國立臺北科技大學
資訊工程系碩士班

碩士學位論文

# 漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

# Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

國立臺北科技大學
資訊工程系碩士班

碩士學位論文

# 漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

# Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

# 摘要

關鍵詞：區塊鏈、聯邦式學習、委員會佔領、驗證者共謀

基於區塊鏈的聯邦式學習 (BCFL) 透過去中心化共識機制解決了信任與隱私問題。現有的 BCFL 系統依賴基於委員會的驗證機制，並假設委員會成員是誠實的或擁有誠實多數。此假設容易受到驗證者共謀的威脅，攻擊者可透過累積權益 (Stake) 來主導委員會。我們識別出一種新型威脅——漸進式委員會佔領攻擊 (PCCA)，理性攻擊者利用激勵機制逐步累積權益，並佔領足夠的委員會席次以發動協同攻擊。一旦攻擊者取得委員會多數席次，現有的委員會架構便無法偵測或防範此類攻擊。為防禦 PCCA，我們提出一種挑戰增強型委員會架構，將安全性與委員會組成解耦：由小型委員會負責例行驗證以提供活性 (Liveness)，而由全域共識支持的挑戰機制提供安全性保證。任何惡意聚合行為都將觸發密碼學驗證、罰沒懲罰，並立即移除惡意驗證者——無論其在委員會中的席次多寡。此機制將安全門檻從委員會多數轉移至全網共識，從而瓦解委員會佔領攻擊。實驗結果顯示，當攻擊發生時，本機制能完全清除惡意委員會成員，而現有最先進的方法則允許攻擊者取得委員會完全控制權並執行不受制衡的攻擊。我們的解耦設計亦允許更小的委員會規模，在不犧牲安全性的前提下提升運算效率。

# ABSTRACT

Keyword: Blockchain, Federated Learning, Committee Capture, Verifier Collusion

Blockchain-based Federated Learning (BCFL) addresses trust and privacy concerns through decentralized consensus. Current BCFL systems rely on committee-based validation assuming honest or honest-majority committees. This assumption is vulnerable to verifier collusion, where attackers accumulate stake to dominate committees. We identify Progressive Committee Capture (PCC), a novel threat where rational attackers exploit incentive mechanisms to gradually accumulate stake and capture sufficient committee seats for coordinated attacks. Existing committee-based architectures cannot detect or prevent such attacks once attackers achieve committee majority. To defend against PCC, we propose an Challenge-Augmented Committee Architecture that decouples security from committee composition: a small committee provides liveness through routine validation, while a challenge mechanism backed by global consensus provides security guarantees. Any malicious aggregation triggers cryptographic verification, slashing penalties, and immediate removal of malicious validators—regardless of their committee representation. This shifts the security threshold from committee majority to global network consensus, neutralizing committee capture attacks. Experimental results demonstrate complete elimination of malicious committee members upon attack attempts, while state-of-the-art approaches allow attackers to achieve full committee control and execute unchecked attacks. Our decoupled design also enables smaller committee sizes, improving computational efficiency without compromising security.

# 誌謝

所有對於研究提供協助之人或機構，作者都可在誌謝中表達感謝之意。

# 目錄

# 圖目錄

# 表目錄

# 第一章　緒論 (Introduction)

隨著人工智慧與分散式運算技術的進步，區塊鏈賦能的聯邦學習 (Blockchain-based Federated Learning, BCFL) 已成為解決多方互不信任情境下協作機器學習的核心技術路徑。在諸如低軌衛星網路 (LEO) [1, 2, 3]、車聯網 (V2X) [4, 5, 6] 以及工業物聯網 (IIoT) [7, 8, 9] 等實際應用場景中，BCFL 展現了其不可替代的重要性。特別是以 LEO 衛星星座為代表的太空 AI 應用場景，星地通訊窗口通常僅約 5 分鐘，且下行頻寬受限於 8Mbps 左右 [2]，使得依賴地面站聚合的傳統模型訓練方案難以實施。BCFL 通過在異質衛星營運商間建立去中心化信任層，成功將收斂時間減少達 30 小時 [3]。同樣地，在工業 4.0 的背景下，BCFL 允許協作工廠在不洩露商業機密的前提下進行預測性維護，實驗資料顯示其通訊開銷可較集中式架構減少約 41% [7]。這些場景共同呈現出「無可信中心」、「資源受限」與「資料高度異質」的特徵，促使 BCFL 成為通用去中心化學習架構的首選方案。

然而，BCFL 在邁向大規模部署時面臨著嚴峻的效率瓶頸，這在業界被稱為「可擴展性兩難」。目前絕大多數 BCFL 系統採用 PBFT (Practical Byzantine Fault Tolerance) 或其變體作為共識機制，其 $O(n^2)$ 的訊息複雜度在節點數增加時會導致效能急劇下降。根據 FLCoin [10] 的實證研究，當參與節點數達到 100 個時，單輪共識產生的訊息量將超過 20,000 條，導致共識延遲攀升至 25 秒以上，此延遲水平已達到模型訓練時間的量級。在極端的車載網路 (VANET) 實測中，100 輛車進行 BCFL 協作會產生 360.57 MB 的巨大資料量，單輪訓練的總通訊開銷高達 19.51 秒 [11]。此外，區塊鏈節點對儲存的高需求 (如比特幣需 200GB，乙太坊超過 465GB) 與邊緣設備 KB 至 MB 級的有限記憶體形成強烈衝突 [12]。這種效能與資源的雙重束縛，使得全節點驗證的傳統架構在實際工業部署中顯得難以維繫。

為了解決上述可擴展性挑戰，學界近年來轉向研究「委員會機制 (Committee Mechanism)」，其核心思想是將驗證責任從全體節點縮減至一組小型驗證者委員會。目前主流的選拔機制包含基於雜湊環的隨機抽樣 [13]、基於幣齡或權益的權重選舉 [14, 10] 以及基於預言機 (VRF) 的 Sortition 機制 [15, 16]。委員會機制的引入立竿見影地改善了系統效能：FLCoin [10] 通過滑動窗口選舉將通訊開銷降低了 90%，並實現了 5.7 倍的訓練加速；BFLC [14] 則利用委員會驗證成功將共識延遲穩定在 3 秒以內。這些最佳化雖成功將通訊複雜度降至與委員會規模 $C$ 相關的 $O(C^2)$ 或 $O(C)$。然而，這種為了效率而進行的「算力與權力集中」也同時引入了新的、尚未被充分探討的安全攻擊面。

最令學界擔憂的危機在於現有委員會防禦機制對「誠實多數假設 (Honest Majority Assumption)」的過度依賴。根據 2024 年針對拜占庭強健聯邦學習的全面調查 [17, 18]，目前超過 93.3% 的 BCFL 研究雖部署了 Krum、Trimmed Mean 或 Median 等防禦演算法，但皆隱含地假設執行這些演算法的實體 (即委員會成員) 是絕對誠實的。現有的威脅模型大多只考慮惡意客戶端上傳毒化梯度，卻忽略了「理性驗證者 (Rational Verifiers)」的危害。最新研究指出，理性對手可以先透過合法行為積累聲譽，一旦在委員中取得超過 33% (針對 BFT 系統) 或 50% (針對一般投票系統) 的主導權，即可輕易繞過所有強健

聚合演算法，甚至偽造聚合結果而不受懲罰。BlockDFL [13] 與 FedBlock [19] 等前沿工作亦坦言，現有機制無法抵禦具備長期策略的委員會共謀攻擊。

上述現象揭示了一個關鍵的「研究缺口 (Research Gap)」：現有 BCFL 缺乏應對「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」的自癒機制。在 PCCA 中，對手並非採取暴力破壞，而是實施「策略性餓死 (Strategic Starvation)」——即在掌控委員會後，優先打包與自身利益相關的更新，並拒絕為誠實參與者提供驗證服務，從而操縱獎勵分配與權益動態。由於缺乏事後的「可追溯審計」與「有效威懾」，一旦誠實多數假設在某一輪次被攻破，系統權力將產生雪崩式的中心化。現有的基於同態加密或權益證明的方案雖然能保護隱私，卻無法在委員會本身已不再可信的情況下，保證模型更新的正確性與資源分配的公平性。如何解耦安全性與共識節點集體信用，成為實現真正去中心化 AI 平台的最後一哩路。

針對這一挑戰，本文提出了一種「挑戰者增強委員會架構 (Challenge-Augmented Committee Architecture, CACA)」，旨在為 BCFL 引入一種全新的安全性保險機制。本研究提出的核心思想是「即時執行、異步審計、罰沒威懾」，這與傳統的「先驗證、後提交」模式有本質區別。我們的主要創新點在於將系統的「活性 (Liveness)」與「安全性 (Security)」進行解耦：即使在委員會不完全可信、甚至被捕獲的情況下，系統仍能通過去中心化的挑戰者網路來檢舉委員會的錯誤決策。具體貢獻概括如下：

- 我們首次定義並模擬量化了漸進式委員會佔領攻擊對 BCFL 長期激勵相容性的破壞力。
- 我們提出了一套基於博弈論設計的「內部罰沒 (Internal Slashing)」協議，確保審計成本低於作惡罰金，從而使得誠實行為成為理性節點的納什均衡。
- 實驗結果顯示，在 30% 惡意共謀的極端環境下，本框架仍能維持超過 98.6% 的模型準確率，並成功將受擊頻率降低約 80%。在 100 節點規模的實驗中，本機制在相同的安全性水平下將日常通訊開銷降低了 44.4%，並將系統最低不可用率從 20% 壓制至 5% 以下。

本論文的組織結構編排如下：第一章為緒論，闡明研究動機、目標與貢獻。第二章介紹聯邦學習、區塊鏈底層架構、拜占庭容錯技術等背景知識，並對現有的去中心化聯邦學習文獻進行分類與批判性評述。第三章定義本研究的系統模型與 PCCA 攻擊者的行為特徵，深入分析其威脅模型。第四章詳細描述 CACA 的具體設計流程、協議設計及安全分析。第五章呈現模擬實驗的參數設定與效能對比結果，驗證所提架構的有效性。第六章對全論文進行總結，並探討本研究在未來的應用前景。

# 第二章　背景知識與相關研究

## 2.1 聯邦式學習基礎 (Federated Learning Fundamentals)

聯邦學習是一種革命性的分散式機器學習典範，其核心創新在於實現「資料不動、模型動」的訓練機制。本節從聯邦學習的起源出發，建立完整的數學框架，深入分析 FedAvg 演算法，並探討其面臨的安全挑戰，為後續章節的區塊鏈整合研究奠定理論基礎。

### 2.1.1 聯邦學習的起源與動機

聯邦學習的概念最早由 Google 於 2017 年正式提出，其動機源於一個關鍵矛盾：現代行動裝置擁有豐富的訓練資料，但這些資料往往具有高度隱私敏感性或資料量龐大，傳統集中式訓練方法不適用 [20]。McMahan 等人在原始論文中明確指出，聯邦學習的設計目標是「將模型訓練與直接存取原始訓練資料的需求解耦」，這體現了**資料最小化原則**（Data Minimization Principle）的核心精神。

**Google Gboard 鍵盤預測**是聯邦學習最具代表性的應用案例。在此應用中，使用者的打字習慣、輸入內容等高度敏感資料完全保留在本地裝置，僅有模型更新以加密形式上傳至雲端進行聚合。Google 報告指出，此系統已部署超過**二十種語言模型**，服務數百萬活躍用戶，實現了下一詞預測、表情符號推薦等功能的持續優化 [20]。值得注意的是，本地訓練僅在裝置閒置、充電中且連接免費 Wi-Fi 時執行，確保對使用者體驗零影響。

聯邦學習與傳統分散式機器學習（如 Parameter Server 架構）存在本質差異。Parameter Server 假設資料為**獨立同分布（IID）**且集中存儲於資料中心，主要目標是透過平行化加速計算。相比之下，聯邦學習面對的是本質上**非獨立同分布（Non-IID）**的資料分布，且資料永遠不離開終端裝置 [20]。McMahan 等人歸納了聯邦優化問題的四大特性：(1) Non-IID：每個使用者的本地資料不代表整體分布；(2) Unbalanced：不同使用者的資料量差異懸殊；(3) Massively Distributed：客戶端數量遠超每個客戶端的樣本數；(4) Limited Communication：裝置經常離線或處於低頻寬環境。這些特性使聯邦學習成為一個獨特的優化問題類別，需要專門的演算法設計 [21]。

從產業背景來看，**資料孤島（Data Silos）**問題日益嚴峻——醫療產業產生全球超過 30% 的資料，但多數資訊仍被鎖在組織邊界內。聯邦學習的出現恰逢歐盟《一般資料保護規則》（GDPR）於 2018 年生效，該法規對違規處理個人資料的企業處以最高全球年營業額 4% 或兩千萬歐元的罰款。聯邦學習「訓練資料不離開裝置」的架構設計，天然符合 GDPR 的同意機制、被遺忘權及資料最小化等要求 [21]。

### 2.1.2 數學框架與優化目標

聯邦學習的優化問題可形式化為一個加權有限和目標函數。設系統中共有 $K$ 個客戶端，第 $k$ 個客戶端擁有 $n_k$ 個訓練樣本，總樣本數為 $n = \sum_{k=1}^{K} n_k$。全域優化目標定義為：

$$\min_{w \in \mathbb{R}^d} F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{2.1}$$

其中 $w \in \mathbb{R}^d$ 為 $d$ 維模型參數，$F_k(w)$ 為第 $k$ 個客戶端的本地目標函數 [20]。本地目標函數定義為該客戶端資料上的經驗風險：

$$F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} \ell(w; x_i, y_i) \tag{2.2}$$

其中 $\mathcal{P}_k$ 為客戶端 $k$ 持有的資料索引集合，$\ell(w; x_i, y_i)$ 為模型在樣本 $(x_i, y_i)$ 上的損失函數。

**加權係數 $p_k = n_k/n$ 的理論依據**在於確保每個訓練樣本對全域目標的貢獻相等，無論該樣本位於哪個客戶端。若資料分布滿足 IID 假設（即 $\mathcal{P}_k$ 為從總體資料隨機均勻抽樣形成），則有 $\mathbb{E}_{\mathcal{P}_k}[F_k(w)] = F(w)$，此時本地優化等價於全域優化 [22]。

然而實際應用中，**Non-IID 資料分布**才是常態。Kairouz 等人 [21] 系統性地歸納了五種 Non-IID 類型：(1) 標籤分布偏斜 $P(y)$：不同客戶端的類別比例不同；(2) 特徵分布偏斜 $P(x)$：相同標籤下的特徵分布差異；(3) 相同標籤不同特徵 $P(x|y)$：如不同地區的手寫風格差異；(4) 相同特徵不同標籤 $P(y|x)$：標註偏好差異；(5) 數量偏斜：各客戶端 $n_k$ 差異懸殊。

為量化資料異質性程度，Li 等人 [22] 引入**異質性度量** $\Gamma$：

$$\Gamma = F^* - \sum_{k=1}^{K} p_k F_k^* \tag{2.3}$$

其中 $F^*$ 和 $F_k^*$ 分別為全域和本地目標函數的最小值。當資料為 IID 時，$\Gamma \to 0$；Non-IID 程度越高，$\Gamma$ 越大，此參數直接影響收斂速度。另一常用度量為**有界散度假設**（Bounded Dissimilarity）：$\mathbb{E}_k[\|\nabla F_k(w)\|^2] \leq B^2 \|\nabla F(w)\|^2$，參數 $B$ 反映本地梯度與全域梯度的偏離程度 [22]。

### 2.1.3 FedAvg 演算法詳解

**FederatedAveraging（FedAvg）**演算法是聯邦學習最基礎且應用最廣泛的優化方法 [20]。其核心思想是讓選定的客戶端在本地執行多步隨機梯度下降（SGD），再由伺服器聚合各客戶端的模型更新。完整演算法如下：

關鍵超參數包括：客戶端選取比例 $C$、本地訓練週期數 $E$、批次大小 $B$、學習率

**Algorithm 1** FederatedAveraging (FedAvg)

**Server executes:**
Initialize global model $w_0$
**for** each round $t = 1, 2, \dots$ **do**
  $m \leftarrow \max(C \cdot K, 1)$ ▷ Select a fraction $C$ of clients
  $S_t \leftarrow$ (randomly select $m$ clients)
  **for** each client $k \in S_t$ **in parallel do**
    $w_k^{t+1} \leftarrow \text{ClientUpdate}(k, w_t)$
  **end for**
  $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{n} w_k^{t+1}$ ▷ Weighted aggregation
**end for**

**function** ClientUpdate($k, w$)
  $\mathcal{B} \leftarrow$ (split local data $\mathcal{P}_k$ into batches of size $B$)
  **for** each local epoch $i = 1, \dots, E$ **do**
    **for** each batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
    **end for**
  **end for**
  **return** $w$ to server
**end function**

$\eta$。McMahan 等人 [20] 的實驗表明，**較小的 $B$ 配合較大的 $E$ 能顯著減少通訊輪數**：在 MNIST 資料集上，設定 $B = 10, E = 20$ 達到 99% 準確率僅需 **18 輪通訊**，相比基準 FedSGD 的 626 輪實現了 **34.8 倍加速**。在 CIFAR-10 實驗中，FedAvg 以 2,000 輪達到 85% 準確率，而標準 SGD 需要 99,000 步，通訊成本降低約 **50 倍** [20]。

**收斂性保證**需要以下標準假設 [22]：(1) $L$-Lipschitz 平滑性：$F_k(v) \leq F_k(w) + \nabla F_k(w)^T(v-w) + \frac{L}{2}\|v-w\|^2$；(2) $\mu$-強凸性；(3) 有界變異數：$\mathbb{E}[\|\nabla F_k(w,\xi) - \nabla F_k(w)\|^2] \leq \sigma_k^2$；(4) 有界梯度：$\mathbb{E}[\|\nabla F_k(w,\xi)\|^2] \leq G^2$。在這些假設下，FedAvg 的收斂速率為 $O(1/T)$，但收斂上界包含異質性項 $\Gamma$ 和本地訓練相關項 $(E-1)^2 G^2$ [22]。

值得注意的是，當 $E > 1$ 時，**學習率必須衰減**才能保證收斂至最優解。Li 等人 [22] 證明若使用固定學習率 $\eta$，最終解與最優解的距離為 $\Omega(\eta(E-1))$。此外，Non-IID 環境下 FedAvg 無法實現與客戶端數量成正比的線性加速，這是其主要理論限制。

### 2.1.4 安全與隱私挑戰

儘管聯邦學習的設計理念是保護資料隱私，但其分散式架構引入了新的安全威脅面。這些威脅可分為**完整性攻擊**（破壞模型效能）和**隱私攻擊**（竊取訓練資料）兩大類 [21]。

**拜占庭攻擊（Byzantine Attacks）**是完整性攻擊的核心威脅。Blanchard 等人 [23] 首次形式化此問題：在 $n$ 個參與者中，最多有 $f$ 個惡意參與者可發送任意更新值。典型攻擊包括**標籤翻轉攻擊**（Label Flipping）：惡意客戶端將本地資料標籤從源類別修改為目標類別；以及**模型投毒攻擊**（Model Poisoning）：直接操縱本地模型參數或梯度。後者的威力顯著更強——Bagdasaryan 等人證明單一惡意參與者可在**一輪內達到 100% 後門任務準確率**，且此攻擊無法被安全聚合機制偵測。

**梯度洩漏攻擊**揭示了聯邦學習中「僅分享梯度」並不能完全保證隱私。Zhu 等人 [24] 提出的 Deep Leakage from Gradients（DLG）攻擊展示了驚人的隱私風險：透過優化隨機初始化的虛擬資料，使其產生的梯度逼近真實梯度，即可**像素級精確還原原始訓練影像**，甚至可逐字元還原文本資料。攻擊僅需約 300 次 L-BFGS 迭代，當批次大小為 1 且影像解析度較低時效果最佳。後續研究進一步提升了攻擊效能，在 64×64 影像上可達 **PSNR > 30 dB** 的高保真還原 [24]。

**Non-IID 資料分布加劇了拜占庭防禦的困難**。傳統防禦方法（如 Krum [23]、Trimmed Mean、Coordinate-wise Median）假設誠實客戶端的更新會聚集在一起，而惡意更新為離群值。然而在 Non-IID 環境下，由於各客戶端本地資料分布差異大，誠實更新本身就呈現高度發散，使得離群值檢測方法失效。研究表明，現有拜占庭容錯方法在極端 Non-IID 情境下可能被完全突破，導致全域模型崩潰。這一觀察直接連結至本研究第三章將探討的威脅模型與防禦機制設計。

# 2.2 區塊鏈聯邦式學習 (Blockchain-based Federated Learning, BCFL)

區塊鏈聯邦式學習透過結合分散式帳本技術與聯邦學習架構, 從根本上解決了傳統聯邦學習的**信任集中化困境**。本節將系統性地闡述 BCFL 的技術動機、架構演進脈絡、委員會共識機制設計, 以及現有方案在委員會安全性上的不足——這些不足正是本研究欲填補的關鍵缺口。

## 2.2.1 BCFL 的動機: 解決信任問題

### 2.2.1.1 中央化架構的脆弱性

傳統聯邦學習雖聲稱「資料不出本地」, 但其架構核心仍仰賴單一中央聚合器, 這導致三類根本性的信任風險。**第一, 聚合器惡意行為風險**: 中央伺服器可執行選擇性聚合 (僅納入特定客戶端更新)、結果篡改 (植入後門或偏差模型), 甚至透過梯度推論攻擊重建原始訓練資料——NeurIPS 2020 研究顯示, 即便是 ImageNet 等級的 ResNet 模型, 攻擊者仍可從梯度中重建訓練影像 [25]。**第二, 單點故障 (SPOF)**: 當中央伺服器因攻擊、故障或網路問題而離線, 整體訓練流程即刻癱瘓, 且缺乏有效的復原機制。**第三, 拜占庭將軍問題**: 惡意客戶端可注入中毒模型更新, 而中央聚合器本身亦可能與攻擊者共謀;USENIX Security 2020 研究指出, 針對性的模型中毒攻擊可使全域模型錯誤率提升達 **90%** [26]。

### 2.2.1.2 區塊鏈特性與信任問題的對應

區塊鏈的三大核心特性恰好對應中央化聯邦學習的信任困境: **不可篡改性**確保一旦聚合結果上鏈即無法竄改, 任何篡改企圖將導致雜湊不符; **透明性**使所有提交的更新與

表 2.1: 中央化聯邦學習的信任風險

| 風險類型 | 描述 | 影響 | 傳統方案局限 |
|---|---|---|---|
| 選擇性聚合 | 伺服器選擇性納入/排除客戶端更新 | 模型偏差、貢獻浪費 | 無法驗證伺服器決策 |
| 結果篡改 | 聚合器修改全域模型 | 後門植入、效能劣化 | 客戶端無法驗證聚合正確性 |
| 梯度推論攻擊 | 從梯度推斷私有資料 | 成員推論、資料重建 | 差分隱私降低準確度 |
| 單點故障 | 中央伺服器不可用 | 訓練中斷、進度遺失 | 冗餘伺服器引入新信任問題 |
| 拜占庭攻擊 | 惡意節點發送中毒更新 | 模型準確度下降達 90% | 穩健聚合於惡意比例 >50% 時失效 |

聚合邏輯對全體參與者可見, 選擇標準編碼於智能合約中; **去中心化**則消除對單一實體的信任依賴, 透過共識機制確保系統持續運作。

表 2.2: 區塊鏈特性對應聯邦學習信任問題

| FL 信任問題 | 對應區塊鏈特性 | 解決機制 |
|---|---|---|
| 結果篡改 | 不可篡改性 | 聚合模型雜湊上鏈, 客戶端使用前驗證 |
| 選擇性聚合 | 透明性 | 智能合約定義可驗證的選擇規則 |
| 單點故障 | 去中心化 | 多節點維護系統狀態,P2P 模型聚合 |
| 缺乏可審計性 | 不可篡改性 + 透明性 | 完整訓練歷程記錄於鏈上 |

BCFL 的三項核心優勢因而顯現: **消除單點故障**——區塊鏈以分散式網路取代中央聚合器, 任一節點失效時其他節點可無縫接續, 研究顯示 BCFL 在節點故障情況下仍可維持 90% 以上準確度; **防篡改的可審計性**——不可變帳本建立永久可驗證的操作紀錄, 智能合約強制執行確定性聚合規則; **拜占庭容錯的激勵對齊**——結合 BFT 共識協議與加密貨幣獎勵機制, 可容忍最多 $f < n/3$ 的惡意節點, 同時透過押金機制懲罰惡意行為。

