



**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，而現有的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 a 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

摘要 . . . . .	i
ABSTRACT . . . . .	ii
誌謝 . . . . .	iii
目錄 . . . . .	iv
圖目錄 . . . . .	vii
表目錄 . . . . .	viii
第一章 緒論 (Introduction) . . . . .	1
第二章 背景知識與相關研究 . . . . .	5
2.1 聯邦學習與去中心化信任需求 . . . . .	5
2.1.1 聯邦學習的核心機制 . . . . .	5
2.1.2 中央化架構的信任困境 . . . . .	6
2.1.3 區塊鏈作為去中心化信任基礎設施 . . . . .	7
2.2 拜占庭容錯的理論基礎 . . . . .	8
2.2.1 拜占庭將軍問題與容錯閾值 . . . . .	8
2.2.2 實用拜占庭容錯協議的核心概念 . . . . .	9
2.2.3 BFT 共識在區塊鏈聯邦學習中的角色 . . . . .	9
2.3 區塊鏈聯邦學習的委員會架構演進 . . . . .	10
2.3.1 從全節點到委員會的效率驅動 . . . . .	10
2.3.2 委員會選舉機制的設計空間 . . . . .	11
2.3.3 委員會規模與安全性的權衡分析 . . . . .	12
2.3.4 基準系統模型：BlockDFL 委員會架構 . . . . .	14
2.4 現有驗證方法與其局限 . . . . .	17
2.4.1 密碼學驗證方法的運算瓶頸 . . . . .	17
2.4.2 樂觀執行方法的架構限制 . . . . .	18
2.4.3 委員會驗證方法的安全假設 . . . . .	19
2.4.4 系統性局限：靜態分析的盲區 . . . . .	20
2.5 研究缺口：從激勵設計到安全隱患 . . . . .	21

2.5.1	獎勵機制的正反饋特性 . . . . .	21
2.5.2	策略性攻擊者的潛在威脅 . . . . .	22
2.5.3	本研究的切入點 . . . . .	22
第三章	威脅模型 (Threat Model) . . . . .	24
3.1	攻擊者模型 . . . . .	24
3.1.1	攻擊者類型：理性攻擊者 . . . . .	24
3.1.2	攻擊者能力與限制 . . . . .	25
3.2	攻擊向量分析 . . . . .	26
3.2.1	資料層攻擊：已有防禦 . . . . .	26
3.2.2	共識層攻擊：本研究重點 . . . . .	27
3.2.3	攻擊層次對比 . . . . .	28
3.3	漸進式委員會佔領攻擊 (Progressive Committee Capture Attack) . . . . .	28
3.3.1	攻擊定義與核心機制 . . . . .	29
3.3.2	攻擊階段詳述 . . . . .	30
3.3.3	權益增長動態分析 (Stake Growth Dynamics Analysis) . . . . .	32
3.3.4	攻擊效果與影響 . . . . .	33
3.3.5	與傳統攻擊的區別 . . . . .	34
3.4	安全目標 . . . . .	36
3.4.1	防止委員會被惡意節點持續控制 . . . . .	36
3.4.2	確保誠實節點的權益公平增長 . . . . .	37
3.4.3	維持模型收斂性與準確性 . . . . .	37
3.4.4	保持系統的去中心化特性 . . . . .	38
3.4.5	激勵相容性 . . . . .	38
3.5	本章小結 . . . . .	39
第四章	挑戰增強型委員會架構 (Challenge-Augmented Committee Architecture) . .	40
4.1	系統架構概覽 . . . . .	41
4.2	異步審計與究責機制 . . . . .	44
4.3	安全性保證 . . . . .	45
4.4	效率分析 . . . . .	47

4.4.1	BlockDFL 的效率瓶頸：安全性與委員會規模的強耦合 . . . . .	47
4.4.2	CACA 的突破：從門檻安全性到經濟安全性 . . . . .	48
4.4.3	通訊複雜度對比分析 . . . . .	48
4.4.4	效率提升的本質：架構層面的解耦創新 . . . . .	49
4.5	激勵機制 . . . . .	50
4.6	本章小結 . . . . .	52
第五章	實驗評估 (Experimental Evaluation) . . . . .	53
5.1	實驗設置 . . . . .	53
5.1.1	資料集與模型配置 . . . . .	53
5.1.2	基準方法與攻擊場景 . . . . .	54
5.1.3	實驗參數配置 . . . . .	54
5.2	實驗結果與分析 . . . . .	55
5.2.1	經濟安全性的核心驗證：權益動態分析 . . . . .	55
5.2.2	攻擊抑制效果：頻率與成功率分析 . . . . .	57
5.2.3	系統穩定性與模型品質的連帶效益 . . . . .	59
5.3	本章小結 . . . . .	60
第六章	結論與未來展望 (Conclusion and Future Work) . . . . .	62
6.1	研究總結 (Summary of Research) . . . . .	62
6.2	研究發現與貢獻 (Research Findings and Contributions) . . . . .	62
6.3	未來展望 (Future Work) . . . . .	63
6.3.1	聯邦學習自癒界限與災難性恢復機制 . . . . .	63
6.3.2	針對多樣化應用情境之自適應委員會設計 . . . . .	63
參考文獻	. . . . .	65



## 圖目錄

4.1	Challenge-Augmented Committee Architecture (CACA) 系統架構與工作流程圖	42
5.1	BlockDFL 與 CACA 權益演化對比（300 輪基礎實驗） . . . . .	56
5.2	2000 輪長期模擬下的權益動態演化 . . . . .	57
5.3	模型準確率收斂曲線比較 . . . . .	59

## 表目錄

2.1	BCFL 驗證方法比較 . . . . .	21
3.1	攻擊層次對比 . . . . .	28
3.2	與傳統攻擊的區別 . . . . .	35
4.1	BlockDFL 與 CACA 在相同安全性水平下的效率對比 ( $N = 100, f = 30\%, p_{risk} < 0.01$ ) . . . . .	49
5.1	實驗參數配置 . . . . .	55

# 第一章 緒論 (Introduction)

隨著人工智慧與分散式運算技術的蓬勃發展，聯邦學習 (Federated Learning) [1] 與區塊鏈技術的深度整合，催生了區塊鏈賦能的聯邦學習 (Blockchain-based Federated Learning, BCFL) 這一創新技術典範。此技術路徑之所以受到學界與產業界的高度重視，根本原因在於其成功回應了多方互不信任情境下協作機器學習的核心挑戰。在低軌衛星網路 (LEO) [2, 3, 4]、車聯網 (V2X) [5, 6, 7] 以及工業物聯網 (IIoT) [8, 9, 10] 等實際應用場景中，BCFL 展現了其獨特且難以替代的技術價值。以 LEO 衛星星座為代表的太空人工智慧應用場景尤具說明性意義，在此類場景中星地通訊窗口通常僅維持約五分鐘，且下行頻寬受限於每秒 8 Mbps 左右的水準 [3]，這種嚴苛的通訊條件使得傳統依賴地面站進行模型聚合的訓練方案在技術上幾乎無法實現。BCFL 透過在異質衛星營運商之間建立去中心化的信任基礎設施，成功將模型收斂所需時間縮減達三十小時之譜 [4]，充分展現了此技術路徑在極端環境下的應用潛力。類似地，在工業 4.0 的發展背景下，BCFL 使得協作工廠能夠在完全不洩露各自商業機密的前提下共同進行預測性維護模型的訓練，實驗數據顯示此種架構可將通訊開銷較傳統集中式方案降低約百分之四十一 [8]。上述多元應用場景共同呈現出三項關鍵特徵：缺乏可信賴的中心化協調者、計算與通訊資源高度受限、以及訓練資料在統計分布上的顯著異質性，正是這些特徵的交織作用，使得 BCFL 逐步確立其作為通用去中心化學習架構首選方案的地位。

然而，當 BCFL 系統嘗試邁向大規模實際部署時，卻面臨著嚴峻的效能瓶頸，此問題在學術文獻中通常被稱為「可擴展性兩難困境」。究其根源，目前絕大多數 BCFL 系統採用實用拜占庭容錯協議 (Practical Byzantine Fault Tolerance, PBFT) [11] 或其各種變體作為底層共識機制，而此類協議固有的  $O(n^2)$  訊息複雜度特性，導致系統效能隨著參與節點數量的增加而呈現急劇下降的趨勢。FLCoin [12] 的實證研究提供了具體的量化證據：當參與節點數量達到一百個時，單一輪次的共識過程所產生的訊息交換量將突破兩萬條，共識延遲因而攀升至二十五秒以上，此延遲水準已經達到甚至超過模型訓練本身所需時間的量級，對於要求即時響應的應用場景而言構成了根本性的障礙。在車載網路 (VANET) 的極端測試條件下，一百輛車輛進行 BCFL 協作訓練將產生高達 360.57 MB 的資料傳輸量，單輪訓練的總通訊時間開銷達到 19.51 秒 [13]，這對於需要頻繁進行模型更新的車載智慧系統而言顯然是無法接受的效能表現。除了通訊效能問

題之外，區塊鏈節點對於儲存空間的高需求同樣構成嚴峻挑戰，例如比特幣全節點需要約 200 GB 的儲存空間，而以太坊更是超過 465 GB，這與邊緣運算設備通常僅具備 KB 至 MB 等級記憶體容量的現實形成了尖銳的衝突 [14]。效能與資源的這種雙重束縛，使得要求所有節點參與驗證的傳統全節點架構在實際工業部署情境中越來越難以為繼，迫切需要創新的技術方案來突破這些根本性的限制。

正是為了化解上述可擴展性挑戰，學術界近年來將研究重心轉向「委員會機制 (Committee Mechanism)」的探索與發展，此機制的核心設計理念是將原本由全體節點共同承擔的驗證責任，集中委派給一個規模相對較小的驗證者委員會來執行。這種架構轉變的根本邏輯在於，透過大幅減少實際參與共識決策過程的節點數量，可以顯著降低系統的通訊複雜度與計算負擔。目前學界提出的主流委員會選拔機制包含數種不同的技術路徑：基於雜湊環結構的隨機抽樣方法 [15]、依據持有代幣數量或權益進行加權選舉的方法 [16, 12]、以及利用可驗證隨機函數 (VRF) 實現的 Sortition 機制 [17, 18, 19]。委員會機制的導入確實為系統效能帶來了立竿見影的改善效果，FLCoin [12] 的研究報告顯示，透過採用滑動窗口選舉策略，系統成功將通訊開銷降低了百分之九十，並實現了 5.7 倍的訓練速度提升；BFLC [16] 則藉由委員會驗證機制將共識延遲穩定控制在三秒以內。這些最佳化成果成功地將通訊複雜度從全網共識的  $O(n^2)$  降至與委員會規模  $C$  相關的  $O(C^2)$  甚至  $O(C)$  等級，極大提升了系統的吞吐量與響應速度。然而，這種為追求效率而將驗證權力集中於少數委員會成員的架構轉變，也在無形中引入了全新的安全攻擊面，對系統的長期安全性與穩定性構成了潛在威脅，而這些威脅至今尚未獲得學術界的充分探討。

深入剖析上述安全隱憂可以發現，現有委員會防禦機制普遍存在一個容易被研究者忽視的結構性弱點，即對「誠實多數假設」的過度依賴。儘管當前的區塊鏈聯邦學習系統已經廣泛導入 Krum、Trimmed Mean 或 Median 等具備拜占庭強健性的聚合演算法 [20, 21]，但這些演算法能夠發揮預期防禦效果的根本前提是：執行聚合運算的委員會成員中，誠實節點必須穩定維持佔多數的地位。當前安全性研究的主要關注焦點集中在惡意客戶端上傳毒化梯度的資料層攻擊，卻在很大程度上忽略了一種更為隱蔽的威脅形式，即委員會成員之間可能形成的策略性共謀行為。一旦攻擊者成功控制委員會中超過三分之二的席位，便能夠串聯起來繞過所有現有的資料層強健聚合邏輯，甚至可以任意偽造聚合結果而無需擔心遭到系統的懲罰或追責。正如 FedBlock [22] 所指出，

當任何參與者都可能成為驗證者時，系統不能僅仰賴誠實多數假設，而必須具備主動檢測並隔離惡意驗證者的能力，然而這項關鍵機制在現有的區塊鏈聯邦學習架構中普遍付之闕如，這揭示了當前安全防禦體系在共識決策層面的本質脆弱性。

上述分析揭示了當前研究領域中存在的一個關鍵「研究缺口」：現有的 BCFL 系統普遍缺乏能夠應對「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」的自我修復與防禦機制。在 PCCA 攻擊模式中，對手並非採取直接的暴力破壞策略，而是實施一種被稱為「策略性餓死 (Strategic Starvation)」的隱蔽手法，具體而言，攻擊者在成功掌控委員會之後，會優先處理與自身經濟利益相關的模型更新請求，同時系統性地拒絕為誠實參與者提供驗證服務，藉此操縱系統的獎勵分配機制與權益動態變化。值得特別關注的是，現有的各種解決方案大多僅聚焦於特定的防禦維度，例如基於同態加密 (Homomorphic Encryption, HE) [23, 24] 或權益證明 (Proof of Stake, PoS) [25] 的技術方案，雖然在隱私保護或准入控制等特定面向上確實有所突破，卻無法在委員會本身已經淪陷而不再可信的情況下，繼續保證模型聚合更新的正確性以及系統資源分配的公平性。由於缺乏事後的「可追溯審計能力」與「有效經濟威懾手段」，一旦誠實多數假設在某一輪次中被攻破，現有系統便缺乏任何機制來識別並懲罰惡意行為者，使得攻擊者得以在後續輪次中持續佔據優勢地位。因此，如何將系統的安全性保障與對共識節點集體誠實性的依賴進行有效解耦，已成為實現真正意義上去中心化人工智慧平台所必須跨越的關鍵技術門檻。

針對這一嚴峻的技術挑戰，本論文提出一套名為「挑戰者增強委員會架構 (Challenge-Augmented Committee Architecture, CACA)」的創新防禦體系，旨在為區塊鏈聯邦學習系統建構一道動態且具備自我修復能力的安全屏障。本研究的核心設計理念在於，透過引入異步審計機制與內部罰沒協議，賦予現有委員會架構檢測並汰除惡意驗證者的能力。這種設計使得系統能夠將「活性 (Liveness)」與「安全性 (Security)」這兩項往往相互牽制的系統屬性進行有效的解耦處理。在本架構下，即使委員會在特定輪次中並非完全可信甚至已經被惡意節點所捕獲，系統仍然能夠透過分散部署於全網的挑戰者網路，對委員會所做出的每一項決策進行事後審計與追責，進而檢舉並懲罰任何偽造或誤導性的聚合結果。這種創新機制不僅大幅提升了系統面對策略性攻擊時的韌性，更為去中心化環境下的模型協作訓練提供了一層額外且獨立的安全性保障，確保了系統在追求效率提升的同時，不會犧牲去中心化架構所應堅守的核心安全原則。

