

激励驱动的联邦学习研究综述：隐私与安全

迟欢欢¹, 熊平¹, 刘恒竹¹, 马霄¹, 朱天清²

(1. 中南财经政法大学信息工程学院, 湖北 武汉 430073; 2. 澳门城市大学数据科学学院, 澳门 999078)

摘要: 联邦学习使多个数据持有者能够在不泄露本地数据的前提下协同完成机器学习任务, 而鼓励参与者加入联邦学习并提供高质量数据是其成功的关键因素之一。然而, 联邦学习在实际应用中面临隐私泄露和安全威胁双重挑战。一方面, 恶意参与者通过主动攻击干扰联邦学习的进程和结果, 破坏其有效性和鲁棒性; 另一方面, 被动攻击造成的隐私泄露风险对参与者加入联邦学习的主观意愿造成负面影响, 使激励机制的实施变得更加困难。近年来, 学术界从隐私保护和安全防御的角度对激励驱动的联邦学习进行了广泛研究, 旨在从安全性和公平性兼顾的视角, 为联邦学习提供整体解决方案。首先, 介绍了激励驱动联邦学习面临的被动攻击和主动攻击, 深入分析这些攻击引发的隐私泄露风险和安全威胁。然后, 分别从隐私保护和安全防御两个方面对激励驱动的联邦学习研究进行综述和分析。在隐私保护方面, 着重介绍了差分隐私和同态加密两种技术在激励驱动联邦学习中的应用; 在安全防御方面, 重点梳理了结合激励机制来防御投毒攻击和搭便车攻击的各种方法。最后, 讨论了激励驱动的联邦学习在应对隐私和安全攻击方面面临的挑战, 并展望了这一领域未来的研究方向。

关键词: 联邦学习; 激励机制; 隐私保护; 差分隐私; 博弈论

中图分类号: TP181

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025035

Survey on incentive-driven federated learning: privacy and security

CHI Huanhuan¹, XIONG Ping¹, LIU Hengzhu¹, MA Xiao¹, ZHU Tianqing²

1. School of Information Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China

2. Faculty of Data Science, City University of Macau, Macau 999078, China

Abstract: Federated learning was enabled to allow multiple data holders to jointly complete machine learning tasks without disclosing local data. Incentivizing participants to engage in federated learning and contribute high-quality data was identified as one of the key factors for its success. However, federated learning was found to face chal-

收稿日期: 2025-01-27; 修回日期: 2025-04-16

通信作者: 熊平, pingxiong@zuel.edu.cn

基金项目: 高等学校学科创新引智基地 (B21038); 湖北省自然科学基金 (2025AFC108, 2024AFB957)

Foundation Items: Innovation and Talent Base for Digital Technology and Finance (B21038); The Natural Science Foundation of Hubei Province (2025AFC108, 2024AFB957)

引用格式: 迟欢欢, 熊平, 刘恒竹, 等. 激励驱动的联邦学习研究综述: 隐私与安全[J]. 网络与信息安全学报, 2025, 11(3): 19-44.

Citation Format: CHI H H, XIONG P, LIU H Z, et al. Survey on incentive-driven federated learning: privacy and security[J]. Chinese Journal of Network and Information Security, 2025, 11(3): 19-44.

allenges in both privacy leakage and security threats in practice. On the one hand, malicious participants were observed to launch active attacks to disrupt the process and results of federated learning, compromising its effectiveness and robustness. On the other hand, privacy leakage risks caused by passive attacks were shown to negatively impact participants' willingness to join federated learning, making the implementation of incentive mechanisms more difficult. In recent years, extensive research was conducted by the community on incentive-driven federated learning from the perspectives of privacy preservation and security defense, aiming to provide comprehensive solutions that balance security and fairness for federated learning. First, various passive and active attacks faced by incentive-driven federated learning were introduced, and the privacy leakage risks and security threats posed by these attacks were analyzed. Subsequently, a comprehensive review and analysis of incentive-driven federated learning research were provided from the perspectives of privacy preservation and security defense. In terms of privacy preservation, the application of differential privacy and homomorphic encryption technologies in incentive-driven federated learning was focused on. In terms of security defense, various methods that combine incentive mechanisms to defend against poisoning attacks and free-riding attacks were specifically outlined. Finally, the challenges still faced by incentive-driven federated learning in addressing privacy leakage and security attacks were discussed, and potential future research directions in this field were explored.

Keywords: federated learning, incentive mechanisms, privacy preservation, differential privacy, game theory

0 引言

联邦学习 (federated learning, FL) 作为一种典型的多方联合学习框架, 已在诸多领域得到广泛应用^[1]。经典 FL 框架采用聚合端与客户端的架构设计, 旨在让多个参与者在泄露本地数据的前提下协同完成机器学习任务, 从而规避数据共享引发的隐私泄露风险。显然, 参与者的数据规模、数据质量及其在学习过程中的参与程度, 均是影响 FL 效果的关键因素。当客户端参与 FL 任务时, 不可避免地会消耗本地设备资源, 包括算力、带宽、电量及数据资源等^[2]。因此, 对 FL 的参与者给予公平、合理的奖励或补偿是十分必要的, 这使得激励驱动的 FL (incentive-driven federated learning, IDFL)^[3]成为 FL 领域的重要研究方向之一。

IDFL 是一种将激励机制深度融入 FL 迭代流程的学习框架, 其核心目标是通过激励机制对参与者及其提供的计算资源、数据资源进行持续优化, 以提升 FL 的效率与模型性能^[4]。与传统 FL 相比, IDFL 的独特性在于将激励功能嵌入 FL 算法内核, 实现对学习过程的微观调控。以 Liu 等^[5]提出的 IDFL 框架为例, 服务器在每轮训练中通过 GTG-Shapley 算法评估每位参与者的贡献度, 并基于不同参与者的本地模型更新组合生成多个 FL 聚合子模型。通过对比这些子模型的性

能指标 (如准确率、损失值等), 服务器筛选出最优子模型, 并聚合该子模型中参与者的更新参数以构建新的全局模型, 从而提升模型的整体准确率。

近年来, 学术界提出了许多应用于 FL 的激励机制, 主流方法以博弈论、拍卖理论和契约理论为理论基石^[6-7]。然而, 一系列针对 FL 框架安全性与鲁棒性的研究表明, 传统 FL 及其激励机制面临诸多威胁。Song 等^[8]的研究表明, 攻击者可通过分析 FL 过程中的中间梯度参数, 推测出参与者本地数据的相关信息。Zhang 等^[9]指出由于服务器无法直接验证数据的真实性, 攻击者可以直接操纵训练数据并将恶意更新发送给服务器, 或者通过本地模型对全局模型的优化过程进行恶意干扰。此外, 好奇的服务器可以通过构建生成对抗网络^[10]来推测客户训练数据的敏感信息, 从而危及用户数据的隐私性和完整性。激励欺诈攻击^[11]会导致奖励分配不公, 降低参与者加入 FL 的积极性。

参考信息安全领域的攻击分类方法, 本文将 IDFL 面临的攻击划分为被动攻击与主动攻击两类。其中, 被动攻击通过截获、推理等手段窃取隐私信息 (如本地数据特征、数据成员隶属关系等)。主动攻击则通过恶意干扰或信息篡改破坏 FL 训练进程。相应地, 本文将被动攻击和主动攻击对 IDFL 产生的威胁定义为隐私威胁和安全威胁。其中, 隐私威胁来自推理攻击、无意的数

据泄露、重构攻击和恶意服务器攻击等；安全威胁则主要来自投毒攻击和搭便车攻击。对于IDFL而言，被动攻击引发的隐私泄露风险，要求其必须强化对数据、模型及各类中间变量的保护，这无疑增加了参与者数据质量评价与贡献评估的复杂性。而不诚实或恶意参与者发起的主动攻击会直接破坏IDFL的训练进程与结果，因此在激励机制中嵌入异常行为识别与惩罚机制至关重要。

面对上述挑战，学者从兼顾安全性和公平性的目标出发，将差分隐私、区块链、同态加密等安全原理和技术应用到FL中，在传统基于博弈论、契约理论、拍卖理论的激励机制中嵌入攻击防御功能，提出了一系列适用于不同场景的IDFL方案，力求同步解决攻击防御与激励驱动问题，构建安全可靠的IDFL系统，为FL提供整体解决方案。本文从隐私保护和安全防御两个维度对IDFL相关研究展开综述，旨在梳理和比较实现安全IDFL的新思路与新方法，并探讨该领域现存的挑战及未来研究趋势。

1 联邦学习与激励机制

1.1 联邦学习

联邦学习是指多个客户端在服务器的协调下协作完成机器学习任务的模式^[12]。在经典的FL框架中，参与者首先在本地上设备上利用自有数据训练本地模型，随后将模型参数或梯度信息上传至服务器。服务器对这些参数或梯度信息进行聚合处理，生成全局模型，并将更新后的全局模型反馈给各参与者。如此往复迭代，直至全局模型达到收敛状态。FL的框架如图1所示。

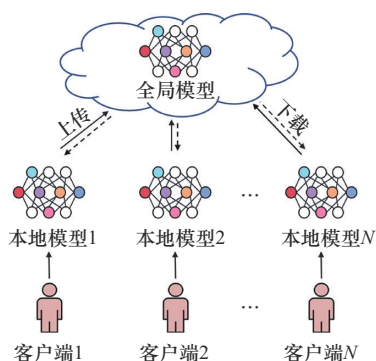


图1 FL的框架

Figure 1 Federated learning framework

设 $\mathcal{N} = \{1, \dots, N\}$ 表示 N 个客户端的集合， D_i 是客户端 i 的本地数据集。典型的FL训练过程包括以下3个阶段。

1) 初始化阶段：服务器首先确定全局模型的架构，并根据训练任务随机地或通过公共数据集上预训练的方式，生成全局模型的初始参数，然后服务器将初始化的模型参数 w_0 分发给选定的客户端。

2) 本地模型训练阶段：在第 t 轮训练中，每个被选定的客户端 i 使用其本地数据集 D_i ，根据接收到的全局模型更新其本地模型 w_t^i ，并将更新后的本地模型参数 w_{t+1}^i 上传至服务器。客户端 i 在第 t 轮中的目标是最小化本地损失 $F(w_t^i)$ 。

3) 全局模型聚合：在每轮更新后，服务器根据客户端的本地更新进行安全聚合操作，然后将更新后的全局模型 w_{t+1} 发送给选定的客户端。

重复步骤2)和步骤3)，直到满足停止准则，目标是最小化全局损失 $F(w_t)$ 。

根据数据分布特性的差异，FL可以分为横向FL、纵向FL和联邦迁移学习等^[13]。当前广泛采用的FL算法包括FedAvg^[14]、FedProx^[15]、FedMA^[16]和Scaffold^[17]等。

1.2 激励机制

1.2.1 激励机制的定义

在FL系统中，激励机制旨在通过公平的奖励分配策略来激励高质量客户端参与模型训练，使资源消耗（如通信和计算成本等）最小化和模型性能（如模型精度和收敛速度等）最大化。

定义1 激励机制

客户端 i 对学习任务的贡献用多维向量 $\mathbf{v}_i$ 表示，其对应的参数为 λ_i 。当 $\mathbf{v}_i = \vec{0}$ 时，表示 i 未参与FL训练。作为一个理性的个体， i 以利润最大化为目标。

$$u_i = \theta_i - c_i(\mathbf{v}_i, \lambda_i) \quad (1)$$

其中， θ_i 为服务器向参与训练的客户端支付的报酬。 $c_i(\cdot)$ 表示 i 的成本函数。服务器（或模型所有者）以收益最大化为目标。

$$u = V(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N) - \sum \theta_i \quad (2)$$

其中， $V(\cdot)$ 表示效用函数。

激励机制的设计需要考虑协作实体之间的信息不对称问题。在FL场景中，信息不对称表现为服务器无法预知客户端的参与意愿，同时对参

与者的信誉、数据质量、计算资源和数据规模等信息也缺乏了解^[18]。同时,激励机制需兼顾激励兼容性,即促使参与者的局部利益与系统的整体目标趋同,实现个体理性与全局理性的统一。另外,激励机制还需要满足公平性、帕累托效率、共谋抵抗、预算平衡、性能改进等方面的需求,以保障激励机制的实际效能^[19]。

1.2.2 激励机制的主要实现方法

(1) 夏普利值 (Shapley value, SV)

SV是指参与者在各种合作组合中的边际贡献^[20]。在FL激励机制中,SV常被用于衡量每个参与者对联邦的贡献度,其分配机制通常基于公平性、个体理性和可加性等原则。给定参与者*i*,其SV的形式化定义为:

$$\phi_i(N, V) = \frac{1}{N} \sum_{S \subseteq N \setminus \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{|N| - 1}{|S|}} \quad (3)$$

其中,*S*表示所有参与者集合*N*的子集,*V*(·)表示FL的效用函数。SV的计算代价通常较高,可以利用蒙特卡罗、分组检验、概率估计和适当的信息共享等抽样技术^[21]来逼近真实值。

(2) Stackelberg 博弈

Stackelberg模型是一种动态博弈模型,用于分析不对称场景下(合作者中存在领导者与跟随者关系)合作者之间的竞争行为^[22]。Stackelberg博弈的解是一种纳什均衡,通常采用逆向归纳法推导跟随者的最优策略,在此基础上选择领导者的最优策略^[23]。在IDFL中,模型拥有者是领导者,数据所有者是跟随者。模型拥有者通过分析跟随者的预期反应为特定的FL任务设置奖励,以最大化全局效用。数据所有者则观察领导者的行动,然后决定其局部训练策略并试图最大化本地效用。

(3) 契约理论

契约理论用于在合作或冲突场景下通过契约设计来激励个体行为,确保各方在契约中得到公平对待^[24]。在FL的语境中,服务器是雇主,数据所有者是雇员。服务器发布合约,数据所有者则选择最适合自己的合约并执行FL任务以获得报酬。由于服务器对数据所有者的数据规模、数据质量等信息掌握有限,即存在信息不对称问题,某些参与者遭受不公平的待遇。利用契约理论设计合理的激励方案,通过确立明确的契约目标来引导参与者的行为,能够有效解决这一问题。

