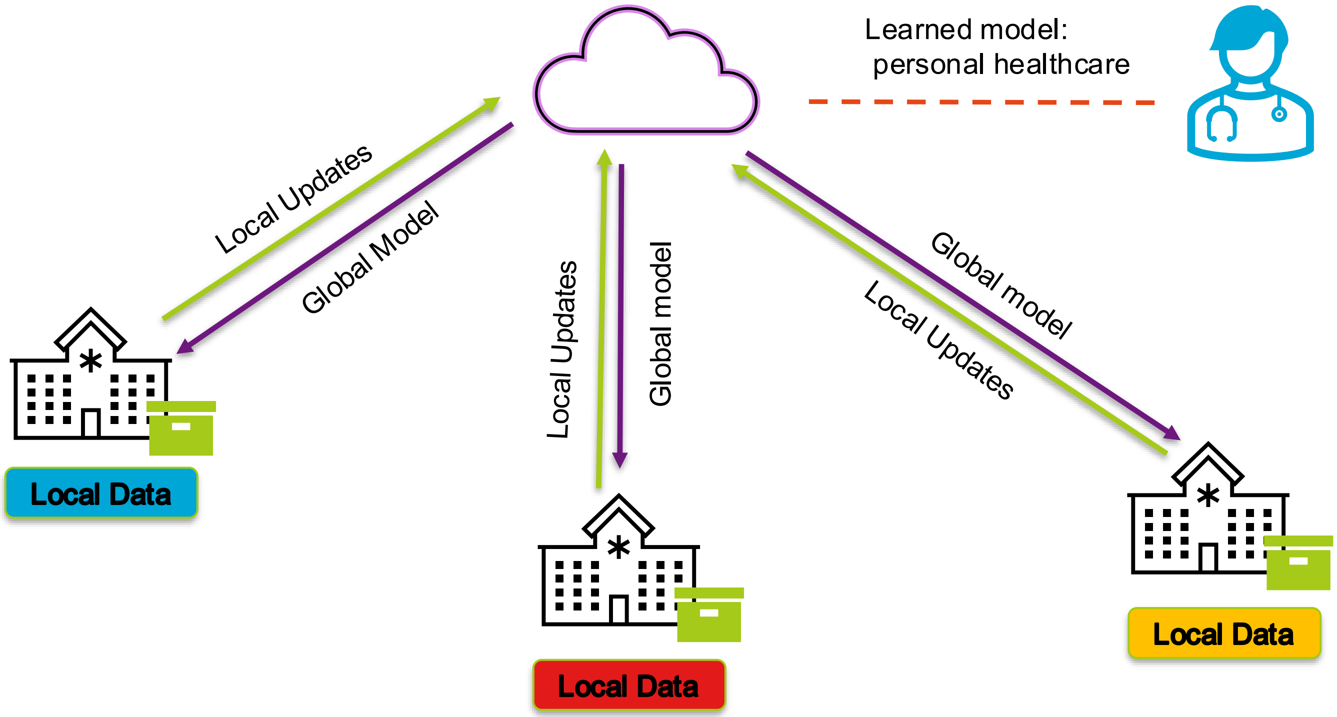
MUDGUARD: Taming Malicious Majorities in Federated Learning using Privacy-Preserving Byzantine-Robust Clustering

Rui Wang, Xingkai Wang, Huanhuan Chen, Jérémie Decouchant, Stjepan Picek, Nikolaos Laoutaris, Kaitai Liang

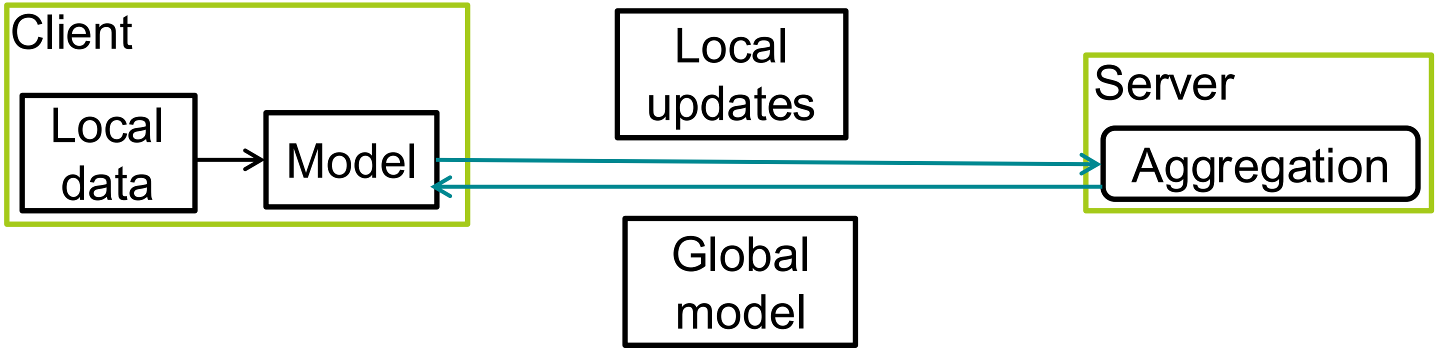
SiMLA 2025

What is Federated Learning?

