RFLPA: A Robust Federated Learning Framework against Poisoning Attacks with Secure Aggregation

Peihua Mai, Ran Yan, Yan Pang

Research Background

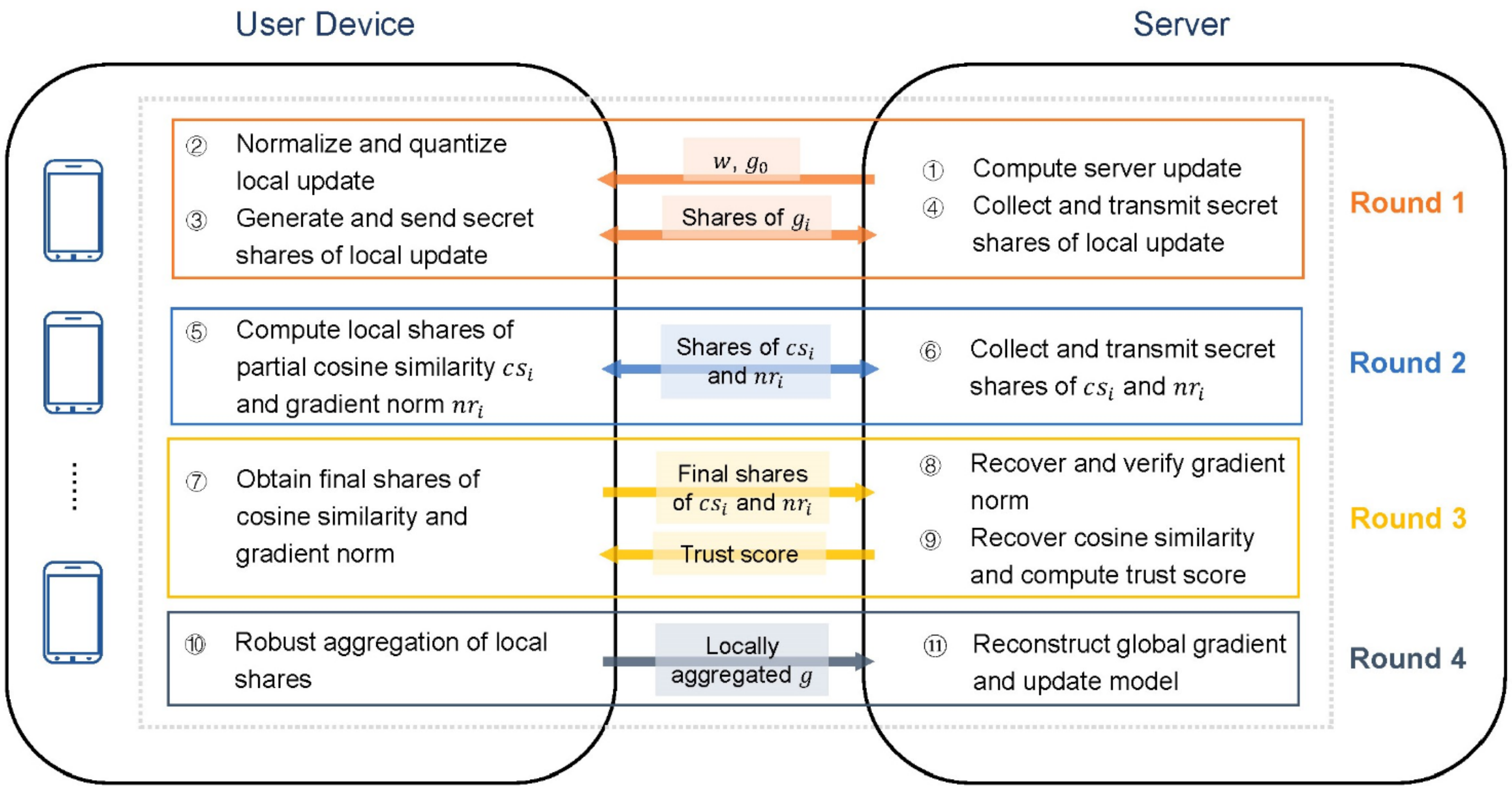
- **Contradiction between privacy and robustness**
  - ✓ SecAgg allows the server to obtain the sum of **gradients without inspecting individual user updates**
  - ✓ Most defense strategies against poisoning attack **require the server to access individual local updates** to detect the attackers

- **Contributions**
  - ✓ We propose a federated learning framework that overcomes privacy and robustness issues with reduced communication cost, especially for high-dimensional models.
  - ✓ To protect the privacy of local gradients, we propose a novel dot product aggregation protocol.
  - ✓ Our framework guarantees the secrecy and integrity of secret shares for a server-mediated network model using encryption and signature techniques.