進一步而言，本研究在理論分析、協議設計與實驗驗證三個相互支撐的維度上均做出了具體而紮實的學術貢獻。在理論分析層面，本研究首次對漸進式委員會佔領攻擊 (PCCA) 模式進行了形式化的數學定義，並透過系統性的模擬實驗量化評估了該攻擊對系統長期激勵相容性所造成的深層破壞效應，藉此揭露了傳統防禦機制在面對具有演化特性的策略性攻擊時所存在的根本局限。基於上述理論發現，本研究進一步結合博弈論模型設計了一套精密的內部罰沒協議 (Internal Slashing Protocol)，此協議透過精心設計的經濟懲罰機制，確保任何節點進行審計驗證所需付出的成本始終低於其從事惡意行為可能獲取的收益，從而促使誠實行為成為所有理性參與者在長期重複博弈下的唯一納什均衡策略 [26]。這些理論層面的創新貢獻在後續的大規模模擬實驗中獲得了充分的實證支持。實驗數據顯示，即使在百分之三十節點進行惡意共謀的極端對抗環境下，採用 CACA 架構的系統仍能將模型最終準確率穩定維持在百分之九十八點六以上的水準，同時在達成相同安全性保障指標的前提下將通訊開銷降低了百分之四十四點四，並將系統的最低不可用率從百分之二十大幅壓制至百分之五以下，顯著提升了系統在對抗性環境中的強健性與運作效率。

本論文後續章節的組織架構安排如下。在本緒論之後，第二章將系統性地奠定本研究所需的理論基礎，內容涵蓋聯邦學習的核心概念、區塊鏈共識機制的運作原理、以及拜占庭容錯技術的數學基礎，並針對現有去中心化聯邦學習領域的相關文獻進行批判性的回顧與評述。以第二章建立的知識基礎為起點，第三章將詳細定義本研究所採用的系統模型與威脅模型，並針對 PCCA 攻擊者的行為特徵、能力邊界與策略空間進行嚴謹的形式化描述。在完成理論模型的建構之後，第四章將提出 CACA 架構的核心設計方案，完整闡明具體的協議運作流程、關鍵演算法的設計細節、以及支撐整體架構的安全性證明邏輯。緊接著，第五章將透過一系列精心設計的針對性實驗與全面的數據比較分析，在多種不同的對抗場景設定下驗證本架構所具備的效能優勢與系統韌性。最終，第六章將總結本研究的主要發現與學術貢獻，並就本機制在未來分散式人工智慧生態系統中的應用潛力與可能的演進方向進行前瞻性的探討。

## 第二章 背景知識與相關研究

本章旨在建立理解區塊鏈聯邦學習委員會安全性所需的理論基礎與技術背景。首先，本章將從聯邦學習的核心價值出發，闡明中央化架構面臨的信任困境，進而說明區塊鏈技術如何作為去中心化信任的基礎設施。接著，本章將介紹拜占庭容錯理論的基本原理，為理解委員會共識機制的安全性閾值提供數學基礎。在此基礎上，本章將探討區塊鏈聯邦學習如何從全節點共識演進至委員會架構，並分析委員會規模與安全性之間的權衡關係。隨後，本章將詳細介紹本研究採用的基準系統模型 BlockDFL 之委員會架構，其內容涵蓋角色定義、運作流程與獎勵機制。最後，本章將回顧現有驗證方法的局限性，指出當前研究在面對策略性攻擊者時的盲區，從而定位本研究欲填補的關鍵缺口。

### 2.1 聯邦學習與去中心化信任需求

#### 2.1.1 聯邦學習的核心機制

聯邦學習是一種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 [1]。在傳統的集中式機器學習中，所有訓練資料必須彙集至中央伺服器進行處理，這種做法在面對隱私敏感資料或資料傳輸成本高昂的場景時顯得力不從心。聯邦學習透過將訓練過程分散至資料所在的終端裝置，僅將模型更新而非原始資料上傳至伺服器進行聚合，從根本上改變了資料與運算的關係。這種架構使得醫療機構能夠在不共享病患紀錄的前提下協同訓練診斷模型，金融機構能夠在不揭露客戶交易資料的情況下建立風險評估系統，行動裝置製造商能夠利用數百萬用戶的使用習慣最佳化輸入法預測，而無需將敏感的打字內容上傳至雲端。

聯邦學習的標準訓練流程可概括為四個階段的迭代循環。在每一輪訓練中，中央伺服器首先將當前的全域模型參數分發給選定的客戶端；各客戶端隨後在本地私有資料上執行若干輪梯度下降，產生反映本地資料特性的模型更新；客戶端將這些更新上傳至伺服器；伺服器執行聚合演算法（最常見的是 FedAvg [1]）將各客戶端的更新整合為新的全域模型。此循環持續進行直至模型收斂或達到預設的訓練輪數。值得注意的

是，聯邦學習面對的資料分布通常具有高度異質性：不同客戶端持有的資料量可能相差懸殊，資料的類別分布也往往呈現顯著差異，這種非獨立同分布（Non-IID）的特性為模型訓練與安全防護帶來了獨特的挑戰 [27]。

## 2.1.2 中央化架構的信任困境

儘管聯邦學習在資料隱私保護上取得了重要進展，其標準架構仍存在一個根本性的信任假設：所有參與者必須信任中央聚合伺服器會誠實地執行聚合運算並正確地分發結果。然而，在缺乏有效驗證機制的情況下，這項假設構成了系統的單點脆弱性。中央伺服器可能因遭受攻擊、內部人員惡意行為或系統故障而偏離預期行為，而客戶端對此幾乎無從察覺，更遑論採取補救措施。

中央化架構面臨的信任風險可歸納為三個層面。第一個層面是聚合正確性的不可驗證性。當伺服器宣稱某一全域模型是由特定客戶端更新聚合而成時，客戶端無法獨立驗證此宣稱的真實性。伺服器可能執行選擇性聚合，意即僅納入部分客戶端的更新而排除具體，或者直接篡改聚合結果以植入後門。研究已證實，透過精心設計的模型修改，攻擊者可在不顯著影響主任務效能的情況下，使模型對特定輸入產生預設的錯誤輸出 [28]。第二個層面是單點故障風險。中央伺服器一旦因攻擊、硬體故障或網路問題而離線，整體訓練流程即刻中斷，且由於缺乏分散式的狀態同步機制，系統難以從中間狀態恢復。第三個層面是隱私保護的局限性。儘管聯邦學習避免了原始資料的直接傳輸，研究表明惡意的聚合伺服器仍可能透過分析客戶端提交的模型更新，推論出關於訓練資料的敏感資訊 [29]。

這些信任風險在跨組織協作的場景中尤為突出。當多個相互獨立甚至存在競爭關係的機構希望聯合訓練模型時，由任何單一機構擔任中央聚合者都難以獲得其他參與者的充分信任。即便引入第三方作為中立的聚合服務提供者，仍無法從根本上消除對該第三方誠實性的依賴。這種信任困境限制了聯邦學習在高價值、高敏感場景中的應用潛力，也促使研究者開始探索去中心化的替代方案。



### 2.1.3 區塊鏈作為去中心化信任基礎設施

區塊鏈技術具備不可篡改性、透明性與去中心化這三項核心特性，恰好對應了中央化聯邦學習面臨的信任困境，使其成為建構去中心化聯邦學習系統的理想基礎設施。不可篡改性源於區塊鏈的鏈式雜湊結構：每個區塊包含前一區塊的雜湊值，任何對歷史資料的修改都將導致後續所有區塊的雜湊值連鎖變化，從而被網路中的其他節點立即偵測。這項特性確保了一旦聚合結果被記錄於區塊鏈，便無法在事後被悄然篡改。透明性則意味著所有被記錄的交易與狀態變更對全體參與者可見，客戶端可以驗證自己的更新是否被納入聚合，也可以審計歷史聚合過程是否遵循預定的規則。去中心化消除了對單一實體的信任依賴：區塊鏈網路由眾多獨立節點共同維護，即使部分節點失效或行為惡意，只要誠實節點佔據多數，系統仍能正確運作。

將區塊鏈整合至聯邦學習架構，可從多個層面強化系統的可信賴性。在聚合正確性方面，智能合約可編碼確定性的聚合規則，確保聚合過程按照預定邏輯執行，而非依賴聚合者的自我約束。聚合結果連同參與者資訊被記錄於區塊鏈，形成永久可查的審計軌跡。在系統可用性方面，區塊鏈的分散式架構天然具備容錯能力：即使部分節點離線，其他節點仍可維持系統運作，避免了中央伺服器故障導致的全面停擺。在激勵對齊方面，區塊鏈原生的代幣機制可用於設計精細的獎懲制度，對誠實貢獻者給予獎勵，對惡意行為者施加經濟懲罰，從而在博弈論意義上引導參與者趨向誠實行為。

然而，區塊鏈並非萬能的信任解決方案。區塊鏈共識機制本身需要假設惡意節點不超過特定比例；以拜占庭容錯協議為例，此閾值通常設定為三分之一。當攻擊者控制的節點超過此閾值時，區塊鏈的安全性保證將不再成立。此外，區塊鏈的共識過程涉及大量的節點間通訊，其延遲與頻寬成本可能與聯邦學習對快速迭代的需求產生張力。這些考量促使研究者發展出委員會架構等效率最佳化方案，但也隨之引入了新的安全性議題。下一節將首先介紹拜占庭容錯的理論基礎，為理解這些安全性議題提供必要的背景知識。

## 2.2 拜占庭容錯的理論基礎

### 2.2.1 拜占庭將軍問題與容錯閾值

**拜占庭將軍問題背景** 拜占庭將軍問題由 Lamport、Shostak 與 Pease 於 1982 年正式提出 [30]，是分散式系統容錯理論的基石。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：如何設計一個協議，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：所有忠誠節點必須就相同的值達成共識（一致性），且若發起者是誠實的，則共識結果必須是發起者提出的值（正確性）。

**數學限制與 Quorum 交叉原理** 拜占庭將軍問題存在一個根本性的數學限制：在僅使用口頭訊息的情況下，問題可解若且唯若誠實節點超過總數的三分之二。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。考慮指揮官向兩位副官發送命令的情境：若指揮官是叛徒，他可能向副官 A 發送「進攻」，向副官 B 發送「撤退」；當兩位誠實副官相互交換收到的命令時，各自都會發現矛盾。然而，若副官 B 是叛徒而指揮官誠實，副官 B 可能向副官 A 謊稱「指揮官說撤退」。關鍵的洞察在於：從副官 A 的視角來看，這兩種情境完全無法區分，因為在兩種情況下，他都收到來自指揮官的「進攻」訊息，以及來自 B 聲稱收到「撤退」的訊息。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。

此三分之一閾值的數學根源在於 Quorum 交叉原理。為確保任何決策都獲得足夠的誠實節點背書，系統需要收集至少  $2f + 1$  個節點的確證。由於最多  $f$  個節點可能是惡意的， $2f + 1$  個確證中必然包含至少  $f + 1$  個來自誠實節點。任意兩個大小為  $2f + 1$  的節點群體，其交集至少包含  $f + 1$  個節點，這確保了至少有一個誠實節點見證了兩次決策，從而防止系統對同一問題做出矛盾的決定。將節點總數代入約束條件  $n \geq 2f + 1 + f$ ，即得  $n \geq 3f + 1$ 。

## 2.2.2 實用拜占庭容錯協議的核心概念

拜占庭將軍問題的早期解法雖然在理論上可行，但其指數級的通訊複雜度使其僅具學術意義。1999 年，Castro 與 Liskov 提出實用拜占庭容錯協議（Practical Byzantine Fault Tolerance, PBFT）[11]，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。

PBFT 協議的運作依賴  $n = 3f + 1$  個副本節點，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：預準備（Pre-prepare）、準備（Prepare）與提交（Commit）。在預準備階段，主節點將請求連同分配的序號廣播給所有副本；在準備階段，收到預準備訊息的副本向其他所有副本廣播準備訊息，當某副本收集到  $2f$  個匹配的準備訊息時，表明系統中有足夠多的節點認可此請求的序號分配；在提交階段，進入準備狀態的副本廣播提交訊息，當收集到  $2f + 1$  個提交訊息時，副本確信此請求已被系統接受，可以執行並回覆客戶端。三階段設計的核心目的是確保即使主節點是惡意的，也無法導致誠實節點對請求順序產生分歧。

PBFT 的通訊複雜度為  $O(n^2)$ ，這是因為在準備與提交階段，每個節點都需要向其他所有節點發送訊息。以  $n = 7$ （可容忍 2 個拜占庭節點）的配置為例，每輪共識約需交換 100 則訊息；當節點數增至  $n = 22$ （可容忍 7 個拜占庭節點）時，訊息數增至約 900 則。這種二次方增長限制了 PBFT 在大規模網路中的直接應用，實務部署通常限制在 10 至 20 個節點的規模。後續研究如 HotStuff [31] 透過流水線化設計與閾值簽章技術，將複雜度進一步降至  $O(n)$ ，但在本研究關注的許可制聯盟鏈場景中，節點數量通常在 PBFT 可承受的範圍內。

## 2.2.3 BFT 共識在區塊鏈聯邦學習中的角色

在區塊鏈聯邦學習系統中，拜占庭容錯共識扮演著確保聚合結果正確性的關鍵角色。與傳統區塊鏈應用（如加密貨幣交易）不同，聯邦學習的「交易」是模型更新，而「帳本狀態」是全域模型參數。當負責聚合的節點可能被攻陷或本身即為惡意參與者時，系統需要一個機制來驗證聚合結果的正確性，並在多個可能存在分歧的結果中達

成共識。BFT 協議正是為此目的而設計：它確保只要惡意節點不超過總數的三分之一，系統就能就聚合結果達成一致，且該結果必然反映誠實多數的判斷。

然而，將 BFT 共識直接應用於大規模聯邦學習系統面臨顯著的效率挑戰。聯邦學習通常涉及數十至數百個參與者，若所有參與者都參與每一輪的 BFT 共識， $O(n^2)$  的通訊複雜度將成為嚴重的效能瓶頸。更重要的是，聯邦學習需要頻繁迭代（典型的訓練過程可能包含數百至數千輪），因此每輪都執行完整的全網共識將導致訓練時間大幅延長。這種效率與安全性之間的張力，促使研究者發展出委員會架構：由一個小型的代表性子集執行共識，以較低的通訊成本達成近似的安全保證。下一節將詳細探討這種架構演進及其伴隨的安全性權衡。

## 2.3 區塊鏈聯邦學習的委員會架構演進

### 2.3.1 從全節點到委員會的效率驅動

區塊鏈聯邦學習的早期研究嘗試將傳統 BFT 共識直接應用於全體參與者，但很快便遭遇了可擴展性的瓶頸。以 BFLC [16] 的實驗配置為例，當參與者數量達到 20 個時，採用完整 PBFT 共識的每輪延遲已超過 100 毫秒；若將參與者擴展至數百個規模，共識延遲將增長至秒級甚至更長，這對於需要快速迭代的聯邦學習訓練而言顯然無法接受。更根本的問題在於通訊頻寬的消耗：每輪共識中，每個節點都需要接收並處理來自其他所有節點的訊息，當節點數量增加時，網路負擔呈平方級增長，這在頻寬受限的邊緣運算環境中尤為致命。

