Reversing Gresham's Law: Exploiting Parasitic Dependencies for Free-Rider Detection in Blockchain Federated Learning

Donghang Duan
Yingcai Honor College
University of Electronic Science and Technology of China
Chengdu, China

I. Problem Definition

In this section, we define our system model first. Then we formally present the problem to be solved in this paper.

A. System Model

In this paper, we consider a blockchain-based federated learning system operating over multiple rounds $t \in \{1, 2, \ldots, T\}$ in an environment potentially containing untrusted participants. The system involves a requester $\mathcal{R}$ and a set of N participants, denoted as $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$.

The participants $\mathcal{P}$ are partitioned into two disjoint sets: honest clients $\mathcal{H} = \{h_1, h_2, \ldots, h_{N_h}\}$ and free-riders $\mathcal{F} = \{f_1, f_2, \ldots, f_{N_f}\}$, where $N = N_h + N_f$. Honest clients genuinely train the model on their local dataset $\mathcal{D}_i$, while free-riders who lacks data or unwilling to expend resources, attempt to obtain rewards and the final model by submitting fabricated updates.

To detect and remove the free-riders, the system employs a reverse auction mechanism coupled with a reputation system. In each round t, participants $p_i \in \mathcal{P}$ will be asked to submit a bid to the requester $\mathcal{R}$. A bid can be denoted as $(P_i^t,\ R_i^t)$, where P_i^t represents the promised contribution to the increase in model performance and R_i^t denotes the reward that the participant p_i expects. The requester $\mathcal{R}$ then selects a subset $C_t \subseteq \mathcal{P}$ of size M_t by ranking the bids based on cost-effectiveness defined by the ratio $\rho_i^t = P_i^t / R_i^t$. In this way, the system forms a reverse auction where participants compete by offering better cost-effectiveness ratios to be selected.

Each participant p_i possesses a reputation score r_i^t, initialized to r_0. After the selected participants C_t submit their model updates, the requester $\mathcal{R}$ publishes a test dataset $\mathcal{D}_{\text{test}}$ and verifies the updates' performance on it, which can be checked on the blockchain. The reputation r_i^t of participants in C_t is updated based on this verification outcome, while participants not selected maintain their reputation. If the verification result aligns with the promised contribution P_i^t, the participant is rewarded with R_i^t and its reputation is increased. Otherwise, the participant is penalized with a decreased on its reputation.

We assume the data distribution across participants is non-IID (Independent and Identically Distributed): participant i's private dataset $\mathcal{D}_i$ may follow a distribution different from others and $\mathcal{D}_{\text{test}}$. The overall goal is to remove all the free-riders from the system while optimize the final global model's performance on $\mathcal{D}_{\text{test}}$ after the least number of rounds T.

B. Problem Formulation

Free-rider $j \in \mathcal{F}$ aims to maximize its reward, while minimizing its computational and data costs. In our paper, we assume that the free-riders don't have any data. They attempt to generate fabricated gradients using an advanced mechanism $\mathcal{A}$, which is based on the Adam (Adaptive Moment Estimation) optimizer, to mimic honest participation. According to previous global models' gradients, the mechanism $\mathcal{A}$ can potentially generate fabricated gradients indistinguishable from the gradients generated by honest clients on statistical grounds, exploiting the non-IID nature of data in federated learning.

However, the mechanism $\mathcal{A}$ fundamentally relies on fresh and real gradient updates from normal clients to generate fabricated gradients that are effective in improving the model performance. If the available history of model updates is out of date or lack of amount, the fabricated gradients' update direction will significantly deviate from the true gradient direction, potentially leading to a decrease in the model performance and revealing the free-rider's nature.

