Improving Intrusion Detection in Distributed Systems with Federated Learning

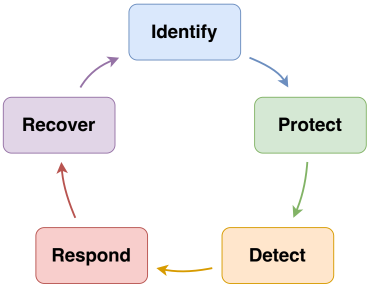
Defense replay at the SuperviZ Workshop

Léo Lavaur

Interdisciplinary Centre for Cybersecurity and Trust (SnT)
University of Luxembourg

Rennes, December 17th, 2024

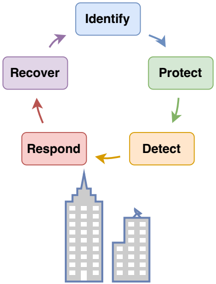
Reviewers:	Anne-Marie Kermarrec · EPFL
	Eric Totel · Télécom SudParis
Examiners:	Sonia Ben Mokhtar · CNRS
	Pierre-François Gimenez · Inria
	Vincent Nicomette · INSA Toulouse
Supervisors:	Fabien Autrel · IMT Atlantique
	Marc-Oliver Pahl · IMT Atlantique
Director:	Yann Busnel · IMT

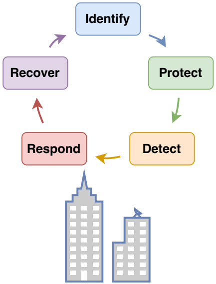


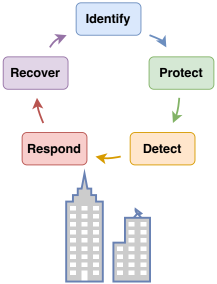
The security life-cycle [1].

[1] National Institute of Standards and Technology. *The NIST Cybersecurity Framework (CSF) 2.0*. 2024

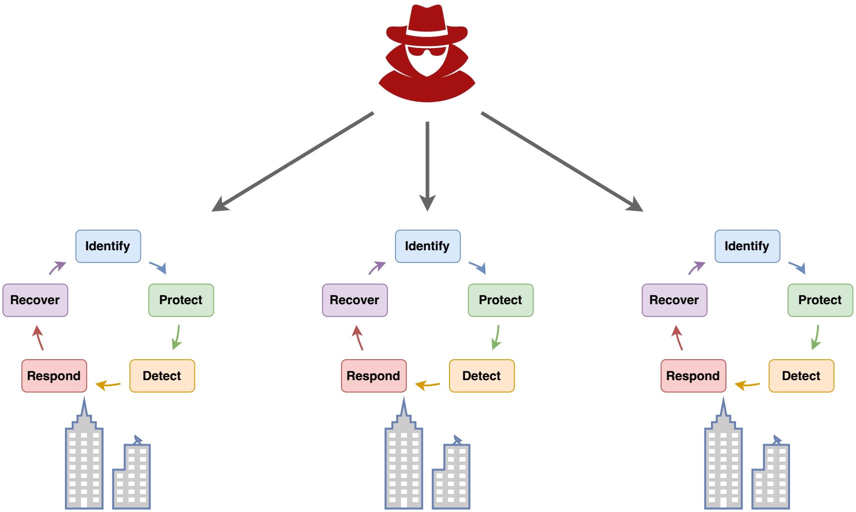
The Need for Collaboration



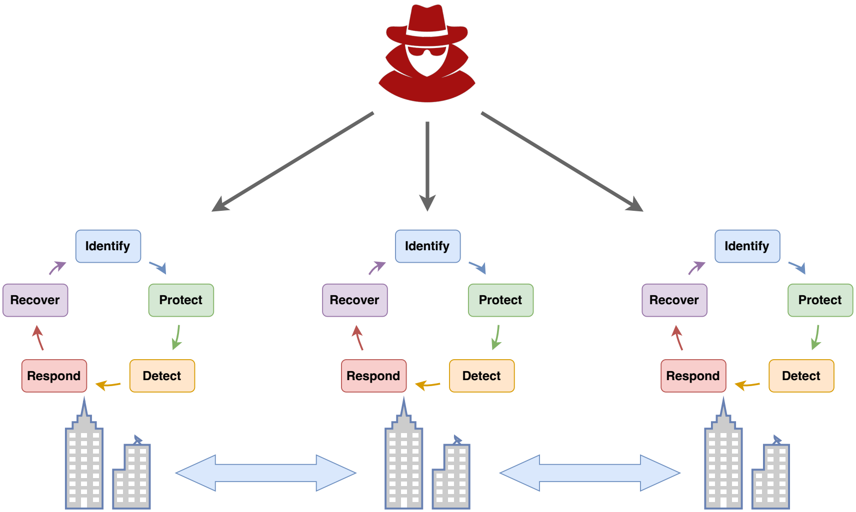




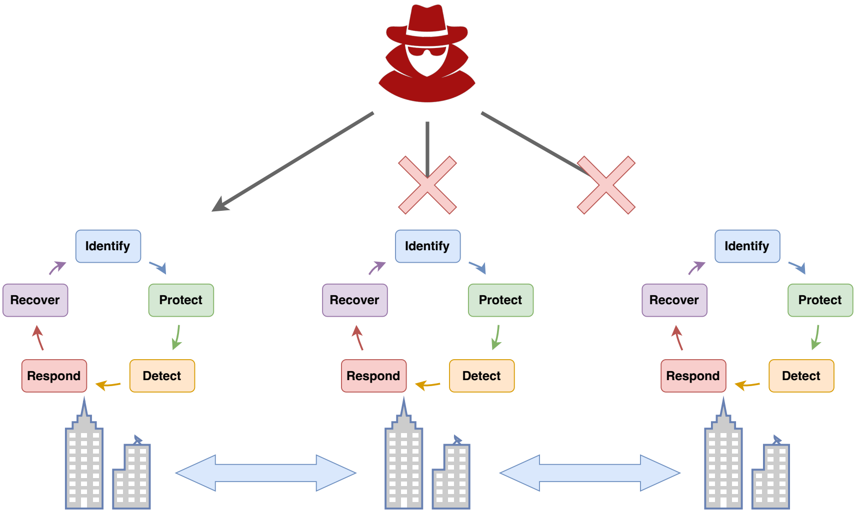
The Need for Collaboration



The Need for Collaboration



The Need for Collaboration



- Collaboration pushed by:
  - common interest (*e.g.*, inter-SOCs[1]);

[1]Security Operational Center.

Collaboration in Cybersecurity

- Collaboration pushed by:
  - common interest (*e.g.*, inter-SOCs[1]);
  - national agencies (*e.g.*, NIST, ENISA, ANSSI);

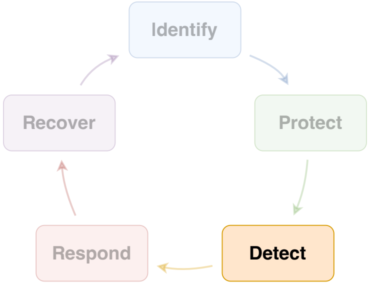
[1] Security Operational Center.

Collaboration in Cybersecurity

- Collaboration pushed by:
  - common interest (*e.g.*, inter-SOCs[1]);
  - national agencies (*e.g.*, NIST, ENISA, ANSSI);
  - regulation (*e.g.*, private-public information sharing in NIS2).

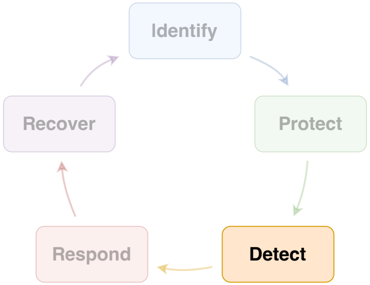
[1] Security Operational Center.

- Collaboration pushed by:
  - common interest (*e.g.*, inter-SOCs[1]);
  - national agencies (*e.g.*, NIST, ENISA, ANSSI);
  - regulation (*e.g.*, private-public information sharing in NIS2).



[1] Security Operational Center.

- Collaboration pushed by:
  - common interest (*e.g.*, inter-SOCs[1]);
  - national agencies (*e.g.*, NIST, ENISA, ANSSI);
  - regulation (*e.g.*, private-public information sharing in NIS2).



Intrusion Detection System (IDS)

IDSs monitor the behavior of a system to detect malicious activities.

[1]Security Operational Center.

- Various types of algorithms: supervised, unsupervised, semi-supervised, reinforcement learning, *etc*.
- Great performance with Deep Learning (DL)...

Machine Learning for Intrusion Detection

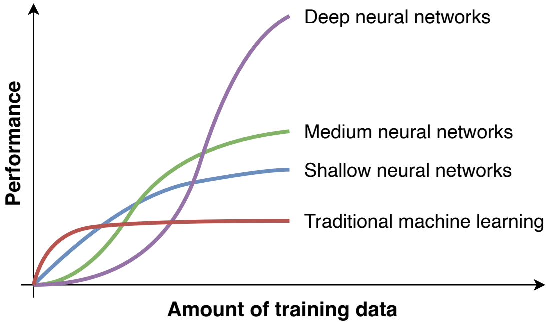
- Various types of algorithms: **supervised**, unsupervised, semi-supervised, reinforcement learning, *etc.*
- Great performance with Deep Learning (DL)...

- Various types of algorithms: supervised, unsupervised, semi-supervised, reinforcement learning, *etc*.
- Great performance with Deep Learning (DL)... *on public datasets at least*.

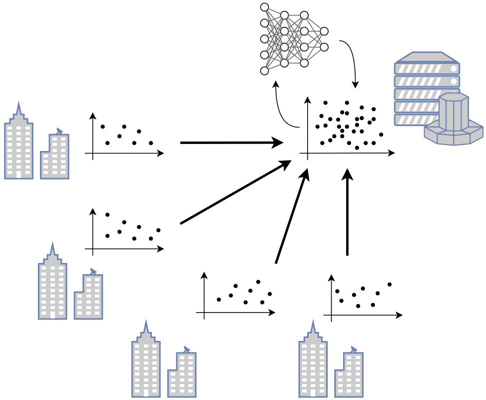
- Various types of algorithms: supervised, unsupervised, semi-supervised, reinforcement learning, *etc*.
- Great performance with Deep Learning (DL)... *on public datasets at least.*

Challenges of local training:

- not enough labelled data;
- risk of local bias or skewed data distribution.