## 2.2.2 BCFL 架構演進

### 2.2.2.1 演進時間線

BCFL 架構經歷了從「全節點共識」到「委員會共識」的關鍵演進。**2018-2020 年的基礎期**以 BlockFL [27] 為代表, 採用工作量證明 (PoW) 作為共識機制, 礦工驗證本地模型更新後打包區塊, 消除了對中央伺服器的依賴; 然而 PoW 的高能耗與長共識延遲使其難以適用於資源受限的邊緣運算場景。**2020-2021 年的委員會共識興起期**以 BFLC [14] 為里程碑, 引入委託式委員會共識, 將共識運算從全網 $O(n^2)$ 降至委員會內 $O(C^2)$, 並採用 K-fold 交叉驗證檢測惡意更新;Lu et al. [7] 更提出將 FL 訓練品質融入共識的「Proof of Training Quality」。**2023-2024 年的優化期**則以 FLCoin [28] 的滑動視窗委員會與 BlockDFL [13] 的雙層評分機制為代表, 前者將通訊複雜度進一步降至線性 $O(s)$, 後者則將拜占庭容忍度提升至 **40%**。

#### 2.2.2.2 代表性系統比較表

表 2.3: 代表性 BCFL 系統比較

| 系統 | 年份 | 共識機制 | 通訊複雜度 | 容錯能力 | 主要創新 | 主 |
|---|---|---|---|---|---|---|
| BlockFL [27] | 2020 | PoW | 依 PoW 難度 | 50% 算力 | 首個 BCFL 框架、延遲模型分析 | 高 |
| BFLC [14] | 2021 | 委員會共識 | $O(C^2)$ | 33% (3f+1) | 委員會共識、K-fold 驗證 | 委 |
| Lu et al. [7] | 2020 | Proof of Training Quality | IIoT 優化 | 33% | 共識與訓練整合 | 需 |
| BlockDFL [13] | 2024 | PBFT 投票 | $O(\text{agg} \times \text{ver})$ | **40%** | 雙層評分、梯度壓縮 | 取 |
| FLCoin [28] | 2024 | 滑動視窗委員會 | $O(s)$ 線性 | <25% | 90% 通訊降低、35% 訓練加速 | 被 |

#### 2.2.2.3 從全節點到委員會的演進動力

全節點共識的核心瓶頸在於**通訊複雜度與可擴展性的矛盾**: 傳統 PBFT 需 $O(n^2)$ 訊息交換, 當節點數達數百時, 共識延遲將嚴重拖累 FL 訓練週期。委員會機制透過選取規模 $C \ll n$ 的代表節點執行共識, 將複雜度降至 $O(C^2)$ 甚至 $O(C)$。FLCoin [28] 實驗顯示, 當 $n = 500$ 時, 採用 $s = 100$ 的滑動視窗可將通訊負擔降低 **90%**, 共識延遲維持在 **3 秒以內**, 單輪迭代時間約 7 秒 (對比 Biscotti 的 40 秒)。這種架構轉變的代價是引入了新的安全假設: 委員會的安全性取決於其組成是否被惡意節點控制。

### 2.2.3 委員會機制的技術細節

#### 2.2.3.1 委員會選擇機制比較

委員會選擇方法直接影響系統的安全性與公平性, 現有方案可分為四類:

表 2.4: 委員會選擇機制比較

| 選擇方法 | 優點 | 缺點 | 代表系統 |
|---|---|---|---|
| **隨機選擇** | 不可預測、防止針對性攻擊 | 可能選到惡意或低品質節點 | BFLC, RapidChain |
| **權益導向** | Sybil 抵抗、經濟安全保障 | 中心化風險 (富者愈富) | Ethereum 2.0, DPoS-based FL |
| **聲譽導向** | 獎勵可靠貢獻者、過濾惡意節點 | 聲譽壟斷、易被長期培養攻擊 | BESIFL, PoQ-based BCFL |
| **貢獻導向** | 與 FL 目標直接對齊 | 新加入者劣勢、指標可被博弈 | FLCoin |

#### 2.2.3.2 委員會大小的安全性分析

委員會安全性依循**超幾何分佈**建模。當從 $n$ 個節點 (含 $m$ 個惡意節點) 中選取 $C$ 個組成委員會時, 恰有 $k$ 個惡意節點被選中的機率為:

$$P(X = k) = \frac{\binom{m}{k}\binom{n-m}{C-k}}{\binom{n}{C}} \tag{2.4}$$

委員會被攻破的機率 (惡意節點超過 BFT 閾值 $\beta = 1/3$) 為:

$$P(\text{compromised}) = \sum_{k=\lceil \beta C \rceil}^{\min(m,C)} P(X = k) \tag{2.5}$$

**數值範例**: 設 $n = 100$、$m = 30$(30% 惡意)、$C = 10$, 計算惡意節點超過 $\frac{1}{3}$ 的機率約為 **3.88%**。此風險對生產系統而言過高。FLCoin [28] 實驗顯示, 將委員會規模提升至 $C = 50$ 可達 **91.3%** 安全機率,$C = 100$ 則達 **98.4%**, 但通訊負擔亦隨之增加——這正是委員會機制的核心權衡。

### 2.2.4 智能合約在 BCFL 中的角色

#### 2.2.4.1 功能模組

BCFL 中的智能合約承擔四項核心功能: **註冊模組**管理客戶端登入, 驗證資格、記錄錢包地址與訓練能力、收取押金, 並維護經驗證的參與者清單; **聚合模組**協調模型更新的彙整, 管理輪次同步、收集模型雜湊、執行聚合演算法 (如 FedAvg), 並儲存結果; **驗證模組**在聚合前驗證模型更新, 評審者以測試資料集評估提交模型, 計算貢獻分數並過濾惡意更新; **獎勵模組**依據貢獻分配激勵, 透過 ERC-20 代幣自動轉帳, 並對惡意行為執行押金沒收。

#### 2.2.4.2 鏈上 vs 鏈下聚合比較

表 2.5: 鏈上與鏈下聚合比較

| 方式 | 優點 | 缺點 | 適用場景 |
|---|---|---|---|
| **鏈上聚合** | 完全透明、確定性執行、防篡改 | Gas 成本極高、擴展性受限 | 小型模型、高審計需求應用 |
| **鏈下聚合** | 低成本、支援大型模型、高效 | 需信任聚合器、透明度較低 | 大規模 FL、生產環境部署 |

研究顯示, 百萬參數級模型的鏈上聚合 Gas 成本將達數百萬單位, 遠超實用範圍。主流方案採用**混合架構**: 鏈下執行聚合運算, 僅將模型雜湊 (約 32 bytes) 上鏈存證, 實際模型存於 IPFS。

#### 2.2.4.3 獎勵機制設計

理論基礎最紮實的獎勵機制採用合作賽局論的 **Shapley 值**。節點 $i$ 的 Shapley 值定義為:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{2.6}$$

其中 $v(S)$ 為聯盟 $S$ 的效用函數 (通常為驗證集準確度)。FedCoin [29] 實作 Proof-of-Shapley 協議, 依據各節點對模型改進的邊際貢獻分配獎勵:Reward$_i$ $= (\phi_i / \sum_j \phi_j) \times$ TotalBudget, 確保公平性、預算平衡與零貢獻零報酬。

### 2.2.5 FedBlock 指出的委員會機制弱點 (Research Gap)

現有委員會共識機制存在若干尚未解決的關鍵弱點, 構成本研究的重要切入點。FedBlock (Nguyen et al., 2024) [19] 在其未來展望章節明確指出:「目前版本中, 智能合約以隨機方式選取客戶端作為驗證者, **但此選擇標準可能並非最佳**」; 更進一步指出「驗證者本身亦可能是惡意的, 結果可能在通訊中遺失, 或參與驗證者不足, 導致客戶端收到錯誤或缺失的驗證分數」, 且「FedBlock 要在實務上可用, **需要一套激勵機制來鼓勵誠實驗證者的參與**」。

BFLC [14] 的委員會滲透弱點更為關鍵。2024 年 Taylor & Francis 的文獻回顧明確指出:「BFLC 採用委員會共識, **然而它容易被惡意節點混入委員會, 從而導致系統偏差**」[30]。FLCoin 的分析亦顯示, 即便採用 $s = 50$ 的滑動視窗, 仍有 **8.7%** 的機率選出不安全委員會。

綜合現有文獻, 委員會機制的核心研究缺口包括: **缺乏針對委員會滲透的穩健防禦**——現有隨機或貢獻導向選擇易被博弈; **缺乏委員會操縱攻擊的形式化安全分析**; **缺乏考量歷史行為、模型品質與拜占庭抵抗能力的適應性選擇機制**; 以及**缺乏可證明阻止惡意驗證者行為的激勵相容驗證機制**。這些缺口為本研究後續章節提出的方法論奠定了基礎。

## 2.3 拜占庭容錯機制

分散式系統在面對節點故障時，必須具備持續運作的能力。然而，當故障不僅止於節點停機，而是節點可能展現任意惡意行為——包括發送矛盾訊息、選擇性沉默或與其他惡意節點協同攻擊——系統便需要更強韌的容錯機制。在區塊鏈驅動的聯邦學習系統中，負責模型聚合的節點若遭攻陷，可能產生錯誤的聚合結果並試圖將其寫入區塊鏈，此類威脅本質上即屬於拜占庭故障。因此，理解拜占庭容錯的基本原理與安全性閾值，是設計可信賴 BCFL 架構的必要前提。

本節首先介紹拜占庭將軍問題的理論基礎，建立容錯閾值的數學基礎；接著說明實用拜占庭容錯協議（PBFT）的運作原理，作為後續挑戰機制設計的安全性後盾；最

後探討委員會架構如何在 BCFL 領域應用，並指出現有方法在面對長期攻擊時的根本性侷限。

## 2.3.1 拜占庭將軍問題與容錯閾值

### 2.3.1.1 問題的起源與形式化定義

拜占庭將軍問題由 Lamport、Shostak 與 Pease 於 1982 年正式提出 [31]。問題的設定源自一個軍事隱喻：拜占庭帝國的數支軍隊包圍敵城，各軍由一位將軍指揮，將軍們僅能透過信使相互通訊。然而，部分將軍可能是叛徒，他們會刻意傳遞錯誤訊息以阻撓忠誠將軍達成一致決策。問題的核心在於：如何設計一個演算法，使得所有忠誠將軍能就「進攻」或「撤退」達成共識，即使存在叛徒試圖破壞協調？

此問題的形式化定義包含兩個交互一致性條件（Interactive Consistency Conditions）：

- **IC1（一致性）**：所有忠誠副官必須執行相同的命令。
- **IC2（正確性）**：若指揮官是忠誠的，則所有忠誠副官必須執行指揮官發出的命令。

這兩個條件共同確保了分散式系統的基本安全性：忠誠節點不會因惡意節點的干擾而產生分歧，且正確的輸入能夠被正確地傳播。將此問題對應到區塊鏈聯邦學習的場景：將軍對應驗證節點，信使對應網路通訊，叛徒對應被攻陷或惡意的聚合器與驗證者。

### 2.3.1.2 三分之一閾值的不可能性證明

拜占庭將軍問題存在一個根本性的數學限制：在僅使用口頭訊息（Oral Messages）的情況下，問題可解若且唯若超過三分之二的參與者是忠誠的。換言之，若系統中有 $n$ 個節點，最多只能容忍 $f$ 個拜占庭節點，其中 $n \geq 3f + 1$。

此限制可透過最簡單的三節點、一叛徒場景直觀理解。考慮以下兩種情境：

**情境一**：指揮官是忠誠的，發送「進攻」命令給兩位副官 A 與 B。副官 B 是叛徒，他向副官 A 謊稱「指揮官說撤退」。此時副官 A 收到兩個矛盾訊息：來自指揮官的「進攻」與來自 B 轉述的「撤退」。

**情境二**：指揮官是叛徒，他向副官 A 發送「進攻」，向副官 B 發送「撤退」。兩位忠誠副官如實向對方轉述各自收到的命令。副官 A 同樣收到兩個矛盾訊息：來自指揮官的「進攻」與來自 B 轉述的「撤退」。

關鍵的洞察在於：從副官 A 的視角來看，情境一與情境二完全無法區分。若叛徒能夠一致地說謊，副官 A 便無法判斷叛徒究竟是指揮官還是另一位副官。任何確定性演算法在此情境下都必然失敗，這從根本上限制了拜占庭容錯系統的設計空間。

此 $n \geq 3f + 1$ 的閾值源於一個簡單的算術事實：當需要確認某個值是否正確時，系統必須能夠獲得足夠多的一致回應以排除惡意節點的干擾。若 $f$ 個節點可能說謊，

則需要至少 $2f+1$ 個節點的確認才能確保至少 $f+1$ 個回應來自誠實節點。由於總節點數必須包含這 $2f+1$ 個確認節點加上 $f$ 個可能的惡意節點，故 $n \geq 3f+1$。

#### 2.3.1.3 故障模型的層級區分

在分散式系統的容錯理論中，故障模型依嚴重程度可分為多個層級。表 2.6 比較了兩種最常見的故障模型——崩潰故障與拜占庭故障——の特性差異。

表 2.6: 故障模型比較：崩潰故障 vs. 拜占庭故障

| 特性 | 崩潰故障 | 拜占庭故障 |
|---|---|---|
| 行為特徵 | 節點停止運作後不再發送訊息 | 節點可能展現任意行為，包括發送矛盾訊息 |
| 可偵測性 | 可透過心跳逾時機制偵測 | 無法透過簡單機制直接偵測 |
| 容錯閾值 | $f < n/2$（少數服從多數） | $f < n/3$（需要超額多數） |
| 典型協議 | Paxos [32]、Raft [33] | PBFT [34]、HotStuff [35]、Tendermint [36] |
| 應用場景 | 可信環境中的分散式資料庫 | 開放或半開放環境的區塊鏈系統 |

崩潰故障假設節點一旦故障便完全停止運作，這在資料中心等可控環境中是合理的假設。然而，在區塊鏈聯邦學習的場景中，節點可能被攻擊者控制而展現惡意行為，單純的崩潰容錯機制無法提供足夠的安全保證。拜占庭容錯需要更嚴格的 $n \geq 3f+1$ 閾值，正是因為惡意節點可能向不同節點發送不同訊息，使得僅憑簡單多數決無法辨別真偽。

### 2.3.2 實用拜占庭容錯協議 (PBFT)

#### 2.3.2.1 PBFT 的設計動機與定位

拜占庭將軍問題的理論解法雖然在 1982 年即已提出，但早期解法的指數級通訊複雜度使其僅具理論意義。1999 年，Castro 與 Liskov 提出實用拜占庭容錯協議（Practical Byzantine Fault Tolerance, PBFT）[34]，首次將 BFT 共識的通訊複雜度降至多項式級別 $O(n^2)$，使其在實際系統中可行。PBFT 的設計目標是在部分同步網路模型下，以合理的效能代價換取對任意惡意行為的容錯能力。

在本研究的架構中，PBFT 扮演的角色是挑戰機制觸發時的安全性後盾。當系統偵測到可疑的聚合行為並發起挑戰時，需要一個能夠抵禦拜占庭攻擊的共識機制來仲裁爭議。PBFT 的成熟理論基礎與經過驗證的安全性保證，使其成為此角色的理想選擇。

#### 2.3.2.2 三階段協議流程

PBFT 協議需要 $n = 3f + 1$ 個副本節點（Replica），其中一個被指定為主節點（Primary），負責為客戶端請求分配序號並發起共識。協議透過三個階段達成共識：

**階段一：Pre-prepare（預準備）**主節點 $p$ 收到客戶端請求 $m$ 後，為其分配一個序號 $n$，並向所有副本廣播預準備訊息：$\langle \text{PRE-PREPARE}, v, n, D(m) \rangle_{\sigma_p}$ 其中 $v$ 為當前視圖編號，$D(m)$ 為請求的摘要，$\sigma_p$ 為主節點的數位簽章。此階段的通訊複雜度為 $O(n)$，因為主節點僅需向所有副本發送一次訊息。

**階段二：Prepare（準備）**副本節點收到預準備訊息後，驗證其有效性：視圖編號是否正確、序號是否在有效範圍內、是否已接受過相同序號但不同摘要的訊息。驗證通過後，副本向所有其他節點廣播準備訊息：$\langle \text{PREPARE}, v, n, d, i \rangle_{\sigma_i}$ 當某節點收集到 $2f$ 個來自不同節點的匹配準備訊息（加上自身持有的預準備訊息，共 $2f + 1$ 票），即進入 prepared 狀態。此階段每個節點都向其他所有節點發送訊息，通訊複雜度為 $O(n^2)$。

**階段三：Commit（提交）**進入 prepared 狀態的副本向所有節點廣播提交訊息：$\langle \text{COMMIT}, v, n, d, i \rangle_{\sigma_i}$ 當節點收集到 $2f + 1$ 個匹配的提交訊息，即進入 committed-local 狀態，可執行請求並回覆客戶端。此階段的通訊複雜度同樣為 $O(n^2)$。

#### 2.3.2.3 通訊複雜度分析

PBFT 的總通訊複雜度由三個階段累加：

$$\text{總複雜度} = O(n) + O(n^2) + O(n^2) = O(n^2) \tag{2.7}$$

以 $n = 4$、$f = 1$ 的最小配置為例：預準備階段產生 3 則訊息，準備階段產生 $4 \times 3 = 12$ 則訊息，提交階段同樣產生 12 則訊息，每輪共識總計 27 則訊息。當節點數增加至 $n = 21$（可容忍 7 個拜占庭節點）時，每輪訊息數增至約 870 則。

此二次方複雜度是 PBFT 的主要效能瓶頸，限制了其在大規模網路中的應用。實務上，傳統 PBFT 部署通常限制在 10 至 20 個節點。然而，在本研究的目標場景——許可制聯盟鏈環境下的聯邦學習——驗證者數量通常在此範圍內，PBFT 的通訊成本是可接受的。更重要的是，本研究採用的樂觀執行機制使得 PBFT 僅在挑戰發生時才被觸發，大幅降低了實際的平均通訊開銷。

#### 2.3.2.4 視圖更換與活性保證

PBFT 的安全性（Safety）保證所有誠實節點對請求序列達成一致，其核心依據是 Quorum 交叉原理。任意兩個大小為 $2f + 1$ 的節點群體，其交集至少包含：

$$\text{交集大小} = (2f + 1) + (2f + 1) - (3f + 1) = f + 1 \tag{2.8}$$

由於最多 $f$ 個節點為拜占庭節點，交集中必包含至少一個誠實節點。這確保了任何兩個 prepared 狀態的決策必然一致，因為它們共享至少一個誠實見證者。

活性（Liveness）保證客戶端請求最終會被執行，這依賴視圖更換（View Change）機制。當副本偵測到主節點故障（例如逾時未收到預準備訊息），將發起視圖更換，選舉新的主節點。新主節點 $p' = (v+1) \mod n$ 需收集 $2f+1$ 個視圖更換訊息，確認先前視圖中已達成的共識狀態，然後在新視圖中繼續處理請求。

視圖更換機制確保了即使主節點被攻陷或離線，系統仍能持續運作。由於拜占庭節點至多 $f$ 個，連續 $f$ 次視圖更換後必然會選出誠實的主節點，系統活性得以恢復。

#### 2.3.2.5 PBFT 的後續發展與本研究的選擇

PBFT 提出後，研究者針對其 $O(n^2)$ 通訊複雜度提出了多種改進方案。HotStuff [35] 透過流水線化的三階段協議與閾值簽章技術，將通訊複雜度降至 $O(n)$，已被 Meta 的 Diem 區塊鏈採用作為共識基礎。Tendermint [36] 將 PBFT 與權益證明機制結合，成為 Cosmos 生態系統的共識協議。Algorand [37] 則透過可驗證隨機函數（VRF）實現無需許可的委員會選舉，在公鏈環境下達成可擴展的拜占庭共識。

然而，這些改進主要針對共識協議本身的效率優化，而非應用層的安全機制設計。本研究的創新聚焦於「何時需要觸發共識」以及「如何設計挑戰機制」，而非「如何優化共識協議」。在本研究的目標場景中，驗證者數量通常在 10 至 20 個範圍內，PBFT 的通訊成本（每輪約數百則訊息）遠低於模型更新的傳輸成本（單個模型更新可達數 MB 至數百 MB），共識效率並非系統瓶頸。

本研究選擇 PBFT 作為挑戰機制的共識協議，基於三個考量：首先，PBFT 的安全性證明經過二十餘年的學術驗證，其理論基礎穩固，便於進行形式化的安全性分析。其次，本研究的框架設計採用模組化架構，挑戰機制與底層共識協議解耦，若未來部署於更大規模的網路，可將 PBFT 替換為 HotStuff 或其他改進協議，而無需修改上層機制。第三，PBFT 支援任意計算的驗證，不受零知識證明或詐欺證明所需的算術電路限制，這對於需要支援多種聚合演算法（如 Krum、FedProx、Median）的聯邦學習場景尤為重要。

### 2.3.3 委員會架構與區塊鏈聯邦學習

#### 2.3.3.1 從全節點共識到委員會機制的演進

傳統 BFT 協議要求所有節點參與每一輪共識，導致通訊成本隨節點數平方增長。當區塊鏈系統需要支援數百甚至數千個節點時，此設計成為不可逾越的效能瓶頸。委員會架構（Committee-based Architecture）的核心理念是將共識責任委派給一個小型代表性子集，由委員會代替全網執行共識協議。

委員會架構的通訊成本可表示為：

$$\text{總通訊成本} = O(c^2) + O(n) \tag{2.9}$$

其中 $c$ 為委員會大小，$n$ 為全網節點數。當 $c \ll n$ 時，此成本遠低於全節點 PBFT 的 $O(n^2)$。委員會內部執行 BFT 共識的成本為 $O(c^2)$，委員會與全網的結果廣播成本為 $O(n)$。

#### 2.3.3.2 BCFL 中的委員會共識應用

委員會架構已被廣泛應用於區塊鏈聯邦學習系統。Li 等人提出的 BFLC 框架 [14] 是最早將委員會共識引入 BCFL 的研究之一。在 BFLC 中，系統從全體參與者中選出一個委員會，負責驗證客戶端提交的模型更新並執行聚合。委員會成員使用自身資料集對更新進行交叉驗證，計算品質分數後透過委員會共識決定是否接受。實驗結果顯示，當委員會大小為 5 時，相較於全節點 PBFT（$n = 20$），共識延遲從 120 毫秒降至 35 毫秒，通訊成本降低約 85%。

BlockDFL [13] 進一步將委員會機制與角色輪替結合。系統根據上一區塊的雜湊值，將參與者隨機分配為三種角色：更新提供者（Update Provider）負責本地訓練、聚合器（Aggregator）負責模型聚合、驗證者（Verifier）組成委員會執行共識。論文建議驗證者數量應「遠小於」總參與者數量以提升效率，實驗配置中使用 4 至 7 個驗證者處理 20 至 60 個參與者的系統。此設計使得 BlockDFL 在處理 166 萬參數的模型時，聚合與驗證時間低於 3 秒，相較於類似系統 Biscotti [15] 處理 7,850 參數需超過 30 秒，效能提升顯著。

FLCoin [28] 採用滑動視窗機制選舉委員會：節點透過提交有效的模型更新獲得「份額」，在固定大小的滑動視窗（通常設為 50 至 100）內持有份額的節點組成當輪委員會。此設計使得通訊複雜度維持在線性 $O(n)$，相較於傳統 PBFT 減少約 90% 的通訊開銷。論文實驗顯示，即使在 500 個節點的規模下，共識延遲仍低於 3 秒。

#### 2.3.3.3 委員會架構的根本性安全隱患

儘管委員會架構顯著提升了 BCFL 系統的效率，其安全性卻建立在一個脆弱的假設之上：委員會成員的誠實多數。BlockDFL 明確指出，其共識機制「僅在超過三分之二的驗證者是誠實的情況下才能產生非空區塊」[13]。當委員會規模較小時，此假設尤其危險。

以 BlockDFL 的典型配置（7 個驗證者）為例，攻擊者若能控制 5 個驗證者，即可掌握超過三分之二的投票權，從而通過任意惡意的聚合結果。即使攻擊者在全網僅佔 30% 的節點，在隨機抽取 7 個驗證者的過程中，攻擊者佔據 5 席以上的機率雖低，但絕非為零。更重要的是，攻擊者可採用「等待策略」：平時潛伏並表現誠實以累積權益（Stake），僅在他們控制的節點「中獎」成為委員會多數時才發動攻擊。在這種情況下，