The Adam-based mechanism $\mathcal{A}$ is defined as follows:

$$\mathcal{A}(\theta^t, \theta^{t-1}) = \begin{cases} \theta^t + \delta & \text{if } t = 0 \\ \theta^t + \Delta\theta_{\text{Adam}} & \text{if } t > 0 \end{cases} \tag{1}$$

$$g_t = \theta^t - \theta^{t-1} \tag{2}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{4}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{5}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{6}$$

$$\Delta\theta_{\text{Adam}} = -\eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{7}$$

Where:

- $\delta \sim \mathcal{N}(0, \sigma^2)$ is a random gaussian noise.
- m_t and v_t are the first and second moment estimates, initialized as $m_0 = 0, v_0 = 0$.
- $\beta_1, \beta_2 \in [0, 1)$ are exponential decay rates for the moment estimates.
- $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates.
- η is the learning rate used by the free-rider's mechanism.
- ϵ is a small constant to prevent division by zero.

What's worth mentioning is that the term g_0 cannot be computed in the first round ($t = 0$) as it requires θ^0 and θ^{-1}, hence the free-rider submits the original model θ^0 with a random noise δ added to it.

We can prove that the free-riders will be removed from the system due to its greedy strategy, no matter what advanced mechanism they use to generate fabricated gradient updates. Therefore, we aim to detect and remove the free-riders from our system with least reward cost while ensuring that the honest clients won't be misjudged as free-riders.

To achieve this, we formulate a multi-objective optimization problem. Let $\mathcal{X}$ represent the set of tunable system parameters, and $\mathbb{X}$ be the feasible search space for these parameters $\mathcal{X}$.

We define the following performance metrics based on a full simulation run of the system with $\mathcal{X}$:

- $T_{elim}(\mathcal{X})$: The total number of rounds required until all free-riders $f \in \mathcal{F}$ have their reputations $r_f^t < r_{thresh}$.
- $C_{total}(\mathcal{X})$: The cumulative sum of rewards R_i^t disbursed by the requester $\mathcal{R}$ to all selected participants $p_i \in C_t$ up to round $T_{elim}(\mathcal{X})$.
- N_h: The total number of honest clients in the set $\mathcal{H}$.
- $N_{h,elim}(\mathcal{X})$: The number of honest clients $h \in \mathcal{H}$ whose reputations r_h^t fall below r_{thresh} by round $T_{elim}(\mathcal{X})$.
- $FPR(\mathcal{X}) = \frac{N_{h,elim}(\mathcal{X})}{N_h}$: The False Positive Rate, representing the misjudgment rate of honest clients.
- $PFM_{final}(\mathcal{X})$: The performance of the final global model $\theta^{T_{elim}(\mathcal{X})}$ evaluated on the test dataset $\mathcal{D}_{\text{test}}$.
- PFM_{min}: A predefined minimum acceptable performance for $Acc_{final}(\mathcal{X})$.

The optimization problem is to find a set of parameters $\mathcal{X}$ that achieve a Pareto optimal trade-off between minimizing the total reward cost and minimizing the misjudgment rate of honest clients, subject to maintaining a satisfactory final model performance:

$$\begin{aligned} \underset{\mathcal{X} \in \mathbb{X}}{\text{minimize}} \quad & (C_{total}(\mathcal{X}), FPR(\mathcal{X})) \\ \text{subject to} \quad & PFM_{final}(\mathcal{X}) \geq PFM_{min} \end{aligned} \tag{8}$$

Solving this multi-objective optimization problem will yield a set of Pareto optimal solutions. Each solution in this set represents a different balance between $C_{total}(\mathcal{X})$ and $FPR(\mathcal{X})$ that cannot be improved in one objective without degrading the other, while respecting the minimum model accuracy constraint. The final selection of parameters from this Pareto set will depend on the specific priorities and risk tolerance of the requester $\mathcal{R}$.

II. Mathcal Proof of Convergence

The proof of convergence unfolds in several dependent stages. We first establish that the selection size, M_t, converges to the actual number of free-riders, N_f. This convergence, driven by the bidding advantages of free-riders and the reputation dynamics, leads to a diminishing participation rate of honest clients in the selected set. Consequently, free-riders, who rely on fresh updates from honest clients, experience a degradation in their performance, causing their success probability to plummet. This sequence of events ensures their continued selection (as they exhibit large reputation changes, albeit negative) and triggers a rapid collapse of their reputations, ultimately leading to their elimination from the system.

