Towards Sybil resilience in Decentralized Learning

Thomas Werthenbach
July 4th, 2023

Student number: 4772466
Thesis committee: Dr. ir. J.A. Pouwelse (supervisor)
Dr. D.M.J. Tax

TU Delft

Introduction
Related work
SybilWall
Evaluation
Conclusion

Introduction

Related work
SybilWall
Evaluation
Conclusion

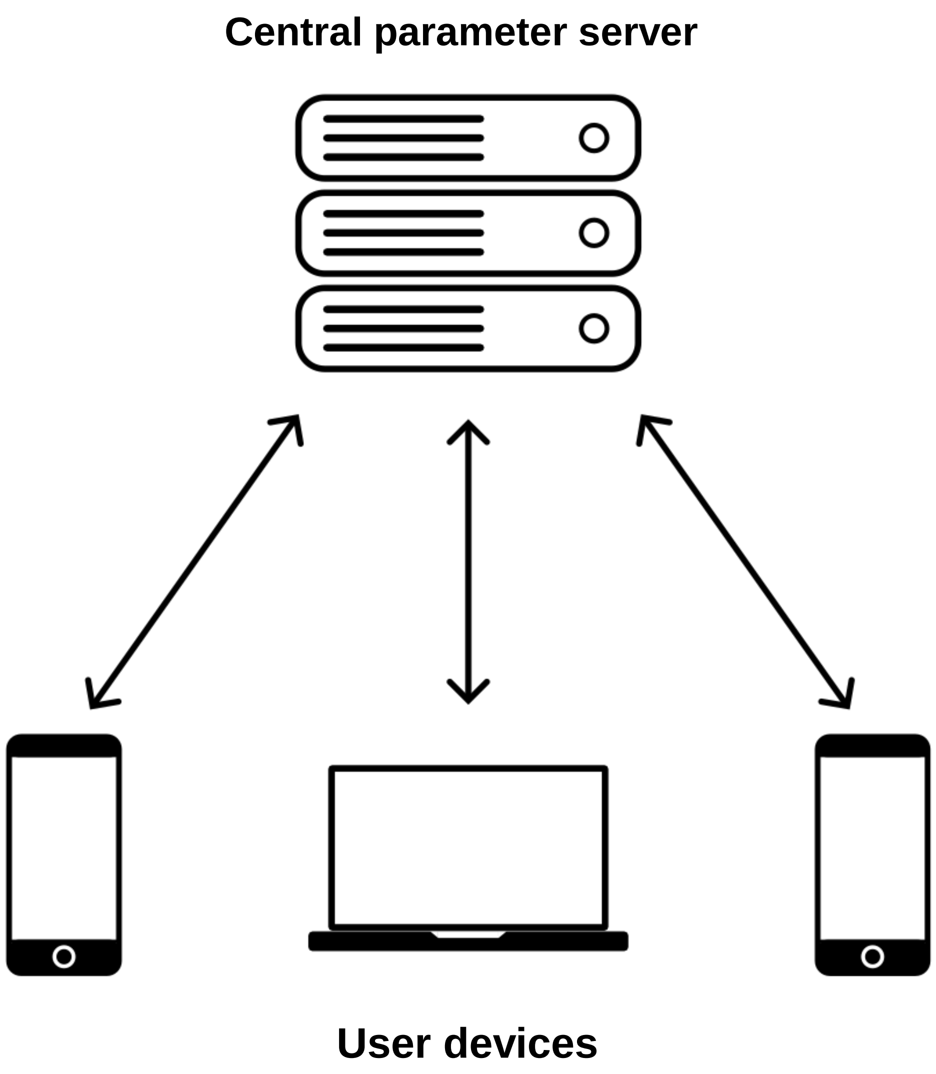
TU Delft

Introduction

- Recent AI developments
- Training requires large datasets
- Privacy law prohibit mass user data collection.
- How does one perform machine learning on comprehensive datasets while respecting privacy rights?

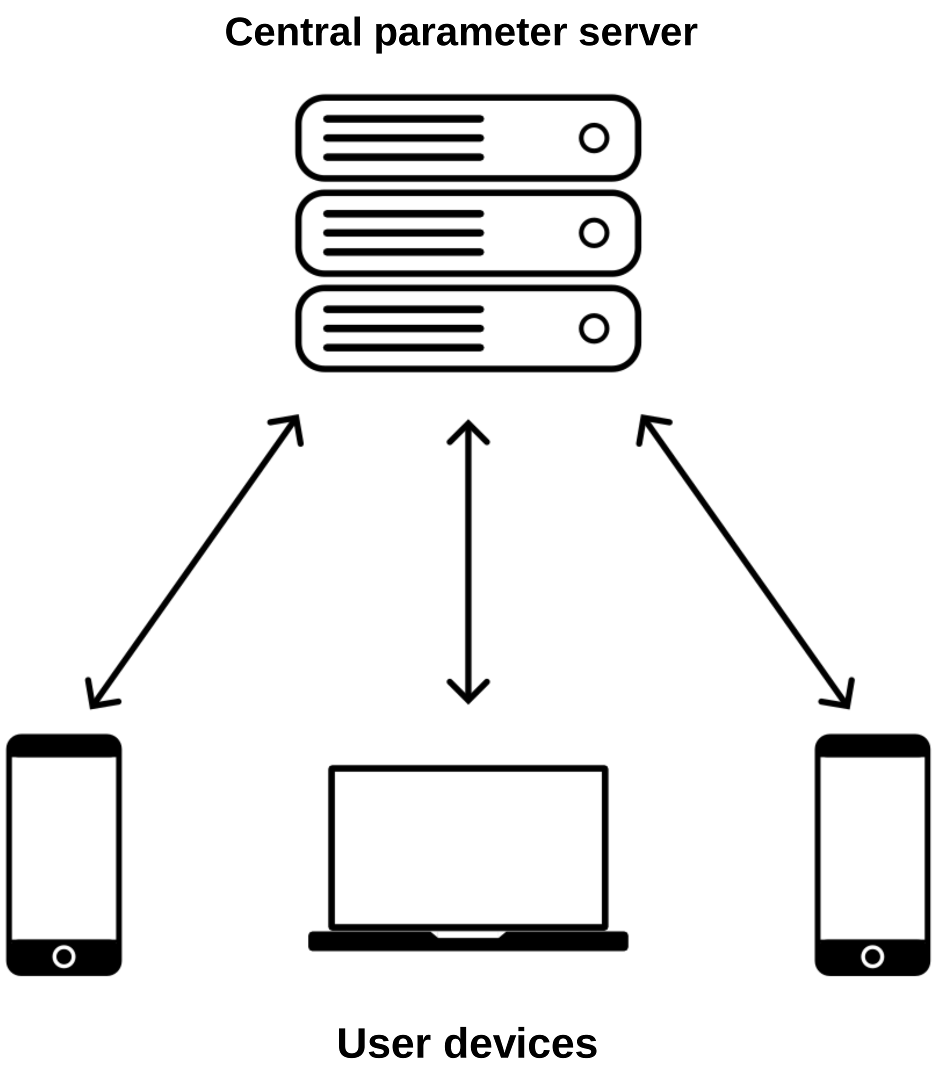
Federated learning

- Training performed on end-user devices
- Real user data
- Centralized model aggregator
- Privacy-enforcing
- Synchronous training rounds



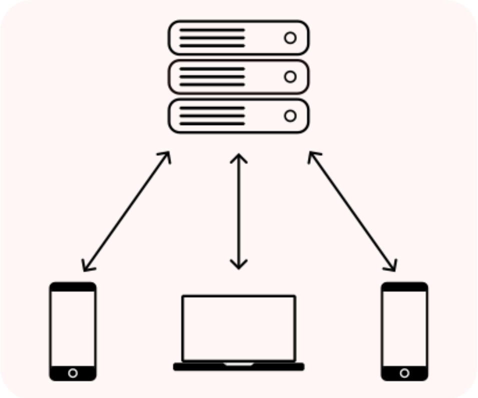
Federated learning training round

1. Train on local data
2. Send gradients to central parameter server
3. Server aggregates
4. Send model to edge devices
5. Repeat



Federated learning

Federated learning



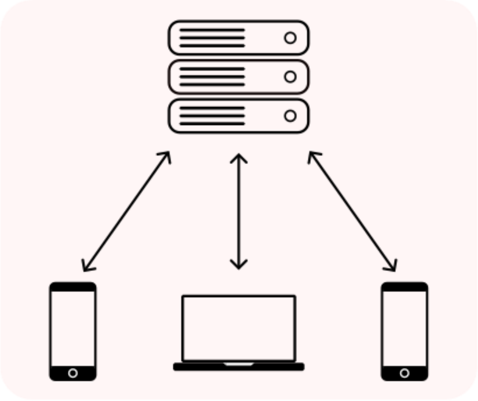
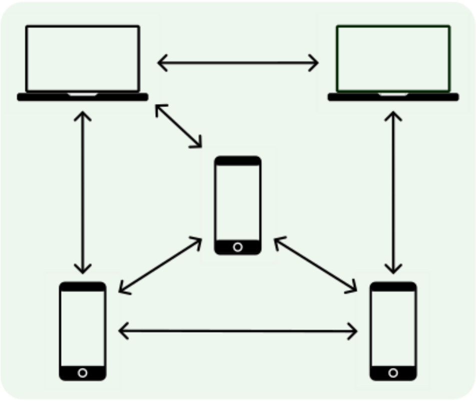
Advantages