委員會架構的核心理念是將共識責任委派給一個規模遠小於全網的代表性子集。令全網節點數為  $n$ ，委員會規模為  $c$ ，其中  $c \ll n$ 。委員會內部執行 BFT 共識的通訊成本為  $O(c^2)$ ，委員會決策結果廣播至全網的成本為  $O(n)$ ，總通訊成本為  $O(c^2 + n)$ 。當  $c$  維持在較小的常數（如 7 至 20）時，此成本近似於線性  $O(n)$ ，相較於全節點 PBFT 的  $O(n^2)$  實現了數量級的改善。FLCoin [32] 的實驗資料印證了這一分析：在 500 個節點的網路中，採用規模為 100 的滑動視窗委員會，相較於全節點共識可減少約 90% 的通訊開銷，共識延遲維持在 3 秒以內。

委員會架構的效率優勢使其迅速成為區塊鏈聯邦學習的主流設計典範。然而，這

種效率的提升並非沒有代價：系統的安全性不再由全網的誠實多數保證，而是取決於委員會的組成是否可信。若攻擊者能夠控制委員會中超過三分之一的席位，便可操控共識結果，通過惡意的聚合提案或拒絕誠實的提案。這種從「全網安全」到「委員會安全」的轉變，將安全性分析的焦點從「全網惡意節點比例」轉移至「委員會選舉機制的抗操控能力」。

## 2.3.2 委員會選舉機制的設計空間

委員會選舉機制決定了哪些節點將被選入委員會，其設計直接影響系統的安全性與公平性。現有方案大致可分為四種取向：隨機選擇、權益導向、聲譽導向與貢獻導向，各有其優勢與潛在風險。

隨機選擇機制透過密碼學隨機數決定委員會組成，其核心優勢在於不可預測性：攻擊者無法提前知曉哪些節點將被選中，因而難以針對性地部署攻擊。RapidChain [33] 採用分散式隨機數生成協議，確保選舉結果對所有參與者而言都是不可預測且可驗證的。然而，純粹的隨機選擇可能將惡意或低品質的節點選入委員會，且無法反映節點過去的行為表現。權益導向機制將選中機率與節點持有的權益（stake）掛鉤，持有越多權益的節點越可能被選入委員會。這種設計的理論基礎是經濟激勵對齊：高權益節點若行為惡意將面臨更大的經濟損失，因此傾向誠實。以太坊 2.0 的驗證者選舉即採用此機制 [34]。然而，權益導向可能導致「富者愈富」的中心化傾向，且無法防範願意承受經濟損失的攻擊者。

聲譽導向機制根據節點的歷史行為表現運算聲譽分數，高聲譽者優先被選入委員會。BESIFL [25] 追蹤各節點提交更新的品質，將聲譽作為委員會選舉的權重。此機制能有效過濾曾有惡意行為記錄的節點，但也面臨兩項挑戰：新加入者缺乏歷史記錄，可能陷入「冷啟動」困境；更重要的是，策略性攻擊者可透過長期的誠實行為累積聲譽，待時機成熟後再發動攻擊。貢獻導向機制以節點對聯邦學習的實質貢獻（如訓練資料量、模型品質）作為選舉依據。FLCoin [32] 的滑動視窗機制即屬此類：節點透過提交有效的模型更新獲得「份額」，在視窗內持有份額的節點組成委員會。這種設計與聯邦學習的目標直接對齊，但貢獻指標可能被博弈；例如，攻擊者可先提交高品質更新以獲取委員會席位，再利用此席位通過惡意提案。

實際系統通常結合多種機制以平衡各方考量。BlockDFL [15] 採用「權益加權的確

定性隨機選擇」：選舉結果由前一區塊雜湊決定（確定性），但各節點被選中的機率與其權益成正比（權益加權）。這種混合設計試圖兼顧不可預測性與經濟激勵，但也繼承了權益導向機制的潛在風險，亦即攻擊者可能透過累積權益逐步提高其影響力。

### 2.3.3 委員會規模與安全性的權衡分析

委員會規模的選擇涉及效率與安全性之間的核心權衡，較小的委員會帶來更低的通訊成本與更快的共識速度，但也更容易被攻擊者滲透，而較大的委員會提供更強的安全保證，卻犧牲了效率優勢。理解這一權衡需要從機率論的角度分析委員會被攻破的風險，而超幾何分佈為此提供了精確的數學工具。當從  $N$  個節點（其中  $fN$  個為惡意節點）中隨機選取  $C$  個組成委員會時，由於委員會成員的選擇是一個無放回抽樣過程，委員會中恰有  $k$  個惡意節點的機率遵循超幾何分佈，其機率質量函數可表示為：

$$P(X = k) = \frac{\binom{fN}{k} \binom{(1-f)N}{C-k}}{\binom{N}{C}} \quad (2.1)$$

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運算了選擇  $k$  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。

對於採用 PBFT 共識的委員會而言，安全性分析需要區分兩種不同的威脅閾值。第一種閾值為三分之一，當惡意節點超過委員會的三分之一時，攻擊者能夠阻止委員會達成任何共識，因為 PBFT 協議要求至少  $2f + 1$  個節點同意才能通過提案，這種攻擊形式本質上是一種拒絕服務攻擊，雖然無法注入惡意內容，但能夠癱瘓系統的正常運作。第二種閾值為三分之二，這是更為嚴重的威脅情境，當惡意節點佔據委員會超過三分之二的席位時，攻擊者不僅能夠阻止誠實提案通過，更能夠強制通過惡意提案，完全控制委員會的決策結果。在本研究所關注的漸進式委員會佔領攻擊情境中，攻擊者的目標正是達成後者，透過控制委員會來通過有利於自身的提案並排除誠實節點的更新，因此後續的風險分析將以三分之二作為委員會被惡意控制的臨界閾值。

基於上述分析，委員會被惡意控制的風險機率  $P_{mal}$  可以表示為惡意節點數量達到

或超過  $\lfloor 2C/3 \rfloor + 1$  的累積機率：

$$P_{mal} = P(X \geq \lfloor 2C/3 \rfloor + 1) = \sum_{k=\lfloor 2C/3 \rfloor + 1}^C \frac{\binom{fN}{k} \binom{(1-f)N}{C-k}}{\binom{N}{C}} \quad (2.2)$$

為了具體理解這個機率模型的實際意涵，以下考察一個具代表性的數值案例。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。在這種條件下，不同委員會規模所對應的被惡意控制風險呈現出明顯的差異。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，透過超幾何分佈的運算，這種情況發生的機率約為 2.74%。當委員會規模增加到  $C = 7$  時，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。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%。這些資料揭示了兩個重要的洞察。其一，委員會規模與安全性之間並非線性關係，存在一個「甜蜜點」區間（約  $C = 9$  至  $C = 13$ ），在此區間內增加委員會規模能夠帶來顯著的安全性提升。其二，即使在相當高的惡意節點比例下，只需要一個規模適中的委員會就能將被惡意控制的風險壓制到相當低的水平。然而，委員會規模的增加直接推高了共識的通訊成本，由於 PBFT 的通訊複雜度為  $O(C^2)$ ， $C = 13$  的委員會其內部通訊量約為  $C = 5$  的 6.76 倍，這種成本增長在頻繁迭代的聯邦學習場景中尤為顯著。

上述分析揭示了委員會架構在傳統設計典範下面臨的根本性困境：在「門檻安全性」的框架內，系統設計者被迫在效率與安全性之間做出取捨，無法同時最大化兩者。然而，這種困境的存在源於一個隱含的假設，即安全性的保障必須依賴於「降低委員會被攻破的機率」。若能改變安全性的實現方式本身，使其不再受制於機率運算，則委員會規模與安全性之間的強耦合關係便有望被解構。這一觀察為第 4 章所提出的挑戰增強型委員會架構提供了理論切入點。

## 2.3.4 基準系統模型：BlockDFL 委員會架構

為深入分析區塊鏈聯邦學習中委員會機制的安全性，本研究採用 BlockDFL [15] 作為基準系統模型。BlockDFL 於 2024 年發表於 WWW 會議，代表當前完全去中心化點對點聯邦學習架構的最新進展。該系統透過角色分離、權益加權選舉與拜占庭容錯共識的結合，在效率與安全性之間取得了當前文獻中的最佳平衡。本節將詳細介紹其系統架構、運作流程與獎勵機制，作為後續威脅分析的基礎框架。

### 2.3.4.1 系統角色與職責定義

BlockDFL 採用角色分離的設計理念，將參與者依據其在每輪訓練中承動的職責劃分為三種角色：更新提供者（Update Provider）、聚合者（Aggregator）與驗證者（Verifier）。這種分工模式源於一個核心洞察：在去中心化環境中，若由單一節點同時負責訓練、聚合與驗證，將難以建立有效的制衡機制。透過將這三項職責分派給不同的參與者群體，系統得以在各環節引入相互監督，降低單點惡意行為對全域模型的影響。

更新提供者構成系統中的多數參與者，其職責是利用本地私有資料執行模型訓練，並將訓練所得的模型更新提交給聚合者。由於訓練資料始終保留在本地裝置，更新提供者的隱私得以保護，這體現了聯邦學習「資料不動、模型動」的核心價值。聚合者的職責則是收集來自多個更新提供者的本地更新，執行篩選與聚合運算，將結果打包為全域更新提案並提交給驗證者。每輪訓練中可能有多個聚合者同時運作，各自獨立收集更新並提交競爭性的提案，這種設計避免了單一聚合者壟斷聚合權的風險。驗證者組成委員會，負責評估各聚合者提交的提案品質，並透過拜占庭容錯共識機制選出最佳提案寫入區塊鏈。驗證者的數量通常遠小於更新提供者，以維持共識效率。

角色的分配並非固定不變，而是在每輪訓練開始時根據上一區塊的雜湊值與各參與者的權益（stake）重新決定。具體而言，區塊雜湊被映射至一個雜湊環，每個參與者依據其權益大小佔據環上相應比例的空間。系統依序從雜湊環上選出聚合者與驗證者，未被選中者則成為更新提供者。這種機制確保了角色分配的確定性與不可預測性：一方面，給定相同的區塊雜湊與權益分布，角色分配結果完全確定，便於驗證；另一方面，由於區塊雜湊在區塊產生前無法預知，參與者難以提前操控自身角色。更重要的



是，權益加權的設計使得持有較多權益的參與者更有可能被選為聚合者或驗證者，這反映了系統對「高權益者傾向誠實」的信任假設。

### 2.3.4.2 訓練輪次的運作流程

BlockDFL 的每輪訓練遵循一個結構化的流程，從角色分配開始，經由本地訓練、聚合、驗證與共識，最終完成全域模型更新與獎勵分配。圖 ?? 呈現了此流程的概念架構，以下將逐一說明各階段的運作邏輯。

訓練輪次始於角色分配階段。當新一輪開始時，所有參與者根據最新區塊的雜湊值與當前權益分布，確定性地運算出本輪的角色分配結果。由於運算過程僅依賴公開可驗證的資訊，任何參與者皆可獨立驗證角色分配的正確性，無需依賴中央協調者。角色確定後，更新提供者隨即進入本地訓練階段，在各自的私有資料集上執行若干輪隨機梯度下降，產生本地模型更新並將其發送給聚合者。

聚合階段是 BlockDFL 架構的關鍵環節。每個聚合者獨立收集來自更新提供者的本地更新，當收集數量達到預設閾值後，開始執行篩選與聚合程序。篩選的目的在於過濾潛在的惡意更新：聚合者首先依據更新提供者的權益進行抽樣，優先納入來自高權益節點的更新；接著透過本地推論測試評估各更新的品質，排除表現異常者。通過篩選的更新被聚合為一個全域更新提案，由聚合者簽署後提交給驗證者委員會。值得注意的是，由於多個聚合者同時運作且各自獨立收集更新，同一輪中將產生多個競爭性的提案，這為後續的驗證階段提供了選擇空間。

驗證與共識階段決定了哪個提案將被接受並寫入區塊鏈。當驗證者委員會收到足夠數量的提案後，驗證程序啟動。每位驗證者獨立使用 Krum 演算法 [20] 對所有提案進行評分，Krum 分數較低者代表與其他提案的整體距離較小，被視為品質較高。基於評分結果，驗證者對各提案進行投票：僅當某提案的 Krum 分數優於三分之二以上的其他提案時，驗證者才會投下贊成票。投票過程遵循簡化的 PBFT 協議，當某提案獲得超過三分之二驗證者的贊成票時，該提案被正式接受。委員會的領導者隨即將接受的提案連同相關資訊打包成新區塊，廣播至全網。所有參與者收到新區塊後，依據其中的全域更新同步更新本地模型，完成本輪訓練。

### 2.3.4.3 獎勵機制與激勵設計

BlockDFL 的獎勵機制遵循「有貢獻才有回報」的設計原則，旨在解決分散式系統中普遍存在的搭便車問題。在傳統聯邦學習中，無論參與者是否真正貢獻高品質的訓練成果，皆可獲得最終全域模型的使用權，這削弱了誠實參與的激勵。BlockDFL 透過將權益獎勵與「被選中」直接綁定，確保只有對本輪全域模型更新有實質貢獻的參與者才能獲得回報，從而建立起正向的激勵結構。

具體而言，當某一提案通過委員會共識並被寫入區塊鏈時，系統將權益增量分配給三類參與者。第一類是提交該提案的聚合者，其承擔了收集更新、執行篩選與聚合運算的工作，並承受提案可能未被選中的風險。第二類是本地更新被納入該提案的更新提供者，他們貢獻了訓練運算資源與本地資料的價值。第三類是對該提案投下贊成票的驗證者，他們執行了驗證運算並參與了共識決策。這三類參與者均分本輪的權益獎勵，其身份被明確記錄於區塊之中，確保獎勵分配的透明與可驗證。

相對地，未對本輪全域模型更新做出貢獻的參與者則不獲得任何獎勵。這包括：提案未被選中的其他聚合者、本地更新未被納入獲選提案的更新提供者、以及對獲選提案投下反對票或未參與投票的驗證者。這種設計創造了明確的激勵導向：聚合者有動機提交高品質的提案以提高被選中的機率；更新提供者有動機提交優質的本地更新以增加被納入提案的可能性；驗證者則有動機投票支持真正優質的提案，因為只有投票與最終結果一致時才能獲得獎勵。

然而，此獎勵機制在激勵誠實行為的同時，也創造了一個具有正反饋特性的動態系統。當參與者獲得權益獎勵時，其總權益增加，這直接提升了該參與者在未來輪次被選為聚合者或驗證者的機率，進而增加其獲得更多獎勵的機會。BlockDFL 的設計者預期此正反饋將使「持續誠實貢獻的參與者逐漸累積優勢，而惡意參與者的影響力則相對削弱」[15]。這項預期建立在一個關鍵假設之上：獲得高權益的參與者必然是長期誠實貢獻者。然而，若攻擊者能夠在潛伏期間偽裝成誠實參與者並成功累積權益，此正反饋機制反而可能成為攻擊者鞏固優勢的工具。這一安全隱患將在本章第 2.5 節進一步探討。

