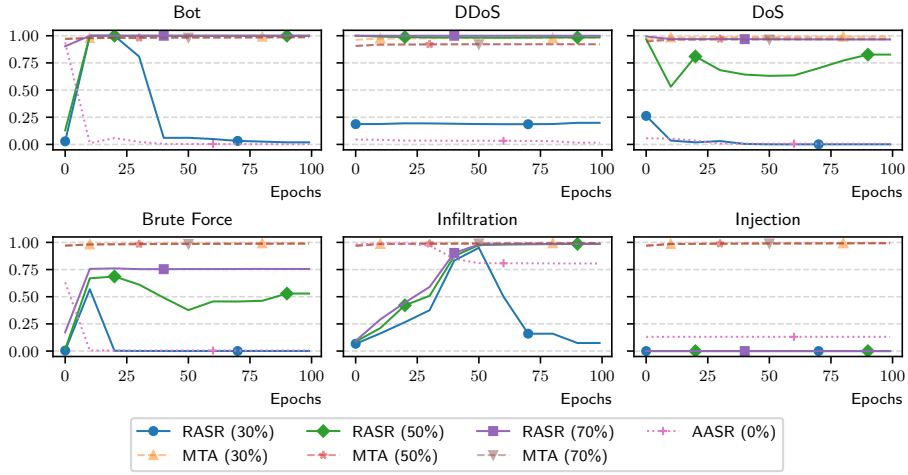


Figure: Backdoor success rate.

Targeted attacks

Can we build an IDS backdoor using label-flipping?

- Yes, but...

Targeted attacks

Can we build an IDS backdoor using label-flipping?

- Yes, but...
- The model's generalization capabilities can mitigate the impact
  - especially with characteristic overlaps between classes
- 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

Conclusion

- A *reproducible evaluation framework* to study the impact of label-flipping attacks in FIDS using FL.
  - reproducible, extendable, and available in open-access
  - first step towards a more comprehensive evaluation and comparison of poisoning attacks and their mitigation strategies.
- A *deeper understanding of the behavior of label-flipping attacks* in FL-based CIDSs.
  - The behavior of poisoning attacks is unpredictable.
  - It is dependent on the hyperparameters, but not on the average performance.
  - Targeted attacks can be effective but are highly dependent on the model's generalization capabilities.

$\rightarrow$ *more detailed results in the paper :)*

Future Work

- Extend the study to other datasets.
- Extend to other feature sets and poisoning attacks.
- Build more appropriate metrics for the evaluation of poisoning attacks in different data distributions.
- Study the impact of the data distribution on the ability to detect attacks using similarity metrics.

Future Work

- Extend the study to other datasets. [DONE]
- Extend to other feature sets and poisoning attacks.
- Build more appropriate metrics for the evaluation of poisoning attacks in different data distributions.
- Study the impact of the data distribution on the ability to detect attacks using similarity metrics. [DONE]

References I

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, PMLR, Apr. 20–22, 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html.

[2] L. Lavaur, M.-O. Pahl, Y. Busnel, and F. Autrel, "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: A Survey," *IEEE Transactions on Network and Service Management*, Special Issue on Network Security Management, Jun. 2022.

[3] L. Lavaur, Y. Busnel, and F. Autrel, "Demo: Highlighting the limits of federated learning in intrusion detection," in *Proceedings of the 44th International Conference on Distributed Computing Systems (ICDCS)*, Jersey City, NJ, USA, Jul. 2024.

[4] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data Poisoning Attacks Against Federated Learning Systems," in *Computer Security – ESORICS 2020*, L. Chen, N. Li, K. Liang, and S. Schneider, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 480–501, ISBN: 978-3-030-58951-6. DOI: 10.1007/978-3-030-58951-6_24.

[5] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane, "Flower: A friendly federated learning research framework," 2020. arXiv: 2007.14390.

[6] O. Yadan, *Hydra - A framework for elegantly configuring complex applications*, Github, 2019. [Online]. Available: https://github.com/facebookresearch/hydra.

[7] E. Dolstra, "The purely functional software deployment model," s.n., S.l., 2006.