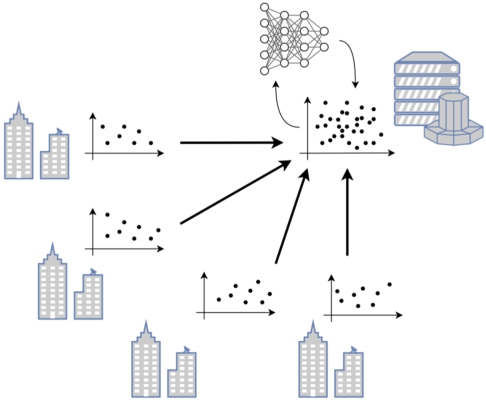


Data Sharing to the Rescue?



Let's pool our data!

Data Sharing to the Rescue?



Let's pool our data! Although...

- Privacy concerns.
- Lack of trust in the data holder.
- Lack of trust in the learning process.
- ...

Federated Learning (FL)

- Novel-*ish* distributed ML paradigm (Google) [2].

[2] McMahan et al. “Communication-Efficient Learning of Deep Networks from Decentralized Data”. Proceedings of Machine Learning Research. 2017

Federated Learning (FL)

- Novel-*ish* distributed ML paradigm (Google) [2].
- Distributed clients can train a common model without sharing their training data.

[2] McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". Proceedings of Machine Learning Research. 2017

Federated Learning (FL)

- Novel-*ish* distributed ML paradigm (Google) [2].
- Distributed clients can train a common model without sharing their training data.
- Privacy-preserving: high level of abstraction for the shared models preventing data leakage.

[2] McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". Proceedings of Machine Learning Research. 2017

FL Fundamentals

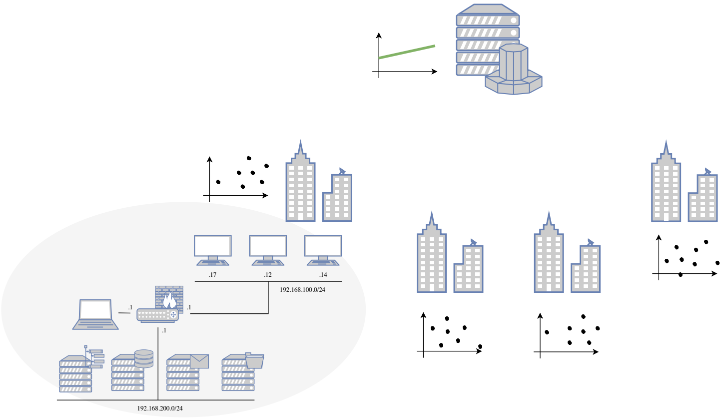


Figure: Typical FL workflow, applied to NIDSs.

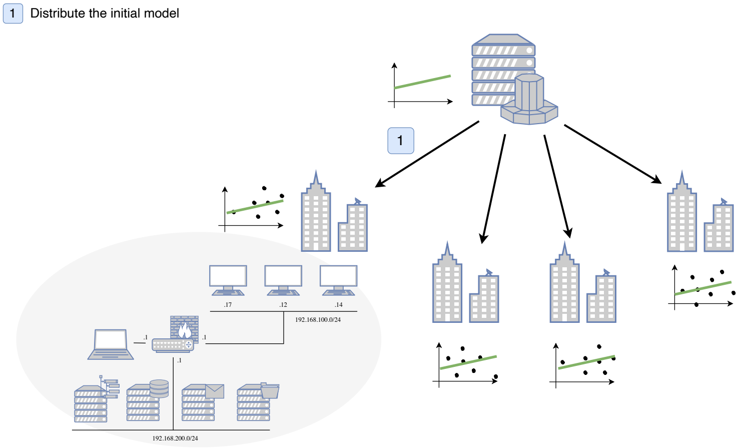
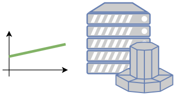


Figure: Typical FL workflow, applied to NIDSs.

1 Distribute the initial model

2 Train the model on local data



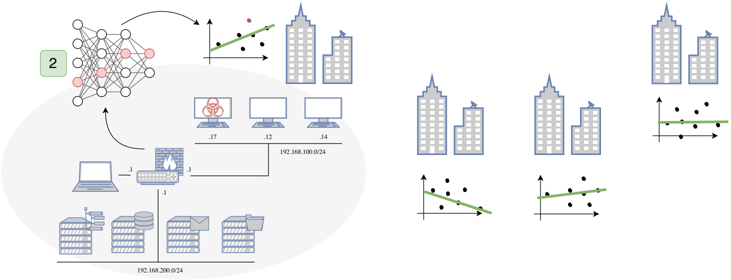


Figure: Typical FL workflow, applied to NIDSs.

FL FUNDAMENTALS

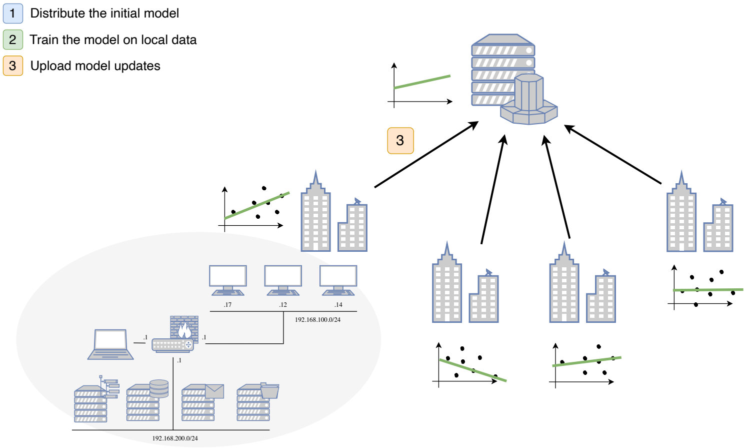


Figure: Typical FL workflow, applied to NIDSs.

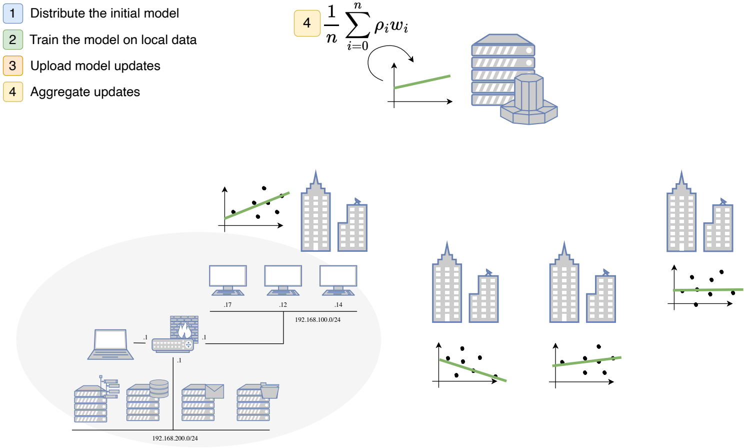


Figure: Typical FL workflow, applied to NIDSs.

1 Distribute the initial model
2 Train the model on local data
3 Upload model updates
4 Aggregate updates
5 Distribute the new model

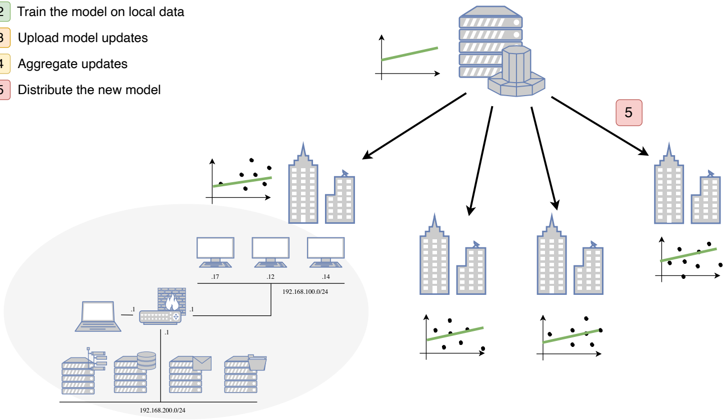


Figure: Typical FL workflow, applied to NIDSs.

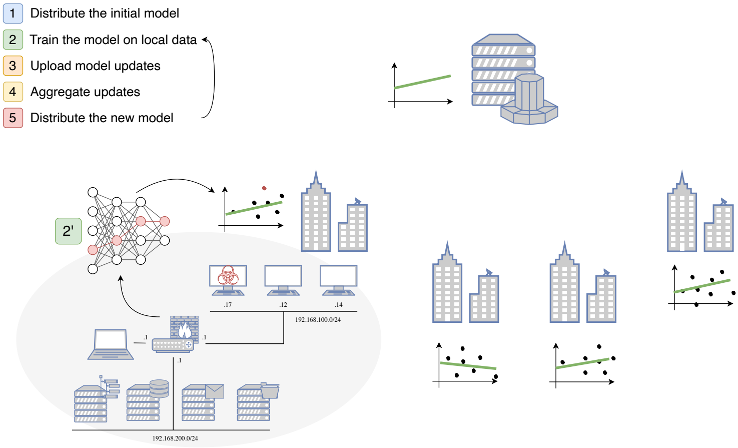


Figure: Typical FL workflow, applied to NIDSs.

Collaborative Intrusion Detection between Distributed Organizations

- Each organization has its own NIDS[2] and monitors an information system.

[2] Network-based IDS

Case Study

Collaborative Intrusion Detection between Distributed Organizations

- Each organization has its own NIDS[2] and monitors an information system.
- Objective: improve their local detection performance.

[2] Network-based IDS

Case Study

Collaborative Intrusion Detection between Distributed Organizations

- Each organization has its own NIDS[2] and monitors an information system.
- Objective: improve their local detection performance.
- Means: **expert knowledge** (*i.e.*, datasets) and **computing resources** (*i.e.*, model training).

[2] Network-based IDS

Collaborative Intrusion Detection between Distributed Organizations

- Each organization has its own NIDS[2] and monitors an information system.
- Objective: improve their local detection performance.
- Means: **expert knowledge** (*i.e.*, datasets) and **computing resources** (*i.e.*, model training).

Figure: Typical workflow for ML-based NIDSs.

[2] Network-based IDS

Case Study – Generalization

A cross-silo use case [3]:

- few clients (*i.e.*, 10–100);
- substantial amount of data, high heterogeneity;
- high availability, significant computing resources.

[3] Kairouz et al. "Advances and Open Problems in Federated Learning". 2021

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.

[4] Lavaur et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey". *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.

[4] Lavaur et al. “The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey”. *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.
- *Security and reliability*: security, privacy, trust, and reputation.

[4] Lavaur et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey". *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.
- *Security and reliability*: security, privacy, trust, and reputation.
- *Experimentation*: evaluation.

[4] Lavaur et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey". *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.
- *Security and reliability*: security, privacy, trust, and reputation.
- *Experimentation*: evaluation.