Lemma 1 (Convergence of Selection Size M_t)**.** *The selection size M_t converges to the number of free-riders N_f as $t \to \infty$.*

Proof. We analyze the convergence behavior of M_t by considering the extended reputation change metric $\delta r_i^t = r_i^t - r_i^{t-q}$, which tracks reputation changes over q rounds. This metric is crucial for understanding the dynamics in both reputation growth and decline phases of free-riders. The value $k_t' = \text{argmax}_j(|\delta r|_j^{top} - |\delta r|_{j+1}^{top})$ identifies the index after the largest gap in sorted absolute reputation changes, and M_t is updated via $M_t = \lfloor \omega M_{t-1} + (1-\omega) k_t' \rceil$.

In the early stages, when free-riders successfully mimic honest clients ($q_f^t \approx 1$), their promised contributions P_f^t are high relative to their requested rewards R_f^t, leading to $\rho_f^t \approx \gamma \bar{\rho}_h^t$ with $\gamma > 1$ (Assumption 2). This results in a higher selection probability $p_f^t > p_h^t$. Consequently, free-riders are more likely to be in C_t and experience reputation increases. The reputation changes for successful free-riders are positive:

$$\delta r_f^t = r_f^t - r_f^{t-q} \approx \sum_{i=t-q}^{t} p_f^i \cdot \alpha \tag{9}$$

Similarly, selected honest clients also experience positive reputation changes:

$$\delta r_h^t = r_h^t - r_h^{t-q} \approx \sum_{i=t-q}^{t} p_h^i \cdot \alpha \tag{10}$$

Due to $p_f^t > p_h^t$, free-riders are more consistently selected and accumulate reputation changes. When clients are ranked by $|\delta r_i^t|$ in descending order, the top positions are predominantly occupied by these N_f free-riders. Thus, the largest gap in $|\delta r_i^t|$ values likely occurs after these N_f clients, leading to:

$$k_t' \approx N_f \tag{11}$$

As the system evolves and the honest participation rate η_h^t potentially drops (as will be shown in Section II-A), free-riders may begin to fail. Even during a transition period where some free-riders succeed and others fail, they, as a group, still constitute the majority of selected clients experiencing significant (positive or negative) reputation changes. The distribution of $|\delta r_i^t|$ is expected to maintain a significant gap after approximately N_f clients, thus keeping:

$$k_t' \approx N_f \tag{12}$$

Eventually, when the honest participation rate is low enough such that $q_f^t \to 0$ for $t > T_2$ (as will be shown in Section II-B), free-riders consistently fail upon selection. For a free-rider $f \in \mathcal{F}$ selected in consecutive rounds, the reputation change over q rounds becomes significantly negative:

$$\delta r_f^t = r_f^t - r_f^{t-q} \approx -\sum_{j=0}^{q-1} \beta^{k_f^{t-j}} < 0 \tag{13}$$

The absolute value $|\delta r_f^t|$ grows exponentially with the number of failures k_f^t since $\beta > 1$. Honest clients, if selected, have $|\delta r_h^t| \approx \alpha$ (if successful), but their selection probability p_h^t becomes very low once free-riders dominate C_t. Unselected clients have $|\delta r_i^t| = 0$. Therefore, the absolute reputation changes of consistently failing free-riders dominate all others:

$$\forall f \in \mathcal{F} \cap C_t, \quad |\delta r_f^t| \gg |\delta r_h^t|, |\delta r_i^t| \text{ for } h \in \mathcal{H}, i \notin C_t \tag{14}$$

When ranking clients by $|\delta r_i^t|$, the top N_f positions will be occupied by these failing free-riders. The largest gap in this ranking occurs after this group, leading to:

$$k_t' \to N_f \tag{15}$$

This robust convergence of k_t' to N_f throughout all phases ensures that M_t, governed by the rule $M_t = \lfloor \omega M_{t-1} + (1-\omega) k_t' \rceil$, also converges:

$$\begin{aligned} \lim_{t\to\infty} M_t &= \lim_{t\to\infty} \lfloor \omega M_{t-1} + (1-\omega) k_t' \rceil \\ &= \lfloor \omega \lim_{t\to\infty} M_{t-1} + (1-\omega) \lim_{t\to\infty} k_t' \rceil \\ &= \lfloor \omega N_f + (1-\omega) N_f \rceil \\ &= N_f \end{aligned} \tag{16}$$