- Privacy-enforcing
- Real-user data

Drawbacks

- Scalability
- Single point of failure

Federated learning vs decentralized learning

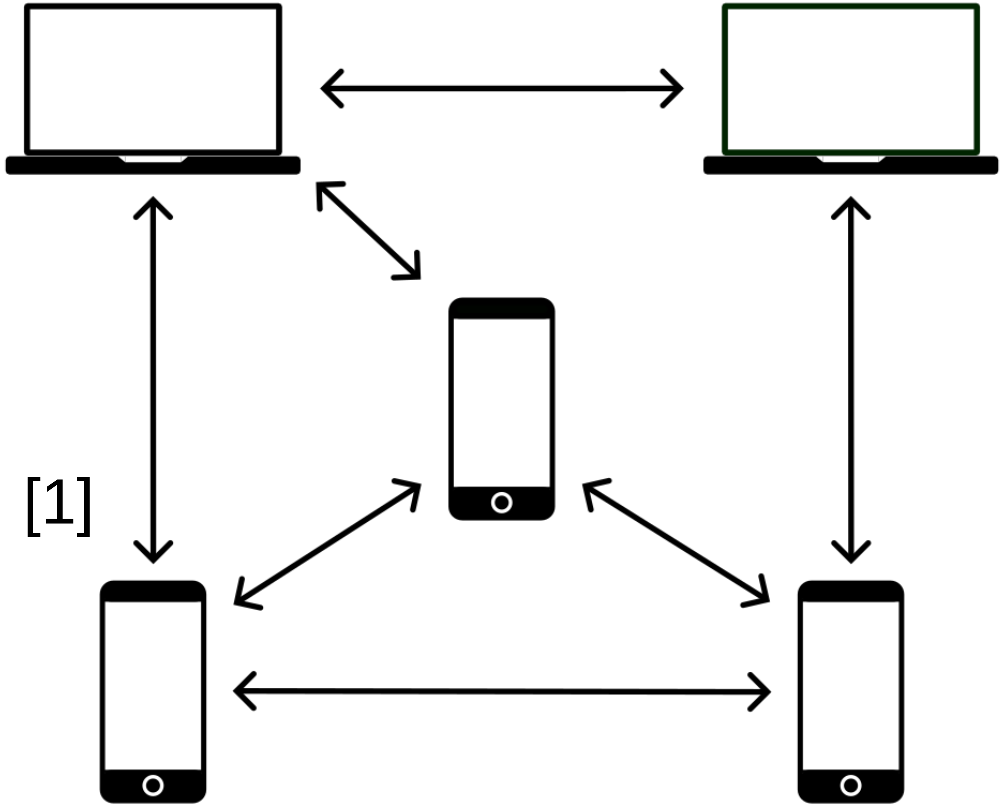
Federated learning	Decentralized learning
	
Advantages ▪ Privacy-enforcing ▪ Real-user data	Advantages ▪ Privacy-enforcing ▪ Real-user data ▪ Boundless scalability ▪ No single point of failure
Drawbacks ▪ Scalability ▪ Single point of failure	Drawbacks ▪ Limited context

TU Delft

Decentralized learning

- Decentralized
- Improved scalability
  - Communication costs
  - Memory capacity
  - Aggregation time
- Performance similar to federated learning [1]

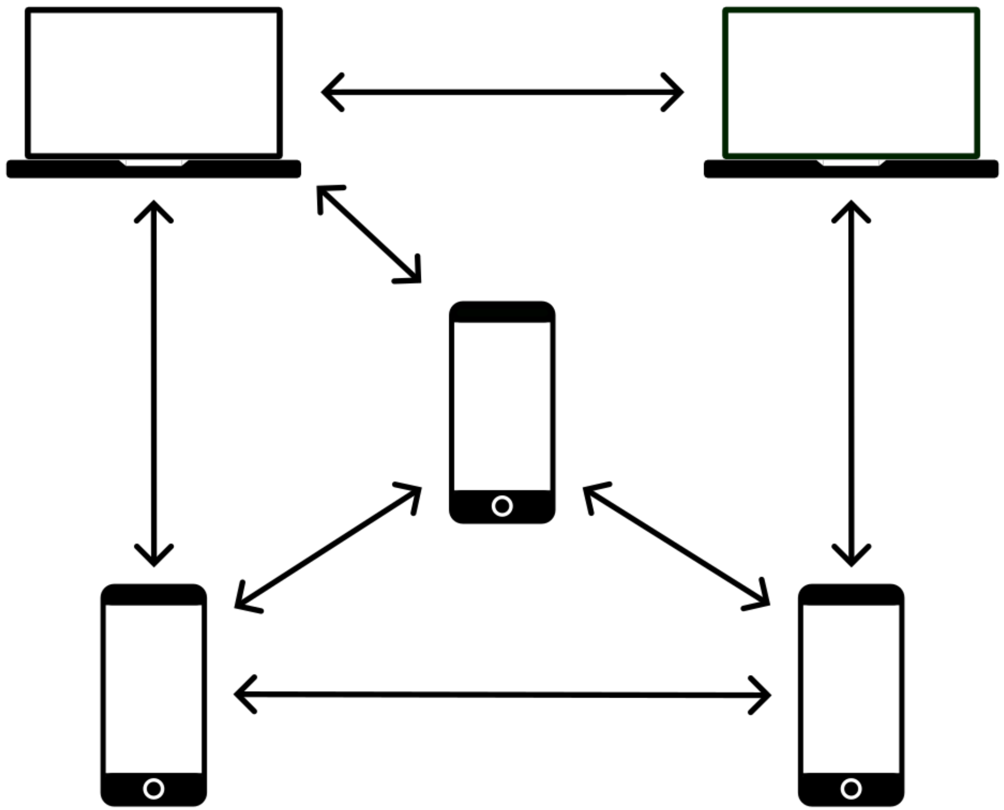
- Limited aggregation context



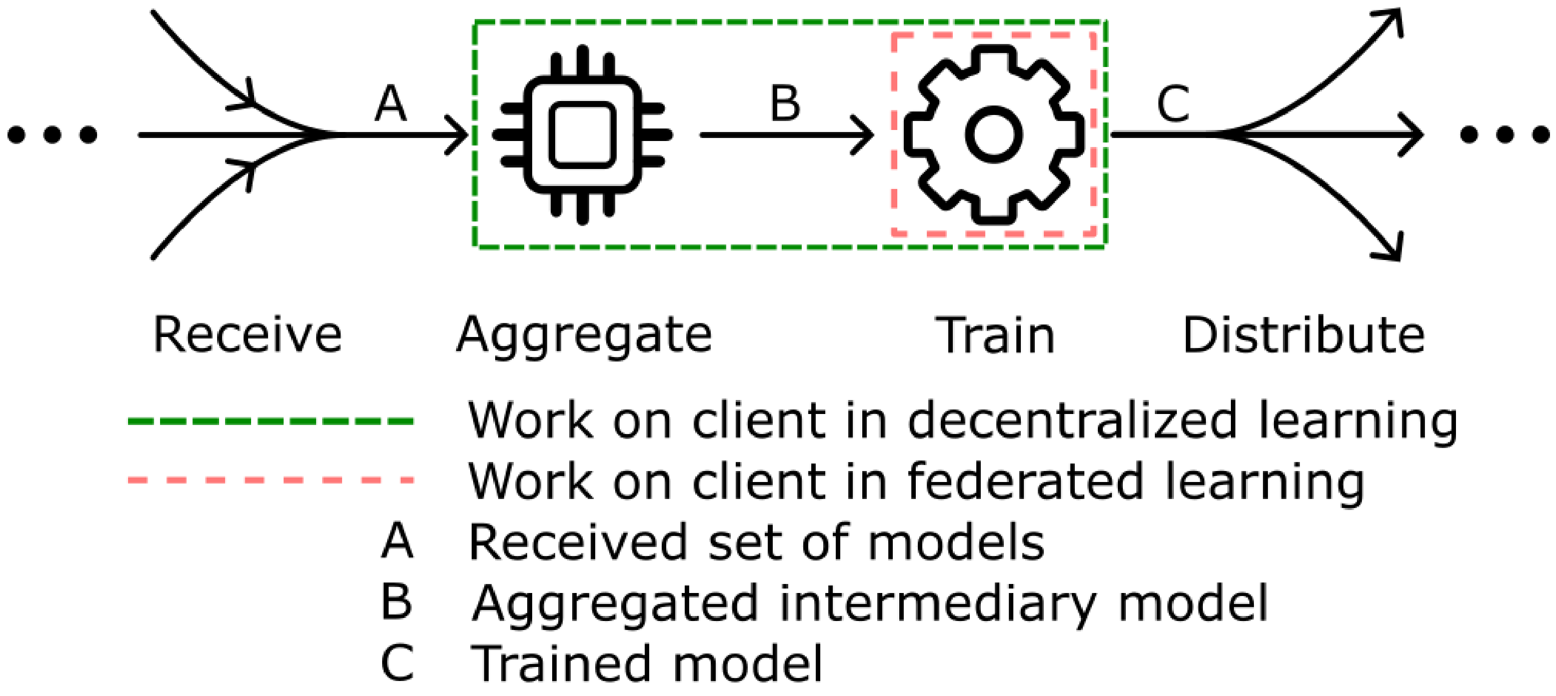
[1] I. Hegedus, G. Danner, and M. Jelasity, "Decentralized learning works: An empirical comparison of gossip learning and federated learning," Journal of Parallel and Distributed Computing, vol. 148, pp. 109–124, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731520303890

Decentralized learning training loop

1. Train on local data
2. Send to neighbors
3. Aggregate
4. Repeat



Federated learning vs decentralized learning



Poisoning attack

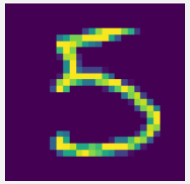
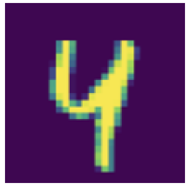
Targeted poisoning attack