- Federated learning (FL)

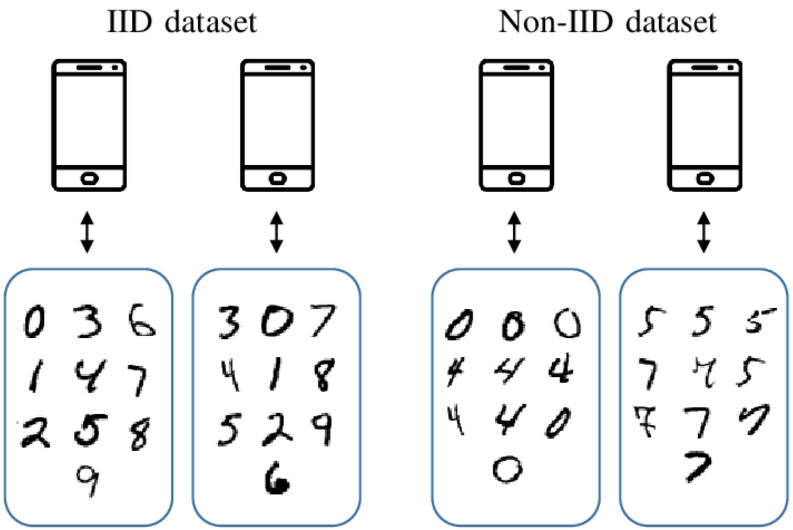


Federated Learning



FL vs. Distributed Learning

- Non – independent and identically distributed dataset



FL vs. Distributed Learning

- FL contains much more clients
- FL training is more complicated: Edge devices like mobile phone could offline at any time

A Categorization of FL

- Horizontal Federated Learning

  same feature space but differ in samples [1]
- Vertical Federated Learning

  same sample ID space but differ in feature space [1]
- Federated Transfer Learning

  data sets differ not only in samples but also in feature space [1]

[1] Yang, Q., Liu, Y., ... Tong, Y., 2019. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology 10. doi:10.1145/3298981

Attacks on FL from semi-honest adversaries

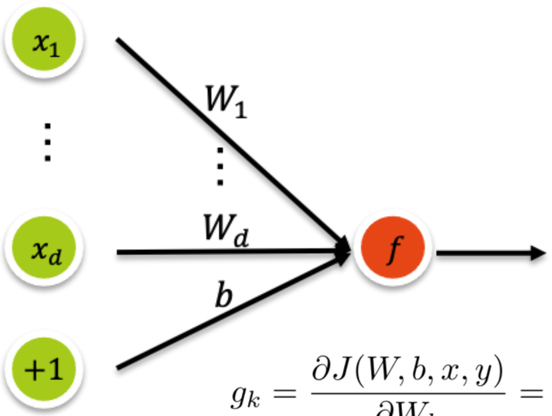
- Semi-honest adversaries: Strictly follow the algorithms but try to infer private information from received messages.
  - Data reconstruction attack [1][2]
  - Membership inference attack[3]
  - Property inference attack
  - …

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." *Advances in neural information processing systems* 32 (2019).
[2] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai. Privacypreserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5):1333–1345, May 2018.
[3] Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017.

Attacks on FL targeting on privacy

Loss-function/ReLu exploitation [2]

First dense layer attack [3]



$$h(x) = f\left(\sum_{i=1}^{d} W_i x_i + b\right)$$

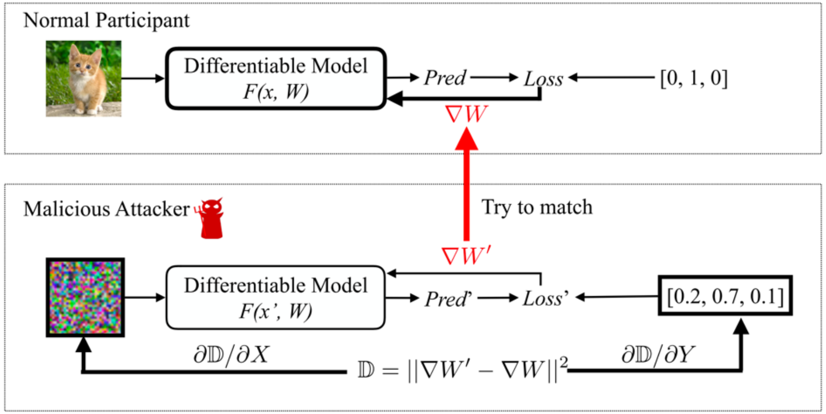
$$g_k = \frac{\partial J(W,b,x,y)}{\partial W_k} = 2(h(x)-y)f'\left(\sum_{i=1}^{d} W_i x_i + b\right) \cdot x_k$$

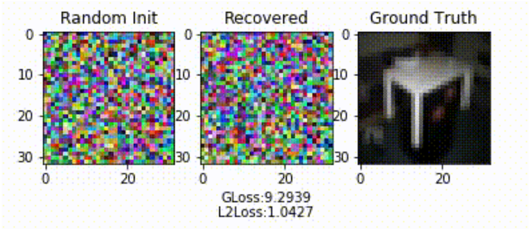
$$g = \frac{\partial J(W,b,x,y)}{\partial b} = 2(h(x)-y)f'\left(\sum_{i=1}^{d} W_i x_i + b\right) \cdot 1$$

$$\frac{g_k}{g} = x_k$$

[2] Akiyoshi Sannai. Reconstruction of training samples from loss functions. arXiv:1805.07337, 2018.
[3] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai. Privacypreserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5):1333–1345, May 2018.

Attacks on FL from semi-honest adversaries





[1] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." *Advances in neural information processing systems* 32 (2019).

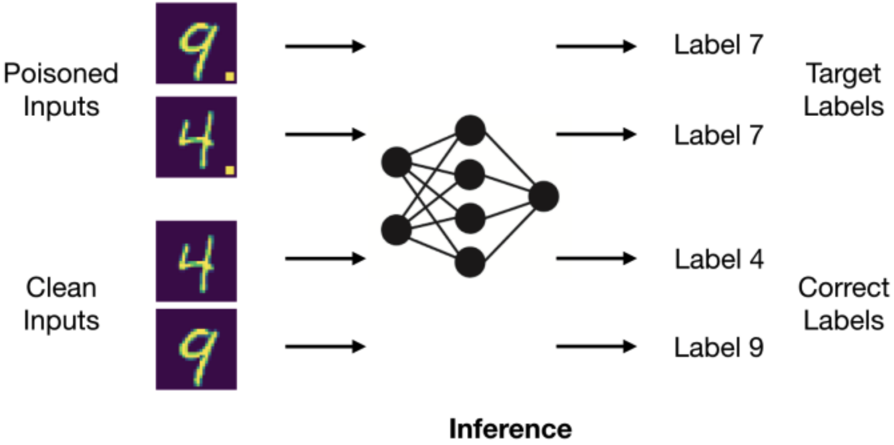
Attacks on FL from malicious adversaries

- Malicous: not only violate privacy, but also arbitrarily deviate from algorithm
  - Untargeted attacks: deteriorate the performance of global model
    a) Gaussian attack[4]
    b) Label flipping[5]
    c) …

[4] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to Byzantine-Robust federated learning. In USENIX Security, pages 1605–1622, 2020
[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In ICML, pages 1467–1474, 2012

Attacks on FL from malicious adversaries

- Targeted attacks (backdoor attacks): behave normally on all inputs except for specific attacker-chosen inputs[6]



[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in AISTATS, 2020.