Design Goals

- **Privacy**
  - ✓ The server learns only the aggregation weights and global gradients.
  - ✓ Leverage secret sharing-based protocol to ensure security.
- **Robustness**
  - ✓ The model accuracy should be robust against model poisonous attack
  - ✓ Compute the similarity between client update and server update
- **Efficiency**
  - ✓ Our framework should maintain computation and communication efficiency even if it is operated on high dimensional vectors
  - ✓ Employ Packed Shamir Secret Sharing to represent multiple secrets by a single polynomial

Overview



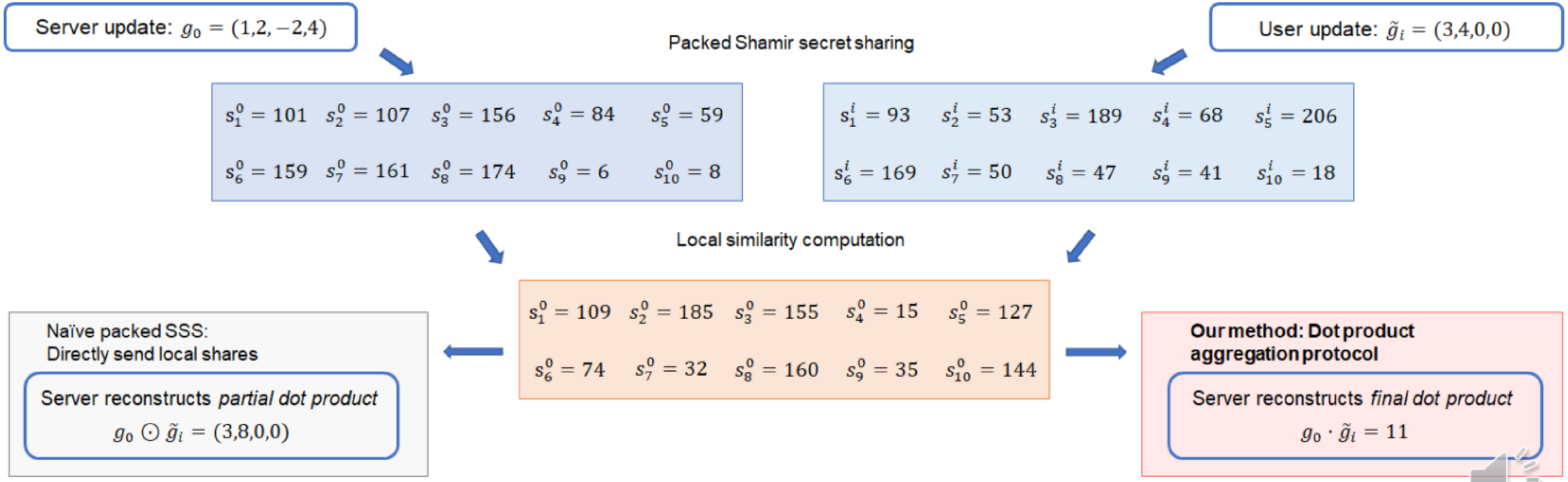
- Secret shares the local gradients with verifiable Packed Shamir Secret Sharing
- Compute the cosine similarity between local updates and server updates
- Aggregate the local updates using the cosine similarity

Aggregation rule

$$TS_i = \max\left(0, \frac{\langle \mathbf{g}_i, \mathbf{g}_0 \rangle}{\|\mathbf{g}_i\|\|\mathbf{g}_0\|}\right) = \max\left(0, \frac{\langle \bar{\mathbf{g}}_i, \mathbf{g}_0 \rangle}{\|\mathbf{g}_0\|^2}\right) \qquad \mathbf{g} = \frac{1}{\sum_{i=1}^{N} TS_i} \sum_{i=1}^{N} TS_i \cdot \bar{\mathbf{g}}_i$$

Dot Product Aggregation Protocol

- Directly applying packed secret sharing may increase the risk of information leakage when calculating cosine similarity and gradient norm.
- Our proposed protocol ensures that **only the single value of dot product is released to the server**.



