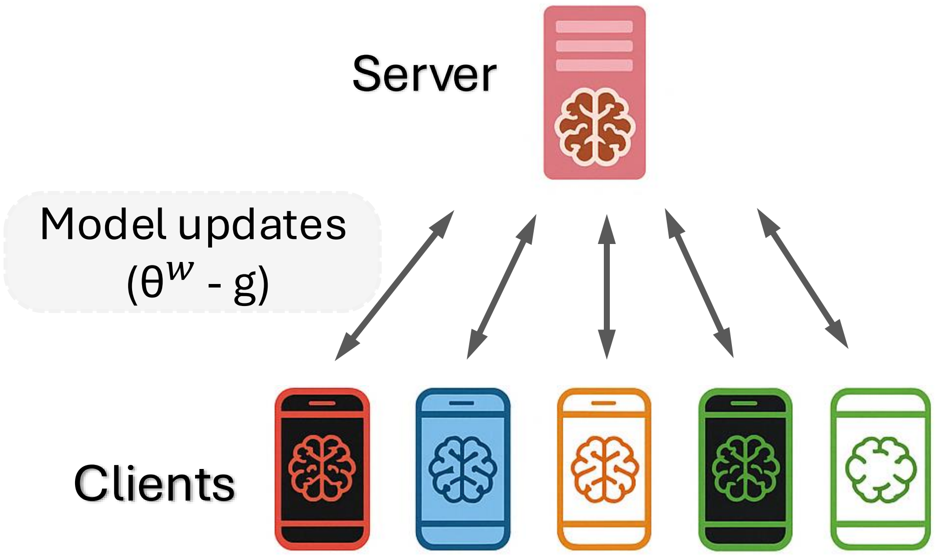
Not All Edges are Equally Robust: Evaluating the Robustness of Ranking-Based Federated Learning

Zirui Gong[1], Yanjun Zhang[2], Leo Yu Zhang[1], Zhaoxi Zhang[2,1], Yong Xiang[3], Shirui Pan[1]

[1] Griffith University
[2] University of Technology Sydney
[3] Deakin University

Contact: zirui.gong@griffithuni.edu.au

Federated Learning and Security Issues



✓ Mitigate data silo issues

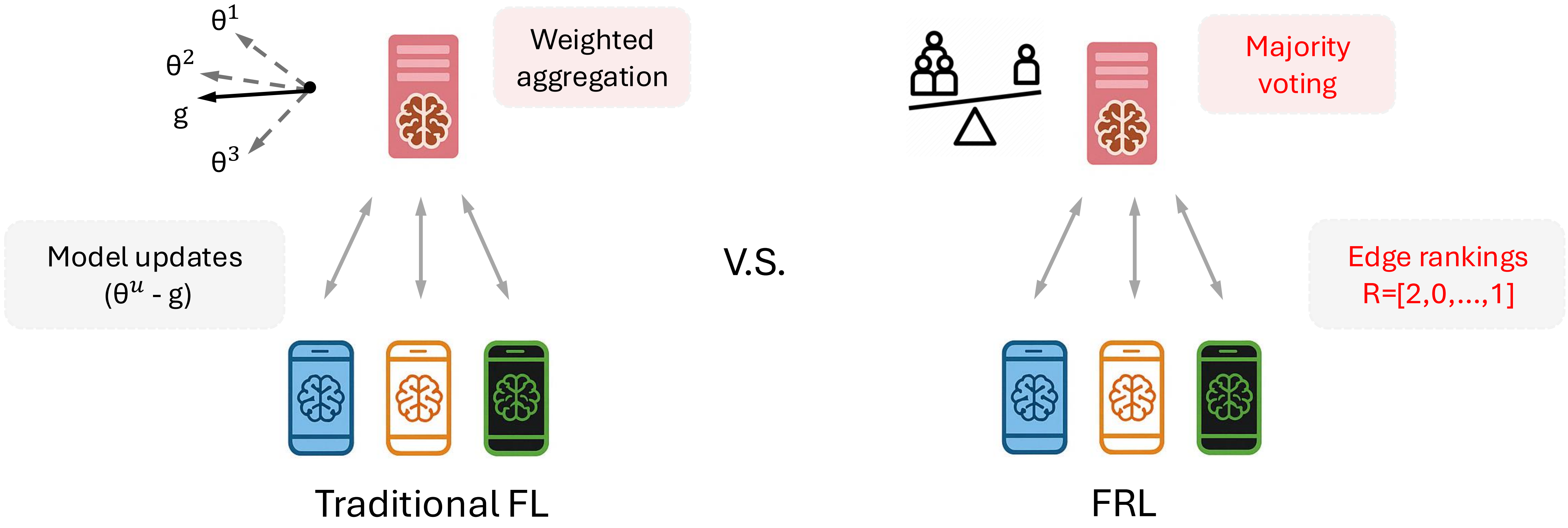
✓ Enable training of ML models on diverse datasets