[8] M. Sarhan, S. Layeghy, and M. Portmann, "Towards a Standard Feature Set for Network Intrusion Detection System Datasets," *Mobile Networks and Applications*, vol. 27, no. 1, pp. 357–370, Feb. 1, 2022, ISSN: 1572-8153. DOI: 10.1007/s11036-021-01843-0. [Online]. Available: https://doi.org/10.1007/s11036-021-01843-0 (visited on 04/23/2024).

References III

[9] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116, ISBN: 978-989-758-282-0. DOI: 10.5220/0006639801080116. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006639801080116 (visited on 10/14/2021).

[10] S. I. Popoola, G. Gui, B. Adebisi, M. Hammoudeh, and H. Gacanin, "Federated Deep Learning for Collaborative Intrusion Detection in Heterogeneous Networks," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Sep. 2021, pp. 1–6. DOI: 10.1109/VTC2021-Fall52928.2021.9625505.

Thank you!

- Consider the use case when designing or choosing mitigations.
- Expect the unexpected: the behavior of poisoning attacks is unpredictable.
- Try it yourself! Our work is *reproducible* and everything is in *open-access*!

Paper

Results

Evaluation framework

Any questions?

Threat model implementation

- A malicious participant can alter its local dataset before training, and start/stop the attack at any round r.
  - Either done by the client itself or by an external attacker.
- The attacker can manipulate labels of its dataset.
  - Untargeted: flip the labels of a proportion of samples.
  - Targeted: associate benign labels to a proportion of samples from a specific attack class.
- An attacker can poison any proportion of its local data
  - Data Poisoning Rate (DPR): $(\alpha) \rightarrow$ proportion of flipped samples.
- Multiple attackers can collude to perform the attack.
  - Model Poisoning Rate (MPR): $(\tau) \rightarrow$ proportion of attackers.

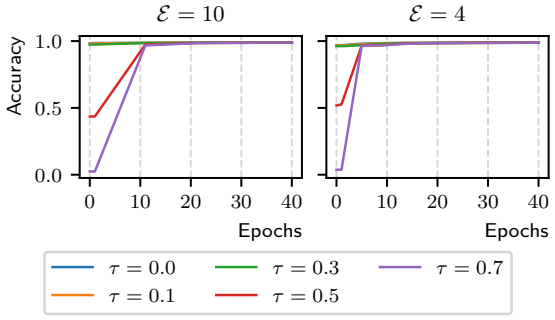


Figure: Recovery of the model after a poisoning attack.

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

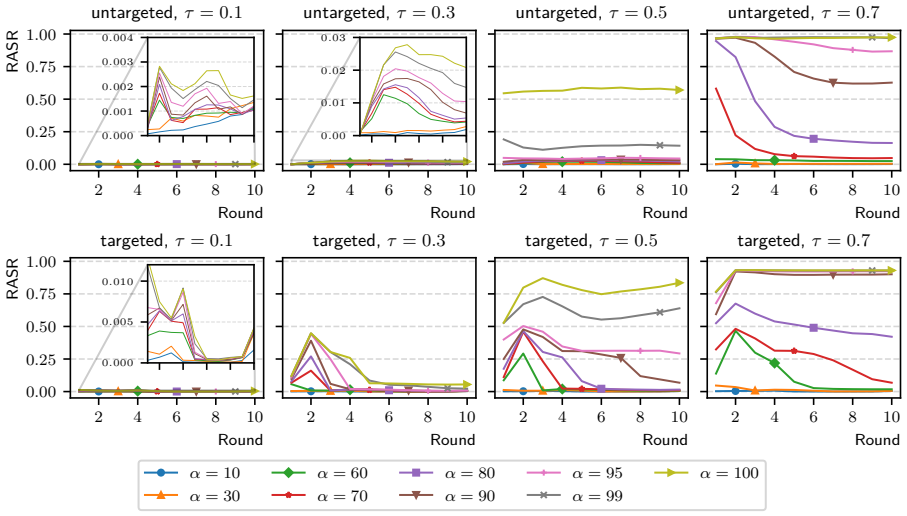


Figure: Impact of τ and α on the attack's effectiveness.