(4) 拍卖理论

拍卖理论在FL中可应用于定价、任务分配和客户端选择等场景^[25]。拍卖理论的核心参与者包括拍卖人与投标人,拍卖的过程可以概括为招标公告、投标收集、中标者确定及其他的辅助程序^[26]。在FL中,模型拥有者是拍卖人,数据所有者是投标人。每个数据所有者根据其训练成本提交出价,模型拥有者根据这些出价来选择优胜者参与FL,如图2所示。利用拍卖理论可以鼓励参与者报告自己的真实信息(如数据特征和隐私成本),从而提高FL的效率,降低计算和通信开销。表1对这些主要实现方法进行了详细对比。

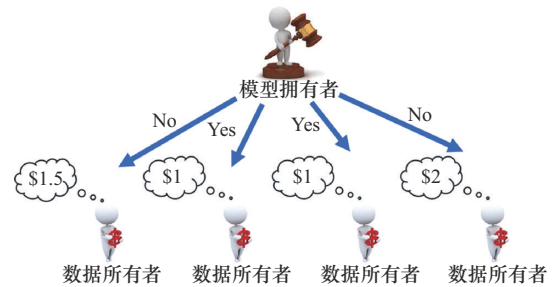


图2 基于拍卖理论>IDFL

Figure 2 IDFL based on auction theory

表1 激励机制的主要实现方法对比

Table 1 Comparison of main implementation methods for incentive mechanisms

实现方法	优势	局限
夏普利值	公平性强,能精准衡量每个参与者的贡献度	计算复杂度高;难以动态调整
Stackelberg 博弈	设计灵活;能有效分析领导者与跟随者之间的决策互动	需要精确预测跟随者反应;领导者可能占据过多收益,导致不公
契约理论	能明确约束和激励参与者行为;有效解决了信息不对称问题	成本较高;难以适应动态环境
拍卖理论	能够有效分配资源;适合竞争性场景	可能引发过度竞争,导致资源分配不均

2 IDFL面临的隐私和安全威胁

FL通过避免参与者本地数据的直接共享，为多方联合学习提供了隐私保护的学习范式。然而，其分布式学习框架与开放的通信环境给模型和数据安全带来了多重威胁，如推理攻击、投毒攻击和生成式对抗网络攻击等^[27]。与传统FL相比，IDFL由于引入激励机制，在一定程度上加剧了隐私与安全风险。一方面，激励机制的实现需涉及参与者的多维信息（如数据规模、数据质量、计算资源及隐私偏好等），为攻击者提供了更多可利用的资源；另一方面，激励机制的设计缺陷可能成为攻击者利用的漏洞（例如，攻击者可通过搭便车攻击破坏激励机制的有效性，导致激励分配不公）。

本节对IDFL中常见的攻击展开介绍，并依据攻击目标将其划分为隐私威胁和安全威胁两大类，如图3所示。隐私威胁是指攻击者利用掌握或截获的资源，推测未被授权的敏感信息（如数据属性、成员隶属关系、模型信息等），但不直接干扰FL的训练过程与结果；安全威胁则指攻击者通过恶意手段主动干扰或误导FL进程，从而导致模型训练效率下降或最终性能受损。

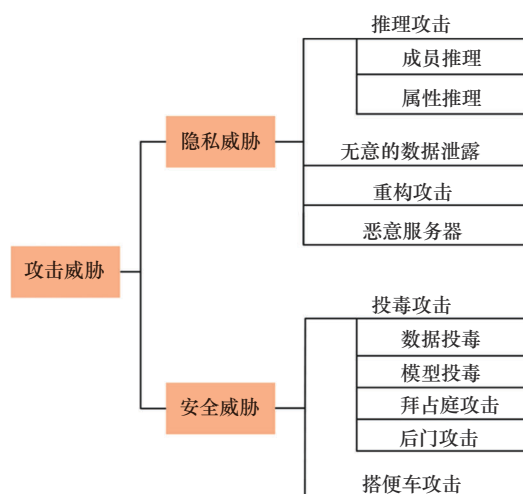


图3 IDFL面临的主要攻击威胁

Figure 3 Main attack threats faced by IDFL

2.1 隐私威胁

(1) 推理攻击

推理攻击是指攻击者通过分析模型的参数或输出，试图推断出训练数据的敏感信息^[28]。根

据不同推理目的，推理攻击可分为成员推理攻击和属性推理攻击两类^[29]。

成员推理攻击的目标是通过分析模型的输出，推断目标样本是否属于模型的训练数据集^[30]。由于某些机器学习任务本身具有一定的敏感性，因此确定一个样本属于训练数据集有可能揭示目标样本的隐私信息。在FL中，成员推理攻击可通过以下两种典型方式实施。

1) 被动成员推理攻击：攻击者通过截获模型参数，在不干扰本地或全局训练过程的前提下，推断目标样本是否属于一方或多方所使用的私有训练数据。

2) 主动成员推理攻击：攻击者主动向目标模型发送特定的查询请求，并通过分析模型的输出进行成员推理，典型的攻击方式是构建影子数据集来训练攻击模型并实施攻击^[31]。在基于生成对抗网络的成员推理攻击^[32]中，攻击者利用生成对抗网络生成具有特定特征的虚假数据，并将其注入FL模型的训练流程，通过观察模型对这些虚假数据的响应，推断目标样本是否用于训练某个模型。

属性推理攻击是指攻击者利用FL过程中服务器与客户端的交互信息及其他先验知识，推断目标训练集是否存在特定的敏感属性，或者推断训练集中敏感属性的数据分布或统计特征等信息的攻击方式^[33]。通常，攻击者通过观察并分析FL过程中模型更新信息，推断参与者本地训练数据中包含的隐私属性，如个人身份、健康状况、性别和住址等敏感信息。此外，攻击者还可能主动介入FL的训练过程，以推断参与者训练数据中的敏感隐私属性。Shen等^[34]在服务器端植入分类器，以每轮聚合后的模型更新为输入，预测可能包含特定目标属性的参与者集合。

(2) 无意的数据泄露

在FL中，无意的数据泄露是指程序或系统配置错误、数据处理不当或模型设计缺陷，导致参与者的敏感数据在非预期的情况下被泄露。这类泄露不仅直接威胁个体隐私，还可能被攻击者应用于模型投毒、成员推理等攻击，进而损害系统的有效性与可信度。Melis等^[35]

的研究表明,攻击者可以利用无意的数据泄露漏洞,通过属性推理攻击成功重建特定客户端的数据。

(3) 重构攻击

重构攻击是指攻击者试图从FL的公开信息(如梯度更新、模型输出)中重建客户端的数据或模型的攻击行为^[36]。根据攻击者重构目标的不同,重构攻击可细分为模型重构攻击^[37]和数据重构攻击^[38]。模型重构攻击的目标是模型本身,攻击者尝试重建整个模型或其关键部分。而数据重构攻击的目标是模型的训练数据,攻击者试图根据模型的输出重构部分或全部训练样本及其真实标签。与推理攻击不同,数据重构攻击旨在恢复完整的训练样本,而推理攻击侧重于推断样本的成员关系或数据统计信息。

(4) 恶意服务器

在FL环境中,服务器作为整个框架的中枢,一旦被未授权用户控制或存在好奇行为,可能通过与参与者的交互信息推理出客户端数据,或操纵全局模型的训练过程并植入恶意任务^[39]。此外,恶意服务器还可能以不连续的方式分析客户端共享的模型,或者监控模型随时间的变化,以进一步挖掘客户端的敏感信息^[36]。

2.2 安全威胁

(1) 投毒攻击

投毒攻击是FL中最常见的主动攻击方式之一。如图4所示,攻击者通过向FL过程注入虚假或有害的数据来破坏模型训练过程,目的是操控模型的输出,使模型产生错误判断^[40]。根据实施阶段不同,投毒攻击可分为数据投毒和模型投毒;根据攻击目的的不同,可以分为拜占庭攻击和后门攻击^[41]。

数据投毒发生在数据收集阶段,其目的是通过向训练数据中插入恶意数据点,在全局模型中制造偏差^[42]。数据投毒通常需要攻击者篡改本地数据集。例如,攻击者通过标签翻转^[43]将样本 x_i 的真实标签 y_i 更改为对抗标签 a ,生成对抗样本 $\{(x_i, a)\}$ 用来训练本地模型,并在多次训练迭代中重复此操作^[44]。由于单次篡改的幅度微小,服务器难以实时识别异常。随着

训练迭代的推进,全局模型会逐步累积偏差,最终导致模型对特定类型样本输出错误的预测结果。

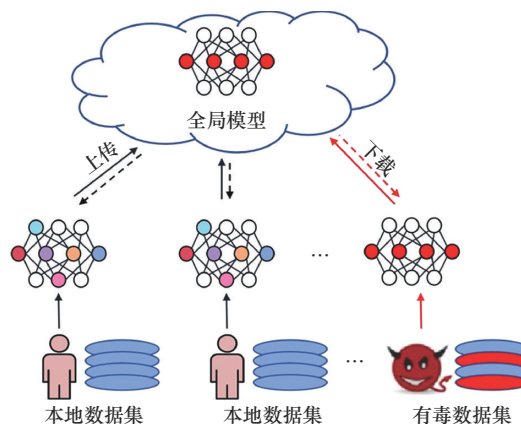


图4 投毒攻击的过程

Figure 4 The process of poisoning attack

模型投毒攻击发生在模型训练阶段^[45]。例如,攻击者训练恶意本地模型 w ,并乘以一个增强因子 λ ,将 λw 提交给服务器,以增强并逐步积累对全局模型的影响。通过这种方式,攻击者可以向全局模型注入特定的错误行为,而无须修改本地数据集。近期的一些研究表明,模型投毒攻击比数据投毒攻击更加有效^[46-47]。

拜占庭攻击属于非定向投毒攻击,其攻击目标并非特定用户或数据样本,而是通过恶意掉线、上传随机或恶意模型更新参数等行为,阻碍模型训练的有效收敛^[48]。在这类攻击中,攻击者通常被假定为能够控制多个参与者的资源,并可上传任意数据的敌手。这使得攻击者可以通过篡改本地数据或模型,实现分布式投毒攻击^[49],通过消耗大量资源或发送大量无效请求来干扰或阻碍FL过程^[50]。

后门攻击属于定向投毒攻击,其目标是在模型中植入隐蔽后门,并在模型部署后通过激活特定输入模式触发恶意行为,导致模型在特定场景下输出错误预测^[51]。这类攻击的典型特征是不影响模型在主任务上的正常性能,因此不易引起系统异常警觉^[52],其隐蔽性使其相较于拜占庭攻击更难被检测。

(2) 搭便车攻击

搭便车攻击是指恶意参与者在FL系统中获

取模型更新或全局模型等收益，却不主动贡献自身数据或计算资源的行为^[53]。攻击者通常伪装成正常客户端，向聚合器发送随机更新的模型权重或与聚合模型近似的权重，误导聚合器将其纳入模型聚合过程。此类攻击不仅破坏了FL系统的公平性与效率，还会显著降低全局模型的收敛速度^[54]。

面对上述隐私和安全威胁，传统IDFL逐步引入一系列隐私保护与安全防御方法，以适应新的安全FL场景。图5展示了这些方法的分类及其依托的理论技术，本文第3节和第4节将分别对隐私保护方法与安全防御方法展开梳理与讨论。

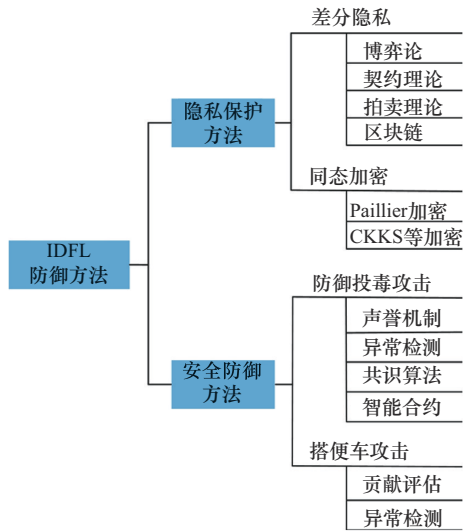


图5 IDFL的隐私保护和安全防御方法

Figure 5 Privacy protection and security defense methods in IDFL

3 面向隐私保护的IDFL

隐私保护是FL最基本的目标，本地数据、模型及其更新信息均属于FL系统需要保护的敏感信息^[55]。在IDFL中，激励机制涉及的客户真实成本、拍卖出价等信息，常因客户需求或环境设定被纳入隐私范畴。当前，隐私保护方法主要分为数据随机化与数据加密两类，对应的，IDFL中多数方案通过差分隐私或同态加密技术实现隐私防护。

3.1 差分隐私应用于IDFL

差分隐私是一种严格且可验证的隐私保护方法^[56]，被广泛应用于在隐私保护的前提下进行

数据发布或数据挖掘。其具体定义如下。

定义2 差分隐私

设有随机算法 A ， P_A 是 A 可能的输出构成的集合， O_A 是 P_A 的任意子集，给定两个相邻数据集 D 和 D' （即 D 和 D' 只有一个数据样本不同）。如果算法 A 满足

$$\Pr \{ A(D) \in O_A \} \leq e^\epsilon \cdot \Pr \{ A(D') \in O_A \} \quad (4)$$

则称算法 A 提供 ϵ 差分隐私保护。其中 ϵ 是隐私预算，用于控制差分隐私的保护程度^[57]。

在IDFL中应用差分隐私，不仅能够对敏感信息实施随机扰动，还可在激励机制中量化每个参与者在训练过程中的隐私损失^[58]，以此作为对参与者进行补偿的依据之一。将差分隐私和传统激励机制相结合，是应对IDFL中隐私威胁的主要方法。

3.1.1 博弈论方法

博弈论是研究多个决策者之间相互作用的数学模型和分析方法^[59]。在FL中，将差分隐私与博弈论结合，可在保护隐私的同时，实现FL各参与者激励分配的公平性。根据博弈模型的不同，此类方法主要分为基于Stackelberg博弈和基于非合作博弈两类。

