联邦学习后门攻击威胁与对抗性防御方法综述

吕晓婷[1),2),9)] 刘敬楷[1),2)] 刘芷辰[1),2)] 陈 政[1),2)] 许光全[3)]
罗文坚[4)] 沈 蒙[5)] 王 滨[6)] 纪守领[7)] 陈 恺[8)] 王 伟[1),2),9)]

[1)](北京交通大学网络空间安全学院 北京 100044)
[2)](智能交通数据安全与隐私保护技术北京市重点实验室 北京 100044)
[3)](天津大学网络安全学院 天津 300072)
[4)](哈尔滨工业大学(深圳)计算机科学与技术学院 广东 深圳 518055)
[5)](北京理工大学网络空间安全学院 北京 100081)
[6)](浙江全省智能物联网络与数据安全重点实验室 杭州 310053)
[7)](浙江大学计算机科学与技术学院 杭州 310007)
[8)](中国科学院信息工程研究所 北京 101408)
[9)](西安交通大学智能网络与网络安全教育部重点实验室 西安 710049)

摘 要 联邦学习作为一种隐私保护的分布式机器学习范式,允许多个参与方在私有数据不出本地的前提下协同训练机器学习模型。然而,联邦学习所面临的安全威胁日益凸显。后门攻击,因其具有隐蔽性强和破坏力大等特点,给联邦学习后门攻击的检测和防御带来了巨大挑战,成为亟待解决的关键问题。本文全面调研、分析和总结了联邦学习后门攻击与防御方法,并展望了未来技术的发展方向。首先,剖析并总结了联邦学习后门攻击方法。基于不同的攻击目标系统,分别针对横向联邦学习、纵向联邦学习、联邦迁移学习以及异构联邦学习四类系统的后门攻击方法进行分析和总结。其次,分类并归纳了联邦学习后门防御方法。基于防御机制所依赖信息来源的数量,分别针对基于多客户端信息和单客户端信息两类后门防御方法进行分析和总结。再次,梳理并分析了实际应用场景中联邦学习所面临的后门攻击和相应的防御方法。基于不同的实际应用场景,分别针对联邦推荐、联邦图以及联邦物联网三种场景的后门攻击和防御进行讨论和分析。最后,探讨并展望了联邦学习后门攻击与防御方法的未来研究方向。基于不同的研究目标,分别针对技术增强的后门攻击、多层次综合性防御、鲁棒性、隐私性以及可用性的平衡,以及大语言模型在联邦训练框架下的后门脆弱性等方面进行探讨和展望。

关键词 隐私保护;联邦学习;后门攻击;后门攻击防御
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02823

A Survey of Backdoor Attack Threats and Adversarial Defense Methods in Federated Learning

LYU Xiao-Ting[1),2),9)] LIU Jing-Kai[1),2)] LIU Zhi-Chen[1),2)] CHEN Zheng[1),2)] XU Guang-Quan[3)]
LUO Wen-Jian[4)] SHEN Meng[5)] WANG Bin[6)] JI Shou-Ling[7)] CHEN Kai[8)] WANG Wei[1),2),9)]

[1)](School of Cyberspace Science and Technology, Beijing Jiaotong University, Beijing 100044)
[2)](Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing 100044)
[3)](School of Cybersecurity, Tianjin University, Tianjin 300072)

收稿日期:2024-10-28;在线发布日期:2025-08-08。本文得到北京市自然科学基金-丰台轨道交通前沿研究联合基金(L221014)、中国国家铁路集团有限公司系统性重大项目(P2023W002,P2024S003, P2024W001-4)和杭州市钱江特聘专家项目资助。**吕晓婷**,博士,助理教授,主要研究领域为联邦学习和后门攻击。 E-mail: xiaoting.lyu@bjtu.edu.cn。**刘敬楷**,博士研究生,主要研究领域为大模型安全。**刘芷辰**,硕士研究生,要研究领域为联邦学习。**陈 政**,博士研究生,主要研究领域为联邦学习。**许光全**,博士,教授,主要研究领域为人工智能安全。**罗文坚**,博士,教授,主要研究领域为人工智能安全。**沈 蒙**,博士,教授,主要研究领域为人工智能安全。**王 滨**,博士,教授,长江学者,主要研究领域为物联网安全。**纪守领**,博士,教授,长江学者,主要研究领域为人工智能安全。**陈 恺**,博士,研究员,万人领军人才,主要研究领域为人工智能安全。**王 伟**(通信作者),博士,教授,长江学者,主要研究领域为人工智能安全和区块链安全等。 E-mail: wei.wang@xjtu.edu.cn。

[4](School of Computer Science and Technology, Harbin Institute of Technology(Shenzhen), Shenzhen, Gungdong 518055)
[5](School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081)
[6](Zhejiang Key Laboratory of Artificial Intelligence of Things(AIoT)Network and Data Security, Hangzhou 310053)
[7](School of Computer Science and Technology, Zhejiang University, Hangzhou 310007)
[8](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 101408)
[9](Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049)

Abstract Federated Learning (FL), as a privacy-preserving distributed machine learning paradigm, allows multiple participants to collaboratively train models without sharing their private data. However, the security threats faced by FL are becoming increasingly prominent. Backdoor attacks, due to their high stealthiness and destructive nature, pose significant challenges to detection and defense in FL, making them a critical issue that needs to be addressed in this field. This paper comprehensively investigates and summarizes backdoor attacks and defense methods in FL, and explores future technological development directions. First, we dissect and summarize backdoor attacks in FL. Based on different target systems, we analyze and summarize backdoor attacks for four types of systems: horizontal FL, vertical FL, federated transfer learning, and heterogeneous FL. Second, we classify and synthesize the defense methods in FL. Based on the number of information sources utilized for defense, we categorize and analyze two main approaches: defenses relying on multiple-client information and those based on single-client information. Third, we review and analyze the backdoor attacks and defense methods that FL faces in practical application scenarios. Based on different practical application scenarios, we discuss and analyze backdoor attacks and defense methods for federated recommendation, federated graph learning, and federated Internet of Things. Finally, we explore future research directions for federated backdoor attacks and defenses. We highlight key challenges and opportunities in several areas, including the development of technically enhanced backdoor attacks, multi-level and comprehensive defense mechanisms, the trade-offs between robustness, privacy, and usability, and the vulnerabilities of large language models in FL.

Keywords privacy-preserving; federated learning; backdoor attacks; backdoor defense methods

1 引 言

人工智能的迅猛发展不仅加速了新一代产业革命的进程，还在应用创新、企业转型以及社会发展的各个层面产生了极为深远的影响，已然成为各国在国家战略层面高度关注的核心技术之一。随着人工智能数据需求的不断增加，数据的隐私与安全问题引起了全球范围内的广泛关注和深入讨论。越来越多的个人和组织意识到，在数据共享的表象之下，隐藏着不容忽视的隐私和安全风险。

在此背景下，联邦学习(Federated Learning，FL)[1-4]作为一种分布式机器学习范式，因其突出的隐私保护特性而受到广泛关注。联邦学习允许多个客户端在不共享各自私有数据的情况下，共同训练机器学习模型。在这种框架下，私有数据存储在各客户端的本地，只有模型信息在客户端与服务器之间传输和聚合。这一框架规避了私有数据在各方间的直接交换，确保了数据的隐私性。

尽管联邦学习在数据隐私保护方面具备显著优势，但在模型安全性与完整性方面仍面临严峻挑战，其中后门攻击(Backdoor Attacks)尤为突出。在传统集中式机器学习中，数据集中存储于单一服务器，攻击者通常通过数据投毒等方式在训练阶段植入后门，使模型在特定触发条件下输出预设结果。已有研究[5-6]对集中式后门攻击进行了系统归纳和总结。相比于拜占庭攻击，后门攻击具有更强的隐蔽性，对现有防御机制提出了更严峻的挑战。

联邦学习框架下的后门攻击展现出独特性与更高的复杂性，主要源于其分布式数据存储与多方协同计算机制。数据分布的异质性及客户端间的复杂交互，使得后门攻击在联邦环境中具备更高的隐蔽

性和更丰富的攻击方式。研究表明，联邦学习中的后门攻击不仅受数据分散特性的影响，还受到模型聚合机制、客户端选择策略等多重因素制约，使其攻击模式与集中式学习存在本质性差异。

因此，针对联邦学习的后门攻击防御，必须结合其分布式架构特性与安全需求，构建系统级安全防护体系。与集中式模型侧重于单点防御（如本地异常检测）不同，联邦学习的防御策略需要涵盖客户端筛选、传输加密、安全聚合等多个环节，形成端到端的安全保障体系。这种差异本质上源于联邦学习的跨设备协同训练特性，以及模型更新过程中潜在的安全威胁。

根据团队的前期研究[7-9]表明，即便联邦学习系统部署了多种防御措施，仍可能遭受精心设计的后门攻击。这类攻击因其高度的隐蔽性和破坏力，对联邦学习系统的安全性构成了严重威胁。因此，构建有效的防御机制已成为当前联邦学习领域的一项紧迫任务。深入探讨联邦学习后门攻击及其防御策略，不仅有助于提升系统的安全性，也为联邦学习技术的进一步发展奠定了重要基础。

目前，国内外已有许多关于联邦学习的综述研究。Abdulrahman等人[10]对联邦学习面临的技术挑战进行了探讨，但并未深入剖析具体的安全威胁。Mothukuri等人[11]则对联邦学习中的安全和隐私问题进行了系统性的回顾，但同样没有对后门攻击进行详尽的分析。Yin等人[12]提供了一个全面的隐私保护分析框架，并讨论了联邦学习框架中隐私泄露的风险。Chen等人[13]专注于联邦学习的安全和隐私保护问题，分析了现有联邦学习隐私和安全漏洞，并对差分隐私、同态加密系统和安全多方计算等技术进行了调研，但其未聚焦于后门风险的讨论。Gao等人[14]详细分析了对抗性和非对抗性攻击对联邦学习的影响，但对后门攻击与防御方法未进行充分的聚焦讨论。Wu等人[15]则专注于横向联邦学习的研究现状、系统应用与挑战，进行了总结与分析。作者团队的前期工作[16]探讨了联邦学习整个生命周期内存在的安全与隐私威胁，然而，该研究并未对联邦学习中的后门攻击及其防御方法进行专门的详细分析和归纳。

Nguyen等人[17]对后门攻击进行了深入分析，涵盖了数据投毒和模型投毒两个主要方面，并从聚合前、中、后三个阶段对现有的防御技术进行了全面的归纳与总结。Gong等人[18]针对联邦学习中的后门攻击进行了系统性的分析与总结，采用数据投毒和模型投毒的分类方式进行讨论，并将防御方法归纳为异常更新检测、鲁棒的联邦训练以及后门模型修复三类。Jialang等人[19]亦基于数据投毒和模型投毒的分类框架，探讨了联邦学习中的后门攻击。

本文在研究分析视角与方法上与现有综述文献有所区别。首先，本文系统性地梳理并深入分析了针对不同受攻击的联邦学习系统框架的后门攻击方法。接着，依据防御者所依赖的不同信息类型，对现行的联邦后门防御策略进行了归类与归纳。此外，本文从实际应用场景出发，对联邦学习中的后门攻击及其防御机制进行了深入探讨和研究。最终，本文对联邦学习后门攻击与防御技术的发展趋势进行了前瞻性探讨，以期促进该领域安全研究的深入。

本文的组织结构安排如图1所示：第1节介绍联邦学习和后门攻击的研究背景；第2节阐述联邦学习的基本定义及其潜在威胁；第3节定义联邦学习后门攻击威胁模型；第4节分析针对不同联邦学习系统的后门攻击方法；第5节分类讨论现有的联邦学习后门防御方法；第6节分析联邦学习在应用场景中面临的后门攻击和防御方法；第7节讨论联邦学习中关于后门攻击与防御方法潜在的未来研究方向；第8节总结全文。

2 联邦学习及其潜在威胁

2.1 联邦学习定义

联邦学习作为一种隐私保护的分布式机器学习范式，其核心思想是：在保证私有数据不出本地的情况下，多个客户端协同训练机器学习模型。在此框架下，客户端之间不直接共享各自的本地数据，而是将基于本地数据训练得到的中间结果（通常为模型参数或模型梯度）上传至服务器。服务器随后对这些上传的中间结果进行聚合，以优化全局模型并广播给各个客户端。

具体地，联邦学习的目标是通过多个客户端的数据分布学习一个全局模型G使得其能够在全体数据上表现良好。数学上，联邦学习的优化目标是最小化全局损失函数：

$$\min_{G} F(G)=\sum_{k=1}^{K} p_k F_k(G) \tag{1}$$

其中，K是客户端的总数；p_k是客户端k的权重，n_k是客户端k的样本数量；$F_k(G)$是客户端k上的损失函数，例如交叉熵或均方误差；$F(G)$是全局的目标损失函数。