The decentralized nature of FL makes it susceptible to client-side poisoning attacks and hinder FL's development and real-world application.

Research scope: this work focus on the security perspective of FL framework aim to evaluating the robustness of existing FL frameworks.

Problem Statement

- SOTA robust FL framework: Federated Ranking Learning (FRL) [1]



[1] Mozaffari, Hamid, Virat Shejwalkar, and Amir Houmansadr. "Every vote counts: Ranking-Based training of federated learning to resist poisoning attacks." *32nd USENIX security symposium (USENIX Security 23)*. 2023.

Problem Statement

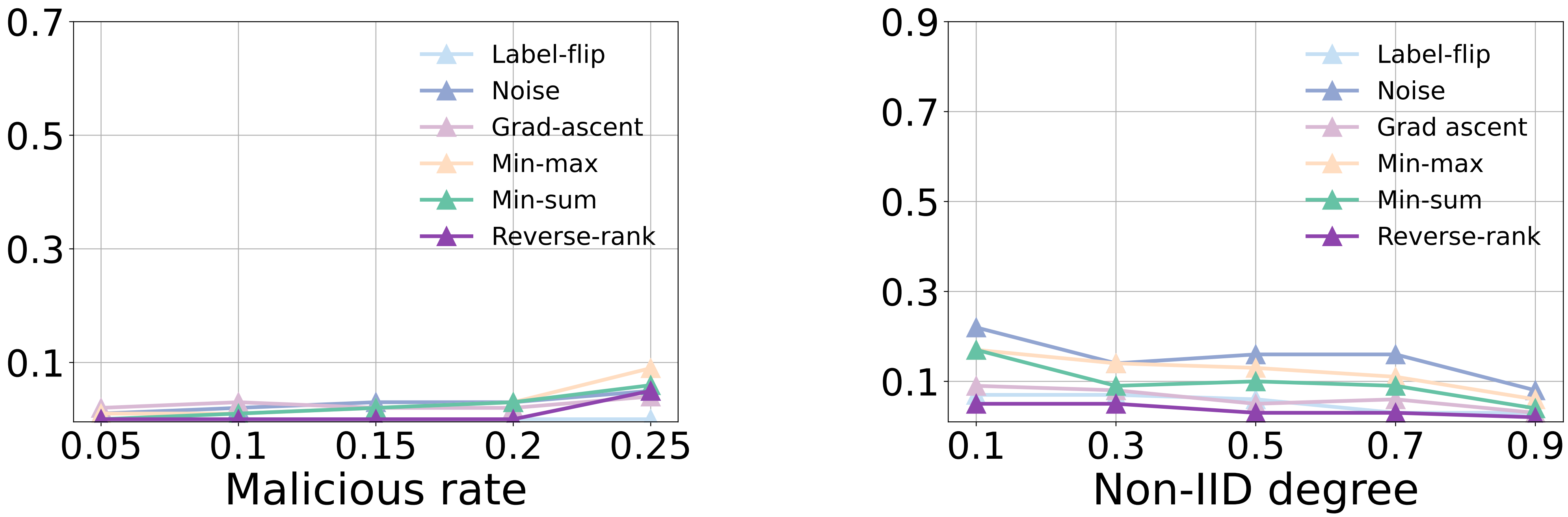
- Why is it robust against client-side poisoning attacks?

  - Ranking format narrows the potential space for malicious updates from an infinite range to n!, effectively bounding the adversary's damage within a defined budget, e.g., (n-1).

  - Server-side majority voting prevents malicious clients from making significant modifications to the global model, as each client only has a single vote.

Research Motivation

How robust is FRL?



Attack impact: $\phi = \mathrm{acc}_{\mathrm{drop}} / \mathrm{acc}_{\mathrm{benign}} \times 100\%$

? Is there any vulnerability inside this framework?