Hu等^[60]将差分隐私与Stackelberg博弈相结合，提出了图6所示的FL数据交易框架。



图6 FL数据交易框架

Figure 6 FL data trading framework

FL数据交易框架的工作流程如下：在隐私预算分配阶段，用户 i 通过差分隐私的变体——零集差分隐私（zero-concentrated differential privacy, ZCDP）^[61]，计算本地学习任务的隐私预算 ϵ_i ， ϵ_i 越大意味着隐私损失越大。在收集到所有用户的隐私预算后，服务器根据总奖励 $R > 0$ 、隐私补偿函数和用户的隐私预算按比例分配支付金额 P_i 。

$$P_i = \frac{\epsilon_i}{\sum_{i \in N} \epsilon_i} R \quad (5)$$

此框架采用两阶段Stackelberg博弈设计激励

机制：第一阶段为领导者制定策略；第二阶段跟随者根据领导者的策略做出最优反应，以最大化自身效用。通过该机制，每个用户可以合理制定最优隐私预算，在保护隐私的同时最大化自身效用。类似地，Liu 等^[62]和 Huang 等^[63]将服务器和客户端之间的交互过程建模为一个两阶段的 Stackelberg 博弈，同时引入差分隐私技术，在保护隐私的前提下实现了博弈双方效用的最大化。

考虑到客户端的隐私需求存在差异，Li 等^[64]通过 Stackelberg 博弈分析服务器与客户端的行为，设计了兼具个性化隐私保护与激励机制的方案。该方案中，差分隐私的应用形式有基于批次（BDP）、基于迭代（EDP）和基于客户端（CDP）3 种。具体而言，BDP 在每个批次结束时为本地模型梯度添加噪声，提供较高强度的隐私保护；EDP 在每轮训练迭代结束时添加噪声，隐私保护强度较低；CDP 则在本地训练结束后添加噪声，提供客户端级的隐私保护。客户端可根据自身隐私保护需求选择合适的隐私保护策略。在 Yuan 等^[65]提出的 IDFL 方案中，将服务器与客户端之间的交互建模为 Stackelberg 博弈：服务器通过优化奖励策略，激励客户端贡献较高隐私预算的数据；客户端则根据提供的奖励调整隐私预算，并通过差分隐私向局部参数注入噪声，降低隐私泄露风险。

Liu 等^[66]提出了一个基于三阶段 Stackelberg 博弈的激励机制，并将其应用于联邦云一边学习场景。在图 7 所示的联邦云一边学习系统中，云中心在第一阶段发布激励机制并预测边缘设备的行为；边缘服务器在第二阶段根据激励机制和对边缘设备隐私偏好的估计，设计包含隐私预算和奖励金额的契约，以满足不同隐私需求；在第三阶段，边缘设备根据接收到的契约和自身的隐私偏好，选择最优的隐私预算参与学习任务，并通过差分隐私在本地模型更新梯度中添加噪声，防止攻击者通过观察更新过程推断原始数据集的特征分布。该机制的特点是，即使边缘服务器无法准确获知边缘设备的隐私预算，也可以通过合理设计的契约激励边缘设备选择合适的隐私预算，从而有效解决信息不对称问题。

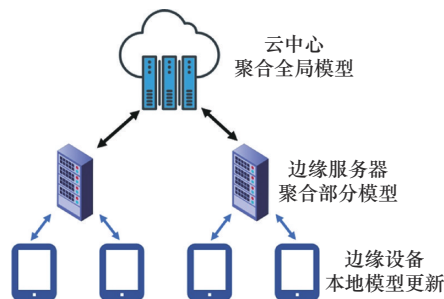


图 7 联邦云一边学习系统

Figure 7 Federated cloud-edge learning system

在非合作博弈方面，Pan 等^[67]假定所有参与者以最大化自身效用为目标，非合作地进行奖励竞争，并设计了相应的激励机制。该机制根据参与者上传的模型价值分配奖励，贡献更多训练数据并添加更少噪声的参与者将获得更多奖励，从而保证奖励分配的公平性。在计算模型价值时，差分隐私根据每个参与者的个性化隐私预算对客户端的模型参数进行干扰与保护。Mao 等^[68]提出了一个多阶段隐私保护博弈模型，在每个训练轮次前，客户端需要权衡隐私损失与模型性能，非合作地选择适当的差分隐私噪声，以最小化自身成本。Zhang 等^[69]设计了两种基于非合作博弈的激励机制，有效地激励参与者真实报告自己的成本。该机制通过联合差分隐私限制单个客户端对聚合结果的影响，为每个客户端提供严格的隐私保护。为了在模型训练准确性、隐私保护和能量消耗之间实现平衡，Yin 等^[70]将 FL 训练过程建模为非合作博弈，分析了各参与者之间的策略选择及收益函数，并通过求解纳什均衡确定最优隐私预算和策略，以帮助用户在保障隐私和提高模型准确性的同时降低能耗。

3.1.2 契约理论

将差分隐私与契约理论相结合，能够在保护隐私的前提下促进契约执行，有效解决信息不对称和道德风险等问题。

Sun 等^[71]提出了一种基于契约的个性化隐私保护激励机制。在该机制中，服务器首先根据给定的隐私预算设计一组个性化契约，并使用零集差分隐私 ρ -zCDP 为不同参与者定制个性化隐私保护级别，并给予相应的隐私成本补偿。为防止攻击者利用共享梯度实施隐私攻击，每个参与者需在传输本地梯度前采用高斯机制对梯度进行混

淆，如式(6)所示。

$$\tilde{g}(\chi_m; \theta) = g(\chi_m; \theta) + b_m \quad (6)$$

其中， $g(\chi_m; \theta)$ 表示参与者 m 使用其本地数据集中的随机小批量样本 χ_m 计算得到的随机梯度； θ 为模型参数； b_m 为注入的高斯噪声，其大小根据所签署合同中指定的隐私保护级别来确定； $\tilde{g}(\chi_m; \theta)$ 表示扰动后的随机梯度。

Wu等^[72]在差分隐私的基础上提出了一种基于多维契约的IDFL机制。该机制考虑了数据所有者的贡献，以及由计算、通信及隐私成本组成的综合成本，并根据数据所有者的成本类型提供相应的奖励，确保每个数据所有者只有在选择与其相匹配的契约时才能获得最大的利润，解决了模型拥有者和数据所有者之间的信息不对称问题。

考虑到FL中客户端的资源异构性及差异化隐私保护需求，Lin等^[73]提出了一种基于异构差分隐私和契约理论的IDFL框架HDP-FL。如图8所示，在该框架中，客户端可以根据市场运营商（market operator, MO）提供的隐私预算契约（ ϵ_i, R_i ）执行本地训练，并在模型中注入差分隐私噪声，将处理后的模型上传给MO。MO从上传的客户端模型中筛选部分进行全局聚合，将生成的全局模型提供给服务请求方（如企业或政府），服务请求方通过与MO的交互支付服务费，获取FL模型的使用权。HDP-FL根据客户的隐私偏好设计了不同的隐私预算分配策略，从而实现了个性化的隐私保护。

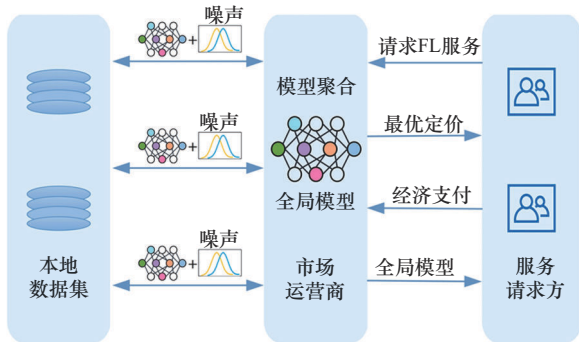


图8 HDP-FL框架

Figure 8 HDP-FL framework

3.1.3 拍卖理论

在拍卖理论中，拍卖主体需要收集参与者的

出价或竞标信息，然后根据特定规则确定资源和奖励的分配^[26]。在IDFL中，将差分隐私与拍卖理论相结合，可在保障拍卖效率与精确性的同时，实现资源的高效配置，提升参与者对激励机制的信任度与参与度。现有研究分为基于正向拍卖和基于反向拍卖两类。

在正向拍卖中，拍卖主体通常是资源的提供者，而参与者则是资源的竞标者。Zheng等^[74]利用拍卖理论和本地差分隐私技术设计了一个能够同时提供激励和隐私保护的IDFL框架。差分隐私被用于对数据所有者上传的本地梯度信息进行扰动。拍卖理论则用于根据数据所有者的出价和系统预算，确定隐私损失程度和支付金额。在该框架中，数据所有者的出价信息被假定是公开的。

Ying等^[75]提出的IDFL机制利用差分隐私的指数机制保护参与者的本地模型隐私和出价隐私。在该机制中，参与者在提交出价后，平台基于指数机制随机选择获胜者，使得攻击者即使知晓选择结果，也难以推断出参与者的真实成本。然而，该机制要求参与者将出价上传至服务器，导致参与者信息对服务器透明，存在潜在的隐私泄露风险。针对这一问题，Wang等^[76]设计了差分隐私出价混淆函数，通过指数机制在本地混淆参与者的出价，从而解决了不可信服务器环境下的投标信息保护问题。此外，Zhu等^[77]提出了一种面向物联网系统的隐私感知双重拍卖机制，利用差分隐私来保护出价信息，同时考虑到频谱资源估值随时间变化的特性，设计了一种基于时间估值的拍卖机制，从而能够更好地反映竞拍者的真实需求和资源价值。

上述方案通常假设服务器可以获得客户端的真实数据质量信息（如数据分布信息、成本信息等），具有一定的局限性。为此，Ren等^[78]引入本地差分隐私框架，通过向客户端的真实数据质量信息（如标签分布信息）注入随机噪声来保护客户端的数据分布信息。服务器可根据加噪后的结果，使用贝叶斯方法计算出每个客户端数据质量的估计值，并选取数据分布与全局数据分布更接近的客户端参与FL。

与正向拍卖不同，反向拍卖的主体是资源需求者，参与者为资源提供者。Li等^[79]设计了一

种基于反向拍卖的激励机制。在该机制中,参与者提交报价,数据请求者根据报价决定是否接纳参与者,旨在激励低报价、高质量数据的用户参与。该机制采用差分隐私来保护参与者的数据,并使用贝叶斯方法评估参与者的数据质量,为表现良好的参与者提供额外奖励。

为防范参与者通过不诚实报告差分隐私噪声量以获取更高奖励的行为,Yang等^[80]验证了不诚实参与者在基于差分隐私的FL方案中产生的负面影响,提出一种基于反向拍卖的激励机制CSRA。该机制通过比较模型更新的方差与参与者声称的差分隐私噪声量,识别不诚实参与者并将其从本轮训练中剔除,同时降低其在后续学习中被选中的概率。

Wang等^[81]利用差分隐私对参与者的出价和本地模型实施双重扰动,实现了对出价和模型的双重差分隐私保护。同时,该研究还综合考虑了参与者的数据量、隐私保护水平、数据非独立同分布(Non-IID)程度等因素,设计了多维评分函数用来评估参与者的质量,并基于扰动后的出价信息和估计质量,采用反向拍卖机制选择获胜者,以最大化所有参与者和服务器的收益。

3.2 同态加密应用于IDFL

同态加密在FL中通常被用于对敏感信息(如用户上传的梯度信息)进行加密,以实现保密通信与安全计算^[82]。它允许在密文上直接进行计算,而无须事先进行解密^[83],如式(7)所示。

$$\text{Dec}(\text{sk}, \text{Enc}(\text{pk}, m_1) \diamond \text{Enc}(\text{pk}, m_2)) = m_1 \circ m_2 \quad (7)$$
其中, $\text{Dec}(\cdot)$ 表示解密运算, $\text{Enc}(\cdot)$ 表示加密运算, m_1 和 m_2 表示明文, sk 表示私钥, pk 表示公钥, $\circ$ 表示明文域上的运算, $\diamond$ 表示密文域上的运算。典型的乘法同态算法有RSA和ElGamal等^[84], Paillier^[85]是最常用的加法同态算法。

Weng等^[86]提出了一种融合智能合约与同态加密的激励机制,通过合约内的超时检查和货币惩罚机制约束不诚实参与者的行为。该机制采用Threshold Paillier算法对参与者的本地梯度信息进行加密。它将私钥分割成多个部分,由不同的参与者持有,可有效防止参与者之间的串谋。

周炜等^[87]通过引入声誉值评估参与者的可靠性,结合参与者的数据贡献、计算贡献和声誉

表现给予相应奖励,以此保障协作学习的公平性与效率;同时,采用区块链和Paillier算法保护中间参数的机密性。Jia等^[88]提出了一种基于Paillier同态加密和区块链技术的多任务联邦学习框架,该框架支持客户端自主选择参与多个学习任务,显著提升了计算资源利用率。通过引入Paillier加密算法保障数据隐私安全,并借助区块链的智能合约实现自动化激励机制,有效激发了参与者的协作积极性,在确保隐私安全的同时,优化了模型训练效率与全局模型性能。

为防止全局模型被窃听,Zhao等^[89]基于Threshold Paillier算法提出了一种安全聚合机制SecAgg。任务发布者将其私钥分割为两个私钥片 sk_1 和 sk_2 ,前者发送给服务器,后者发送给各参与者。各参与者使用任务发布者提供的公钥 pk 对本地模型参数进行加密后将其提交至服务器,服务器完成聚合得到加密的全局模型;随后,服务器利用 sk_1 对加密的全局模型进行部分解密,并将结果发送给所有参与者,参与者再使用 sk_2 对其进行完全解密,然后开展新一轮的本地训练。此机制确保本地模型与全局模型在传输过程中始终以密文形式存在,有力保障了模型的机密性。此外,该框架融合出价机制与定价机制,设计了一种混合激励机制HyInc。借助逆向Vickrey拍卖机制,参与者仅能提交自身数据量作为出价,且无法知晓其他参与者的出价情况,从而避免参与者通过虚假出价获取不公平高回报的现象。

除主流的Paillier算法外,部分IDFL方案也尝试采用其他同态加密算法保护参与者的隐私。Chen等^[90]采用RSA同态加密算法对参与者提交的多维信息(如数据质量、计算性能指标)及激励机制中生成的参与者得分信息等进行加密处理,确保敏感信息在传输和计算过程中的机密性。He等^[91]在基于博弈论的区块链激励机制中引入了全同态加密算法CKKS^[92]。该机制采用分布式密钥管理方案,防止密钥泄露并实现对模型参数的协作解密。Xiong等^[93]提出了一种基于矩阵的同态加密方案,利用加密矩阵和噪声矩阵对客户端模型参数进行加密,从而防止数据泄露;同时,结合智能合约,根据模型评估结果计算每个客户端的贡献度并完成奖励分配,确保激励的公平性。