#### 2.3.4.4 本研究採用此模型的理由

本研究選擇 BlockDFL 作為基準系統模型，基於以下四項考量。首先，BlockDFL 代表了區塊鏈聯邦學習委員會架構的最新技術水準，其於 2024 年發表於頂級網路研討會，融合了角色分離、權益加權選舉、雙層評分機制與拜占庭容錯共識等多項先進設計，具有高度的代表性。其次，BlockDFL 的系統定義清晰完整，論文詳細說明了角色職責、運作流程與獎勵規則，這為形式化的威脅分析提供了堅實基礎。相較於部分僅提供概念性描述的研究，BlockDFL 的明確定義使得安全性分析能夠建立在具體的系統行為之上，而非抽象的假設。

第三，BlockDFL 的委員會機制與獎勵設計具有廣泛的通用性。儘管不同 BCFL 系統在具體實現上存在差異，但「小型委員會執行共識」與「權益驅動的角色選舉」已成為此領域的主流設計典範。因此，針對 BlockDFL 所發現的安全問題與提出的防禦機制，在原理上可推廣至採用類似架構的其他系統。最後，BlockDFL 已有公開的實驗資料與效能基準，這為本研究後續的防禦機制評估提供了可比較的參照點。綜合以上考量，BlockDFL 是本研究進行威脅建模與防禦設計的理想分析對象。

## 2.4 現有驗證方法與其局限

### 2.4.1 密碼學驗證方法的運算瓶頸

零知識機器學習（Zero-Knowledge Machine Learning, zkML）代表了密碼學驗證方法在機器學習領域的前沿探索 [35]。其核心理念是將機器學習運算轉換為算術電路，並生成零知識證明，使驗證者無需重新執行運算即可確認結果的正確性。這種方法在理論上提供了最強的安全保證：證明的正確性完全依賴密碼學假設，無需信任任何參與方。若能將 zkML 應用於聯邦學習的聚合驗證，系統將能在不揭露個別更新內容的前提下，證明聚合結果確實是由指定的本地更新按照預定規則運算而得。

然而，zkML 面臨嚴峻的運算效能挑戰，這使其在當前技術條件下難以應用於實際的聯邦學習系統。將神經網路運算轉換為算術電路的過程會產生大量的多項式約束，約束數量隨模型複雜度急劇膨脹。根據現有基準測試，即使是相對簡單的 LeNet [36] 模型，其約束數量也可達數億級別；對於 ResNet-18 [37] 等級的模型，證明生成時間需要

近一分鐘；而對於 VGG16 [38] 或更大規模的模型，證明生成可能耗時數十分鐘甚至數小時，且需要數百 GB 乃至 TB 級別的記憶體 [35]。考慮到聯邦學習通常需要數百至數千輪迭代，每輪都執行如此耗時的證明生成顯然不切實際。

更關鍵的局限在於 zkML 難以支援拜占庭容錯聚合演算法。Krum [20] 與 Multi-Krum 等防禦性聚合方法需要運算所有客戶端更新之間的成對距離，這在零知識電路中會產生  $O(n^2 \cdot d)$  的約束爆炸，其中  $n$  為客戶端數量， $d$  為模型參數維度。排序與中位數運算在算術電路中同樣極度昂貴。現有的 zkFL 方案如 RiseFL [39] 僅能支援 L2-norm 有效性檢查等簡單驗證，而無法實現完整的拜占庭容錯聚合驗證。這意味著即使克服了效能瓶頸，zkML 仍無法為採用 Krum 等防禦機制的系統（如 BlockDFL）提供聚合正確性的密碼學證明。

## 2.4.2 樂觀執行方法的架構限制

樂觀機器學習（Optimistic Machine Learning, opML）採用與 zkML 截然不同的設計哲學：預設所有運算結果都是正確的，僅在有參與者提出質疑時才啟動驗證程序 [40]。這種「樂觀執行」的模式大幅降低了正常情況下的運算成本，因為絕大多數時候驗證程序不會被觸發。當爭議發生時，系統透過互動式的二分協議（Bisection Protocol）逐步縮小爭議範圍，最終定位至單一運算步驟，由鏈上的欺詐證明虛擬機（Fraud Proof Virtual Machine, FPVM）進行仲裁。ORA Protocol 已展示此方法可支援 LLaMA 2 [41] 等數十億參數規模的模型在以太坊上運行 [42]。

然而，opML 的架構設計與聯邦學習的需求存在根本性的衝突。首先是挑戰期的問題。為確保驗證者有充足時間偵測並提交欺詐證明，主流的樂觀執行系統如 Optimism [43] 與 Arbitrum [44] 採用長達一週的挑戰期。這種設計對於區塊鏈交易的最終確認或許可以接受，但對於需要快速迭代的聯邦學習訓練而言則完全不可行。若每輪聚合都需等待一週才能確認，數百輪的訓練將耗時數年。即使大幅縮短挑戰期，仍會顯著拖慢訓練進度，且可能因驗證者反應時間不足而削弱安全保證。

其次，opML 的信任模型與 BCFL 的多驗證者場景存在落差。opML 建立在「AnyTrust」假設之上：只要存在至少一個誠實的驗證者願意監控並挑戰錯誤結果，系統就是安全的。這本質上是一個單一提交者與單一挑戰者之間互相對抗的兩方爭議模型。然而，BCFL 的委員會共識涉及多個驗證者對多個提案的集體決策，這種多

方參與的結構難以直接套用 opML 的爭議解決框架。此外，opML 的設計假設運算輸入（如模型更新）是公開可見的，以便驗證者能夠重新執行運算並發現錯誤。這與聯邦學習對更新隱私的保護需求存在張力。

### 2.4.3 委員會驗證方法的安全假設

相較於密碼學方法與樂觀執行方法，基於委員會共識的驗證方法在效率與實用性之間取得了較佳的平衡，因而成為當前 BCFL 系統的主流選擇。這類方法的核心思想是由一個小型委員會代全網執行驗證與共識，透過拜占庭容錯協議確保只要委員會中的誠實成員佔據多數，驗證結果就是可信的。前文介紹的 BlockDFL 即屬此類，其他代表性系統包括 FLCoin [32] 與 BFLC [16]。

FLCoin 提出基於滑動視窗的動態委員會機制，將聯邦學習的貢獻歷史作為委員會成員資格的依據。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內持有份額的節點組成當輪委員會。這種設計使委員會組成與 FL 目標直接對齊，且透過視窗的滑動實現成員的動態更替。FLCoin 的安全性分析顯示，在全網惡意節點比例不超過 25% 且視窗大小為 100 的條件下，委員會安全的機率可達 98.4% [32]。然而，論文並未深入分析惡意節點透過策略性參與逐步累積份額的可能性，其實驗也假設無惡意節點參與，未驗證對抗性累積策略的防禦效果。

BFLC 開創性地將委員會共識引入 BCFL，採用聲譽機制決定委員會組成 [16]。系統追蹤各節點提交更新的品質歷史，高聲譽者優先被選入委員會。委員會成員使用 K-fold 交叉驗證評估提交的模型更新，透過共識決定是否接受。這種設計能有效過濾曾有不良記錄的節點，但也面臨冷啟動問題：新加入者缺乏歷史記錄，難以建立初始信任。更重要的是，後續研究明確指出 BFLC「容易被惡意節點混入委員會，從而導致系統偏差」[15]，當惡意節點佔據委員會半數席位時即可發動攻擊。

綜觀這些委員會驗證方法，它們共享一個根本性的安全假設：委員會的誠實多數。無論採用隨機抽樣、權益加權、聲譽評分或貢獻歷史等任何選舉機制，所有方案都假設某種形式的誠實多數能夠在委員會層級得到維持。這一假設在面對靜態的、比例固定的惡意節點時或許成立，但當攻擊者採取動態的、策略性的行為時，其有效性便值得商榷。下一小節將進一步分析這種靜態假設的盲區。

## 2.4.4 系統性局限：靜態分析的盲區

回顧上述三類驗證方法，可以發現一個貫穿其中的系統性局限：現有的安全性分析幾乎都建立在「攻擊者資源固定」的靜態假設之上。zkML 的安全性證明假設攻擊者的運算能力無法突破密碼學難題；opML 假設至少存在一個持續監控的誠實驗證者；委員會方法則假設惡意節點在全網中的比例  $\alpha$  是一個固定的常數，並據此運算委員會被攻破的機率。這種「快照式」的分析視角忽略了一個關鍵的動態因素：理性的攻擊者會根據系統狀態調整其策略，而非機械地執行固定行為。

以委員會方法為例，第 2.3.3 節的超幾何分佈分析假設每輪選舉都是從固定的惡意節點比例中獨立抽樣。然而，若系統的獎勵機制存在正反饋特性（例如在 BlockDFL 架構中，獲得獎勵的節點將累積更多權益，進而提高未來被選中的機率），則各輪選舉之間並非獨立事件。攻擊者可利用此特性採取「耐心策略」：在初期表現完全誠實以累積權益與聲譽，僅在其控制的節點佔據委員會優勢時才發動攻擊。這種行為模式無法被靜態的機率分析所捕捉。

更根本的問題在於，現有分析未區分「惡意節點的存在比例」與「惡意節點的有效影響力」。在權益加權的選舉機制中，一個持有 10% 權益的惡意節點，其對委員會組成的影響力可能遠大於十個各持有 1% 權益的惡意節點。若攻擊者能夠透過合法途徑（累積獎勵、建立聲譽）集中權益於少數節點，即使其控制的節點數量佔全網比例不高，仍可能在委員會層級取得超額的影響力。FLCoin 的分析雖然考慮了惡意節點比例與委員會安全機率的關係，但未分析惡意節點比例本身可能隨時間演變的動態過程。BlockDFL 假設「持有大量權益的參與者傾向誠實」，但未充分論證此假設在面對長期潛伏的策略性攻擊者時是否仍然成立。

表 2.1 總結了三類驗證方法在安全性、效率與通用性三個維度上的特性。可以看出，沒有任何一種方法能夠同時滿足所有需求：zkML 提供最強的安全保證但效率過低且無法支援複雜聚合；opML 效率較高但挑戰期過長且架構不適用於多驗證者場景；委員會方法在效率與實用性上表現最佳，但其安全性依賴可能被違反的誠實多數假設。這種「安全性-效率-通用性」的三難困境，構成了本領域研究的核心挑戰。

表 2.1: BCFL 驗證方法比較

方法類別	安全性基礎	主要效率瓶頸	對 BCFL 的適用性
zkML	密碼學證明，無需信任假設	證明生成耗時數分鐘至數小時	低：無法支援 Krum 等複雜聚合
opML	經濟激勵與 AnyTrust 假設	挑戰期長達數天	低：架構不適用於多驗證者場景
委員會共識	委員會誠實多數假設	共識通訊成本 $O(c^2)$	高：主流選擇，但假設可能被違反

## 2.5 研究缺口：從激勵設計到安全隱患

前述分析揭示了現有 BCFL 驗證方法的共同盲區：靜態的安全性假設無法應對動態的攻擊策略。本節將聚焦於當前最具實用性的委員會驗證方法，深入探討其獎勵機制如何在激勵誠實行為的同時，也為策略性攻擊者創造了可利用的漏洞。這一分析將指向本研究欲填補的關鍵缺口，並為第三章的威脅模型建構奠定基礎。

### 2.5.1 獎勵機制的正反饋特性

第 2.3.4 節詳細介紹了 BlockDFL 的獎勵機制：當某一提案通過委員會共識時，提交該提案的聚合者、更新被納入的更新提供者、以及投贊成票的驗證者共同分享權益獎勵。這種設計的初衷是解決搭便車問題，確保只有實質貢獻者才能獲得回報。然而，從系統動態的角度審視，此機制創造了一個具有正反饋特性的循環結構。

正反饋的運作邏輯可描述如下：節點獲得權益獎勵後，其總權益增加；由於角色選舉採用權益加權機制，權益增加直接提升該節點在未來輪次被選為聚合者或驗證者的機率；作為聚合者或驗證者，節點更有機會參與被接受的提案，從而獲得更多獎勵。這種「獎勵 → 權益增加 → 選中機率提升 → 更多獎勵」的循環，在數學上構成一個正反饋系統。BlockDFL 的設計者預期此特性將使誠實參與者逐漸累積優勢，因為他們持續提交高品質更新並誠實驗證，從而獲得穩定的獎勵流入。相對地，惡意參與者若因提交低品質更新或投票與最終結果不一致而錯失獎勵，其相對權益份額將逐漸稀釋。

然而，這一預期建立在一個隱含假設之上：系統能夠有效區分誠實行為與惡意行為，並據此差異化地分配獎勵。問題在於，當惡意節點選擇「偽裝誠實」，意即在攻擊

發動前完全模仿誠實節點的行為，系統便無法將其與真正的誠實節點區分開來。在這種情況下，惡意節點同樣能夠獲得獎勵、累積權益、提升影響力，正反饋機制反而成為攻擊者滲透系統的工具。

## 2.5.2 策略性攻擊者的潛在威脅

基於上述分析，可以設想一種策略性的攻擊模式：攻擊者控制的節點在初期階段完全表現為誠實參與者，正常參與訓練、提交高品質更新、在驗證時投票支持真正優質的提案。透過這種「潛伏」行為，攻擊節點逐步累積權益與聲譽，提高其被選為聚合者或驗證者的機率。當攻擊者評估其控制的節點已在委員會中佔據足夠優勢時（例如在某一輪選舉中恰好佔據超過三分之一的驗證者席位），才發動實際攻擊。

這種攻擊模式的危險性在於其隱蔽性與自我強化性。在潛伏階段，攻擊節點的行為與誠實節點完全一致，現有的異常偵測機制無從識別。更重要的是，一旦攻擊者在某一輪成功控制委員會，他們可以操控共識結果，將獎勵僅分配給自己控制的節點（透過接受包含攻擊節點更新的提案、拒絕僅包含誠實節點更新的提案），同時確保攻擊節點作為「投贊成票的驗證者」獲得驗證獎勵。這種操控使攻擊者的權益份額在成功攻擊後加速增長，進一步提高其在未來輪次控制委員會的機率，形成惡性循環。

現有文獻已零星地意識到委員會滲透的風險。FedBlock 在其未來展望中指出：「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」，並警告「驗證者本身亦可能是惡意的」[22]。BFLC 的後續評論指出該系統「容易被惡意節點混入委員會」[15]。然而，這些觀察多停留在定性描述的層次，尚未發展出系統性的威脅模型來形式化分析此類攻擊的具體機制、成功條件與潛在危害。特別是，現有研究未充分探討獎勵機制的正反饋特性如何與委員會選舉機制交互作用，使得攻擊者能夠透過「合法」途徑逐步擴大影響力。

## 2.5.3 本研究的切入點

綜合本章的分析，可以明確界定當前 BCFL 委員會安全性研究的核心缺口：

第一，**缺乏針對策略性攻擊者的形式化威脅模型**。現有安全性分析假設惡意節點比例固定且行為模式單一，未考慮攻擊者可能採取的動態策略，如長期潛伏、選擇性



攻擊時機、以及利用正反饋機制累積優勢。

第二，**缺乏對獎勵機制安全性意涵的深入分析**。現有研究將獎勵機制視為激勵誠實行為的工具，但未系統性地審視其正反饋特性可能被攻擊者利用的風險。

第三，**缺乏能夠偵測與抵禦漸進式滲透的防禦機制**。無論是隨機選擇、權益加權或聲譽導向，現有委員會選舉機制都未針對「透過合法途徑累積影響力」的攻擊模式設計相應的防護措施。