- Label-flipping
- Backdoor

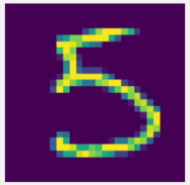
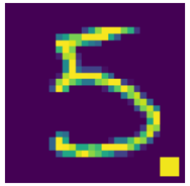
Untargeted poisoning attack

- A little is enough [1]
- Static optimization attack [2]

Label-flipping attack

Training sample	Label
	4
	5

Backdoor attack

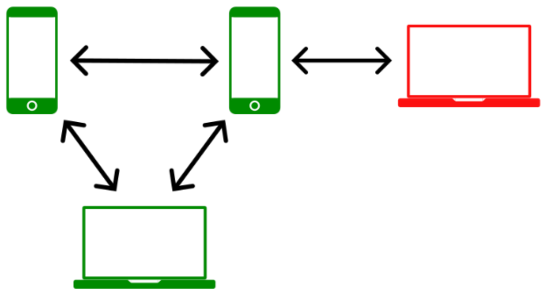
Training sample	Label
	5
	7

From [3]

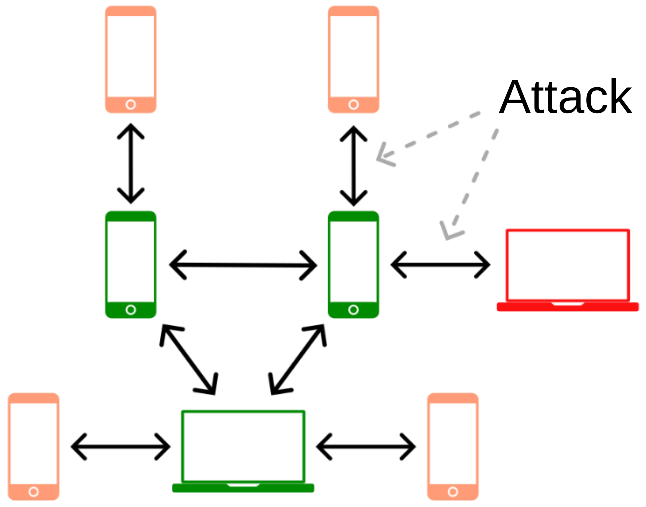
[1] G. Baruch, M. Baruch, and Y. Goldberg, “A Little Is Enough: Circumventing Defenses For Distributed Learning,” in Advances in Neural Information Processing Systems, 2019, vol. 32. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/ec1c59141046cd1866bbbcdfb6ae31d4-Paper.pdf
[2] M. Fang, X. Cao, J. Jia, and N. Z. Gong, “Local model poisoning attacks to byzantine-robust federated learning,” in Proceedings of the 29th USENIX Conference on Security Symposium, 2020, pp. 1623–1640.
[3] S. Udeshi, S. Peng, G. Woo, L. Loh, L. Rawshan, and S. Chattopadhyay, Model Agnostic Defence against Backdoor Attacks in Machine Learning. 2022.

Sybil attack

- Adversary creates fake identities (Sybils)
- Adversary increases its influence in the network
- Benign nodes cannot distinguish between benign and Sybil
- Amplifies poisoning attack



Single attacker



Sybil attack

Problem statement

- Federated learning does not scale
- Federated learning has a single point of failure
- Unstudied Sybil poisoning resilience of decentralized learning

- Contributions:
  - Demonstration of inscalability of federated learning
  - Effective adversarial strategy
  - SybilWall
  - Empirical evaluation

Introduction

Related work

SybilWall

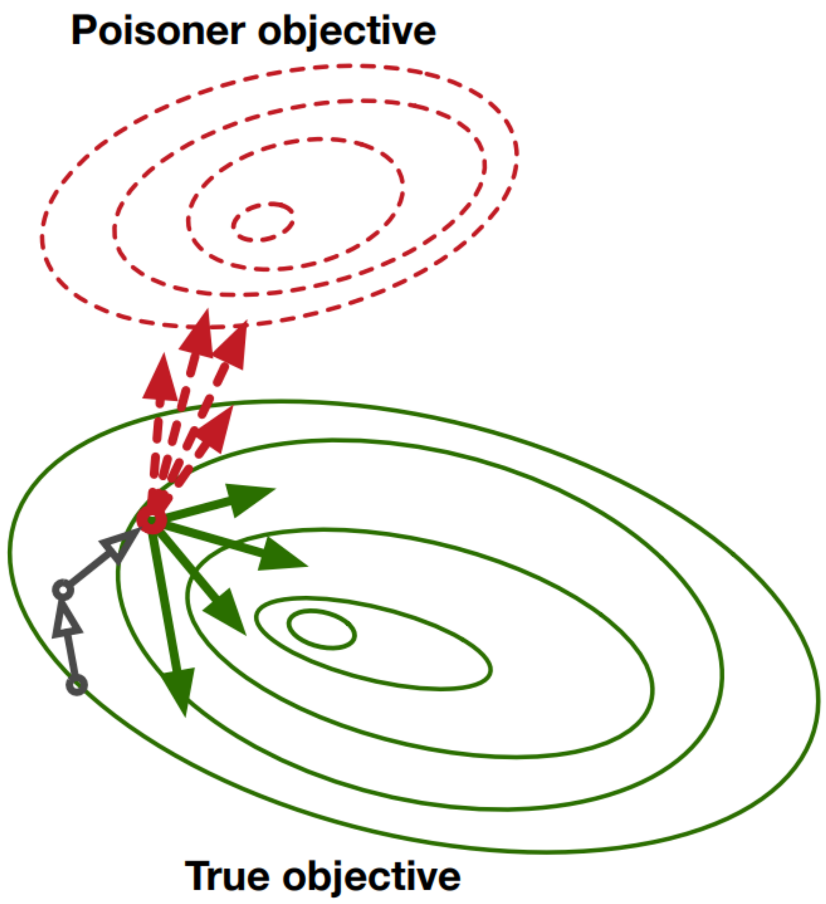
Evaluation

Conclusion

FoolsGold

- Primary inspiration for SybilWall
- Designed for federated learning

- High similarity between Sybils
- Low similarity between honest nodes
- Assign lower weight to similar models

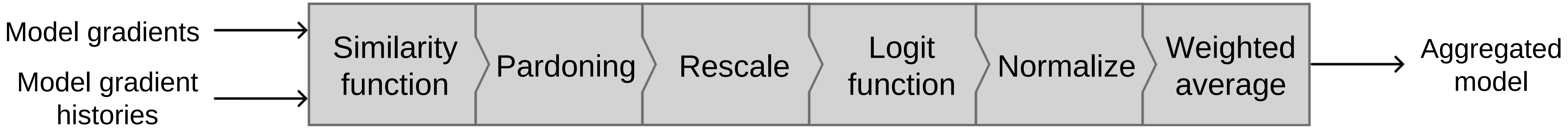


From [1]

[1] C. Fung, C. J. M. Yoon, and I. Beschastnikh, “Mitigating Sybils in Federated Learning Poisoning,” CoRR, vol. abs/1808.04866, 2018, [Online]. Available: http://arxiv.org/abs/1808.04866

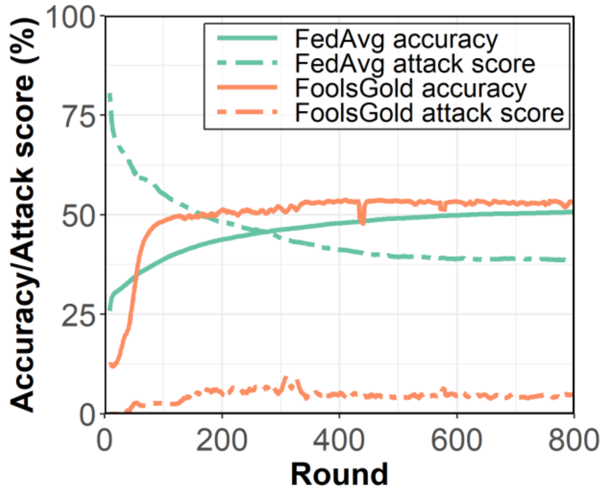
FoolsGold

- Input for aggregation in round T for every node $i \in N$:
  - Model gradient: Δw_i^T
  - Model gradient history: $\sum_{t=0}^{T} \Delta w_i^t$