3.3 分析与总结

从隐私保护角度来看，差分隐私具备数据保护强度可灵活调节的特性，兼具灵活性与高效性优势；同态加密则可提供数据机密性保护，然而往往伴随较高的计算与通信成本。将差分隐私和同态加密应用到IDFL中，实现方式各有所异，表2对这些方法的优缺点进行了归纳和比较。

差分隐私与博弈论的结合，赋予了IDFL更强的灵活性和可扩展性，有助于协调多主体决策并保障机制的公平性。然而，服务器与客户端之间存在的信息不对称问题，会对激励机制的有效性和可持续性产生影响；同时，博弈论方法通常伴随较高的计算代价，尤其在实时场景中，如何提升激励机制的计算效率成为该方法亟待解决的问题。

契约理论能够激励客户如实反映其真实成本与隐私保护需求，在一定程度上缓解服务器与客户端之间的信息不对称。当契约理论与差分隐私相结合时，系统可根据不同客户端的隐私偏好，为其分配个性化的隐私预算并提供相应补偿，从而在隐私保护强度与模型学习效率之间实现动态平衡。然而，该模式也存在局限性：数据所有者需严格遵循契约条款行事，这在一定程度上限制了其自主决定数据规模与敏感资源投入的灵活性，导致其难以实现本地利益的最大化。

拍卖理论与差分隐私相结合，使IDFL能够依据隐私预算动态筛选参与学习的客户，并根据参与者的出价灵活分配资源与奖励，从而有效控制成本。同时，差分隐私通过保护参与者的出价信息，可防止隐私泄露。然而，拍卖理论易受合谋攻击，即参与者可能达成协议降低出价或操纵

拍卖结果，进而破坏激励机制的有效性。同态加密技术的应用能够确保数据在整个计算和通信过程中的机密性与完整性，提供比差分隐私更强的安全保护。但基于同态加密的FL通常涉及大量双线性点对点复杂运算，显著增加了计算开销。此外，在迭代训练中，过大的通信代价可能导致本地模型在聚合阶段出现时延和参数过期，影响全局模型的训练效率及客户的参与体验。

4 面向安全防御的IDFL

本节从防范主动攻击的角度，对面向安全防御的IDFL进行梳理和分析，包括用于抵御投毒攻击的声誉机制、异常检测、共识算法、区块链智能合约等方法，以及用于抵御搭便车攻击的贡献评估和攻击检测方法。

4.1 防御投毒攻击

防御投毒攻击的关键在于及时识别攻击者并约束其行为。在IDFL中，常用的防御技术包括声誉机制、异常检测、共识算法、智能合约，以及这些技术的组合。例如，通过建立声誉机制实施异常检测^[94]，系统可综合参与者的历史表现与实时行为分析来识别异常个体；将共识算法与智能合约结合^[95]，则有助于确保参与者行为的一致性并规范其操作。这些技术的组合应用，使IDFL能够更有效地应对复杂的应用环境。

4.1.1 声誉机制

在IDFL中，声誉机制用于评估参与者的贡献，从而有效区分表现优异的参与者与潜在的攻击者。

Xu等^[96]通过计算参与者上传的本地梯度和全局梯度之间的余弦相似度来评估参与者贡献，并依此动态更新参与者的声誉值。在模型聚合过

表2 IDFL中隐私保护方法对比

Table 2 Comparison of privacy protection methods in IDFL

文献	主要方法	优点	缺点
[60, 62-70]	差分隐私+博弈论	隐私保护与策略优化兼顾	计算复杂；噪声影响博弈策略效果
[71-73]	差分隐私+契约理论	明确隐私保护责任与激励条款，提高了参与者的信任度；适用于长期合作场景	契约设计复杂，需平衡隐私保护与激励效果
[74-81]	差分隐私+拍卖理论	激励参与者提供真实隐私预算，提高了系统的公平性和效率	容易受到合谋攻击
[86-89]	Paillier加密算法	无须密钥共享，支持部分解密	加解密速度较慢，影响模型的性能
[90-91, 93]	CKKS等加密算法	支持浮点数计算，效率更高	密钥管理复杂，计算成本高

程中, 服务器根据参与者的声誉权重进行加权聚合, 从而增强系统对投毒攻击的防御能力, 确保只有贡献良好的参与者才能从全局模型的更新中受益。Ouyang 等^[97]首先计算参与者本地模型与全局模型间的相似度, 据此得到参与者的相对声誉值; 随后, 利用区块链技术提取参与者的历史声誉值, 并将其与相对声誉值融合形成参与者的最终声誉值, 据此确定模型聚合中各参与者的权重分配, 有效降低恶意参与者对全局模型的不良影响。Zhao 等^[98]通过计算参与者本地模型参数的欧氏距离, 选择距离中心最近的参数子集用于全局聚合, 动态调整参与者的声誉得分 (被选中的参与者声誉值增加, 未入选者声誉值减少), 以此筛选恶意行为。所有声誉评分变动均通过区块链进行透明记录, 确保数据不可篡改。同时, 结合差分隐私技术, 在特征提取阶段添加可控噪声, 进一步防范潜在的推理攻击。

考虑到单纯依赖声誉作为激励因素存在局限性, Wang 等^[99]提出一种双重激励机制, 该机制结合 SV 计算的参与者贡献值与声誉值, 对参与者表现进行全面评估。在该机制中, 参与者声誉主要来源于注册时的初始化声誉值和每轮训练中的得分。当参与者声誉评分低于设定阈值时, 将被禁止参与后续学习。Yang 等^[100]提出的 IDFL 框架依据参与者的贡献和声誉为其提供奖励。其中, 训练者的贡献由委员会根据其上传的模型质量进行评估, 与提交的模型质量呈正相关关系, 委员会成员的贡献则与其有效判断数量成正比。此外, 该框架采用加权平均指数法综合考虑参与者的模型质量和历史表现, 动态评估其声誉值, 以此有效避免声誉垄断, 激励高质量参与者积极参与。Feng 等^[101]设计了一个由本地模型更新链和全局模型更新链组成的 IDFL 框架, 前者旨在有效地跟踪所有本地模型更新并建立声誉系统, 后者则专注于训练和更新全局模型, 确保学习过程的稳定性。在此基础上, 研究人员将参与者的奖励分为来自 FL 的奖励和来自区块链的奖励。其中, FL 的奖励综合考虑参与者的数据规模、数据质量、任务满意度和历史声誉等因素, 区块链的奖励则主要考虑参与者的计算能力和通信成本。

然而, 上述方案的声誉评分主要来自单一的

评分者, 容易受到恶意评分者的影响。针对这一问题, Kang 等^[102]另辟蹊径, 引入参与者的直接声誉和推荐声誉, 并借助多权重主观逻辑模型^[103]计算参与者的声誉。直接声誉源自交互历史, 积极互动能够提升候选人的声誉; 推荐声誉则由其他任务发布者 (即推荐者) 提供, 推荐者依据与候选参与者之间的交互经验, 赋予相应的推荐声誉。任务发布者不仅可以挑选高声誉候选者参与 FL 模型训练, 还可以将他们对参与者的声誉评价上传至声誉区块链, 作为推荐声誉。该声誉计算方案采用双重声誉维度和区块链技术, 既能精准识别恶意候选人, 又能有效防范任务发布者之间的串谋行为。

为解决任务发布者难以知晓本地数据质量和参与者的准确声誉值等信息不对称问题, Zhang 等^[104]提出了基于区块链的去中心化 IDFL 框架 BDFL。该框架去除了用于模型聚合的服务器, 但引入了一个用于模型验证的审核委员会, 具体架构如图 9 所示。在 BDFL 中, 客户端以 P2P 方式与邻居交换模型更新, 并在本地完成聚合操作。审核委员会由一组可访问区块链的节点 (审核员) 组成, 其职责是对模型更新进行验证, 并据此构建每个客户端的声誉体系。这使得高声誉客户端的模型更新更易被邻居接受。实验数据显示, 即便恶意客户端占比达 30%, BDFL 依然能够实现快速收敛, 且保持较高的准确率。同样秉持去中心化理念, Domingo-Ferrer 等^[94]提出一种声誉管理机制 Co-Utility, 该机制下每个参与者的声誉由随机选取的 m 个其他参与者进行管理, 从而防止恶意参与者破坏声誉系统。

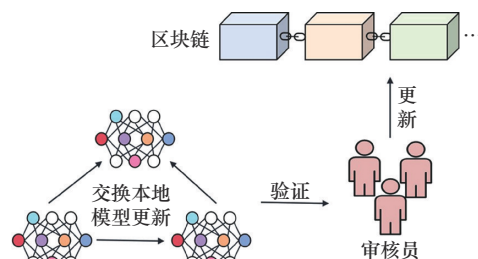


图 9 基于区块链的去中心化 FL 框架
Figure 9 Decentralized FL framework based on blockchain

同时, 方案结合数据所有者与任务发布者的历史交互记录, 以及其他任务发布者对数据所有

者的间接声誉意见，多维度评估数据所有者的声誉，以此激励其提供高质量的数据和模型参数。Rehman等^[106]提出一种针对移动边缘计算场景的IDFL框架，充分考虑边缘设备、雾节点和云服务器3个实体的声誉。在该框架中，各实体之间可相互进行声誉评分，声誉分数的哈希值最终提交给智能合约，经智能合约计算得出每个实体的声誉。通过这种方式，有效激励诚实参与者积极投入计算任务，同时抑制恶意参与者的不良行为。

表3对基于声誉机制的激励方法进行了总结。声誉机制在IDFL中的应用能够有效识别和过滤恶意模型更新，从而在一定程度上抵御投毒攻击，并激励参与者提供高质量的模型更新。但是，声誉机制的有效性一方面依赖验证数据集能否准确反映整体数据的特征和分布，另一方面依赖声誉信息的及时更新。

4.1.2 异常检测

在IDFL中，异常检测技术通常用于监测参

与者的模型更新，以识别异常或恶意的参与者，提前发现可能的投毒攻击。异常检测的依据主要有两种：基于本地更新的统计特征^[94, 107-112]和基于模型的性能指标^[102, 113-114]，即如果某个本地更新相对于其他更新在统计特征上是异常的，或者将某本地更新聚合到全局模型中会降低模型的某些性能指标，则该更新会被标记为不良更新。表4对这两种异常检测方法进行了对比分析。

Domingo-Ferrer等^[94]提出的IDFL方案采用基于距离的异常检测方法^[47]识别异常的模型更新。具体而言，对于给定的一批模型更新，首先计算其质心 C （如平均值或其他中心性度量），然后计算每个更新与质心的距离，将计算结果与预设阈值进行比较。若某个更新 U 与 C 之间的距离值大于阈值，则将该更新标记为不良更新，检测结果作为参与者的声誉评分与奖惩机制的依据。

Wu等^[107]提出的基于区块链的可信FL系统采用欧氏距离度量本地模型更新间的差异。如图10所示，系统中每个参与者兼具训练者和验

表3 基于声誉机制防御投毒攻击的IDFL总结

Table 3 Summary of IDFL defending against poisoning attacks based on reputation mechanism

文献	描述	奖励方式	主要技术或原理
[96]	根据参与者上传梯度的相似性更新声誉	根据参与者的声誉	余弦相似性
[97]	历史声誉用于逆向拍卖	根据参与者的出价和贡献	拍卖理论
[98]	通过欧氏距离筛选模型参数并动态调整声誉	根据参与者的声誉	区块链和差分隐私
[99]	声誉来源于初始化声誉值和每轮训练中的得分	根据参与者的SV和声誉	SV和智能合约
[100]	声誉来源于参与者的模型质量和历史表现	根据参与者的贡献和声誉	区块链
[101]	声誉的来源主要有FL任务表现和区块链奖励	来自FL和区块链两部分	区块链
[102]	声誉来自直接声誉和推荐声誉	根据参与者的资源和声誉	契约理论
[104]	审核员根据验证结果动态调整参与者的声誉	根据声誉和模型验证的准确性	委员会机制
[94]	对声誉机制进行去中心化的管理	根据参与者的声誉	协同效用
[105]	考虑了参与者和任务发布者两方的声誉	根据参与者的声誉	共识算法
[106]	考虑了边缘设备、雾节点和云服务器三方的声誉	根据参与者的声誉	智能合约

表4 基于异常检测防御投毒攻击的IDFL对比

Table 4 Comparison of IDFL defending against poisoning attacks based on anomaly detection

文献	检测方法	统计量/模型指标	优缺点
[94, 107-112]	基于更新的统计特征	相似度、欧氏距离等	能够快速识别异常行为，对动态变化环境的适应性强；但无法准确判断模型质量，且对于复杂的攻击模式可能失效
[102, 113-114]	基于模型的性能指标	模型准确率、模型损失等	能够更加准确地评估模型质量，且具有较高的检测能力；但需要较多的计算资源，无法及时响应动态变化的环境

证者双重角色。每轮训练后，参与者计算自身模型与其他参与者模型间的欧氏距离，若该值超过预设阈值，则认为该模型可能被攻击或篡改，并向系统报告。当某个模型被半数以上参与者报告为异常时，智能合约自动判定其为恶意模型，并取消该模型的聚合资格。对于未被判定为恶意的模型，参与者通过本地数据测试其准确率，并将结果上传至区块链，作为声誉评估的重要依据。

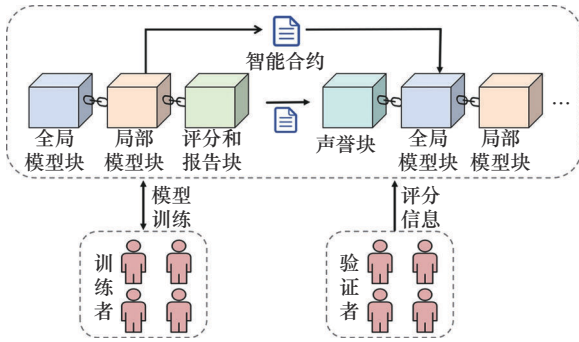


图 10 基于区块链的可信 FL 系统
Figure 10 Blockchain based trusted FL system

Zhao 等^[108]利用每个客户端与其余所有客户端模型更新的欧氏距离之和来确定聚合权重，距离和越大，则该客户端的聚合权重越小，从而降低异常客户端的影响。Zeng 等^[109]采用余弦相似度计算本地模型与上一轮聚合模型之间的距离（相似度低于阈值则判定为异常），并基于相似度生成置信度，将此作为本轮聚合模型的权重和分配奖励的依据。