針對上述缺口，本研究將在後續章節展開以下工作。第三章將形式化定義「漸進式委員會佔領攻擊」(Progressive Committee Capture Attack, PCCA) 的威脅模型，基於第 2.3.4 節建立的 BlockDFL 系統模型，分析攻擊者如何利用獎勵機制的正反饋特性逐步滲透委員會，並量化攻擊在不同參數設定下的成功機率與所需時間。第四章將提出針對 PCCA 的防禦架構設計，透過打破正反饋循環與引入動態挑戰機制，在維持系統效率的前提下提升對策略性攻擊者的抵抗能力。第五章將透過實驗評估所提出方法的有效性，驗證其在多種攻擊場景下的防護效能。

## 第三章 威脅模型 (Threat Model)

基於第二章所建立的委員會架構系統模型，本章將深入剖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並分析「漸進式委員會佔領攻擊」(Progressive Committee Capture Attack, PCCA)，這是一種專門針對權益機制設計缺陷的隱蔽性攻擊手法。透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。值得特別強調的是，這種攻擊與傳統的模型投毒攻擊存在本質性差異，其危險性並非體現在對單一模型品質的破壞，而是從根本上顛覆了去中心化系統的安全假設，能夠將表面上維持去中心化形態的聯邦學習系統，實質上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分散式系統理論中對最壞情況的假設。然而，在實際的區塊鏈應用場景中，攻擊者往往具有明確的經濟動機而非單純追求破壞，他們的行為模式遵循經濟理性原則，首要目標是利益最大化。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，而如果能夠透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，這種複雜的目標結構使得攻擊行為更加隱蔽且難以偵測。在最直接的層面，攻擊者追求經濟利益的最大化，具體表現為透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，透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，他們將能夠更頻繁地控制委員會的組成，進而掌握聯邦學習過程中的關鍵決策權，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，因為它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。

### 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，因為控制更高比例的節點需要投入大量的經濟資源與協調成本。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，這些約束為防禦機制的設計提供了重要的切入點。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，區塊鏈的不可篡改性為系統提供了可靠的審計基礎。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，這使得攻擊者必須採用更為精細的策略來實現其目標。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，不同層次的攻擊具有截然不同的目標、手法與防禦需求。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，為後續的防禦機制設計提供清晰的問題定位。

### 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，透過污染訓練資料或模型更新來破壞最終模型的品質，這類攻擊在聯邦學習安全研究中已經得到廣泛的關注與深入的探討。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布，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 [20]、Trimmed Mean [21]、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，這種層次間的依賴關係構成了現有防禦體系的結構性弱點。

## 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，這類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透過事先協調，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 2 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」，統計資料顯示約 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 類系統中，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或篡改；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；第三是自我強化特性，一旦攻擊成功，攻擊者將獲得更多獎勵，導致其權益進一步增加，形成正反饋循環。

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，這種對比有助於理解本研究選擇聚焦於共識層攻擊的理論依據。

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最佳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。這種認識促使本研究將焦點放在共識層的安全性分析與防禦機制設計上，而非繼續在資料層防禦的技術細節上進行增量式的改進。

3.3 漸進式委員會佔領攻擊 (Progressive Committee Capture Attack)

本節將詳細定義本研究針對的核心威脅：漸進式委員會佔領攻擊 (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，其獨特之處在於透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顯著優勢地位的漸進式轉變，最終建立起對委員會決策的持續性影響力。

### 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累積優勢的工具。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，而是透過操縱投票結果來控制獎勵分配。具體而言，惡意委員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。透過持續執行這種排他性策略，惡意節點能夠獲得相對於誠實節點更高比例的系統獎勵，逐步擴大其權益優勢。隨著攻擊者權益佔比的提升，其在未來委員會選舉中獲得多數席位的機率也會相應提高，形成自我強化的正反饋循環。演算法 1 以形式化的方式呈現了 PCCA 的決策邏輯，清晰展示了攻擊者如何根據當前控制比例動態調整其行為模式。

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，透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。

---

**Algorithm 1** 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

- 1: **Check Phase:** Calculate control ratio  $r = \frac{|\mathcal{V} \cap \mathcal{C}_{adv}|}{|\mathcal{V}|}$
  - 2: **if**  $r \leq 2/3$  **then**  
    *State 1: Shadow Mode (Lurking)*
  - 3:     Follow the protocol honestly to accumulate stake and await majority.
  - 4: **else**  
    *State 2: Capture Mode (Occupying)*
  - 5:     **if** Aggregator is Adversarial **then**
  - 6:         **Full Stack Poisoning:** Force approve malicious proposal.
  - 7:     **else**
  - 8:         **Strategic Starvation:** Force reject honest proposal.
  - 9:     **end if**
  - 10: **end if**
- 

## 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。根據第 2.3.3 節的超幾何分布分析，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，這種低頻率的攻擊窗口使得潛伏階段可能持續相當長的時間。

在這漫長的等待期間，攻擊者必須維持完美的誠實表現以避免被識別為可疑節點。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，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。其核心策略是透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，透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。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。

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，這種攻擊的精妙之處在於其高度的隱蔽性。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，訓練過程並未完全停滯，只是收斂速度相對放緩，這使得攻擊行為不易被外部觀察者識別為明顯的惡意行為。然而，從經濟激勵的角度看，這種攻擊造成了顯著的後果：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，無法獲得應得的經濟回報。這種「付出努力但得不到回報」的狀態會導致兩種效應：一方面，誠實節點因為無法獲得獎勵而使其權益增長停滯，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，惡意節點透過獲取更高比例的獎勵實現權益的相對增長，其在下一輪委員會中的佔比優勢進一步擴大。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。

**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，透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；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，即使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，從模型品質角度看，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透過正常的選舉機制實現自我恢復。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，攻擊者透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。

然而，值得特別注意的是，即使在佔領階段，惡意節點的權益增長也並非呈現指數式的無限擴張。這是因為在 BlockDFL 的獎勵機制中，誠實節點仍然能夠透過擔任 Update Provider 角色獲得部分獎勵，即使他們的提案被惡意委員會拒絕，他們作為被選中提案的 Update Provider 時仍可分得相應的獎勵份額。這種機制設計意味著惡意節點無法完全壟斷系統的全部獎勵，而是會與誠實節點形成一種動態的權益分配平衡。具體而言，假設系統每輪分配的總獎勵為  $R$ ，惡意委員會雖然能夠透過操縱投票使獎勵更傾向於流向惡意節點，但誠實節點作為 Update Provider 的貢獻仍會獲得部分補償，這使得雙方的權益比例會趨向於某個穩定的比值而非無限分化。

基於上述分析，權益演化的動態可以表示為一個有界的增長模型。設  $\alpha$  為惡意節點在成功控制委員會時能夠獲得的獎勵比例優勢係數，由於誠實節點仍能透過 Update Provider 角色獲得部分獎勵，這個係數  $\alpha$  具有上界，通常在 1.1 至 1.2 之間，這意味著惡意節點每輪獲得的獎勵大約是誠實節點的 1.1 到 1.2 倍，而非無限倍數。這種有界的優勢係數導致雙方的權益比例會收斂到一個穩定值：

$$\lim_{t \rightarrow \infty} \frac{S_{mal}(t)}{S_{hon}(t)} = \alpha \cdot \frac{S_{mal}(0)}{S_{hon}(0)} \quad (3.1)$$

從系統動力學的角度看，這是一個具有穩定平衡點的正反饋系統，而非傳統理解中的發散系統。這種平衡的存在並不意味著 PCCA 攻擊不具威脅性，相反地，它揭示了攻擊的另一種危險形態：攻擊者能夠建立並維持一種持久的「領先者優勢」，即使無法實現完全壟斷，也能長期保持對系統治理的顯著影響力。這種持續性的權益優勢使得攻擊者能夠更頻繁地控制委員會，形成一種「常態化」的治理失衡狀態，而現有系統缺乏打破這種平衡的內在機制。

### 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。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の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 實現了權力結構的顯著傾斜。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透過隨機選舉產生的，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並穩定在優勢水平，這種表面上的去中心化逐漸演變為實質上的寡頭主導。當攻擊者的權益佔比穩定在較高水平時，他們獲得委員會多數席位的機率將顯著高於隨機分布的預期值，意味著從統計意義上他們能夠在更多輪次中控制委員會，形成一種「軟性壟斷」的治理狀態。這種從分散到集中的權力轉移過程，雖然不會達到完全壟斷的程度，但已經足以嚴重損害區塊鏈系統的核心價值主張。

在經濟激勵層面，PCCA 造成了激勵機制的扭曲與部分失靈。對於誠實節點而言，他們會發現一個令人沮喪的現實：即使投入大量計算資源進行本地訓練、提交高品質的模型更新，在委員會投票環節仍然面臨被系統性排斥的風險，獲得的經濟回報明顯低於預期。這種「付出與回報不成比例」的狀態會逐漸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，當持續的低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，部分節點可能選擇降低參與程度或退出系統。這種節點流失會形成另一層負面效應：誠實節點的退出會進一步提高惡意節點的相對權益佔比，使得系統更容易被控制，這又會加速更多誠實節點的離開，形成一種緩慢但持續的惡性循環。

### 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，這種對比有助於理解為何現有的防禦機制難以有效應對 PCCA。

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

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

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。從攻擊者動機角度來看，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；而 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。

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。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，攻擊者的相對權益優勢就會擴大，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，就會沿著權益失衡的軌道持續發展，直到達到某個穩定的不平衡狀態。

從防禦策略的角度來看，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，這些方法的有效性已經在大量實驗中得到驗證。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、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，以下將從五個層面逐一闡述這些安全目標的具體內涵與實現要求。

### 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，例如透過罰沒 (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，而應該透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，並透過鏈上的驗證過程來證明其合理性。透過這種多層次的防禦設計，系統能夠確保攻擊者無法透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，確保他們透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審視獎勵分配機制，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，仍然可以透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持續下降，這可能需要引入反壟斷機制，例如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，例如透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強健性。

### 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透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，而是要透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  $E[\text{Payoff}] = P_{\text{success}} \cdot G_{\text{attack}} - P_{\text{caught}} \cdot L_{\text{slash}} < 0$ ，其中  $P_{\text{success}}$  是攻擊成功的機率， $G_{\text{attack}}$  是攻擊成功時獲得的經濟收益， $P_{\text{caught}}$  是攻擊被檢測到的機率， $L_{\text{slash}}$  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  $P_{\text{caught}}$ ，透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  $L_{\text{slash}}$ ，使其遠大於潛在的攻擊收益  $G_{\text{attack}}$ ，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  $G_{\text{attack}}$ ，例如透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。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：漸進式委員會佔領攻擊 (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 的目標在於透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持續性影響力。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，這種攻擊者模型更貼近現實世界中的威脅場景。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，攻擊者透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。前者透過系統性地拒絕誠實提案來阻止誠實節點獲得獎勵，後者則在同時控制 Aggregator 和 Verifier 的情況下直接注入惡意更新。

透過權益增長動態分析，我們釐清了 PCCA 攻擊的實際影響特徵。值得注意的是，由於誠實節點仍能透過擔任 Update Provider 角色獲得部分獎勵，惡意節點的權益增長並非呈現指數式的無限擴張，而是會與誠實節點形成一種動態平衡，惡意節點的權益通常穩定在誠實節點的 1.1 至 1.2 倍左右。這種平衡的存在並不意味著攻擊不具威脅性，相反地，它揭示了攻擊者能夠建立並維持一種持久的「領先者優勢」，形成常態化的治理失衡狀態。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。下一章將介紹本研究提出的防禦機制，展示如何透過挑戰增強型委員會架構，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過程中，始終面臨執行效率的嚴峻挑戰。傳統拜占庭容錯共識機制雖然能夠提供強健的安全保證，但其伴隨的高昂通訊成本卻難以適應機器學習場景中頻繁迭代更新的需求。第 3 章的威脅分析已經揭示了現有委員會機制的根本缺陷：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理性攻擊者有機會透過漸進式的權益累積來逐步掌控驗證權力，而現有的防禦機制過度依賴「誠實多數假設」，缺乏對策略性攻擊者的有效威懾手段。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 2.3.4 節所定義的 BlockDFL 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。

本架構的核心設計哲學源於對聯邦學習與金融交易系統本質差異的深刻認識。金融交易系統對每一筆交易都要求即時且不可逆的正確性保證，因為任何錯誤都可能導致資產的永久損失，這種特性迫使傳統區塊鏈系統必須在每次狀態變更前達成全網共識。然而，機器學習過程本身具備天然的抗噪性與自我修復能力，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。CACA 正是基於這一洞察，將安全性驗證從同步的阻塞式流程轉變為異步的非阻塞式審計機制，成功地將效率最佳化與安全保障解耦。這種設計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，本架構從根本上重塑了攻擊者的理性決策空間，使得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，從而消除了發動攻擊的經濟誘因。

本章的結構安排如下：首先在第 4.1 節中概述 CACA 相對於 BlockDFL 的架構創新，闡明挑戰者角色與異步審計機制如何嵌入現有的委員會流程；接著在第 4.2 節深入探討異步審計與究責機制的運作原理，包括挑戰流程的觸發條件、仲裁機制的執行邏輯，以及「僅懲罰不回滾」策略的設計考量；隨後在第 4.3 節論證雙層信任模型如何提

供等同於全網共識的安全保障；在第 4.4 節透過通訊複雜度分析與概率模型推導，量化評估本架構的效率優勢；最後在第 4.5 節探討激勵機制的經濟學基礎，說明如何透過罰沒與獎勵的精心設計實現激勵相容性。

## 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，而這一目標的實現建立在對現有 BlockDFL 架構的繼承與創新之上。如第 2.3.4 節所詳述，BlockDFL 透過角色分離的設計理念，將參與者劃分為更新提供者、聚合者與驗證者三種角色，並透過權益加權的隨機選舉機制決定每輪的角色分配。這種設計在效率與基本安全性之間取得了當時文獻中的最佳平衡，為後續的架構創新奠定了堅實的基礎。CACA 完整保留了 BlockDFL 的訓練流程與角色定義，包括更新提供者的本地訓練職責、聚合者的提案生成流程，以及驗證委員會的 Krum 評分與 PBFT 共識機制，這些經過驗證的設計元素構成了本架構運作的基礎框架。

CACA 的核心創新在於引入了第四種角色，即挑戰者，以及與之配套的異步審計機制。這一創新從根本上改變了系統的安全性保障方式，將防禦策略從「事前預防」轉向「事後追責」。圖 4.1 展示了 CACA 的完整運作流程，清晰呈現了挑戰機制如何嵌入現有的委員會共識流程。在每一輪次的正常運作中，系統首先根據前一區塊的雜湊值進行動態角色分配，隨後更新提供者執行本地訓練並將結果提交給聚合者，聚合者生成提案後交由驗證委員會進行 Krum 評分與 PBFT 投票，這一流程與 BlockDFL 完全一致。關鍵的差異出現在共識達成之後：在 BlockDFL 中，委員會的決策即為最終決策，系統缺乏對委員會潛在惡意行為的事後追責能力；而在 CACA 中，委員會達成共識後系統立即執行模型更新，但同時開啟了一個異步的審計窗口，允許任何持有足夠質押的節點作為挑戰者對委員會的決策進行事後驗證。