This convergence of M_t to N_f is critical as it creates a self-reinforcing mechanism: it ensures that free-riders are continuously targeted for selection, whether they are in a reputation growth phase (due to high ρ_f^t) or a decline phase (due to large negative δr_f^t). □

A. Phase 1: Decline of Honest Client Participation Rate

By Assumption 2, free-riders $\mathcal{F}$ submit bids with an enhanced cost-effectiveness ratio $\rho_f^t \approx \gamma \bar{\rho}_h^t$ where $\gamma > 1$. According to Definition 1 (Selection based on Reverse Auction), this ensures that free-riders have a higher selection probability than honest clients $\mathcal{H}$, i.e., $p_f^t > p_h^t$.

With the selection size M_t converging to N_f (as established in Lemma 1), and free-riders being preferentially selected, the $M_t \approx N_f$ slots in the selected set C_t will be predominantly filled by free-riders. Consequently, the proportion of honest clients in C_t, denoted by $\eta_h^t = \frac{|C_t \cap \mathcal{H}|}{M_t}$, will diminish over time:

$$\eta_h^t = \frac{|C_t \cap \mathcal{H}|}{M_t} \approx \frac{|C_t \cap \mathcal{H}|}{N_f} \to 0 \tag{17}$$

This implies that there exists a time T_1 such that for all subsequent rounds $t > T_1$, the average honest participation rate over the recent τ rounds (as relevant for Assumption 2) falls below the critical threshold θ:

$$\exists T_1 \text{ s.t. } \forall t > T_1, \quad \frac{1}{\tau} \sum_{j=t-\tau}^{t-1} \eta_h^j < \theta \tag{18}$$

During the initial part of this phase, while M_t is still stabilizing or if η_h^t has not yet dropped significantly, both free-riders and selected honest clients might experience reputation gains. However, the preferential selection ensures free-riders are more consistently chosen, contributing to the conditions for M_t's convergence as described in Lemma 1.

B. Phase 2: Free-Rider Performance Degradation

Let $T_2 = T_1 + \tau$. The time T_2 accounts for the window τ needed for the reduced honest client participation (established in Section II-A) to impact the free-riders' ability to generate effective updates. For rounds $t > T_2$, the condition $\frac{1}{\tau}\sum_{j=t-\tau}^{t-1} \eta_h^j < \theta$ holds. According to Assumption 2 (Free-Rider's Greedy Behavior), the success probability q_f^t for free-riders $f \in \mathcal{F}$ drops significantly:

$$\forall f \in \mathcal{F}, \quad q_f^t \to 0 \quad \text{for } t > T_2 \tag{19}$$

At this stage, free-riders can no longer reliably mimic honest updates due to the lack of fresh, genuine gradient information from a sufficient number of honest participants in the recent selection history.

C. Phase 3: Ensured Selection and Reputation Collapse of Free-Riders

Consider the system dynamics for $t > T_2$. Free-riders now consistently fail when selected ($q_f^t \approx 0$) due to the reasons outlined in Section II-B. Crucially, Lemma 1 established that $M_t \approx N_f$. The mechanism for M_t adjustment relies on k_t', which identifies the group of N_f clients with the most significant absolute reputation changes $|\delta r_i^t|$. As free-riders begin to fail, their reputation r_f^t decreases by $\beta^{k_f^t}$ per failure (Definition 2), leading to large negative values for $\delta r_f^t \approx -\sum \beta^{k_f^j}$. The magnitude of these changes ensures that free-riders continue to dominate the top N_f positions in the sorted list of $|\delta r_i^t|$, thus $k_t' \approx N_f$ is maintained, and $M_t \approx N_f$ persists. This means the system continues to select these N_f free-riders, who are now trapped in a cycle of being selected and then failing. This persistent selection and failure leads to their reputation collapse, as formalized in Lemma 2.