Li 等^[110]提出了一种融合横向评估与纵向评估的双维度检测机制，用于识别 FL 中的恶意参与者。横向评估是指服务器在每轮聚合阶段计算参与者提交的本地模型之间的相似度，以识别异常模型。纵向评估是指服务器使用 Transformer 模型^[115]对每个参与者之前轮次提交的本地模型进行时间序列预测，并将预测结果与当前轮次提交的模型进行比较，若两者差异过大，则表明参与者修改了他们的模型。最后，机制通过加权平均综合横向与纵向评估结果，生成各参与者的当前轮次置信度评分。若置信度低于阈值，则认为该模型存在异常，并降低其聚合权重和声誉值。

Chen 等^[111]提出了一种基于随机欧氏距离投票的检测方法，两个模型之间的距离被定义为

$$O_{ij} = \|m_i^2 - m_j^2 - 2m_i m_j^T\| \quad (8)$$

其中， m_i 和 m_j 分别表示参与者 i 和 j 上传的模型参数向量， $\|\cdot\|$ 表示范数。服务器随机检测本地模型之间的欧氏距离，并将其平均值作为阈值。如果 m_i 和 m_j 之间的距离大于阈值，则 j 将 i 判定为恶意参与者。服务器汇总所有对 i 的投票结果，若超过 1/3 的用户将 i 判定为恶意参与者，则服务器最终判定 i 为恶意参与者，其奖励被其他诚实用户均分。He 等^[112]基于客户端之间模型参数的欧氏距离进行筛选，仅选择距离最近的一定数量客户端参与当前轮的聚合，而未入选的客户端被认为可能存在恶意行为，其恶意行为计数加 1。一旦用户的恶意行为计数超过预设的风险阈值，该用户将被永久移出后续训练过程。此外，恶意行为计数也被用来计算客户端的风险系数，以更新其声誉值和分配奖励。

在 FL 场景中，模型精确度、模型损失和模型准确率等性能指标，为异常检测提供了重要的量化依据。Gao 等^[113]基于性能指标构建了一套高效的异常检测机制，通过对参与者提交的本地更新进行深度分析，实现对潜在恶意更新的精准识别与过滤。具体而言，模型损失差被定义为

$$S(\theta, G_i) = L_i(\theta) - L_i(\theta - G_i) \quad (9)$$

其中， G_i 表示参与者 i 的本地更新， $L_i(\theta)$ 表示当前的全局模型测试损失， $L_i(\theta - G_i)$ 表示不包含 i 时的测试损失，损失差 $S(\theta, G_i)$ 反映了 i 的更新对全局模型性能的改善程度。若 $S(\theta, G_i)$ 为负，说明 i 的更新会降低模型性能，该参与者则被认定为恶意参与者。

Xu 等^[114]提出了一种基于区块链的 IDF 框架 BESIFL，该框架利用模型准确率检测并移除恶意参与者。在每轮模型聚合环节，BESIFL 系统会从所有参与者中随机选取部分参与者作为验证者，并向其分发测试数据集。验证者会将每个参与者的本地模型准确率与上轮聚合模型准确率进行对比分析，若发现某个参与者的本地模型准确率出现显著下降，且下降幅度超过预设阈值，BESIFL 系统便会判定该本地模型对应的参与者

为恶意参与者，并将其从后续模型聚合流程中移除。

Kang 等^[102]提出融合 RONI 和 FoolsGold^[116]的双技术检测方案，实现对本地模型更新的异常检测。其中，RONI 技术针对独立同分布数据场景的投毒攻击检测，通过比较分析单个本地模型更新被聚合前后全局模型的性能变化，判定该更新是否可靠。FoolsGold 技术则面向非独立同分布数据场景，通过比较参与者上传的本地梯度向量与全局平均梯度的相似度实施异常检测。针对检测出的不可靠参与者，主要的惩罚包括降低声誉评分、扣留奖励、模型更新被移除等。

将异常检测机制引入 IDFL 系统，能够通过量化指标有效评估参与者提交的本地模型质量，实时识别潜在投毒者并实施分级惩罚措施，必要时将其从系统中移除。然而，当前异常检测方法在判定指标选择与阈值设定上仍依赖人工经验，存在一定主观性，可能会导致误报或漏报。

4.1.3 共识算法

共识算法的主要功能是确保分布式网络中各参与者在交易历史和决策上达成统一共识^[117]。常用的共识算法包括委员会共识机制（committee consensus mechanism, CCM）^[118]、工作量证明（proof of work, PoW）^[119]、权益证明（proof of stake, PoS）^[120]、拜占庭容错（Byzantine fault tolerance, BFT）^[121]和模型解释选择证明（proof of interpretation and selection, PoIS）^[95]等。在 IDFL 中，共识算法通过约束参与者行为等，有效抵御投毒攻击。表 5 对基于共识算法抵御投毒攻击的 IDFL 方案进行了总结和比较。

Li 等^[122]在其设计的 IDFL 中引入了 CCM 算

法实现异常检测。该机制通过智能合约驱动的动态委员会选举委员会成员。在每轮训练中，委员会成员将各自持有的本地数据作为验证集，对其他参与者提交的更新梯度进行测试并评分，只有评分达标的更新梯度才会被记录至区块链。得益于 CCM 算法的设计，系统将共识过程中的计算与通信任务集中由委员会成员完成，大幅减少了全网节点参与共识的资源消耗，显著提升了训练效率。此外，委员会依据参与者的贡献分数进行奖励分配，既确保了激励体系的公平透明，又能够有效过滤低质量或恶意更新，进而提升系统的安全性和可靠性。

为应对 FL 中的投毒攻击威胁，研究者基于不同共识机制设计了多样化防御方案。Qu 等^[123]利用 PoW 共识算法设计了一个 IDFL 框架。PoW 共识算法要求参与者通过解决计算难题来证明其工作成果。这意味着攻击者需要控制大量的计算资源才能发起投毒攻击，从而提高了投毒攻击的成本。Xu 等^[105]提出了一种适合资源约束的 FL 共识算法，即贡献证明（proof of contribution, PoC）算法。PoC 算法以参与者的贡献（如在线时间、本地模型质量及数据贡献）作为共识的关键依据，确保参与者只有提供实际贡献才能获得奖励。

Boualouache 等^[124]基于 BFT 共识算法，设计了基于准确率的委托拜占庭容错共识算法（A-DBFT）用于恶意攻击检测。该算法首先筛选出模型准确率较高的参与者作为共识成员，以此保障共识过程的可靠性和模型精度；随后通过比较本地模型准确率与全局模型准确率的差值，判断是否存在恶意攻击行为，若差值超出动态阈值，相应的本地模型将被判定为可疑并予以剔除。

表 5 基于共识算法防御投毒攻击的 IDFL 方案对比

Table 5 Comparison of IDFL defending against poisoning attacks based on consensus algorithms

文献	采用的共识算法	优点	缺点
[122]	委员会共识机制	决策效率高、支持更复杂的决策规则	存在委员会成员作恶的风险、可能导致中心化
[123, 105]	工作量证明/贡献证明	可抵抗中心化、防止无效劳动	需要验证模型质量、存在参数隐私泄露风险
[101, 124]	拜占庭容错	允许节点之间存在一定的故障和作恶行为	效率低、需要大量的通信资源
[95, 125-126]	模型解释选择证明/模型性能的共识	支持更加公平的贡献评价、激励机制更加透明	计算复杂度较高、对数据异构的适应性差

Feng 等^[101]提出的 IDFL 框架采用了 BFT 算法。该算法具备容忍恶意参与者的能力。只要恶意参与者数量不超过总参与者的 1/3, 系统就能达成有效共识。这意味着即便攻击者控制部分参与者发动投毒攻击, 系统仍可维持正常运行。Boualouache 等^[124]基于 BFT 共识算法, 设计了基于准确率的委托拜占庭容错共识算法 (A-DBFT) 用于恶意攻击检测。该算法首先筛选出模型准确率较高的参与者作为共识成员, 以此保障共识过程的可靠性和模型精度; 随后比较本地模型准确率与全局模型准确率, 根据其差值判断是否存在恶意攻击行为, 若差值超出动态阈值, 相应的本地模型将被判定为可疑并予以移除。

Kasyap 等^[95]为 IDFL 框架开发了 PoIS 共识算法。PoIS 利用 SV 计算各特征对类别标签的影响程度, 生成特征归因矩阵, 并对不同参与者同一类别标签的特征归因矩阵进行标准化处理与对比。若某参与者的特征归因矩阵与其他参与者存在显著差异, 则认定该参与者存在恶意行为, 其模型更新将被拒绝参与聚合。此外, PoIS 还依据参与者的贡献度和声誉值设计奖励函数, 实现公平的激励分配。米波等^[125]融合区块链与模型水印技术, 提出基于模型性能的去中心化共识机制 PFconsensus。各参与者在本地训练后, 将嵌入水印的梯度模型广播至网络, 随后基于自身数据对其他参与者的模型性能进行评估, 通过投票选出性能最优的一批本地模型进行聚合, 并将结果写入区块链, 以此确保共识过程的可追溯性。同时, 该机制根据模型性能与参与者贡献度实施奖励分配, 激励参与者共同提升模型质量。

Ying 等^[126]提出基于循环的分片共识算法。该算法将验证节点划分为多个并行工作的分片组, 每个分片独立使用公开数据集评估模型性能。当首个分片无法达成共识时, 系统会按序激活后续分片进行补充验证。这种循环验证机制规定, 只要累计超过 2/3 的验证者达成一致, 或所有分片均完成验证, 即可终止验证流程, 在保证与非分片共识协议同等安全性的前提下, 显著提高了验证效率。

综上所述, 共识算法在去中心化 FL 框架中

发挥着关键作用, 能够有效促进参与者在模型更新与聚合上达成一致, 降低恶意参与者注入有害数据的风险, 同时通过去中心化结构消除中央聚合端带来的单点故障隐患。然而, 在设计共识算法时, 需谨慎权衡系统效率与安全性。过于复杂的共识机制可能导致计算成本飙升, 而过度简化的共识机制易引发安全漏洞。此外, 由于这类算法对参与者的协作程度要求较高, 少数参与者的不稳定或失效, 都可能对整个系统的共识过程产生显著负面影响。

4.1.4 智能合约

智能合约是一种被多方认可、运行在区块链上、能够根据预设条件自动处理算法的程序^[127]。它能够在没有可信第三方的条件下实现可信交易, 且交易具有可追溯性和不可逆性。将智能合约引入 IDFL (去中心化联邦学习框架), 能够借助其透明化、自动化的规则体系, 为系统构建更可靠的安全与激励机制, 确保公正奖励或惩罚, 实现模型更新的验证, 有效抵御投毒攻击。

Wang 等^[99]和 Feng 等^[101]借助智能合约对收入分配及支付条款实施自动化管理, 以此保障 IDFL 的透明性与运行效率。当数据需求方为训练模型进行资金支付时, 智能合约会依据预先设定的分配方案自动完成资金划拨, 确保分配过程兼具公平性与不可篡改性。这一自动化特性省去了中间环节, 进而降低了交易成本。此外, 智能合约能够对参与者的贡献及行为进行记录, 自动更新参与者的声誉信息, 并据此对诚实且有价值的参与者给予奖励; 同时, 针对恶意参与者, 智能合约会自动执行惩罚措施, 如扣除奖励、实施暂时或永久的参与禁令等^[104]。

Wang^[128]提出了一种基于区块链的 FL 系统, 通过基于检查点 (即模型参数) 的智能合约记录并追踪各参与者的行为。智能合约会定期 (如每完成一轮训练) 存储模型检查点, 以便后续需要时恢复模型。一旦检测到模型因遭受攻击而出现性能下降, 智能合约将立即调用最近存储的检查点, 使模型恢复至被攻击前的状态, 从而保障模型的完整性。Jaberzadeh 等^[129]提出了一个基于智能合约抵押金机制的 IDFL (交互式与分布式联

邦学习) 框架。在该框架中, 参与者需预先缴纳保证金; 若检测到某参与者提交的本地模型导致测试模型性能下降, 其抵押存款将被智能合约自动没收, 以此惩罚恶意行为。Xu 等^[130]提出的 IDFL 方案中, 数据请求者首先将预设的模型保证金与奖励机制上传至智能合约。智能合约根据数据提供者的竞价和保证金支付情况筛选参与者——数据提供者需支付保证金方可访问模型参数并提交梯度更新。随后, 智能合约会评估参与者提交的梯度质量, 并基于评估结果分配奖励和退还保证金, 以此激励数据提供者贡献高质量的模型更新, 提升系统的可信度与公平性。

智能合约具备自动执行合约条款的特性, 无须第三方干预。在 IDFL 中应用智能合约, 能够有效提升学习效率, 增强系统透明度, 同时提高抵御投毒攻击的能力。不过, 智能合约存在一个显著特点, 即一旦完成部署, 便难以进行调整。此时, 任何设计层面的漏洞都可能引发一系列严重问题, 如数据丢失、信誉损害等。鉴于此, 在设计智能合约时, 必须开展充分的测试与审核工作, 以此确保合约逻辑准确无误。

4.2 防御搭便车攻击

搭便车攻击是 IDFL 面临的常见威胁之一。其典型表现为: 攻击者参与训练过程, 却未有实质性贡献, 反而免费获取其他客户端的劳动成果。这种行为不仅会破坏激励机制的公平性, 还会降低诚实参与者的积极性, 进而影响 FL 的效率与可靠性。

目前, 抵御搭便车攻击的有效手段是通过公平的贡献评估机制对攻击者进行识别和惩罚, 其中常用的方法是利用 SV 来度量参与者在 FL 中的贡献。Liu 等^[5]提出了一种改进的 SV 计算方法 GTG-Shapley^[131]。该方法利用梯度更新重建联邦学习模型, 避免了直接训练不同参与者组合模型的烦琐过程, 显著降低了计算贡献所需的成本。基于贡献评估结果, 系统会选择贡献较大的模型进行聚合并给予相应奖励, 从而有效将搭便车者排除在外。类似地, Shi 等^[132]利用 SV 来衡量每个参与者在每轮学习中的贡献。他们将局部梯度和全局梯度的余弦相似性作为计算 SV 的效用函数, 使贡献评估能够在线性时间内完成, 然后根

据贡献高低筛选出贡献较高的本地模型进行聚合。Salarbashishahri 等^[133]提出的方法利用参与者的 SV 值来确定参与者模型的聚合权重, 实现加权聚合。对于 SV 值显著低于平均水平的客户端, 系统将禁止其在后续迭代中参与全局模型的更新, 以此避免搭便车攻击的影响。