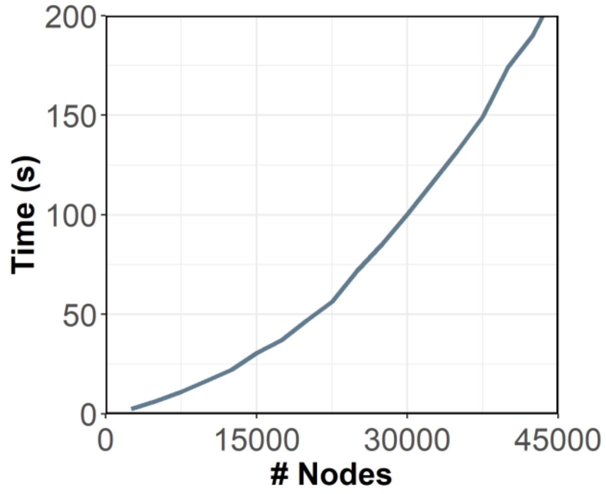


FoolsGold's aggregation function

FoolsGold

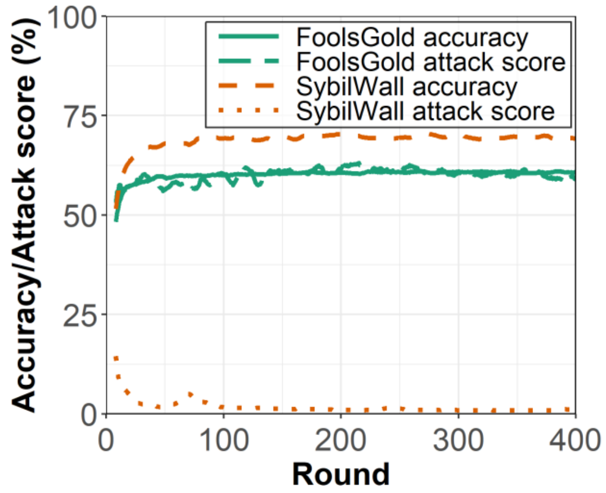


FoolsGold compared to FedAvg

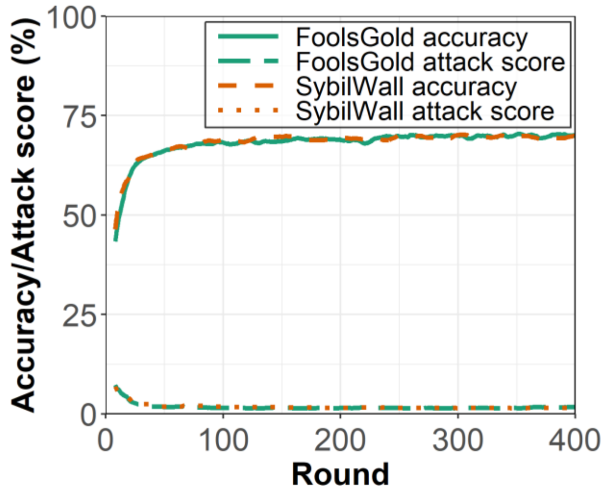


Aggregation time against number of nodes

Federated learning



Network topology 1:
SybilWall compared to FoolsGold



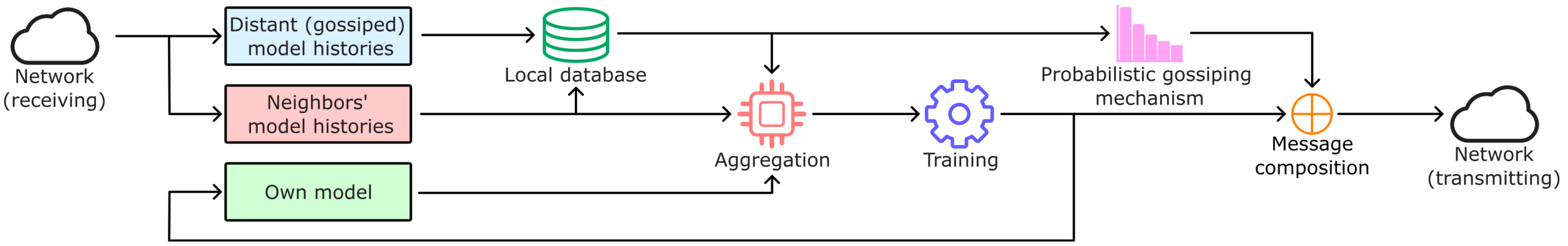
Network topology 2:
SybilWall compared to FoolsGold

Decentralized learning

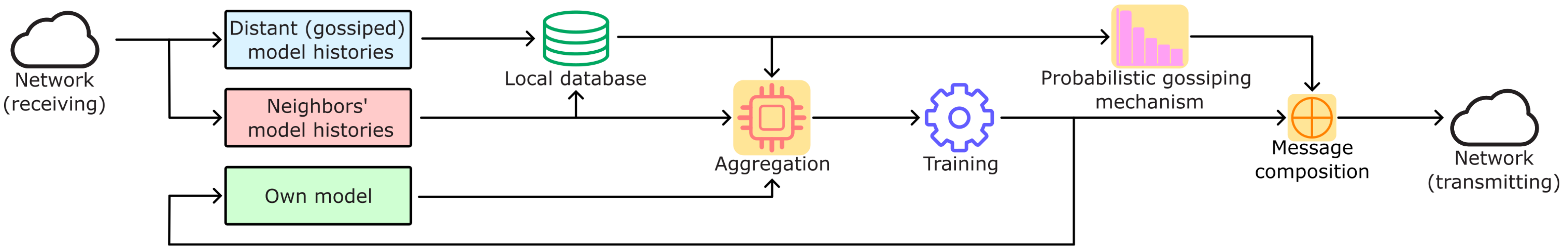
TU Delft

TU Delft

SybilWall architecture

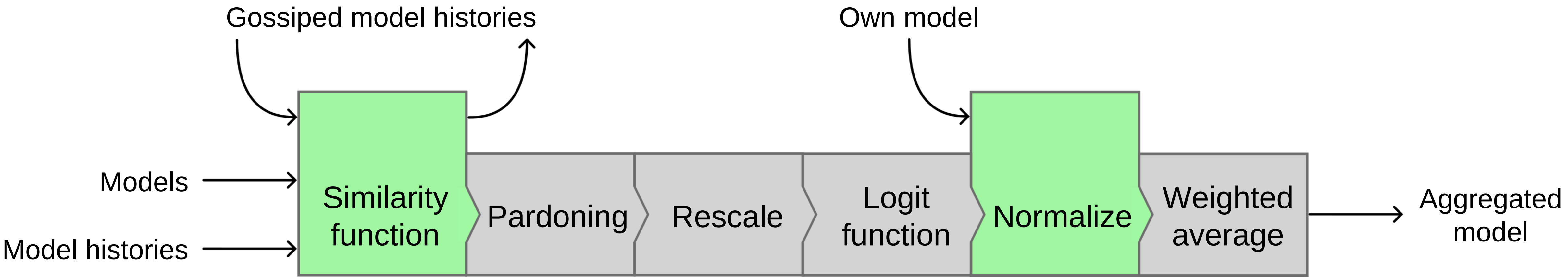


SybilWall architecture



1. Aggregation function

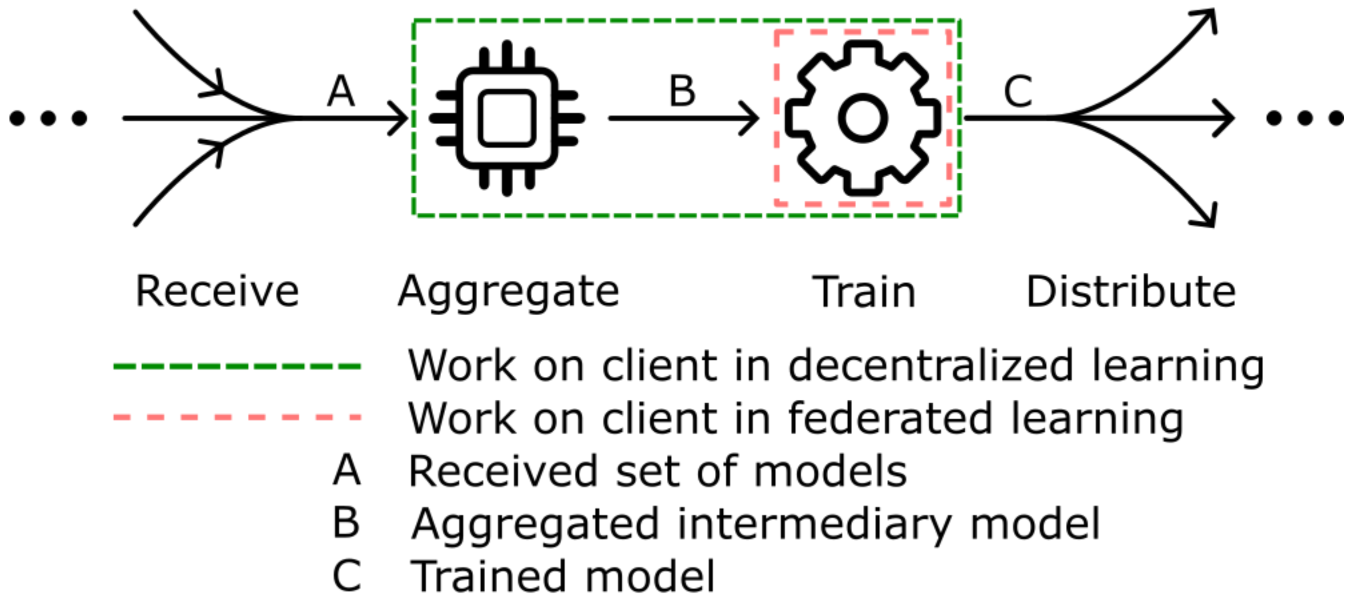
- FoolsGold-inspired
- 2 improvements:
  - Support for gossiped model histories
  - Nodes trust themselves



SybilWall's aggregation function

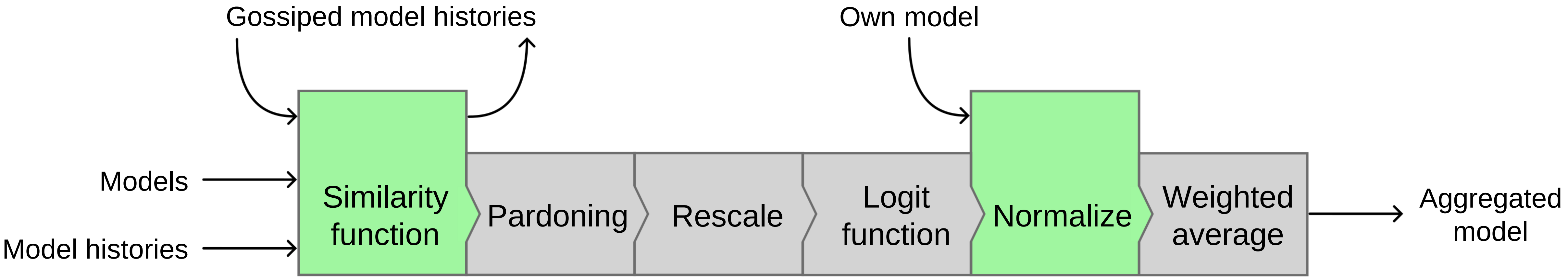
1. Aggregation function

- Uses model history rather than model gradient history



1. Aggregation function

- Input for aggregation in round T for every neighbouring node $i \in N$:
  - Model: w_i^T
  - Model history: $\sum_{t=0}^{T} w_i^t$