联邦学习可以通过考量数据特征与样本的重叠

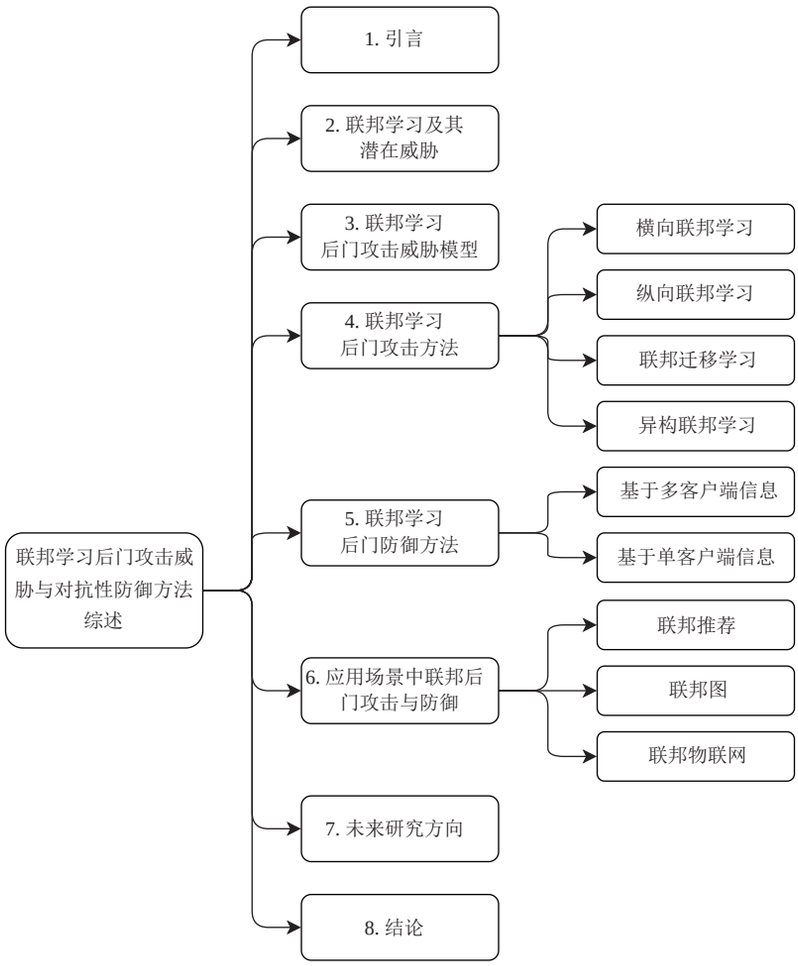

图1 本文内容组织框架图

程度、通信架构以及参与客户端的特性来进行细致的分类。本文主要聚焦于基于数据特征和样本重叠度的分类方法。具体而言，依据数据特征和样本索引之间的重叠情况，联邦学习可以被划分为三种主要形式：横向联邦学习（Horizontal FL，HFL）、纵向联邦学习（Vertical FL，VFL）以及联邦迁移学习（Federated Transfer Learning，FTL）[2]。横向联邦学习适用于客户端拥有相同或相似的数据特征，但样本索引不重叠的情况。例如，不同地区的银行可能收集到的客户数据在特征上类似，但客户身份并不相同。纵向联邦学习则适用于客户端拥有相同样本索引但数据特征不同的场景。例如，银行和电商可能服务相同客户，但银行掌握客户的财务交易数据，而电商则拥有客户的购物浏览和购买记录。联邦迁移学习适用于客户端在数据特征空间和样本索引上都存在差异的场景。例如，不同国家的医疗机构可能在某些疾病分类特征上有交集，但每个机构也拥有独特的特征和病例数据。除此之外，异构联邦学习也是一类特殊的联邦学习框架，专门应对参与方设备或数据在特征、模型架构、计算能力等方面存在差异的情况。

2.2 联邦学习潜在威胁

联邦学习作为一种分布式机器学习范式，虽然在保护数据隐私方面具有显著优势，但其分布式架构和协作训练机制也引入了多种潜在威胁。这些威胁不仅限于模型层面的攻击，还涉及系统安全性和数据隐私性的多方面挑战。根据攻击目标和影响范围，联邦学习面临的潜在威胁主要可分为两大类：安全威胁和隐私威胁。

2.2.1 安全威胁

针对联邦学习的安全威胁，主要关注于破坏模型的完整性和可用性。攻击者通过操控恶意客户端的训练数据或模型更新来降低全局模型的性能或者在全局模型中植入后门，以实现其特定目标。这类安全威胁可以根据攻击者的目标分为有目标投毒攻击和无目标投毒攻击。

联邦学习中的无目标投毒攻击（Untargeted Poisoning Attacks）[20-21]通过在恶意客户端的本地训

练数据中引入投毒样本，或直接操纵本地模型的更新，对全局模型施加负面影响，降低模型性能。这种攻击的目的在于削弱全局模型在处理正常数据时的准确性和有效性，而非针对特定的输入样本。无目标投毒攻击对联邦学习系统的可用性构成了严重威胁。

联邦学习中的有目标投毒攻击（Targeted Poisoning Attacks）[22-23]通过在恶意客户端的本地数据中植入恶意样本或篡改模型更新，旨在操纵全局模型对特定任务或输入的响应，使其输出攻击者预定的结果。此类攻击在不影响模型整体性能的情况下，针对性地破坏特定目标，严重威胁了联邦学习系统的安全与完整。

联邦学习后门攻击[24]是一类特殊的有目标投毒攻击。其核心目的是在全局模型中植入隐秘的后门，使得模型在接收到含有攻击者预设触发器的输入时，能够按照攻击者的意图进行输出。而在面对不含触发器的正常样本时，模型的性能则保持正常，与未受攻击的模型表现相同。后门攻击的显著特点是其高度的隐秘性，仅在特定触发器激活时显现，这使得防御者难以察觉和防范。

除了上述提到的安全威胁，联邦学习在安全方面还面临着一系列其他挑战，尤其是在客户端与服务器之间的通信环节。在联邦学习的框架下，客户端与服务器之间需要频繁交换中间训练结果。这一通信过程若缺乏有效的安全防护，便会暴露出安全隐患。外部攻击者可发动中间人攻击[25]，截获并恶意篡改传输的信息，威胁模型的安全性。

2.2.2　隐私威胁

联邦学习隐私威胁主要源于攻击者通过逆向工程（Reverse Engineering）或推理攻击（Inference Attacks）等手段，试图从共享的模型更新或中间梯度信息中推测并恢复客户端的私有数据。这类攻击通常包括成员推断攻击[26-28]、数据重构攻击[29-32]以及属性推断攻击[33-36]等，严重威胁联邦学习的隐私保护能力。成员推断攻击的目的在于揭示特定数据是否被用于模型的训练过程。数据重构攻击则通过深入分析模型更新，尝试重建出与客户端原始样本相仿的数据。与此不同，属性推断攻击并不致力于重现整个数据样本，而是通过解析模型更新的细节，推测出样本中包含的某些敏感属性，例如年龄、性别等。这些攻击手段的存在，极大地增加了数据主体敏感信息泄露的风险。

在上述联邦学习系统面临的众多安全和隐私威胁中，后门攻击因其独特的隐蔽性和危害性而尤为突出。这种攻击能够在保持模型整体性能不变的情况下，破坏模型的完整性，使得模型在特定的触发条件下按照攻击者的意图行事。更为棘手的是，后门的存在往往对联邦学习系统的防御者而言是难以察觉的，这无疑大大提升了检测和防御的复杂性。鉴于此，本文将集中探讨联邦学习中的后门威胁，全面审视和分析现有的后门攻击及其防御方法，旨在深化对联邦学习系统后门漏洞的认识，进而推动联邦学习系统安全性和完整性的提升。

3　联邦学习后门攻击威胁模型

3.1　问题定义

在传统机器学习的后门攻击中，攻击者在训练阶段直接向模型植入特定的后门，导致模型在接收到含有特定触发器的输入时，产生攻击者预设的输出；而对于不包含触发器的正常输入，模型的表现正常。这种攻击在安全敏感的应用领域尤为危险。

而在联邦学习系统中，后门攻击呈现出更为复杂和隐蔽的形态。攻击者无法直接篡改全局模型，而是通过操纵恶意客户端的本地模型，间接地影响全局模型的更新，以此在全局模型中悄无声息地植入后门。如图2所示，攻击者通过在恶意客户端的数据集中植入含有特定触发器的后门样本，生成带有恶意的本地模型更新，并将其发送至服务器。在联邦学习模型的聚合阶段，这些恶意更新对全局模型的更新方向产生影响，从而将后门嵌入到全局模型中。

3.2　威胁模型

在联邦学习后门攻击中，攻击者操控少数恶意客户端，通过在这些客户端的数据中植入特定触发模式来达成攻击目的。这些模式可能包括图像的特定像素布局或文本中的关键短语等。攻击者促使本地模型学习触发器与目标输出间的映射，从而在全局模型中隐蔽地植入后门。

在横向联邦学习中，令G为全局模型，C_{total}代表参与联邦学习的客户端集合，其中$C_{mal} \subset C_{\text{total}}$为攻击者控制的恶意客户端子集。每个恶意客户端$c_i \in C_{mal}$基于其本地数据集$D_i^{cln}$来构造后门样本集$D_i^{bkd}$。后门样本的构建是通过向正常样本添加特定的触发器$\Delta$，并改其标签为攻击者期望的标签$y^{tar}$来实现的。给定原始样本$x$、触发器模式$\Delta$和掩码$m$，

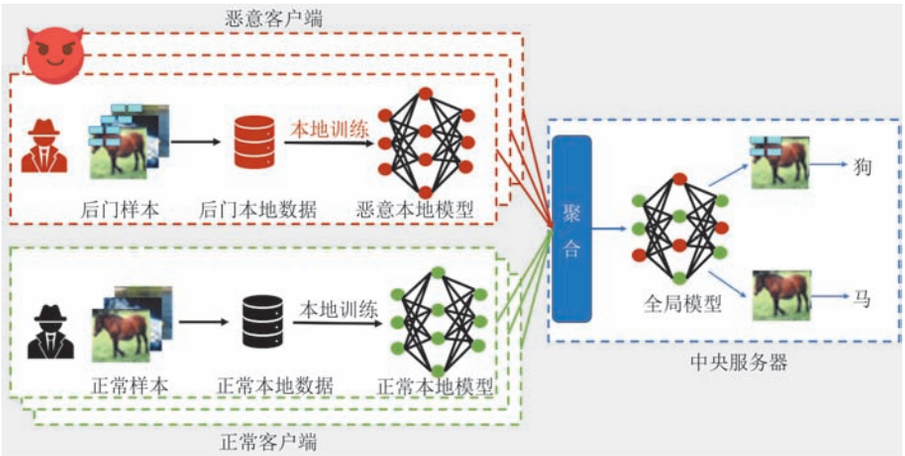

图2 联邦学习中后门攻击示意图

后门样本表示为 $x^{bkd}=(1-m)\odot x+m\odot\Delta$，其中$\odot$表示逐元素乘法。

在联邦学习的第t轮训练中，每个客户端基于其本地数据集D_i^{cln}训练本地模型L_i^t，并将模型更新$\Delta L_i^t=L_i^t-G^{t-1}$发送至服务器。服务器汇聚所有客户端的模型更新，以更新全局模型G^t，其更新公式为

$$G^t=G^{t-1}+\sum_{i=1}^{K}p_k\Delta L_i^t \tag{2}$$

在这一过程中，攻击者控制的恶意客户端通过其生成的恶意模型更新参与到聚合操作中，借此将后门嵌入到全局模型。一旦后门成功植入，全局模型将对带有触发器的后门样本做出错误分类，将其误判为攻击者预设的目标类别y^{tar}。然而，对于正常样本，全局模型仍能保持正确分类。

纵向联邦学习与横向联邦学习的后门攻击的基本原理是相似的。攻击者通过控制的客户端或参与方，将恶意模型更新引入到全局模型中，从而植入后门。不同的是，对于纵向联邦学习的情况，数据特征由多个客户端分布式存储，每个客户端拥有其独特的特征子集。恶意客户端通过在其特征子集上注入特定触发器。每个恶意客户端在训练时利用正常数据集D_i^{cln}和后门数据集D_i^{bkd}优化其本地模型L_i^t，并生成包含后门信息的模型更新ΔL_i^t。

接下来，本文将从攻击者的目标、知识和能力三个方面详细分析联邦学习中的后门攻击威胁模型。

3.2.1 攻击者的目标

联邦学习后门攻击的核心目标涵盖三个主要方面。首先，基本目标是在模型中植入后门，使得后门模型G_{mal}在接收到包含特定触发器的输入x^{bkd}时，能够输出攻击者预定的输出结果y^{tar}。其次，攻击者需确保模型在遇到不包含触发器的正常输入x时，其性能与未受攻击的原始模型G_{ori}一致。即

$$\begin{aligned}&G_{mal}(x^{bkd})=y^{tar},\\&G_{mal}(x)=G_{ori}(x),\\&x^{bkd}=(1-m)\odot x+m\odot\Delta\end{aligned} \tag{3}$$

最后，攻击者控制的恶意客户端生成的模型更新需要能够绕过联邦学习系统所部署的各类防御方法，以成功地将后门植入全局模型之中。

3.2.2 攻击者的知识

后门攻击者所掌握的知识对实施攻击的难度和复杂性有着重要的影响。如表1所示，本文依据攻击者掌握的关键信息，从恶意客户端、正常客户端以及服务器三个角度对其知识范围进行分析。