由於共識僅在小型委員會內部達成，攻擊者可直接操控投票結果，繞過所有資料層防禦機制。

此問題的根源在於委員會架構將「效率」與「安全性」綁定在同一個元件上：委員會既負責提供系統活性（持續處理更新），也負責提供安全性保證（驗證更新正確性）。當委員會被攻陷時，兩者同時失效。

#### 2.3.3.4 漸進式委員會佔領攻擊

現有委員會架構面臨的更深層威脅是漸進式佔領攻擊（Progressive Committee Capture Attack）。此攻擊利用權益累積機制的正回饋特性，透過兩階段策略逐步控制委員會：

**潛伏階段**：攻擊者控制的節點在初期表現完全誠實，正常參與訓練並提交高品質的模型更新。透過持續的誠實行為，攻擊節點累積權益與聲譽，提高被選入委員會的機率。

**佔領階段**：當攻擊節點首次在某一輪佔據委員會多數席位時，他們可以操控共識結果，將獎勵僅分配給自己控制的節點，同時拒絕誠實節點的更新。由於委員會選舉通常基於權益或聲譽，此操作使攻擊者的相對權益份額持續增加，進一步提高其在未來輪次佔據委員會多數的機率，形成自我強化的惡性循環。

此攻擊的危險性在於其隱蔽性：攻擊者無需在任何時刻控制全網多數節點，僅需耐心等待並利用概率波動。一旦成功佔領委員會，現有系統缺乏有效的偵測與清除機制。BlockDFL 指出惡意領導者「只能拒絕投票並廣播空區塊以延遲迭代」[13]，但未分析多個被攻陷驗證者協同作惡的場景。當被攻陷的委員會成為「合法」權威時，系統無法區分正當權威與被佔領的權威。

#### 2.3.3.5 現有方法的侷限性總結

綜合以上分析，現有 BCFL 委員會架構存在三個根本性侷限：

第一，**誠實多數假設的脆弱性**。無論採用隨機抽樣、權益加權或聲譽評分，所有委員會選舉機制都假設某種形式的誠實多數——無論是機率意義上的（隨機選中的委員會大概率誠實）還是經濟意義上的（持有較多權益的節點傾向誠實）。然而，這些假設在面對策略性攻擊者時並不穩固。

第二，**效率與安全性的耦合設計**。現有架構將委員會同時用於提供活性與安全性，使得攻擊者一旦控制委員會即可同時破壞兩者。這種耦合設計源於傳統 BFT 共識的思維慣性，但在 BCFL 的應用場景中並非必要。

第三，**缺乏事後偵測與清除機制**。一旦惡意節點透過合法途徑（累積權益、建立聲譽）獲得委員會席位，現有系統無法事後偵測其惡意行為，也無法在發現惡意行為後將其清除。被攻陷的狀態成為新的「合法」狀態，系統缺乏自我修復能力。

本研究針對上述侷限，提出將安全性保證從委員會層級提升至全網層級的架構設計。透過將「活性」與「安全性」解耦——由小型委員會負責日常的樂觀執行以提供活

性，由挑戰機制配合全網 PBFT 共識提供安全性保證——系統可在維持效率的同時抵禦委員會佔領攻擊。具體機制設計詳見第四章。

## 2.4 區塊鏈聯邦學習驗證機制的相關研究

現有區塊鏈聯邦學習 (BCFL) 驗證方法可分為兩大類: 基於密碼學證明的驗證方法與基於委員會的共識方法。前者追求數學上可證明的正確性但面臨嚴重的效能瓶頸, 後者透過經濟激勵達成共識但依賴誠實多數假設。本節系統性分析這些方法的技術原理、效能數據與固有局限, 以定位本研究的貢獻。

### 2.4.1 基於驗證的方法:zkML 的計算瓶頸

零知識機器學習 (zkML) 透過將 ML 計算轉換為算術電路, 使驗證者無需重新執行即可確認計算正確性 [38]。其技術堆疊包括 Groth16(證明大小最小, 約 **200 bytes**)、PLONK(通用可更新設置) 與 zk-STARK(無需信任設置, 具量子抗性)[39]。轉換過程需經歷三階段: 首先將浮點數量化為有限域整數, 接著將每個運算分解為多項式約束, 最後生成密碼學證明。

然而, 約束數量隨模型複雜度急劇膨脹。根據 ZEN 編譯器的基準測試 [40],ShallowNet-MNIST 需要 **4.31M** 個約束, 而 LeNet-Face-large-ORL 則暴增至 **263M** 個約束。Chen 等人在 EuroSys 2024 的實驗顯示 [38],ResNet-18 的證明生成需 **52.9 秒**,VGG16 需 **637 秒**,DistillGPT-2 更高達 **3,651 秒** (約一小時), 且需要 **1TB RAM** 的高規格硬體。框架效能差異顯著:ezkl 比 RISC Zero 快 **65.88 倍**, 記憶體使用減少 **98.13%** [41]。

zkML 的核心局限在於難以支援拜占庭容錯聚合演算法。Krum 與 Multi-Krum 需計算所有客戶端更新間的成對距離, 產生 $O(n^2 \cdot d)$ 的約束爆炸; 排序與中位數運算在零知識電路中極度昂貴。現有 zkFL 方案如 RiseFL [42] 僅支援 L2-norm 有效性檢查, 將密碼學成本從 O(d) 降至 O(d/log d), 但仍無法實現完整的距離計算。與本研究相比,zkML 提供密碼學安全性但犧牲了聚合演算法的通用性, 而本研究透過委員會機制在保持演算法靈活性的同時達成可驗證性。

### 2.4.2 基於驗證的方法:opML 的架構限制

樂觀機器學習 (opML) 採用「預設正確」的執行模式, 僅在爭議發生時才啟動驗證 [43]。其運作流程為: 服務提供者於鏈下執行 ML 推論並提交結果, 驗證者在挑戰期內可發起欺詐證明, 透過二分協議 (Bisection Protocol) 逐步縮小爭議範圍至單一計算步驟, 最終由 FPVM(欺詐證明虛擬機) 在鏈上仲裁。ORA Protocol 是首個開源 opML 實現, 支援 LLaMA 2 等 **7B+ 參數**模型直接於以太坊運行 [44]。

挑戰期設計反映安全性與效率的權衡。Optimism 採用 **7 天**、Arbitrum 採用 **6.4 天**的挑戰期 [45], 以確保驗證者有充足時間偵測並提交欺詐證明, 同時容納網路延遲、時區差

異與潛在的共識失效。然而, 這種設計與 FL 訓練動態根本衝突——聯邦學習需要快速迭代更新與聚合, 每輪等待 7 天驗證將使訓練完全不可行。

opML 的 AnyTrust 假設 (「至少一個誠實驗證者」) 與 BCFL 的需求存在本質差異。opML 設計為單一提交者與單一挑戰者間的兩方爭議, 而非多方參與者間的共識達成。FPVM 的記憶體限制需採用延遲載入設計, 當 FL 模型涉及大量參與者更新時可能超出實際限制。此外,opML 假設「數據與模型非敏感」, 與 FL 的隱私保護需求相悖。雖然 opp/ai 透過整合 zkML 元件增強隱私, 但仍維持單一證明者架構, 無法滿足多驗證者場景需求。

### 2.4.3 基於委員會的方法:FLCoin 的滑動窗口機制

FLCoin 提出基於滑動窗口的動態委員會選舉機制 [28], 將聯邦學習過程本身作為委員會成員資格的依據。每個有效更新區塊代表一個委員會成員份額, 窗口大小固定為 s, 隨新區塊附加而滑動更新。節點的貢獻值計算為 $C_k = \alpha \times |D_k|$, 其中 $\alpha$ 為預定義係數,$|D_k|$ 為訓練數據規模; 貢獻值最高者成為委員會領導者。

拜占庭安全機率透過超幾何分佈計算:$P[X \leq s/3]$ 表示窗口內惡意節點數不超過容錯閾值的機率。在網路規模 $n = 500$、惡意節點比例 $\leq 25\%$、窗口大小 $s = 100$ 的條件下, 安全機率達 **98.4%** ;s=150 時提升至 99.8% ,s=50 時降至 91.3% [28]。驗證採用兩步驟: 誠實訓練檢查 (驗證訓練時間與算力的一致性) 與準確度檢查 (委員會成員使用本地數據驗證模型品質)。

效能方面,FLCoin 相較 PBFT 實現通訊開銷降低 **90%**、訓練時間縮短 **35%**。在 100 節點配置下, 共識延遲僅 **3.05 秒** (PBFT 為 25.11 秒), 且隨網路規模擴大保持穩定 [28]。然而,FLCoin 未明確處理長期權益累積風險——惡意節點可透過持續參與逐步增加委員會影響力。身份鏈依賴「預定義的可信管理者群組」, 引入中心化風險。論文亦承認實驗假設無惡意節點, 未驗證對抗性累積策略的防禦效果。

### 2.4.4 基於委員會的方法:BlockDFL 的權益加權選舉

BlockDFL 採用完全去中心化的點對點架構 [13], 透過最新區塊雜湊值與權益加權實現委員會選舉的隨機性與可驗證性。系統定義三種角色: 更新提供者 (UP) 負責本地訓練、聚合者負責收集與篩選更新、驗證者透過 PBFT 投票達成共識。其核心假設為「持有大量權益的參與者傾向誠實行為, 因為他們能從貨幣獎勵中獲益更多」。

BlockDFL 採用兩層評分機制: 第一層由聚合者透過本地推論評估更新品質並篩選; 第二層由驗證者使用 **Krum 演算法**過濾異常值。這使系統能容忍 **40% 惡意參與者**, 優於多數現有框架的 30% 閾值 [13]。獎勵連鎖機制將權益均等分配給被選中全局更新的聚合者、更新提供者與支持驗證者, 區塊內容完整記錄所有獲獎身份。

與 FLCoin 的關鍵差異在於選舉基礎:BlockDFL 依賴經濟權益,FLCoin 依賴 FL 貢獻歷史。這產生不同的安全特性——BlockDFL 在對手取得 >50% 權益、協調 >40% 惡意節點、或願意犧牲權益發動攻擊時失效。後者尤其值得關注: 國家級攻擊者或競爭對手

可能接受經濟損失以達成外部目標。此外,Sybil 攻擊者可跨多重身份逐步累積權益, 最終達成多數影響力。

### 2.4.5 基於委員會的方法:BFLC 與其他方案

BFLC 開創性地將委員會共識引入 BCFL [14], 採用雙區塊儲存設計: 模型區塊儲存聚合後的全局模型, 更新區塊儲存經驗證的本地梯度。每輪約 **40% 活躍節點**被選為下輪委員會成員, 透過 K-fold 交叉驗證評估提交更新的品質。實驗於 FISCO 區塊鏈系統上進行, 使用 FEMNIST 數據集與 AlexNet 模型, 證明在正常與對抗場景下均維持較高準確度。

然而,BFLC 的聲譽機制存在冷啟動問題——新節點缺乏歷史數據建立信任, 使惡意節點易於滲透委員會。當惡意節點佔據 50% 委員會席位時攻擊即可成功。後續研究指出 BFLC「易受惡意節點混入委員會的影響」[13], 且委員會共識機制可能導致節點間大量通訊開銷。

VBFL 提出 PoS 啟發的去中心化驗證機制 [46], 個別驗證者使用準確度差異 (VAD) 指標評估更新品質, 連續多輪被識別為惡意的裝置將被列入黑名單。實驗顯示在 15% 惡意裝置下達 **87% 準確度**, 比 Vanilla FL 高 **7.4 倍**。VFChain 則首創結合可驗證性與可審計性的框架 [47], 其雙跳鏈 (DSC) 數據結構支援高效的委員會輪換搜尋與歷史追溯。這些方案共同面臨 50% 拜占庭閾值限制與資源受限裝置的驗證計算負擔。

### 2.4.6 現有方法的系統性局限分析

綜合分析揭示現有方法在「安全性-效率-通用性」三維度上的 Pareto 前沿權衡。zkML 提供最強的密碼學安全性 (無需信任假設), 但證明生成時間與模型規模呈超線性增長, 且無法支援 Krum 等複雜聚合;opML 透過經濟激勵大幅降低計算成本, 但 7 天挑戰期與單一證明者架構使其不適用於多驗證者 FL 場景。

委員會方法在效率與實用性間取得平衡, 但均依賴某種形式的誠實多數假設——無論是 FLCoin 的 25% 資源閾值、BlockDFL 的 50% 權益閾值, 或 BFLC 的 50% 委員會閾值。

表 2.7: BCFL 驗證方法比較

| 方案 | 安全性保證 | 效率 (典型延遲) | 聚合通用性 |
|---|---|---|---|
| zkML | 密碼學證明 | 分鐘至小時 | 僅 FedAvg |
| opML | 經濟安全 (AnyTrust) | 7 天挑戰期 | 受 FPVM 限制 |
| FLCoin | 98.4% 機率 (s=100) | 3.05 秒共識 | 支援 |
| BlockDFL | 40% 容錯 | <3 秒驗證 | 支援 Krum |
| BFLC | 50% 委員會閾值 | 中等 | 支援 |

更關鍵的是, 所有委員會方案均未充分處理**長期權益累積**導致的委員會滲透風險。FLCoin 的滑動窗口基於即時貢獻而非累積權益, 但未建立權益衰減機制;BlockDFL 的權

益直接影響選舉機率, 惡意方可透過長期參與逐步控制系統。這揭示了現有研究的核心缺口: 靜態的安全性分析假設對手資源固定, 忽略了對手策略性累積影響力的動態過程。

## 2.5 系統模型與前置定義 (System Model and Preliminaries)

本節定義本研究所採用的基準系統模型。此模型基於 BlockDFL 委員會架構並進行擴展, 作為後續威脅分析與防禦設計的基礎。

### 2.5.1 網路模型

本研究考慮一個去中心化的區塊鏈聯邦學習系統，由以下三種核心角色構成：

1. **Update Providers (UP)**：原為客戶端 (Clients)，集合記為 $\mathcal{U} = \{u_1, u_2, ..., u_N\}$。每個 Update Provider 持有本地私有資料集 $\mathcal{D}_i$，負責在本地進行模型訓練並提交更新。

2. **Aggregators (AG)**：集合記為 $\mathcal{A} = \{a_1, a_2, ..., a_K\}$。負責收集 UP 的更新，執行初步聚合生成提案。Aggregator 的選擇基於權益。

3. **Verifier Committee (VC)**：集合記為 $\mathcal{V} = \{v_1, v_2, ..., v_M\}$。Verifiers 組成委員會，負責驗證 Aggregator 的提案。委員會成員通過共識機制批准提案並上鏈。

### 2.5.2 聚合與共識流程

在每個訓練輪次 $r$，系統執行以下流程：

1. **本地訓練**：UP 訓練 model update $\Delta w_i$ 並發送給 AG。

2. **初步聚合**：AG 生成聚合更新 $\Delta w_{agg}$ 並提交提案交易。

3. **委員會驗證**：委員會 $\mathcal{V}_r$ 執行驗證邏輯（如 Krum 檢驗）。

4. **共識決策**：委員會通過 BFT 共識對提案投票。

5. **獎勵分配**：若提案通過，AG、UP 與投票贊成的 Verifiers 共同瓜分系統獎勵。

### 2.5.3 權益動態與攻擊面

權益（Stake）在系統中扮演核心角色，既是選擇權重的依據，也是經濟獎勵的來源。這種「贏家通吃」的正反饋特性雖然激勵了誠實行為，但也創造了攻擊面：若攻擊者能策略性地累積權益，便能逐步掌控委員會。與傳統 PoS 不同，BCFL 中的攻擊者不

僅能破壞共識，還能透過投毒模型永久損害全域模型的效能，且難以被傳統 BFT 機制偵測。

# 第三章　威脅模型 (Threat Model)

基於第二章定義的委員會架構系統模型（詳見 2.5 節），本章將深入分析該架構在面對理性攻擊者時所呈現的安全脆弱性。本章的核心任務在於定義「漸進式委員會佔領攻擊」(Progressive Committee Capture Attack, PCCA)，這是一種針對權益機制的隱蔽性攻擊手法。通過揭示攻擊者如何利用權益機制內建的正反饋特性逐步實現網路控制權的轉移，本章為後續章節的防禦機制設計提供明確的安全目標與理論基礎。值得注意的是, 這種攻擊不同於傳統的模型投毒攻擊, 而是從根本上顛覆了去中心化系統的安全假設, 其危險性在於能夠將表面上去中心化的聯邦學習系統重新集權化至攻擊者手中。

## 3.1　攻擊者模型

### 3.1.1　攻擊者類型：理性攻擊者

本研究所考慮的攻擊者屬於理性攻擊者 (Rational Adversary) 範疇, 這與傳統區塊鏈安全研究中常見的拜占庭攻擊者存在本質性差異。拜占庭攻擊者的行為動機往往是純粹的破壞性, 他們可能採取任意惡意行為來癱瘓系統, 即使這些行為會導致自身利益受損也在所不惜。這種攻擊模型源自於最壞情況的假設, 但在實際應用中, 攻擊者往往具有明確的經濟動機, 而非單純追求破壞。相比之下, 理性攻擊者的行為模式遵循經濟理性原則, 他們的首要目標是利益最大化而非系統破壞。這意味著理性攻擊者會仔細評估每次攻擊行為的預期收益與成本, 只有當預期收益明顯大於成本時才會採取行動。更重要的是, 如果能夠通過機制設計使得攻擊的預期收益為負, 理性攻擊者將自發地選擇誠實行為, 無需依賴傳統的誠實多數假設。這種區分為基於博弈論的防禦機制提供了理論基礎, 也是本研究設計激勵相容機制的關鍵前提。

理性攻擊者的目標體系呈現出多層次性與長期性的特徵。在最直接的層面, 攻擊者追求經濟利益的最大化, 具體表現為通過操縱委員會來獨佔訓練獎勵, 將誠實節點排除在獎勵分配機制之外。然而, 這種短期經濟收益只是攻擊者目標體系的表層, 更深層的目標在於權益壟斷與網路控制。通過系統性地阻止誠實節點的權益增長, 攻擊者能夠逐步提高自身在整個系統中的權益佔比, 這種權益佔比的提升會直接轉化為在委員會選擇

過程中的優勢地位。當攻擊者的權益佔比達到某個臨界點後,他們將能夠持續控制委員會的組成,進而完全掌握聯邦學習過程,包括決定哪些模型更新會被接受,哪些會被拒絕。這種從經濟收益到網路控制的轉變,體現了攻擊者策略的長期性與系統性,也是PCCA攻擊之所以危險的根本原因。值得強調的是,這種攻擊並非僅僅影響模型品質,而是從根本上顛覆了去中心化系統的權力結構,將名義上的去中心化系統實質上轉變為由攻擊者集中控制的體系。

### 3.1.2 攻擊者能力與限制

在能力方面,本研究假設攻擊者能夠控制系統中一定比例的驗證者節點,這個比例記為 $f$。在典型的威脅場景下,我們假設 $f \leq 0.3$,即攻擊者最多控制全網 30% 的節點。這個假設並非任意設定,而是基於實際區塊鏈系統中攻擊者資源有限的現實考量。被攻擊者控制的這些節點並非孤立運作,而是能夠相互協調並共同執行精心設計的攻擊策略。例如,當多個惡意節點同時被選入同一個委員會時,它們可以串通一致地投票,形成協同作惡的局面。更值得注意的是,攻擊者具備策略性調整能力,能夠根據系統的動態狀態靈活改變行為模式。在權益積累的早期階段,攻擊者可能完全表現誠實以建立信譽並累積資源,而一旦獲得委員會的多數席位,便會立即切換至攻擊模式。此外,攻擊者擁有完整的觀察能力,可以追蹤區塊鏈上的所有公開資訊,包括其他節點的權益分布、歷史行為記錄、委員會組成變化等,並基於這些資訊進行精確的策略規劃。

然而,攻擊者的能力並非無限,其行為同時受到多個維度的約束。從密碼學角度來看,攻擊者無法突破系統所採用的密碼學原語,這意味著他們既無法偽造其他節點的數位簽章,也無法篡改已經寫入區塊鏈的歷史資料。在網路控制層面,攻擊者的節點數量受到經濟成本的限制,無法達到發動傳統 51% 攻擊所需的絕對多數。更關鍵的是,理性攻擊者的行為受到經濟激勵的根本性約束,如果精心設計的防禦機制能夠確保攻擊的預期成本大於潛在收益,那麼理性攻擊者將不會嘗試發動攻擊。此外,系統的可驗證性特徵為防禦提供了重要基礎:攻擊者無法阻止其他節點獨立驗證聚合結果的正確性,任何參與者都可以重新執行聚合演算法並檢測委員會是否正確遵守協議規則。這種透明性與可驗證性為後續設計挑戰機制奠定了技術可行性基礎。

## 3.2 攻擊向量分析

區塊鏈聯邦學習系統作為一個多層次的複雜架構, 其安全威脅同樣呈現出層次化的特徵。本節的目標是系統性地分析不同層次的攻擊向量, 釐清各層防禦的現狀與局限, 進而明確本研究的關注焦點。這種層次化的分析框架不僅有助於理解 PCCA 攻擊的獨特性, 也能揭示現有研究在安全分析上存在的系統性盲點。

### 3.2.1 資料層攻擊：已有防禦

資料層攻擊主要針對聯邦學習的訓練階段, 通過污染訓練資料或模型更新來破壞最終模型的品質。具體而言, 惡意客戶端可能採用資料投毒 (Data Poisoning) 手段, 在本地訓練時刻意使用被污染的資料集, 導致產生的模型更新偏離正常分佈, 從而影響全域模型的收斂方向。另一種更直接的攻擊方式是模型投毒 (Model Poisoning), 惡意客戶端不經過真實的訓練過程, 而是直接構造精心設計的惡意模型更新向量, 這些更新可能包含後門觸發器或導向特定的錯誤分類行為。針對這類資料層威脅, 現有的聯邦學習研究已經發展出相對成熟的防禦框架, 其中最具代表性的是拜占庭強健聚合演算法, 如 Krum、Trimmed Mean、Median 等方法。這些演算法的核心思想是利用統計學方法識別並過濾異常的模型更新, 即使在存在一定比例惡意客戶端的情況下, 仍能保證全域模型朝著正確的方向收斂。

然而, 這些看似完備的防禦方法實際上建立在一個關鍵但往往被忽視的假設之上: 執行這些防禦演算法的驗證者本身是誠實的。這個假設在傳統的中心化聯邦學習場景中或許是合理的, 因為中心化伺服器的可信度通常由組織層面的信任保證。但在去中心化的區塊鏈聯邦學習系統中, 驗證者同樣是由網路中的普通節點擔任, 並沒有任何外部的信任背書。如果驗證者本身受到攻擊者控制, 他們完全可以選擇不執行這些拜占庭強健演算法, 或者更隱蔽地篡改演算法的執行結果, 宣稱執行了防禦措施但實際上接受了惡意更新。在這種情況下, 無論資料層的防禦演算法設計得多麼精妙, 都將完全失去效力。這揭示了一個根本性的問題: 資料層防禦的有效性依賴於共識層的安全性, 如果共識層本身被攻陷, 資料層的所有防線都將不攻自破。

### 3.2.2 共識層攻擊: 本研究重點

相較於已經得到充分研究的資料層攻擊, 針對共識層的攻擊則構成了本研究的核心關注對象。共識層攻擊的目標不是訓練資料或模型更新本身, 而是負責執行聚合和驗證工作的委員會機制。驗證者共謀 (Verifier Collusion) 是這類攻擊的典型形式, 多個惡意驗證者可以通過事先協調, 在投票環節協同作惡, 共同通過明顯包含錯誤或惡意特徵的聚合結果。更具威脅性的是委員會佔領 (Committee Capture) 攻擊, 攻擊者不滿足於偶然的共謀機會, 而是試圖系統性地操縱委員會選擇機制, 逐步增加惡意節點在委員會中的席位佔比, 最終實現對委員會的持續性控制。

如第三章的文獻回顧所揭示的, 現有區塊鏈聯邦學習研究在這個層面存在系統性的「驗證層盲點」。統計數據顯示, 約 93% 的相關研究在設計系統時隱含地假設驗證者是誠實的, 或者至少滿足誠實多數的條件。僅有極少數研究, 如 KFC 等, 明確考慮了惡意驗證者可能存在的場景, 並嘗試設計相應的防禦機制。更值得關注的是, 即使在引入了 Verifier 機制的 BlockDFL 類論文中, 大多數研究仍然假設 Aggregator 和 Verifier 之間在利益上是相互獨立的, 或者至少 Verifier 群體內部維持著誠實多數。本研究指出了一個被普遍忽視的風險:Verifier 和 Aggregator 完全可能形成利益集團 (Cartel), 攻擊者可以同時滲透委員會與聚合節點, 形成從上游到下游的完整控制鏈。這種「全棧控制」的風險是對現有 BlockDFL 架構安全分析的重要補充, 也是 PCCA 攻擊得以成功的關鍵條件之一。