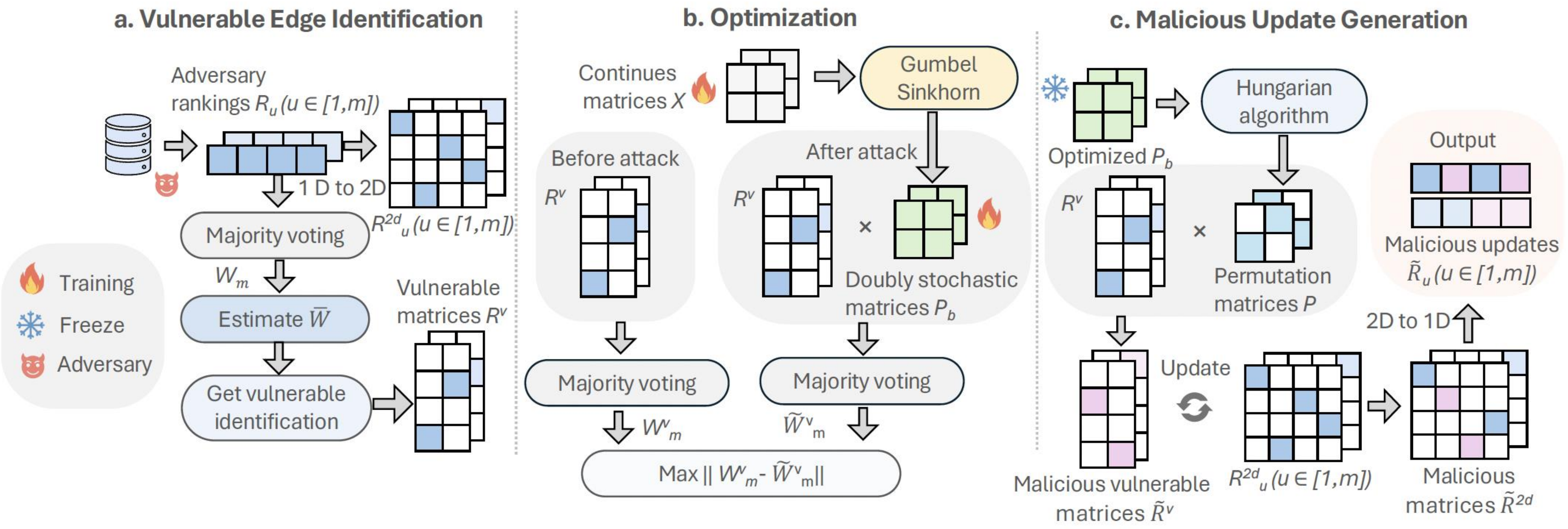
Our work

- We conduct the first systematic analysis of FRL's robustness, uncovering a critical vulnerability within the framework.

- Based on the results of the analysis, we design and implement a new attack (VEM) that targets and effectively manipulates the vulnerable edges.

- Extensive experiments across different network architectures and datasets demonstrate that our VEM significantly outperforms SOTA attacks.