Lemma 2 (Reputation Collapse of Free-Riders)**.** *For any free-rider $f \in \mathcal{F}$, there exists a finite time T_f such that its reputation $r_f^{T_f} < r_{thresh}$.*

Proof. Consider $t > T_2$, where $q_f^t \approx 0$ for any $f \in \mathcal{F}$ (from Section II-B). The expected change in reputation for a free-rider $f \in \mathcal{F}$, if selected, is (from Definition 2):

$$\mathbb{E}[\Delta r_f^t | r_f^t, k_f^t] = p_f^t [q_f^t \alpha - (1-q_f^t)\beta^{k_f^t}] \tag{20}$$

Since $q_f^t \approx 0$, this simplifies to:

$$\mathbb{E}[\Delta r_f^t] \approx -p_f^t \beta^{k_f^t} \tag{21}$$

As established in Lemma 1 and discussed in Section II-C, $M_t \to N_f$, and the selected clients are almost exclusively free-riders. Thus, the selection probability $p_f^t \approx 1$ for any $f \in \mathcal{F}$ (assuming it has not yet been eliminated and is among the N_f "most notable" clients in terms of $|\delta r_i^t|$). Let's assume a lower bound $p_f^t \geq p_{min} > 0$ for free-riders that are candidates for selection within $M_t \approx N_f$ before their reputation drops below r_{thresh}.

Let $\Delta r = r_f^{T_2} - r_{thresh}$ be the amount of reputation a free-rider f (active at T_2) needs to lose to fall below r_{thresh}. We need the sum of expected decreases from T_2 until some time $T_f - 1$ to exceed Δr:

$$\sum_{t=T_2}^{T_f-1} \mathbb{E}[-\Delta r_f^t] \approx \sum_{t=T_2}^{T_f-1} p_f^t \beta^{k_f^t} > \Delta r \tag{22}$$

Using the lower bounds p_{min} for p_f^t and $k_{min} = k_f^{T_2}$ for k_f^t (since k_f^t is non-decreasing):

$$\sum_{t=T_2}^{T_f-1} p_f^t \beta^{k_f^t} \geq p_{min} \sum_{t=T_2}^{T_f-1} \beta^{k_f^t} \tag{23}$$

Since k_f^t is non-decreasing and $\beta > 1$, the terms $\beta^{k_f^t}$ are themselves non-decreasing and positive. A simple lower bound for the sum is:

$$p_{min} \sum_{t=T_2}^{T_f-1} \beta^{k_f^t} \geq p_{min} \beta^{k_{min}} (T_f - T_2) \tag{24}$$

To ensure the reputation drops below the threshold r_{thresh}, we require:

$$p_{min} \beta^{k_{min}} (T_f - T_2) > \Delta r \tag{25}$$

Solving for T_f, we get:

$$T_f > T_2 + \frac{\Delta r}{p_{min} \beta^{k_{min}}} \tag{26}$$

Since Δr is finite, $p_{min} > 0$, $\beta > 1$, and k_{min} is finite, the right-hand side is a finite value. Thus, a finite time T_f must exist for each free-rider $f \in \mathcal{F}$ by which its reputation falls below r_{thresh}. □

Theorem 1 (Free-Rider Elimination)**.** *The proposed mechanism ensures that all free-riders are eventually eliminated from the system.*

Proof. According to Lemma 2, for every free-rider $f \in \mathcal{F}$, there exists a finite time T_f at which its reputation $r_f^{T_f}$ drops below the threshold r_{thresh}. Let $T_{max} = \max_{f \in \mathcal{F}} \{T_f\}$. Since each T_f is finite, their maximum, T_{max}, is also finite. For any time $t > T_{max}$, it holds that for all free-riders $f \in \mathcal{F}$, their reputation $r_f^t < r_{thresh}$. Participants whose reputation scores are below r_{thresh} are effectively eliminated or ignored by the system's selection process (as they would not be selected or their contributions would be deemed unreliable). Therefore, after a finite time T_{max}, the system contains no active free-riders. □