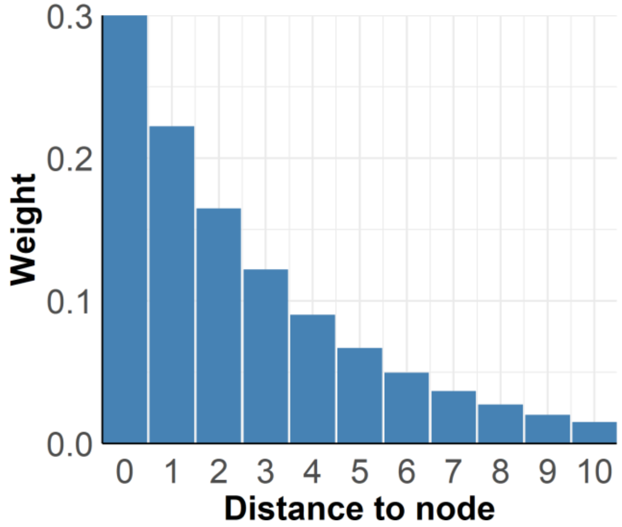


SybilWall's aggregation function

2. Probabilistic gossiping mechanism

- In each round, every node transmits:
  - Its own trained model
  - A probabilistically selected model history from its local database (gossip)
- The gossiped model is selected using a weighted random selection
  - The weights correspond to the exponential distribution, where the distance to the originating node serves as the parameter d

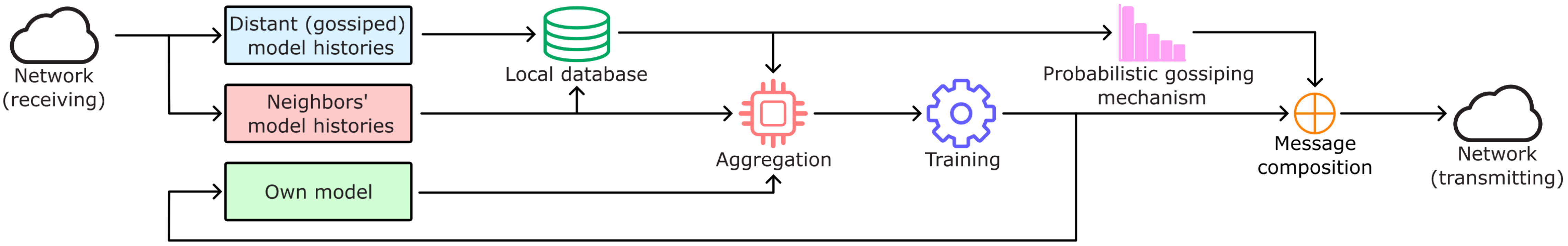
$$P(d) = \lambda e^{-\lambda d}$$



3. Message composition

- Omit trained model, as it can be inferred from subsequent model histories
- Messages are composed of:
  - h_i: model history of sender i
  - g_k: gossiped model history of distant node k
  - r_i: round number from which model history h_i originates
  - r_k: round number from which gossiped model history g_k originates
- Each message component is signed by the corresponding node
- Downtime and unreachability support

SybilWall



Network
(receiving)
Distant (gossiped)
model histories
Neighbors'
model histories
Own model
Local database
Aggregation
Training
Probabilistic gossiping
mechanism
Message
composition
Network
(transmitting)

Introduction
Related work
SybilWall

Evaluation

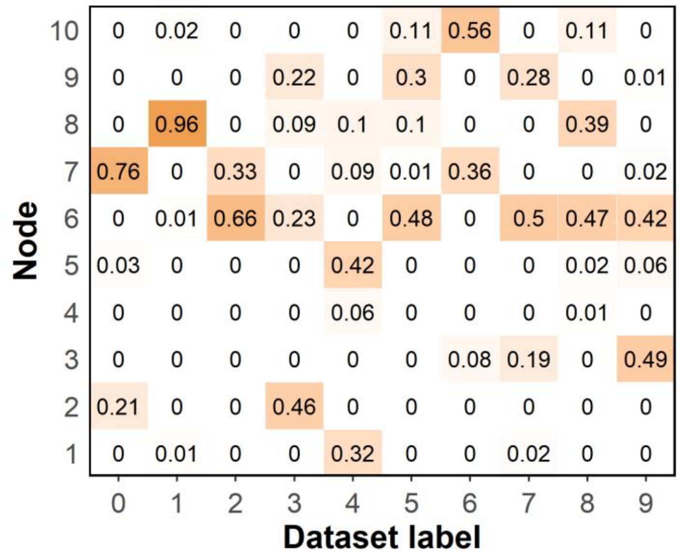
Conclusion

Experimental setup

- Python-based IPv8 implementation
- 100 nodes simulation on DAS-6
- 4 datasets
- Dirichlet-based data distribution

Dataset	Model	Learning rate
MNIST	Single soft-max layer	$\eta = 0.01$
FashionMNIST	Single soft-max layer	$\eta = 0.01$
SVHN	LeNet-5	$\eta = 0.004$
CIFAR-10	LeNet-5	$\eta = 0.004$

Evaluated datasets



Example Dirichlet distribution

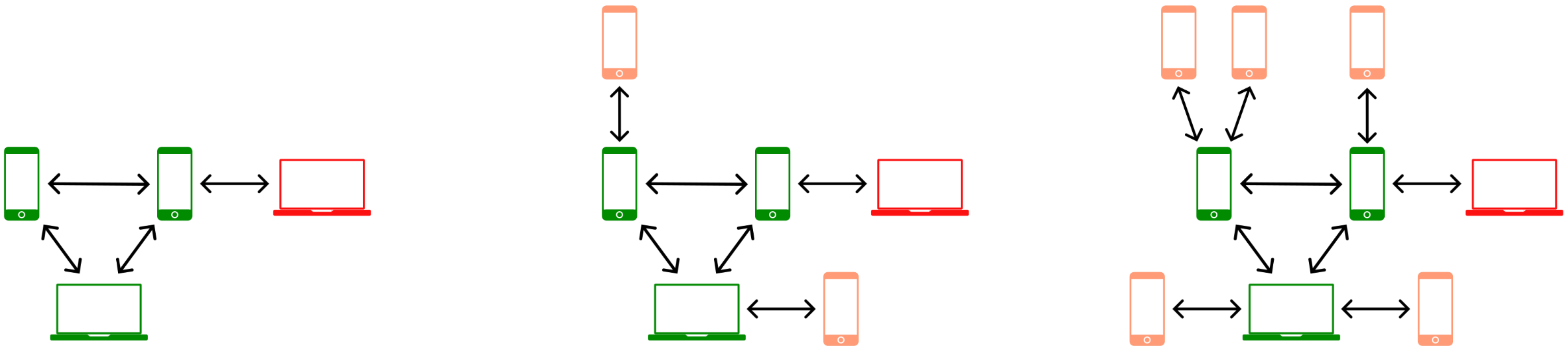
TU Delft

Experimental setup

- Network topology
  - Random geometric graphs
- Evaluation metrics
  - Accuracy: percentage of correctly classified samples of the original dataset
  - Attack score: percentage of correctly classified samples of the maliciously altered segment of the dataset