共識層攻擊之所以比資料層攻擊更加危險, 在於其具有三個顯著特徵。首先是防禦繞過能力: 一旦委員會被惡意節點控制, 所有的資料層防禦機制都可以被直接忽略, 惡意委員會可以選擇不執行 Krum 等防禦演算法, 或者即使執行也可以篡改結果。其次是隱蔽性: 攻擊者在權益積累的早期階段可以完全表現誠實, 不會觸發任何異常檢測機制, 只有在獲得足夠優勢時才發動攻擊, 這使得傳統的基於行為監測的防禦方法難以發揮作用。第三是自我強化特性: 一旦攻擊成功, 攻擊者將獨佔系統獎勵, 導致其權益進一步增加, 這又會提高其在未來委員會中的佔比, 形成正反饋循環。這種自我強化機制使得系統一旦陷入被攻擊狀態, 將很難通過內部的自我修復機制恢復正常。

### 3.2.3 攻擊層次對比

為了更清晰地呈現不同層次攻擊的特徵差異與防禦現狀, 表 3.1 提供了系統性的對比分析。

表 3.1: 攻擊層次對比

| 攻擊層次 | 攻擊者 | 攻擊目標 | 現有防禦 | 防禦假設 | 本研究關注 |
|---|---|---|---|---|---|
| 資料層 | 惡意客戶端 | 模型品質 | Krum, Trimmed Mean | 驗證者誠實 | 否 |
| 共識層 | 惡意驗證者 | 網路控制 | 誠實多數假設 | 多數驗證者誠實 | 是 |

從表中可以清楚地看到, 資料層攻擊已經發展出相對完善的防禦方法體系, 但這些方法的有效性建立在驗證者誠實執行協議的假設之上。相比之下, 共識層攻擊的防禦仍然停留在依賴誠實多數假設的階段, 缺乏針對理性攻擊者的激勵相容機制。這種防禦上的不對稱性, 正是本研究需要填補的關鍵空白。更深層次地看, 資料層防禦與共識層防禦之間存在著依賴關係: 前者的有效性完全取決於後者的可靠性。因此, 即使投入再多的研究資源去優化資料層的拜占庭強健演算法, 如果不能從根本上解決共識層的安全問題, 整個防禦體系仍然建立在不穩固的基礎之上。

## 3.3 漸進式權益佔領攻擊 (Progressive Committee Capture Attack)

本節將詳細定義本研究針對的核心威脅: 漸進式權益佔領攻擊 (Progressive Committee Capture Attack, PCCA)。這是一種專門針對基於權益的委員會選擇機制的隱蔽性攻擊手法, 其獨特之處在於通過精心設計的兩階段策略, 利用權益機制內在的正反饋特性, 實現從小規模滲透到全面控制的漸進式轉變。

### 3.3.1 攻擊定義與核心機制

PCCA 的本質是一種針對權益衍生系統的經濟攻擊, 其核心在於利用「權益-選舉-獎勵-權益」這一閉環機制中存在的正反饋特性。在正常運作的權益證明系統中, 節點的權益決定了其被選入委員會的機率, 而成功參與委員會工作又會獲得獎勵, 從而增加權

**Algorithm 2** High-Level Strategy of Progressive Committee Capture Attack (PCCA)

**Require:** Current Committee $\mathcal{V}$, Adversary Controlled Nodes $\mathcal{C}_{adv}$
**Ensure:** Action for the current round
**Check Phase:** Calculate control ratio $r = \frac{|\mathcal{V} \cap \mathcal{C}_{adv}|}{|\mathcal{V}|}$
**if** $r \leq 2/3$ **then**
*State 1: Shadow Mode (Lurking)*
    Follow the protocol honestly to accumulate stake and await majority.
**else**
*State 2: Capture Mode (Occupying)*
    **if Aggregator is Adversarial then**
        **Full Stack Poisoning**: Force approve malicious proposal.
    **else**
        **Strategic Starvation**: Force reject honest proposal.
    **end if**
**end if**

益。這種設計的初衷是激勵節點誠實參與, 但攻擊者可以將這一機制轉化為權益壟斷的工具。PCCA 的攻擊策略分為兩個明確的階段: 在潛伏階段, 攻擊者控制的節點完全遵守協議規則, 表現得與誠實節點無異, 目的是積累初始權益並建立良好的信譽記錄。這個階段的持續時間取決於攻擊者的初始資源與委員會的隨機選擇結果, 攻擊者會持續觀察系統狀態, 等待一個關鍵的時機窗口: 當多個惡意節點恰好同時被選入同一個委員會, 且其席位數超過委員會總席位的三分之二時, 攻擊便進入第二階段。

在佔領階段, 攻擊者利用在委員會中的多數優勢, 啟動「戰略性餓死」(Strategic Starvation) 策略。這種策略的核心不是直接破壞模型品質, 而是通過操縱投票結果來控制獎勵分配。具體而言, 惡意委員會會系統性地拒絕由誠實節點主導的聚合提案, 即使這些提案包含高品質的模型更新。由於區塊鏈聯邦學習系統通常採用「提案-投票-獎勵」的連動機制, 被拒絕的提案意味著相關的 Aggregator 和 Update Providers 都無法獲得本輪獎勵。通過持續執行這種排他性策略, 惡意節點不僅獨佔了系統獎勵, 還造成了誠實節點的權益停滯。隨著時間推移, 攻擊者的權益呈現指數增長趨勢, 而誠實節點的權益佔比相對下降, 這進一步提高了攻擊者在未來委員會選舉中的優勢, 形成自我強化的正反饋循環。最終, 當攻擊者的權益佔比達到某個臨界值後, 他們將能夠持續控制委員會的組成, 完全掌握網路的治理權。

演算法 2 以形式化的方式呈現了 PCCA 的決策邏輯。攻擊者在每一輪開始時都會計算其在當前委員會中的控制比例 $r$, 這個比例決定了攻擊者採取的行為模式。當控制比例未超過三分之二時, 攻擊者進入「影子模式」, 嚴格遵守協議規則以避免暴露身份並持續積累權益。一旦控制比例超越臨界值, 攻擊者立即切換至「佔領模式」, 此時的具體

策略取決於當輪 Aggregator 的身份。如果 Aggregator 本身也受攻擊者控制, 那麼整個提案-驗證鏈條都在攻擊者掌握之中, 此時可以執行更激進的「全棧投毒」策略, 直接將包含惡意內容的模型更新寫入區塊鏈。如果 Aggregator 為誠實節點, 攻擊者則採用相對保守的「戰略性餓死」策略, 通過拒絕誠實提案來實現經濟層面的打擊, 同時避免在技術層面留下明顯的攻擊痕跡。

## 3.3.2 攻擊階段詳述

### 3.3.2.1 階段一: 潛伏階段 (Latent Phase)

潛伏階段是 PCCA 攻擊成功的關鍵前提, 其核心目標是在不引起任何懷疑的情況下, 為後續的佔領階段創造必要條件。在這個階段, 攻擊者面臨的主要挑戰是如何在誠實行為與權益積累之間取得平衡。由於委員會的選擇基於權益加權的隨機抽樣, 攻擊者的初始權益佔比直接決定了其節點被選入委員會的機率, 進而影響多個惡意節點同時入選的可能性。假設攻擊者控制全網 $f = 0.3$ 的節點, 而委員會大小為 $C = 7$, 那麼要形成超過三分之二的多數優勢, 至少需要 5 個惡意節點同時被選中。根據超幾何分佈的計算, 這種情況發生的機率約為 2.4%, 這意味著攻擊者平均需要等待約 42 輪才能獲得一次發動攻擊的機會。

在這漫長的等待期間, 攻擊者必須維持完美的誠實表現。當攻擊者控制的節點被選為 Update Provider 時, 它們會基於本地資料集進行真實的模型訓練, 提交符合協議規範的高品質更新。當被選為 Aggregator 時, 它們會正確執行聚合演算法, 包括運行 Krum 等拜占庭強健機制來過濾異常更新。當被選為 Verifier 時, 它們會認真驗證聚合結果的正確性, 對誠實的提案投贊成票, 對存在問題的提案投反對票。這種全方位的誠實表現不僅能夠幫助攻擊者積累權益, 更重要的是建立起良好的歷史記錄, 使得其他節點和監督機制都將其視為可信的誠實參與者。潛伏階段的持續時間是彈性的, 攻擊者會根據權益積累的速度與委員會組成的隨機結果動態調整策略, 在確保安全的前提下耐心等待最佳的攻擊時機。

#### 3.3.2.2 階段二: 佔領階段 (Capture Phase)

當攻擊者在系統中累積了足夠的權益並成功控制了某一輪委員會的超過三分之二席位時,PCCA 進入最關鍵的佔領階段。與傳統攻擊採取單一破壞模式不同,PCCA 在佔領階段展現出高度的策略彈性, 根據攻擊者對系統不同組件的控制程度, 採取不同層次的攻擊手法。這種分層策略設計使得攻擊既能最大化經濟收益, 又能根據實際情況控制暴露風險。

**場景一: 戰略性餓死 (Strategic Starvation via Committee Capture)** 在第一種場景中, 攻擊者成功控制了 Verifier 委員會的絕對多數席位 ($|\mathcal{V}_{mal}| > \frac{2}{3}|\mathcal{V}_{committee}|$), 但當輪的 Aggregator 角色仍由誠實節點擔任或未完全受攻擊者控制。這種非對稱的控制狀態為攻擊者提供了一種獨特的攻擊機會, 其核心策略是通過操縱投票結果來重新分配系統的經濟激勵。基於 BlockDFL 架構中普遍採用的獎勵連鎖機制, 只有當聚合提案獲得委員會的批准並成功寫入區塊鏈時, 相關的 Aggregator 和 Update Providers 才能獲得本輪的獎勵分配。攻擊者正是利用這一機制設計的關鍵環節, 通過控制委員會的投票權來決定誰能獲得獎勵, 誰將被排除在外。

具體而言, 惡意委員會會採取系統性的差別對待策略。對於由誠實 Aggregator 提交的聚合提案, 即使這些提案基於高品質的模型更新並且聚合過程完全正確, 惡意委員會仍然會協同投出反對票, 使其無法達到所需的三分之二多數支持。這種拒絕行為在表面上可能被包裝為「品質不達標」或「驗證失敗」, 但其真實目的是阻止誠實節點獲得應得的獎勵。與此同時, 如果存在一個包含較多惡意 Update Providers 的 Aggregator, 即使其提交的聚合結果在技術上屬於次優 (Sub-optimal) 而非最優, 惡意委員會也會優先批准該提案。這種策略的精妙之處在於, 次優更新雖然會影響模型收斂速度, 但不會導致模型完全崩潰, 因此具有較強的隱蔽性, 不易被外部觀察者識別為明顯的攻擊行為。

戰略性餓死攻擊的破壞力主要體現在經濟層面而非技術層面。從模型品質的角度看, 由於系統仍然接受了某種形式的模型更新 (雖然是次優的), 訓練過程並未完全停滯, 只是收斂速度相對放緩。然而, 從經濟激勵的角度看, 這種攻擊造成了災難性的後果。誠實節點發現無論自己多麼努力地訓練模型、提交高品質更新, 最終都會在委員會投票環節被系統性地排除, 無法獲得任何經濟回報。這種「付出努力但得不到回報」的狀態會導致兩種嚴重後果: 一方面, 誠實節點因為無法獲得獎勵而使其權益陷入停滯, 在未來

的委員會選舉中, 其被選中的機率相對下降; 另一方面, 惡意節點通過獨佔獎勵實現權益的持續增長, 其在下一輪委員會中的佔比進一步擴大。這種馬太效應形成了正反饋循環, 使得攻擊者的優勢隨時間推移而不斷鞏固, 最終導致權益分布完全失衡, 系統的去中心化特性名存實亡。

**場景二: 全棧投毒 (Full Stack Poisoning)** 第二種場景代表了 PCCA 攻擊的最極端形態, 攻擊者不僅控制了委員會的絕對多數, 同時也成功滲透了當輪的 Aggregator 角色。這種「全棧控制」狀態意味著從模型聚合到結果驗證的整個流程都處於攻擊者的掌控之下, 系統原本設計的多層防禦機制完全失效。在這種情況下, 攻擊者的目標從經濟打擊轉向直接的技術破壞, 通過向區塊鏈中注入惡意的模型更新來破壞全域模型的性能。

全棧投毒攻擊的執行過程展現了多層防禦失效的連鎖反應。在聚合層面, 惡意 Aggregator 可以選擇性地接收來自惡意 Update Providers 的投毒更新, 這些更新可能採用標籤翻轉 (Label Flipping)、梯度反轉或後門注入等多種投毒技術。正常情況下,Aggregator 應該執行 Krum、Trimmed Mean 等拜占庭強健聚合演算法來過濾這些異常更新, 但由於 Aggregator 本身已被攻陷, 這些防禦機制要麼被完全跳過, 要麼被刻意誤用以保留惡意更新。更隱蔽的做法是, 惡意 Aggregator 可以宣稱執行了防禦演算法, 但實際上修改了演算法的參數或執行邏輯, 使其失去過濾效果。在驗證層面, 由惡意委員會對這個明顯包含問題的聚合結果進行投票表決。儘管任何具備計算能力的節點都可以重新執行聚合演算法並發現結果的異常, 但由於委員會成員超過三分之二都是惡意的, 他們會協同投出贊成票, 強制使該提案達到共識所需的支持門檻。

全棧投毒攻擊的後果是全方位的。從模型品質角度, 被污染的更新一旦寫入區塊鏈並被全網採用, 將直接導致全域模型的準確率大幅下降, 在某些精心設計的後門攻擊場景下, 模型甚至可能在特定輸入下表現出完全違背預期的行為。從經濟層面看, 由於惡意 Aggregator 和惡意 Update Providers 瓜分了本輪的全部獎勵, 攻擊者不僅成功破壞了模型, 還進一步鞏固了其經濟優勢, 使得系統越來越難以通過正常的選舉機制實現自我恢復。從系統信任角度看, 一旦發生全棧投毒, 即使只是單次事件, 也會嚴重損害用戶對整個區塊鏈聯邦學習系統的信心, 可能引發大規模的節點退出, 加速系統的崩潰。值得強調的是, 全棧投毒場景的出現揭示了一個被廣泛忽視的系統性風險: 在現有的 BlockDFL 架構中,Aggregator 和 Verifier 雖然在協議設計上被視為相互制約的獨立角色, 但在實際攻

擊場景下, 它們完全可能被同一利益集團所控制, 形成合謀關係, 這是對現有安全分析框架的重要挑戰。

### 3.3.3 權益增長動態分析 (Stake Growth Dynamics Analysis)

為了更精確地理解 PCCA 攻擊的長期影響, 我們需要建立權益演化的數學模型, 量化分析在沒有外部干預的情況下, 攻擊者的權益佔比如何隨時間推移而變化。假設系統初始狀態下, 攻擊者控制的節點總權益為 $S_{mal}(0)$, 誠實節點的總權益為 $S_{hon}(0)$, 攻擊者的初始權益佔比為 $f_0 = \frac{S_{mal}(0)}{S_{mal}(0)+S_{hon}(0)} = 0.3$。在潛伏階段, 雙方的權益都保持正常增長, 攻擊者通過誠實參與獲得獎勵, 權益佔比維持在初始水平附近。關鍵的轉折點出現在攻擊者首次獲得委員會超過三分之二席位的時刻, 此時戰略性餓死策略開始生效。

在每一輪成功的攻擊中, 假設系統分配的總獎勵為 $R$, 由於惡意委員會系統性地拒絕誠實提案, 這些獎勵將完全流向攻擊者控制的節點。因此, 在第一次成功攻擊後, 惡意節點的權益增加至 $S_{mal}(1) = S_{mal}(0) + R$, 而誠實節點的權益則停滯在 $S_{hon}(1) = S_{hon}(0)$。更重要的是, 隨著惡意節點權益的增加, 其在下一輪委員會選舉中獲得超過三分之二席位的機率也相應提高, 這意味著攻擊的成功頻率會隨時間遞增。假設攻擊者平均每 $k$ 輪能夠成功發動一次攻擊, 且這個頻率 $k$ 本身也是權益佔比的遞減函數, 那麼經過 $t$ 輪訓練後, 惡意節點的累積權益可以近似表示為:

$$S_{mal}(t) = S_{mal}(0) + \frac{t}{k(f(t))} \cdot R \tag{3.1}$$

其中 $f(t) = \frac{S_{mal}(t)}{S_{mal}(t)+S_{hon}(0)}$ 是動態變化的權益佔比,$k(f)$ 表示在權益佔比為 $f$ 時, 平均多少輪能成功攻擊一次。由於 $k(f)$ 隨 $f$ 增加而減小, 這意味著攻擊頻率在加速, 形成指數增長的趨勢。與此同時, 誠實節點的權益保持不變 $S_{hon}(t) = S_{hon}(0)$, 導致雙方的權益比例關係變為:

$$\frac{S_{mal}(t)}{S_{hon}(t)} = \frac{S_{mal}(0) + \frac{t}{k(f(t))} \cdot R}{S_{hon}(0)} \tag{3.2}$$

當 $t$ 趨向無窮大時, 這個比例也趨向無窮, 數學上意味著攻擊者最終將實現權益的絕

對壟斷。從系統動力學的角度看,這是一個典型的正反饋系統,一旦被觸發就會持續自我強化,直到達到某種飽和狀態 (例如攻擊者控制 100% 權益) 或外部干預介入。這種權益集中化的趨勢從根本上違背了區塊鏈系統去中心化的設計初衷,將一個理論上應該由全網節點共同治理的系統,轉變為由單一利益集團實質控制的中心化架構。

### 3.3.4 攻擊效果與影響

PCCA 攻擊對區塊鏈聯邦學習系統造成的破壞是多維度且層層遞進的,其影響範圍涵蓋了技術性能、經濟激勵、系統治理等多個關鍵層面。在模型品質層面,即使攻擊者採取相對溫和的戰略性餓死策略,系統的訓練效能也會受到明顯影響。由於惡意委員會傾向於批准次優更新而拒絕最優更新,每一輪訓練對全域模型的改進幅度都會小於正常情況,導致收斂速度顯著放緩。在某些情況下,如果被批准的次優更新與全域模型的最佳改進方向存在較大偏差,甚至可能出現訓練震盪或陷入局部最優的情況。更極端的情況下,如果惡意委員會完全拒絕所有誠實更新而只接受包含惡意內容的更新,模型將無法正常收斂,準確率持續低迷甚至出現退化。在全棧投毒場景下,模型品質的損害更加直接和嚴重,被注入的惡意更新可能包含精心設計的後門觸發器或針對特定類別的偏差,使得模型在大部分正常輸入上表現正常,但在特定條件下產生攻擊者預期的錯誤行為。

從網路治理權的角度看,PCCA 實現了權力結構的根本性轉移。在攻擊的初期階段,系統表面上仍然維持著去中心化的形態,委員會的組成看起來是通過隨機選舉產生的,各個節點都有機會參與。但隨著攻擊者權益佔比的持續上升,這種表面上的去中心化逐漸演變為實質上的寡頭壟斷。當攻擊者的權益佔比超過某個臨界值 (例如 50%) 後,他們獲得委員會多數席位的機率將超過 50%,意味著從統計意義上,他們能夠在大多數輪次中控制委員會。進一步地,當權益佔比達到更高水平 (例如 70% 或 80%) 時,惡意節點獲得超過三分之二席位的機率接近 100%,此時系統已經完全喪失了自我修復能力,每一輪的委員會都將被攻擊者控制,去中心化的承諾淪為空談。這種從分散到集中的權力轉移過程,徹底顛覆了區塊鏈系統的核心價值主張,使得系統在功能上退化為由攻擊者單方面控制的中心化架構。

在經濟激勵層面,PCCA 造成了激勵機制的嚴重扭曲與失靈。對於誠實節點而言,他們會發現一個令人沮喪的現實: 無論投入多少計算資源進行本地訓練,無論提交的模型更新品質有多高,最終都會在委員會投票環節被系統性地排除,獲得的經濟回報為零。

這種「努力與回報脫鉤」的狀態會迅速瓦解誠實節點的參與動機。理性的節點會進行成本效益分析, 當持續的零回報無法覆蓋參與系統所需的計算成本、網路成本和時間成本時, 退出系統成為理性選擇。這種節點流失會形成另一層正反饋: 誠實節點的退出進一步提高了惡意節點的權益佔比, 使得系統更容易被控制, 這又會加速更多誠實節點的離開。最終, 系統可能陷入「死亡螺旋」, 參與者數量持續下降, 網路活躍度大幅萎縮, 即使從技術上系統仍在運轉, 但已經失去了作為去中心化平台的實質意義。

### 3.3.5 與傳統攻擊的區別

為了更清晰地凸顯 PCCA 攻擊的獨特性與威脅性, 表 3.2 提供了與傳統拜占庭攻擊和資料投毒攻擊的系統性對比。

表 3.2: 與傳統攻擊的區別

| 特徵 | 傳統攻擊 | PCCA |
|---|---|---|
| 攻擊目標 | 模型品質 | 網路控制權 |
| 攻擊者動機 | 破壞 | 利益最大化 |
| 攻擊策略 | 直接投毒 | 漸進式滲透 |
| 隱蔽性 | 低 (立即可檢測) | 高 (初期表現誠實) |
| 自我強化 | 無 | 有 (權益正反饋) |
| 防禦方法 | 資料層防禦 | 需要激勵相容機制 |

從攻擊目標來看, 傳統的資料投毒或模型投毒攻擊主要關注破壞機器學習模型的性能指標, 例如降低分類準確率、植入後門、造成特定類別的誤判等。這類攻擊的影響主要局限在機器學習的技術層面, 即使攻擊成功, 系統的治理結構和參與者組成並不會發生根本改變。相比之下,PCCA 的目標是奪取系統的治理權, 控制決定模型演化方向的委員會機制。一旦攻擊成功, 攻擊者不僅能夠影響模型品質, 更能決定哪些節點可以參與、哪些提案會被接受, 實質上控制了系統的未來走向。從攻擊者動機角度, 傳統拜占庭攻擊者的行為模式往往基於最壞情況假設, 他們可能出於意識形態、惡意競爭或純粹的破壞慾望而發動攻擊, 即使這些行為會導致自身經濟利益受損也在所不惜。PCCA 則建立在理性經濟人的假設之上, 攻擊者的每一步行動都經過精心計算, 目標是最大化長期的經濟收益。這種基於理性的攻擊模型更貼近現實世界中的威脅場景, 因為大多數攻擊者確實具有明確的經濟動機。

從攻擊策略的時間維度來看, 傳統攻擊通常採取直接而迅速的方式, 惡意節點從一

開始就提交明顯異常的更新或投票, 試圖在短時間內對系統造成最大破壞。這種「一次性」的攻擊模式雖然可能在短期內造成嚴重影響, 但也使得攻擊行為容易被檢測系統識別, 被發現後攻擊者將失去繼續作惡的能力。PCCA 則採用漸進式的長期策略, 攻擊者願意在潛伏階段投入大量時間和資源來建立信譽, 只在時機成熟時才發動攻擊。這種耐心的策略使得攻擊具有極強的隱蔽性, 因為在攻擊的大部分時間裡, 惡意節點的行為與誠實節點完全無法區分。更關鍵的是,PCCA 具有傳統攻擊所不具備的自我強化特性。傳統攻擊即使成功也不會改變攻擊者與誠實節點之間的力量對比, 下一輪攻擊仍然面臨同樣的難度。但 PCCA 每成功一次, 攻擊者的權益就會增加, 未來攻擊的成功率也隨之提高, 形成滾雪球效應。這種正反饋機制使得系統一旦開始被滲透, 就會沿著權益集中化的軌道持續滑落, 直到完全失去去中心化特性。

從防禦策略的角度, 傳統攻擊已經發展出相對成熟的應對方法, 主要集中在資料層面的統計檢測與過濾。Krum、Trimmed Mean、Median 等拜占庭強健聚合演算法能夠有效識別並排除異常的模型更新, 即使在存在一定比例惡意客戶端的情況下, 仍能保證模型朝著正確方向收斂。這些方法的有效性已經在大量實驗中得到驗證, 成為聯邦學習安全研究的標準工具。然而,PCCA 攻擊完全繞過了這些資料層防禦, 因為它直接攻擊的是執行這些防禦演算法的驗證者本身。當驗證者被攻陷後, 無論資料層的防禦設計得多麼精妙, 都可以被選擇性地忽略或篡改。這揭示了一個層次化的依賴關係: 資料層防禦的有效性完全依賴於共識層的安全性。要應對 PCCA, 需要從根本上改變防禦思路, 不能再依賴誠實多數假設, 而是必須設計激勵相容的機制, 使得理性攻擊者發現誠實行為才是其利益最大化的最優策略。這需要引入經濟懲罰、聲譽機制、挑戰驗證等新的防禦維度, 構建一個多層次的安全框架。