表1 联邦学习中后门攻击威胁模型下攻击者知识总结

知识持有者	关键信息	获取难度
恶意客户端	本地训练数据（特征分布、类别分布、数据量等）	低
	本地训练过程（本地训练损失函数、优化算法、超参数等）	中
	本地模型更新	中
正常客户端	本地训练数据（特征分布、类别分布、数据量等）	高
	本地训练过程（本地训练损失函数、优化算法、超参数等）	高
	本地模型更新	高
中央服务器	聚合规则、检测和防御方法	高

攻击者通常能够全面访问其所控制的恶意客户端的本地训练数据，包括数据的特征分布、类别分布和数据量等详细信息。此外，恶意客户端在本地训练过程中所使用的损失函数、优化算法、超参数配置，以及生成的模型更新，均保存在攻击者控制的客户端中。因此，一旦这些客户端被完全操控，攻击者便可以轻松掌握所有关键的训练细节。

相比之下，获取不受攻击者控制的正常客户端的相关信息则极具挑战性。在联邦学习架构中，每个客户端的数据和模型信息都是独立存储的，并且不对外共享，这使得攻击者难以直接访问其他正常客户端的本地数据、训练过程和结果。

此外，服务器的聚合规则和防御机制也是攻击者想要获取的重要信息。掌握这些信息有助于攻击者设计出能够绕过系统检测和过滤的恶意模型更新，从而影响全局模型。然而，由于服务器不在攻击者的控制范围内，且其聚合规则和防御方法通常不会公开，获取这些信息的难度极高。

3.2.3　攻击者的能力

在深入探讨联邦学习后门攻击的防御方法之前，理解攻击者的能力范围至关重要。表2详细列出了后门攻击中攻击者的能力范围，涉及仅发布后门数据集、控制恶意客户端、控制正常客户端、控制中央服务器以及攻击频率等多个维度。

表2　联邦学习中后门攻击威胁模型下攻击者能力总结

攻击者能力	恶意行为	执行难度
仅发布后门数据集	篡改数据集（添加触发器，修改标签），并在互联网公开发表	低
控制恶意客户端	篡改本地数据集（添加触发器，修改标签）	中
	控制本地训练过程（修改损失函数、修改训练算法、调整超参数）	中
	篡改本地模型更新	中
控制正常客户端	篡改本地数据集（添加触发器，修改标签）	高
	控制本地训练过程（修改损失函数、修改训练算法、调整超参数）	高
	篡改本地模型更新	高
控制中央服务器	篡改聚合规则以及防御检测机制	高
攻击频率	单轮攻击，攻击效果短暂	低
	多轮攻击，攻击效果持久	高

能力有限的攻击者仅能在互联网上公开发布嵌入后门的数据集，当联邦学习系统中的客户端使用该数据集进行训练时，模型可能会在无意间学习到后门行为。相比之下，能力较强的攻击者通常能够控制少量客户端，这些客户端可能是攻击者注入的虚假节点，或者是已被攻陷的真实客户端。在这些恶意客户端中，攻击者可以篡改本地数据集、操纵训练过程，甚至直接修改模型更新，以在全局模型中植入后门。相较之下，攻击者通常无法在未受其控制的正常客户端和服务器上篡改数据、干扰训练过程或修改模型更新，更无法直接影响模型的聚合与防御机制。

此外，攻击者在联邦学习系统中发动后门攻击的频率也是一个关键能力因素。单轮攻击对攻击者而言较为简单和易行，但其所产生的效果通常是短暂的，且面临快速遗忘的难题。相比之下，多轮攻击能够带来更为持久和显著的影响，然而其执行难度较高，攻击者需持续介入多轮训练，并维持对恶意客户端的操控。在客户端数量庞大的联邦学习系统中，由于每轮训练参与的客户端是随机挑选的，攻击者难以确保恶意客户端能够稳定参与训练。

3.2.4　性能评估指标

以分类任务为例，在评估联邦学习后门攻击的性能时，通常会考虑以下几个指标。

攻击成功率（Attack Success Rate，ASR）是衡量后门攻击有效性的指标之一，它反映了后门样本被识别为目标类别的概率。攻击成功率具有以下两种计算方式。

第一种评估攻击成功率的方法计算时包含了所有类别的测试样本，不论其真实标签是否为目标类别，其定义如下：

$$ASR=\frac{\sum_{i=1}^{M}I\left(G_{mal}\left(x_i^{\mathrm{bkd}}\right)=y^{tar}\right)}{M} \tag{4}$$

其中，y^{tar}为攻击的目标类别；x_i^{bkd}是后门测试集中的样本；M是测试集的样本总数，I是指示函数，当括号内条件成立时取值为1，否则为0。

第二种评估攻击成功率的方法在计算时排除那些真实标签为目标类别的样本，其定义如下：

$$ASR=\frac{\sum_{i=1}^{M}I(G_{mal}(x_i^{bkd})=y^{tar}\wedge y_i\neq y^{tar})}{\sum_{i=1}^{M}I(y_i\neq y^{tar})} \tag{5}$$

其中，y_i是x_i^{bkd}的真实标签。

主任务准确率(Accuracy，ACC)是评估模型在主任务上的性能指标，反映了模型在正常测试数据集上的分类能力。

$$ACC=\frac{\sum_{i=1}^{M}I(G_{mal}(x_i)=y_i)}{M} \tag{6}$$

其中，x_i表示不含有后门触发器的正常测试样本。

主任务准确率下降值(Accuracy Decline，AD)表示在后门攻击前后，模型在主任务上的分类准确率的下降程度。

$$AD=ACC_{ori}-ACC_{mal} \tag{7}$$

其中，ACC_{ori}表示未被攻击的模型的主任务准确率，ACC_{mal}表示嵌入后门的模型的主任务准确率。该指标用于评估后门攻击对模型正常分类性能的影响，反映了后门攻击的目标：在确保攻击成功率(ASR)尽可能高的同时，尽量保持模型主任务准确率下降值(AD)尽可能小。

后门持久性(Durability)是评估后门攻击持续效果的指标，主要反映后门在全局模型部署后的存续时间。在实验设置中，通常通过监测在攻击停止后，经过多轮联邦学习训练，攻击成功率的降低程度来评估后门的持久性。如果攻击成功率的降幅很小，这表明即便在多个正常客户端的持续训练过程中，后门仍能保留在模型中，从而表明该后门具有较高的持久性。

4 联邦学习后门攻击方法

后门攻击利用联邦学习分布式架构的特点，通过恶意客户端在训练过程中向全局模型植入后门，对全局模型的行为进行操控。鉴于联邦学习有多种系统框架，包括横向联邦学习、纵向联邦学习、联邦迁移学习以及异构联邦学习，它们的数据结构和模型训练机制各有特点，因此，后门攻击的策略和执行方法也会随之而异。本章将深入探讨不同的联邦学习框架下后门攻击方法。通过对现有文献和研究成果的细致分析，揭示各种框架下后门攻击的共性和差异，以及它们对系统安全性和完整性的影响。

4.1 面向横向联邦学习的后门攻击

在横向联邦学习框架中[37-39]，各客户端的数据在特征空间上保持高度相似，但样本索引不同。客户端基于本地数据独立训练模型，并向中央服务器提交更新，服务器通过参数聚合形成全局模型。

如图3所示，攻击者可利用这一分布式训练机制，在恶意客户端中植入后门，并通过上传的模型更新逐步污染全局模型，使其在特定触发输入下产生预设的错误预测。这种攻击方式隐蔽性强、难以检测，对联邦学习系统的安全性构成严重威胁。

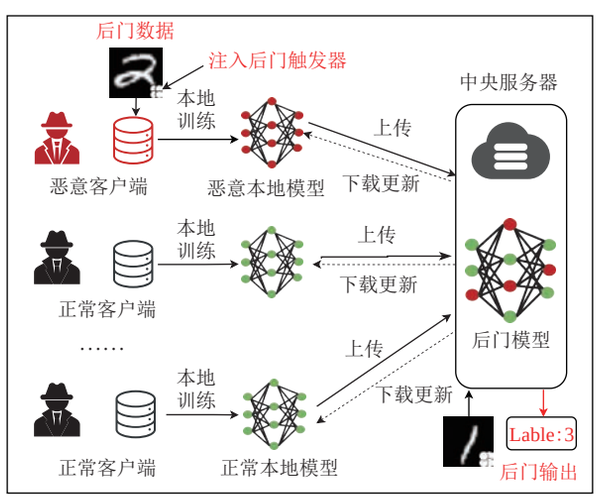

图3 横向联邦学习中后门攻击示意图

更具体地，在通用横向联邦学习中，客户端数据通常服从独立同分布，这一特性使得攻击者能够更精准地实施后门攻击。然而，数据异质性仍然是横向联邦学习的现实挑战，而个性化联邦学习旨在应对这一问题。本文将分别探讨通用横向联邦学习与个性化联邦学习中的后门攻击策略，分析其攻击方式及影响。

4.1.1 通用横向联邦学习的后门攻击

在通用横向联邦学习中，各个客户端的数据特征完全一致，且遵循独立同分布(IID)原则。这种一致性虽然便于模型的训练和测试，但也使得攻击者更容易找到漏洞并实施针对性的后门攻击。

Tolpegin等人[21]提出了一种标签翻转攻击方法。攻击者通过在本地数据集上进行标签翻转操作，将部分数据的标签修改为目标类别，从而在不直接篡改全局模型参数的前提下，间接影响了全局模型的决策。

Sun等人[40]则提出了一种为模型更新投毒攻击(MUPA)。在训练过程的后期，随着全局模型逐渐趋于收敛，客户端上传的模型更新幅度逐渐降低。

在这一阶段，恶意客户端利用其本地的后门数据集进行训练，并将带有后门特征的模型更新上传至服务器，进而操纵全局模型的更新方向。为了提高后门攻击的效能，尤其是在单轮训练中成功植入后门，Bagdasaryan等人[41]提出了一种模型替换攻击。该攻击将原始的全局模型替换为攻击者精心设计的恶意本地模型，以此实现后门的植入。

分布式后门攻击DBA[42]则利用了联邦学习分布式的特点，通过将一个完整的后门触发器分解为多个局部触发器，并分别嵌入到各个客户端的训练数据中。Gong等人[43]进一步发展了这一概念，提出了一种名为CBA的协调后门攻击方法。CBA利用基于模型信息的策略对局部触发器进行优化，进一步提升了攻击效果。Ren等人[44]提出了一种多强度后门攻击方法SBA。该方法通过全局触发器和局部触发器的协同作用，来进行有效攻击。

Tong等人[45]提出了名为SemSBA的后门攻击方法。该方法利用未标记训练样本中的语义特征作为后门触发器，并添加对抗性扰动以生成恶意样本。为了提高攻击成功率，攻击者在模型分配伪标签之前对未标记样本进行扰动，来诱导模型错误地分配目标伪标签。

Zhuang等[46]人的研究聚焦于后门关键层。后门关键层是模型中一小部分层，攻击这些层可以达到与攻击整个模型相当的效果，但更难以被防御机制检测到。攻击者通过替换模型中的不同层来评估每一层对后门攻击成功的影响，变化较大的更可能是后门关键层。在确定后门关键层后，可以通过层级毒化或层级翻转来植入后门。

为了在联邦学习系统中提升后门攻击的隐蔽性并有效规避现有防御措施，Lyu等人[8]提出了一种名为CerP的合谋隐匿攻击策略。通过微调后门触发器、控制攻击者的本地模型偏差，以及增加恶意模型之间的多样性，以实现隐蔽且成功的后门攻击。Fang等人[47]则提出了F3BA攻击方法，该方法通过识别并翻转对主任务影响最小的模型参数符号，增强了模型对后门触发器的敏感性，并联合优化触发器模式与客户端模型，使得攻击更加隐蔽且持久，从而有效规避现有的防御措施。

在联邦学习中，后门攻击的持久性是攻击者在设计攻击策略时必须考虑的关键因素。研究[48]提出了一种名为Chameleon的攻击方法，其核心在于利用对比学习技术来增大后门样本与干扰样本之间的嵌入距离，同时减小与后门样本共享相同目标标签的正常样本之间的嵌入距离。Zhang等人[49]提出了名为Neurotoxin的攻击方法。攻击者通过正常数据集确定更新幅度小的参数作为约束集。然后，攻击者基于后门数据集更新约束集内参数。由于这些参数更新频率低，Neurotoxin植入的后门能够在全局模型中长期存活。与之相似的是，Shi等人[50]通过分析模型更新的历史数据识别出更新幅度极小的冗余神经元，并仅针对这些冗余神经元进行参数调整，以植入后门触发器。为了增强后门持久性，攻击者结合反向梯度优化和正向激活路径增强技术，确保后门在多轮聚合中不易被覆盖。Lyu等人[7]基于投影梯度下降的技术设计了一种名为CoBA的后门攻击方法，该方法可以实现稀疏攻击效果。

Nguyen等人[51]提出了IBA攻击方法。首先，利用生成模型学习并创建难以被检测的触发器，并通过对抗性训练来预测和适应全局模型的动态变化。该方法仅需要控制少量客户端即可实现高攻击成功率，有效规避了主流的防御机制，实现了更持久的后门效果。Zhang等人提出的A3FL[52]利用对抗性训练技术，预测并模拟全局模型可能采取的消除触发器影响的方法，以确保后门在全局模型中持久有效。其还通过选择性的后门模型参数和限制模型更新的范围，增强后门的隐蔽性和持久性。