除使用 SV 进行贡献评估外, Richardson 等^[134]提出使用影响力^[135]衡量 IDFL 中参与者的贡献。影响力被定义为评估对象对模型损失函数的影响程度。系统根据参与者的影响力大小进行奖励分配, 激励参与者根据预期的奖励调整自己的行为策略, 使诚实提交数据和模型成为其最佳选择, 从而抑制搭便车行为。Huang 等^[136]提出了 COFL 和 CAFL 两种博弈模型, 根据参与者提供的计算资源和训练数据批量大小来衡量参与者的贡献。其中, COFL 模型不对参与者强加贡献要求, 由参与者自行决定计算资源投入和数据提供量。但作者证明, 在无约束条件下该模型无法杜绝搭便车现象。于是, CAFL 模型对训练过程中的参与者贡献提出明确要求, 若参与者承诺的训练批量大小未达到最小贡献阈值, 则将被排除在 FL 之外。

Li 等^[137]提出了一种综合评估体系, 将参与者的数据规模、数据质量及计算能力综合定义为“训练能力”, 并划分为不同等级, 作为评估参与者贡献的核心指标。服务器根据参与者的训练能力等级设计契约内容, 使高等级参与者获得更多收益, 同时拒绝向低等级参与者支付奖励。

除贡献评估机制外, 异常检测同样是识别搭便车行为的有效手段。Wang 等^[138]提出了一种名为 FRAD 的搭便车攻击检测机制。该机制通过综合考量参与者的计算资源、通信成本和数据质量计算其贡献值, 再结合参与者的贡献度与声誉筛选可疑节点, 最后利用深度自编码高斯混合模型 (DAGMM)^[139]完成检测。实验结果显示, 该机制能够有效识别约 80% 的搭便车者, 即便在搭便车者占多数的复杂场景下, 依然保持着较高的检测性能。

Zhang 等^[140]提出了一种基于聚类的方法来检测搭便车者。服务器依据参与者的本地模型与训练损失, 运用 K-means 聚类算法将参与者划分

为多个集合,然后计算集合内参与者模型的相似度。若某个参与者的本地模型与其他参与者本地模型之间的相似度低于阈值,则将该参与者标记为搭便车者。与之类似,Chen等^[141]同样借助K-means算法,基于参与者本地模型参数进行分组,以组内参与者模型参数的余弦相似度为衡量标准,将低于预设阈值的参与者判定为搭便车者。Zheng等^[142]将任务竞争、长期合作和搭便车等因素纳入考量,将模型训练过程构建为动态重复博弈模型,通过推导纳什均衡获取最优策略,将每轮训练中参与者上传的模型更新对整体性能的提升效果作为搭便车行为的判定依据。

综上所述,公平的贡献评估机制与异常检测方法相辅相成,均能在不同层面抵御搭便车攻击,共同维护系统的公平性与安全性。具体防御方法的对比与总结见表6。

5 研究展望

当前,IDFL领域仍面临诸多待解挑战。特别是在复杂应用场景、新型安全攻击威胁和动态演进的物理环境下,激励机制的有效性与FL的安全性等方面仍需开展更深入的探索与研究。基于对现有研究的分析和总结,本节提出以下几个未来研究方向。

5.1 异构场景下的IDFL

在实际应用场景中,异构FL正受到越来越多的关注。FL参与者在数据分布、模型结构、网络环境及计算资源等方面的异构特性,对IDFL的隐私保护机制和抗攻击鲁棒性提出了更高要求^[143]。

在隐私保护层面,参与者的异构性往往催生个性化的隐私保护诉求,这需要更精细化的隐私约束机制来满足多样化的隐私保护需求^[144],同时也为激励机制的设计带来了新的挑战。一方面,异构的数据分布与计算资源特性,导致难以

用统一标准衡量参与者的贡献度;另一方面,个体对隐私保护的差异化期望,使得用于计算激励的参与者信息无法实现统一。从攻击威胁角度来看,异构环境的客观存在为恶意参与者提供了更隐蔽的操作空间,使得攻击行为更加难以被识别。因此,在异构场景下,设计适应个性化隐私保护需求且能够对参与者进行有效甄别的IDFL,仍是亟待突破的研究方向。

5.2 多任务IDFL

当前,多数IDFL研究聚焦于单任务学习场景,即由一个任务发布者与多个参与者协作完成一个全局模型的训练任务。在多任务学习场景中,FL的各参与者并非围绕一个共同的学习任务,而是面向各自本地的学习目标,以集中式^[145]或分布式^[146]模式通过FL训练本地模型。

尽管学术界已针对多任务FL提出一些解决方案^[147-149],但针对该场景的激励机制研究仍较为匮乏。Zhao等^[150]针对多任务联邦边缘学习场景,提出一种基于多Agent深度强化学习的激励机制。该机制通过Stackelberg博弈模型刻画边缘服务器与终端设备的交互过程,并从理论上证明了Stackelberg均衡的存在性与唯一性。Xuan等^[151]针对多任务发布者场景提出了一种基于契约理论的激励机制,解决了多任务发布者场景下的最优契约设计问题,能够满足个体理性约束和激励兼容性约束需求。

从隐私保护与安全攻击视角分析,多任务场景下的IDFL参与者为实现本地利益最大化往往存在竞争关系,且在实践中常因所属领域差异而存在内在利益冲突。因此,多任务IDFL中的参与者存在充当恶意参与者并实施投毒攻击的潜在动机。可见,构建适应多任务环境的安全IDFL体系,仍是该领域亟待突破的重要研究方向。

表6 IDFL中防御搭便车攻击的方法对比

Table 6 Comparison of methods to defend against free-riding attacks in IDFL

文献	防御方法	优点	缺点
[5, 132-134, 136-137]	贡献评估	能够根据贡献调整奖励或权重,提高合作效率;优化资源分配	需要消耗额外的计算资源和能源、需要公平且可接受的评估方式
[138, 140-142]	异常检测	及时发现攻击、精确识别异常个体	异常检测可能会产生误报、无法完全准确地识别所有搭便车行为

5.3 面向新型攻击的IDFL

作为主流攻击技术之一,后门攻击近年来衍生出多种新形态。Xie等^[152]提出了一种分布式后门攻击方法,将全局触发器拆解为多个单独训练的局部触发器,然后将分散的后门信息聚合,显著提高了攻击的持久性和隐蔽性。Liu等^[153]针对多恶意参与者的FL场景,提出一种全组合后门攻击方法,利用组合理论设计触发策略,强化了受攻击模型的后门效应,展现出更强的攻击持久性与鲁棒性。当前,针对此类分布式后门攻击的防御研究尚未取得实质性进展。有针对性地设计新型激励机制,或许能通过参与者筛查机制阻断多个攻击者之间的恶意协同。

作为一种新兴攻击技术,Label-only攻击近年来逐渐受到学术界关注^[154]。该攻击仅需利用模型对样本的预测标签(不需要预测分数向量)即可实施,使得对该攻击的防御面临巨大挑战。目前,学者已陆续提出基于Label-only的模型偷取攻击^[155]、成员推理攻击^[156-157]、模型反转攻击^[158]、对抗样本攻击^[159]等新型攻击方式。从Label-only的攻击原理来看,防御此类攻击可能需通过修改或扰动模型输出标签来解决。差分隐私方法显然能够为实现这一目的提供可行思路^[160],但其基本原理决定了在提供强防护(即隐私预算较小)时,必然以牺牲较大的模型性能为代价。当前,针对Label-only攻击的防御研究仅有零星成果^[161],整体仍处于起步阶段。从激励机制视角出发,通过设计合理策略鼓励参与者提供高质量本地数据与模型更新,缓解FL模型过拟合问题,有助于增强对Label-only攻击的防御能力。

总体而言,FL过程中的风险点覆盖所有参与者,防御攻击与激励机制在实现安全FL目标中具有相辅相成的作用。面对持续涌现的新型攻击手段,将两者有机结合以设计高鲁棒性的IDFL框架,是未来极具价值的研究方向。

5.4 面向下一代网络的IDFL

第六代(6G)通信网络的应用使得依赖高速数据传输、低时延和大规模连接的各种服务成为可能,这恰好迎合了FL的发展需求。然而,在6G通信中应用FL也面临着新的挑战,包括更

加普遍的数据异构问题、更高的计算和通信代价、移动客户端的可扩展性、6G通信中数据的无缝聚合等,将激励机制融入FL的框架中进行整体设计,能够在参与者选择、提高资源利用率和抵御安全攻击等方面发挥积极作用。

在6G网络中,数据通常分布在边缘设备上,这些设备在能耗、计算资源、网络时延等方面存在巨大的异构性。通过激励机制来优化客户端资源和任务的分配,以满足6G网络低时延需求,能够有效保障FL的训练效率。Bouzinis等^[162]提出了一种基于Stackelberg博弈的激励机制,通过鼓励用户贡献无线资源和计算资源,同时检测并阻止恶意用户,促使FL快速收敛。Zhu等^[163]针对FL中客户的动态性和加入联邦的自主性,提出了一种基于声誉的动态激励机制,利用合作博弈模型评估客户端在声誉体系中的关联贡献,以提升6G环境下FL的能量效率与模型性能。Cai等^[164]利用区块链技术实现去中心化,避免单点故障,同时结合博弈论设计激励机制,激励诚实参与者积极参与、防范恶意参与者,从而解决6G无线网络中FL面临的安全和效率问题。

此外,6G网络的高带宽与低时延特性使大规模分布式FL成为可能。为提升学习效率并降低隐私泄露风险,联邦分割学习(split federated learning, SFL)^[165]被提出。SFL融合了FL的并行计算优势与分割学习^[166]对全局模型和本地模型的隐私保护能力,成为执行大规模FL的新范式。然而,稳定的参与者数量是SFL成功的关键因素^[167],因此在SFL方案中引入激励机制对保障学习效率具有重要的作用。

6 结束语

FL通过激励机制吸引高质量用户持续参与协同学习,为提升学习效率与模型性能奠定基础。而缓解和消除参与者对隐私泄露风险和安全攻击的担忧,是实现这一目标的核心环节。自FL概念被提出以来,其安全性与鲁棒性便备受学术界关注。学者通过模拟和构建多样化攻击模型,系统分析FL面临的安全威胁及隐私保护目标的达成程度,并针对性地提出解决方案。与之相应,能够适配这些安全FL方案的激励机制,近年来也成为研究热点,推动了FL基础理论与

技术的持续完善。

本文将 IDFL 面临的安全攻击划分为被动攻击与主动攻击,并将两类攻击产生的威胁分别定义为隐私威胁和安全威胁。基于此,本文从面向隐私保护的 IDFL 和面向安全防御的 IDFL 两个维度,对现有研究进行分类梳理,深入剖析各类激励机制的基本原理、应用场景及应对不同攻击的优缺点。

随着 FL 技术的迭代与应用场景的拓展,新型隐私风险与安全威胁将持续涌现,对 IDFL 方案的设计提出更高要求。本文最后针对不同应用场景下 IDFL 的待解决问题展开讨论,并对未来研究方向进行前瞻性展望。

参考文献:

- [1] ALEDHARI M, RAZZAK R, PARIZI R M, et al. Federated learning: a survey on enabling technologies, protocols, and applications[J]. IEEE Access, 2020, 8: 140699-140725.
- [2] LUOS, CHEN X, WU Q, et al. HFEL: joint edge association and resource allocation for cost-efficient hierarchical federated edge learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6535-6548.
- [3] ALI A M, ILAHI I, QAYYUM A, et al. Incentive-driven federated learning and associated security challenges: a systematic review[C]// Proceedings of the 2021 ACM Conference on Incentive-Driven Federated Learning. New York: ACM Press, 2021.
- [4] ZHAN Y, ZHANG J, HONG Z, et al. A survey of incentive mechanism design for federated learning[J]. IEEE Transactions on Emerging Topics in Computing, 2021, 10(2): 1035-1044.
- [5] LIU Z, CHEN Y, ZHAO Y, et al. Contribution-aware federated learning for smart healthcare[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2022: 12396-12404.
- [6] GUO W, WANG Y, JIANG P. Incentive mechanism design for federated learning with stackelberg game perspective in the industrial scenario[J]. Computers & Industrial Engineering, 2023, 184: 109592.
- [7] LIM W Y B, XIONG Z, KANG J, et al. When information freshness meets service latency in federated learning: a task-aware incentive scheme for smart industries[J]. IEEE Transactions on Industrial Informatics, 2022, 18(1): 457-466.
- [8] SONG C, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 587-601.
- [9] ZHANG J, CHEN J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets[C]// Proceedings of the 18th IEEE International Conference on Trust. Piscataway: IEEE Press, 2019: 374-380.
- [10] ZHU L, LIU Z, SONG H. Deep leakage from gradients[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2020: 14774-14784.
- [11] 乐紫莹, 陈珂, 寿黎但, 等. 基于梯度放大的联邦学习激励欺诈攻击与防御[J]. 软件学报, 2025, 36(5): 2254-2269.
- [12] YUE Z Y, CHEN K, SHOU L D, et al. Reward fraud attack and defense for federated learning based on gradient scale-up[J]. Journal of Software, 2025, 36(5): 2254-2269.
- [12] ZHANG C, XIE Y, BAI H, et al. A survey on federated learning[J]. Knowledge-Based Systems, 2021, 216: 106775.
- [13] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.
- [14] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proceedings of Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273-1282.
- [15] HUANG Y, CHU L, ZHOU Z, et al. Personalized cross-silo federated learning on non-IID data[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Hong Kong: AAAI Press, 2021: 7865-7873.
- [16] WANG H, YUROCHKIN M, SUN Y, et al. Federated learning with matched averaging[J]. arXiv preprint arXiv: 2002.06440, 2020.
- [17] KARIMIREDDY S P, KALE S, MOHRI M, et al. SCAFFOLD: stochastic controlled averaging for federated learning[C]// Proceedings of the 37th International conference on machine learning. New York: PMLR, 2020: 5132-5143.
- [18] GAO J, ZHAO L, SHEN X. Network utility maximization based on an incentive mechanism for truthful reporting of local information[J]. IEEE Transactions on Vehicular Technology, 2018, 67(8): 7523-7537.
- [19] ZENG R, ZENG C, WANG X, et al. A comprehensive survey of incentive mechanism for federated learning[J]. arXiv preprint arXiv: 2106.15406, 2021.
- [20] SOLAN E. Stochastic games[C]// Proceedings of the Meyers RA. 2012: 3064-3074.
- [21] JIA R, DAO D, WANG B, et al. Towards efficient data valuation based on the shapley value[C]// Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2019: 1167-1176.
- [22] SARIKAYA Y, ERCETIN O. Motivating workers in federated learning: a stackelberg game perspective[J]. IEEE Networking Letters, 2019, 2(1): 23-27.
- [23] KHAN L U, PANDEY S R, TRAN N H, et al. Federated learning for edge networks: resource optimization and incentive mechanism[J]. IEEE Communications Magazine, 2020, 58(10): 88-93.
- [24] WEMPE B. Contract theory[M]. Berlin: Springer-Verlag, 2020.
- [25] ERASLAN B, GOZUPEK D, ALAGOZ F. An auction theory based algorithm for throughput maximizing scheduling in centralized cognitive radio networks[J]. IEEE Communications Letters, 2011, 15(7): 1167-1170.