Resist malicious clients

- Krum: selects one of the n local updates as the global model updates based on smallest square-distance score.

$$s_i = \sum_{\boldsymbol{g}_j \in \Gamma_{i,n-f-2}} \|\boldsymbol{g}_j - \boldsymbol{g}_i\|_2^2$$

- Median: instead of using mean value of FedAvg (original FL), it considers the median value of each parameter.

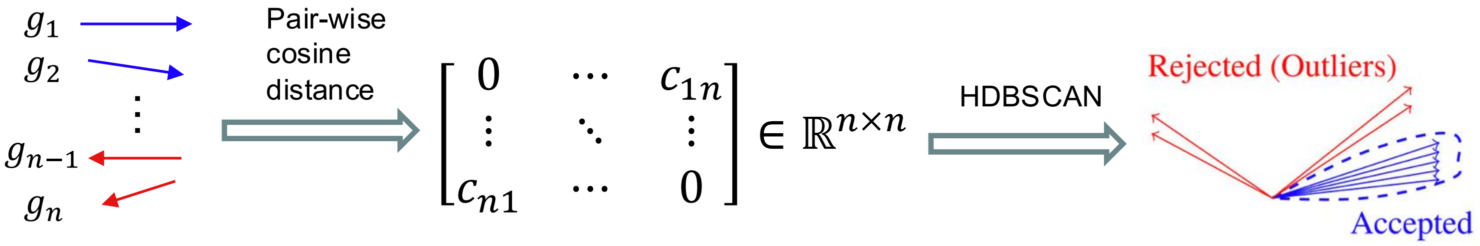
Resist malicious clients

- Trimmed Mean: for each model parameter, server removes the largest k and the smallest k values, and then computes the mean of the remaining $n - 2k$ values as global updates.

- Weakness
  - Honest majority setting

Resist malicious adversaries

- FLAME[11]:

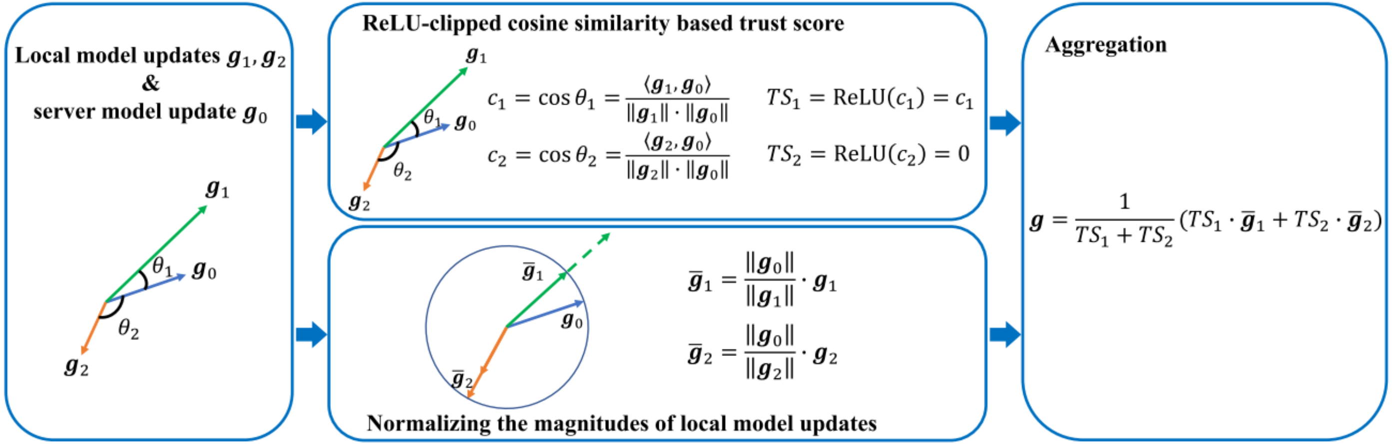


- Weakness:
  - Honest majority setting
  - Not works in non-iid

[11] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen M¨ollering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. Flame: Taming backdoors in federated learning. In USENIX Security, 2022

Resist malicious adversaries

- Fltrust[10]:



- Weakness:
  - Require an auxiliary dataset

[10] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In NDSS, 2021.

Goals

- Defense against malicious majority of clients
  - Without the root of trust dataset
  - Without the assumption of honest majority
  - With Privacy preservation