Overall Framework

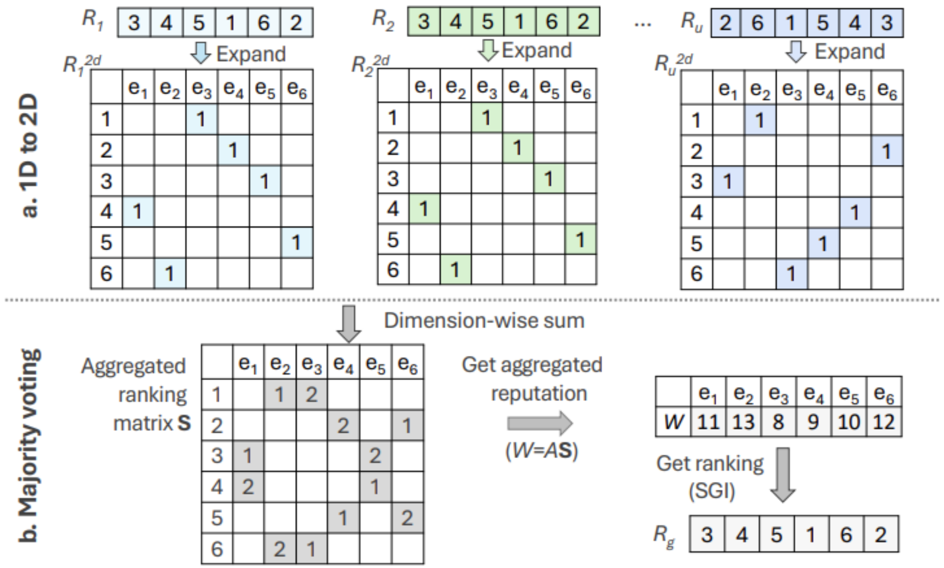


Vulnerable Edges Identification

Definition 1 (Permutation Matrix). Given a permutation R of n elements, the corresponding permutation matrix R^{2d} is an $n \times n$ matrix defined as follows:

$$R^{2d}[i,j] = \begin{cases} 1 & \text{if } R[i] = j, \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

In R^{2d}, the column j indicates the edge ID, and the row i indicates the *reputation*. For instance, as shown in Fig.



- If the importance difference between edge and the selection boundary is smaller than the maximum damage the adversary can cause in one round, we call the that edge a vulnerable edge.

Theorem 1. *Give the aggregated reputation of $U-m$ benign users, i.e., $\overline{W} = [\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_n]$, the reputation of a vulnerable edge e_v is bounded by*

$$\overline{w}_{\max} - m(a_n - a_1) < \overline{w}_v < \overline{w}_{\min} + m(a_n - a_1), \qquad (5)$$

where $\overline{w}_{\min} = \min(\overline{W}^{\text{in}})$ *and* $\overline{w}_{\max} = \max(\overline{W}^{\text{out}})$.

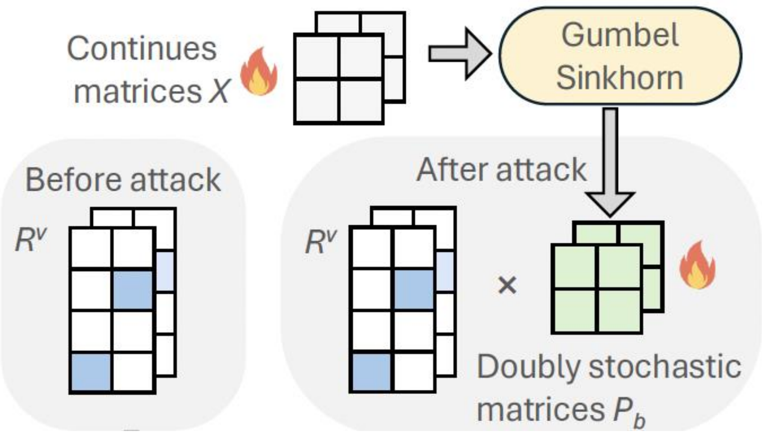
Optimization

- Overall objective: The global model's importance of those vulnerable edges after attack deviates significantly from their original values.

$$\arg \max_{\widetilde{R}_1^v, \widetilde{R}_2^v, \ldots, \widetilde{R}_m^v} \left\| W^v - \widetilde{W}^v \right\|,$$

$$\text{s.t. } W^v = A \sum_{u=1}^{U} R_u^v,$$

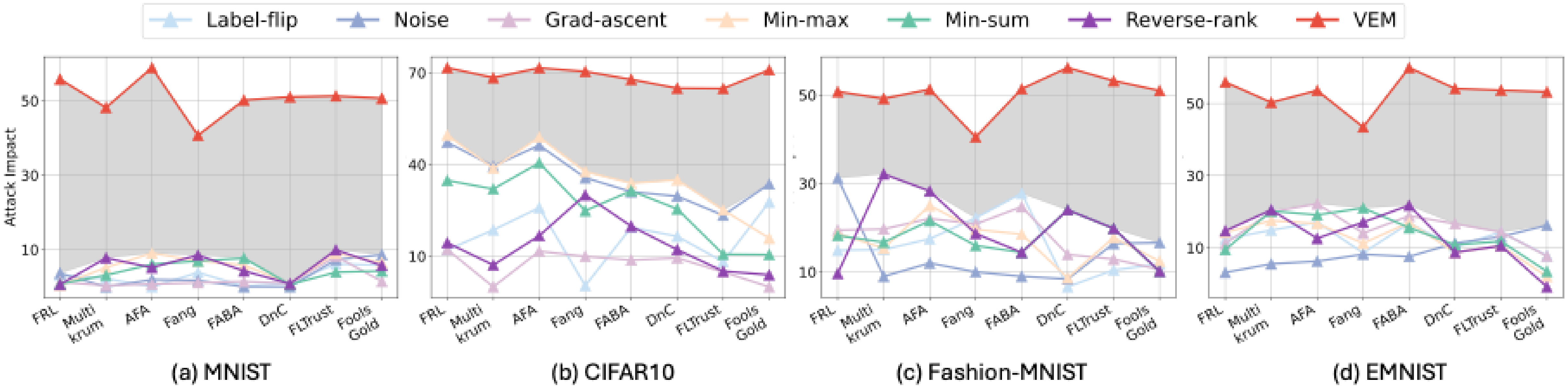
$$\widetilde{W}^v = A\left(\sum_{u=1}^{m} \widetilde{R}_u^v + \sum_{u=m+1}^{U} R_u^v\right),$$