## 3.4 安全目標

基於前述對 PCCA 攻擊機制與影響的深入分析, 本節將明確提出本研究所設計的防禦機制需要達成的安全目標。這些目標不僅要能夠有效防禦 PCCA 攻擊, 更要在防禦過程中保持系統的去中心化特性與經濟激勵的合理性, 避免引入新的安全風險或中心化依賴。

### 3.4.1 防止委員會被惡意節點持續控制

防禦機制的首要目標是破壞 PCCA 攻擊的自我強化循環，確保即使攻擊者在某一輪成功獲得委員會的超過三分之二席位，也無法將這種優勢轉化為長期的控制權。這個目標的實現需要從多個維度入手。首先，系統必須具備檢測惡意委員會行為的能力，能夠識別出委員會是否在系統性地拒絕高品質提案或批准次優提案。其次，一旦檢測到可疑行為，必須有相應的懲罰機制能夠迅速介入，對參與作惡的委員會成員進行經濟制裁，例如通過罰沒 (Slashing) 機制沒收其部分或全部權益。這種懲罰的力度必須足夠大，使得攻擊者即使成功獲得短期經濟利益，也會因為被懲罰而遭受更大的長期損失。第三，懲罰機制的執行不能依賴中心化的仲裁者，而應該通過去中心化的挑戰與驗證流程來實現，任何節點都應該有權利對可疑的委員會決策提出質疑，並通過鏈上的驗證過程來證明其合理性。通過這種多層次的防禦設計，系統能夠確保攻擊者無法通過單次成功攻擊建立起持久的優勢地位。

### 3.4.2 確保誠實節點的權益公平增長

第二個核心目標是保護誠實節點的經濟利益，確保他們通過正常參與系統能夠持續獲得應得的獎勵，權益能夠穩定增長而不會被惡意委員會的排他性策略所剝奪。這個目標的達成需要重新設計獎勵分配機制，打破 PCCA 攻擊依賴的「提案被拒絕則所有相關節點零獎勵」的連動關係。一種可能的設計思路是引入備選獎勵通道，即使誠實節點的提案在某一輪被惡意委員會拒絕，但只要能夠證明其提案的品質確實優於被批准的提案，仍然可以通過挑戰機制獲得補償性獎勵。另一種思路是設計基於長期表現的獎勵平滑機制，使得單輪的獎勵分配不是全有或全無，而是基於節點的歷史貢獻與聲譽進行累積評估。此外，系統還需要確保即使在面臨攻擊的情況下，誠實節點的相對權益佔比不會下降。這可能需要引入反壟斷機制，例如限制單一節點或節點群體的權益上限，或者對權益增長速度過快的節點進行額外審查。長期而言，只有當誠實行為能夠獲得穩定且可預期的經濟回報，理性節點才會選擇持續誠實參與，系統才能維持健康的參與者生態。

### 3.4.3 維持模型收斂性與準確性

儘管 PCCA 攻擊的主要目標是奪取網路控制權而非直接破壞模型, 但防禦機制仍然需要確保在存在攻擊的情況下, 聯邦學習的核心功能不受影響, 模型能夠正常收斂並達到預期的準確率。這個目標的實現依賴於防禦機制能夠有效識別並拒絕次優或惡意的更新。具體而言, 系統需要建立多層次的品質檢測機制, 不僅在 Aggregator 層面執行拜占庭強健聚合, 更要在 Verifier 層面引入獨立的品質驗證流程, 例如通過在驗證集上測試聚合結果的性能表現, 或者對比多個獨立聚合的一致性。當檢測到當輪的聚合結果明顯劣於歷史水平或存在異常模式時, 系統應該有能力觸發特殊處理流程, 例如要求重新聚合、延長驗證期或啟動社區投票。即使部分輪次受到攻擊影響, 只要大多數輪次的更新品質能夠得到保證, 整體訓練過程仍然能夠朝著正確方向推進。從長期收斂性的角度看, 防禦機制應該確保最終模型的準確率與無攻擊場景相當, 或至少在可接受的誤差範圍內, 證明系統具備抵禦攻擊的魯棒性。

### 3.4.4 保持系統的去中心化特性

在設計防禦機制時, 一個容易陷入的誤區是為了提高安全性而引入中心化的信任假設或特權節點。本研究強調, 防禦機制本身不應成為新的中心化風險來源, 必須始終保持系統的去中心化本質。這意味著防禦機制不能依賴任何可信第三方或中心化仲裁者來判斷節點行為的善惡, 也不能設置擁有特殊權限的超級節點來監督其他節點。所有的檢測、驗證與懲罰流程都應該通過去中心化的協議來實現, 任何普通節點都應該有平等的權利參與挑戰與驗證過程。這種設計理念要求我們不能簡單地依賴誠實多數假設, 而是要通過精巧的激勵機制設計, 利用博弈論的原理使得理性節點自發選擇誠實行為。密碼學技術如零知識證明、可驗證計算等可以在這個過程中發揮重要作用, 它們允許節點在不暴露私有資訊的前提下證明自己的計算正確性, 為去中心化驗證提供了技術基礎。只有當防禦機制本身也是去中心化的, 系統才能真正實現端到端的安全性, 而不會在解決一個問題的同時創造新的安全隱患。

### 3.4.5 激勵相容性

最後但也是最根本的安全目標是實現激勵相容性 (Incentive Compatibility), 這是應對理性攻擊者的核心策略。激勵相容性的含義是, 系統的機制設計應該使得理性節點的最優策略就是誠實行為, 發動攻擊不僅不能帶來額外收益, 反而會導致預期的經濟損失。從數學上表達, 攻擊的預期收益 $E[\text{Payoff}]$ 必須為負, 即 $E[\text{Payoff}] = P_{success} \cdot G_{attack} - P_{caught} \cdot L_{slash} < 0$, 其中 $P_{success}$ 是攻擊成功的機率,$G_{attack}$ 是攻擊成功時獲得的經濟收益,$P_{caught}$ 是攻擊被檢測到的機率,$L_{slash}$ 是被懲罰時損失的權益數量。要確保這個不等式成立, 有幾種設計策略。第一種是提高檢測機率 $P_{caught}$, 通過設計更敏感的異常檢測機制和更廣泛的挑戰參與機制, 使得惡意行為難以逃脫監督。第二種是大幅增加懲罰力度 $L_{slash}$, 使其遠大於潛在的攻擊收益 $G_{attack}$, 即使攻擊成功機率較高, 但一旦被抓住就會損失慘重, 理性節點不願意承擔這種風險。第三種是降低攻擊收益 $G_{attack}$, 例如通過限制單輪獎勵的上限或將獎勵分散到多個輪次, 使得單次成功攻擊的收益不足以覆蓋長期的作惡成本。與此同時, 獎勵機制應該確保誠實行為能夠獲得穩定且豐厚的回報, 使得誠實節點的長期累積收益明顯高於嘗試攻擊的預期收益。只有當這種激勵結構被成功建立, 系統才能從根本上消除理性攻擊者的作惡動機, 實現自我維持的安全性。

## 3.5 本章小結

本章系統性地構建了針對區塊鏈聯邦學習委員會架構的威脅模型, 核心聚焦於一種新型的共識層攻擊: 漸進式權益佔領攻擊 (Progressive Committee Capture Attack, PCCA)。與傳統的資料層投毒攻擊著眼於破壞模型品質不同,PCCA 的野心在於通過經濟手段逐步奪取系統的治理權, 最終實現對整個網路的實質性控制。這種攻擊之所以危險, 不僅在於其隱蔽性和自我強化特性, 更在於它揭示了現有區塊鏈聯邦學習研究中普遍存在的一個系統性盲點: 絕大多數研究在設計驗證機制時, 隱含地假設驗證者是誠實的或至少滿足誠實多數, 但這個假設在去中心化環境下並沒有可靠的保證機制。

本章首先定義了理性攻擊者模型, 明確了攻擊者以利益最大化而非單純破壞為目標的行為特徵。在此基礎上, 我們詳細剖析了 PCCA 的兩階段攻擊策略。在潛伏階段, 攻

擊者通過完美的誠實表現積累權益與信譽, 耐心等待多個惡意節點同時被選入委員會的時機窗口。一旦獲得超過三分之二的席位優勢, 攻擊立即進入佔領階段, 根據對系統組件的控制程度採取戰略性餓死或全棧投毒策略。前者通過系統性地拒絕誠實提案來阻止誠實節點獲得獎勵, 造成權益停滯; 後者則在同時控制 Aggregator 和 Verifier 的情況下直接注入惡意更新。無論採用哪種策略, 核心目的都是利用權益機制的正反饋特性, 實現權益的指數增長與治理權的持續壟斷。

通過權益增長動態分析, 我們從數學上證明了在沒有有效防禦機制的情況下,PCCA 將不可避免地導致權益集中化, 最終將去中心化系統轉變為由攻擊者單方面控制的寡頭結構。這種演化過程不僅破壞了模型訓練的效能, 更從根本上顛覆了區塊鏈系統的核心價值承諾。基於這一威脅分析, 本章提出了五個層次化的安全目標: 防止委員會持續控制、確保誠實節點權益公平增長、維持模型收斂性與準確性、保持系統去中心化特性, 以及實現激勵相容性。這些目標不僅要能夠有效抵禦 PCCA 攻擊, 更要在防禦過程中避免引入新的中心化風險或過度依賴傳統的誠實多數假設。下一章將介紹本研究提出的防禦機制, 展示如何通過挑戰增強委員會架構與混合式 Optimistic-PBFT 安全聚合框架, 在不依賴誠實多數假設的前提下, 構建激勵相容的防禦體系, 實現上述安全目標。

# 第四章　挑戰增強型委員會架構

# (Challenge-Augmented Committee Architecture)

區塊鏈聯邦學習系統在追求去中心化安全性的同時，往往面臨著執行效率的巨大挑戰。傳統的拜占庭容錯共識機制雖然能夠提供強大的安全保證，卻因其高昂的通訊成本而難以應用於需要頻繁更新的機器學習場景。為了突破這一困境，本章提出「挑戰增強型委員會架構」(Challenge-Augmented Committee Architecture, CACA)，該架構建立在第二章 2.5 節所定義的基準委員會模型之上, 透過引入異步審計機制與內部罰沒協議，實現了從傳統「門檻安全性」向「經濟安全性」的典範轉移。此設計哲學的核心在於認識到聯邦學習與金融交易系統在本質上的差異：機器學習過程具備天然的抗噪性與自我修復能力，這使得我們能夠在不犧牲長期安全性的前提下，優先保障系統的即時執行效率。

本架構的設計理念源自於對現有委員會機制根本缺陷的深刻洞察。小規模委員會雖然能夠顯著降低通訊複雜度，但其固有的集中化特性使得攻擊者能夠透過漸進式的權益累積來逐步控制驗證權力。CACA 透過將安全性驗證從同步的、阻塞式的流程轉變為異步的、非阻塞式的審計機制，成功地將效率優化與安全保障解耦。這種解耦使得系統能夠在正常情況下維持極高的執行效率，同時保留了在異常情況下動員全網資源進行仲裁的能力。更重要的是，透過引入經濟懲罰機制，本架構將攻擊者的理性決策空間重新塑造：任何試圖操縱委員會共識的行為都將面臨遠超其潛在收益的經濟損失，從而從根本上消除了發動攻擊的經濟誘因。

本章的結構安排如下：首先在 4.1 節中概述 CACA 的整體架構與設計哲學，詳細闡明各組件之間的協作關係；接著在後續各節中深入探討異步審計機制的運作原理、雙層信任模型的安全性論證、通訊複雜度的理論分析，以及激勵機制的經濟學基礎。透過將理論分析與概率模型相結合，本章將論證 CACA 如何在維持極高執行效率的同時，提供具備激勵相容性的強大安全保障，從而為第三章所提出的五項安全目標提供完整的技術實現路徑。

## 4.1 系統架構概覽

挑戰增強型委員會架構的設計目標在於建立一個既具備經濟安全性又能保持高執行效率的去中心化學習平台。相較於傳統的區塊鏈共識機制需要在每次狀態更新時達成全網共識，本架構採用了更為靈活的分層驗證策略，將日常的效率需求與極端情況下的安全需求巧妙地分離開來。圖 4.1 展示了 CACA 的完整運作流程，該流程涵蓋了從初始的角色分配、本地模型訓練、聚合結果驗證，直到潛在的挑戰仲裁等各個階段。這種設計使得系統能夠在絕大多數正常情況下以最小的通訊開銷快速完成模型更新, 同時保留了在檢測到異常行為時啟動全網仲裁的能力。

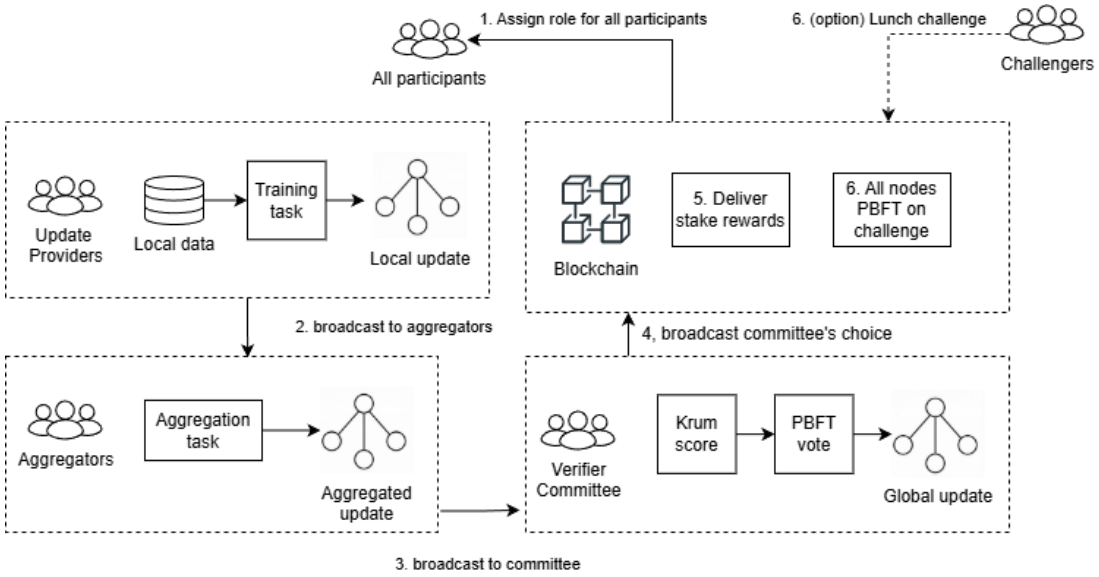



圖 4.1: Challenge-Augmented Committee Architecture (CACA) 系統架構與工作流程圖

本系統的運作依賴於四個核心角色之間的精密協作，每個角色都承擔著特定的職責並受到相應的激勵約束。訓練者 (Update Provider, UP) 是系統中持有本地私有資料的參與節點，他們的主要任務是在本地執行模型訓練並將運算出的本地更新提交給指定的聚合者。這些訓練者構成了聯邦學習系統的基礎，其資料隱私性透過本地訓練的方式得到保護，無需將原始資料暴露於網路之中。聚合者 (Aggregator, AG) 則負責收集來自多個訓練者的本地更新，執行初步的彙整運算並生成聚合更新，隨後將此聚合結果作為「提案」提交給驗證委員會。聚合者的角色設計旨在減少驗證委員會需要處理的資料量, 同時透過多個聚合者之間的競爭機制來降低單點故障的風險。

驗證委員會 (Verifier Committee, VC) 是整個架構中最為關鍵的組件, 由透過質押權

重選出的小型委員會組成。委員會的核心職責在於針對多個聚合者提交的提案運行 Krum 評分演算法, 並透過 PBFT 共識機制投票決定其中哪一份提案應被採納為該輪次的全域更新。這種設計使得驗證過程能夠在保持高效率的同時, 具備一定程度的拜占庭容錯能力。值得注意的是, 委員會的規模被刻意保持在較小的範圍內, 這是基於對通訊複雜度與安全性之間權衡的深思熟慮。雖然小委員會在理論上更容易被攻擊者控制, 但透過後續將介紹的異步挑戰機制, 這種看似的安全性劣勢實際上轉化為了效率優勢。最後, 挑戰者 (Challenger) 這一角色向所有持有足夠質押的節點開放, 他們在背景中異步監聽鏈上資料並重新執行 Krum 演算法的運算。一旦發現委員會選定的全域更新與正確的 Krum 運算結果不符, 任何挑戰者都可以發起挑戰程序, 從而觸發全網仲裁機制。這種開放式的監督設計確保了即使委員會被惡意控制, 攻擊行為也能夠被及時發現並受到懲罰。

系統的工作流程被精心設計為一個連貫且高效的循環過程, 每個階段都與前後階段緊密銜接。在每一輪次開始時, 區塊鏈系統根據前一區塊的哈希值進行動態角色抽選, 這種基於隨機性的分配機制能夠有效防止攻擊者預先布局。抽選機率與節點的質押權益成正比, 且嚴格遵循驗證者、聚合者、更新提供者的優先順序進行分配。這種優先序設計確保了最重要的驗證角色能夠由權益最大的節點擔任, 從而提高了系統的整體安全性。角色分配完成後, 被選定為更新提供者的節點使用其本地資料進行模型訓練, 並將運算結果傳遞給當輪選定的聚合者。聚合者收集到足夠數量的本地更新後, 執行初步的聚合運算並將結果作為提案提交給驗證委員會。

驗證委員會收到所有聚合提案後, 針對每一份提案運行 Krum 演算法進行評分, 該演算法能夠有效識別出偏離正常分布的異常更新。委員會成員隨後透過 PBFT 共識機制對評分最優的提案進行投票, 一旦達成共識, 該提案即被確定為最終的全域更新。這裡的關鍵設計在於, 系統採用了即時更新策略而非傳統的等待確認機制。一旦委員會達成共識, 區塊鏈立即更新全域模型並根據貢獻度向更新提供者、聚合者與驗證委員會成員發放獎勵。此過程完全非阻塞, 下一輪訓練可以立即基於新模型開始, 從而確保了系統的持續高效運作。這種設計哲學的核心在於認識到聯邦學習系統具備天然的容錯能力, 短暫的不完美更新能夠透過後續輪次逐步修正, 因此無需為了追求絕對的即時正確性而犧牲整體執行效率。

在大多數情況下, 系統的運作將在此階段順利完成並進入下一輪次。然而,CACA 的

**Algorithm 3** CACA Execution Protocol (Instant Update)

**Require:** Current Round $r$, Total Stake Weighted Nodes $\mathcal{N}$
**Ensure:** Updated Global Model $w_{r+1}$

**Role Assignment:**
Blockchain selects $\mathcal{V}$ (Committee), $\mathcal{A}$ (Aggregators), $\mathcal{U}$ (Update Providers) from $\mathcal{N}$ based on stake and randomness.
**Training & Aggregation:**
Each $u \in \mathcal{U}$ trains using $w_r$, broadcasts updates to $\mathcal{A}$.
Each $a \in \mathcal{A}$ aggregates updates into proposal $p_a$, sends to $\mathcal{V}$.
**Consensus & Update:**
$\mathcal{V}$ runs Krum on all proposals $\{p_a\}$.
$\mathcal{V}$ votes on the best proposal via PBFT.
Commit $w_{r+1}$ to blockchain **immediately**.
Distribute rewards to $\mathcal{U}, \mathcal{A}, \mathcal{V}$.

**Algorithm 4** Asynchronous Challenge Mechanism (Slash-Only)

**Require:** Challengers $\mathcal{C}$
**Ensure:** Punishment for Malicious Acts

**for** each Challenger $c \in \mathcal{C}$ **do**
    $c$ retrieves committee inputs and re-executes Krum.
    **if** $c$ detects outcome mismatch with $w_{r+1}$ **then**
        $c$ posts **Challenge Transaction** with deposit.
        **Arbitration Triggered:** All nodes re-verify.
        **if** Malicious Consensus Confirmed **then**
            **Burn/Slash** stake of malicious $\mathcal{V}$.
            Reward Challenger $c$ and all nodes.
            *// Note: Model $w_{r+1}$ is NOT reverted.*
        **end if**
        **Exit Loop**.
    **end if**
**end for**

創新之處在於引入了異步挑戰這一選用階段, 該階段作為系統的安全後盾, 在背景中持續運作而不影響正常流程。任何擔任挑戰者角色的節點都可以持續監控鏈上資料, 重新執行 Krum 演算法並驗證委員會的決策是否正確。若挑戰者發現委員會選定的結果與正確的 Krum 運算答案存在不一致, 便可以質押一定金額的押金發起挑戰。挑戰一旦成立, 區塊鏈系統將啟動全參與者的 PBFT 仲裁程序, 調動全網資源重新執行 Krum 運算以判定真實的正確答案。這種設計巧妙地將效率與安全性解耦：在正常情況下, 系統以小委員會的效率運行；在異常情況下, 系統能夠迅速升級至全網共識的安全等級。透過這種分層設計,CACA 成功地在維持高執行效率的同時, 為系統提供了等同於全網共識的安全保障。

## 4.2 異步審計與究責機制

異步審計機制是 CACA 架構中最具創新性的設計要素, 其核心理念在於將傳統區塊鏈系統中同步驗證與即時執行之間的緊密耦合關係予以解構。傳統的拜占庭容錯系統要求在每次狀態變更之前必須達成全網共識, 這種設計雖然能夠提供強大的即時正確性保證, 卻也導致了系統吞吐量與延遲性能的嚴重退化。然而, 當我們深入審視聯邦學習系統的本質特性時, 會發現這種對即時正確性的執著追求實際上並非必要。機器學習過程本身具備顯著的抗噪性, 模型參數在訓練過程中的微小偏差通常不會導致災難性的後果, 而是能夠透過後續的訓練迭代逐步修正。基於這一洞察,CACA 採用了「先執行後審計」的設計哲學, 允許系統在委員會達成共識後立即更新模型, 而將嚴格的正確性驗證推遲到異步的背景審計流程中進行。

這種即時執行策略的實施機制相當直接但極具威力。當驗證委員會對某一聚合提案達成共識後, 該提案所對應的模型更新會立即被視為有效並寫入區塊鏈。全域模型參數隨即更新, 所有訓練者節點都可以基於這個最新的模型狀態開始下一輪的本地訓練。這個過程不需要等待任何額外的確認期或審計結果, 從而確保了系統的端到端延遲能夠降至最低。在理想情況下,CACA 的執行效率幾乎與完全無防禦機制的中心化系統相當, 這是因為正常流程中唯一的額外開銷僅來自於小規模委員會內部的 PBFT 共識, 而這個開銷相較於全網共識而言幾乎可以忽略不計。這種設計選擇體現了對系統「活性」(Liveness) 的優先保障：只要委員會能夠達成共識, 系統就能夠持續前進, 而不會因為等待完整的安全驗證而陷入停滯。

然而, 即時執行策略的採用並不意味著系統放棄了對安全性的追求, 而是將安全保障從同步的阻塞式驗證轉移到了異步的非阻塞式審計。挑戰機制的設計確保了即使委員會的決策存在問題, 這些問題也能夠被及時發現並受到適當的懲罰。挑戰流程的核心在於其所依賴的「數學確定性」, 這是一個至關重要的設計要素。由於 Krum 演算法是一個完全確定性的數學運算, 給定相同的輸入必然產生相同的輸出, 因此委員會無法透過資訊不對稱來掩蓋其惡意行為。所有參與挑戰的節點都能夠獨立地重新執行 Krum 運算, 並驗證委員會的選擇是否符合演算法的正確結果。這種基於數學證明的驗證方式消除了傳統審計機制中常見的主觀判斷空間, 使得挑戰過程具備了客觀性和不可辯駁性。