- Solution: Design a detector
  - No judgment criteria

Recap

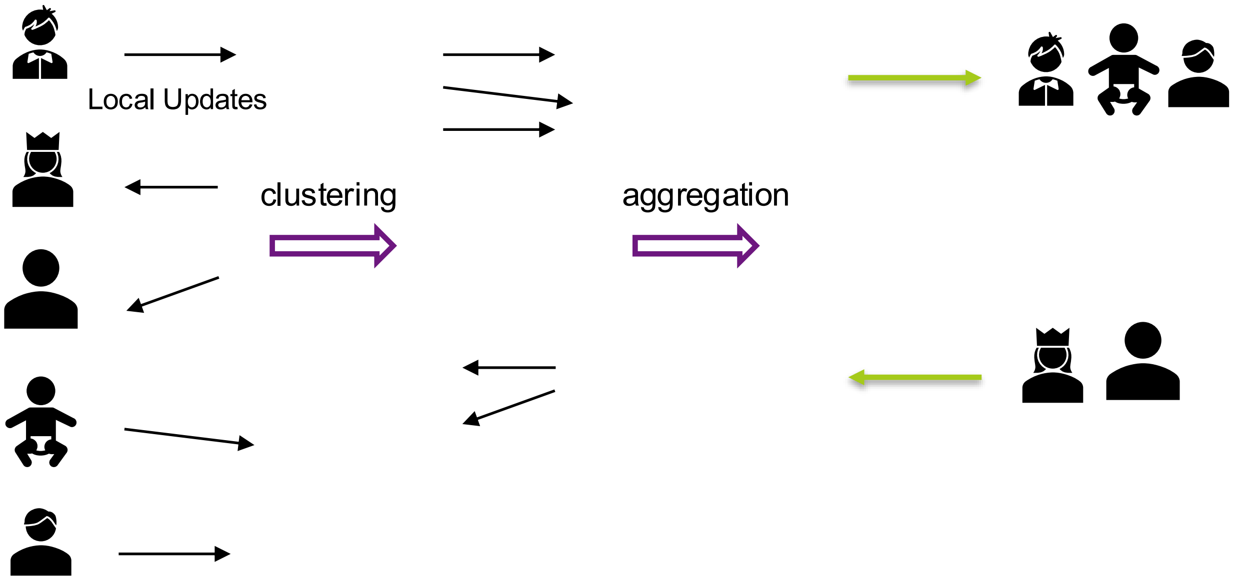
- Byzantine attacks aim to manipulate the FL model training process and degrade the model performance.

of honest clients

of malicious clients

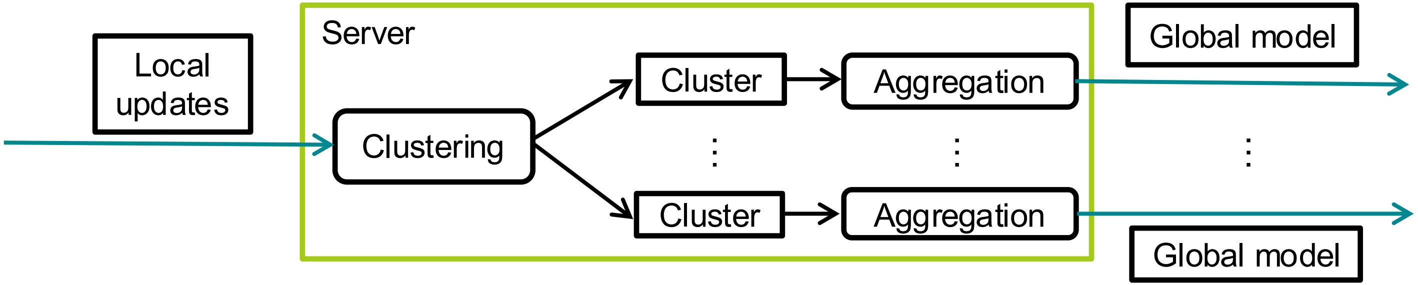
Solution: robustness

- Server's perspective



Solution: robustness

- Model segmentation

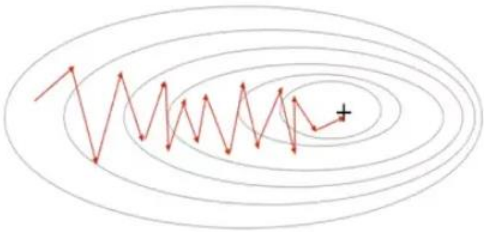


Solution: robustness

- Goal
  - Design a detector -> do not mix clients with different behavior (no need criteria)
- Chanllenge
  - Non-iid

Solution: robustness

- feature extraction with adjusted cosine similarity



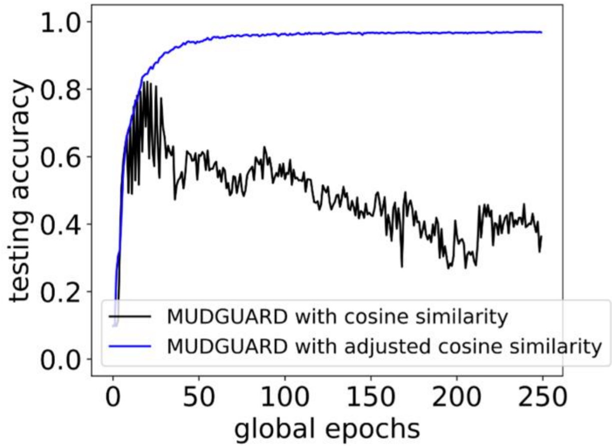
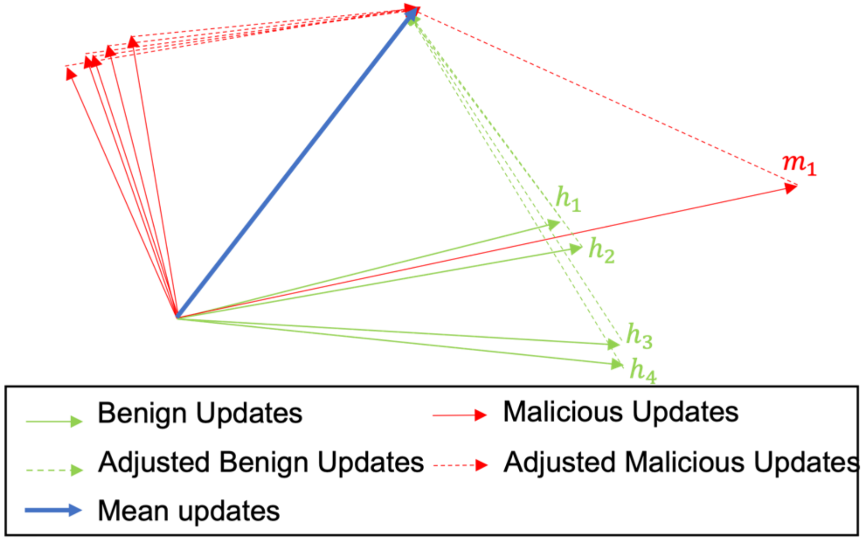


Fig. 12: Comparison of MUDGUARD with cosine similarity and adjusted cosine similarity under GA.