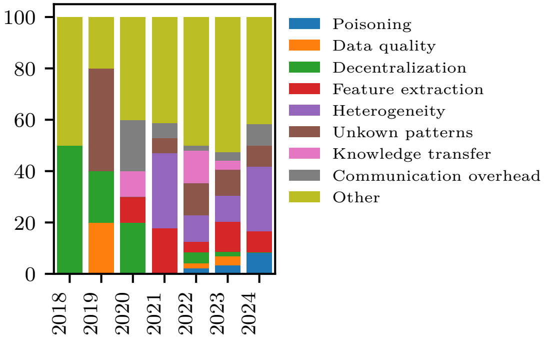


Figure: Challenges addressed by the literature (until 2024-04).

[4] Lavaur et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey". *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.
- *Security and reliability*: security, privacy, trust, and reputation.
- *Experimentation*: evaluation.

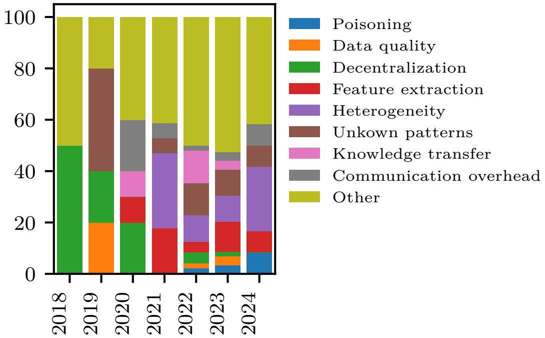


Figure: Challenges addressed by the literature (until 2024-04).

[4] Lavaur et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey". *IEEE Transactions on Network and Service Management*. 2022

Challenges from the Literature [4]

- *Functionality*: performance, heterogeneity, transferability, self-defense, and self-healing.
- *Deployment*: adaptability and scalability.
- *Security and reliability*: security, privacy, trust, and reputation.
- *Experimentation*: evaluation.

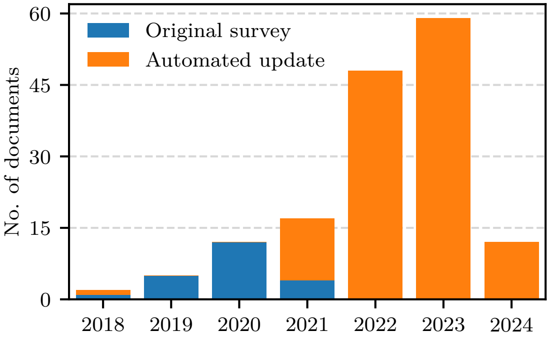


Figure: Publications on FL & IDS (until 2024-04).

[4] Lavaur et al. “The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey”. *IEEE Transactions on Network and Service Management*. 2022

Addressed Challenges

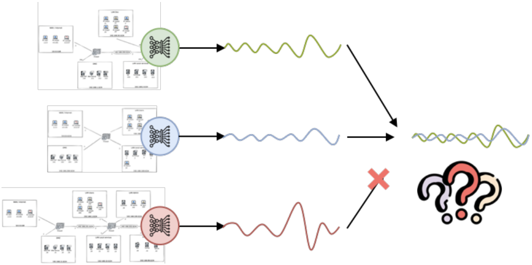


Figure: Heterogeneity headaches.

Challenge I: *Too much heterogeneity leads to poor performance...* [5]

Challenge II: *Difficult to identify malicious contributions when models are different...*

Challenge III: *No representative dataset of heterogeneous distributed intrusion detection...*

[5] Lavaur, Busnel, and Autrel. "Demo: Highlighting the Limits of Federated Learning in Intrusion Detection". *Proceedings of the 44th International Conference on Distributed Computing Systems (ICDCS)*. 2024

Addressed Challenges

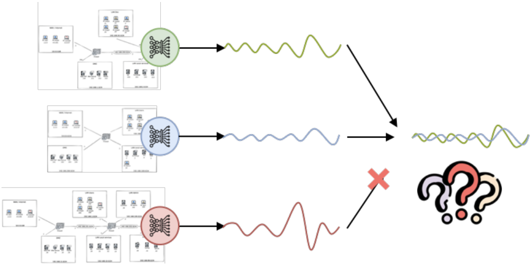


Figure: Heterogeneity headaches.

Challenge I: *Too much heterogeneity leads to poor performance...*

Challenge II: *Difficult to identify malicious contributions when models are different...*

Challenge III: *No representative dataset of heterogeneous distributed intrusion detection...*

Addressed Challenges

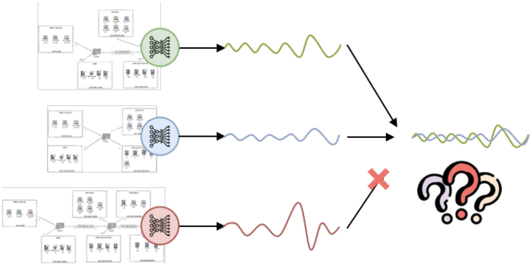


Figure: Heterogeneity headaches.

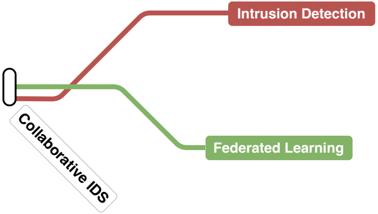
Challenge I: *Too much heterogeneity leads to poor performance...*

Challenge II: *Difficult to identify malicious contributions when models are different...*

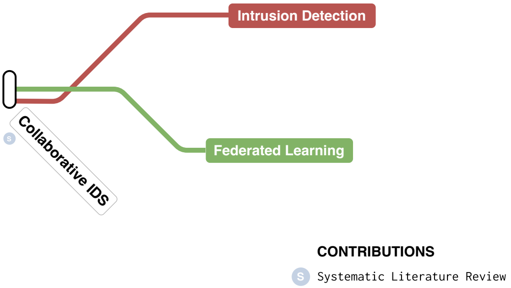
Challenge III: *No representative dataset of heterogeneous distributed intrusion detection...* [6]

[6] Lavaur et al. "Federated Learning as Enabler for Collaborative Security between Not Fully-Trusting Distributed Parties". *Proceedings of the 29th Computer & Electronics Security Application Rendezvous (C&ESAR)*. 2022

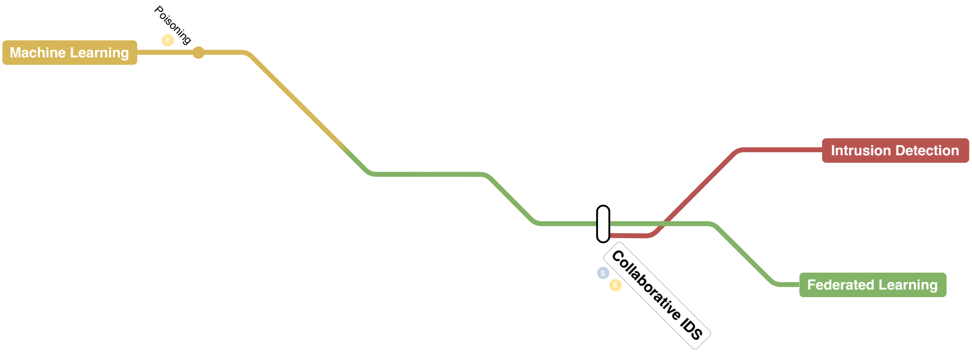
CONTRIBUTIONS SUMMARY



CONTRIBUTIONS SUMMARY



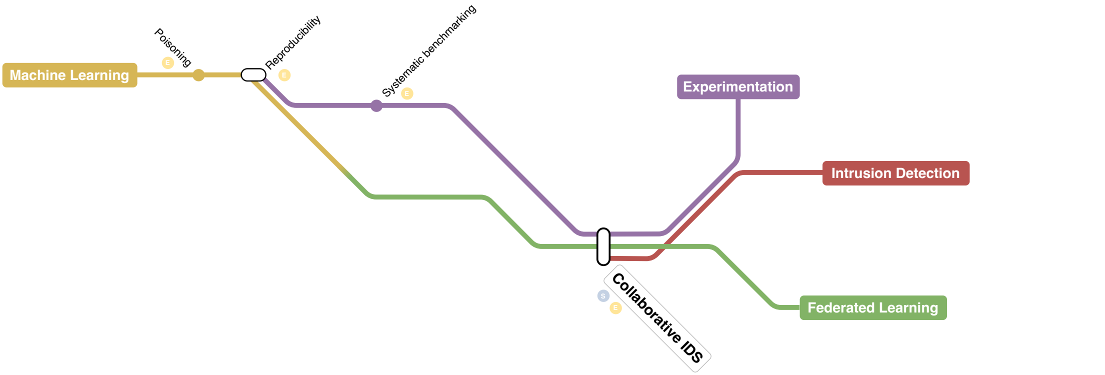
Contributions Summary



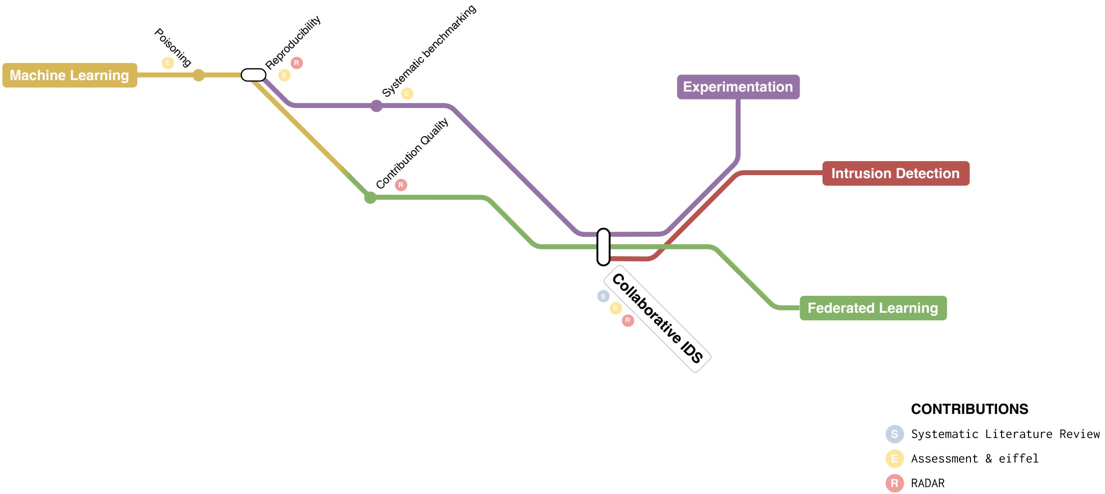
CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel

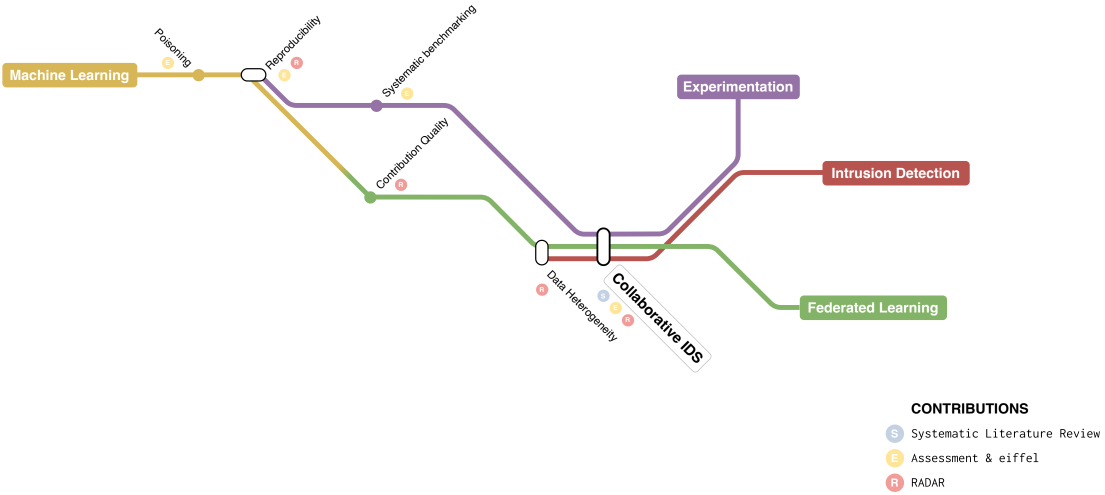
Contributions Summary



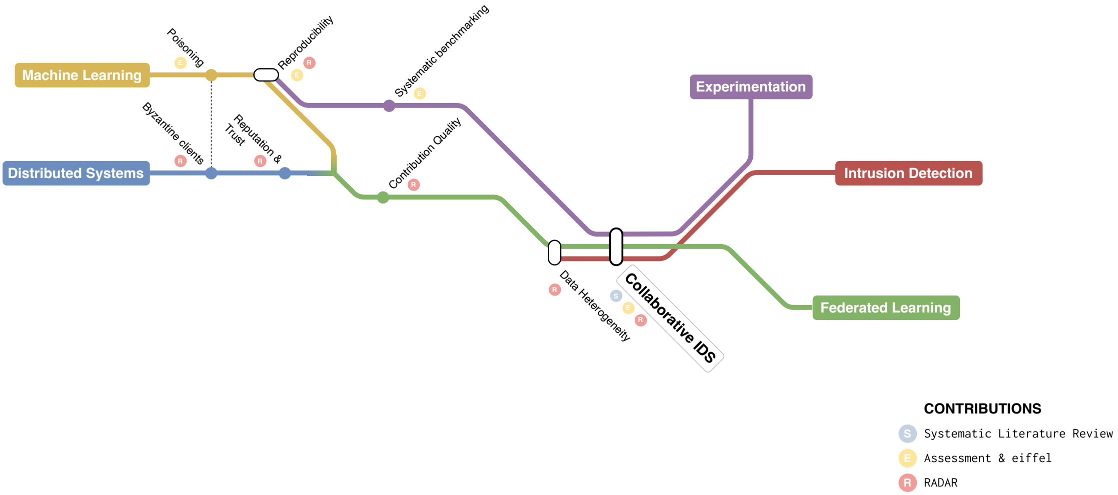
Contributions Summary



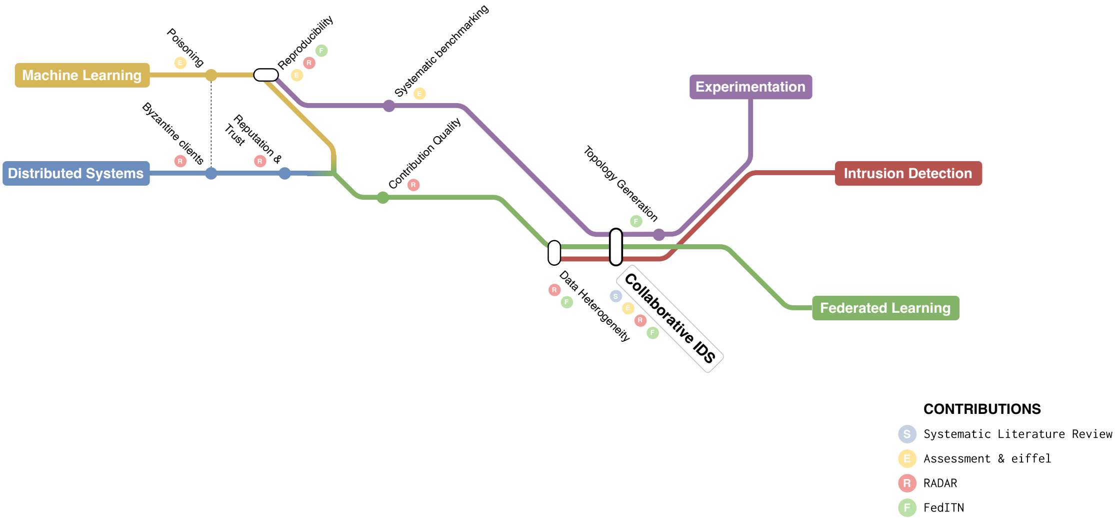
Contributions Summary



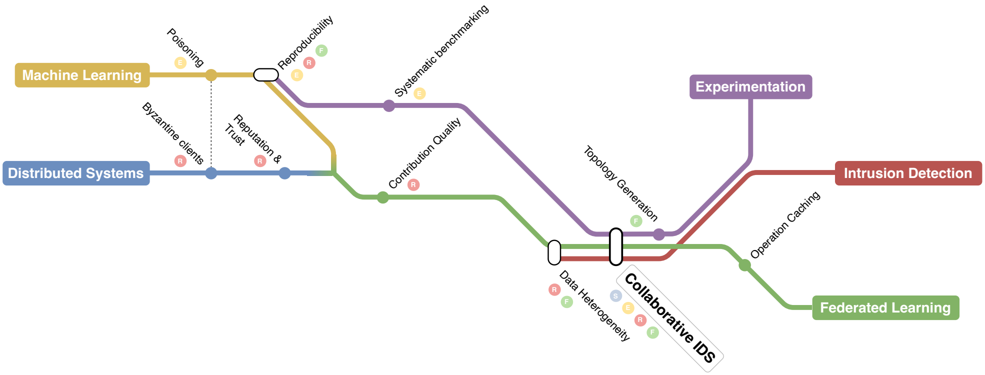
Contributions Summary



Contributions Summary



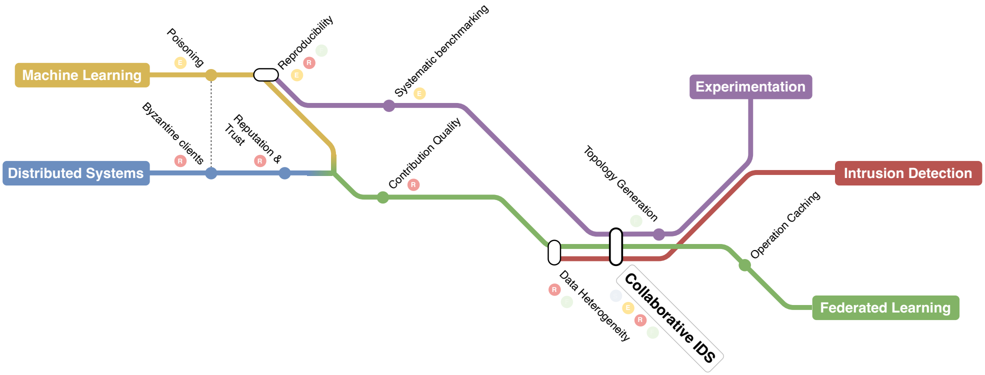
Contributions Summary



CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

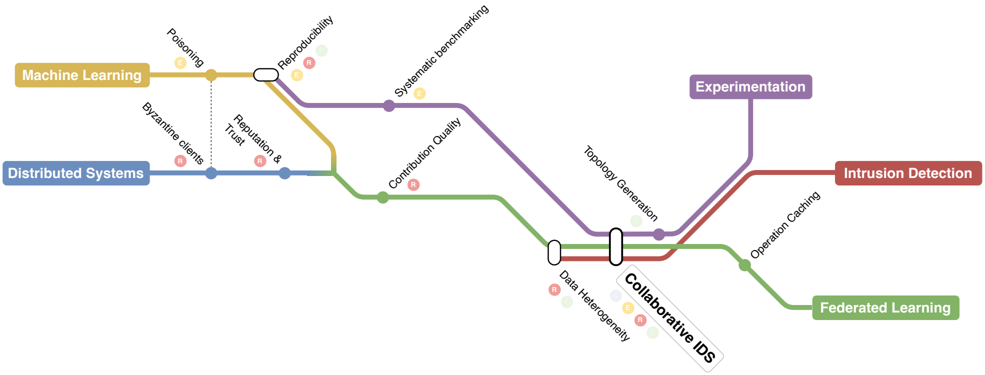
CONTRIBUTIONS SUMMARY



CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Contributions Summary



CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

E Assessing the Impact of Label-Flipping Attacks

[7] Lavaur, Busnel, and Autrel. “Systematic Analysis of Label-flipping Attacks against Federated Learning in Collaborative Intrusion Detection Systems”. *Proceedings of the 19th International Conference on Availability, Reliability and Security (ARES)*. 2024

The Problem of Poisoning Attacks

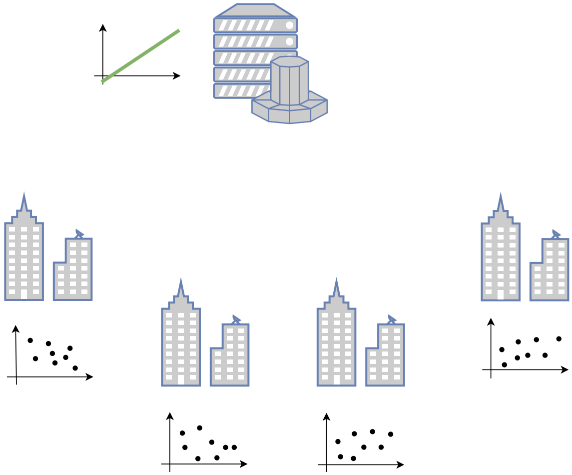


Figure: Poisoning attacks on FL.