上述的攻击方法中都假定客户端是攻击者，但Sun等人[53]提出了一种数据无关后门攻击方法DABS，该方法考虑了服务器可能存在恶意行为的情况。DABS允许恶意服务器直接修改全局模型，通过在公共未标记数据集上训练一个带有特定触发器的后门子网络，并用这个子网络替换全局模型的一部分，从而植入后门。

4.1.2　个性化联邦学习的后门攻击

个性化联邦学习[54-58]是横向联邦学习的一个分支，它致力于解决通用横向联邦学习在处理数据异质性方面的不足. 在通用横向联邦学习中，所有客户端共同训练并使用一个统一的全局模型。但这种方法可能无法适应每个客户端独特的数据分布和个性化需求。个性化联邦学习通过为每个客户端引入定制化的本地模型来应对这一挑战。尽管个性化联邦学习提供了更高的灵活性和适应性，但它也面临着后门攻击的挑战。个性化联邦学习中的后门攻击示意如图4所示。与通用横向联邦学习相比，个性化联邦学习中的后门攻击面临着后门在个性化本地模型上灾难性遗忘的问题，这要求攻击者开发出新的策略来确保后门的持久性和有效性。除此之外，

由于个性化联邦学习中客户端的数据分布差异性，增加了攻击者仅通过恶意客户端数据来攻击其他正常客户端的难度。

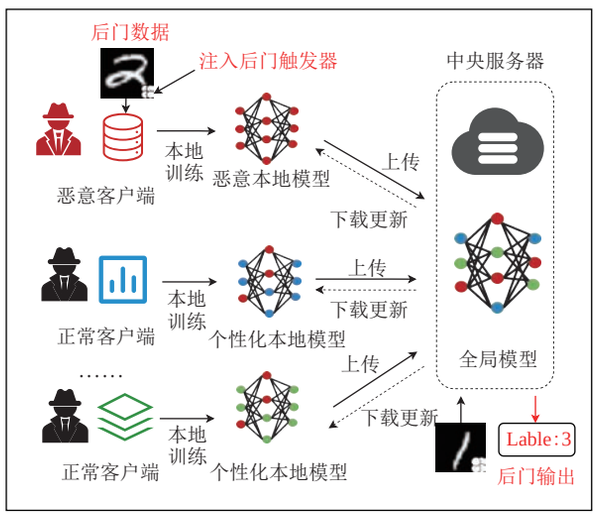

图 4 个性化联邦学习中后门攻击示意图

Qin 等人[59]研究发现，部分共享模型参数的个性化联邦学习能有效提升对后门攻击的鲁棒性。Ye 等人[60]进一步揭示其仍存安全隐患，并提出 BapFL 攻击，仅针对特征编码器施加后门，以规避分类器暴露，同时通过噪声注入提升攻击泛化能力。为了解决后门在个性化本地模型的灾难性遗忘问题，Lyu 等人[9]通过主任务和后门任务的梯度对齐和损失对齐来优化触发器和恶意本地模型。通过主任务和后门任务的对齐使得后门模型与正常模型的决策边界一致，从而导致后门保证持久性，同时可以绕过服务器和客户端部署的各种防御方法。

针对攻击者不同类别数据量不足的问题，Mei 等人[61]提出了一种隐私推理增强的隐蔽后门攻击（PI-SBA）。该方法主要利用生成式对抗网络构建了一个多样化的数据重构机制，生成补充数据集。PI-SBA 独立的后门数据生成流程使得其能够灵活地与其他现有的后门攻击策略进行集成，极大地增强了攻击的适应性和有效性。

为解决数据异质性问题，超网络（Hypernetworks）技术也被引入个性化联邦学习中。研究[62]提出了联邦超网络框架 HyperNetFL，该框架能够为每个客户端定制生成其本地模型参数。然而，该框架仍然面临着后门攻击的威胁。Lai 等人[63]针对 HyperNetFL 提出了 HNTroj 模型转移后门攻击。在此方法中，攻击者通过向超网络注入后门，使得超网络为客户端生成带有后门的本地模型参数，从而达到攻击者预定的攻击效果。

4.2 面向纵向联邦学习的后门攻击

纵向联邦学习[64-68]是一种独特的联邦学习框架，其特点是参与的客户端各自拥有不同的数据特征集，但却共享相同的样本索引集合。如图 5 所示，这种数据的纵向分割为攻击者提供了独特的机会，他们可能利用数据分割的特性和模型更新过程中的不对称性，来实施难以察觉的后门攻击。

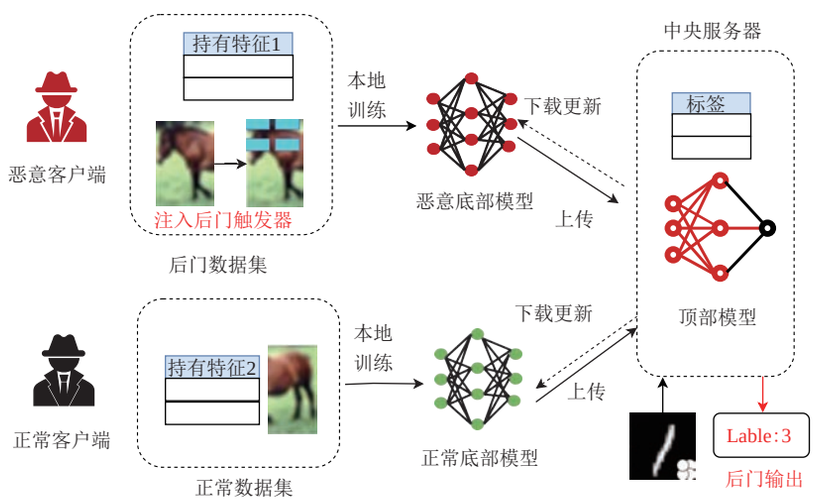

图 5 纵向联邦学习中后门攻击示意图

在纵向联邦学习环境中，拥有标签信息和部分数据特征的参与者被视作主动方，而仅掌握部分数据特征、缺失标签信息的参与者则被归类为被动方。鉴于主动方掌握标签信息，其实施后门攻击相对容易。研究普遍将被动方视为潜在攻击者，这无疑增加了攻击的难度。由于缺乏标签信息，被动方难以构建包含特定触发器和目标标签的后门样本。此外，由于无法获取标签信息，被动方无法通过调整

损失函数来实施攻击，这对攻击者造成了困难。

SplitNN[69-71]因其独特的模型架构而受到广泛关注。该框架设计中，客户端运行底部模型以提取本地数据的特征嵌入，而服务器则使用顶部模型进行分类决策。He等人[72]针对SplitNN框架下的纵向联邦学习提出了一种隐蔽的后门攻击策略DDPA，该策略首先获取一小部分目标类别标记样本并输入底部模型以获得目标类别的局部嵌入向量，再找出最相关的向量作为触发向量，来替换目标样本嵌入。

为弥补攻击者在样本标签上的缺失，逆向工程可用于重建标签信息。BadVFL[73]通过标签推理重构数据标签，并精心选择源类和目标类以嵌入后门触发器。该方法利用VFL数据按特征分割的特性，结合代理模型和敏感性分析设计触发器，分阶段实施攻击，从而在不直接访问训练标签的情况下成功植入后门。同样地，Gu等人[74]针对纵向联邦学习全局模型对客户端潜在表征的依赖性，提出LR-BA攻击。该方法利用少量带标签的辅助数据构建分类器，推断样本标签，并生成恶意表征，以增强后门标签的预测能力。此外，Liu等人[23]提出了被动方梯度替换攻击，攻击者通过接收的中间梯度信息推断标签，并利用目标标签样本的训练梯度替换后门样本的梯度，从而成功植入后门。

然而，现有后门策略常依赖于标签逆向技术来生成伪标签，在缺乏大规模数据支持的实际应用中存在局限。针对二分类任务，Chen等人[75]提出了UAB对抗性后门攻击方法，该方法利用VFL中边缘节点保留数据局部性的特点，通过插入通用触发器和执行无标签后门注入，在特定迭代中生成触发后门的通用触发器。UAB攻击不依赖额外数据，而是基于二元分类任务中的类别倾斜特性，推断少量非目标类别的伪标签。

此外，在纵向联邦学习中，客户端普遍采用加密技术保护私有信息，增加了攻击者推断标签信息的难度。然而，Liu等人[76]发现，即便在同态加密保护下的通信过程中，攻击者仍可高精度重建私有标签信息。攻击者通过训练梯度反转模型，在不直接访问或修改标签的情况下重建标签，并通过替换加密通信消息的方式植入后门。

4.3　面向联邦迁移学习的后门攻击

联邦迁移学习[77-81]允许模型在不同的领域或任务之间实现知识的迁移。该框架不仅提升了数据的利用效率，还增强了模型在多样化场景下的泛化能力。然而，联邦迁移学习跨领域的知识共享特性，同时也为攻击者提供了后门攻击机会。在这种框架下，攻击者面临的主要挑战是在模型迁移到新的环境和任务时，如何保持后门功能的稳定性。如图6所示，在联邦迁移学习场景下，后门攻击可能通过以下方式实现：下游用户将在源域数据上训练得到的、已被植入后门的模型应用于本地数据集，从而导致模型在特定输入下表现出攻击者预期的恶意行为。

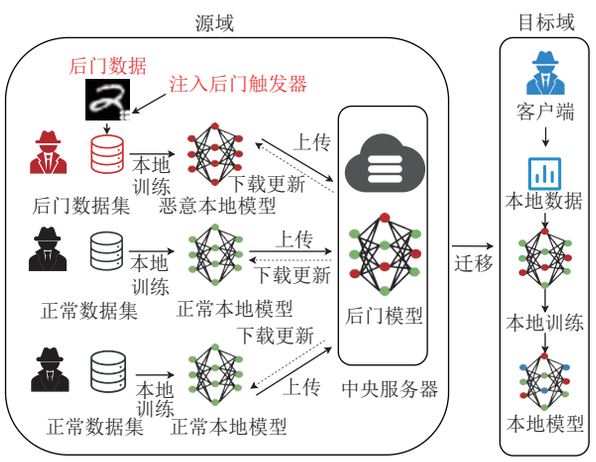

图6　联邦迁移学习中后门攻击示意图

Arazzi等人[82]认为联邦迁移学习通常分为两部分：由服务器在公开数据集上执行的特征提取器学习，以及由客户端使用私有本地数据执行的分类器学习。基于此，他们提出了一种名为FB-FTL的后门攻击方法。该方法使用Grad-CAM[83]技术来识别图像中对模型决策最为关键的区域。攻击者设定阈值，根据热图的响应强度创建掩码，明确了触发器需要注入的图像区域。随后采用数据集蒸馏技术，从目标后门类别中提取特征作为触发器。通过恶意客户端的模型训练实现后门的植入。

4.4　面向异构联邦学习的后门攻击

异构联邦学习（Heterogeneous Federated Learning，HFL）与上述几种框架不同，允许参与训练的各个客户端在数据分布、模型结构、通信能力和硬件资源等方面存在显著的异质性。这种全面的异质性为攻击者提供了更多的攻击面和潜在的攻击策略，但同时也增加了后门攻击的难度。例如，在异构联邦学习中，客户的数据可能存在标签不同、特征不同和质量不同等问题，攻击者可能需要对每个客户端定制后门。因此，对其进行后门攻击需要对这些异质性因素有深入的理解。

Li等人[84]的研究发现使用预训练的模型生成

用于联邦学习的公共数据集时，可能会引入安全漏洞，并提出了Fed-EBD。攻击者首先使用一个已经预训练好的后门基础模型生成有毒公共数据集，其中包括了正常数据与后门数据。然后通过联邦学习过程中将后门传递给各个客户端模型。

4.5 联邦后门攻击方法总结

本章对不同联邦学习架构下的后门攻击策略进行了深入分析，并通过表3详细比较和总结了各类后门攻击方法。此外，表4进一步汇总了不同攻击方法的实验评估，涵盖其适用的模型类型、实验数据集及攻击策略。同时，根据相应研究论文的原文介绍，本文在表4的“性能提升”一列中总结了各攻击方法在对应论文中主要关注和改进的方向，即该方法的优化重点在于提升后门攻击的有效性、隐匿性或持久性。