Solution: robustness

- feature extraction with adjusted cosine similarity



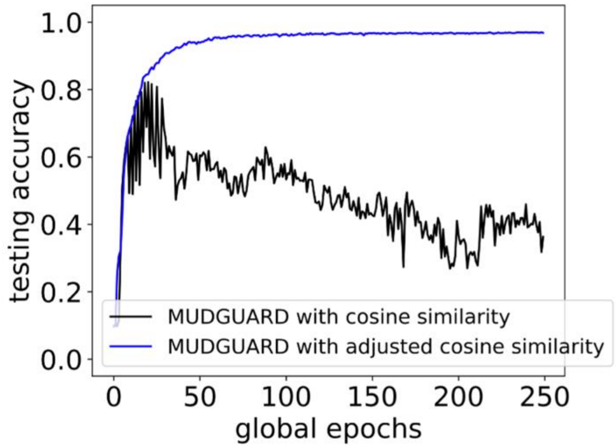


Fig. 12: Comparison of MUDGUARD with cosine similarity and adjusted cosine similarity under GA.

Solution: robustness

- Clustering with L2 distance

0.00000	1.01169	0.97561	1.00121	0.92856	0.87801	1.16170	1.16078	1.16186	1.16193
1.01169	0.00000	1.10076	1.05509	1.11455	0.82658	1.22054	1.21734	1.21974	1.21842
0.97561	1.10076	0.00000	0.98400	0.91487	1.04535	1.15020	1.15120	1.15062	1.15113
1.00121	1.05509	0.98400	0.00000	0.98145	1.00986	1.03675	1.03595	1.03597	1.03605
0.92856	1.11455	0.91487	0.98145	0.00000	1.04619	1.14841	1.15347	1.14990	1.15133
0.87801	0.82658	1.04535	1.00986	1.04619	0.00000	0.98727	0.98368	0.98629	0.98514
1.16170	1.22054	1.15020	1.03675	1.14841	0.98727	0.00000	0.00147	0.00061	0.00089
1.16078	1.21734	1.15120	1.03595	1.15347	0.98368	0.00147	0.00000	0.00146	0.00097
1.16186	1.21974	1.15062	1.03597	1.14990	0.98629	0.00061	0.00146	0.00000	0.00050
1.16193	1.21842	1.15113	1.03605	1.15133	0.98514	0.00089	0.00097	0.00050	0.00000

(a) Pairwise cosine distance

0.00000	0.86338	0.87512	0.94522	0.89751	0.85723	1.53722	1.53583	1.53751	1.53762
0.86338	0.00000	0.90396	0.91755	1.01084	0.74407	1.51221	1.50796	1.51127	1.50955
0.87512	0.90396	0.00000	0.90117	0.86601	0.99187	1.49603	1.49717	1.49665	1.49729
0.94522	0.91755	0.90117	0.00000	0.96466	1.01181	1.41758	1.41633	1.41656	1.41667
0.89751	1.01084	0.86601	0.96466	0.00000	1.05767	1.49820	1.50532	1.50040	1.50247
0.85723	0.74407	0.99187	1.01181	1.05767	0.00000	1.41875	1.41345	1.41742	1.41578
1.53722	1.51221	1.49603	1.41758	1.49820	1.41875	0.00000	0.00297	0.00123	0.00179
1.53583	1.50796	1.49717	1.41633	1.50532	1.41345	0.00297	0.00000	0.00295	0.00196
1.53751	1.51127	1.49665	1.41656	1.50040	1.41742	0.00123	0.00295	0.00000	0.00101
1.53762	1.50955	1.49729	1.41667	1.50247	1.41578	0.00179	0.00196	0.00101	0.00000

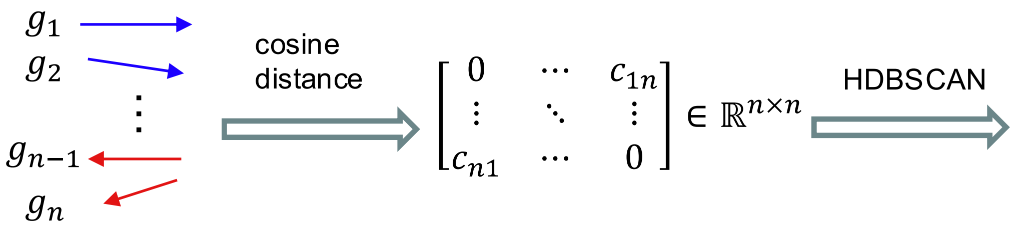
(b) Pairwise adjusted cosine distance (CosM)

0.00000	1.23327	1.24935	1.36995	1.29557	1.25933	3.67430	3.67257	3.67472	3.67485
1.23327	0.00000	1.31066	1.34136	1.46532	1.07769	3.62484	3.62202	3.62484	3.62442
1.24935	1.31066	0.00000	1.29037	1.23302	1.43819	3.58309	3.58226	3.58366	3.58403
1.36995	1.34136	1.29037	0.00000	1.37996	1.44992	3.39557	3.39397	3.39566	3.39585
1.29557	1.46532	1.23302	1.37996	0.00000	1.53537	3.55512	3.55600	3.55615	3.55695
1.25933	1.07769	1.43819	1.44992	1.53537	0.00000	3.41839	3.41513	3.41827	3.41778
3.67430	3.62484	3.58309	3.39557	3.55512	3.41839	0.00000	0.01106	0.00355	0.00665
3.67257	3.62202	3.58226	3.39397	3.55600	3.41513	0.01106	0.00000	0.00868	0.00568
3.67472	3.62484	3.58366	3.39566	3.55615	3.41827	0.00355	0.00868	0.00000	0.00370
3.67485	3.62442	3.58403	3.39585	3.55695	3.41778	0.00665	0.00568	0.00370	0.00000