挑戰流程的觸發條件設計得相當明確且易於驗證。挑戰者透過持續監控鏈上的公

開資料,獲取每一輪次中所有聚合者提交的提案以及委員會最終選定的全域更新。挑戰者在本地重新執行 Krum 演算法,計算出理論上應該被選中的最優提案,並將其與委員會實際選定的結果進行比對。若兩者不一致,則意味著委員會的決策過程存在問題,無論是由於計算錯誤還是惡意操縱,都構成了發起挑戰的充分理由。挑戰者此時可以提交一筆挑戰交易並附帶規定金額的質押金。這筆質押金的設計具有雙重目的:一方面,它能夠防止惡意節點透過大量無效挑戰來發動拒絕服務攻擊,因為錯誤的挑戰會導致質押金的損失;另一方面,它也為成功的挑戰者提供了經濟激勵,使得監督委員會行為成為一項有利可圖的活動。

當挑戰交易被提交到區塊鏈後,系統進入仲裁階段,這是整個挑戰機制中最為關鍵的環節。智能合約會立即鎖定相關的質押金,包括挑戰者的押金以及被挑戰的委員會成員的質押。隨後,合約調取該輪次中鏈上緩存的所有聚合提案資料,這些資料在委員會共識階段就已經被完整地記錄在區塊鏈上,確保了仲裁過程的資料完整性。接下來,系統觸發全網仲裁機制,所有驗證節點都被要求重新執行 Krum 演算法的運算。這個過程本質上是將原本由小委員會執行的驗證任務擴展到了全網範圍,從而將安全性等級提升到了與全網 PBFT 共識相當的高度。全網驗證者透過 PBFT 協議對仲裁結果進行投票,若超過三分之二的節點確認委員會的決策確實存在錯誤,則挑戰成立,系統將執行相應的懲罰措施。這種從小委員會到全網共識的動態升級機制,巧妙地平衡了日常運作的效率需求與極端情況下的安全需求。

當仲裁確認委員會存在惡意行為時,系統的處置策略展現出了與傳統分散式系統截然不同的設計哲學。CACA 採用「僅懲罰不回滾」(Slash-Only) 的政策,這一決策基於對聯邦學習系統特性的深刻理解以及對系統整體效益的全面考量。這種處置方式的第一個重要考量在於算力效率與機器學習系統的自癒特性。若選擇回滾模型狀態,則意味著從被攻擊的輪次開始,之後所有輪次的訓練成果都將被作廢,這將造成極為嚴重的運算資源浪費。考慮到聯邦學習通常需要經歷數百甚至數千個訓練輪次,即使僅回滾數十個輪次也將導致大量誠實節點的貢獻付之東流。更重要的是,機器學習模型具備顯著的自我修復能力,即使某一輪次的更新受到惡意操縱而包含了有偏差的梯度資訊,後續輪次中來自誠實節點的正確更新也能夠逐步抵銷這種負面影響,使模型重新收斂到正確的方向。

第二個關鍵考量涉及仲裁機制的時效性問題。全參與者的 PBFT 仲裁雖然能夠提供

最高等級的安全保證,但其通訊複雜度和時間延遲都顯著高於小委員會共識。在實際運作中,從挑戰發起到仲裁完成,往往需要經歷相當長的時間窗口,在此期間系統可能已經完成了數十個新的訓練輪次。模型參數在這個過程中持續演進,當仲裁最終判定某個早期輪次存在問題時,該輪次的影響很可能已經透過後續的正常訓練被大幅稀釋。在這種情況下,強行回滾不僅缺乏實質意義,反而會破壞系統訓練過程的連續性,導致更大的效率損失。因此,CACA 選擇接受這種由時間延遲帶來的不完美性,將重點放在對惡意行為的懲罰而非對歷史狀態的修正上。這種選擇體現了對系統整體效益的優先考量,認識到在快速演進的機器學習過程中,持續前進往往比追求歷史的完美更為重要。

第三個支持「僅懲罰不回滾」策略的理論基礎來自於機器學習領域的正規化效應。從更廣闊的視角來看,偶爾出現的次優更新 (Sub-optimal Updates) 實際上可以被視為向訓練過程中引入的隨機噪音。機器學習理論告訴我們,適度的噪音注入在特定情境下能夠起到正規化的作用,幫助模型避免過度擬合訓練資料,從而提升在未見資料上的泛化能力。雖然這種「意外的正規化」效果不應被視為系統設計的主要目標,但它確實提供了一個有趣的理論視角,說明了為何少量的非最佳更新未必會對最終的模型品質造成災難性影響。這種認知進一步支持了不回滾策略的合理性,因為它暗示著聯邦學習系統具備足夠的韌性來容忍偶發的偏差。

基於上述三點深入的學術考量,CACA 的處置方式聚焦於經濟懲罰而非狀態回退。當仲裁確認委員會的惡意行為後,系統立即執行罰沒 (Slashing) 操作,沒收惡意委員會成員以及涉案聚合者的全額質押金。這種懲罰的力度是極為嚴厲的,因為質押金的規模通常被設定在足夠高的水平,以確保攻擊的潛在收益遠小於被發現後的損失。被罰沒的資金並非簡單地銷毀或歸入系統金庫,而是被精心分配以維持激勵相容性。挑戰者作為揭露惡意行為的功臣,將獲得其中相當一部分作為獎勵,這確保了監督機制具備持續的經濟驅動力。剩餘的資金則分配給全體誠實參與者,包括那些在被攻擊輪次中提供了正確更新的訓練者,以及參與了仲裁過程的驗證節點,作為對他們所承受風險和付出努力的補償。

與此同時,系統對於受影響的模型更新採取了保留而非回退的策略。被操縱的更新紀錄會完整地保存在區塊鏈上,連同相關的懲罰記錄一起成為系統歷史的一部分。這種透明化處理方式不僅有助於學術研究和系統審計,也為其他參與者提供了寶貴的警示資訊。更重要的是,系統明確依賴聯邦學習演算法自身的強健性,透過後續輪次中誠實更新

的持續累積, 逐步稀釋並覆蓋惡意更新所帶來的負面影響。這種做法雖然在短期內可能允許模型參數存在一定程度的偏差, 但從長期來看, 模型將在誠實節點的主導下重新收斂到正確的狀態。這種設計體現了對機器學習過程本質的深刻理解, 認識到模型訓練是一個持續優化的過程而非一次性的精確計算, 因此能夠容忍和吸收過程中的局部擾動。

這種「僅懲罰不回滾」策略的最終效果在於建立了極為強大的經濟威懾力。從攻擊者的理性決策視角來看, 即使成功控制了某一輪次的委員會並注入了惡意更新, 其所能獲得的收益是相當有限的。單次的模型操縱充其量只能影響一個訓練輪次的參數更新, 而這種影響很快就會被後續的正常訓練所稀釋。相對地, 一旦攻擊行為被挑戰者發現並經仲裁證實, 攻擊者將面臨全額質押金的損失, 並被永久性地從治理委員會中除名, 失去未來獲得驗證獎勵的機會。這種極度不對稱的風險收益比使得發動攻擊在經濟上變得完全不理性。更重要的是, 這種懲罰機制成功打破了漸進式委員會佔領攻擊 (PCCA) 所依賴的正反饋循環。在沒有罰沒機制的系統中, 攻擊者可以透過操縱委員會來獲取不當獎勵, 進而增加其質押權重, 最終逐步掌控整個系統。而在 CACA 中, 任何作惡嘗試都會導致質押的減少而非增加, 從而從根本上切斷了這種惡性循環的可能性, 確保了系統長期治理的穩定性與公正性。

## 4.3 安全性保證

CACA 架構的安全性建立在一個精心設計的雙層信任模型之上, 該模型透過巧妙地分配不同層級的安全職責, 成功地在維持高效率的同時提供了等同於全網共識的安全保障。這種設計的核心洞察在於認識到, 在去中心化系統中, 不同類型的安全威脅需要不同程度的防禦機制, 而將所有安全職責都交給同一層級的共識機制既沒有必要也不符合效益。傳統的區塊鏈系統通常採用單一層級的信任假設, 要求每一次狀態變更都必須經過全網共識的嚴格驗證。這種設計雖然能夠提供強大的安全保證, 但其高昂的通訊成本使其難以應用於需要頻繁更新的應用場景。CACA 透過引入分層信任的概念, 將效率優化與安全保障解耦, 使得系統能夠根據實際威脅的性質動態調整其安全等級。

雙層信任模型的第一層是檢測層 (Detection Layer), 其採用了極為寬鬆但極其有效的「1-of-N 誠實假設」。這個假設的含義是, 只要全網 $N$ 個參與節點中存在至少一個誠實節點願意擔任挑戰者的角色, 任何委員會層級的惡意行為就能夠被成功揭露。這種假設的

寬鬆程度遠超過傳統拜占庭容錯系統所要求的「三分之二誠實節點」假設, 因為它僅需要單一誠實節點的存在而非多數誠實節點的協調行動。從概率角度來看, 在一個擁有數百或數千個參與者的大型網路中, 所有節點同時選擇沉默或串謀的可能性極其微小, 幾乎可以視為不可能事件。更重要的是, 這個誠實節點可以是任何類型的參與者, 無論是被選入委員會的候補成員、未被選中的閒置節點, 還是專門從事監督工作的獨立審計者, 只要他們能夠訪問鏈上的公開資料並執行 Krum 演算法的驗證, 就具備了發起挑戰的能力。

檢測層的設計巧妙地利用了區塊鏈系統的資料透明性特質。由於所有聚合提案都被完整地記錄在鏈上, 任何節點都能夠獨立地重新執行驗證計算, 這使得委員會的惡意行為無法被隱藏或掩蓋。攻擊者即使成功控制了當前輪次的整個委員會, 也無法阻止其他節點訪問相同的資料並發現異常。這種設計本質上將監督權力從少數特權節點民主化到了整個網路, 創造了一個「人人都是潛在監督者」的環境。值得注意的是, 檢測層並不要求挑戰者必須在攻擊發生的當下立即發現問題, 而是允許在一個合理的時間窗口內進行事後審計。這種靈活性進一步降低了監督的門檻, 因為挑戰者可以在方便的時候批次處理多個輪次的驗證工作, 而不需要持續保持實時監控的高強度狀態。

雙層信任模型的第二層是仲裁層 (Arbitration Layer), 其採用了更為嚴格但同樣標準的「全網三分之二誠實假設」。當挑戰被發起並進入仲裁階段後, 最終的判決權力從小委員會回歸到全網範圍或至少是一個大規模的陪審團。這個階段的安全假設要求網路中誠實節點的數量 $N_{honest}$ 必須超過總節點數 $N_{total}$ 的三分之二, 即 $N_{total} > 3f$, 其中 $f$ 為惡意節點的上限數量。這是幾乎所有拜占庭容錯共識協議的標準假設, 也是區塊鏈系統普遍依賴的安全基礎。在仲裁階段, 所有參與驗證的節點透過 PBFT 協議對挑戰的正當性進行投票, 只有當超過三分之二的節點確認委員會確實存在錯誤時, 挑戰才會被判定為成立。這種高門檻的設計確保了仲裁結果的可靠性, 防止了錯誤挑戰或惡意挑戰對系統造成的干擾。

這兩層信任機制的結合創造了一個強大而靈活的安全框架。在正常運作情況下, 系統主要依賴檢測層的低門檻監督來威懾潛在的攻擊行為。攻擊者明確知道, 即使只有一個誠實節點存在, 其惡意行為也有被揭露的風險, 而這種風險所對應的懲罰是全額質押金的損失。這種認知極大地提高了發動攻擊的心理門檻, 使得多數理性的攻擊者在權衡利弊後選擇誠實行為。當異常情況真的發生並觸發挑戰時, 系統能夠迅速升級到仲裁層,

透過全網共識來確保判決的公正性。這種設計的優雅之處在於,它將小委員會的效率優勢與大網路的安全優勢完美結合。小委員會負責日常的快速決策,其可能的錯誤或惡意行為由檢測層持續監督;大網路則作為最終的裁判,在需要時提供不可辯駁的仲裁結果。

從攻擊成本的角度來分析,雙層信任模型顯著提高了成功攻擊所需的資源投入。若攻擊者希望發動一次完整的攻擊並確保不被懲罰,其必須同時滿足兩個極為苛刻的條件。第一個條件是收買當前輪次委員會中超過三分之二的成員,以確保其惡意提案能夠透過委員會的 PBFT 共識。假設委員會規模為 $C$,且成員的選擇基於質押權重,攻擊者需要控制的質押金額至少為全體委員會成員質押總額的三分之二以上。這已經是一筆相當可觀的投資,考慮到質押金的設定通常較高以增加攻擊門檻。然而,僅僅控制委員會還遠遠不夠,因為攻擊者還必須防止其惡意行為被檢測和懲罰。這就引出了第二個更為嚴苛的條件:攻擊者需要收買或壓制全網足夠數量的節點,確保沒有任何誠實節點會發起挑戰,或者即使有挑戰發起,也能在仲裁階段控制超過三分之一的投票權以阻擋共識達成。

這第二個條件的達成難度遠超第一個。在檢測層面,攻擊者面臨的是「1-of-N 誠實假設」的挑戰,這意味著只要有一個節點保持誠實並願意發起挑戰,攻擊就會被揭露。要確保沒有任何節點發起挑戰,攻擊者理論上需要控制或買通全部 $N$ 個可能的挑戰者,這在大型網路中幾乎是不可能完成的任務。即使退一步假設攻擊者無法阻止挑戰的發起,而是選擇在仲裁階段透過操縱投票來逃避懲罰,其所需控制的資源也極為龐大。根據 PBFT 的安全假設,攻擊者需要控制全網至少三分之一以上的節點才能阻止仲裁達成正確的共識。若設全網節點總數為 $N_{total}$,攻擊者需要控制的節點數量至少為 $\lfloor N_{total}/3 \rfloor + 1$。在一個擁有數百個驗證者的網路中,這意味著攻擊者需要同時操控數十個甚至上百個獨立的節點,所需的質押金總額將達到天文數字。

將這兩個條件的成本累加,我們可以得出總攻擊成本的數學表達。設單個委員會成員的平均質押額為 $s_c$,委員會規模為 $C$,則控制委員會所需的成本約為 $\frac{2}{3}C \cdot s_c$。設全網單個節點的平均質押額為 $s_n$,全網節點總數為 $N_{total}$,則在仲裁階段阻擋共識所需的成本約為 $\frac{1}{3}N_{total} \cdot s_n$。總攻擊成本為這兩者之和,即 $Cost_{total} = \frac{2}{3}C \cdot s_c + \frac{1}{3}N_{total} \cdot s_n$。關鍵的觀察在於,雖然 CACA 使用了小委員會來提升效率,但其安全性並未隨之降低到僅依賴小委員會的水平。相反,透過異步挑戰機制的引入,系統的安全性實質上由全網規模

$N_{total}$ 決定而非委員會規模 $C$。這意味著攻擊成本從原本單純控制小委員會的 $O(C)$ 量級, 大幅提升到了需要控制全網的 $O(N_{total})$ 量級, 實現了安全性的顯著擴展。這種設計使得 CACA 能夠在保持小委員會的效率優勢的同時, 享有等同於全網共識的安全保障, 從而優雅地解決了去中心化系統中效率與安全之間的經典兩難困境。

## 4.4 效率分析

為了全面評估 CACA 架構的實際運作效率, 本節透過嚴格的通訊複雜度分析與概率模型推導, 論證該架構如何在理論層面實現效率與安全的最佳平衡。通訊複雜度是分散式系統性能的核心指標之一, 它直接決定了系統的吞吐量、延遲以及可擴展性。在區塊鏈聯邦學習的場景中, 通訊成本的重要性尤為突出, 因為模型參數的傳輸往往涉及大量的資料交換, 而共識協議又要求多輪的訊息往返。因此, 任何試圖在去中心化環境中實現高效機器學習的系統, 都必須在通訊複雜度上做出創新性的優化。

傳統的全網 PBFT 共識機制雖然能夠提供強大的拜占庭容錯能力, 但其通訊複雜度呈現二次方增長的特性, 這在大規模網路中成為了嚴重的性能瓶頸。在標準的 PBFT 協議中, 每個驗證節點都需要向其他所有節點廣播其提案或投票訊息, 並接收來自其他所有節點的回應。若網路中有 $N$ 個驗證節點, 則每一輪共識所需傳遞的訊息數量大致為 $N \times (N-1)$, 其漸近複雜度為 $O(N^2)$。這種二次方的增長意味著, 當網路規模從 10 個節點擴展到 100 個節點時, 通訊成本將增長約 100 倍, 而非線性的 10 倍。在聯邦學習場景中, 這種指數級的通訊成本增長將嚴重限制系統的實用性, 使得全網 PBFT 僅適用於小規模的封閉環境而難以擴展到真正的去中心化網路。

為了緩解這一問題,BlockDFL 等先前研究提出了固定小委員會的解決方案。透過將驗證職責限制在一個規模為 $C$ 的小型委員會內, 共識過程僅需在委員會成員之間進行, 從而將通訊複雜度降低到 $O(C^2)$。由於 $C$ 通常遠小於全網節點總數 $N$, 這種優化能夠帶來顯著的效率提升。然而, 這種方法的問題在於, 委員會規模的縮小直接導致了安全性的降低。較小的委員會更容易被攻擊者透過累積質押權重來逐步控制, 而一旦委員會被控制, 整個系統的安全性就蕩然無存。因此,BlockDFL 面臨著一個兩難困境：若要維持足夠的安全性, 就必須使用較大的委員會, 但這又會削弱效率優勢；若要最大化效率, 就必須使用極小的委員會, 但這又會帶來不可接受的安全風險。這種固有的矛盾使得固定

小委員會方案難以在實際應用中取得理想的效果。

CACA 架構透過引入異步挑戰機制, 成功地突破了這一兩難困境。本架構的通訊複雜度特性需要分兩種情況來討論。在正常運作情況下, 當委員會誠實執行其職責且沒有挑戰發起時, 系統的通訊複雜度與固定小委員會方案完全相同, 均為 $O(C^2)$。這是因為驗證過程僅在委員會內部進行, 無需全網參與。關鍵的區別在於,CACA 能夠安全地使用比 BlockDFL 更小的委員會規模, 因為異步挑戰機制提供了額外的安全保障, 使得小委員會的脆弱性不再是致命缺陷。在異常情況下, 當挑戰被發起並觸發全網仲裁時, 系統的通訊複雜度會暫時上升到 $O(C^2) + O(N^2)$, 其中 $O(C^2)$ 代表原始的委員會共識成本, 而 $O(N^2)$ 代表全網 PBFT 仲裁的成本。表面上看, 這似乎比全網 PBFT 的成本更高, 但關鍵在於這種高成本狀態僅在極少數情況下出現, 而非每一輪次都必須承擔。

為了量化分析系統的期望通訊複雜度, 我們引入挑戰發生的概率 $p$。這個概率代表了在任意給定輪次中, 系統需要進行全網仲裁的可能性。在理性行為假設下, 由於挑戰機制所帶來的高額經濟懲罰, 潛在的攻擊者會意識到發動攻擊的期望收益為負值, 因此傾向於選擇誠實行為。這意味著在均衡狀態下, 挑戰發生的概率 $p$ 應該趨近於零。即使考慮到偶發的系統錯誤或非理性攻擊者的存在, 在一個運作良好的系統中,$p$ 的數值也應該保持在極低的水平, 例如 0.1% 到 1% 之間。基於這個概率, 我們可以計算系統的期望通訊複雜度。每一輪次的通訊成本要麼是正常情況下的 $O(C^2)$(發生概率為 $1-p$), 要麼是挑戰情況下的 $O(C^2) + O(N^2)$(發生概率為 $p$)。因此, 期望通訊複雜度可表示為：

$$E[Comm] = (1-p) \cdot O(C^2) + p \cdot (O(C^2) + O(N^2)) = O(C^2) + p \cdot O(N^2) \tag{4.1}$$

當 $p \to 0$ 時, 上式中的第二項趨近於零, 整體的期望複雜度近似於 $O(C^2)$。這個結果表明, 在絕大多數時間裡,CACA 的通訊效率與最優化的小委員會方案相當, 但同時享有由異步挑戰機制所提供的全網級別安全保障。這種設計實現了一個重要的經濟學原理：將罕見但嚴重的風險事件 (委員會作惡) 的處理成本推遲到該事件實際發生時才支付, 而不是預先在每一輪次中都為此付出代價。這種「按需付費」的安全模式使得系統能夠在正常運作中保持極高的效率, 同時具備在需要時迅速升級安全等級的能力。

除了通訊複雜度分析, 我們還需要透過概率模型來論證小委員會在配備異步挑戰機制後的安全性。這個分析的核心問題是：在給定網路規模 $N$ 和惡意節點比例 $f$ 的

情況下, 最小的委員會規模 $C$ 應該設定為多少, 才能確保惡意節點控制委員會的機率低於可接受的風險閾值。這個問題的數學建模需要使用超幾何分佈 (Hypergeometric Distribution), 因為委員會成員的選擇是一個無放回抽樣過程。假設驗證者總池包含 $N$ 個節點, 其中惡意節點的數量為 $f \cdot N$, 誠實節點的數量為 $(1-f) \cdot N$。當我們從這個池中隨機抽取 $C$ 個節點組成委員會時, 委員會中惡意節點數量 $X$ 服從超幾何分佈。

超幾何分佈的概率質量函數描述了在無放回抽樣中獲得特定數量成功樣本的機率。在我們的場景中,「成功」被定義為抽到一個惡意節點。因此,$X = k$ 的機率可以表示為：

$$P(X = k) = \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{4.2}$$

其中 $\binom{n}{m}$ 表示二項式係數, 代表從 $n$ 個元素中選擇 $m$ 個元素的方式數量。這個公式的分子部分計算了選擇 $k$ 個惡意節點和 $C-k$ 個誠實節點的所有可能組合方式, 而分母則是從 $N$ 個節點中選擇 $C$ 個節點的總組合數。我們關心的安全性指標是惡意節點在委員會中佔據超過三分之二席位的機率, 因為這是 PBFT 共識機制的臨界點。若惡意節點數量 $X$ 達到或超過 $\lfloor 2C/3 \rfloor + 1$, 則攻擊者能夠控制委員會的共識結果。因此, 委員會被惡意控制的風險概率 $P_{mal}$ 可以表示為：

$$P_{mal} = P(X \geq \lfloor 2C/3 \rfloor + 1) = \sum_{k=\lfloor 2C/3 \rfloor + 1}^{C} \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{4.3}$$

為了具體理解這個概率模型的含義, 讓我們考察一個實際的數值案例。假設驗證者總池規模 $N = 100$, 網路中惡意節點的比例 $f = 0.3$, 即存在 30 個惡意節點和 70 個誠實節點。這是一個相對極端的假設, 因為 30% 的惡意比例已經接近大多數拜占庭容錯系統所能容忍的上限。在這種情況下, 我們可以計算不同委員會規模下被惡意控制的風險。當委員會規模 $C = 5$ 時, 惡意節點需要至少佔據 4 個席位才能達到控制閾值。透過超幾何分佈的計算, 這種情況發生的機率約為 2.74%。雖然這個風險不算高, 但在某些對安全性要求極為嚴格的應用中可能仍不夠理想。若將委員會規模增加到 $C = 7$, 惡意節點需要至少 5 個席位才能控制, 此時風險機率約為 2.42%, 略有下降但改善不明顯。

當委員會規模進一步增加到 $C = 9$ 時, 情況出現了顯著變化。此時惡意節點需要佔據至少 7 個席位才能達到三分之二的控制閾值, 而這種情況發生的機率驟降至約 0.28%。

這個數值已經低於許多實際系統所設定的風險容忍度 (通常為 1%)。繼續增加委員會規模, 當 $C = 11$ 時風險進一步降至約 0.25%, 當 $C = 13$ 時則降至約 0.21%。這些數據揭示了一個重要的洞察：即使在相當高的惡意節點比例 (30%) 下, 只需要一個規模適中的委員會 (如 9 到 13 個成員) 就能將被惡意控制的風險壓制到極低的水平。更重要的是, 這個風險水平是在沒有考慮異步挑戰機制的情況下計算的。當我們將挑戰機制納入考量後, 即使這低於 1% 的概率事件真的發生, 攻擊者也將在事後面臨全額質押金的罰沒, 從而使得攻擊在經濟上變得不可行。