Comparison with Existing Frameworks

	Robustness against malicious users	Privacy Protection against server	Collusion threshold during model training	MPC techniques
FedAvg	Yes	No	/	/
Bulyan	Yes	No	/	/
Trim-mean	Yes	No	/	/
KRUM	Yes	No	/	/
Central DP	Yes	No	/	/
Local DP	Not effective	Yes	/	/
RFA	No	Yes	/	/
PEFL	Yes	Yes	1	HE (Paillier)
PBFL	Yes	Yes	1	HE (CKKS)
ShieldFL	Yes	Yes	1	HE (Paillier)
SecureFL	Yes	Yes	1	MPC & HE (BFV)
RoFL	Yes	Yes	$O(N)$	ZKP
ELSA	Yes	Yes	1	MPC
BREA	Yes	Yes	$O(N)$	Secret sharing
RFLPA	Yes	Yes	$O(N)$	Secret sharing

Compared with existing methods that achieve the robust and privacy goals, RFLPA:

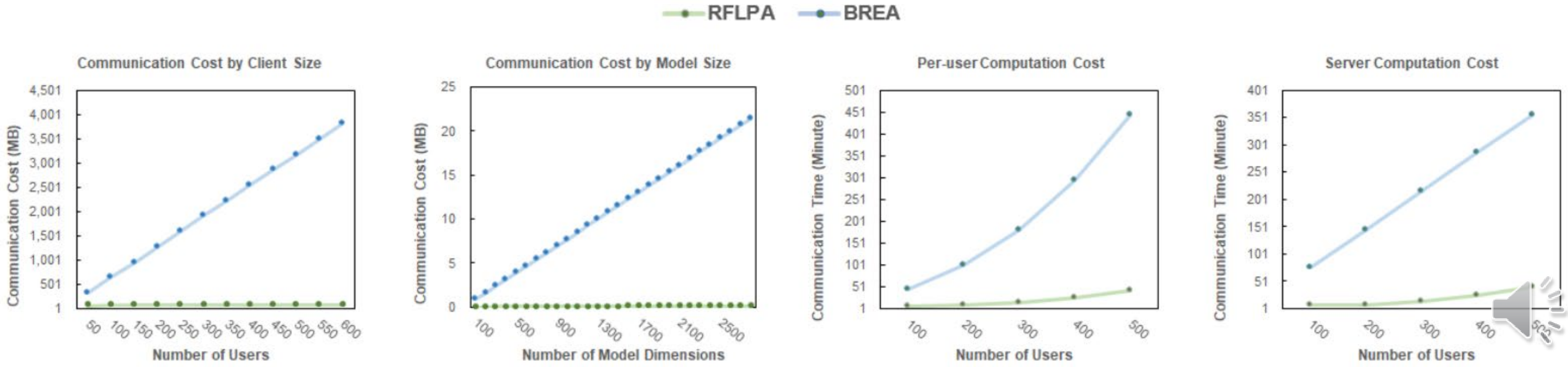
- Get rid of the assumption of **two non-colluding parties;**
- Mitigate the **heavy computation overhead** caused by HE and ZKP methods.

Efficiency Analysis

- Communication complexity of our protocol reduces from $O(MN + N)$ to $O(M + N)$.
- The server-side computation overhead is reduced to $O((M + N)\log^2 N \log\log N)$.

	RFLPA		BERA	
	Computation	Communication	Computation	Communication
Server	$O((M+N)\log^2 N \log\log N)$	$O((M+N)N)$	$O((N^2+MN)\log^2 N \log\log N)$	$O(MN+N^2)$
User	$O((M+N^2)\log^2 N)$	$O((M+N))$	$O(MN\log^2 N + MN^2)$	$O(MN+N)$

- Our framework reduces the communication and computation cost by over 75% compared with BREA.



Accuracy Analysis

- RFLPA demonstrates more stable performance for up to 30% adversaries compared to other baselines.