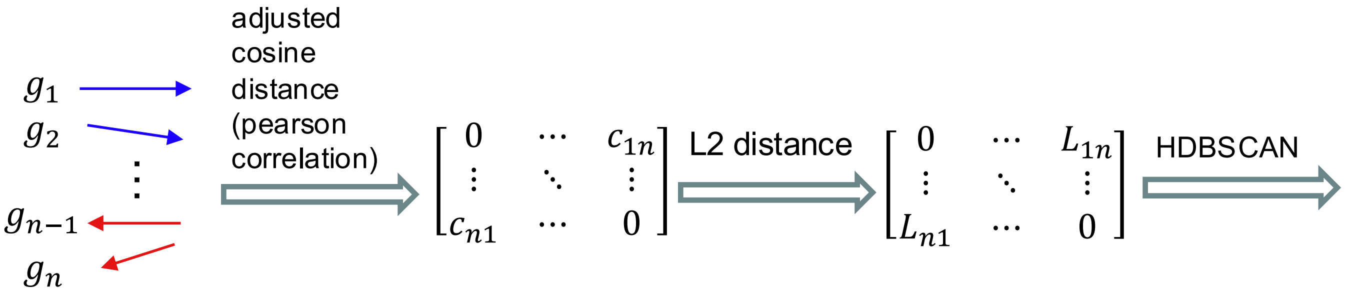
(c) Pairwise L_2 distance for CosM

Solution: robustness

- FLAME[11]:



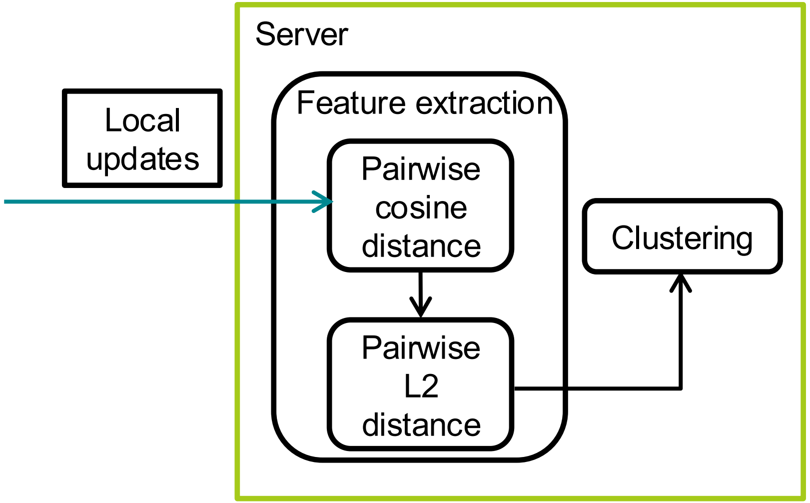
- MUDGUARD



[11] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen M¨ollering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. Flame: Taming backdoors in federated learning. In USENIX Security, 2022