The Problem of Poisoning Attacks

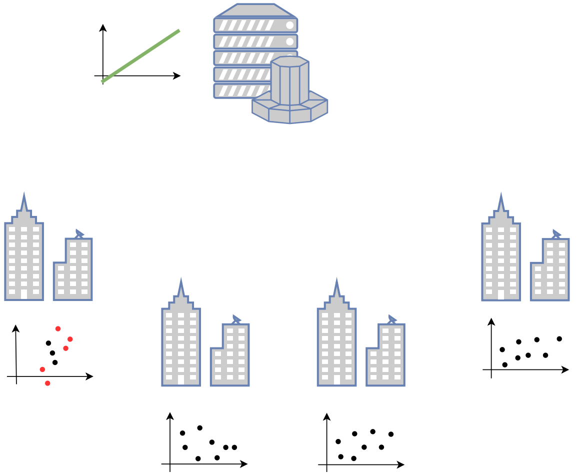


Figure: Poisoning attacks on FL.

The Problem of Poisoning Attacks

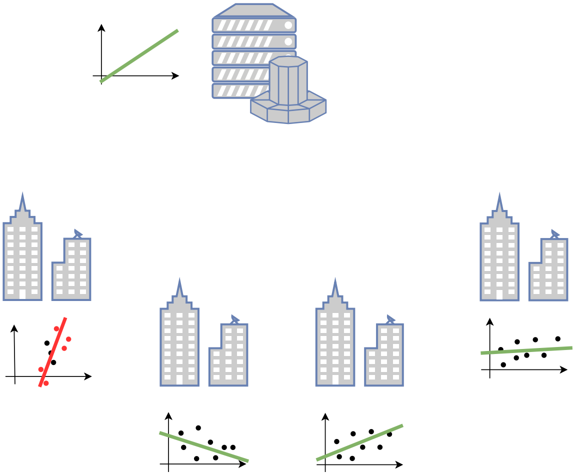


Figure: Poisoning attacks on FL.

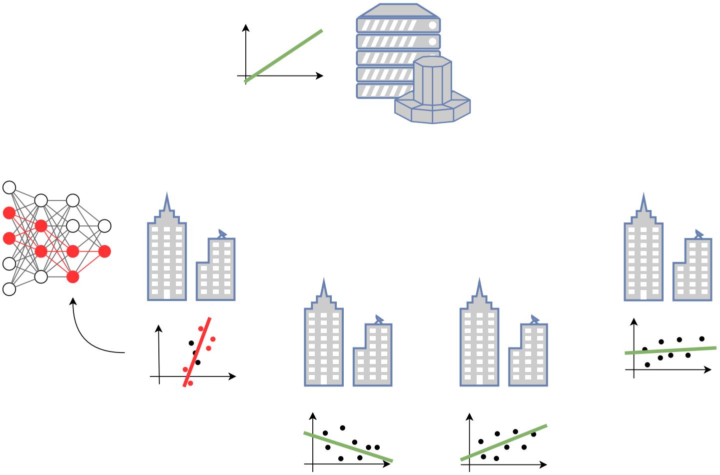


Figure: Poisoning attacks on FL.

The Problem of Poisoning Attacks

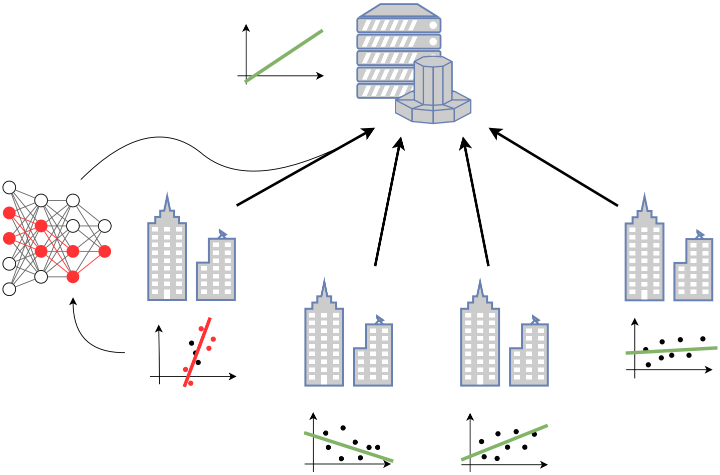


Figure: Poisoning attacks on FL.

The Problem of Poisoning Attacks

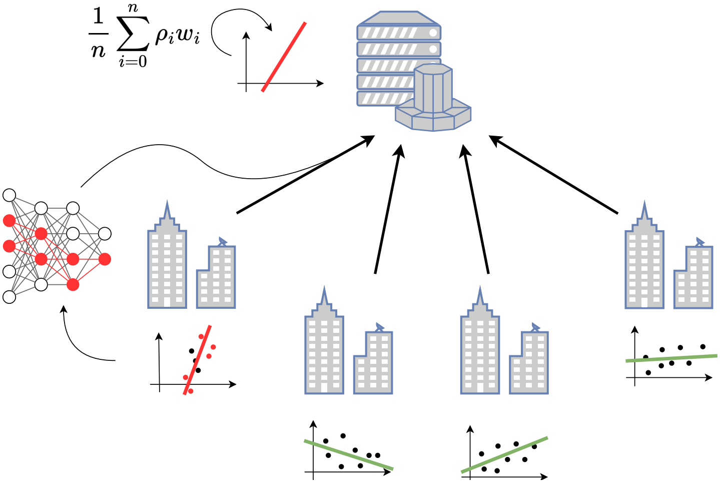


Figure: Poisoning attacks on FL.

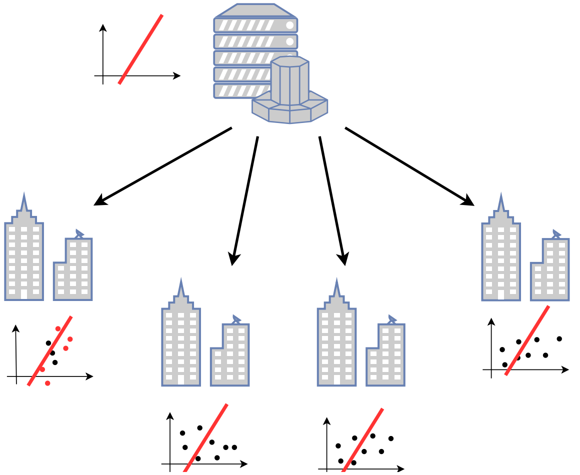


Figure: Poisoning attacks on FL.

Poisoning attacks

Poisoning attacks

Component

- Data poisoning (*e.g.*, label-flipping, clean-label)
- Model poisoning (*e.g.*, gradient boosting)

Poisoning attacks

Component	Objective
▶ Data poisoning (*e.g.*, label-flipping, clean-label) ▶ Model poisoning (*e.g.*, gradient boosting)	▶ Untargeted: impact model performance ▶ Targeted: modify behavior for specific samples

Poisoning attacks

Component

- Data poisoning (*e.g.*, label-flipping, clean-label)
- Model poisoning (*e.g.*, gradient boosting)

Objective

- Untargeted: impact model performance
- Targeted: modify behavior for specific samples

Proportion

- Single attacker
- Colluding attackers: multiple coordinated adversaries

Characterizing Poisoning Attacks

Poisoning attacks

Component

- Data poisoning (*e.g.*, label-flipping, clean-label)
- Model poisoning (*e.g.*, gradient boosting)

Objective

- Untargeted: impact model performance
- Targeted: modify behavior for specific samples

Proportion

- Single attacker
- Colluding attackers: multiple coordinated adversaries

Existing studies

- Often partial, focusing on challenging a specific defense mechanism.
- Lack of reproducibility and comparability (different datasets, models, and attacks).
- No targeted attacks binary classification.

Existing studies

- Often partial, focusing on challenging a specific defense mechanism.
- Lack of reproducibility and comparability (different datasets, models, and attacks).
- No targeted attacks binary classification.

Research Questions

1. Is the behavior of poisoning attacks predictable?
2. Do hyperparameters influence the impact of poisoning attacks?
3. Are IDS backdoors realistic using label-flipping attacks?
4. Is there a critical threshold where label-flipping attacks begin to impact performance?
5. Is gradient similarity enough to detect label-flipping attacks?

Existing studies

- Often partial, focusing on challenging a specific defense mechanism.
- Lack of reproducibility and comparability (different datasets, models, and attacks).
- No targeted attacks binary classification.

Research Questions

1. Is the behavior of poisoning attacks predictable?
2. Do hyperparameters influence the impact of poisoning attacks?
3. Are IDS backdoors realistic using label-flipping attacks?
4. Is there a critical threshold where label-flipping attacks begin to impact performance?
5. Is gradient similarity enough to detect label-flipping attacks?

RQ5: Is Gradient Similarity Enough to Detect Label-Flipping Attacks?

- Known technique to detect poisoning attacks [8].

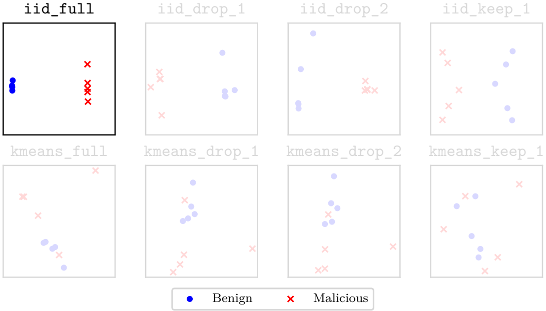


Figure: PCA projection of the gradients in 2D (CICIDS).

[8] Tolpegin et al. “Data Poisoning Attacks Against Federated Learning Systems”. Lecture Notes in Computer Science. 2020

RQ5: Is Gradient Similarity Enough to Detect Label-Flipping Attacks?

- Known technique to detect poisoning attacks [8].
- High heterogeneity makes it harder to detect attackers.

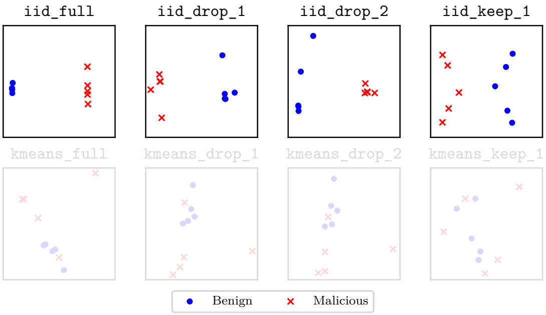


Figure: PCA projection of the gradients in 2D (CICIDS).

[8] Tolpegin et al. "Data Poisoning Attacks Against Federated Learning Systems". Lecture Notes in Computer Science. 2020

RQ5: Is Gradient Similarity Enough to Detect Label-Flipping Attacks?

- Known technique to detect poisoning attacks [8].
- High heterogeneity makes it harder to detect attackers.