Proportion of Attackers		Gradient Manipulation No	10%	20%	30%	Label Flipping No	10%	20%	30%
FedAvg	MNIST	**0.98 ±0.0**	0.46 ±0.1	0.40 ±0.1	0.32 ±0.0	**0.98 ±0.0**	**0.96 ±0.0**	0.92 ±0.0	0.82 ±0.0
	F-MNIST	**0.88 ±0.0**	0.55 ±0.0	0.51 ±0.0	0.45 ±0.1	**0.88 ±0.0**	0.82 ±0.0	0.73 ±0.0	0.69 ±0.0
	CIFAR-10	0.76 ±0.3	0.14 ±0.2	0.13 ±0.8	0.13 ±0.2	0.76 ±0.3	0.72 ±1.1	0.68 ±2.7	0.59 ±0.8
Bulyan	MNIST	0.98 ±0.0	0.92 ±0.0	0.89 ±0.0	0.87 ±0.0	0.98 ±0.0	0.91 ±0.0	0.90 ±0.0	0.87 ±0.0
	F-MNIST	0.86 ±0.0	0.73 ±0.0	0.71 ±0.1	0.69 ±0.0	0.86 ±0.0	0.76 ±0.0	0.70 ±0.1	0.68 ±0.0
	CIFAR-10	**0.77 ±1.0**	**0.73 ±0.8**	0.45 ±1.2	0.27 ±0.6	**0.77 ±1.0**	**0.72 ±0.2**	0.62 ±1.8	0.40 ±0.9
Trim-mean	MNIST	0.98 ±0.0	0.95 ±0.0	0.93 ±0.0	0.91 ±0.0	0.98 ±0.0	0.95 ±0.0	0.92 ±0.0	0.90 ±0.0
	F-MNIST	0.86 ±0.0	0.81 ±0.0	0.74 ±0.0	0.71 ±0.0	0.86 ±0.0	0.78 ±0.0	0.74 ±0.0	0.73 ±0.0
	CIFAR-10	0.76 ±1.0	0.57 ±2.1	0.51 ±1.1	0.47 ±2.2	0.76 ±1.0	0.71 ±1.3	0.68 ±0.7	0.56 ±1.1
LDP	MNIST	0.87 ±0.1	0.13 ±0.0	0.10 ±0.0	0.10 ±0.0	0.87 ±0.1	0.87 ±0.3	0.83 ±1.2	0.77 ±2.1
	F-MNIST	0.74 ±0.1	0.59 ±0.4	0.53 ±1.2	0.12 ±0.0	0.74 ±0.1	0.63 ±0.5	0.62 ±0.2	0.59 ±1.2
	CIFAR-10	0.14 ±0.2	0.14 ±0.2	0.12 ±0.3	0.12 ±0.1	0.14 ±0.2	0.14 ±0.2	0.14 ±0.3	0.13 ±0.1
CDP	MNIST	0.96 ±0.0	0.96 ±0.0	0.95 ±0.0	0.94 ±0.0	0.96 ±0.0	0.96 ±0.0	0.95 ±0.3	0.91 ±0.2
	F-MNIST	0.83 ±0.1	0.51 ±0.1	0.41 ±0.0	0.34 ±0.1	0.83 ±0.1	0.81 ±0.5	0.79 ±0.0	0.78 ±0.7
	CIFAR-10	0.71 ±1.2	0.12 ±0.5	0.12 ±0.3	0.12 ±0.3	0.71 ±1.2	0.68 ±0.7	0.66 ±1.5	0.63 ±1.3
BREA	MNIST	0.94 ±0.0	0.93 ±0.0	0.93 ±0.0	0.93 ±0.0	0.94 ±0.0	0.94 ±0.0	0.93 ±0.0	0.93 ±0.0
	F-MNIST	0.84 ±0.0	0.83 ±0.0	0.82 ±0.0	0.81 ±0.0	0.84 ±0.0	0.84 ±0.0	0.82 ±0.0	0.81 ±0.0
	CIFAR-10	0.70 ±1.0	0.69 ±1.1	0.68 ±1.9	0.68 ±0.7	0.70 ±1.0	0.70 ±2.2	0.67 ±0.9	0.65 ±2.7
RFLPA	MNIST	0.96 ±0.0	**0.96 ±0.0**	**0.95 ±0.0**	**0.95 ±0.0**	0.96 ±0.0	0.96 ±0.0	**0.95 ±0.0**	**0.95 ±0.0**
	F-MNIST	0.84 ±0.0	**0.84 ±0.0**	**0.83 ±0.0**	**0.82 ±0.0**	0.84 ±0.0	**0.83 ±0.0**	**0.83 ±0.0**	**0.82 ±0.0**
	CIFAR-10	0.74 ±2.3	0.70 ±1.8	**0.70 ±1.9**	**0.69 ±1.8**	0.74 ±2.3	0.71 ±1.7	**0.70 ±1.6**	**0.69 ±0.8**