這個概率分析的結論具有深遠的實踐意義。它證明了 CACA 能夠安全地使用極小的委員會規模, 例如 9 到 15 個成員, 而不會顯著增加安全風險。相比之下, 若要達到相同的安全保障水平, 傳統的全網 PBFT 需要所有 100 個節點參與共識, 其通訊複雜度為 $O(100^2) = O(10000)$。而 CACA 在使用 9 個成員的委員會時, 通訊複雜度僅為 $O(9^2) = O(81)$, 效率提升超過 100 倍。這種巨大的效率差異使得 CACA 能夠在實際應用中達到接近中心化系統的性能, 同時保持去中心化架構所帶來的安全性和抗審查性。更進一步地, 當網路規模擴大時, 這種效率優勢會變得更加顯著。若驗證者池增長到 $N = 1000$, 全網 PBFT 的複雜度將膨脹到 $O(1000^2) = O(1000000)$, 而 CACA 依然可以使用相同規模的小委員會 (因為概率分析顯示, 在更大的池中抽取相同規模的委員會, 風險反而會進一步降低), 其複雜度保持在 $O(81)$ 的量級。這種可擴展性特質使得 CACA 特別適合應用於大規模的去中心化聯邦學習平台, 為數千甚至數萬參與者的協作學習提供了理論基礎。

## 4.5 激勵機制

激勵機制是維持去中心化系統長期穩定運行的根本動力, 其設計的優劣直接決定了系統能否在沒有中心化權威的情況下自發形成良性的治理秩序。在 CACA 架構中, 激勵機制的設計遵循博弈論與機制設計理論的核心原則, 旨在創造一個激勵相容 (Incentive Compatible) 的環境, 使得誠實行為成為所有理性參與者的最優策略。與傳統的區塊鏈系統依賴持續增發代幣來支付安全成本不同,CACA 採用了一種更為可持續且經濟高效的方法, 即透過對違規者的資產罰沒 (Slashing) 來支付審計與仲裁的相關費用。這種「懲罰驅動」的激勵模式具有多重優勢, 既避免了通貨膨脹對代幣價值的長期侵蝕, 又確保了安

全成本由真正造成風險的行為者承擔，而非由全體參與者分攤。

罰沒機制的核心設計理念在於建立一個極度不對稱的風險收益結構，使得攻擊行為在經濟上變得完全不理性。每個願意擔任驗證者或聚合者角色的節點，都必須預先質押一定數量的代幣作為其誠實行為的保證金。這個質押金的規模被精心設定在一個足夠高的水平，確保其價值遠超過任何單次攻擊所能獲得的潛在收益。當節點被證實存在惡意行為，例如驗證委員會成員串謀選擇了錯誤的聚合結果，或者聚合者提交了惡意構造的虛假提案，系統將立即沒收其全額質押金。這種懲罰的嚴厲程度傳遞了一個明確的訊號：在 CACA 系統中，任何作惡嘗試都將導致災難性的經濟損失，而這種損失是即刻的、確定的且不可逆轉的。相對地，誠實參與者雖然需要承擔質押金被鎖定的機會成本，但能夠獲得穩定且可預期的區塊獎勵，這種穩定收益的累積在長期內將遠超過任何一次性攻擊所能帶來的非法所得。

被罰沒的資金並非簡單地從系統中移除或銷毀，而是透過精心設計的分配機制來強化激勵相容性。資金分配的首要受益者是成功發起挑戰的挑戰者，他們將獲得罰沒金額中相當可觀的一部分作為獎勵。這種設計確保了監督委員會行為成為一項有利可圖的經濟活動，從而吸引足夠數量的節點願意投入資源進行持續的審計工作。挑戰者的獎勵必須足夠高，以覆蓋其進行驗證計算的運算成本、質押挑戰押金的機會成本，以及承擔錯誤挑戰被反向懲罰的風險溢價。在實際設計中，挑戰成功後的獎勵通常被設定為罰沒金額的 30% 到 50%，這個比例確保了挑戰活動具備充分的經濟激勵。剩餘的罰沒資金則分配給全體誠實參與者，特別是那些在被攻擊輪次中提供了正確更新的訓練者，以及積極參與了仲裁過程的驗證節點。這種廣泛的獎勵分配機制不僅補償了誠實節點因系統遭受攻擊而承受的潛在損失，更重要的是創造了一種集體監督的文化，使得每個參與者都有動力關注系統的整體健康狀況。

激勵機制的另一個關鍵設計要素是動態調整能力，使得系統能夠根據實際運作情況自適應地優化參數設定。若系統在一段長時間內未發生任何挑戰事件，這可能暗示著兩種截然不同的情況：一種可能是威懾機制運作良好，所有參與者都選擇了誠實行為；另一種可能是挑戰門檻設置過高，抑制了潛在挑戰者的監督意願。為了區分這兩種情況並確保監督機制的活躍性，系統可以在長期無挑戰的情況下適當降低挑戰者的質押門檻，或者增加挑戰成功後的獎勵比例，從而鼓勵更多節點參與到監聽工作中。這種調整應該是漸進且謹慎的，避免因過度降低門檻而引發惡意挑戰的濫用問題。相反地，若系統在某

一時期內挑戰頻發,這可能意味著當前的懲罰力度不足以威懾攻擊者,或者質押金要求過低使得攻擊成本可以被接受。在這種情況下,系統可以提高委員會成員和聚合者的最低質押要求,同時增加罰沒比例,以強化經濟威懾效果。這種動態調整機制使得 CACA 能夠在面對不斷演變的威脅環境時保持適應性,確保激勵結構始終處於最優配置狀態。

從長期均衡的角度來分析,CACA 的激勵機制創造了一個穩定且可持續的經濟生態。對於誠實節點而言,參與系統的期望收益來自於兩個管道:一是擔任驗證者、聚合者或訓練者時獲得的常規區塊獎勵,這是一種穩定且可預測的收入流;二是在極少數情況下,當系統遭受攻擊時,透過參與挑戰或仲裁而獲得的額外獎勵。由於攻擊事件在均衡狀態下極為罕見,後者僅能視為偶發的獎金而非穩定收入。然而,即使只依賴前者,誠實參與的長期回報率也足以吸引理性的節點持續參與系統。關鍵在於,這種誠實參與是低風險的,節點只需按照協議規則執行其職責,就能確保獲得獎勵而不會面臨質押金損失的風險。相對地,對於潛在的攻擊者,其決策邏輯則完全不同。發動一次成功的攻擊能夠帶來的收益是有限的,主要體現在該輪次中對模型更新方向的控制權,而這種控制權的價值在聯邦學習場景中往往並不高,因為單次的模型偏移很快就會被後續的誠實更新所修正。然而,攻擊的成本卻是極為高昂的,不僅包括控制委員會所需的大量質押金投入,更包括一旦攻擊被發現後全額質押金的損失。

更進一步地,攻擊者還必須考慮長期的機會成本。在 CACA 系統中,被罰沒的節點將永久失去其驗證者資格,這意味著他們不僅失去了當前的質押金,也失去了未來所有輪次中獲得驗證獎勵的機會。若我們假設系統將長期穩定運行,這種未來收益流的淨現值可能遠超一次性攻擊所能獲得的短期利益。因此,任何理性的攻擊者在進行成本效益分析時,都會得出攻擊在經濟上完全不划算的結論。這種極度不對稱的風險收益結構,是 CACA 激勵機制設計的核心成就。它不依賴於對參與者道德水平的假設,而是透過純粹的經濟邏輯來引導行為,使得即使是完全自利且缺乏道德約束的理性行為者,也會選擇誠實參與而非發動攻擊。這種激勵相容性確保了系統能夠在沒有中心化監管的情況下實現自我治理,為去中心化聯邦學習平台的長期穩定運作奠定了堅實的經濟基礎。

## 4.6 本章小結

本章提出的挑戰增強型委員會架構代表了區塊鏈聯邦學習系統設計理念的一次重要轉變, 其核心創新在於透過異步審計與經濟懲罰機制的引入, 成功地將傳統上互相衝突的效率與安全性目標統一到一個連貫的框架之中。這種統一並非透過在兩者之間尋求妥協而達成, 而是透過重新思考安全性的實現方式, 將同步驗證的即時成本轉化為異步審計的條件成本, 從而在不犧牲長期安全保障的前提下, 最大化了系統的執行效率。透過移除傳統區塊鏈系統中無所不在的確認等待期,CACA 使得聯邦學習訓練過程能夠以接近中心化系統的速度持續推進, 每一輪模型更新都能夠在委員會達成共識後立即生效, 而不需要等待冗長的全網確認。

CACA 的安全性保障建立在雙層信任模型的堅實基礎之上, 該模型巧妙地利用了聯邦學習系統固有的容錯特性以及區塊鏈網路的資料透明性。透過將檢測職責開放給所有願意參與的節點, 系統將監督門檻降低到了極致, 只需存在單一誠實節點願意執行挑戰, 任何委員會層級的惡意行為都將無所遁形。與此同時, 透過保留全網仲裁作為最終判決機制, 系統確保了在真正需要時能夠動員全網資源來確保判決的公正性與不可辯駁性。本章透過嚴格的超幾何分佈分析證明, 即使在相當高的惡意節點比例下, 只需要一個規模極小的委員會配合異步挑戰機制, 就能夠將系統被攻擊的風險控制在可接受的範圍內, 而一旦這種小概率風險真的發生, 經濟懲罰機制將確保攻擊者付出遠超其收益的代價。

通訊複雜度分析進一步揭示了 CACA 架構的效率優勢。在絕大多數正常運作的情況下, 系統的通訊成本維持在小委員會共識的 $O(C^2)$ 量級, 相較於全網 PBFT 的 $O(N^2)$ 複雜度實現了數量級的降低。即使在極少數需要觸發全網仲裁的異常情況下, 由於這種情況的發生概率趨近於零, 其對系統整體期望性能的影響也微乎其微。這種設計使得 CACA 能夠在保持極高執行效率的同時, 享有等同於全網共識的安全保障, 從而優雅地解決了去中心化系統設計中的經典難題。激勵機制的創新設計則確保了這種架構不僅在理論上可行, 在實踐中也能夠長期穩定運作。透過建立極度不對稱的風險收益結構,CACA 使得誠實行為成為所有理性參與者的最優策略, 而任何試圖操縱系統的行為都將面臨災難性的經濟後果。

總體而言, 本章所提出的 CACA 架構為區塊鏈聯邦學習系統提供了一條突破效率與

安全兩難困境的可行路徑。它不僅解決了現有系統面臨的技術挑戰，更重要的是提供了一套完整的理論框架，可以指導未來更多去中心化機器學習應用的設計。然而，理論分析終究需要實證驗證來支撐其有效性。下一章將透過多維度的模擬實驗，在各種攻擊場景下驗證 CACA 架構的實際性能表現，特別是其在面對第三章所描述的漸進式委員會佔領攻擊時的防禦能力與系統穩健性，從而為本研究的理論主張提供實證基礎。

# 第五章　實驗評估 (Experimental Evaluation)

本章旨在驗證所提出的「基於異步審計與即時執行的防禦架構」在防禦「權益佔領攻擊」方面的有效性，並評估其在維持去中心化安全性的同時，是否能顯著提升系統效率。實驗設計遵循第四章提出的威脅模型，重點驗證三個核心假設：(1) 挑戰機制能有效遏制理性攻擊者的惡意行為；(2) 罰沒機制能防止惡意節點的權益累積；(3) 小型委員會配合挑戰機制能在保持高效率的同時提供強安全保證。

## 5.1　實驗設置

為了公平比較，我們在相同的實驗環境下模擬了本研究提出的方法與目前主流的基於委員會的防禦方案。

### 5.1.1　資料集與模型

我們採用 MNIST 手寫數字資料集作為基準測試任務。模型架構為一個標準的捲積神經網路，包含兩個捲積層與兩個全連接層。

資料分佈設置：為了全面評估系統性能，本研究考量了獨立同分佈 (IID) 與非獨立同分佈 (Non-IID) 兩類環境。在 IID 設置中，資料被均勻地隨機分配給所有客戶端。而在 Non-IID 設置中，我們採用基於 Dirichlet 分佈 (Dir($\alpha$)) 的資料劃分，並將濃度參數設定為 $\alpha = 0.5$。這種設定會導致每個客戶端持有的類別分佈呈現高度異質性，模擬了真實場景中資料分佈極度不均的情況，從而增加模型聚合與抗攻擊的挑戰。

### 5.1.2　基準方法與攻擊場景

基準方法 (BlockDFL)：採用固定大小委員會的主流區塊鏈聯邦學習方案。該方案依賴誠實多數假設，使用 BFT 共識機制進行模型聚合驗證。委員會大小設定為 $C = 7$，這是 BlockDFL 論文中建議的配置，能在效率與基本安全性之間取得平衡。我們設定 BFT 的共識門檻為 $2/3$，即必須有超過 $2/3$ 的成員同意才能通過提案。

本研究方法 (Ours)：同樣採用 $C = 7$ 的委員會大小，但引入了事後挑戰機制。在正

常情況下，系統採用即時執行模式，僅由單一聚合器執行聚合；當檢測到異常時，任何節點都可以發起挑戰，觸發完整的 BFT 驗證流程。

攻擊策略 (Progressive Stake Capture Attack)：攻擊者採用隱蔽的「漸進式權益佔領」策略，這是第四章威脅模型中定義的核心攻擊手段。攻擊分為兩個階段：

1. 潛伏階段 (Latent Phase)：只要攻擊者尚未獲得委員會的控制權 (即未達 2/3 席位)，皆會維持潛伏狀態並表現誠實，透過提交正常的模型更新來穩定積累權益。此階段的目的是建立信譽並增加權益佔比，從而提升未來被選入委員會的機率，為發動攻擊奠定基礎。

2. 佔領階段 (Capture Phase)：一旦攻擊者在委員會中獲得超過 2/3 席位，立即根據控制情況啟動攻擊策略。具體包含兩種場景：

    - 場景一：戰略性餓死 (Strategic Starvation)。當攻擊者僅控制委員會超過 2/3 席位時，拒絕打包誠實節點的更新，僅接受包含攻擊者更新的提案，從而獨佔獎勵並使誠實節點權益停滯。
    - 場景二：全棧投毒 (Full Stack Poisoning)。當攻擊者同時控制委員會超過 2/3 席位與 Aggregator 時，直接繞過檢測機制提交「標籤翻轉」(Label Flipping) 的惡意更新，並利用委員會多數強制達成共識，從而直接破壞模型品質。

### 5.1.3 實驗參數

本研究實驗採用的系統參數配置如表 5.1 所示。這些參數的設定遵循了 BlockDFL [13] 等主流 BCFL 研究的標準配置，確保實驗結果的可比性。

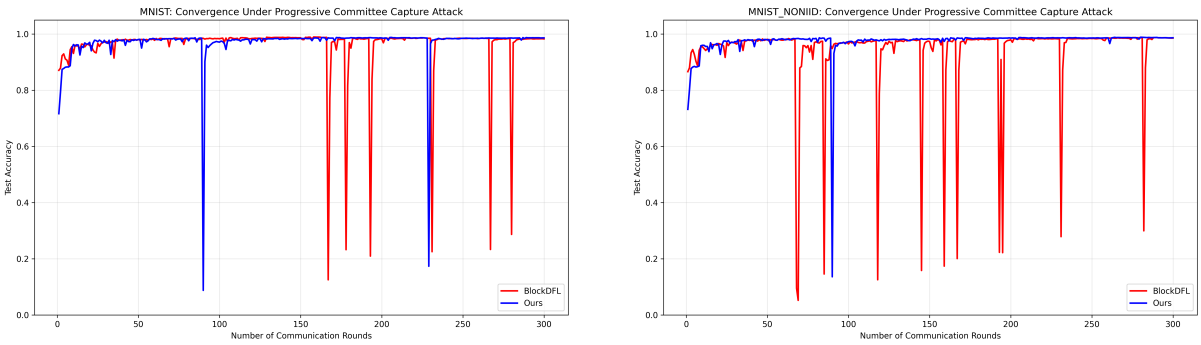
## 5.2 實驗結果與分析

### 5.2.1 模型效能與攻擊表現分析

本節針對系統在不同資料分佈下的收斂性與遭受攻擊的頻率進行量化分析。圖 5.1a 至圖 5.1b 分別展示了 IID 與 Non-IID 環境下，BlockDFL 與本研究方法（Ours）的表現。

表 5.1: 實驗參數配置 (Experimental Parameter Configurations)

| **參數名稱** | **設定值** |
|---|---|
| 訓練輪數 | $R = 300$ |
| 客戶端總數 | $N = 100$ (Verifier Pool Size) |
| 委員會大小 | $C = 7$ |
| 攻擊者數量 | $M = 30$ (初始權益佔比 30%) |
| 初始權益分配 | 所有參與節點初始均分配 100 單位 |
| 設備池分配 | Aggregator: 4 位，Provider: 其餘節點 |
| 獎勵機制 (每輪) | Verifier: 1.0, Aggregator: 1.0, Provider: 0.05 |
| 罰沒機制 | 挑戰成功時，惡意委員全額罰沒 (Full Slashing) |
| 學習率 | $\eta = 0.01$ (衰減率 0.99) |
| 本地訓練參數 | Epochs = 1, Batch Size = 32 |
| 資料分佈環境 | IID 及 Dirichlet-based Non-IID ($\alpha = 0.5$) |




(a) IID 環境 (均勻分佈)　　(b) Non-IID 環境 ($\alpha = 0.5$)

圖 5.1: 模型準確率收斂比較。(a) 為 IID 環境，(b) 為 Non-IID 環境。

### 1) 顯性攻擊影響與收斂穩定性

實驗結果顯示，BlockDFL 在兩類環境下均展現出明顯的安全性漏洞。

**攻擊頻率：**在 300 輪訓練中，BlockDFL 分別遭受了 10 次（IID）與 12 次（Non-IID）成功的委員會佔領。相較之下，本研究方法透過異步審計機制，在 IID 中僅遭受 2 次佔領，在更具挑戰性的 Non-IID 環境中也僅遭受 3 次佔領，顯示出極強的韌性。

**瞬時破壞力：**以圖 5.1b（Non-IID）為例，BlockDFL 於第 68 輪遭受標籤翻轉（Label Flipping）攻擊時，準確度由正常水平瞬間崩潰至 9.55%。這證明了在傳統 BCFL 框架下，單次成功的委員會佔領即可對全球模型造成致命打擊。

**顯性攻擊與聯邦學習的自癒性：**觀察圖 5.1b（Non-IID）可以發現，BlockDFL 在第 68 輪遭受標籤翻轉攻擊後，準確度雖瞬間崩潰至 9.55%，但隨後幾輪呈現快速回升。這印證了聯邦學習具備顯著的自我修復能力（Self-healing capacity）：只要攻擊者無法持續佔領委員會，後續輪次的誠實更新即可逐步抵銷惡意梯度產生的噪聲。因此，單次的標籤翻轉攻擊雖會造成系統震盪，但通常不會導致模型不可逆的毀滅。

**Non-IID 強健性解釋：**值得注意的是，即便在 $\alpha = 0.5$ 的高度異質資料分佈下，本系統仍能維持與 IID 相似的收斂速度。此現象源於系統採用的「基於驗證的選優機制」（Selection-based mechanism），透過全局驗證集有效過濾了 Non-IID 引起的權重發散（Client Drift）。

### 2) 系統穩定性與最低不可用率分析

為了進一步量化攻擊對系統運行的實質衝擊，本研究定義「最低不可用率」（Minimum Unavailability Rate）為指標。我們保守地假設每次受擊後的恢復期僅需 5 輪（此為實驗觀測結果 5–25 輪之最小值），並據此運算系統處於效能崩潰狀態的比例。

**下限估計與效能鴻溝：**根據實驗資料的量化分析，在 Non-IID 環境下，BlockDFL 由於遭受了 12 次成功的委員會佔領攻擊，即便採用最為樂觀的 5 輪恢復期進行運算，系統在 300 輪的訓練過程中仍有至少 20%（即 60 輪）的時間處於不可用狀態。若進一步考量到實驗中實際觀察到的最大恢復期（25 輪），其實際癱瘓時間將遠超此比例。

相比之下，本研究提出的方法憑藉「異步審計機制」，將成功受擊次數大幅壓制在 3 次以內。在同樣的保守估計準則下，本系統的最低不可用率僅為 5%（15/300 輪）。這

項數據對比清晰地證明：儘管聯邦學習具有「自癒性」，但頻繁的受擊仍會使傳統框架在訓練過程中陷入極大的不穩定；而本方法則能確保系統在 95% 以上的訓練時間內，始終維持高品質的服務能力。

**連續受擊的連鎖反應：**此外，BlockDFL 的高受擊頻率（平均每 25 輪一次）與恢復期（5–25 輪）在時間軸上高度重疊。這意味著在 Non-IID 較複雜的收斂過程中，BlockDFL 極易在尚未從前次攻擊完全恢復時再次受擊，導致模型準確度長期在低位震盪，無法累積有效的全局知識。

### 3) 最終準確率對比

在經歷 300 輪的攻防博弈後，兩者的最終訓練結果如下：

- **IID 環境：**本研究方法最終準確率達到 98.63%，BlockDFL 為 98.26%。
- **Non-IID 環境：**本研究方法達到 98.67%，BlockDFL 為 98.57%。

誠然，BlockDFL 展現了聯邦學習的自癒特性，但在 Non-IID 環境下，每次受擊後的恢復期至少需要 5 輪。保守估計，BlockDFL 在訓練過程中有超過 20% 的時間處於不可用狀態。本研究方法透過異步審計將攻擊頻率降低了 80%，確保了模型在整個週期內維持高水準的服務能力。這種「過程穩定性」在需要實時部署的關鍵任務中，其價值遠超最終 0.1% 的準確率增益。

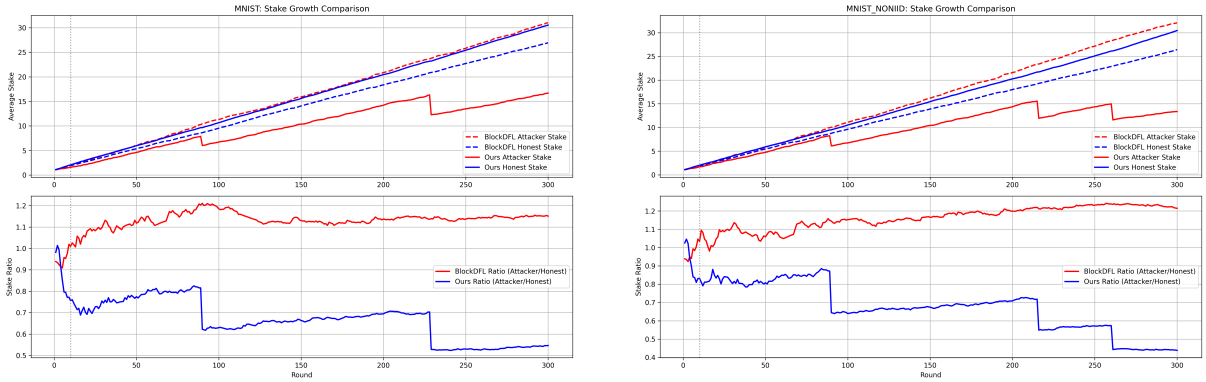
## 5.2.2 安全動態與治理風險深層分析

本節進一步探討權益演化與隱蔽攻擊的內在邏輯，揭示基於權益選拔（Stake-based Selection）系統中的固有治理風險。

### 1) 權益優勢的建立與自我強化機制

透過對原始權益資料的追蹤發現，在 BlockDFL 中，攻擊者平均持有的權益穩定維持在誠實節點的 1.1 至 1.2 倍。這種優勢地位的建立具有其系統必然性：

**任務價值差異：**系統中執行運算量較大或關鍵性較高的任務（如 Aggregator 或 Committee 成員）所獲取的獎勵遠高於普通 Provider。




(a) IID 環境　　(b) Non-IID 環境

圖 5.2: 權益演化比較。(a) 為 IID 環境，(b) 為 Non-IID 環境。

**正向回饋循環：**由於角色分配機制與權益掛鉤，一旦節點獲得初步權益優勢，其未來被選中擔任重要角色的機率隨之增加，進而獲得更多獎勵。

**增長上限分析：**攻擊者權益比未能呈現指數級成長，是因為其無法完全操控隨機的角色分配邏輯。即便惡意委員會策略性地選擇有利於惡意節點的更新，系統中仍有部分誠實節點（UP 或 AG）會獲得獎勵，從而形成了 1.1–1.2 倍的動態平衡區間。然而，只要「高貢獻任務獲得高獎勵」的分配邏輯不變，這種**領先者優勢（Leader Advantage）**便會轉化為長期的治理威脅。