表3 联邦学习中后门攻击方法对比总结

模型框架	攻击方法	攻击方法		触发类型		修改标签		核心机制
		服务端	客户端	语义	人工	否	是	
横向联邦学习	Label Flipping[21]		√	√			√	篡改样本标签
	MUPA[40]		√	√			√	全局模型收敛后攻击
	ModelRe[41]		√	√	√		√	放大模型更新，实现模型替代
	Chameleon[48]		√	√			√	利用干扰与协助样本
	SemSBA[45]		√	√			√	利用对抗性扰动生成有毒样本
	SBA[44]		√	√			√	全局触发器与局部触发器的协同
	DBA[42]		√		√		√	全局触发器分解
	CBA[43]		√		√		√	全局触发器分解并优化
	Cerp[8]		√		√		√	恶意模型更新多维控制
	F3BA[47]		√		√		√	不重要模型参数符号反转
	Neurotoxin[49]		√		√		√	模型更新的稀疏性利用
	RNBA[50]		√	√			√	冗余神经元利用
	BCL[46]		√		√		√	模型关键层毒害
	IBA[51]		√		√		√	学习生成最优触发器
	A3FL[52]		√		√		√	对抗适应优化后门触发器
	DABS[53]	√			√	√		子网络替换
	Bapfl[60]		√		√		√	优化特征提取器
	pFedBA[9]		√		√		√	主任务和后门任务对齐
	PI-SBA[61]		√	√			√	辅助数据集生成
	HNTroj[84]		√		√		√	联邦超网络后门注入
纵向联邦学习	DDPA[72]		√		√	√		表征替换
	BadVFL[73]		√		√	√		样本标签逆向还原
	LR-BA[74]		√		√	√		表征替换
	GRA[22]		√		√	√		样本标签逆向还原
	GRA-HE[76]		√		√	√		样本标签逆向还原
	UAB[75]		√		√	√		通用触发器生成
联邦迁移学习	FB-FTL[82]		√		√	√		基于Grad-CAM的触发器设计
异构联邦学习	FED-EBD[85]		√		√		√	生成带有恶意触发器的合成数据集

从整体分析来看，联邦学习中的后门攻击主要围绕两个核心机制展开：(1) 数据特征操控，即通过设计触发器或篡改数据分布(如标签翻转)来改变模型的决策边界；(2) 模型更新干扰，包括参数替换、梯度扰动等策略，利用联邦学习的分布式聚合机制将恶意更新融入全局模型。然而，现有研究仍然存在局限性，主要体现在过于依赖简化假设(如固定攻击者比例、理想通信环境)，较少考虑动态防御机制及复杂网络拓扑对攻击效果的影响。

此外，不同联邦学习架构在数据处理方式和模型训练流程上的差异，不仅影响全局模型的整体性能，也为攻击者提供了多样化的攻击路径，同时增加了攻击实施的挑战。例如，在横向联邦学习中，由于各客户端数据特征高度一致，攻击者可以利用精心设计的样本诱导模型学习错误特征，从而在全局模型中成功植入后门。而在纵向联邦学习中，由于不

表4　联邦学习中后门攻击方法实验对比总结

攻击框架	攻击方法	模型结构	数据集	数据分布	防御方法出处	性能提升
横向联邦学习	Label Flipping[21]	CNN	CIFAR10, Fashion-Mnist	IID	文献[21]	E
	MUPA[40]	CNN	EMNIST	Non-IID	文献[40]	E,S
	ModelRe[41]	ResNet, LSTM	CIFAR10, Reddit	Non-IID	未提及	E,P,S
	Chameleon[48]	ResNet, VGG	CIFAR10, CIFAR100, EMNIST	Non-IID	文献[40,86,87]	E,P
	SemSBA[45]	WideResNet	CIFAR10, CIFAR100	IID	未提及	E
	SBA[44]	ResNet, MLP	CIFAR10, MNIST, Tiny-ImageNet	Non-IID	文献[88,89]	E,P,S
	DBA[42]	ResNet, CNN	LOAN, MNIST, Cifar10, Tiny-ImageNet	Non-IID	文献[88,89]	E,P,S
	CBA[43]	ResNet, LeNet	CIFAR10, MNIST	Non-IID	文献[89]	E,P
	Cerp[8]	ResNet	CIFAR100, Fashion-Mnist, LOAN	Non-IID	文献[86,87,89-97]	E,P,S
	F3BA[8]	ResNet, CNN	CIFAR10, Tiny-ImageNet	Non-IID	文献[90,94,98-102]	E,P, S
	Neurotoxin[49]	ResNet, LSTM, LeNet	CIFAR10, EMNIST, Reddit, Sentiment140	Non-IID	文献[40,86,90,91]	E,P
	RNBA[50]	CNN, FCNN	Retina Disease Detection, Face Recognition	IID, Non-IID	文献[89,92,96,103-105]	E,P,S
	BCL[46]	ResNet, CNN, VGG	CIFAR10, Fashion-Mnist	Non-IID	文献[86,87,92,101,106,107]	E,P,S
	IBA[51]	ResNet, LeNet, VGG	CIFAR10, MNIST, Tiny-ImageNet	Non-IID	文献[40,86-89,101]	E,P,S
	A3FL[52]	ResNet, LeNet, VGG	CIFAR10, MNIST, Tiny-ImageNet	Non-IID	文献[40,86,87,90,91,94,98,99,101,102,108]	E,P,S
	DABS[53]	VGG	CIFAR10, Tiny-ImageNet	IID, Non-IID	未提及	E
	Bapfl[60]	VGG, CNN	CIFAR10, MNIST, Tiny-ImageNet	Non-IID	文献[86,91]	E
	pFedBA[9]	ResNet, LeNet, CNN	CIFAR10, CIFAR100, Fashion-MNIST, N-BaIoT	Non-IID	文献[40,86,87,91,93,109]	E,P
	PI-SBA[61]	AlexNet	CIFAR10, MNIST, YouTube Aligned Face	Non-IID	文献[40,110]	E,S
	HNTroj[84]	LeNet, CNN	CIFAR10, Fashion MNIST	Non-IID	文献[40]	E,S
纵向联邦学习	DDPA[72]	SplitNN	CIFAR10, Epsilon, Criteo, CovType	Non-IID	文献[111]	E,S
	BadVFL[73]	ResNet, CNN	CIFAR10, CIFAR100, CINIC-10, Criteo	Non-IID	文献[40]	E
	LR-BA[74]	ResNet, CNN	CIFAR10, CIFAR100, NUS-WIDE, CINIC-10, BHI	Non-IID	文献[40,95,112]	E
纵向联邦学习	GRA[22]	未提及	NUS-WIDE, MNIST	Non-IID	文献[40,112]	E,P,S
	GRA-HE[76]	ResNet, MLP	CIFAR10, CIFAR100, NUS-WIDE, MNIST	Non-IID	文献[40,112,113]	E,P,S
	UAB[23]	MLP	Zhongyuan, Lending-Club	Non-IID	文献[113-115]	E,S
联邦迁移学习	FB-FTL[82]	ResNet, VGG, CNN	CIFAR10, CINIC-10, SVHN, GTSRB	Non-IID	文献[86,87,89,91,116]	E,P
异构联邦学习	FED-EBD[85]	ResNet, DistilBERT	CIFAR10, AG-News, SST-2	IID, Non-IID	未提及	E,S

注：E代表有效性(Effectiveness)，S代表隐匿性(Stealthiness)，P代表持久性(Persistence)。

同参与方的数据维度不同，攻击者可能会利用逆向还原技术，分析数据关联性推测并赋值标签，从而实现有效攻击。对于联邦迁移学习，攻击的迁移性至关重要，攻击者需要确保后门在不同任务和场景中仍然有效，以提升攻击成功率。最后，在异构联邦学习中，由于客户端可能使用不同的设备、数据格式或模型架构，攻击者必须设计具有跨环境适应性的后门，以增强攻击的隐蔽性和通用性。

综上所述，不同联邦学习架构的特点为后门攻击提供了多样化的实施路径。攻击者需要根据架构特点定制攻击策略，以最大化攻击成功率，同时规避已有的防御机制。

5 联邦学习后门防御方法

与集中式学习范式相比，联邦学习的分布式特性使其更容易受到灵活多变的外部攻击者的威胁。后门攻击的隐蔽性，尤其是当触发器信息对防御者未知时，使得仅依赖常规样本测试来检测模型中的后门极为困难，从而给防御工作带来了前所未有的挑战。然而，在联邦学习系统中，防御者可以利用更丰富的信息资源来构建和优化防御机制，例如来自多个客户端的训练中间信息或单个客户端的连续历史信息。因此，本文依据防御方法所依赖的客户端信息来源，将其划分为基于多客户端信息和基于单客户端信息的两类防御策略。

值得注意的是，无论联邦学习采用横向、纵向、迁移学习或异构架构，由于其本质上都涉及多个客户端的协作，防御策略的分类主要取决于信息来源的数量，而非具体的联邦学习架构。因此，这种基于客户端信息的分类方法具有普适性，能够适用于不同类型的联邦学习系统。

5.1 基于多客户端信息的后门防御

如图7所示，基于多客户端信息的后门防御方法主要通过分析多个客户端上传的中间结果来检测并消除后门，从而降低后门攻击的影响。根据所依赖的技术机制，本文进一步将其细分为单体系防御方法和多体系防御方法，以更系统地总结现有的后门防御策略。

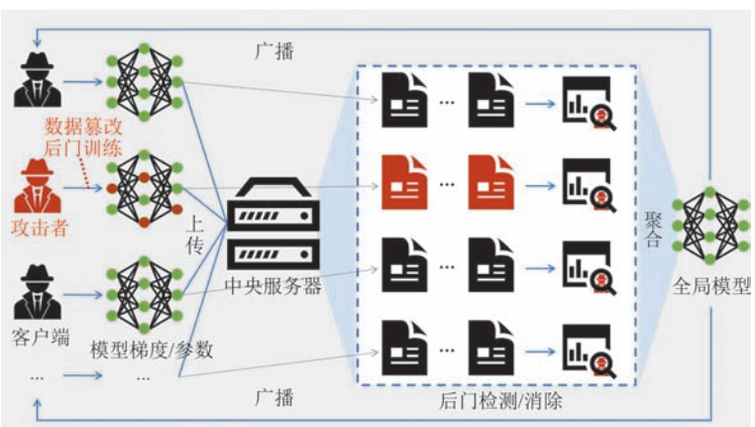

图7 基于多客户端信息的后门防御

5.1.1 单体系防御

单体系防御指的是服务器运用单一核心技术机制，对来自众多客户端的信息进行汇总、分析和检测，以有效防御后门攻击。这种策略因其实现和维护的简便性，成为联邦学习系统的基础防御措施。

Krum[85]是一种基于平方距离的异常检测方法，用于联邦学习中的鲁棒聚合。计算每个客户端的模型更新与其他客户端更新之间的欧式距离之和，作为异常得分，并选择异常得分最小的一个客户端的模型更新作为该轮全局模型的模型更新。Multi-Krum[87]在Krum的基础上进行了扩展。在计算完异常得分后，该方法选择得分较低的n个模型更新，并对其进行平均，以此作为全局模型的最终更新。

Median[91]和Trimmed-Mean[92]在针对模型参数的坐标级别进行聚合。在每轮聚合时，对客户端提交的每个模型参数维度的更新值取中位数，以此作为全局模型在该维度的更新。而Trimmed-Mean算法则在每轮聚合时，去除每个维度参数中的β个极值，然后计算剩余参数的均值作为聚合结果。

Bulyan[90]结合了Krum和Trimmed-Mean的优点，首先利用Krum筛选出得分最低的$n-2m$个客户端更新(n表示参与的客户端数量，m表示估计的恶意客户端数量)，再应用Trimmed-Mean进行聚合，以增强对恶意更新的抵抗力。

对于多客户端提交的本地模型更新，结合部分参数的视角进行分析是检测后门模型的有效方法。Liu等人[117]提出了一种融合模型层参数与整体参数检测的后门防御方法，称为PnA。在聚合阶段，对于每个客户端的模型更新，计算与其他客户端相同层参数更新之间的欧式距离和以及全部参数更新之

间的欧式距离和。对于层参数距离和与全部参数距离和，进行归一化，并统计异常值的数量。最后，选择异常值数量最小本地模型更新进行均值聚合。

利用额外的干净数据集来指导模型训练的方向是一种防御思路。FLTrust[92]采用了一种基于信任分数的方法，通过维护一个辅助模型来指导鲁棒聚合。服务器使用一个干净数据集训练辅助模型，通过计算辅助模型更新与客户端更新的相似度，作为信任分数。根据信任分数对客户端的模型更新进行加权平均，更新全局模型。然而，依赖于额外的干净数据集的做法在一定程度上会影响模型的泛化性。

在联邦训练过程中运用差分隐私(Differential Privacy，DP)也可以实现干扰后门攻击的效果。通过添加噪声，实现减少后门模型更新与正常更新之间的差异，破坏模型中的后门。FedCDP[96,118]要求服务器端对聚合后的全局模型注入随机噪声。Sun等人[97]研究了弱差分隐私的后门防御方法，表明通过向模型中添加少量高斯噪声，可以限制后门攻击。

基于多客户端的模型更新，在训练中动态调整学习率可以实现对后门攻击的防御。Safa等人[40]提出的RLR方法通过识别客户端模型更新数值和方向的差异，调整聚合时不同模型更新的学习率，以抵御后门攻击。RLR要求对全局模型更新的每个维度进行符号总和的计算。当某维度的符号总和低于阈值θ时，该维度的学习率被设为-1。这种方法的优势在于其与各种聚合算法的兼容性，提供了一种灵活的防御策略。

差异测试是一种软件测试技术，通过比较两个或多个程序在相同输入下的输出结果，来识别它们之间的行为差异。Gill等人[119]借鉴这一技术，提出了名为FedDefender的联邦后门防御方法。FedDefender基于一种假设：在联邦学习环境中，基于相似数据训练出的本地模型应该具有近似的神经元激活模式，如果本地模型在给定输入下的激活模式与其他模型显著不同，那么模型很可能植入了后门。该方法通过生成随机测试样本，收集每个本地模型的神经元激活值，根据激活值的偏差计算恶意分数。当恶意分数超过预设阈值，则忽略该本地模型的贡献；当客户端的恶意分数低于阈值则通过调整聚合权重来减少其潜在影响。