- 734-736.
- [26] CHANDRASHEKAR T S, NARAHARI Y, ROSA C H, et al. Auction-based mechanisms for electronic procurement[J]. IEEE Transactions on Automation Science and Engineering, 2007, 4(3): 297-321.
- [27] FENG J, WU Y, SUN H, et al. Panther: practical secure two-party neural network inference[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1149-1162.
- [28] LYU L J, YU H, MA X J, et al. Privacy and robustness in federated learning: attacks and defenses[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 8726-8746.
- [29] 陈晋音, 李荣昌, 黄国瀚, 等. 纵向联邦学习方法及其隐私和安全综述[J]. 网络与信息安全学报, 2023, 9(2): 1-20.
- CHEN J Y, LI R C, HUANG G H, et al. Survey on vertical federated learning: algorithm, privacy and security[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 1-20.
- [30] 顾晋豪, 白跃彬. 联邦学习模型安全与隐私研究进展[J]. 软件学报, 2023, 34(6): 2833-2864.
- GU Y H, BAI Y B. Survey on security and privacy of federated learning models[J]. Journal of Software, 2023, 34(6): 2833-2864.
- [31] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]// Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 739-753.
- [32] KRICHEN M. Generative adversarial networks[C]// Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2023: 1-7.
- [33] 王坤庆, 刘婧, 李晨, 等. 联邦学习安全威胁综述[J]. 信息安全研究学报, 2022, 8(3): 223-234.
- WANG K Q, LIU J, Li C, et al. A survey on threats to federated learning[J]. Journal of Information Security Research, 2022, 8(3): 223-234.
- [34] SHEN M, WANG H, ZHANG B, et al. Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing[J]. IEEE Internet of Things Journal, 2021, 8(4): 2265-2275.
- [35] MELIS L, SONG C, DE CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]// Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 691-706.
- [36] SONG M, WANG Z, ZHANG Z, et al. Analyzing user-level privacy attack against federated learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430-2444.
- [37] DRIOUICH I, XU C, NEGLIA G, et al. Local model reconstruction attacks in federated learning and their uses[J]. arXiv preprint arXiv: 2210.16205, 2022.
- [38] ZHANG X, LI J, ZHANG J, et al. Data reconstruction from gradient updates in federated learning[C]// Proceedings of the Machine Learning for Cyber Security: 4th International Conference. Berlin: Springer, 2023: 586-596.
- [39] MOTHUKURI V, PARIZI R M, POURIYEH S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619-640.
- [40] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]// Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.
- [41] 高莹, 陈晓峰, 张一余, 等. 联邦学习系统攻击与防御技术研究综述[J]. 计算机学报, 2023, 46(9): 1781-1805.
- GAO Y, CHEN X F, ZHANG Y Y, et al. A survey of attack and defense techniques for federated learning systems[J]. Chinese Journal of Computers, 2023, 46(9): 1781-1805.
- [42] WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: yes, you really can backdoor federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084.
- [43] XIE C, CHEN M, CHEN P Y, et al. CRFL: certifiably robust federated learning against backdoor attacks[C]// Proceedings of International Conference on Machine Learning. New York: PMLR, 2021: 11372-11382.
- [44] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems[C]// Proceedings of Computer Security-ESORICS 2020: 25th European Symposium on Research in Computer Security. 2020: 480-501.
- [45] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031-2063.
- [46] FANG M, CAO X, JIA J, et al. Local model poisoning attacks to Byzantine-Robust federated learning[C]// Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). 2020: 1605-1622.
- [47] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]// Proceedings of International Conference on Machine Learning. 2019: 634-643.
- [48] FARHADKHANI S, GUERRAOURI R, VILLEMAUD O. An equivalence between data poisoning and Byzantine gradient attacks[C]// Proceedings of International Conference on Machine Learning. 2022: 6284-6323.
- [49] WEN J, ZHANG Z, LAN Y, et al. A survey on federated learning: challenges and applications[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(2): 513-535.
- [50] BLANCHARD P, MHAMDI E M E, GUERRAOURI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 118-128.
- [51] LI Y, LI Y, WU B, et al. Invisible backdoor attack with sample-specific triggers[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 10-17.
- [52] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning[J]. arXiv preprint arXiv: 1911.07963, 2019.
- [53] LIN J, DU M, LIU J. Free-riders in federated learning: attacks and defenses[J]. arXiv preprint arXiv: 1911.12560, 2019.

- [54] FRABONI Y, VIDAL R, LORENZI M. Free-rider attacks on model aggregation in federated learning[C]// International Conference on Artificial Intelligence and Statistics. 2021: 1846-1854.
- [55] FENG J, YANG L T, ZHU Q, et al. Privacy-preserving tensor decomposition over encrypted data in a federated cloud environment[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 17(4): 857-868.
- [56] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Proceedings of the Theory of Cryptography. 2006: 3876.
- [57] HAEBERLEN A, PIERCE B C, NARAYAN A. Differential privacy under fire[C]// Proceedings of the 20th USENIX Conference on Security. Berkeley: USENIX Association, 2011: 33.
- [58] HU R, GUO Y, LI H, et al. Personalized federated learning with differential privacy[J]. IEEE Internet of Things Journal, 2020, 7(10): 9530-9539.
- [59] LEE J, KIM D, NIYATO D. Market analysis of distributed learning resource management for internet of things: a game-theoretic approach[J]. IEEE Internet of Things Journal, 2020, 7(9): 8430-8439.
- [60] HU R, GONG Y. Trading data for learning: incentive mechanism for on-device federated learning[C]// Proceedings of the 2020 IEEE Global Communications Conference. Piscataway: IEEE Press, 2020: 1-6.
- [61] BUN M, STEINKE T. Concentrated differential privacy: simplifications, extensions, and lower bounds[C]// Theory of Cryptography Conference. 2016: 635-658.
- [62] LIU Y, LIU P, JING W, et al. PD2S: a privacy-preserving differentiated data sharing scheme based on blockchain and federated learning[J]. IEEE Internet of Things Journal, 2023, 10(24): 21489-21501.
- [63] HUANG G, WU Q, SUN P, et al. Collaboration in federated learning with differential privacy: a stackelberg game analysis[J]. IEEE Transactions on Parallel and Distributed Systems, 2024, 35(3): 455-469.
- [64] LI M, TIAN Y, ZHANG J, et al. IMFL: an incentive mechanism for federated learning with personalized protection[J]. IEEE Internet of Things Journal, 2024, 11(13): 23862-23877.
- [65] YUAN W, WANG X. QI-DPFL: quality-aware and incentive-boosted federated learning with differential privacy[J]. arXiv preprint arXiv: 2404.08261, 2024.
- [66] LIU T, DI B, AN P, et al. Privacy-preserving incentive mechanism design for federated cloud-edge learning[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(3): 2588-2600.
- [67] PAN Q, WU J, BASHIR A K, et al. Joint protection of energy security and information privacy federated learning approach[J]. IEEE Transactions on Industrial Informatics, 2022, 18(5): 3473-3483.
- [68] MAO W, MA Q, LIAO G, et al. Game analysis and incentive mechanism design for differentially private cross-silo federated learning[J]. IEEE Transactions on Mobile Computing, 2024, 23(10): 9337-9351.
- [69] ZHANG L, ZHU T, XIONG P, et al. A robust game-theoretical federated learning framework with joint differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 3333-3346.
- [70] YIN L, LIN S, SUN Z, et al. A game-theoretic approach for federated learning: a trade-off among privacy, accuracy and energy[J]. Digital Communications and Networks, 2024, 10(2): 389-403.
- [71] SUN P, CHE H, WANG Z, et al. Pain-FL: personalized privacy-preserving incentive for federated learning[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(12): 3805-3820.
- [72] WU M, YE D, DING J, et al. Incentivizing differentially private federated learning: a multidimensional contract approach[J]. IEEE Internet of Things Journal, 2021, 8(13): 10639-10651.
- [73] LIN X, WU J, LI J, et al. Heterogeneous differential-private federated learning: trading privacy for utility truthfully[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(6): 5113-5129.
- [74] ZHENG S, CAO Y, YOSHIKAWA M, et al. FL-Market: trading private models in federated learning[C]// Proceedings of the 2022 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2022: 1525-1534.
- [75] YING C, JIN H, WANG X, et al. Double insurance: incentivized federated learning with differential privacy in mobile crowdsensing[C]// Proceedings of the 2020 International Symposium on Reliable Distributed Systems (SRDS). Piscataway: IEEE Press, 2020: 81-90.
- [76] WANG Z, LI J, HU J, et al. Towards privacy-driven truthful incentives for mobile crowdsensing under untrusted platform[J]. IEEE Transactions on Mobile Computing, 2023, 22(2): 1198-1212.
- [77] ZHU K, HUANG L, NIE J, et al. Privacy-aware double auction with time-dependent valuation for blockchain-based dynamic spectrum sharing in IoT systems[J]. IEEE Internet of Things Journal, 2023, 10(8): 6756-6768.
- [78] REN K, LIAO G C, MA Q, et al. Differentially private auction design for federated learning with Non-IID data[C]// Proceedings of the 2022 International Conference on Service Science (ICSS). 2022: 305-312.
- [79] LI D, GUO Q, YANG C, et al. Trusted data sharing mechanism based on blockchain and federated learning in space-air-ground integrated networks[J]. Wireless Communications and Mobile Computing, 2022, 1: 5338876.
- [80] YANG Y, HU M, ZHOU Y, et al. CSRA: robust incentive mechanism design for differentially private federated learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 892-906.
- [81] WANG D, REN J, WANG Z, et al. PrivAim: a dual-privacy preserving and quality-aware incentive mechanism for federated learning[J]. IEEE Transactions on Computers, 2023, 72(7): 1913-1927.
- [82] LIU H, MING Y, WANG C, et al. Flexible selective data sharing with fine-grained erasure in VANETs[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 9582-9597.
- [83] ZHANG C, LI S, XIA J, et al. BatchCrypt: efficient homomorphic encryption for cross-silo federated learning[C]// Proceedings of the 2020 USENIX Annual Technical Conference. Berkeley: USENIX

- Association, 2020: 493-506.
- [84] GAMAL T E. A public key cryptosystem and a signature scheme based on discrete logarithms[J]. *IEEE Transactions on Information Theory*, 1984, 31(4): 469-472.
- [85] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[C]// *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques*. Berlin: Springer, 1999: 223-238.
- [86] WENG J, WENG J, ZHANG J, et al. DeepChain: auditable and privacy-preserving deep learning with blockchain-based incentive[J]. *IEEE Transactions on Dependable and Secure Computing*, 2021, 18(5): 2438-2455.
- [87] 周炜, 王超, 徐剑, 等. 基于区块链的隐私保护去中心化联邦学习模型[J]. *计算机研究与发展*, 2022, 59(11): 2423-2436.
- ZHOU W, WANG C, XU J, et al. Privacy-preserving and decentralized federated learning model based on the blockchain[J]. *Journal of Computer Research and Development*, 2022, 59(11): 2423-2436.
- [88] JIA Y, XIONG L, FAN Y, et al. Blockchain-based privacy-preserving multi-tasks federated learning framework[J]. *Connection Science*, 2024, 36(1): 2299103.
- [89] ZHAO B, LIU X, CHEN W N, et al. CrowdFL: privacy-preserving mobile crowdsensing system via federated learning[J]. *IEEE Transactions on Mobile Computing*, 2023, 22(8): 4607-4619.
- [90] CHEN Q, YAO L, WANG X, et al. SecMDGM: federated learning security mechanism based on multi-dimensional auctions[J]. *Sensors*, 2022, 22(23): 9434.
- [91] HE Y, LUO M, WU B, et al. A game theory-based incentive mechanism for collaborative security of federated learning in energy blockchain environment[J]. *IEEE Internet of Things Journal*, 2023, 10(24): 21294-21308.
- [92] CHEON J H, HAN K, KIM A, et al. Bootstrapping for approximate homomorphic encryption[C]// *Proceedings of the Advances in Cryptology -EUROCRYPT 2018*. Berlin: Springer, 2018.
- [93] XIONG R, REN W, ZHAO S, et al. CoPiFL: a collusion-resistant and privacy-preserving federated learning crowdsourcing scheme using blockchain and homomorphic encryption[J]. *Future Generation Computer Systems*, 2024, 156: 95-104.
- [94] DOMINGO-FERRER J, BLANCO-JUSTICIA A, MANJÓN J, et al. Secure and privacy-preserving federated learning via coultility[J]. *IEEE Internet of Things Journal*, 2022, 9(5): 3988-4000.
- [95] KASYAP H, MANNA A, TRIPATHY S. An efficient blockchain assisted reputation aware decentralized federated learning framework[J]. *IEEE Transactions on Network and Service Management*, 2023, 20(3): 2771-2782.
- [96] XU X, LYU L. A reputation mechanism is all you need: collaborative fairness and adversarial robustness in federated learning[J]. *arXiv preprint arXiv: 2011.10464*, 2020.
- [97] OUYANG J, WANG W. FedCrowdSensing: incentive mechanism for crowdsensing based on reputation and federated learning[C]// *Proceedings of the 2023 IEEE Symposium on Computers and Communications (ISCC)*. Piscataway: IEEE Press, 2023: 1454-1461.
- [98] ZHAO Y, ZHAO J, JIANG L, et al. Privacy-preserving blockchain-based federated learning for IoT devices[J]. *IEEE Internet of Things Journal*, 2021, 8(3): 1817-1829.
- [99] WANG Z, YAN B, DONG A. Blockchain empowered federated learning for data sharing incentive mechanism[J]. *Procedia Computer Science*, 2022, 202: 348-353.
- [100] YANG X H, LI T C. A blockchain-based federated learning framework for secure aggregation and fair incentives[J]. *Connection Science*, 2024, 36(1): 2316018.
- [101] FENG L, YANG Z, GUO S, et al. Two-layered blockchain architecture for federated learning over the mobile edge network[J]. *IEEE Network*, 2022, 36(1): 45-51.
- [102] KANG J, XIONG Z, NIYATO D, et al. Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory[J]. *IEEE Internet of Things Journal*, 2019, 6(6): 10700-10714.
- [103] LIU Y, LI K, ZHANG Y, et al. A novel reputation computation model based on subjective logic for mobile ad hoc networks[C]// *Proceedings of the 2009 Third International Conference on Network and System Security*. Piscataway: IEEE Press, 2009: 294-301.
- [104] ZHANG X, HUA Y, QIAN C. Secure decentralized learning with blockchain[C]// *Proceedings of the 2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. Piscataway: IEEE Press, 2023: 116-124.
- [105] XU G, ZHOU Z, DONG J, et al. A blockchain-based federated learning scheme for data sharing in industrial internet of things[J]. *IEEE Internet of Things Journal*, 2023, 10(24): 21467-21478.
- [106] REHMAN M H, SALAH K, DAMIANI E, et al. Towards blockchain-based reputation-aware federated learning[C]// *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. Piscataway: IEEE Press, 2020: 183-188.
- [107] WU Y, CHEN G, LIU Y, et al. TBFL: a trusted blockchain-based federated learning system[C]// *Proceedings of the 2022 IEEE 24th Int Conf on High Performance Computing & Communications*. Piscataway: IEEE Press, 2022: 311-318.
- [108] ZHAO S, PU J, FU X, et al. Byzantine-robust federated learning with ensemble incentive mechanism[J]. *Future Generation Computer Systems*, 2024, 159: 272-283.
- [109] ZENG H, LI J, LOU J, et al. BSR-FL: an efficient Byzantine-robust privacy-preserving federated learning framework[J]. *IEEE Transactions on Computers*, 2024, 73(8): 2096-2110.
- [110] LI H, SHI K. Incentive mechanism for federated learning participants based on statistical analysis features[C]// *Proceedings of the 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE 2023)*. Piscataway: IEEE Press, 2023: 158-161.
- [111] CHEN X, LAN P, ZHOU Z, et al. Toward federated learning with Byzantine and inactive users: a game theory approach[J]. *IEEE Access*, 2023, 11: 34138-34149.