SSP Attack

- Adversarial strategy
- Average attack edge density ϕ



$\phi = 1/3$ $\phi = 1$ $\phi = 2$

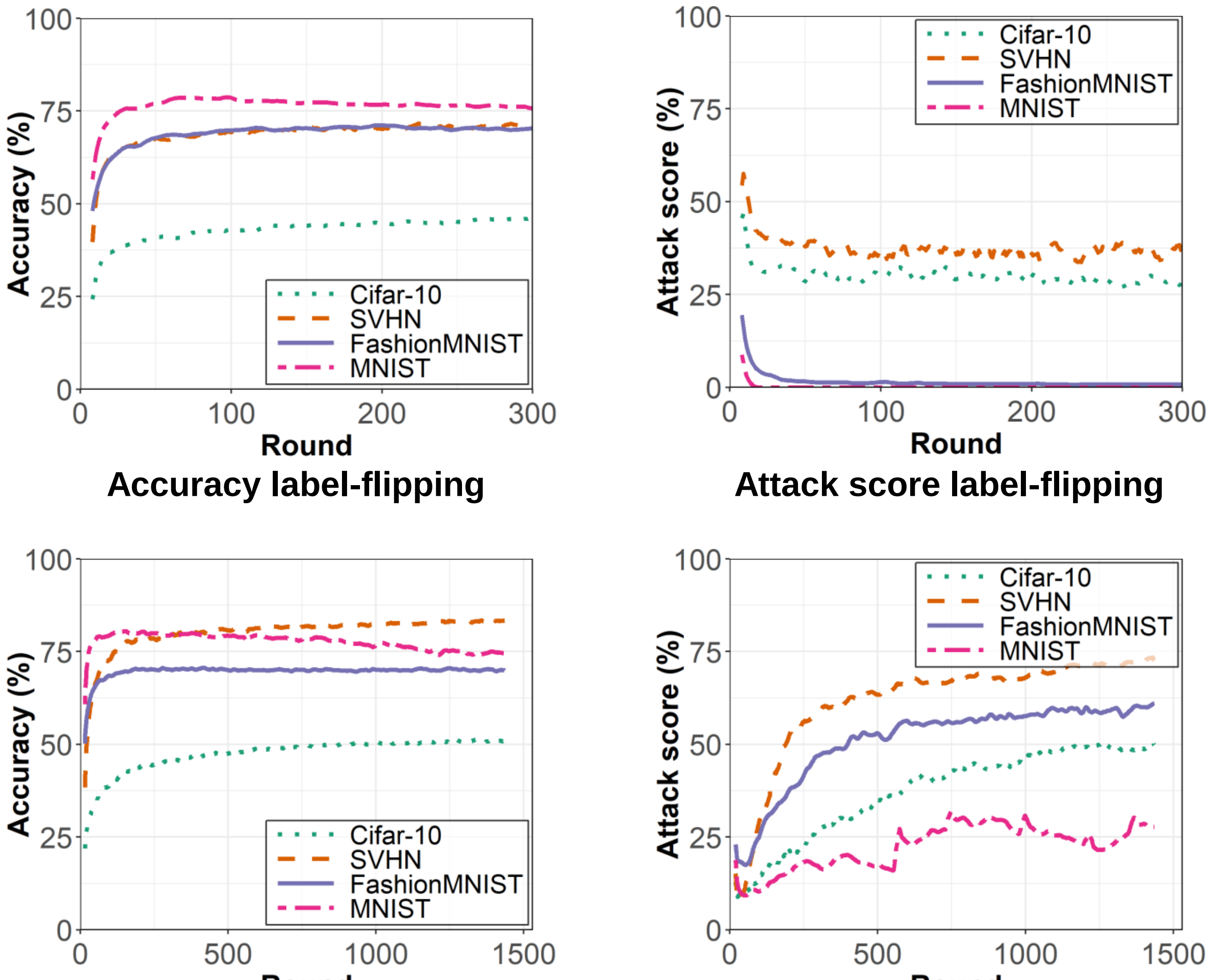
Effect of dataset

We evaluated SybilWall on numerous datasets:

- MNIST
- FashionMNIST
- SVHN
- CIFAR-10

Attack edge density: $\phi = 1$

Results



Accuracy label-flipping

Attack score label-flipping

Accuracy backdoor

Attack score backdoor

Comparison with existing techniques (1/2)

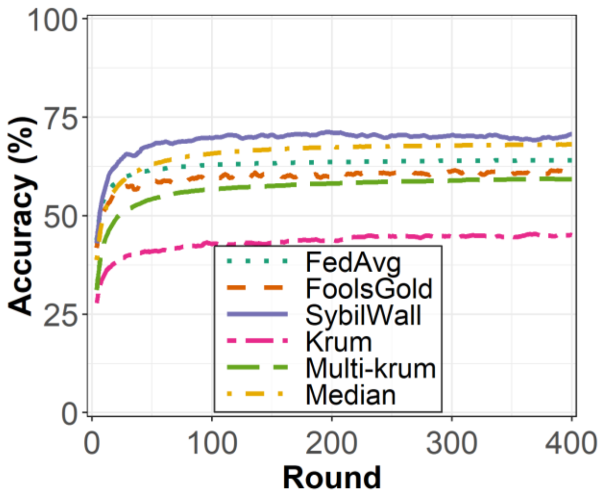
We compare SybilWall with existing techniques:

- FedAvg
- FoolsGold
- Krum
- Multi-Krum
- Median

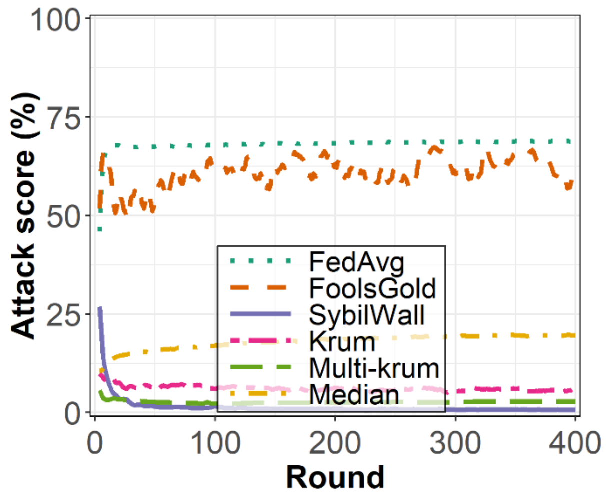
Dataset: FashionMNIST

Attack edge density: $\phi \in \{1, 4\}$

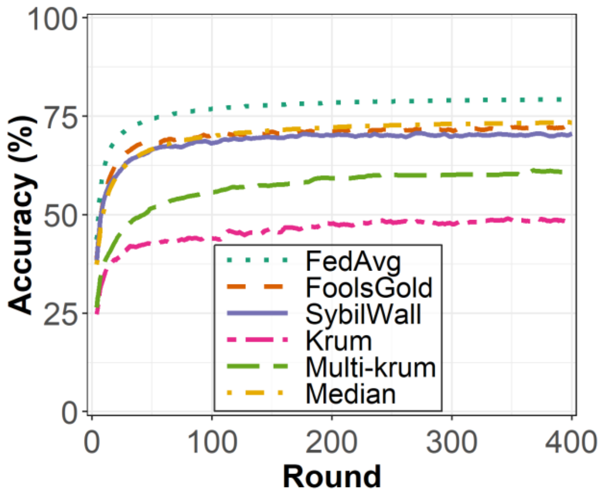
Results: $\phi = 1$



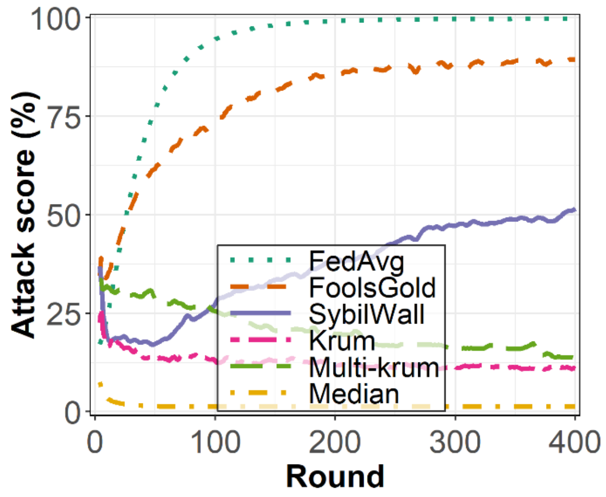
Accuracy label-flipping



Attack score label-flipping



Accuracy backdoor



Attack score backdoor

TU Delft

Comparison with existing techniques (2/2)

We compare SybilWall with existing techniques:

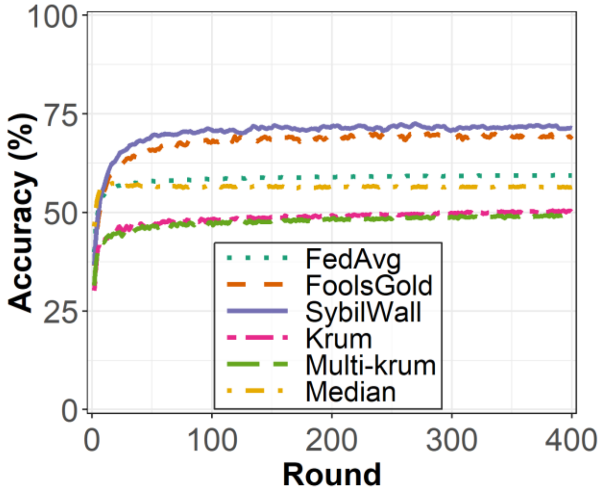
- FedAvg
- FoolsGold
- Krum
- Multi-Krum
- Median

Dataset: FashionMNIST

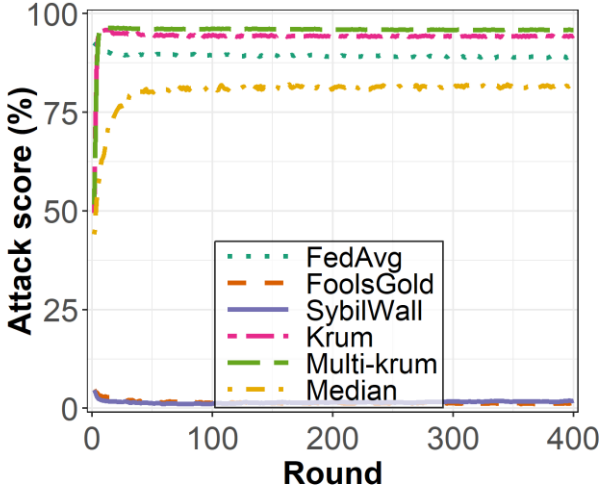
Attack edge density: $\phi \in \{1, 4\}$