- Challenges: The optimization function is not continuous, so it cannot be solved directly.

Main Results

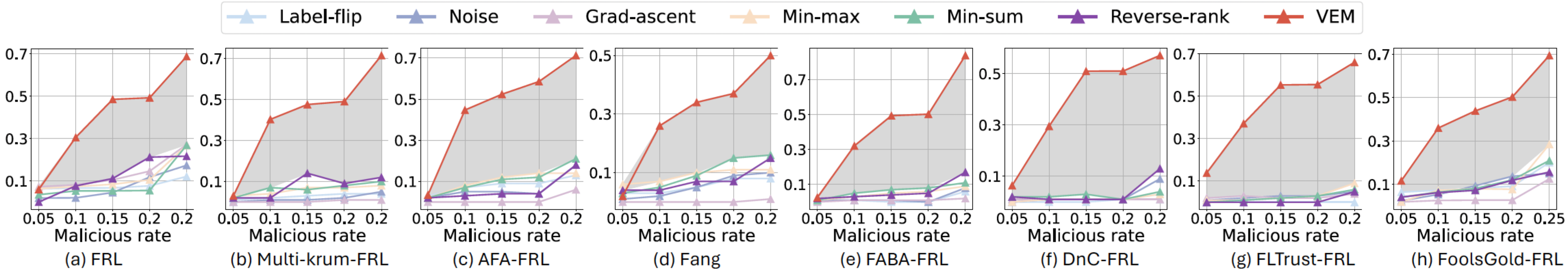
Comparison with the State-of-the-art Attacks under different defenses.



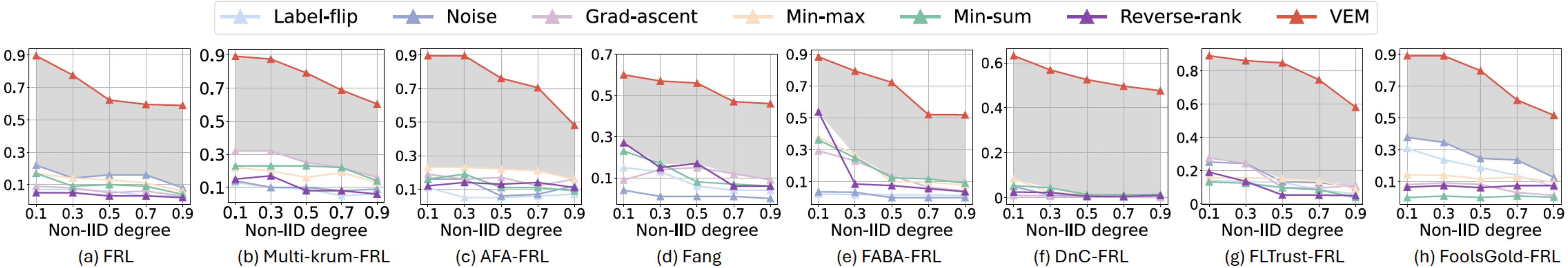
It achieves 53.23% attack impact and is 3.7x more impactful than others.

Ablation study

Impact of the percentage of malicious clients



Impact of non-IID degree



Discussion and Further Work

- Investigate targeted poisoning attacks under ranking-based FL.
- Certified robustness evaluation.
- Design more robust FL framework with less information sharing.

Thank You!

Paper

Linkedin