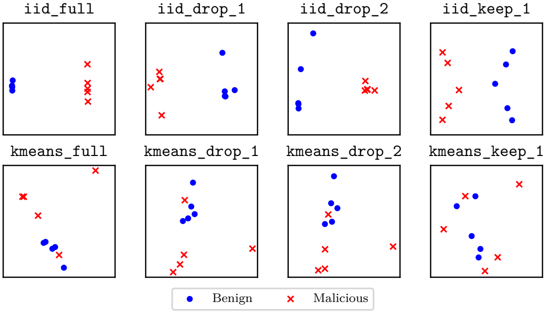


Figure: PCA projection of the gradients in 2D (CICIDS).

[8] Tolpegin et al. “Data Poisoning Attacks Against Federated Learning Systems”. Lecture Notes in Computer Science. 2020

Takeaways

1. A *deeper* understanding of the behavior of label-flipping attacks in FL-based CIDSs.
    - Similarity-based detection techniques show limitations in detecting poisoning attacks.
    - Limited by the models' generalization capabilities and the characteristic overlap between classes.
    - Hyperparameter dependencies, but not on the average performance impact.

Takeaways

1. A *deeper* understanding of the behavior of label-flipping attacks in FL-based CIDSs.
   - Similarity-based detection techniques show limitations in detecting poisoning attacks.
   - Limited by the models' generalization capabilities and the characteristic overlap between classes.
   - Hyperparameter dependencies, but not on the average performance impact.
2. A reproducible evaluation framework to study the impact of label-flipping attacks in FIDS using FL.
   - Reproducible, extendable, and available in open-access[3].
   - Calls to be extended to other poisoning attacks, datasets, and partitioning strategies.

[3] https://github.com/leolavaur/eiffel

R

Fighting Byzantine Contributions in Heterogeneous Settings

[9] Lavaur et al. "RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning". *Proceedings of the 43rd International Symposium on Reliable Distributed Systems (SRDS)*. 2024

Case study reminder

- Multiple organizations collaborating on a federated Intrusion Detection System.
- Partial heterogeneity in the datasets: different data distributions but existing similarities.

Case study reminder

- Multiple organizations collaborating on a federated Intrusion Detection System.
- Partial heterogeneity in the datasets: different data distributions but existing similarities.

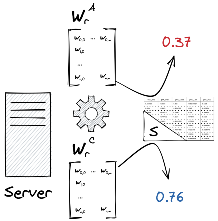
Byzantine contributions:

- data quality issues (*e.g.*, labelling, noise);
- distribution mismatches; and
- adversaries, *possibly colluding*.

Quality Assessment in Heterogeneous Settings

For n participants p_i and their local datasets d_i of unknown similarity, each participant uploads a model update w_i^r at each round r. Given $P = \{p_1, p_2, \ldots, p_n\}$ and $W = \{w_1^r, w_2^r, \ldots, w_n^r\}$, how can one assess the quality of each participant's contribution without making assumptions on the data distribution across the datasets d_i?

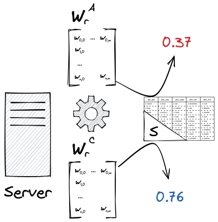
Server-side evaluation [10]



- Only applicable in IID settings.
- Single source of truth.

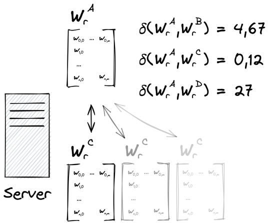
[10] Zhou et al. "A Differentially Private Federated Learning Model against Poisoning Attacks in Edge Computing". *IEEE Transactions on Dependable and Secure Computing*. 2022

Server-side evaluation [10]



- Only applicable in IID settings.
- Single source of truth.

Server-side comparison [11]

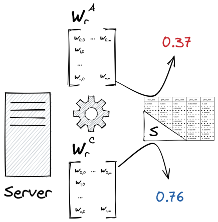


- Less related to client data.

[10] Zhou et al. "A Differentially Private Federated Learning Model against Poisoning Attacks in Edge Computing". *IEEE Transactions on Dependable and Secure Computing*. 2022

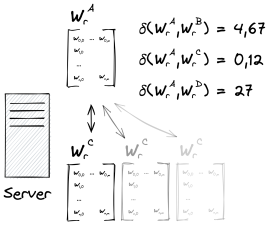
[11] Briggs, Fan, and Andras. "Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data". *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020

Server-side evaluation [10]



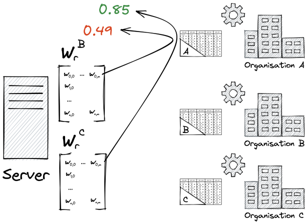
- Only applicable in IID settings.
- Single source of truth.

Server-side comparison [11]



- Less related to client data.

Client-side evaluation [12]



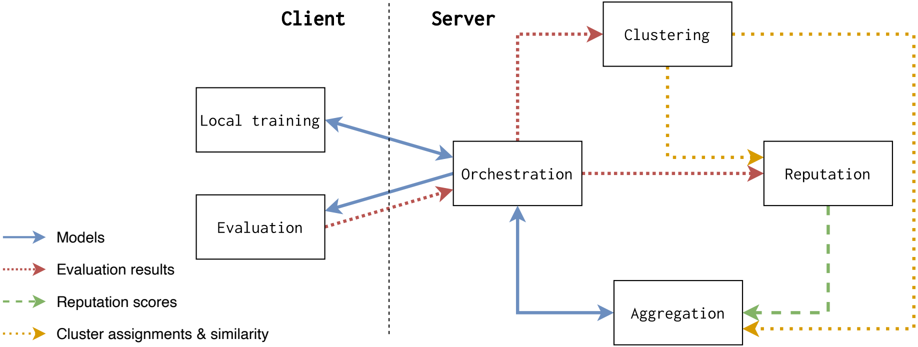
- High cost in cross-device.
- More susceptible to badmouthing.

[10] Zhou et al. "A Differentially Private Federated Learning Model against Poisoning Attacks in Edge Computing". *IEEE Transactions on Dependable and Secure Computing*. 2022

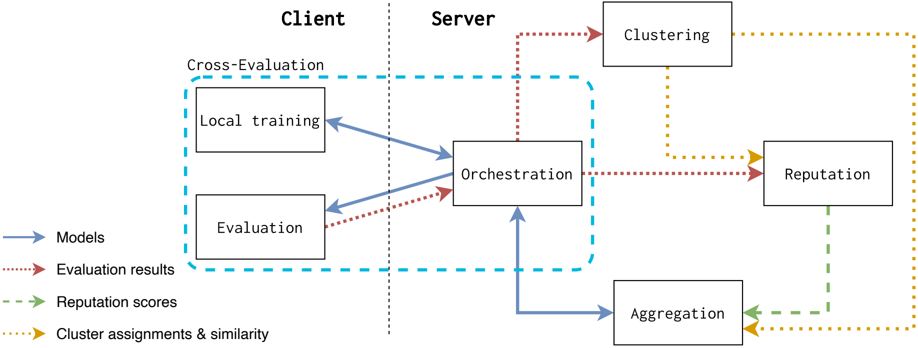
[11] Briggs, Fan, and Andras. "Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data". *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020

[12] Zhao et al. *Shielding Collaborative Learning: Mitigating Poisoning Attacks through Client-Side Detection*. 2020

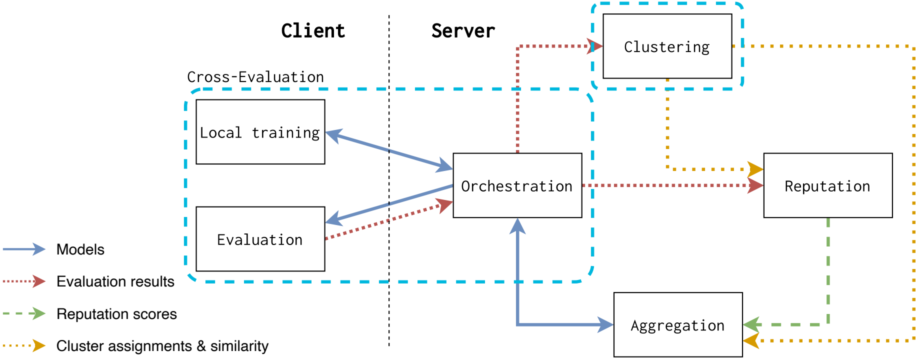
Introducing RADAR



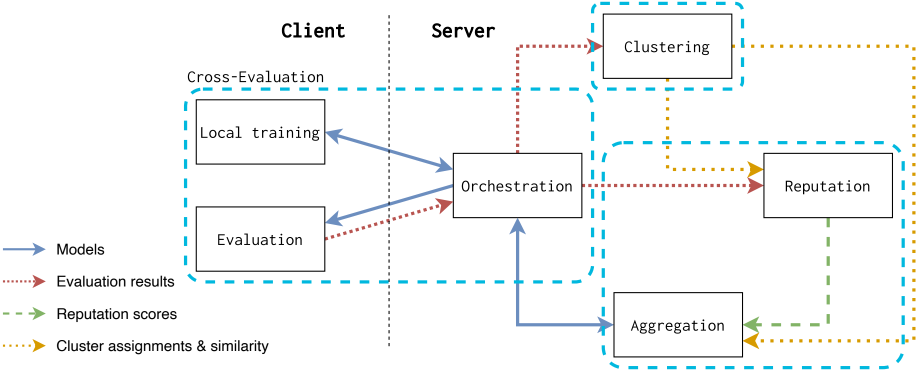
RADAR architecture.



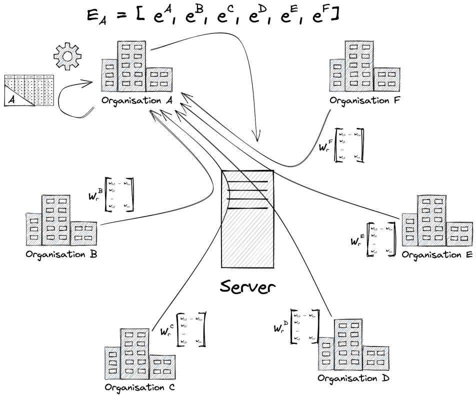
RADAR architecture.



RADAR architecture.

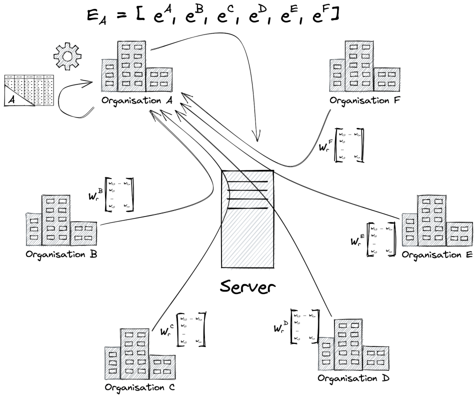


RADAR architecture.