這種「先執行後審計」的設計哲學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。挑戰者角色的設計體現了 CACA 對去中心化監督的核心承諾，這一角色向所有持有足夠質押的節點開放，而非僅限於特定的特權群體。挑戰者的職責在於持續監聽鏈上資料，獨立重新執行 Krum 演算法的運算，並將運算結果與委員會選定的全域更新進行比對。

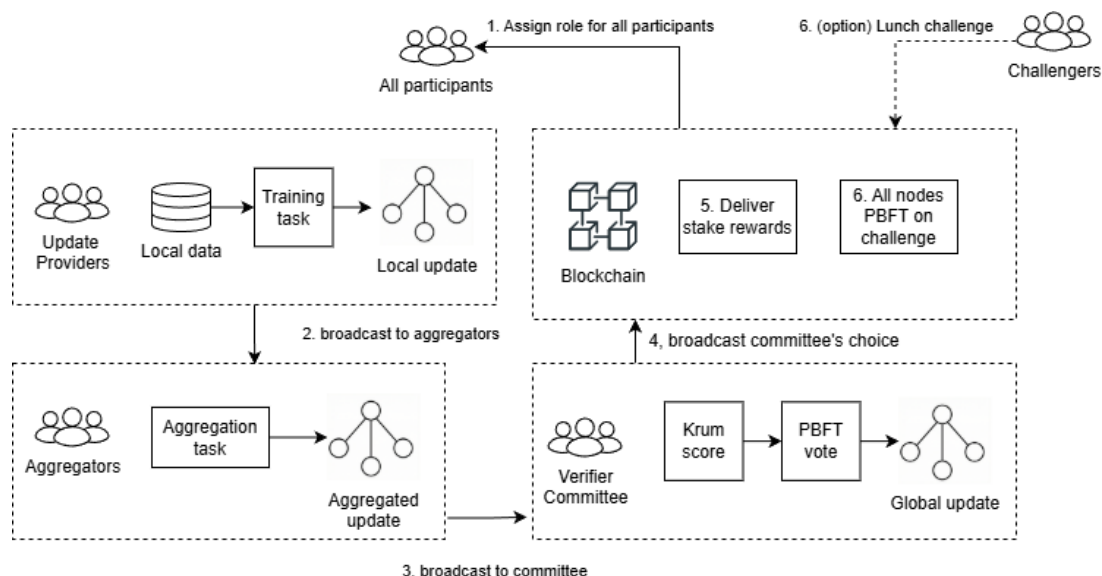


圖 4.1: Challenge-Augmented Committee Architecture (CACA) 系統架構與工作流程圖

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，任何偏離正確結果的決策都將被挑戰者精確識別。一旦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，便可以質押規定金額的押金發起挑戰交易，觸發全網仲裁機制。

這種開放式的監督設計本質上將監督權力從少數委員會成員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。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，這正是 CACA 能夠在採用較小委員會的同時維持高安全性的關鍵所在。演算法 2 與演算法 3 分別以形式化的方式呈現了 CACA 的即時執行協議與異步挑戰機制。即時執行協議描述了從角色分配到模型更新的完整流程，其核心特徵在於委員會達成共識後立即提交全域模型更新，無需等待任何額外的確認期。這種設計選擇體現了對系統「活性」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。

異步挑戰機制則作為系統的安全後盾在背景中持續運作，其核心邏輯在於：挑戰者持續驗證委員會決策的正確性，一旦發現異常便發起挑戰，觸發全網重新驗證；若惡意行為被確認，系統執行罰沒操作沒收惡意節點的質押金。值得特別注意的是，已經提交的模型更新不會被回滾，這種「僅懲罰不回滾」的策略是 CACA 設計中的重要考量，其理論基礎將在下一節詳細闡述。

---

**Algorithm 2** CACA Execution Protocol (Instant Update)

---

**Require:** Current Round  $r$ , Total Stake Weighted Nodes  $\mathcal{N}$

**Ensure:** Updated Global Model  $w_{r+1}$

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

---

**Algorithm 3** Asynchronous Challenge Mechanism (Slash-Only)

---

**Require:** Challengers  $\mathcal{C}$

**Ensure:** Punishment for Malicious Acts

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

## 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「悲觀併發控制」的設計哲學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。CACA 則採用了「樂觀執行」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。這種設計選擇的理論基礎在於認識到聯邦學習與金融交易在容錯需求上的本質差異：金融交易的錯誤是不可逆的資產損失，而機器學習的偶發偏差則能夠透過後續訓練迭代逐步修正。這種固有的自我修復能力為樂觀執行策略提供了安全邊際，使得系統能夠在不犧牲長期安全性的前提下最大化執行效率。

挑戰流程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，而這種事後究責能力正是 CACA 實現經濟安全性的關鍵。挑戰流程的觸發條件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，隨後在本地重新執行 Krum 演算法，運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，無論是由於運算錯誤還是惡意操縱，都構成了發起挑戰的充分理由。挑戰者提交挑戰交易時必須附帶規定金額的質押金，這筆質押金的設計具有雙重目的：一方面防止惡意節點透過大量無效挑戰來發動拒絕服務攻擊，另一方面為成功的挑戰者提供經濟激勵，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。智能合約首先鎖定相關的質押金，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與不可篡改性。系統隨即觸發全網仲裁機制，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，系統將執行罰沒操作沒收惡意委員會成員的全額質押金，並將部分罰沒資金分配給挑戰者作為獎勵。

當仲裁確認委員會存在惡意行為時，CACA 採用「僅懲罰不回滾」的處置策略，這一設計選擇基於對聯邦學習系統特性的深刻理解。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。這種特性在第 5 章的實驗中將得到驗證。相對地，若選擇回滾模型狀態，則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。考慮到聯邦學習通常需要經歷數百甚至數千個訓練輪次，這種回滾將造成極為嚴重的運算資源浪費。更重要的是，全網仲裁的時間延遲意味著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，此時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。因此，CACA 將處置重點放在對惡意行為的經濟懲罰而非對歷史狀態的修正上，透過高額的質押金罰沒來建立強大的經濟威懾力，使得攻擊在經濟上變得完全不理性。

## 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。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，將效率最佳化與安全保障解耦，使得系統能夠根據實際威脅的性質動態調整其安全等級。在正常情況下，系統以小委員會的效率運行；而在異常情況下，系統能夠迅速升級至全網共識的安全等級，這種彈性是傳統單層信任模型所無法提供的。

雙層信任模型的第一層是檢測層，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。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運算。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。

雙層信任模型的第二層是仲裁層，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。這個階段的安全假設要求網路中誠實節點的數量必須超過總節點數的三分之二，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，其核心優勢在於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。第二個條件是收買或壓制全網足夠數量的節點，以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。第二個條件的達成難度遠超第一個，因為在檢測層面攻擊者面臨的是「1-of- $N$  誠實假設」的挑戰。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。

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。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，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

第 2.3.3 節的分析揭示了傳統委員會架構面臨的根本性困境：在 BlockDFL 等現有系統中，安全性的保障完全依賴於「委員會中誠實節點佔據多數」這一機率性條件，而要提高此條件成立的機率，唯一的途徑便是擴大委員會規模，這又直接導致通訊成本的攀升。本節將論證 CACA 如何透過將安全性保障從「門檻安全性」轉移至「經濟安全性」，成功打破委員會規模與安全性之間的強耦合關係，從而在維持等效安全保證的前提下實現顯著的效率提升。

### 4.4.1 BlockDFL 的效率瓶頸：安全性與委員會規模的強耦合

BlockDFL 的安全性論證建立在超幾何分佈的機率運算之上，其核心邏輯可概括為：若要將委員會被惡意控制的風險壓制在可接受的水準之下，系統必須維持足夠大的委員會規模。以第 2.3.3 節的數值分析為例，在全網節點數  $N = 100$ 、惡意節點佔比  $f = 30\%$  的威脅環境下，若將「委員會被惡意節點佔據超過三分之二席位」的風險閾值設定為  $p < 0.01$ ，則委員會規模至少需要達到  $c = 9$  才能滿足此安全性要求。這意味著系統在每一輪訓練中都必須執行規模為 9 的 PBFT 共識，其通訊複雜度為  $O(c^2) = O(81)$ 。

這種設計邏輯的深層問題在於其「悲觀併發控制」的本質。BlockDFL 預設每一輪都可能遭受攻擊，因此必須在每一輪都部署足以抵禦攻擊的防禦資源。然而，在實際運作中，攻擊者成功控制委員會的情況畢竟屬於少數輪次，絕大多數時候系統處於正常運作狀態，此時維持大型委員會所付出的通訊成本便成為一種「預防溢價」。換言之，為了應對可能但並非必然發生的威脅，系統在每一輪都承擔了高昂的固定開銷。

這種將安全成本均攤至每一輪的做法，在需要頻繁迭代的聯邦學習場景中顯得尤為低效，因為數百甚至數千輪的訓練過程會將這種單輪的效率損失累積放大。

## 4.4.2 CACA 的突破：從門檻安全性到經濟安全性

CACA 對效率問題的回應並非追求「更好的機率保證」，而是從根本上改變了安全性的實現方式。傳統的門檻安全性思維聚焦於「如何降低委員會被攻破的機率」，這種思路必然導向更大的委員會規模。CACA 則採取截然不同的策略：與其執著於將被攻破的機率壓制至趨近於零，不如確保即使委員會被攻破，攻擊者也無法從中獲取正向收益。這種從「預防攻擊發生」到「消除攻擊誘因」的視角轉換，構成了經濟安全性的理論基礎。

在經濟安全性的框架下，委員會被攻破的機率不再是唯一的安全性指標，因為異步挑戰機制確保了任何惡意行為都將面臨全額質押金的罰沒。對於理性攻擊者而言，發動攻擊的決策取決於預期收益與預期成本的比較：即使成功控制委員會的機率存在，但一旦被挑戰者揭露並經全網仲裁確認，攻擊者將損失遠超其潛在收益的質押資產。這種不對稱的風險收益結構，使得「不發動攻擊」成為理性攻擊者的最優策略，從而在行為層面消除了攻擊的實際發生。由此，委員會規模的選擇便不再完全受制於安全性的機率運算，系統得以在滿足基本安全閾值的前提下採用相對較小的委員會來獲取效率優勢，而額外的安全性則由經濟懲罰機制另行保障。

## 4.4.3 通訊複雜度對比分析

基於上述設計哲學的差異，BlockDFL 與 CACA 在相同安全性要求下展現出不同的通訊成本特徵。表 4.1 呈現了兩種架構在關鍵效率維度上的系統性對比，該對比以「委員會被惡意控制的風險低於 1%」作為統一的安全性基準，並假設全網節點數  $N = 100$ 、惡意節點佔比  $f = 30\%$  的威脅環境。

BlockDFL 為達到  $p < 0.01$  的安全性閾值，必須採用  $c = 9$  的委員會規模，其每輪的通訊複雜度固定為  $O(81)$ 。相對地，CACA 透過經濟安全性的補充保障，得以採用  $c = 7$  的委員會規模，常態通訊複雜度降至  $O(49)$ ，實現了約 39.5% 的通訊成本削減。這種效率提升的關鍵並非完全取消機率性的安全保證，而是透過經濟懲罰機制提供了

表 4.1: BlockDFL 與 CACA 在相同安全性水平下的效率對比

$$(N = 100, f = 30\%, p_{risk} < 0.01)$$

評估維度	BlockDFL	CACA	差異分析
安全性實現方式	門檻安全性	經濟安全性	機率保證 vs. 激勵相容
所需委員會規模	$c = 9$	$c = 7$	規模縮減 22.2%
常態通訊複雜度	$O(c^2) = O(81)$	$O(c^2) = O(49)$	通訊成本降低 39.5%
安全性維護模式	每輪固定開銷	條件式觸發開銷	預防性 vs. 響應性

額外的安全層，使得較小委員會在面對理性攻擊者時仍能維持等效的實質安全性。

從系統運作的動態視角來看，CACA 的通訊成本呈現條件式的特徵。在正常運作情況下，系統僅需支付  $O(c^2) = O(49)$  的委員會共識成本；唯有當挑戰被觸發並進入全網仲裁時，才會產生額外的  $O(N^2)$  通訊開銷。然而，由於經濟懲罰機制有效消除了理性攻擊者的作惡誘因，挑戰觸發的機率  $p$  在長期均衡中將趨近於零。據此，系統的期望通訊複雜度可表示為：

$$E[Comm] = (1 - p) \cdot O(c^2) + p \cdot (O(c^2) + O(N^2)) = O(c^2) + p \cdot O(N^2) \quad (4.1)$$

當  $p \rightarrow 0$  時，期望複雜度近似於常態值  $O(c^2)$ ，這意味著 CACA 在絕大多數情況下享有較小委員會的效率優勢，而全網仲裁的高昂成本僅作為威懾手段存在，實際上鮮少被觸發。這種「按需付費」的安全模式，相較於 BlockDFL 每輪都必須支付的固定「預防溢價」，在資源利用上顯然更為經濟。

#### 4.4.4 效率提升的本質：架構層面的解耦創新

綜合上述分析，CACA 相對於 BlockDFL 的效率優勢並非源自共識協議本身的改進，而是源自架構層面的根本性創新，即將安全性與委員會規模的強耦合關係予以弱化。在 BlockDFL 的設計中，委員會規模是安全性的唯一保障手段，兩者之間存在不可調和的強耦合關係：追求更高的安全性必然要求更大的委員會，而更大的委員會必然帶來更高的通訊成本。CACA 透過引入異步挑戰機制與經濟懲罰協議，為安全性開闢了獨立於委員會規模的第二條保障路徑，從而打破了這種強耦合。

這種解耦的實踐意義在於，系統設計者得以在滿足基本安全閾值的前提下，根據效率需求選擇較小的委員會規模，而無需過度顧慮安全性的機率運算。雖然  $c = 7$  的委員會在純機率意義上的安全性略低於  $c = 9$ ，但經濟懲罰機制提供的額外威懾力足以彌補這一差距。攻擊者或許更容易獲得控制委員會的機會，但每一次攻擊嘗試都面臨著災難性的經濟後果，這種威懾足以使理性攻擊者放棄攻擊意圖。最終，系統在實際運作中達成了一種新的均衡：較小的委員會提供了效率優勢，而幾乎不會發生的攻擊確保了這種效率優勢不會被全網仲裁的開銷所侵蝕。

## 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來支付審計與仲裁的相關費用。這種「懲罰驅動」的激勵模式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時，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運算的運算成本、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%。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。

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；二是在極少數系統遭受攻擊時透過參與挑戰或仲裁而獲得的額外獎勵。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以及永久失去驗證者資格所帶來的未來收益流損失。