基于博弈分析，Jia等人[120]提出了一种在服务器端的交互式联邦后门防御方法FedGame。FedGame将后门问题构建为最大最小博弈。服务器利用模型更新创建辅助模型，用于逆向识别可能的后门触发器和目标类别。对于每个模型更新，根据其对特定触发器的响应获得得分，并根据得分调整聚合过程中各客户端贡献的权重。

概率分布与检测[121]的概念在数据分析和机器学习中被广泛应用，用于识别和区分正常与异常行为。Kumari等人[122]基于此提出了一种名为BayBFed的防御方法。BayBFed通过贝塔过程表示每个客户端模型更新概率，对于每个客户端生成一个层次化的贝塔过程先验，用于生成客户端的模型更新。将模型更新的每一维参数都认为是独立的伯努利过程，获取当前轮次的贝塔后验，来决定权重是否更新。检测步骤是通过一种自适应的中餐厅过程来进行检测和过滤恶意的本地模型更新。

选举机制也能够与联邦学习后门防御相结合。Qin等人[123]提出了Snowball框架，一种结合两种选举策略的服务器防御方法。Snowball框架的聚合阶段融合了两种选举策略。自底而上策略允许客户端基于k-means选择与本地模型更新最相似的更新，形成候选集。自顶而下策略则关注候选集中模型更新的差异，通过计算模型更新两两之间的差值，构建差异集合。利用差异集合训练变分自编码器，以识别并选择重构误差最小的模型更新，并将其加入到候选集中。重复进行，直到候选集中的模型更新数量达到预设目标，再聚合模型。

持久同调理论(Persistence Homology)是一种代数拓扑工具，用于分析拓扑空间中的持久特征。Ma和Gao[124]基于这一理论，设计了一种部署在服务器的检测方法，PH-Detect。根据输入样本，计算本地模型每一层的激活值，生成距离矩阵。对于每一层的激活值矩阵，构造对应的Vietoris-Rips复形。将每个单纯复形出现和消失的时间作为点对添加到持久图中。在归一化后进行主成分分析或其他降维技术提取主要特征，并训练一个分类器，以判断客户端的本地模型是否恶意。

集成学习技术在处理数据时，能够有效忽略少数异常结果的影响。Cao等人[125]基于集成学习提出了一种名为FLCert的联邦防御方法。FLCert将所有客户端按照特定或随机的方式分组。每个组独立运用联邦学习算法，分别训练出自己的全局模型。在预测阶段，每个小组的全局模型独立进行预测并记录结果。最终，通过统计各模型预测结果的频率，选择出现最频繁的标签作为最终预测。

利用对比学习的方法区分干净模型与后门模型的特征也是一种防御后面攻击的思想。GANcrop[126]

通过对比学习的思想检测本地客户端中的后门模型。通过生成式对抗网络(Generative Adversarial Networks,GANs)[127]恢复后门触发器,并利用这些触发器重新训练模型以消除后门攻击的影响。这种通过对比学习的方法利用多个客户端的信息进行后门检测的难点是需要有效地构建正负样本对,才能够有效地实现后门检测。

BackdoorIndicator[128]是一种基于离群样本测试的联邦后门防御方法。服务器在广播全局模型前,利用一个离群样本集通过训练将一个指示任务注入到全局模型中。对于每个经过客户端本地训练的模型更新,替换批量归一化层的值,并评估指示任务的准确性。将指示任务准确性高的模型更新标记为可疑,并将其从聚合的模型更新列表中排除。

Cao等人[129]基于重构思想提出了一种纵向联邦框架下针对预测阶段后门样本的防御方法,称为VFLIP。在训练阶段,服务器通过利用其他客户端提交的嵌入向量训练一个掩蔽自编码器,实现对客户端嵌入的预测。在预测阶段,对于每个客户端的嵌入向量,基于预测嵌入与实际嵌入的欧氏距离,计算异常得分。基于训练阶段的数据分布构建阈值,并基于该阈值识别并去除异常嵌入。

FedPD[130]是一种基于原型学习的联邦学习后门防御方法。这种方法要求服务端为全局模型生成全局原型并进行广播,客户端需要根据全局原型在本地数据集上训练模型,并为本地模型生成对应的本地原型,上传给服务端。服务端通过计算相同类别本地原型的相似性与全局原型的相似性,更新全局原型。在客户端训练本地模型的过程中补充一个对比损失,通过这个损失约束在本地训练中迫使来自客户的每个样本接近其类别的全局原型,同时远离其他课程的全局原型,实现后门防御。

5.1.2 多体系防御

多体系防御方法整合多种技术策略来检测和防御后门攻击。与单体系防御方法相比,多体系防御涉及的技术更加全面和复杂,包括但不限于过滤、裁剪、加噪以及聚类等多种策略。这些策略先后作用于来自多个客户端的信息,以消除后门攻击的影响。

FLAME[87]采用过滤、裁剪和加噪三个关键步骤来抵御后门攻击。计算所有本地模型更新之间的余弦距离,然后利用HDBSCAN算法识别并过滤恶意模型更新。设定本地模型更新的中位数作为剪切阈值,动态地裁剪每轮过滤后的模型更新。通过估计本地模型与全局模型之间的距离来确定噪声的大小,进而向聚合后的全局模型中添加噪声。

DeepSight[102]是一种结合了过滤和裁剪的联邦后门防御方法。评估模型间的差异度、识别训练数据相似的本地模型以及全局模型输出层之间的余弦距离,作为特征进行聚类,并通过聚类结果与分类器,衡量训练数据的同质性。同时考虑分类器的结果与聚类结果,构建本地模型更新的候选集合。通过计算本地模型更新L2范数的中值,动态调整那些超过中值的模型更新,以缩小其影响。采用FedAvg对裁剪后的同一集群内的本地模型更新进行聚合,并在聚合后将结果发送回相应集群内的客户端。

RFBDS[131]是一种包括稀疏化、聚类和自适应裁剪三个步骤的后门防御方法。对于每个模型更新,通过保留每层参数绝对值的最大值和其符号的乘积(称为主要梯度)来进行稀疏化。通过计算主要梯度间的欧氏距离,选择将前k个主要梯度的中位数作为阈值,将阈值作为正常客户端模型更新的界限、k作为最小正常客户端的数量进行Optics聚类[132],得到聚类结果I。通过计算每个模型更新的L2范数,进行排序并选择最小的$|I|$个本地模型,实现对模型更新集合的裁剪。将裁剪后的模型更新集合进行均值聚合得到该轮的全局模型。

RoseAGG[133]是一种包括自适应部分聚合、干净成分提取和信任分计算三个步骤的联邦后门防御方法。通过正则化只考虑模型更新的方向,实现数值的自适应,以消除更新幅度不同引起的偏差。利用DBSCAN[134]聚类将模型更新按照方向进行分组,并进行均值聚合。对不同组的聚合结果再次归一化,形成一个更新矩阵。通过PCA对更新矩阵进行提取,以获取干净的成分。计算各组聚合结果在主成分上的投影距离,作为信任分,并将信任分作为权重进行加权平均,以获取全局模型更新。

MESAS[135]是一种结合多种指标检测的服务端联邦后门防御方法。MESAS需要计算六个指标,包括本地模型参数之间的欧氏距离、余弦距离、本地训练后模型参数数值增加的参数数量、本地模型参数的方差、本地模型中所有参数之间的最大距离和最小距离。通过确定每个指标的中值,检测计算出的客户端六个指标是否在其中值周围均匀分布,并检查大于中值的指标值与小于中值的指标值是否遵循相同的分布进行过滤,并利用欧式距离对经过过滤的指标进行两个簇的聚类。最终将较大

的集群作为正常的模型更新，裁剪剩余数据，进行聚合。

Lockdown[136]是一种基于孤立子空间训练和剪枝的联邦学习后门防御方法。该方法通过随机图模型为每层参数分配不同的稀疏性，使得参数较多的层具有更高的稀疏性，并为每个客户端进行相同的随机子空间初始化，使用二进制掩码表示模型参数是否存在于子空间中。客户端仅在其子空间内进行本地更新，通过动态剪枝裁去子空间内梯度较小的参数，并恢复梯度较大的参数。服务器则比较客户端子空间，去除出现频率较低的参数。Lockdown通过客户端和服务器两阶段剪枝有效去除恶意参数。

CRFL[94]是一种包括参数裁剪、参数扰动和参数平滑的可证明联邦后门防御框架。服务器通过裁剪全局模型并加入随机的高斯噪声，来破坏恶意本地模型中的后门。在推理阶段，服务器通过添加高斯噪声的形式进行参数平滑，并基于最终的全局模型生成多个噪声模型。对于每个输入样本，通过计算多个噪声模型的结果，得到相应的经验估计。最终使用Hoeffding不等式计算经验估计的边界，以确定预测的置信度。此外，参数平滑能够为每个测试样本提供一个可证明半径，以表明模型预测能够保持稳定和一致的扰动范围。

CrowdGuard[137]通过余弦距离和欧式距离结合统计显著性检验和迭代剪枝筛选异常模型，并对客户端投票结果进行聚类，过滤恶意反馈。该方法要求服务端将本地模型分发给其他客户端进行交叉验证，客户端在可信执行环境下推理并记录隐藏层输出，进行后门检测。最终，服务器根据投票结果在聚合前移除可疑本地模型。

FLPurifier[138]是一种结合本地训练阶段对比学习和聚合阶段动态过滤的联邦学习后门防御方法。客户端将本地模型解耦为编码器和分类器：编码器利用无标签数据和非对称Siamese网络进行对比学习，并通过类似触发器的数据增强方法破坏恶意标签与后门触发器的关联；分类器则在固定编码器参数的基础上利用本地数据进行训练。服务器通过余弦相似度对分类器进行加权平均，从而削弱恶意更新的影响。

5.2　基于单客户端信息的后门防御

基于单客户端信息的后门防御方法提供了一种独特的安全防护视角。与基于多客户端信息后门防御方法不同，如图8所示，这种方法专注于分析单个客户端的本地模型更新，以识别和防御后门攻击。无论是由服务器还是客户端实施，这类方法都利用了单个客户端的信息来增强系统的安全性。尽管联邦学习的优势在于多客户端的协作，但基于单客户端的防御策略突显了单一数据源在安全检测中的潜力。这种方法以其独到的优势和效果，为联邦学习的安全架构增添了一层额外的保护。

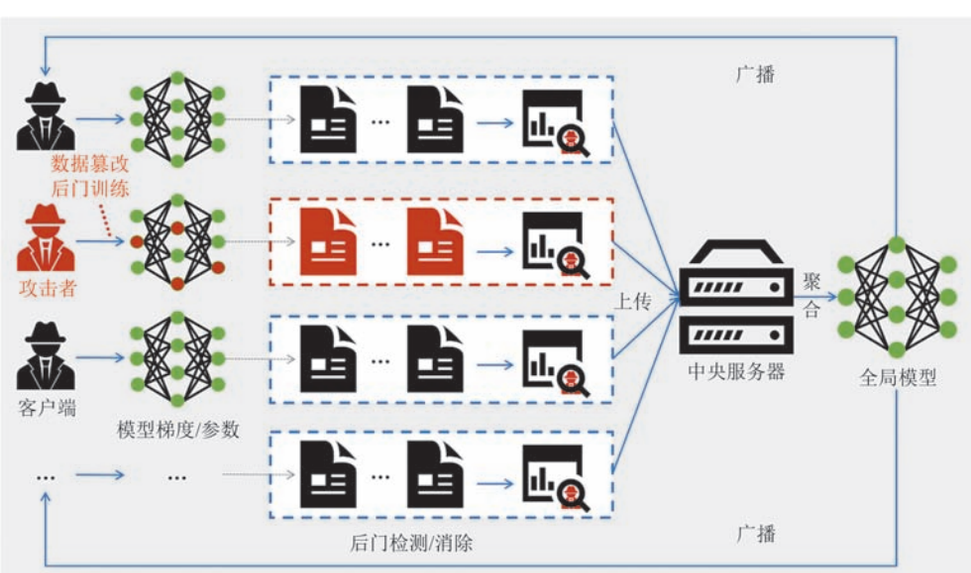

图8　基于单客户端信息的后门防御

基于单客户端信息也同样可以结合差分隐私技术实现后门防御。FedLDP[139]要求客户端使用差分隐私随机梯度下降(DP-SGD)[140]在本地数据集上进行训练。这种训练方式在计算模型更新时引入由差分隐私的预算参数决定的随机噪声，实现对客户端隐私的保护，并且在一定程度上能够抵御后门攻

击。Naseri等人[141]通过实验验证了FedCDP与FedLDP在缓解后门攻击方面的有效性。但由于添加的噪声是随机的，后门攻击的防御效果与联邦训练的稳定性可能存在冲突。