TUDelft

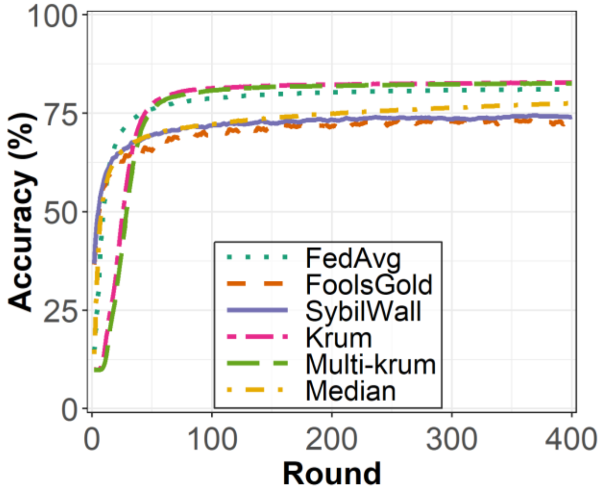
Results: $\phi = 4$



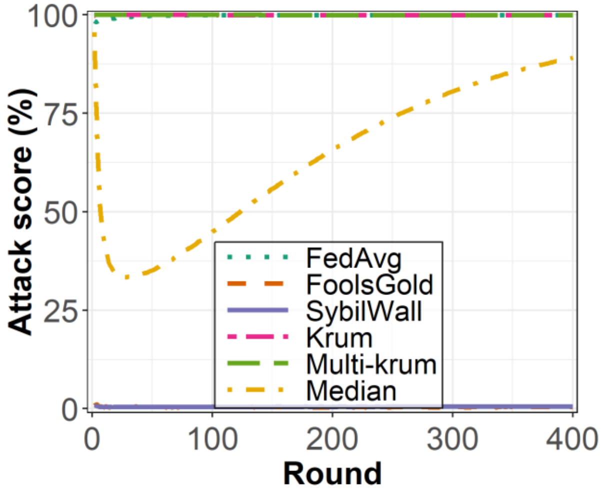
Accuracy label-flipping



Attack score label-flipping



Accuracy backdoor



Attack score backdoor

Effect of attack edge density

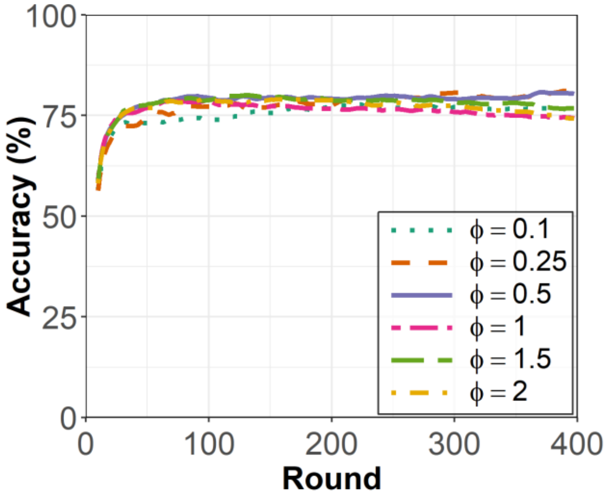
We evaluate SybilWall's defensive capabilities against various attack edge densities:

- $\phi = 0.1$
- $\phi = 0.25$
- $\phi = 0.5$
- $\phi = 1$
- $\phi = 1.5$
- $\phi = 2$

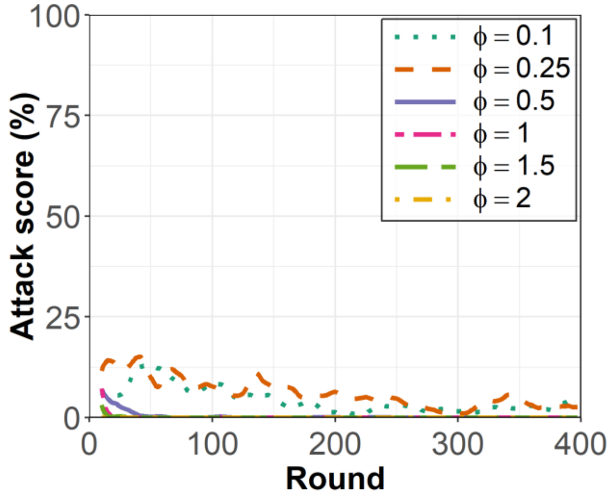
Dataset: MNIST

TUDelft

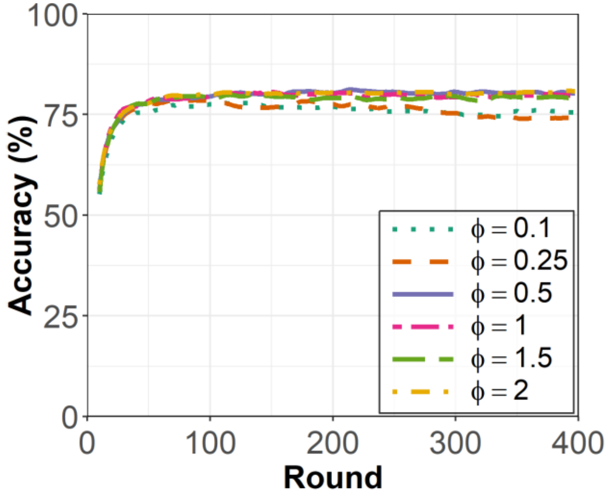
Results



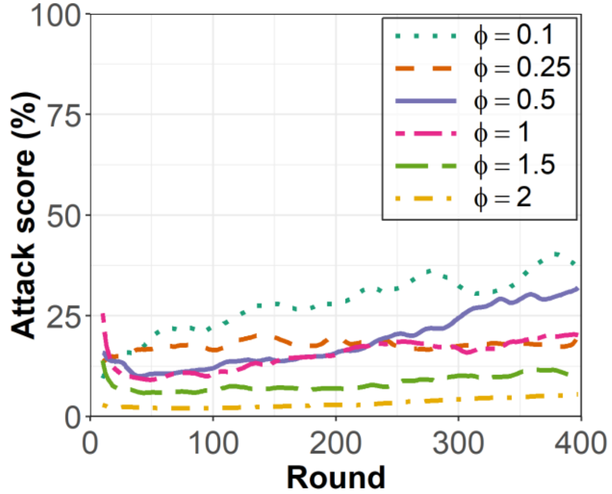
Accuracy label-flipping



Attack score label-flipping



Accuracy backdoor



Attack score backdoor

Effect of data distribution

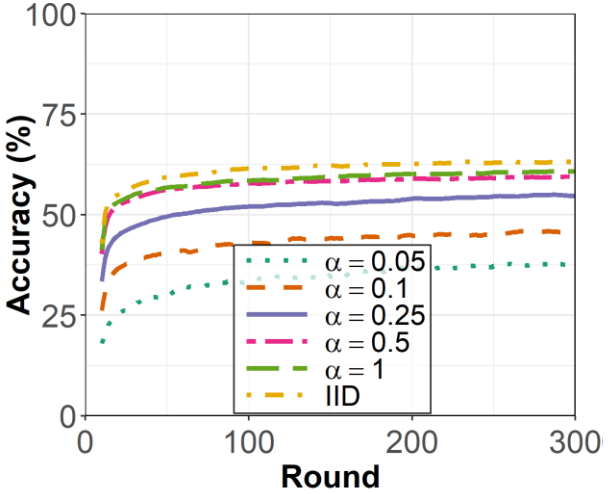
We evaluate SybilWall's peformance on numerous data distributions:

- $\alpha = 0.1$
- $\alpha = 0.25$
- $\alpha = 0.5$
- $\alpha = 1$
- $\alpha = 1.5$
- IID