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。更重要的是，這種懲罰機制成功打破了第 3 章所描述的漸進式委員會佔領攻擊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。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。CACA 建立在第 2.3.4 節所定義的 BlockDFL 委員會模型之上，完整保留了其經過驗證的訓練流程與角色定義，同時透過引入挑戰者角色與異步審計機制，實現了從傳統「門檻安全性」向「經濟安全性」的典範轉移。這種轉移並非透過在效率與安全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。

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。通訊複雜度分析揭示了 CACA 架構的效率優勢：在相同的安全性要求下，CACA 得以採用  $c = 7$  的委員會規模，相較於 BlockDFL 所需的  $c = 9$  減少了 22.2% 的規模，通訊成本相應降低約 39.5%。這種效率提升源自經濟安全性對機率安全性的補充，使得系統在絕大多數正常運作的情況下能夠維持在小委員會共識的  $O(C^2)$  量級。

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，從根本上打破了漸進式委員會佔領攻擊所依賴的正反饋循環。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 3 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)

本章旨在透過系統性的實驗設計，驗證挑戰增強型委員會架構在應對漸進式委員會佔領攻擊時的防禦效能。有別於傳統聯邦學習研究以模型準確率作為主要評估指標，本研究的核心關注點在於經濟安全性機制能否有效重塑理性攻擊者的決策空間，使得誠實行為成為長期均衡下的最優策略。具體而言，實驗設計圍繞三個層次遞進的研究問題展開：罰沒機制能否打破惡意節點的權益累積循環、經濟威懾能否將攻擊頻率抑制至可忽略的水平、以及上述防禦效果能否在不犧牲模型品質的前提下達成。透過對權益動態演化、攻擊成功頻率與系統可用性等多維度指標的追蹤分析，本章將從實證角度檢驗第 4 章所提出的理論主張。

### 5.1 實驗設置

#### 5.1.1 資料集與模型配置

本研究採用 MNIST 手寫數字資料集作為聯邦學習任務的基準測試平台，該資料集包含 60,000 張訓練影像與 10,000 張測試影像，每張影像為 28×28 像素的灰階圖像，對應 0 至 9 共十個數字類別。選擇此資料集的考量在於其廣泛的可比較性，MNIST 已成為聯邦學習安全性研究的標準基準，採用相同的測試平台有助於與現有文獻進行公平的效能對照。模型架構方面，本研究採用一個標準的卷積神經網路，該網路包含兩個卷積層與兩個全連接層，其參數規模足以展現聯邦學習的典型特徵，同時保持實驗的運算可行性。

在資料分佈的設置上，本研究採用獨立同分佈的配置方式，將訓練資料均勻地隨機分配給所有客戶端。這項設計選擇源於本研究防禦機制的架構特性。如第 4 章所闡述，挑戰增強型委員會架構的防禦機制運作於共識層而非資料層，其核心功能在於透過經濟懲罰來威懾惡意的委員會行為，而攻擊能否成功取決於委員會的組成結構，與底層的資料分佈特性相互獨立。換言之，無論資料呈現獨立同分佈或高度異質的 Non-IID 分佈，攻擊者控制委員會的機率計算以及罰沒機制的觸發條件都維持不變。資料分佈影響的是攻擊成功後對模型造成的傷害程度，而非攻擊本身能否被阻止或攻擊

者是否會因此受到懲罰。由於本研究的核心目標在於驗證經濟安全性機制對攻擊行為的威懾效果，而此效果與資料分佈無關，因此 IID 環境已足以充分驗證本架構的核心效能。

## 5.1.2 基準方法與攻擊場景

為了公平評估本研究所提出方法的效能，實驗將挑戰增強型委員會架構與 BlockDFL 進行對照比較。BlockDFL 代表了當前基於委員會機制的區塊鏈聯邦學習系統的主流設計，其採用固定規模的驗證委員會並依賴誠實多數假設來確保系統安全性。在本實驗中，兩種方法均採用相同的委員會規模  $C = 7$ ，這是 BlockDFL 原始論文中建議的配置，旨在效率與基本安全性之間取得平衡。共識機制方面，兩者皆採用  $2/3$  的 BFT 共識門檻，意即提案必須獲得超過三分之二委員會成員的同意方能通過。本研究方法與 BlockDFL 的關鍵差異在於引入了事後挑戰機制：在正常運作情況下，系統採用即時執行模式以維持高效率；當任何節點偵測到異常時，可透過質押押金發起挑戰，觸發全網仲裁流程並對確認的惡意行為執行罰沒。

攻擊場景的設計嚴格遵循第 3 章所定義的漸進式委員會佔領攻擊模型。攻擊者採取引性的兩階段策略：在潛伏階段，攻擊者控制的節點完全模仿誠實行為，正常參與訓練並提交高品質的模型更新，目的在於累積權益並建立信譽，為後續攻擊奠定資源基礎；一旦攻擊者在某輪委員會選舉中獲得超過  $2/3$  的席位，便立即進入佔領階段並根據當時的控制情況啟動攻擊。佔領階段的攻擊行為分為兩種場景：當攻擊者僅控制委員會多數但未同時控制聚合者時，採取戰略性餓死策略，系統性地拒絕包含誠實節點更新的提案，藉此獨佔該輪獎勵並阻止誠實節點的權益增長；當攻擊者同時控制委員會與聚合者時，則採取更激進的全棧投毒策略，直接將包含標籤翻轉等惡意內容的模型更新強制寫入區塊鏈。這兩種攻擊場景分別對應了經濟層面的治理操縱與技術層面的模型破壞，全面涵蓋了理性攻擊者可能採取的策略空間。

## 5.1.3 實驗參數配置

表 5.1 彙整了本研究實驗所採用的系統參數配置。這些參數的設定參考了 BlockDFL [15] 等主流區塊鏈聯邦學習研究的標準配置，以確保實驗結果的可比較性與



可重現性。

表 5.1: 實驗參數配置

參數名稱	設定值
訓練輪數	$R = 300$ (基礎實驗) / $R = 2000$ (長期實驗)
驗證者池規模	$N = 100$
委員會大小	$C = 7$
惡意節點數量	$M = 30$ (初始權益佔比 30%)
初始權益分配	所有節點均分配 100 單位
角色配置	聚合者 4 位，其餘為更新提供者
每輪獎勵	驗證者 1.0，聚合者 1.0，更新提供者 0.05
罰沒規則	挑戰成功時全額沒收惡意節點質押
學習率	$\eta = 0.01$ (衰減率 0.99)
本地訓練	每輪 1 個 epoch，批次大小 32

在參數設定的考量上，惡意節點初始佔比設定為 30% 代表了一個相當嚴峻的威脅情境，這個比例接近大多數拜占庭容錯系統所能容忍的理論上限。委員會規模  $C = 7$  的選擇則反映了效率與安全性的權衡，根據第 2.3.3 節的超幾何分佈分析，在此配置下惡意節點獲得委員會超過 2/3 席位的單輪機率約為 2.4%，這個機率雖然不高但在數百輪的訓練過程中足以產生多次成功攻擊的機會，為驗證防禦機制提供了充分的測試場景。獎勵機制的設定遵循 BlockDFL 的原始設計，驗證者與聚合者因承擔較高的運算與協調責任而獲得較高的單輪獎勵，更新提供者則獲得與其貢獻相稱的基礎獎勵，這種差異化的獎勵結構是權益累積動態的重要驅動因素。

## 5.2 實驗結果與分析

### 5.2.1 經濟安全性的核心驗證：權益動態分析

本研究的核心論點在於，挑戰增強型委員會架構能夠透過經濟懲罰機制打破漸進式委員會佔領攻擊所依賴的正反饋循環。為驗證此論點，本節首先分析兩種架構下惡意節點與誠實節點之間的權益演化動態，這是評估經濟安全性機制有效性的最直接指標。

在缺乏挑戰機制的 BlockDFL 架構中，實驗數據清楚揭示了權益失衡的系統性趨勢。追蹤原始權益數據可以發現，惡意節點的平均權益在訓練過程中穩定維持在誠實

節點的 1.1 至 1.2 倍，這種優勢並非偶然的統計波動，而是源於系統獎勵機制的結構性特徵。由於聚合者與驗證者角色所獲得的單輪獎勵顯著高於一般更新提供者，而角色分配機制又與節點權益直接掛鉤，一旦某些節點透過早期的隨機優勢累積了較高權益，其在後續輪次被選為高獎勵角色的機率便隨之提升，進而獲得更多獎勵並進一步鞏固優勢地位。這種「權益優勢帶來更多獎勵，更多獎勵鞏固權益優勢」的正反饋循環，正是第 3 章所描述的漸進式委員會佔領攻擊得以成立的經濟基礎。值得注意的是，惡意節點的權益優勢並未呈現指數級的無限膨脹，而是穩定在 1.1 至 1.2 倍的區間，這是因為攻擊者無法完全操控角色分配的隨機性，且系統中仍有部分誠實節點會在各輪次中獲得獎勵。然而，即便是這種「有上界」的優勢，在長期運作中仍構成顯著的治理威脅。

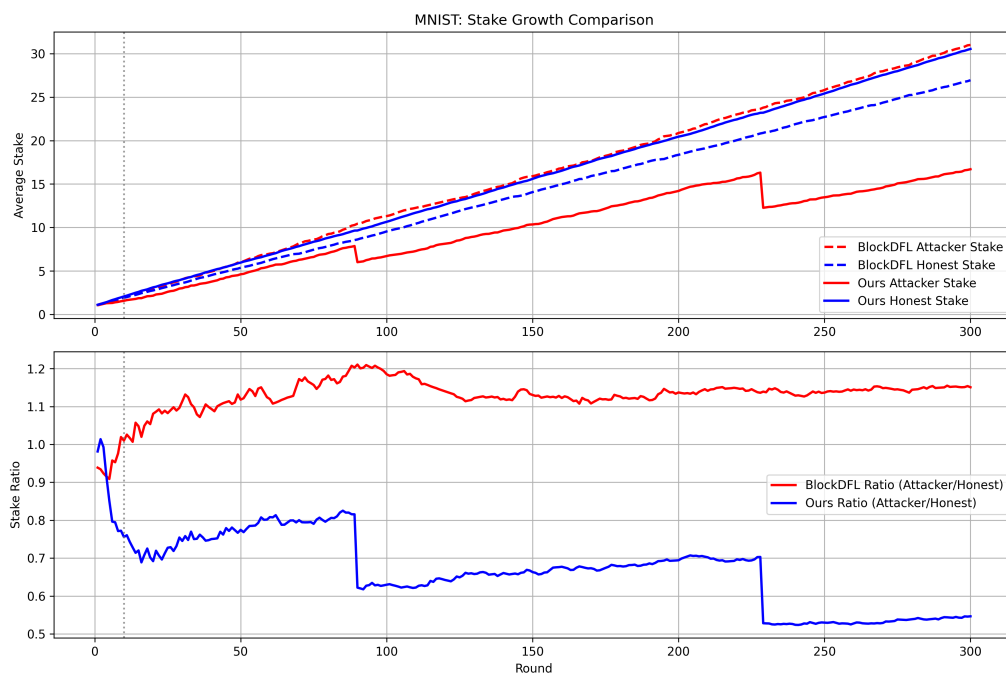


圖 5.1: BlockDFL 與 CACA 權益演化對比（300 輪基礎實驗）

挑戰增強型委員會架構的權益演化則呈現出截然不同的動態特徵。觀察圖 5.1 可以發現，惡意節點的平均權益並非呈現平滑的線性變化，而是展現出明顯的「台階式下降」模式。這種階梯狀的軌跡對應了異步審計機制觸發罰沒的具體時點：每當惡意節點嘗試利用委員會控制權發動攻擊時，挑戰者便會偵測到聚合結果與正確 Krum 運算之間的不一致，隨即發起挑戰並觸發全網仲裁，一旦惡意行為被確認，涉事節點的全額質押將被沒收。在 300 輪的觀察期內，這種罰沒事件在第 90 輪與第 229 輪各發生一次，導致惡意節點的平均權益在實驗結束時降至誠實節點的 0.56 倍。這個數值的意涵相當

深遠：惡意節點不僅未能如在 BlockDFL 中那樣建立起權益優勢，反而陷入了相對於誠實節點的劣勢地位，其在未來委員會選舉中獲得多數席位的機率將顯著降低。

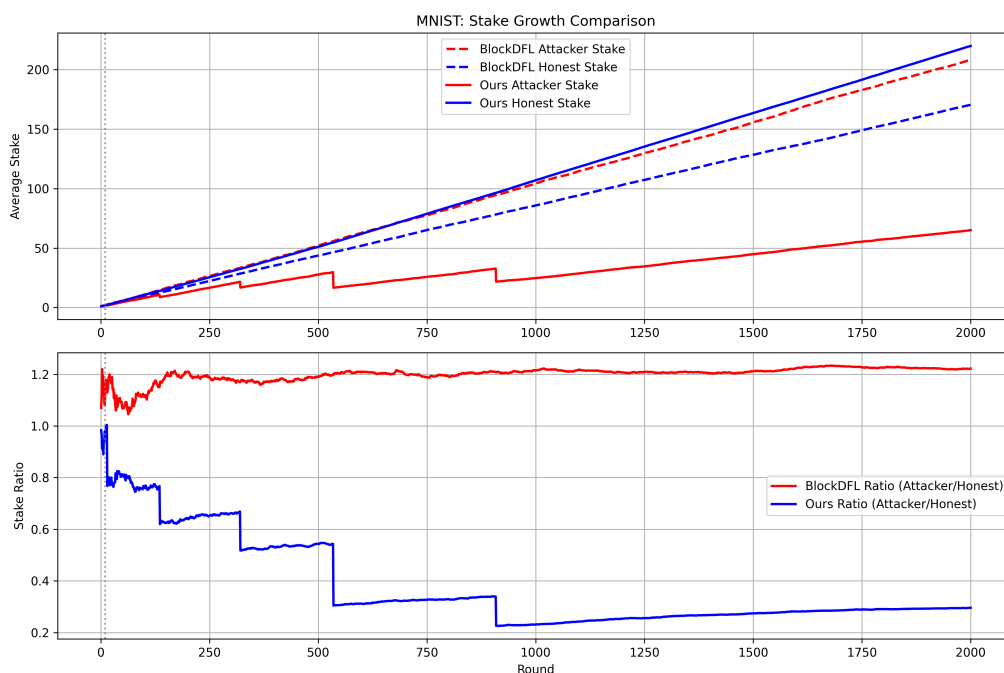


圖 5.2: 2000 輪長期模擬下的權益動態演化

將觀察期延伸至 2000 輪的長期實驗進一步印證了經濟懲罰機制的威懾效果。如圖 5.2 所示，在 BlockDFL 架構中，惡意節點的權益優勢在約 250 輪後便穩定維持在誠實節點的 1.2 倍左右，這種「強者恆強」的馬太效應使得攻擊者能夠長期維持其治理影響力。相較之下，本研究方法中的惡意節點權益軌跡在第 909 輪出現了決定性的轉折：該輪次發生的罰沒事件導致惡意節點的平均權益瞬間暴跌至誠實節點的 22.6%，這次經濟重創產生了持久的邊緣化效果。在隨後長達 1091 輪的觀察期內，惡意節點因權益基礎過於薄弱而徹底喪失了競爭委員會多數席位的能力，再也未能成功發動任何一次佔領攻擊。這項實驗結果有力地證實了經濟安全性機制的核心效能：罰沒機制能夠將一次性的攻擊失敗轉化為永久性的治理排除，從根本上打破了漸進式委員會佔領攻擊所依賴的正反饋循環。