FLIP[107]是一种在客户端进行部署的联邦后门防御方法。通过对每个类的数据进行触发器优化，以逆向出将正常标签数据翻转为目标标签的最小输入（通用触发器），并获取通用触发器作用的类间距离排序。根据是否具有最有可能存在触发器的源类和目标类（类间距低）的数据，进行双向触发器优化或单向触发器优化。在推理阶段，通过设定阈值来过滤掉置信度较低的样本，从而有效地排除那些可能包含嵌入触发器的后门样本。

FLDetector[106]是一种基于历史模型更新一致性的检测方法，来检测投毒攻击与后门攻击中的恶意客户端。基于单客户端在多轮联邦训练中上传的模型更新的历史信息，计算当前轮客户端的模型更新与之前每轮模型更新的差值，构成蕴含本地训练历史信息的矩阵，并利用L-BFGS算法近似Hessian矩阵，估计当前轮客户端的模型更新。通过计算估计的模型更新与实际的模型更新之间的欧式距离，进行归一化，并将r轮训练之后的欧氏距离的均值作为可疑分数。基于可疑分数进行k-means聚类，将客户端聚成两个集群，将处于平均可疑分较大集群内的客户端视为恶意客户端。

Sun等人[97]基于恶意客户端对参数攻击效果隐藏于Hessian矩阵核内的发现，提出了一种名为FL-WBC的客户端后门防御方法。基于单客户端联邦训练过程中的信息，比较连续两轮本地训练的梯度变化来估算损失函数的二阶偏导数（Hessian矩阵）。研究发现在大多训练轮次中，受到后门攻击的全局模型的Hessian矩阵是高度稀疏的。基于这个发现，FL-WBC要求当Hessian矩阵是高度稀疏时，在本地模型参数的Hessian矩阵的对角线上小幅度元素（接近于0的元素）的位置添加均值为0，标准值为s的拉普拉斯噪声，以扰动Hessian的零空间，实现防御的效果。

类似地，Zhu等人[142]基于Hessian矩阵估计攻击效应，提出了一种部署在客户端的联邦后门防御方法，称为LeadFL。LeadFL在客户端模型的本地训练过程中引入了一个旨在减少Hessian矩阵稀疏性的正则化项。通过计算连续两轮训练中梯度的差值，基于差值与学习率的比值近似Hessian矩阵对角线上的数值，并利用这些数值来构造正则化项。为了保持模型更新过程的稳定性，LeadFL还引入了梯度裁剪技术来控制正则化项的范围，确保其不会超过设定的阈值。

MCFL[143]是一种基于模式连接性的客户端后门防御机制。在客户端将训练的本地模型作为曲线的一端，将服务端的全局模型作为曲线的另一端。通过优化过程找到曲线的中间点参数，利用插值的方法计算曲线上的点，并使用本地数据计算损失梯度，更新中间点参数。最终，客户端利用定义的曲线参数值从训练好的曲线上恢复模型的权重，并作为本地模型参数，发送给服务端参与全局模型更新。

5.3 联邦学习后门防御方法总结

本章对联邦学习系统中针对后门攻击的防御方法以及对应的实验进行了分析和总结，并在表5中精炼概括了这些方法的特点与核心机制；在表6中统计了防御方法对应实验使用的联邦范式、数据集、数据分布以及评估的攻击方法。近些年大多联邦后门防御方法都会在横向联邦学习的范式下对图像分类任务进行实验验证，且通常会考虑Non-IID的数据。而对于其他结构的联邦学习框架以及序列化数据的任务和模型，则较少进行实验上的验证。

本文将基于多客户端信息的后门防御分为多体系防御和单体系防御。单体系防御依赖核心技术机制，简单但可能不足以应对复杂多变的攻击。多体系防御融合多种技术（如聚类、加噪、裁剪）实现更全面的防御，但在实际部署中面临三个挑战：一是结构复杂、部署难度高，尤其在多中心服务器的复杂联邦系统中；二是可能影响训练稳定性，降低全局模型的泛化性；三是不同技术间的相互作用尚未充分研究，限制了防御设计理论的深入理解与后续优化的灵活性。

基于单客户端信息的防御方法利用客户端自身的数据和训练历史来识别和减轻后门攻击。这类方法的优势在于可以不依赖其他客户端的信息，能够独立地减少后门攻击的影响。然而，由于客户端设备算力通常受到限制，效率往往是这类方法在实际应用中面临的一个挑战。另一个挑战是基于单客户端信息的防御方法对全局模型的影响力有限。

表5　联邦学习中后门防御方法对比总结

方法类别	后门防御方法	部署位置		作用时间		实验数据分布		核心机制
		服务端	客户端	聚合前	聚合后	IID	Non-IID	
多客户端信息：单体系防御	Krum[86]	√		√		√		根据模型更新距离进行一次筛选
	Multi-Krum[86]	√		√		√		根据模型更新距离进行多次筛选
	Median[91]	√		√		√		每个维度的中位数作为结果
	Trimmed-Mean[91]	√		√		√		剪切模型更新每个维度的尾部值
	Bulyan[90]	√		√		√		Krum ＋Trimmed-Mean
	PnA[117]	√		√		√		检测网络各个层参数的异常程度
	FedCDP[141]	√		√			√	全局模型注入随机噪声
	FLTrust[92]	√		√			√	维护干净数据集的模型作为标杆
	RLR[101]	√		√		√		根据参数差异动态变化聚合的学习率
	FedDefender[119]	√		√		√		构建随机样本进行差异测试
	FedGame[120]	√		√		√		构建辅助全局模型，逆向触发器
	BayBFed[122]	√		√			√	动态表示本地模型分布，聚类筛选
	PH-Detect[124]	√		√			√	构建拓扑空间的特征进行降维分类
	Snowball[123]	√		√			√	对多中心聚类交集再进行差值聚类
	FLCert[125]	√			√		√	集成学习
	GANcrop[126]	√		√			√	利用逆向GAN触发器
	BackdoorIndicator[128]	√		√			√	在全局模型中构建指示任务
	VFLIP[129]	√			√		√	训练中构建MAE模型学习嵌入特征
	FedPD[130]	√		√		√		利用全局原型作为全局知识纠正本地训练
多客户端信息：多体系防御	FLAME[90]	√		√	√	√	√	聚类＋裁剪＋加噪
	DeepSight[102]	√		√		√	√	聚类＋裁剪＋均值聚合
	RFBDS[131]	√		√			√	稀疏化＋聚类＋均值聚合
	RoseAGG[133]	√		√			√	聚类＋提取主成分＋计算聚合权重
	MESAS[135]	√		√		√	√	多指标统计测试＋分布检验＋聚类
	Lockdown[136]	√		√		√	√	构建子空间＋客户端剪枝＋服务器剪枝
	CRFL[94]	√		√	√	√	√	裁剪＋加噪＋推理认证
	CrowdGuard[137]	√	√	√		√	√	距离指标＋显著性检验＋剪枝＋聚类
	FLPurifier[138]	√	√	√		√	√	对比学习＋分类器加权平均
单客户端信息	FedLDP[140]		√	√			√	本地模型更新注入随机噪声
	FLIP[107]		√		√		√	触发器逆向＋对抗训练＋样本过滤
	FLDetector[106]	√		√			√	估计本地模型更新，进行相似度判断
	FL-WBC[97]		√	√		√	√	客户端在本地Hessian矩阵中添加噪声
	LeadFL[97]		√	√		√	√	本地训练添加正则化项
	MCFL[143]		√	√			√	基于模式连接性重构模型参数

注：聚合前指在生成全局模型之前，对客户端或其模型参数进行筛选、加噪等操作；聚合后 指在全局模型生成后，对其实施裁剪、加噪等处理，或在推理阶段进行相应操作。＋表示两种机制或思想的结合使用。

表6　联邦学习后门防御方法实验内容对比总结

方法类别	防御方法	联邦范式	模型结构	数据集	数据分布	评估攻击方法
多客户端信息：单体系防御	Krum[86]	HFL	MLP，CNN	Spambase，MNIST	IID	未提及
	Multi-Krum[86]	HFL	MLP，CNN	Spambase，MNIST	IID	未提及
	Median[91]	HFL	LR，CNN	MNIST	IID	文献[21]
	Trimmed-Mean[91]	HFL	LR，CNN	MNIST	IID	文献[21]
	Bulyan[90]	HFL	FCN，CNN	MNIST，CIFAR-10	IID	未提及
	PnA[117]	HFL	CNN，VGG	MNIST，Fashion-MNIST，CIFAR-10	Non-IID	文献[21，42，93，144]

续表

方法类别	防御方法	联邦范式	模型结构	数据集	数据分布	评估攻击方法
多客户端信息：单体系防御	FedCDP[141]	HFL	RNN, LSTM, ResNet	EMNIST, CIFAR-10, Reddit, Sentiment140	Non-IID	文献[41,141]
	FLTrust[92]	HFL	LR, CNN, ResNet	MNIST, Fashion-MNIST, CH-MNIST, CIFAR-10, HAR	Non-IID	文献[21,41]
	RLR[101]	HFL	CNN	Fashion-MNIST, Federated EMNIST	IID, Non-IID	文献[42]
	FedDefender[119]	HFL	CNN	MNIST, Fashion-MNIST	IID	文献[119]
	FedGame[120]	HFL	CNN, ResNet	MNIST, CIFAR-10	IID, Non-IID	文献[41,42,49]
	BayBFed[122]	HFL	CNN, LSTM	MNIST, Fashion-MNIST, CIFAR-10,Reddit, IoT	IID, Non-IID	文献[41,145]
	PH-Detect[124]	HFL	LeNet, ResNet, STN	MNIST, CIFAR-10, GTSRB	Non-IID	文献[42,145,146]
	Snowball[123]	HFL	CNN, RNN	MNIST, Fashion-MNIST, Federated EMNSIT, CIFAR-10, Sentiment140	IID, Non-IID	文献[43,44]
	FLCert[125]	HFL	FCN, CNN, ResNet, LSTM	MNIST, CIFAR-10, HAR, Reddit	IID, Non-IID	文献[41,147]
	GANcrop[126]	HFL	ResNet	CIFAR-10	Non-IID	文献[148]
	BackdoorIndicator[128]	HFL	VGG, ResNet,	CIFAR-10, CIFAR-100, EMNIST	IID, Non-IID	文献[42,48,49,147]
	VFLIP[129]	VFL	FCN, VGG	CIFAR-10, CINIC10, Imagenet, NUS-WDE, Bank Marketing	Non-IID	文献[73,150]
	FedPD[130]	HFL	CNN, ResNet	MNIST, Fashion-MNIST, CIFAR-10	IID, Non-IID	文献[24,147,151]
多客户端信息：多体系防御	FLAME[90]	HFL	RNN, LSTM, ResNet	MNIST, CIFAR-10, Tiny-ImageNet, IoT-Traffic, Reddit	IID, Non-IID	文献[41,42,145]
	DeepSight[102]	HFL	RNN, LSTM	MNIST, CIFAR-10, IoT, Reddit	IID, Non-IID	文献[21,145]
	RFBDS[131]	HFL	未提及	MNIST, Fashion-MNIST, CIFAR-10	IID, Non-IID	文献[42]
	RoseAGG[133]	HFL	未提及	Fashion-MNIST, EMNIST, CIFAR-10, CIFAR-100	Non-IID	文献[8,41,42,145]
	MESAS[135]	HFL	CNN, ResNet, SqueezeNet, DistilBERT	MNIST, CIFAR-10, GTSRB, SST-2	IID, Non-IID	文献[21,41,145,147,152]
	Lockdown[136]	HFL	LeNet, ResNet	Fashion-MNIST, CIFAR-10, CIFAR-100, Tiny-ImageNet	IID, Non-IID	文献[41,42,49,147,151]
	CRFL[94]	HFL	LR	MNIST, EMNIST, LOAN	Non-IID	文献[41,145]
	CrowdGuard[137]	HFL	CNN, VGG, DenseNet, ResNet	MNIST, CIFAR-10	IID, Non-IID	文献[41]
	FLPurifier[138]	HFL	CNN, ResNet	MNIST, Fashion-MNIST, EMNIST, CIFAR-10, CIFAR-100, GTSRB	IID, Non-IID	文献[41,42,145]
单客户端信息	FedLDP[140]	HFL	RNN, ResNet, LSTM	EMNIST, CIFAR-10, Reddit, Sentiment140	Non-IID	文献[41,42],
	FLIP[107]	HFL	FCN, ResNet	MNIST,Fashion-MNIST	Non-IID	文献[41,42]
	FLDetector[106]	HFL	CNN, ResNet	MNIST, Federated EMNIST, CIFAR-10	Non-IID	文献[41,42,144]
	FL-WBC[97]	HFL	CNN	Fashion-MNIST, CIFAR-10	IID, Non-IID	文献[153]
	LeadFL[97]	HFL	CNN	Fashion-MNIST, CIFAR-10, CIFAR-100	IID, Non-IID	文献[154]
	MCFL[143]	HFL	CNN	MNIST, Fashion-MNIST, Federated EMNIST, CIFAR-10	IID, Non-IID	文献[41,42,145]