Advantages

- Exhaustive overview of the entire system at each round *r*. **No need of prior knowledge!**
- Evaluations (*e.g.*, accuracy, F1 score) representative of participants' data.

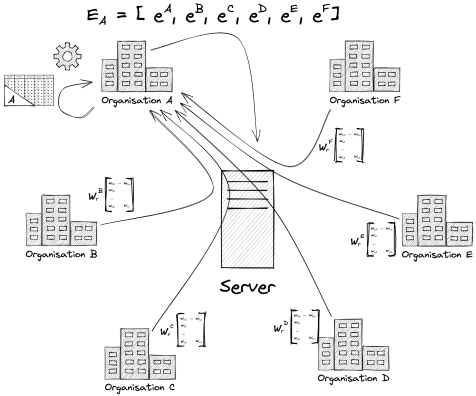


Advantages

- Exhaustive overview of the entire system at each round *r*. No need of prior knowledge!
- Evaluations (*e.g.*, accuracy, F1 score) representative of participants' data.

Drawbacks

- High communication and computation costs.
- Does not scale well.



Advantages

- Exhaustive overview of the entire system at each round r. No need of prior knowledge!
- Evaluations (*e.g.*, accuracy, F1 score) representative of participants' data.

Drawbacks

- High communication and computation costs.
- Does not scale well.

But...

- Cross-silo use case: few clients, with reasonable computing capacity.
- Slow workflow: long time between rounds.

Objective

- Build *more* homogeneous communities of participants to facilitate model aggregation.

Objective

- Build *more* homogeneous communities of participants to facilitate model aggregation.

- Distance metric
  - Based on **cross-evaluation** results.
  - **Cosine similarity** [11].

[11] Briggs, Fan, and Andras. "Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data". *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020

Objective

- Build *more* homogeneous communities of participants to facilitate model aggregation.

- Distance metric
  - Based on cross-evaluation results.
  - Cosine similarity [11].
- Algorithm
  - Hierarchical clustering. [11]
  - Dynamic aggregation threshold.

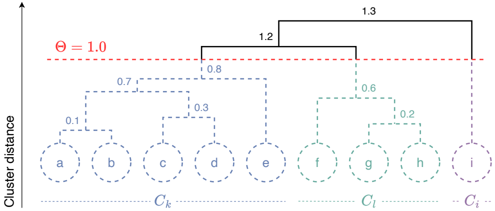


Figure: Hierarchical clustering.

[11] Briggs, Fan, and Andras. "Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data". *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020

Definition: Reputation Systems [13]

- Long-lived entities expecting future interaction.
- Capture and distribution of feedback about current interactions (such information must be visible in the future).
- Use of feedback to guide trust decisions.

[13] Resnick et al. "Reputation Systems". *Communications of the ACM*. 2000

Reputation-aware Aggregation

Definition: Reputation Systems [13]

- Long-lived entities expecting future interaction.
- Capture and distribution of feedback about current interactions (such information must be visible in the future).
- Use of feedback to guide trust decisions.

- Votes weighted by the similarity inside each cluster.
- Exponential decay for potential redemption.

[13] Resnick et al. “Reputation Systems”. *Communications of the ACM*. 2000

Datasets

- Heterogeneous datasets, but some participants can share similarities.
- 4 datasets: CIC-CSE-IDS2018, UNSW-NB15, Bot-IoT, ToN_IoT.
- NF-V2 [14] feature set (*i.e.*, NetFlow V9).

[14] Sarhan, Layeghy, and Portmann. *Towards a Standard Feature Set for Network Intrusion Detection System Datasets*. 2021

Setup: Byzantine Scenarios

Parameters

- *Target*: Affected classes.
- *Data Poisoning Rate (DPR)*: proportion of targeted data with flipped labels.
- *Model Poisoning Rate (MPR)*: number of attackers in the cluster.

`colluding minority 100T`

(*i.e.*, 2 attackers, 100% DPR on Reconnaissance class).

Table: *Effect of different attack configurations (untargeted) on all baselines.* **RA** is RADAR, **FG** is `FoolsGold`, **FA** is `FedAvg` (on *all* participants), and **FC** is `FedAvg` ideally clustered per dataset.

Scenario	ASR (%)			
	RA	FG	FA	FC
Targeted (`100T`)				
`Benign`	**0.00**	5.17	5.10	0.09
`Lone`	**0.00**	93.82	6.73	0.45
`Collud. min.`	**0.00**	2.97	9.99	53.40
`Collud. maj.`	73.39	**8.10**	17.65	59.36
Untargeted (`100U`)				
`Benign`	0.09	0.39	33.30	**0.06**
`Lone`	**0.08**	99.89	54.70	0.12
`Collud. min.`	0.10	**0.04**	44.53	6.26
`Collud. maj.`	**0.08**	38.98	59.49	94.36

lower is better

Table: *Effect of different attack configurations (untargeted) on all baselines.* **RA** is RADAR, **FG** is `FoolsGold`, **FA** is `FedAvg` (on *all* participants), and **FC** is `FedAvg` ideally clustered per dataset.

Scenario	ASR (%) RA	FG	FA	FC
Targeted (`100T`)				
`Benign`	**0.00**	5.17	5.10	0.09
`Lone`	**0.00**	93.82	6.73	0.45
`Collud. min.`	**0.00**	2.97	9.99	53.40
`Collud. maj.`	73.39	**8.10**	17.65	59.36
Untargeted (`100U`)				
`Benign`	0.09	0.39	33.30	**0.06**
`Lone`	**0.08**	99.89	54.70	0.12
`Collud. min.`	0.10	**0.04**	44.53	6.26
`Collud. maj.`	**0.08**	38.98	59.49	94.36

lower is better

Table: *Effect of different attack configurations (untargeted) on all baselines.* **RA** is RADAR, **FG** is `FoolsGold`, **FA** is `FedAvg` (on *all* participants), and **FC** is `FedAvg` ideally clustered per dataset.

Scenario	ASR (%) RA	FG	FA	FC
Targeted (`100T`)				
`Benign`	**0.00**	5.17	5.10	0.09
`Lone`	**0.00**	93.82	6.73	0.45
`Collud. min.`	**0.00**	2.97	9.99	53.40
`Collud. maj.`	73.39	**8.10**	17.65	59.36
Untargeted (`100U`)				
`Benign`	0.09	0.39	33.30	**0.06**
`Lone`	**0.08**	99.89	54.70	0.12
`Collud. min.`	0.10	**0.04**	44.53	6.26
`Collud. maj.`	**0.08**	38.98	59.49	94.36

lower is better

Takeways

1. RADAR can:
   - leverage cross-evaluation, clustering and reputation to address heterogeneity and Byzantine contributions;
   - adjust rapidly to changes in behavior; and
   - mitigate most tested scenarios (limiting case handled up to 80% of poisoned data).

Takeways

1. RADAR can:
   - leverage cross-evaluation, clustering and reputation to address heterogeneity and Byzantine contributions;
   - adjust rapidly to changes in behavior; and
   - mitigate most tested scenarios (limiting case handled up to 80% of poisoned data).
2. How generic?
   - Only few conditions: parametric models, locally owned evaluation set, a small-scale use case, and a trusted central server.

Takeways

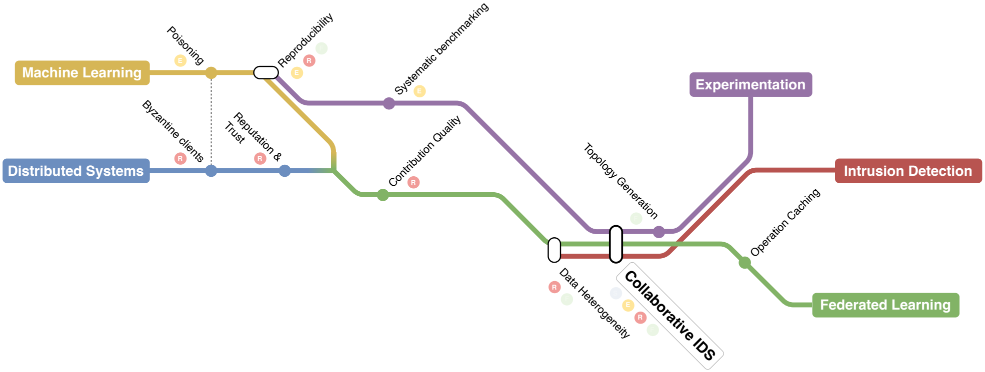
1. RADAR can:
    - leverage cross-evaluation, clustering and reputation to address heterogeneity and Byzantine contributions;
    - adjust rapidly to changes in behavior; and
    - mitigate most tested scenarios (limiting case handled up to 80% of poisoned data).
2. How generic?
    - Only few conditions: parametric models, locally owned evaluation set, a small-scale use case, and a trusted central server.

Takeways

1. RADAR can:
   - leverage cross-evaluation, clustering and reputation to address heterogeneity and Byzantine contributions;
   - adjust rapidly to changes in behavior; and
   - mitigate most tested scenarios (limiting case handled up to 80% of poisoned data).
2. How generic?
   - Only few conditions: parametric models, locally owned evaluation set, a small-scale use case, and a trusted central server.
3. Future works:
   - Remove the central server dependency for increased trust and scalability.
   - Test the approach in more realistic heterogeneous settings.