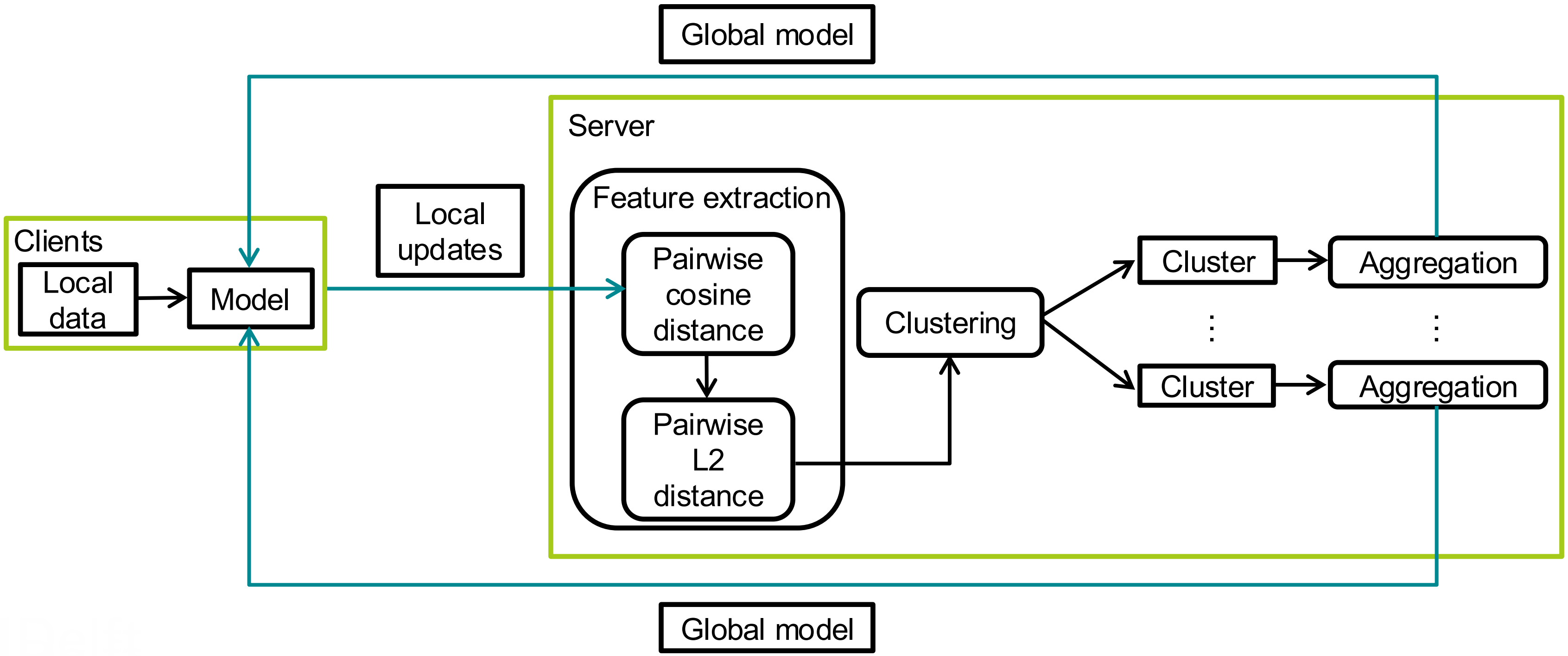
Solution: robustness

- Feature extraction



Solution: robustness

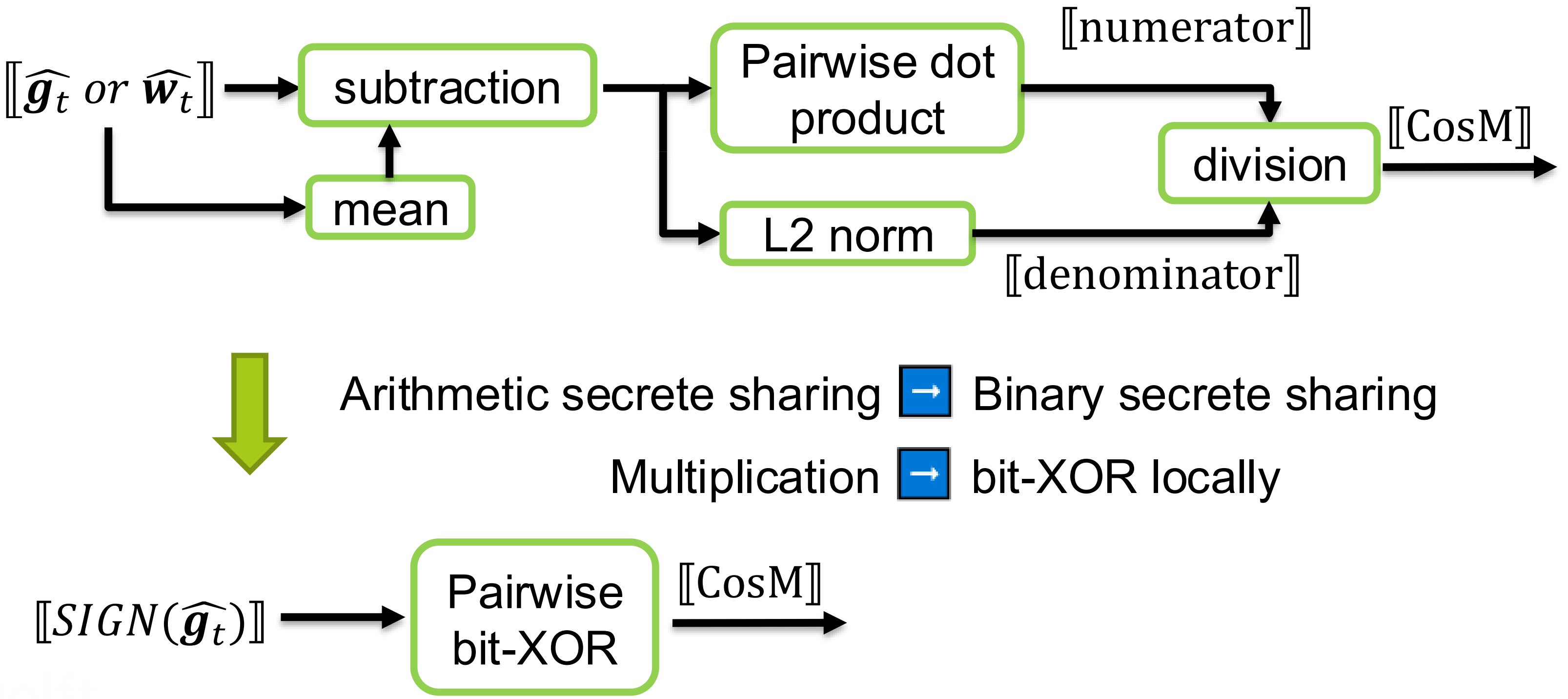
- Overview



Solution: privacy

- Secure Multi-party computation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Solution: privacy

- Differential attack
  - A small number of honest clients are (single client is) clustered together with malicious clients.
  - Malicious clients behave honestly to gain honest aggregations.
- Differential Privacy

$$\widetilde{\boldsymbol{g}}_t^i \leftarrow \boldsymbol{g}_t^i / \max(1, \|\boldsymbol{g}_t^i\|_2/\Delta) + \mathcal{N}(0, \Delta^2\sigma^2)$$

$$\hat{\boldsymbol{g}}_t^i \leftarrow \mathsf{KS}(\widetilde{\boldsymbol{g}}_t^i, \mathcal{N}) \cdot \widetilde{\boldsymbol{g}}_t^i$$

Solution: privacy

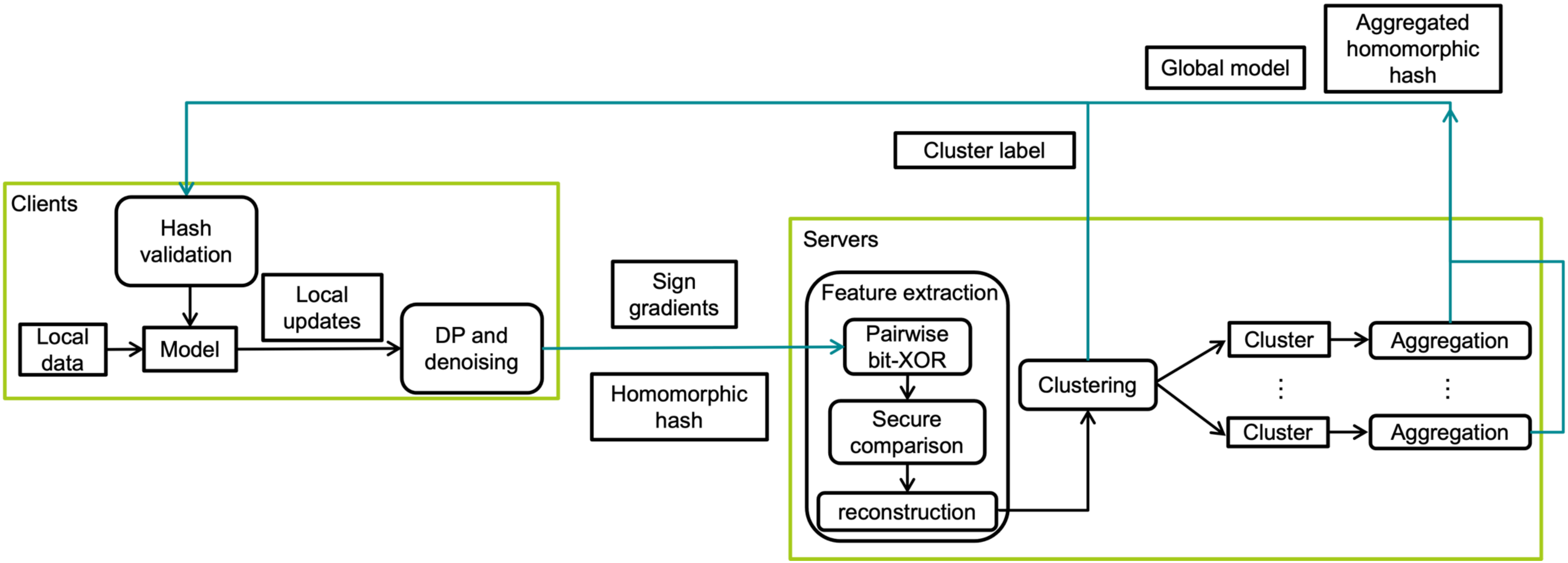
- Defend against malicious server
  - Sending wrong secret shares to the clients