## 2) 隱蔽投毒 (Covert Poisoning) 攻擊的普遍性與隱蔽性

進一步分析揭示，隱蔽投毒攻擊的隱蔽性並非僅限於 Non-IID 環境，而是系統層面的普遍風險。

**模型指標的局限性：**如實驗資料所示（例如 Non-IID 第 239 輪），即便委員會已被惡意佔領且正在執行隱蔽投毒攻擊，全球模型的準確度仍可能維持上升。這是因為攻擊者可透過保留部分高質量更新來偽裝其行為。

**解耦威脅：**這種現象顯示了**「模型效能」與「系統誠信」的解耦**。若缺乏本研究提出的罰沒機制（Slashing），攻擊者可以長期隱藏在系統中累積權益，直到達成「全棧共謀」（Full-stack Collusion）的條件。

### 3) 罰沒機制與權益抑制的動態演化

圖 5.2 記錄了 300 輪內節點權益的動態變化，這不僅反映了系統的獎懲邏輯，更揭示了惡意節點在攻擊過程中的資源損耗特徵。

**1. 台階式下降的制裁特徵：**觀察圖 5.2 可以發現，惡意節點的平均權益並非線性遞減，而是呈現顯著的「台階式下降」。這種現象對應了本研究異步審計機制觸發 Slashing 的具體時點：

- **IID 環境：**在第 90 輪與第 229 輪發生兩次大幅度的權益減損，最終降至誠實節點的 0.56 倍。
- **Non-IID 環境：**在第 90、216 與 261 輪分別觸發制裁，導致其權益在第 300 輪時僅剩誠實節點的 0.43 倍。

每一次「台階」的出現，都代表一次成功的惡意行為攔截與經濟懲罰。

**2. 經濟資本的不可逆損耗：**雖然在 300 輪的觀測期內，攻擊發生的頻率未呈現明顯的早晚期差異，但惡意節點的經濟資源（Stake）已處於持續枯竭狀態。由於本系統採用基於權益的角色選拔機制，攻擊者每次發動攻擊都面臨著喪失「治理資本」的風險。

**3. 長期治理安全性的推論：**儘管短期內攻擊者仍能憑藉剩餘權益參與競爭，但 0.43–0.56 倍的權益差距已構成實質性的進入門檻。

- **先行者優勢轉移：**誠實節點透過穩定訓練持續累積權益，擴大了與惡意節點的貧富差距。
- **攻擊難度遞增：**隨著訓練輪數繼續增加，惡意節點若要再次達成「委員會佔領」所需的席位，其權益權重將顯得捉襟見肘。

這種「台階式」的權益縮減證明了本機制能有效剝奪攻擊者的治理資源，從經濟層面限制了惡意行為的擴張潛力。

## 5.2.3 長期賽局中的經濟嚇阻力分析

為了驗證本研究提出的防禦機制在長期運作下的穩定性與嚇阻效果，我們將實驗模擬輪數擴展至 2000 輪。圖 5.3 展示了長期賽局下的權益動態變化，這些數據揭示了兩種機制在經濟誘因設計上的根本差異。

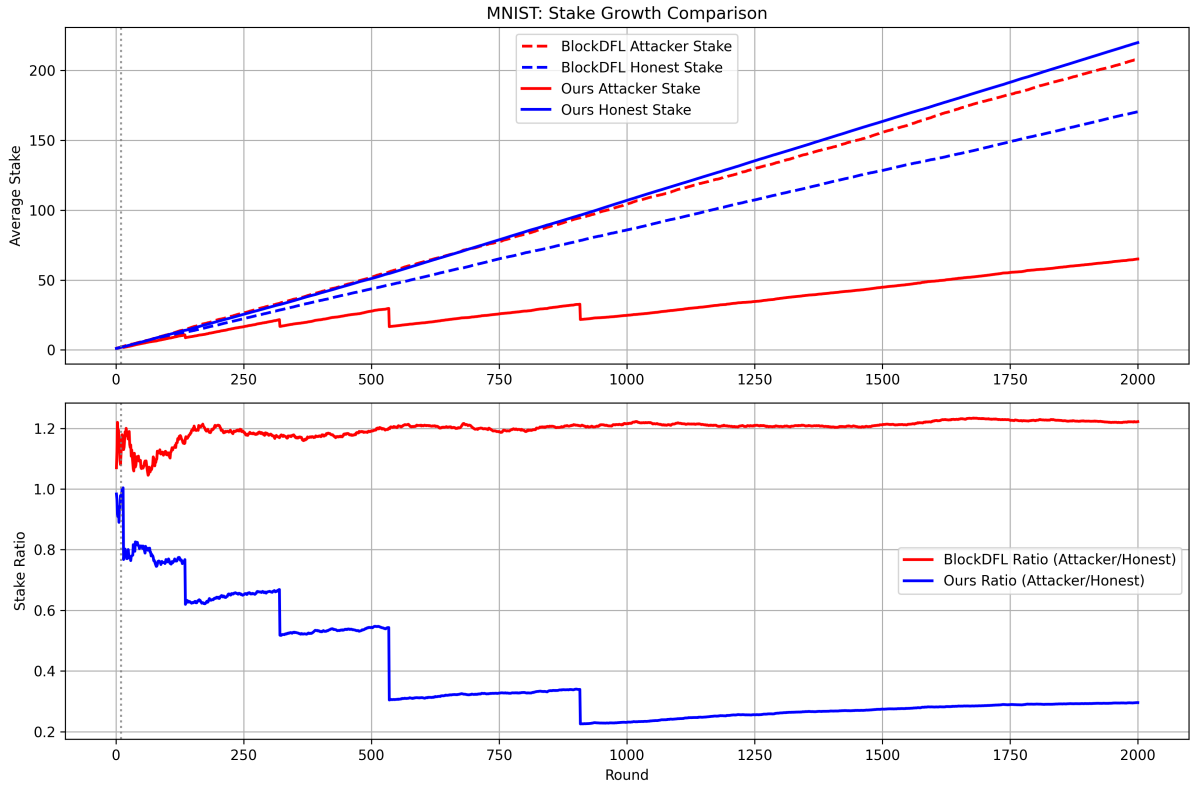



圖 5.3: 2000 輪長期模擬下的權益動態比較

## 1) BlockDFL 的財富固化與持續威脅

**強者恆強的馬太效應：**在 BlockDFL 的長期模擬中，我們觀察到顯著的財富固化現象。數據顯示，攻擊者的平均權益在約 250 輪後，便穩定維持在誠實節點的 1.2 倍左右。這種 20% 的權益優勢源於該機制缺乏有效的負向反饋迴路（Negative Feedback Loop）。一旦攻擊者透過初期優勢累積了較高的權益，其被選入委員會並獲得獎勵的機率便隨之提升，進而鞏固其經濟地位。

**高頻率的治理失效：**這種權益優勢直接轉化為對系統治理權的掌控。在總計 2000 輪的模擬中，惡意節點成功攻佔委員會多數高達 84 次。這意味著在 BlockDFL 架構下，攻擊者不僅能長期存活，更能平均每 24 輪就發動一次成功的委員會劫持，形成持續性的安全漏洞。

**2) 本研究方法的經濟嚇阻與邊緣化效應**

**不對稱的攻擊風險：**相較之下，本研究方法展現了極強的經濟嚇阻力。在相同的 2000 輪測試中，惡意節點僅成功佔領委員會 5 次。這巨大的差異（84 次對 5 次）證明了引入罰沒機制後，攻擊者的期望收益被大幅壓縮，迫使其在大部分時間必須保持誠實以避免資產歸零。

**攻擊者的經濟致死螺旋：**觀察圖 5.3 的第 909 輪可發現一個具決定性的轉折點：攻擊者在發動第五次攻擊後隨即受到異步審計機制的制裁 (Slashing)，導致其平均權益瞬間暴跌至誠實節點的 22.6%。

**永久性的治理排除：**這一經濟重創產生了長期的邊緣化效果。在隨後的 1091 輪（超過總時長的一半）中，攻擊者因權益基礎過低，徹底失去了競爭委員會席次的能力，再也無法成功發動任何一次佔領攻擊。這項結果有力地證實了本系統能有效將一次性的攻擊失敗轉化為永久性的治理排除，從而確保系統在長期演化中趨向於「誠實者主導」的穩定態。

# 5.3 效率與可擴展性分析

## 5.3.1 系統開銷與安全性需求對比

為了評估系統在極端壓力下的效能表現，我們設定基準安全性要求為「受攻擊頻率（成功被劫持機率 $p$）低於 1%」。在總節點數 $N = 100$、惡意節點佔比 $f = 30\%$ 的環境下，對比 BlockDFL 與本研究所提出的方案。

表 5.2: 不同防禦機制在相同安全性水平 ($p < 0.01$) 下的複雜度對比 ($N = 100, f = 30\%$)

| **評估維度** | **傳統方案 (BlockDFL)** | **本研究方法 (CACA)** | **差異性質分析** |
|---|---|---|---|
| 核心安全性模型 | 門檻安全性 | 經濟安全性 | 信任多數 vs. 激勵相容 |
| 設計哲學 | 悲觀併發控制 | 樂觀執行與異步審計 | 預防 vs. 治理 |
| 委員會大小 ($c$) | $c = 9$ | $c = 5$ | 資源佔用降低 44.4% |
| 常態通訊複雜度 | $O(c^2) = 81$ | $O(c^2) = 25$ | 顯著降低日常頻寬負載 |
| 安全維護成本 | 固定開銷 | 條件式開銷 | 靜態冗餘 vs. 動態防禦 |
| 挑戰觸發代價 | 無 | $O(p \cdot N^2)$ | 僅在檢測異常時觸發 |
| 長期穩定狀態 | 固定於 $O(81)$ | 趨近於 $O(25)$ | 基於經濟嚇阻的博弈均衡 |

### 5.3.2 複雜度差異與經濟安全性分析

本小節針對表 5.2 中的複雜度模型進行深度分析，揭示兩者在處理安全性威脅時的本質差異：

#### 1. 預防溢價與資源冗餘 (Pessimistic Overhead)

BlockDFL 採用的是一種「預防性策略」。為了將攻擊成功率壓制在 0.01 以下，系統必須在每一輪都維持高達 $c = 9$ 的大型委員會進行 BFT 共識。即便在系統完全誠實運行的狀態下，這份 $O(81)$ 的高昂通訊代價也是不可減免的「預防溢價」。這種設計雖然安全，但缺乏對實際威脅程度的自適應性，導致資源長期處於冗餘狀態。

#### 2. 基於罰沒機制的博弈均衡 (Economic Deterrence)

相較之下，本研究方法將安全性保證由「事前攔截」轉為「事後追責」。透過引入罰沒機制 (Slashing Mechanism)，我們成功改變了攻擊者的收益預期：

- **激勵不相容 (Incentive Incompatibility)：**雖然本研究採用的 $c = 5$ 委員會在理論上被佔領的風險較高，但由於存在 $O(N^2)$ 的全量審計與全額罰沒風險，對於「理性攻擊者」而言，發動攻擊的期望收益將遠低於潛在的經濟損失。
- **$p$ 值的動態演化：**雖然在模擬環境中我們考慮了 $p < 0.01$ 的頻率，但在真實部署環境中，一旦首位攻擊者遭到處罰並被剔除，後續節點將因「經濟致死螺旋」的威懾而選擇誠實策略。因此，實際的挑戰觸發頻率 $p$ 將隨時間迅速遞減，使得系統的攤銷成本 (Amortized Cost) 極度趨近於 $O(c_{low}^2)$，從而在極低開銷下實現了與大型委員會等效的安全等級。

## 5.4 本章小結

本章透過多維度的實驗設計與複雜度建模，全面驗證了所提出的「挑戰增強型委員會架構」在動態攻防環境下的優越性。實驗結果不僅支持了本文的核心假設，更揭示了該架構在去中心化治理中的深層潛力，具體總結如下：

- **動態防禦的韌性：**在 MNIST 任務的 IID 與 Non-IID 環境下，本研究方法均展現了極強的抗攻擊能力。數據顯示，相較於傳統固定委員會方案（BlockDFL），本架構將受擊頻率降低了約 80%，並將系統的「最低不可用率」從 20% 大幅壓制至 5% 以下。這證明了挑戰機制能有效彌補小型委員會在即時防禦上的不足，確保模型訓練過程的連續性與穩定性。

- **經濟治理的有效性：**長期賽局實驗（2000 輪）證實，引入罰沒機制（Slashing）能對惡意行為產生實質性的經濟嚇阻。透過追蹤權益演化發現，攻擊者的治理資本會因挑戰觸發而陷入「致死螺旋」，最終其權益佔比降至誠實節點的 22.6%，達成永久性的治理排除。此結果說明了「挑戰增強」不只是技術層面的補救，更是一種從經濟誘因上根除惡意行為的治理手段。

- **效率與安全性的雙贏：**複雜度對比分析顯示，在相同的安全性邊界（$p < 0.01$）要求下，本架構成功打破了安全性與通訊開銷的強耦合關係。透過解耦共識流程，系統在常態下僅需維持 $c = 5$ 的輕量級運作（$O(c^2) = 25$），相較於必須維持 $c = 9$ 的傳統方案（$O(c^2) = 81$），顯著降低了系統整體的通訊冗餘。

綜上所述，實驗數據有力地支撐了本文論點：「挑戰增強型委員會架構」能以極低的常態通訊成本，換取等同甚至優於大型委員會的安全保證，為大規模區塊鏈聯邦學習的部署提供了一條具備高擴展性的技術路徑。

# 第六章　結論與未來展望 (Conclusion and Future Work)

## 6.1　研究總結 (Summary of Research)

本研究針對區塊鏈聯邦學習 (Blockchain-based Federated Learning, BCFL) 在委員會架構下過度依賴「誠實多數假設」的安全漏洞進行了系統性分析。我們識別出一種針對權益機制缺陷的「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」，揭示了理性攻擊者如何透過累積治理資源，規避傳統的資料層防禦。為了彌補這一安全性缺口，本論文提出「挑戰增強型委員會架構 (Challenge-Augmented Committee Architecture, CACA)」，其核心設計哲學在於安全性與治理規模的解耦。透過引入異步審計與內部罰沒協議，我們將系統的安全防禦從「門檻安全性 (Threshold Security)」轉向「經濟安全性 (Economic Security)」，確保系統在面對具備策略性的理性對手時，仍能維持高度的活性與模型聚合的正確性。

## 6.2　研究發現與貢獻 (Research Findings and Contributions)

本研究的主要發現與貢獻總結如下：

- 定義並驗證 PCCA 的威脅演化：本研究首次定義了漸進式委員會佔領攻擊的兩階段模型（潛伏與佔領），並量化了權益機制正反饋如何加速網路控制權的轉移。實驗證實，傳統架構（如 BlockDFL）在長期運行中存在顯著的財富固化與治理失效風險。

- 強化系統在極端環境下的服務能力：透過 CACA 的挑戰機制，系統在遭受 30% 惡意共謀的壓力下，能有效將成功受擊頻率壓制在極低水平。數據顯示，本架構不僅能將最低不可用率從 20% 降至 5% 以下，更能在 Non-IID 資料分佈下維持與 IID 環境相近的收斂穩定性。

- 重塑理性攻擊者的誘因結構 (Incentive Realignment)：長期賽局實驗顯示，罰沒機制能有效打破惡意節點的「權益累積循環」。數據指出，攻擊失敗導致的治理權益驟降（至誠實節點的 22.6%），實質上內部化了作惡的外部性成本，使得攻擊的預期收益遠低於潛在損失。這種經濟上的不對稱性，迫使理性節點趨向誠實策略，從而實現了無須依賴中心化仲裁的去中心化治理平衡。

- 打破安全性與通訊開銷的強耦合：本研究證明了「事前預防」轉向「事後追責」的效率優勢。在維持相同安全性邊界的前提下，CACA 允許系統在常態下僅維持

輕量級的小型委員會運作（如 $c = 5$），成功減少了約 44.4% 的通訊冗餘，為資源受限的邊緣運算場景提供具擴展性的防禦方案。

## 6.3 未來展望 (Future Work)

本研究提出的挑戰增強型委員會架構 (CACA) 在應對理性攻擊者時展現了優越的經濟防禦力。基於現有成果，未來研究可朝以下兩個方向進一步延伸：

### 6.3.1 聯邦學習自癒界限與災難性恢復機制

本研究目前仰賴聯邦學習本身的自癒能力來抵銷惡意梯度，並對攻擊者實施「僅懲罰不回滾」的策略以維持系統活性。然而，未來研究可進一步探討在更極端的攻擊行為（如旨在徹底毀滅模型的非理性拜占庭攻擊）下，自癒能力的失效界限。當「全棧投毒」場景注入的更新足以導致模型發生不可逆的發散時，如何設計一套高效的「模型回溯復原機制」將成為核心課題。此機制的挑戰在於，如何在偵測到災難性損害後，精準且低開銷地將模型狀態回溯至受攻擊前的檢查點，同時避免因頻繁回溯導致誠實節點的算力嚴重浪費。

### 6.3.2 針對多樣化應用情境之自適應委員會設計

本研究證實了小規模委員會配合挑戰機制能在常態下提供極高的效率。但在實際應用中，如低軌衛星網路 (LEO) 的通訊窗口限制、或是工業物聯網 (IoT) 中邊緣設備的異質資源約束，其面臨的威脅水平與環境壓力各不相同。未來研究可探討如何建構一套「自適應委員會」機制，根據當前網路的威脅監控數據與應用場景特徵，動態調整委員會的規模或選拔權重門檻。此方向的主要挑戰在於，如何在動態變化的環境中，始終維持足夠的經濟安全性 (Economic Security) 閾值，並確保效率優化不會因過度縮減委員會而產生不可預見的安全缺口。

# 參考文獻


[1] S. R. Pokhrel. “Blockchain Brings Trust to Collaborative Drones and LEO Satellites: An Intelligent Decentralized Learning in the Space”. In: *IEEE Sensors J.* 21.22 (2021), pp. 25331–25339.

[2] W. Wu, Z. Shen, et al. “A Sharded Blockchain-Based Secure Federated Learning Framework for LEO Satellite Networks”. In: *arXiv preprint arXiv:2411.06137* (2024).

[3] M. Elmahallawy and A. J. Akbarfam. “Decentralized Trust for Space AI: Blockchain-Based Federated Learning Across Multi-Vendor LEO Satellite Networks”. In: *arXiv preprint arXiv:2501.00000* (2025).

[4] Y. Lu et al. “Blockchain Empowered Asynchronous Federated Learning for Secure Data Sharing in Internet of Vehicles”. In: *IEEE Trans. Veh. Technol.* 69.4 (2020), pp. 4298–4311.

[5] H. Liu et al. “Blockchain and Federated Learning for Collaborative Intrusion Detection in Vehicular Edge Computing”. In: *IEEE Trans. Veh. Technol.* 70.6 (2021), pp. 6073–6084.

[6] S. R. Pokhrel and J. Choi. “Federated Learning With Blockchain for Autonomous Vehicles: Analysis and Design Challenges”. In: *IEEE Trans. Commun.* 68.8 (2020), pp. 4734–4746.

[7] Y. Lu et al. “Blockchain and federated learning for privacy-preserved data sharing in industrial IoT”. In: *IEEE Trans. Ind. Informat.* 16.6 (2020), pp. 4177–4186.

[8] Y. Qu et al. “Decentralized privacy using blockchain-enabled federated learning in fog computing”. In: *IEEE Internet Things J.* 7.6 (2020), pp. 5171–5183.

[9] W. Li et al. “EPP-BCFL: Efficient and Privacy-Preserving Blockchain-Based Federated Learning”. In: *Sci. Rep.* (2025).

[10] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.

[11] M. Wang et al. “A Blockchain-Based Federated Learning Framework for Vehicular Networks”. In: *Sci. Rep.* (2024).

[12] J. Zhang et al. “FedChain: A blockchain-based federated learning framework with adaptive client selection”. In: *Proc. VLDB Endow.* (2024).

[13] Z. Qin et al. “BlockDFL: A blockchain-based fully decentralized peer-to-peer federated learning framework”. In: *Proc. Web Conf. (WWW)*. Singapore, 2024, pp. 2914–2925.

[14] Y. Li et al. “A blockchain-based decentralized federated learning framework with committee consensus”. In: *IEEE Netw.* 35.1 (2021), pp. 234–241.

[15] M. Shayan et al. “Biscotti: A Blockchain System for Private and Secure Federated Learning”. In: *IEEE Trans. Parallel Distrib. Syst.* 32.7 (2021), pp. 1513–1525.

[16] J. Weng et al. “DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-Based Incentive”. In: *IEEE Trans. Dependable Secur. Comput.* 18.5 (2021), pp. 2438–2455.

[17] X. Li et al. “Enhancing Byzantine robustness of federated learning via tripartite adaptive authentication”. In: *J. Big Data* (2025).

[18] Z. Xing et al. “Zero-Knowledge Proof-based Verifiable Decentralized Machine Learning: A Comprehensive Survey”. In: *arXiv preprint arXiv:2312.00000* (2023).

[19] D. H. Nguyen et al. “FedBlock: A Blockchain Approach to Federated Learning against Backdoor Attacks”. In: *Proc. IEEE Big Data*. 2024.

[20] B. McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[21] Peter Kairouz et al. “Advances and open problems in federated learning”. In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210.

[22] Tian Li et al. “Federated learning: Challenges, methods, and future directions”. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60.

[23] P. Blanchard et al. “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent”. In: *NeurIPS*. 2017.

[24] L. Zhu, Z. Liu, and S. Han. “Deep Leakage from Gradients”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[25] Jonas Geiping et al. “Inverting Gradients - How easy is it to break privacy in federated learning?” In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 16937–16947.

[26] M. Fang et al. “Local Model Poisoning Attacks to Byzantine-Robust Federated Learning”. In: *Proc. USENIX Security*. 2020.

[27] H. Kim et al. “Blockchained on-device federated learning”. In: *IEEE Commun. Lett.* 24.6 (2020), pp. 1279–1283.

[28] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.

[29] Yi Liu et al. “Fedcoin: A peer-to-peer payment system for federated learning”. In: *arXiv preprint arXiv:2002.11711* (2020).

[30] A. Author et al. “A Blockchain-based Federated Learning Framework for Secure Aggregation and Fair Incentives”. In: *Taylor & Francis* (2024).

[31] Leslie Lamport, Robert Shostak, and Marshall Pease. “The Byzantine generals problem”. In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 4.3 (1982), pp. 382–401.

[32] Leslie Lamport. “Paxos made simple”. In: *ACM SIGACT News* 32.4 (2001), pp. 51–58.

[33] Diego Ongaro and John Ousterhout. “In search of an understandable consensus algorithm”. In: *2014 USENIX Annual Technical Conference (USENIX ATC 14)*. 2014, pp. 305–319.

[34] Miguel Castro and Barbara Liskov. “Practical Byzantine fault tolerance”. In: *OSDI*. Vol. 99. 1999. 1999, pp. 173–186.

[35] Maofan Yin et al. “HotStuff: BFT consensus with linearity and responsiveness”. In: *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing*. 2019, pp. 347–356.

[36] E. Buchman, J. Kwon, and Z. Milosevic. “The latest gossip on BFT consensus”. In: *arXiv preprint arXiv:1807.04938* (2018).

[37] Yossi Gilad et al. “Algorand: Scaling byzantine agreements for cryptocurrencies”. In: *Proceedings of the 26th Symposium on Operating Systems Principles*. 2017, pp. 51–68.

[38] B. J. Chen et al. “ZKML: An Optimizing System for ML Inference in Zero-Knowledge Proofs”. In: *Proc. EuroSys*. 2024.

[39] Ariel Gabizon, Zachary Williamson, and Oana Ciobotaru. *PLONK: Permutations over lagrange-bases for oecumenical noninteractive arguments of knowledge*. Cryptology ePrint Archive, Report 2019/953. 2019.

[40] B. Feng et al. “ZEN: An optimizing compiler for verifiable, zero-knowledge neural network inferences”. In: *Cryptology ePrint Archive* (2021).

[41] EZKL. *Benchmarking ZKML frameworks*. EZKL Blog. 2024. URL: `https://blog.ezkl.xyz/`.

[42] Y. Zhu et al. “RiseFL: Secure and Verifiable Data Collaboration with Low-Cost Zero-Knowledge Proofs”. In: *Proc. VLDB Endow.* 17.9 (2024), pp. 2321–2334.

[43] K. Conway et al. “opML: Optimistic Machine Learning on Blockchain”. In: *arXiv preprint arXiv:2401.00000* (2024).

[44] ORA Protocol. *opML documentation*. docs.ora.io. 2024.

[45] Optimism Foundation. *Rollup protocol overview*. docs.optimism.io. 2024.

[46] H. Chen et al. “Robust blockchained federated learning with model validation and proof-of-stake inspired consensus”. In: *arXiv preprint arXiv:2101.03300* (2021).

[47] Z. Peng et al. “VFChain: Enabling verifiable and auditable federated learning via blockchain systems”. In: *IEEE Trans. Netw. Sci. Eng.* 9.1 (2022), pp. 173–186.