- [112] HE Q, TIAN Y, WANG S, et al. RRFL: a rational and reliable federated learning incentive framework for mobile crowdsensing[J]. *Journal of King Saud University-Computer and Information Sciences*, 2024, 36(3): 101977.
- [113] GAO L, LI L, CHEN Y, et al. FIFL: a fair incentive mechanism for federated learning[C]// *Proceedings of the 50th International Conference on Parallel Processing*. New York: ACM Press, 2021: 1-10.
- [114] XU Y, LU Z, GAI K, et al. BESIFL: blockchain-empowered secure and incentive federated learning paradigm in IoT[J]. *IEEE Internet of Things Journal*, 2023, 10(8): 6561-6573.
- [115] WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: a survey[C]// *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023: 6778-6786.
- [116] FUNG C, YOON C J M, BESCHASTNIKH I. Mitigating sybils in federated learning poisoning[J]. *arXiv preprint arXiv: 1808.04866*, 2020.
- [117] ZHOU S, LI K, XIAO L, et al. A systematic review of consensus mechanisms in blockchain[J]. *Mathematics*, 2023, 11(10): 2248.
- [118] PASS R, SHI E. Hybrid consensus: efficient consensus in the permissionless model[C]// *Proceedings of the International Symposium on Distributed Computing (DISC 2017)*. 2017: 91.
- [119] JAKOBSSON M, JUELS A. Proofs of work and bread pudding protocols[C]// *Proceedings of the IFIP TC6/TC11 Joint Working Conference on Secure Information Networks: Communications and Multimedia Security*. 1999: 258-272.
- [120] KING S, NADAL S. PPCoin: peer-to-peer crypto-currency with proof-of-stake[C]// *King2012PPCoinPC*. 2012.
- [121] GRAMOLI V. From blockchain consensus back to Byzantine consensus[J]. *Future Generation Computer Systems*, 2020, 107: 760-769.
- [122] LI Y, CHEN C, LIU N, et al. A blockchain-based decentralized federated learning framework with committee consensus[J]. *IEEE Network*, 2021, 35(1): 234-241.
- [123] QU Y, POKHREL S R, GARG S, et al. A blockchained federated learning framework for cognitive computing in industry 4.0 networks[J]. *IEEE Transactions on Industrial Informatics*, 2021, 17(4): 2964-2973.
- [124] BOUALOUACHE A, BRIK B, SENOUCI S M, et al. On-demand security framework for 5GB vehicular networks[J]. *IEEE Internet of Things Magazine*, 2023, 6(2): 26-31.
- [125] 米波, 翁渊, 黄大荣, 等. 基于区块链共识激励机制的新型联邦学习系统[J]. *信息安全学报*, 2024, 9(1): 15-32.
- MI B, WENG Y, HUANG D R, et al. A novel FL system based on consensus motivated blockchain[J]. *Journal of Cyber Security*, 2024, 9(1): 15-32.
- [126] YING C, XIA F, WEI D S L, et al. BIT-FL: blockchain-enabled incentivized and secure federated learning framework[J]. *IEEE Transactions on Mobile Computing*, 2025, 24(2): 1212-1229.
- [127] BUTERIN V. A next-generation smart contract and decentralized application platform[C]// *Buterin2015ANG*. 2015.
- [128] WANG S. BlockFedML: blockchained federated machine learning systems[C]// *Proceedings of the International Conference on Intelligent Computing, Automation and Systems (ICICAS)*. 2019: 751-756.
- [129] JABERZADEH A, SHRESTHA A K, KHAN F A, et al. Blockchain-based federated learning: incentivizing data sharing and penalizing dishonest behavior[J]. *arXiv preprint arXiv: 2307.10492*, 2023.
- [130] XU G, KONG D, ZHANG K, et al. A model value transfer incentive mechanism for federated learning with smart contracts in aiOT[J]. *IEEE Internet of Things Journal*, 2024, 12(3): 2530-2544.
- [131] LIU Z, CHEN Y, YU H, et al. GTG-Shapley: efficient and accurate participant contribution evaluation in federated learning[J]. *ACM Transactions on Intelligent Systems and Technology*, 2022, 13(4): 1-21.
- [132] SHI Z, ZHANG L, YAO Z, et al. FedFAIM: a model performance-based fair incentive mechanism for federated learning[J]. *IEEE Transactions on Big Data*, 2022, 1-13.
- [133] SALARBASHISHAHRI M, OKEGBILE S D, CAI J. A shapley value-enhanced evaluation technique for effective aggregation in federated learning[C]// *Proceedings of the 2022 IEEE Future Networks World Forum (FNWF)*. Piscataway: IEEE Press, 2022: 88-93.
- [134] RICHARDSON A, FILOS-RATSIKAS A, FALTINGS B. Budget-bounded incentives for federated learning[C]// *Proceedings of the Federated Learning*. Berlin: Springer, 2020: 176-188.
- [135] COOK R D, WEISBERG S J T. Characterizations of an empirical influence function for detecting influential cases in regression[J]. *Technometrics*, 1980, 22: 495-508.
- [136] HUANG G, CHEN X, OUYANG T, et al. Collaboration in participant-centric federated learning: a game-theoretical perspective[J]. *IEEE Transactions on Mobile Computing*, 2022, 22(11): 6311-6326.
- [137] LI L, YU X, CAI X, et al. Contract-theory-based incentive mechanism for federated learning in health crowdsensing[J]. *IEEE Internet of Things Journal*, 2023, 10(5): 4475-4489.
- [138] WANG B, LI H, LIU X, et al. FRAD: free-rider attacks detection mechanism for federated learning in AIoT[J]. *IEEE Internet of Things Journal*, 2023, 11(3): 4377-4388.
- [139] CHEN Y, ZHANG J, YEO C. Privacy-preserving knowledge transfer for intrusion detection with federated deep autoencoding gaussian mixture model[J]. *Information Sciences*, 2022, 609: 1204-1220.
- [140] ZHANG C, XU Y, ELAHI H, et al. A blockchain-based model migration approach for secure and sustainable federated learning in IoT systems[J]. *IEEE Internet of Things Journal*, 2023, 10(8): 6574-6585.
- [141] CHEN Z, ZHANG H, LI X, et al. FDFL: fair and discrepancy-aware incentive mechanism for federated learning[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 8140-8154.
- [142] ZHENG J, LI G, WANG W, et al. Long-term incentive mechanism

- for federated learning: a dynamic repeated game approach[C]// Proceedings of the 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). Piscataway:IEEE Press, 2023: 1-6.
- [143] YE M, FANG X, DU B, et al. Heterogeneous federated learning: state-of-the-art and research challenges[J]. ACM Computing Surveys, 2023, 56(3): 1-44.
- [144] ZHAO P, LAI L. Minimax optimal Q learning with nearest neighbors[J]. IEEE Transactions on Information Theory, 2024, 71(2): 1300-1322.
- [145] SMITH V, CHIANG C K, SANJABI M, et al. Federated multi-task learning[C]// Proceedings of the International Conference on Neural Information Processing Systems. 2017: 4427-4437.
- [146] MARFOQ O, NEGLIA G, BELLET A, et al. Federated multi-task learning under a mixture of distributions[J]. Advances in Neural Information Processing Systems, 2021, 34: 15434-15447.
- [147] HE C, CEYANI E, BALASUBRAMANIAN K, et al. Spreadgnn: decentralized multi-task federated learning for graph neural networks on molecular data[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 6865-6873.
- [148] WU Z, SUN S, WANG Y, et al. FedICT: federated multi-task distillation for multi-access edge computing[J]. IEEE Transactions on Parallel and Distributed Systems, 2024, 35(6): 952-966.
- [149] CHEN Y Q, ZHANG T, JIANG X L, et al. FedBone: towards large-scale federated multi-task learning[J]. Journal of Computer Science and Technology, 2024, 39(5): 1040-1057.
- [150] ZHAO N, PEI Y, LIANG Y C, et al. Multi-agent deep reinforcement learning based incentive mechanism for multi-task federated edge learning[J]. IEEE Transactions on Vehicular Technology, 2023, 72: 13530-13535.
- [151] XUAN S, WANG M, ZHANG J, et al. An incentive mechanism design for federated learning with multiple task publishers by contract theory approach[J]. Information Sciences, 2024, 664: 120330.
- [152] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning[C]// Proceedings of the 8th International Conference on Learning Representations(ICLR). 2020: 1-19.
- [153] LIU T, ZHANG Y, FENG Z, et al. Beyond traditional threats: a persistent backdoor attack on federated learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Hong Kong: AAAI Press, 2024: 21359-21367.
- [154] KAHLA M, CHEN S, JUST H A, et al. Label-only model inversion attacks via boundary repulsion[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE Press, 2022: 15025-15033.
- [155] SANYAL S, ADDEPALLI S, BABU R V. Towards data-free model stealing in a hard label setting[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2022: 15284-15293.
- [156] LI Z, ZHANG Y. Membership leakage in label-only exposures[C]// Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2021: 880-895.
- [157] CHOQUETTE-CHOO C A, TRAMER F, CARLINI N, et al. Label-only membership inference attacks[C]// Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 1964-1974.
- [158] ZHU T, YE D, ZHOU S, et al. Label-only model inversion attacks: attack with the least information[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 991-1005.
- [159] VO V, ABBASNEJADE, RANASINGHE D C J A. RamBoAttack: a robust query efficient deep neural network decision exploit[J]. arXiv preprint arXiv: 2112.05282, 2021.
- [160] WEI J, BAO E, XIAO X, et al. DPIS: an enhanced mechanism for differentially private SGD with importance sampling[C]// Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2022: 2885-2899.
- [161] YE D, ZHU T, GAO K, et al. Defending against label-only attacks via meta-reinforcement learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3295-3308.
- [162] BOUZINIS P S, DIAMANTOULAKIS P D, KARAGIANNIDIS G K. Incentive-based delay minimization for 6G-enabled wireless federated learning[C]// Proceedings of the Frontiers in Communications and Networks. 2022.
- [163] ZHU Y, LIU Z, WANG P, et al. A dynamic incentive and reputation mechanism for energy-efficient federated learning in 6G[J]. Digital Communications and Networks, 2022, 9: 817-826.
- [164] CAI L, DAI Y, HU Q, et al. Bayesian game-driven incentive mechanism for blockchain-enabled secure federated learning in 6G wireless networks[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(5): 4951-4964.
- [165] THAPA C, ARACHCHIGE P C M, CAMTEPE S, et al. SplitFed: when federated learning meets split learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2022: 8485-8493.
- [166] VEPAKOMMA P, SWEDISH T, RASKAR R, et al. No peek: a survey of private distributed deep learning[J]. arXiv preprint arXiv: 1812.03288, 2018.
- [167] HAFI H, BRIK B, FRANGOUDIS P A, et al. Split federated learning for 6G enabled-networks: requirements, challenges, and future directions[J]. IEEE Access, 2024, 12: 9890-9930.

[作者简介]



迟欢欢 (1995-), 女, 河南周口人, 中南财经政法大学博士生, 主要研究方向为联邦学习、隐私保护、激励机制。



熊平 (1974-), 男, 湖北荆州人, 中南财经政法大学教授、博士生导师, 主要研究方向为机器学习、信息安全、联邦学习、隐私保护。



马霄 (1989-), 女, 河南新乡人, 中南财经政法大学副教授, 主要研究方向为推荐系统、联邦学习、图算法、人工智能。



刘恒竹 (1996-), 女, 河南平顶山人, 中南财经政法大学博士生, 主要研究方向为人工智能、隐私保护、机器遗忘。



朱天清 (1979-), 女, 湖北武汉人, 澳门城市大学教授、博士生导师, 主要研究方向为信息安全、人工智能安全、数据隐私保护。