Conclusion

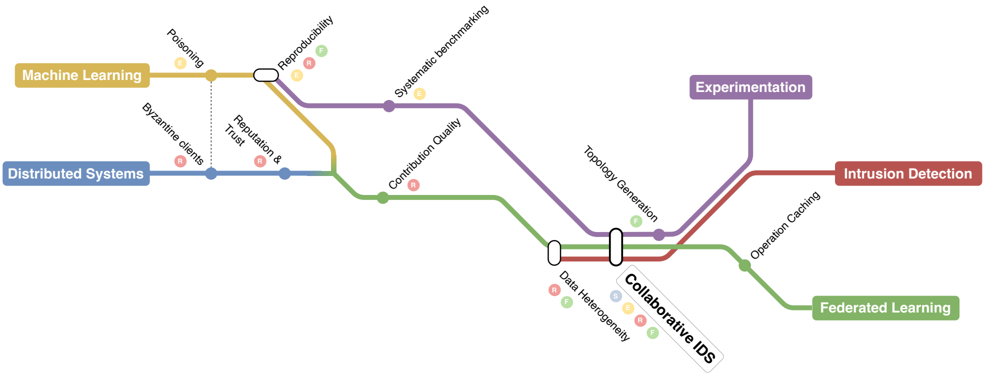
CONTRIBUTIONS



CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Future Work



CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Future Work

Machine Learning
Distributed Systems
Experimentation
Intrusion Detection
Federated Learning
Poisoning
Reproducibility
Systematic benchmarking
Byzantine clients
Reputation & Trust
Contribution Quality
Topology Generation
Operation Caching
Decentralized Reputation Systems
Decentralized Quality Assessment
Decentralized Community Clustering
Data Heterogeneity
Collaborative IDS
CONTRIBUTIONS
S Systematic Literature Review
E Assessment & eiffel
R RADAR
F FedITN

Future Work


Machine Learning
Distributed Systems
Experimentation
Intrusion Detection
Federated Learning
Comparable Evaluations
Poisoning
Reproducibility
Systematic benchmarking
Byzantine clients
Reputation & Trust
Contribution Quality
Topology Generation
Operation Caching
Decentralized Reputation Systems
Decentralized Quality Assessment
Decentralized Community Clustering
Data Heterogeneity
Collaborative IDS
CONTRIBUTIONS
S Systematic Literature Review
E Assessment & eiffel
R RADAR
F FedITN

Future Work

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Perspectives: Going beyond FIDSs

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Perspectives: Going beyond FIDSs

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Perspectives: Going beyond FIDSs

Perspectives: Going beyond FIDSs

Perspectives: Going beyond FIDSs

Perspectives: Going beyond FIDSs

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Perspectives: Going beyond FIDSs

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

PERSPECTIVES: GOING BEYOND FIDSS

CONTRIBUTIONS

- S Systematic Literature Review
- E Assessment & eiffel
- R RADAR
- F FedITN

Perspectives: Going beyond FIDSs

Thank you for your attention!

Improving Intrusion Detection in Distributed Systems with Federated Learning

- Three publications in international conferences: ICDCS 2024, ARES (BASS) 2023, and SRDS 2024.
- One article in an international journal: IEEE TNSM.
- National and international tutorials on Federated Learning for Intrusion Detection: EUR CyberSchool's Spring Research School 2023, NoF 2023 and ICDCS 2024.

[1] National Institute of Standards and Technology. *The NIST Cybersecurity Framework (CSF) 2.0.* NIST CSWP 29. Gaithersburg, MD: National Institute of Standards and Technology, Feb. 26, 2024, NIST CSWP 29. DOI: 10.6028/NIST.CSWP.29. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf (visited on 05/23/2024).

[2] Brendan McMahan et al. **"Communication-Efficient Learning of Deep Networks from Decentralized Data".** In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, Apr. 20–22, 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[3] Peter Kairouz et al. **"Advances and Open Problems in Federated Learning".** Mar. 8, 2021. arXiv: 1912.04977 [cs, stat]. URL: http://arxiv.org/abs/1912.04977 (visited on 04/01/2022).

[4] Léo Lavaur et al. **"The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey".** In: *IEEE Transactions on Network and Service Management*. Special Issue on Network Security Management (June 2022).

[5] Léo Lavaur, Yann Busnel, and Fabien 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, July 2024.

[6] Léo Lavaur et al. **"Federated Learning as Enabler for Collaborative Security between Not Fully-Trusting Distributed Parties".** In: *Proceedings of the 29th Computer & Electronics Security Application Rendezvous (C&ESAR)*. Rennes, France, Oct. 2022.

[7] Léo Lavaur, Yann Busnel, and Fabien Autrel. **"Systematic Analysis of Label-flipping Attacks against Federated Learning in Collaborative Intrusion Detection Systems".** In: *Proceedings of the 19th International Conference on Availability, Reliability and Security (ARES). Workshop on Behavioral Authentication for System Security (BASS)*. Vienna, Austria, Aug. 2024.

[8] Vale Tolpegin et al. **"Data Poisoning Attacks Against Federated Learning Systems".** In: *Computer Security – ESORICS 2020*. Ed. by Liqun Chen et al. 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.

[9] Léo Lavaur et al. **"RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning".** In: *Proceedings of the 43rd International Symposium on Reliable Distributed Systems (SRDS)*. Charlotte, NC, USA, Sept. 2024.

[10] Jun Zhou et al. **"A Differentially Private Federated Learning Model against Poisoning Attacks in Edge Computing".** In: *IEEE Transactions on Dependable and Secure Computing* (2022), pp. 1–1. ISSN: 1941-0018. DOI: 10.1109/TDSC.2022.3168556.

[11] Christopher Briggs, Zhong Fan, and Peter Andras. **"Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data".** In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020 International Joint Conference on Neural Networks (IJCNN). July 2020, pp. 1–9. DOI: 10.1109/IJCNN48605.2020.9207469.

[12] Lingchen Zhao et al. ***Shielding Collaborative Learning: Mitigating Poisoning Attacks through Client-Side Detection.*** Mar. 9, 2020. arXiv: 1910.13111 [cs]. URL: http://arxiv.org/abs/1910.13111 (visited on 08/28/2022). Pre-published.

[13] Paul Resnick et al. **"Reputation Systems".** In: *Communications of the ACM* 43.12 (Dec. 1, 2000), pp. 45–48. ISSN: 0001-0782. DOI: 10.1145/355112.355122. URL: https://doi.org/10.1145/355112.355122 (visited on 02/01/2023).

[14] Mohanad Sarhan, Siamak Layeghy, and Marius Portmann. ***Towards a Standard Feature Set for Network Intrusion Detection System Datasets.*** May 14, 2021. arXiv: 2101.11315 [cs]. URL: http://arxiv.org/abs/2101.11315 (visited on 09/12/2022). Pre-published.

[15] Rafael Uetz et al. **“Reproducible and Adaptable Log Data Generation for Sound Cybersecurity Experiments”.** In: *Annual Computer Security Applications Conference*. ACSAC '21: Annual Computer Security Applications Conference. Virtual Event USA: ACM, Dec. 6, 2021, pp. 690–705. ISBN: 978-1-4503-8579-4. DOI: 10.1145/3485832.3488020. URL: https://dl.acm.org/doi/10.1145/3485832.3488020 (visited on 08/08/2022).

[16] ACM. ***Artifact Review and Badging v1.1.*** Aug. 24, 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current (visited on 08/17/2022).

[17] Daniel J Beutel et al. **“Flower: A Friendly Federated Learning Research Framework”.** 2020. arXiv: 2007.14390.

[18] Eelco Dolstra. **“The Purely Functional Software Deployment Model”.** S.l.: s.n., 2006.

Extra Slides

Assessment

Sound experiments [15]; [16]:

- *valid* (i.e., well-defined and unrefutable);
- *controllable* (e.g., parameterized); and
- *reproducible* (i.e., the same results can be obtained by another group using the author's artefact).

[15] Uetz et al. "Reproducible and Adaptable Log Data Generation for Sound Cybersecurity Experiments". *Annual Computer Security Applications Conference*. 2021
[16] ACM. *Artifact Review and Badging v1.1*. 2020

Experimental Setup

Experiment orchestration using `Eiffel` [5].

- `Flower` simulation framework [17] for Federated Learning (FL).
- `Hydra` for experiment generation and configuration.
- Custom-made poisoning engine with different attack strategies.
- Nix [18] and Poetry to fix system and Python dependencies, enabling reproducibility.

1,067 experiments $\times$ 10 seeds (1,613 hours of computation.)

[17] Beutel et al. “Flower: A Friendly Federated Learning Research Framework”. 2020
[18] Dolstra. “The Purely Functional Software Deployment Model”. 2006

RQ1: Are Poisoning Attacks Predictable?

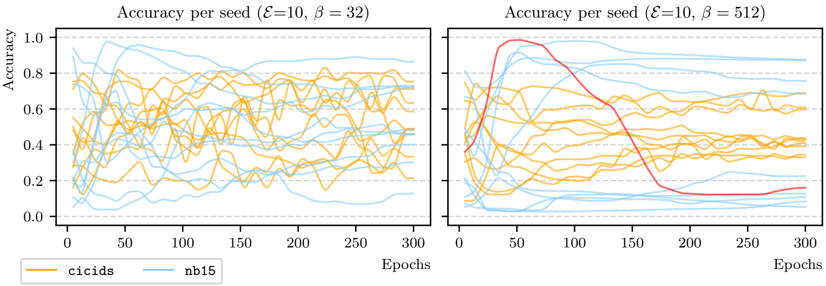


Figure: Predictability of label-flipping attacks.

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

RQ2: Do Hyperparameters Influence the Impact of Poisoning Attacks?

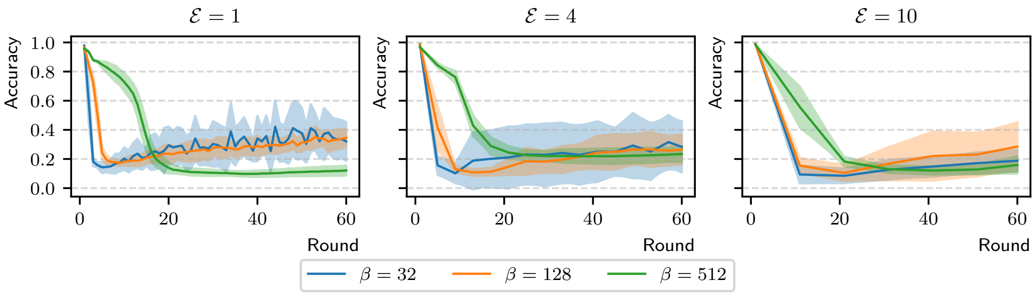


Figure: Effect of the hyperparameters on the accuracy of the poisoned model in the late scenario (50% attackers, CICIDS).

- `late-3` scenario: attackers start poisoning after 3 rounds
- High batch size leads to more inertia, less instantaneous impact
  $\rightarrow$ More impactful in constrained environments

Extra Slides

RADAR

Results

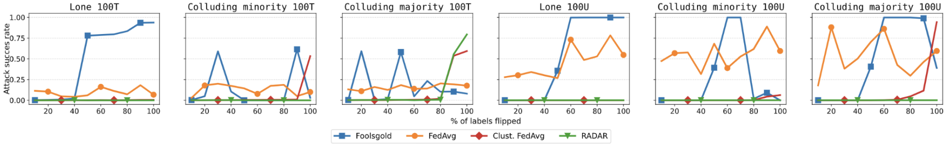


Figure: Baseline comparison.

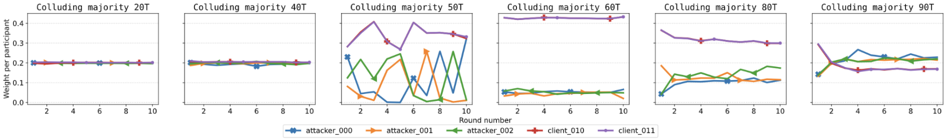


Figure: RADAR's limiting scenario.