## 5.2.2 攻擊抑制效果：頻率與成功率分析

權益動態的變化最終會反映在攻擊成功頻率這一更直觀的安全性指標上。本節分析兩種架構在不同實驗規模下的攻擊事件記錄，以量化經濟威懾對攻擊者行為的實質

影響。

在 300 輪的基礎實驗中，BlockDFL 共遭受了 10 次成功的委員會佔領攻擊，平均約每 30 輪便會發生一次安全事件。這個頻率與第 2.3.3 節的理論分析大致吻合：在惡意節點佔全網 30% 且委員會規模為 7 的配置下，單輪攻擊成功的機率約為 2.4%，300 輪訓練中的期望攻擊次數約為 7.2 次，實際觀測的 10 次略高於期望值但仍在合理的統計波動範圍內。這些攻擊事件中，部分為戰略性餓死類型，攻擊者透過拒絕誠實提案來獨佔獎勵；部分為全棧投毒類型，攻擊者在同時控制委員會與聚合者的情況下直接注入標籤翻轉的惡意更新。後者對模型造成的傷害尤為顯著，在實驗過程中曾導致模型準確率從正常水平瞬間大幅下降，展現了單次成功攻擊可能造成的災難性後果。

相較之下，挑戰增強型委員會架構在相同的 300 輪實驗中僅遭受 2 次成功攻擊，攻擊頻率降低了 80%。這項改善並非源於委員會組成機制的改變（兩種架構採用相同的權益加權選舉機制），而是來自兩個相互關聯的因素。其一是經濟威懾的直接效應：理性的攻擊者在評估攻擊決策時會將罰沒風險納入成本計算，當預期的懲罰損失超過潛在收益時，攻擊便不再是理性選擇，即使攻擊者恰好獲得了委員會的控制權，也可能選擇放棄攻擊以保全質押資產。其二是權益邊緣化的間接效應：如前節所述，罰沒機制導致惡意節點的權益持續萎縮，這直接降低了其在後續輪次被選入委員會的機率，從根本上減少了攻擊機會的出現頻率。

2000 輪的長期實驗更戲劇性地展現了上述效應的累積作用。在 BlockDFL 架構中，惡意節點在整個實驗期間共成功發動了 84 次委員會佔領攻擊，平均每 24 輪便會發生一次安全事件。這個高頻率的攻擊成功記錄意味著系統長期處於不穩定狀態，惡意節點能夠反覆利用權益優勢來操控委員會決策。本研究方法的表現則形成鮮明對比：2000 輪中僅發生 5 次成功攻擊，且這些攻擊全部集中在實驗的前 909 輪。如前節所述，第 909 輪的罰沒事件導致惡意節點權益暴跌至誠實節點的 22.6%，在此之後的 1091 輪中，攻擊者再也未能成功佔領委員會，形成了事實上的永久性治理排除。這項結果的意涵超越了單純的頻率降低：它證明了經濟安全性機制能夠在長期演化中將系統導向「誠實者主導」的穩定均衡，使得攻擊不僅在單次決策中不划算，在長期賽局中更是不可行。

### 5.2.3 系統穩定性與模型品質的連帶效益

攻擊頻率的降低會自然轉化為系統穩定性的提升，這種連帶效益雖然並非本研究的核心貢獻，但對於評估防禦機制的實務價值仍具重要參考意義。本節從系統可用性與模型收斂兩個面向分析這些連帶效益。

為量化攻擊對系統運作的實質衝擊，本研究定義「最低不可用率」作為評估指標，該指標衡量系統因攻擊而處於效能崩潰狀態的時間比例。根據實驗觀測，每次成功的全棧投毒攻擊會導致模型準確率急劇下降，而聯邦學習的自我修復機制需要約 5 至 25 輪才能使模型恢復至正常水平。採用最為保守的 5 輪恢復期估計，BlockDFL 在 300 輪實驗中因 10 次攻擊而產生至少 50 輪的不可用時間，對應約 16.7% 的最低不可用率；若採用實際觀測中較常見的 15 輪恢復期，不可用率將攀升至 50% 左右。本研究方法憑藉僅 2 次的攻擊記錄，將最低不可用率控制在 3.3% 以下。這種差異在需要持續運作的關鍵任務場景中具有重要意義：醫療診斷、自動駕駛決策或金融風險評估等應用不僅要求最終模型的準確性，更要求訓練過程中模型狀態的穩定性，頻繁的效能崩潰即使最終能夠恢復，仍會嚴重影響系統的可靠性與可用性。

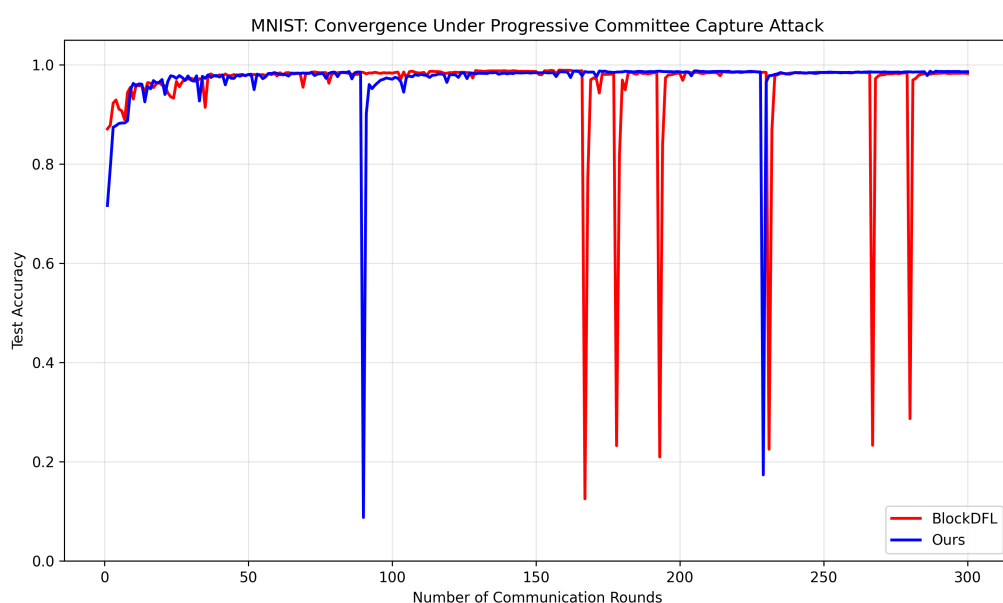


圖 5.3: 模型準確率收斂曲線比較

在模型收斂品質方面，圖 5.3 呈現了兩種架構在 300 輪實驗中的準確率變化軌跡。實驗結束時，本研究方法達成 98.63% 的最終準確率，BlockDFL 則為 98.26%，差異僅

約 0.37 個百分點。這個結果並不令人意外，因為聯邦學習本身具備顯著的自我修復能力：只要攻擊者無法持續佔領委員會，後續輪次中來自誠實節點的正確更新便能逐步抵銷惡意梯度所造成的偏差，使模型重新收斂至正確方向。從圖中可以觀察到，BlockDFL 的準確率曲線在某些輪次出現明顯的下降與回升波動，這些波動對應了全棧投毒攻擊造成的瞬時衝擊以及後續的自我修復過程。相較之下，本研究方法的準確率曲線則呈現更為平滑的上升趨勢，這反映了較低的攻擊頻率所帶來的訓練穩定性優勢。聯邦學習的自癒特性解釋了為何兩種架構的最終準確率差異不大，同時也凸顯了本研究關注經濟安全性而非單純模型品質的理論動機：在聯邦學習場景中，偶發的模型偏差可以被自然修復，真正需要擔憂的是攻擊者透過權益累積逐步掌控系統治理權的長期威脅。

值得強調的是，最終準確率的相近並不意味著兩種架構的效能等價。如前所述，BlockDFL 在訓練過程中經歷了遠為頻繁的效能崩潰與恢復週期，這種「過程不穩定性」在實務應用中可能造成嚴重後果。此外，聯邦學習的自癒能力存在其固有的界限：當攻擊頻率過高導致恢復期與下一次攻擊重疊時，模型可能陷入長期的低位震盪而無法有效累積學習成果。本研究方法透過將攻擊頻率抑制在極低水平，從根本上避免了這種連鎖惡化的風險，確保了模型能夠在絕大多數訓練時間內維持穩定的學習軌跡。

## 5.3 本章小結

本章透過系統性的實驗設計，從權益動態、攻擊頻率與系統穩定性三個層面驗證了挑戰增強型委員會架構的防禦效能。實驗結果有力地支持了本研究的核心論點：經濟安全性機制能夠有效替代傳統的門檻安全性假設，在不擴大委員會規模的前提下達成等效甚至更優的安全保證。

在經濟安全性的核心驗證方面，權益動態分析揭示了兩種架構的本質差異。BlockDFL 中惡意節點能夠穩定維持 1.1 至 1.2 倍的權益優勢，這種優勢在 2000 輪的長期實驗中持續存在並轉化為對系統治理權的顯著影響力。本研究方法則透過罰沒機制成功打破了這種正反饋循環：惡意節點的權益軌跡呈現「台階式下降」的特徵，在第 909 輪的關鍵轉折點後權益暴跌至誠實節點的 22.6%，達成了事實上的永久性治理排除。這一結果從實證角度印證了第 4.5 節的理論分析，證明經濟懲罰機制確實能夠重塑

理性攻擊者的決策空間，使得誠實行為成為長期均衡下的最優策略。

在攻擊抑制效果方面，實驗數據展現了經濟威懾的實質影響力。300 輪實驗中攻擊成功次數從 BlockDFL 的 10 次降至本研究方法的 2 次，降幅達 80%；2000 輪長期實驗中的對比更為懸殊，BlockDFL 的 84 次對比本研究方法的僅 5 次，且後者的攻擊全部集中在前 909 輪，此後再無成功攻擊的記錄。這種攻擊頻率的大幅降低不僅源於經濟威懾的直接效應，更源於權益邊緣化所產生的間接效應：惡意節點因質押萎縮而逐漸喪失競爭委員會席位的能力，攻擊機會本身便大幅減少。

在系統穩定性的連帶效益方面，攻擊頻率的降低自然轉化為可用性的提升，最低不可用率從 BlockDFL 的約 16.7% 降至本研究方法的 3.3% 以下。同時，兩種架構的最終模型準確率相近（98.63% 對 98.26%），這證實了經濟安全性機制在強化防禦的同時並未犧牲聯邦學習的核心功能。這一發現呼應了第 4 章的設計哲學：聯邦學習本身具備自我修復能力，真正需要防禦的不是偶發的模型偏差，而是攻擊者透過權益累積逐步奪取治理權的長期威脅。

綜上所述，實驗結果有力地驗證了挑戰增強型委員會架構的核心價值主張。透過將安全性保障從機率運算轉移至激勵相容，本架構成功解耦了安全性與委員會規模之間的傳統強耦合關係，開闢了一條兼顧效率與安全的技術路徑。這種基於經濟理性的自我維持安全機制，為大規模去中心化機器學習平台的實際部署奠定了可行的技術基礎。

## 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。資料顯示，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」。資料指出，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資料與應用場景特徵，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，並確保效率最佳化不會因過度縮減委員會而產生不可預見的安全缺口。

## 参考文献

- [1] 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.
- [2] 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.
- [3] 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).
- [4] 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).
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] W. Li et al. “EPP-BCFL: Efficient and Privacy-Preserving Blockchain-Based Federated Learning”. In: *Sci. Rep.* (2025).
- [11] Miguel Castro and Barbara Liskov. “Practical Byzantine fault tolerance”. In: *OSDI*. Vol. 99. 1999. 1999, pp. 173–186.
- [12] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.
- [13] M. Wang et al. “A Blockchain-Based Federated Learning Framework for Vehicular Networks”. In: *Sci. Rep.* (2024).
- [14] J. Zhang et al. “FedChain: A blockchain-based federated learning framework with adaptive client selection”. In: *Proc. VLDB Endow.* (2024).

- [15] 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.
- [16] Y. Li et al. “A blockchain-based decentralized federated learning framework with committee consensus”. In: *IEEE Netw.* 35.1 (2021), pp. 234–241.
- [17] 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.
- [18] 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.
- [19] Yossi Gilad et al. “Algorand: Scaling byzantine agreements for cryptocurrencies”. In: *Proceedings of the 26th Symposium on Operating Systems Principles*. 2017, pp. 51–68.
- [20] P. Blanchard et al. “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent”. In: *NeurIPS*. 2017.
- [21] D. Yin et al. “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates”. In: *ICML*. 2018.
- [22] D. H. Nguyen et al. “FedBlock: A Blockchain Approach to Federated Learning against Backdoor Attacks”. In: *Proc. IEEE Big Data*. 2024.
- [23] C. Zhang et al. “BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning”. In: *Proc. USENIX Annu. Tech. Conf. (ATC)*. 2020.
- [24] Y. Miao et al. “Privacy-Preserving Byzantine-Robust Federated Learning via Blockchain Systems”. In: *IEEE Trans. Inf. Forensics Security* 17 (2022), pp. 2848–2861.
- [25] H. Chen et al. “Robust blockchained federated learning with model validation and proof-of-stake inspired consensus”. In: *arXiv preprint arXiv:2101.03300* (2021).
- [26] J. Chiu and T. V. Koepl. *Incentive Compatibility on the Blockchain*. Tech. rep. 2018-34. Bank of Canada, 2018.
- [27] 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.
- [28] E. Bagdasaryan et al. “How to backdoor federated learning”. In: *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*. 2020, pp. 2938–2948.
- [29] L. Zhu, Z. Liu, and S. Han. “Deep Leakage from Gradients”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.
- [30] 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.

- [31] 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.
- [32] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.
- [33] Mahdi Zamani, Mahnush Movahedi, and Mariana Raykova. “RapidChain: Scaling Blockchain via Full Sharding”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Toronto, Canada: ACM, 2018, pp. 931–948. doi: 10.1145/3243734.3243853.
- [34] V. Buterin et al. “Combining GHOST and Casper”. In: *arXiv preprint arXiv:2003.03052* (2020).
- [35] B. J. Chen et al. “ZKML: An Optimizing System for ML Inference in Zero-Knowledge Proofs”. In: *Proc. EuroSys*. 2024.
- [36] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [37] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [38] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *International Conference on Learning Representations*. 2015.
- [39] 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.
- [40] K. Conway et al. “opML: Optimistic Machine Learning on Blockchain”. In: *arXiv preprint arXiv:2401.00000* (2024).
- [41] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [42] ORA Protocol. *opML documentation*. docs.ora.io. 2024.
- [43] Optimism Foundation. *Rollup protocol overview*. docs.optimism.io. 2024.
- [44] Harry Kalodner et al. “Arbitrum: Scalable, private smart contracts”. In: *27th USENIX Security Symposium (USENIX Security 18)*. 2018, pp. 1353–1370.