- Homomorphic hash function

$$\mathsf{H}(x) = (g^{\mathsf{H}'_{\delta,\phi}(x)}, h^{\mathsf{H}'_{\delta,\phi}(x)})$$

$$\mathsf{H}(x_1 + x_2) \leftarrow (g^{\mathsf{H}'_{\delta,\phi}(x_1)+\mathsf{H}'_{\delta,\phi}(x_2)}, h^{\mathsf{H}'_{\delta,\phi}(x_1)+\mathsf{H}'_{\delta,\phi}(x_2)})$$

Solution: privacy

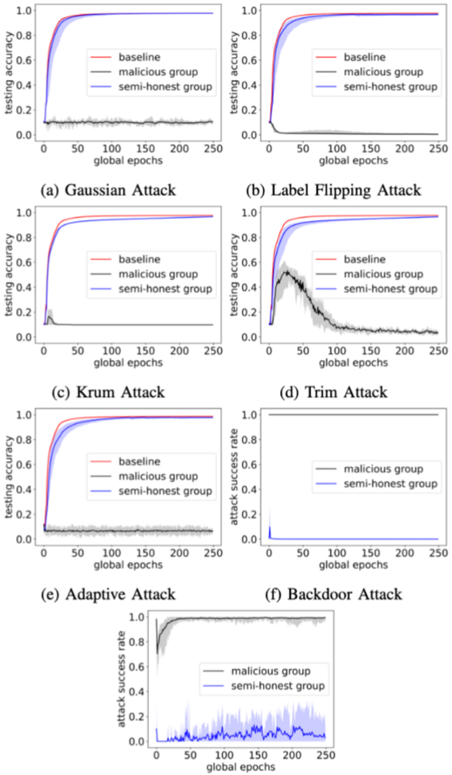


Results

ξ =0.6		MNIST		FMNIST		CIFAR-10	
		TPR	TNR	TPR	TNR	TPR	TNR
GA	FLAME	0.821	0.846	0.848	0.847	0.879	0.928
	weights-MUDGUARD	1	1	1	1	1	1
	MUDGUARD	0.957	1	0.94	1	0.966	1
LFA	FLAME	0.653	0.612	0.634	0.655	0.742	0.711
	weights-MUDGUARD	0.974	0.987	0.975	0.977	0.98	0.985
	MUDGUARD	0.929	0.924	0.927	0.916	0.943	0.967
Krum	FLAME	0.587	0.622	0.521	0.63	0.527	0.578
	weights-MUDGUARD	0.974	0.953	0.973	0.968	0.971	0.966
	MUDGUARD	0.916	0.929	0.96	0.933	0.967	0.959
Trim	FLAME	0.691	0.679	0.699	0.664	0.646	0.615
	weights-MUDGUARD	0.976	0.964	0.975	0.965	0.973	0.988
	MUDGUARD	0.938	0.944	0.927	0.913	0.964	0.958
AA	FLAME	0.591	0.573	0.612	0.625	0.766	0.719
	weights-MUDGUARD	0.998	0.982	0.99	0.982	0.984	0.982
	MUDGUARD	0.971	0.943	0.941	0.935	0.943	0.96
BA	FLAME	0.777	0.763	0.794	0.83	0.856	0.897
	weights-MUDGUARD	0.957	0.969	0.965	0.97	0.963	0.979
	MUDGUARD	0.936	0.928	0.926	0.931	0.947	0.928
EA	FLAME	0.313	0.32	_	_	0.248	0.288
	weights-MUDGUARD	0.899	0.903	_	_	0.893	0.921
	MUDGUARD	0.856	0.876	_	_	0.827	0.83

TABLE IV: Effectiveness of clustering among FLAME method, weights-MUDGUARD, and MUDGUARD.

Results

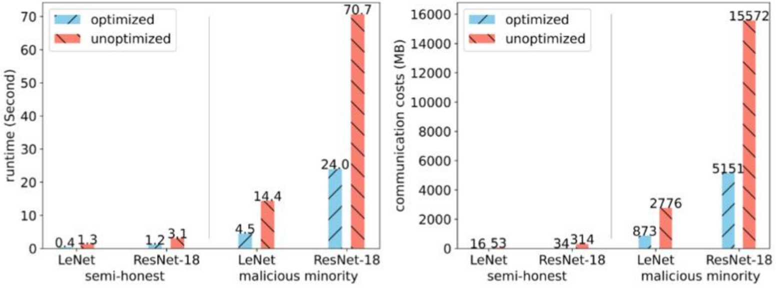


(a) Gaussian Attack (b) Label Flipping Attack

(c) Krum Attack (d) Trim Attack

(e) Adaptive Attack (f) Backdoor Attack

(g) Edge-case Attack



(a) Runtime (b) Communication costs

Thanks!

Q & A