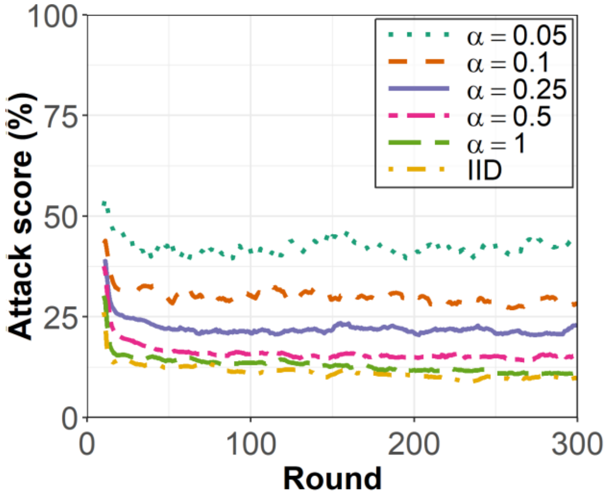
Dataset: CIFAR-10

Attack edge density: $\phi = 1$

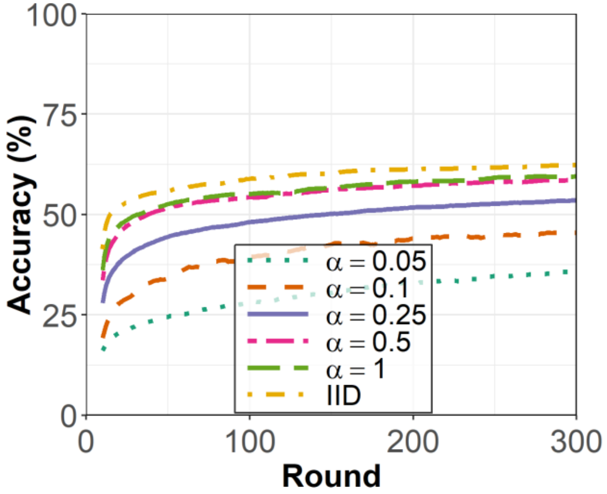
Results



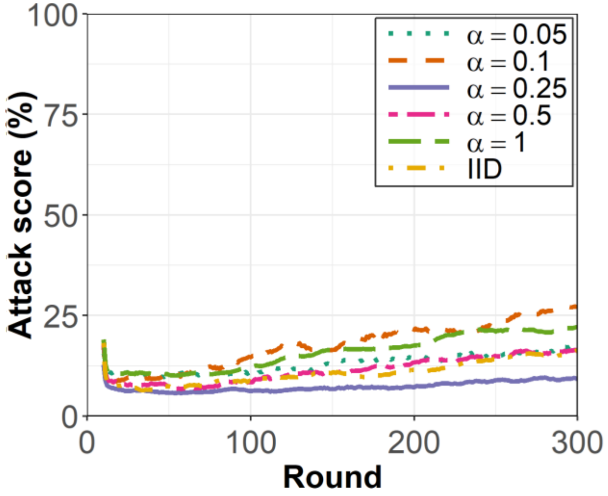
Accuracy label-flipping



Attack score label-flipping



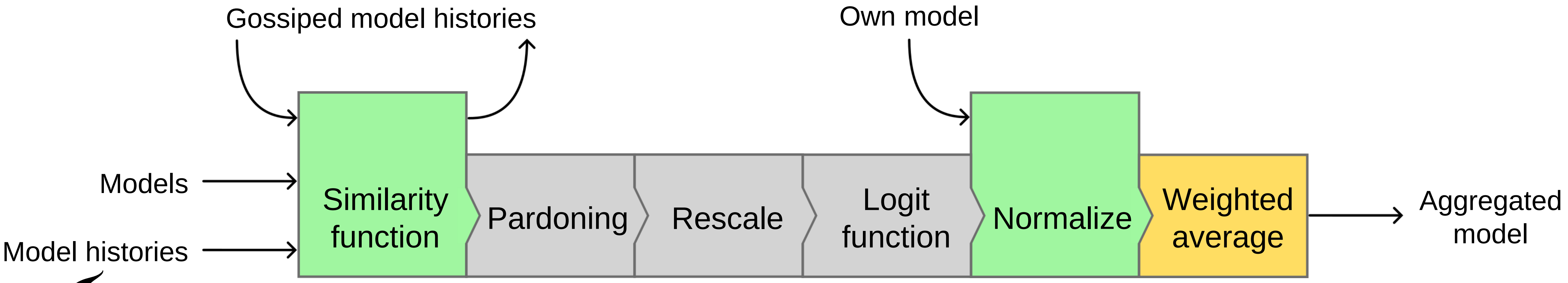
Accuracy backdoor



Attack score backdoor

Further enhancing SybilWall

- SybilWall does not fully mitigate backdoor attacks for low values of ϕ
- We further enhance SybilWall by replacing the weighted average with:
  - Weighted median
  - Median
  - Krum-based filter



SybilWall's aggregation function

Further enhancing SybilWall

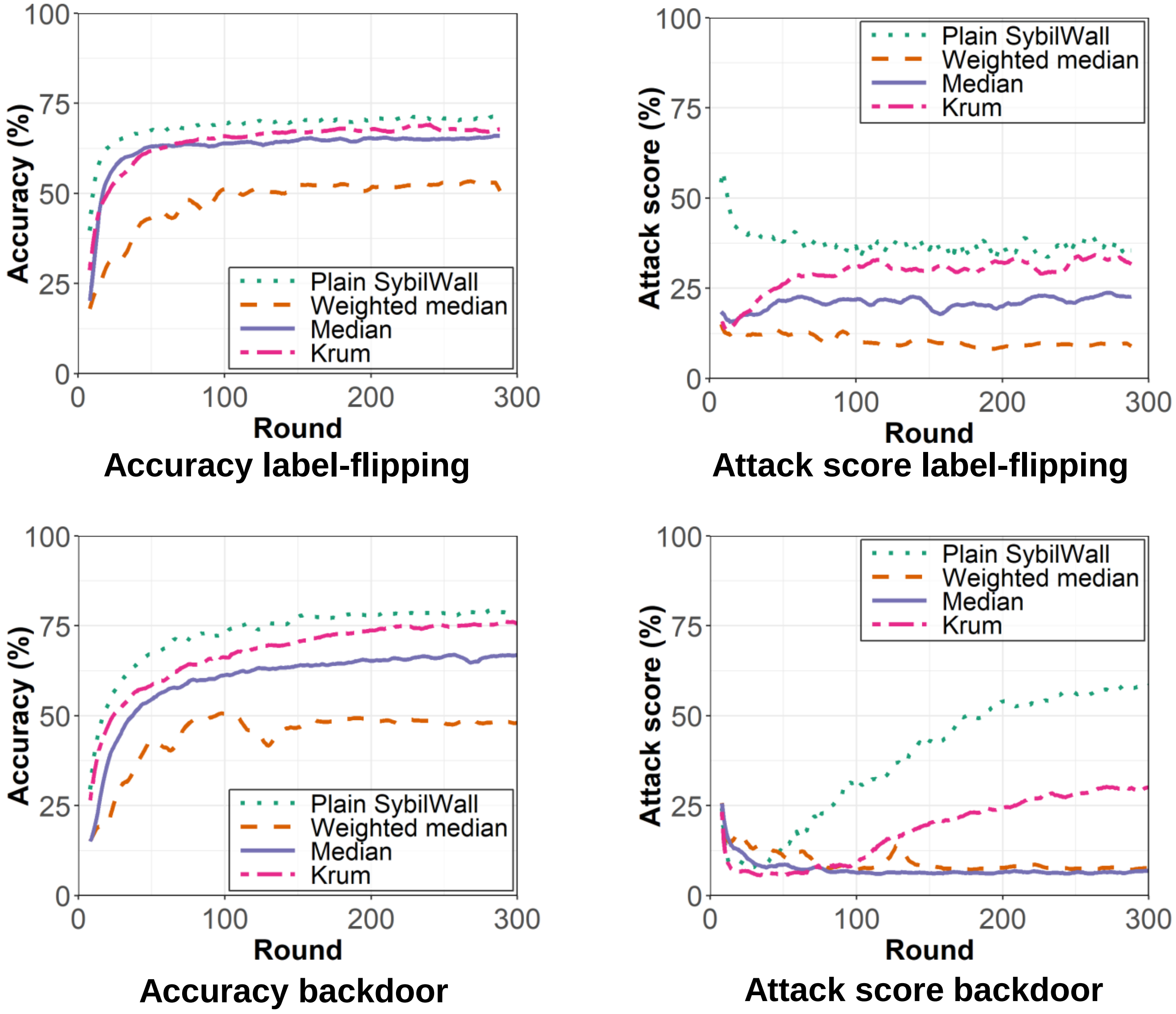
We evaluate possible enhancements of SybilWall:

- Weighted median
- Median
- Krum-based filter

Dataset: SVHN

Attack edge density: $\phi = 1$

Results



Introduction
Related work
SybilWall
Evaluation

Conclusion

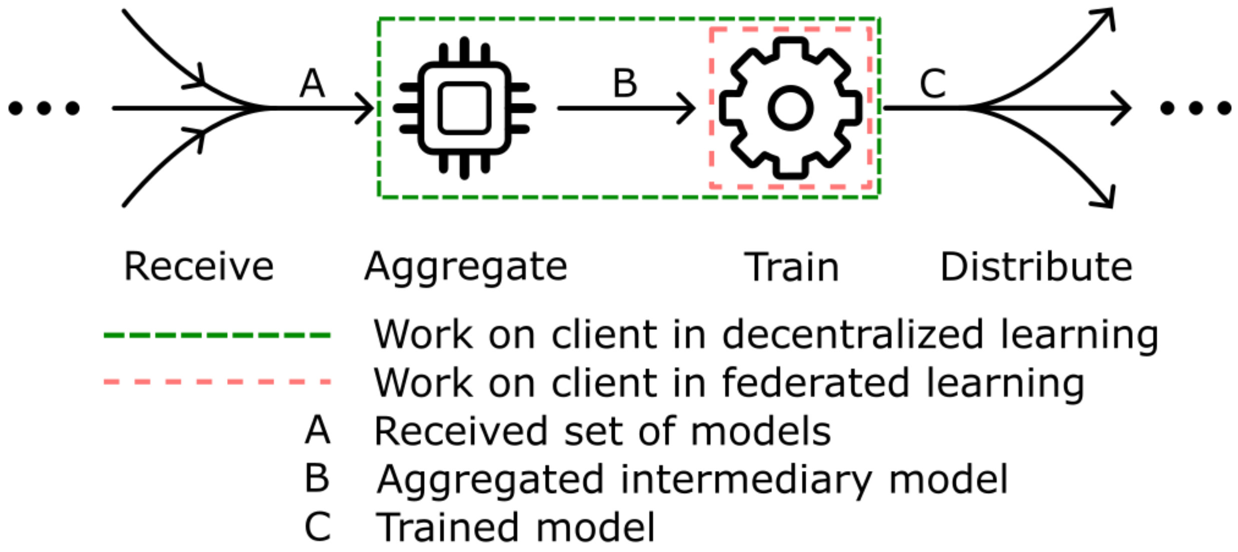
TU Delft

Conclusion

- SybilWall
  - Aggregation function
  - Probabilistic gossiping mechanism
- Satisfactory performance on 4 datasets
- Stronger Sybil resilience over other defensive algorithms
  - Mitigates the label-flipping attack
  - Slows down the backdoor attack

Future work

- Further enhancement of SybilWall
- Filtering for relevant weights during aggregation
- Improving SybilWall's resilience against backdoor attacks
  - e.g. employing gradient history rather than model history



Thank you for your attention

TU Delft

Sources

- Image on cover from https://www.bing.com/images/create powered by https://openai.com/dall-e-2
- News article “Accenture Makes a $3 Billion Bet on A.I.” from https://www.nytimes.com/2023/06/13/business/dealbook/accenture-ai-billion-consulting.html
- News article “Germany Could Block ChatGPT if Needed, Says Data Protection Chief” from https://www.voanews.com/a/germany-could-block-chatgpt-if-needed-says-data-protection-chief-/7034099.html
- News article “ChatGPT: Are Europeans afraid that Generative AI will take away their jobs?” from https://www.euronews.com/next/2023/06/13/chatgpt-are-europeans-afraid-that-generative-ai-will-take-away-their-jobs
- News article “AI Unlocks Mysteries of Brain Fluid Flow: A Leap Forward in Alzheimer’s Research” from https://neurosciencenews.com/ai-alzheimers-brain-fluid-23462/
- News article “ChatGPT banned in Italy over privacy concerns” from https://www.bbc.com/news/technology-65139406