注：LR表示逻辑回归模型；MLP表示多层感知机；CNN表示由卷积层和全连接层组成的卷积神经网络。

6　应用场景中联邦后门攻击与防御

联邦学习作为一种隐私保护的分布式机器学习范式，在多个应用场景中展现出巨大潜力。然而，随着应用的广泛，联邦后门攻击的威胁也日益突出，严重影响系统的安全性。不同场景下的后门攻击差异显著，主要源于数据特性、模型架构和应用需求的不同。例如，在联邦推荐系统中，后门攻击可能导致特定物品获得额外的曝光机会，从而削弱推荐系统的公正性和准确性。这种攻击不仅会影响用户体验，还可能被恶意利用进行商业竞争或舆论操控。在联邦图学习领域，后门攻击可能扭曲关键节点或边的连接关系，进而影响图算法的预测准确性和可靠性。考虑到图数据在社交网络分析、生物信息学等关键领域的广泛应用，此类攻击可能带来严重影响。在联邦物联网中，后门攻击可能引发广泛的系统故障和数据泄露，对智能家居、自动驾驶等关键领域的安全性构成直接威胁。这些场景下的安全漏洞可能导致重大经济损失甚至危及人身安全。本章将深入探讨与分析这些应用场景中的后门攻击方法与防御策略。通过对比分析不同场景下的攻击特征和防御需求，期望为联邦学习的安全应用提供系统性指导。

6.1　联邦推荐

联邦推荐（Federated Recommendation）[155-157]是一种融合联邦学习与推荐系统的技术框架，在保护用户隐私的同时提供个性化推荐服务。该框架通过多客户端协同训练，使数据始终保留在本地，从而有效规避传统推荐系统因数据集中存储而引发的隐私泄露风险和数据共享限制。联邦推荐在金融、电子商务等领域展现出显著优势，既能保护数据隐私，又能提升推荐效果[158-159]。

然而，在联邦推荐场景中，后门攻击成为严重的安全威胁。攻击者可以通过篡改本地数据或模型更新，向系统注入恶意信息，破坏推荐结果的公平性，并显著降低全局推荐模型的准确性和可靠性。如图9所示，攻击者可以在本地数据中注入虚假的用户-物品交互记录，使全局模型发生偏差，从而操控特定物品的曝光率，使其异常提升或降低。

不同于第四章讨论的后门攻击方法，针对联邦推荐系统的后门攻击，攻击者需充分考虑系统特有的用户数据结构和特征矩阵，并精心设计适配的触发机制。此外，由于推荐系统对数据的实时性要求较高，攻击者必须确保后门能够迅速适应用户行为的动态变化，这无疑增加了攻击的复杂性和挑战性。

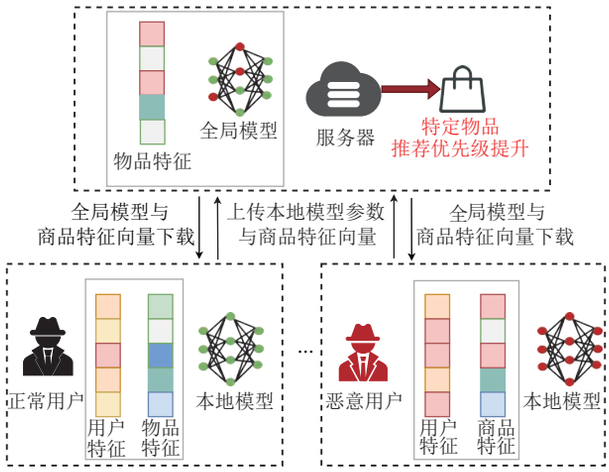

图9　联邦推荐系统中的后门威胁

Rong等人[160]提出了一种专门针对联邦推荐系统的后门攻击方法。在该方法中，攻击者利用公开的用户交互数据估计用户特征向量，并模拟生成恶意中间结果。这使得攻击者能够误导推荐系统，将某个目标物品识别为热门选择，提升该物品在推荐列表中的排名。实验结果显示，即便恶意客户端所占比例低于5%，攻击成功率依然较高，凸显了后门攻击在联邦推荐系统中的潜在威胁和现实可行性。这一发现揭示了该系统在后门攻击方面的脆弱性。

因此，在设计和部署联邦推荐系统时，用户需考虑并采取相应的防御措施以确保系统安全。然而，由于联邦推荐系统中客户端生成的中间结果在形式和结构上与第五章讨论的防御方法中的模型更新存在差异，相关防御方法在该场景的应用面临挑战。首先，物品特征矩阵的梯度通常不具有模型更新的结构特征，这使得一些依赖于特定模型结构的联邦后门防御方法难以直接使用。其次，物品特征矩阵包含用户偏好行为的隐私信息，实际应用中常会结合差分隐私、同态加密或秘密共享等技术进行隐私保护，这可能影响后门防御方法的有效性。最后，物品特征矩阵的稀疏性也会对防御效果产生影响。

为了有效抵御联邦推荐系统中的后门攻击，Yuan等人[161]提出了一种名为HiCS的联邦推荐后门防御方法。HiCS通过结合分层梯度裁剪和稀疏更新两种策略来提高系统的鲁棒性。在分层梯度裁剪中，所有上传的梯度都会被裁剪，以确保每个恶意用户最多只能对模型更新贡献有限的影响。而在稀疏更新策略中，服务器会从聚合的梯度中选择幅度

最大的进行更新，同时将其他梯度置零。

6.2 联邦图学习

联邦图学习（Federated Graph Learning）[162-163]融合了联邦学习与图神经网络[164]的优势，旨在分布式环境下对图结构数据进行深入学习与分析。该技术适用于社交网络分析[165]、生物信息学[166]等领域的大规模敏感图数据，既确保了数据的隐私与安全性，又为复杂图结构数据提供了高效且标准化的分析手段，充分挖掘并利用数据的价值。

如图10所示，联邦图学习同样面临着后门攻击的挑战。在联邦图学习中，后门攻击可通过注入恶意节点或边、篡改本地模型更新或在节点特征或子图中植入触发器来实现。

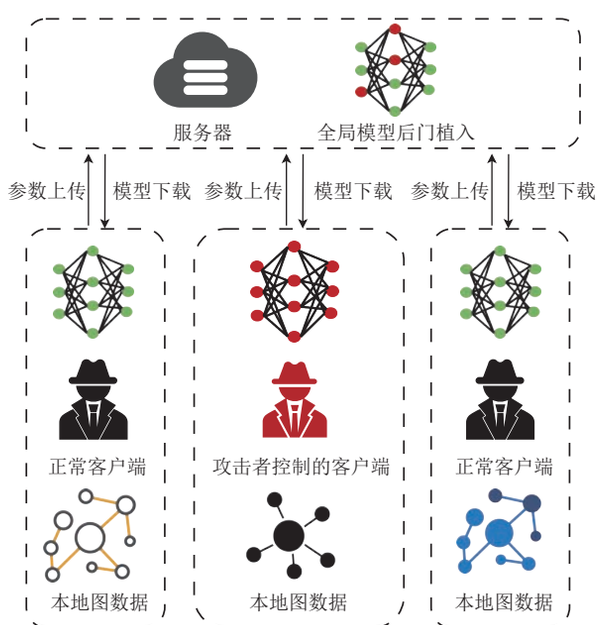

图10 联邦图学习中的后门威胁

联邦图学习主要包含两种架构：第一种是各客户端独立持有结构化的图数据[167-168]，并在此基础上执行联邦学习过程；第二种则是图数据的结构化特征分布在各个客户端之间，客户端所持有的数据相互之间形成了图的连接关系[169-170]。

在第一种架构下，训练数据为本地图数据，这与常见联邦学习框架中的数据模态（如文本、图像等）存在差异。因此，第四章所述的后门攻击技术不能直接应用于此场景。Xu等人[171]针对联邦图神经网络中的后门攻击有效性进行了系统测试。他们通过实验比较了分布式和集中式两种后门攻击策略，并在多种图学习攻击场景及不同数据集上进行了深入分析。研究结果显示，在恶意后门样本数量一致的情况下，分布式后门攻击不仅成功率更高，而且对主任务准确性的影响较小，展现出更佳的隐蔽性。此外，即使在部署了通用防御措施的情况下，联邦图学习场景下的后门攻击依然能够发挥作用。

Yang等人[172]提出了一种针对联邦图学习的精细化后门攻击方法，相较于传统随机触发器构造方法，设计了一种触发器生成器，通过优化触发器的节点选择与拓扑结构，实现了对后门图关键触发器位置与形态的自适应学习，显著提升了攻击效果。同时，为应对此类攻击，研究提出了一种基于子图划分的防御机制，该机制通过将输入图分解为多个子图，基于模型对各子图的预测结果进行多数表决，从而实现对输入图的鲁棒性分析。

针对这种架构的联邦图学习系统，第五章提出的联邦后门防御策略需要进行更深入的验证。如表6所示，虽然这些防御方法在图像数据、表格数据和序列数据上已经证明了其有效性，但在图结构数据这一特殊领域，这些防御机制的有效性和适用性仍需要进一步的实验验证和评估。

在联邦图学习的第二种架构中，目前尚未有针对性的后门攻击策略。第五章提出的联邦后门防御方法是否可以直接应用于此架构，并且这些方法的作用机制及其对原始任务的影响仍需经过严格的验证。有必要针对联邦学习的第二种架构，进行后门攻击脆弱性的挖掘，并开展相应的防御策略设计。

6.3 联邦物联网

联邦物联网使得众多分散的物联网设备能够在不泄露本地数据的情况下，协同参与全局模型的训练过程[172-179]。联邦物联网的应用领域极为广泛，包括智能家居、智慧城市、工业自动化、健康监测等多个重要行业。在这些应用场景中，联邦物联网系统的安全性至关重要，因为它不仅关乎数据的隐私保护，还涉及系统的完整性与服务可用性。

联邦物联网系统面临着后门攻击的严峻挑战。设备之间的异构性（如计算资源、通信能力和操作系统差异）使得后门攻击部署更加复杂。同时，物联网设备的数据实时流动和更新要求后门攻击策略具备在动态数据环境中的稳定性与持续性。攻击者可能篡改传感器数据、伪造设备行为或在本地训练中传递带偏差的模型更新，进而影响全局模型性能。因此，第四章中的后门攻击方法需要在真实联邦物联网环境中进行测试与验证，以确保其有效性。另一方面，在实现联邦物联网场景下的后门防御时，必须充分考虑客户端数据和模型的异质性带来的挑战。特别是

在面对客户端设备多样性所导致的计算环境差异与效率不一致时，还需解决同步通信和聚合效率降低的问题。这些因素是确保防御方法有效性的关键。

在轨道交通领域，确保铁路设备的平稳运行对维持高效服务至关重要。由于轨道交通系统包括大量物联网设备，这些设备共同构成了一个复杂的网络，负责实时监控并确保铁路运行的安全与效率。为了提高对设备异常状态的检测准确性，需要利用分布在不同站段的设备数据来训练检测模型。联邦学习技术能够有效整合来自各火车站的物联网设备数据，不仅提升了异常检测能力，还确保了数据的隐私性和安全性。

如图11所示，攻击者在本地设备训练阶段向图像数据中植入隐蔽触发器，导致全局模型在部署阶段对部分设备状态产生错误判断。这种攻击可能引发安全隐患，如错误的信号识别或列车状态误判，从而对轨道交通系统的安全运行构成潜在威胁。该攻击不仅影响模型的可靠性，还可能危及整个交通网络的稳定性。

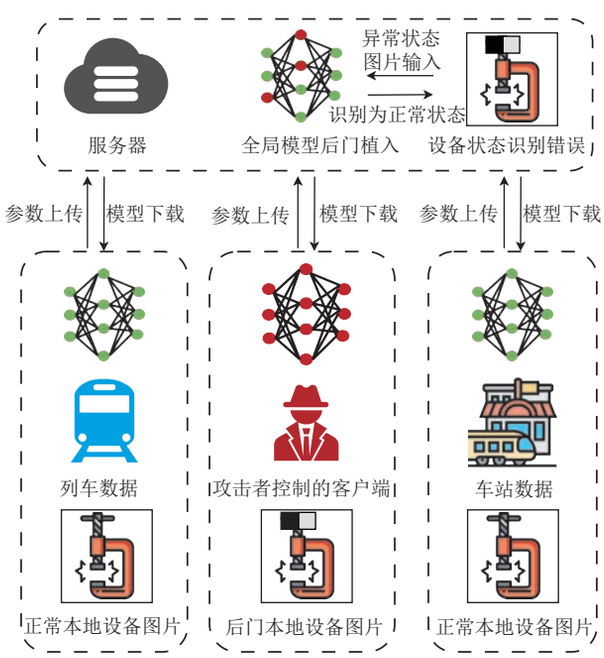

图11　轨道交通领域的后门攻击威胁

Zhu等人[180]针对联邦铁路物联网系统中的后门攻击问题，提出了一系列后门攻击与防御方法。他们通过层放缩技术精心构造恶意模型更新，以绕过服务器的防御措施，潜在地损害全局模型的性能。此外，他们设计了一种基于层聚类的防御方法，该方法通过逐层对模型更新进行聚类分析，再完成全局模型的聚合，有效提高了防御能力，保障了铁路系统的安全稳定运行。

6.4　总　结

本章围绕联邦推荐、联邦图学习和联邦物联网这三个典型应用场景，深入探讨了联邦后门攻击的潜在威胁及其应对策略。在这些场景中，已有研究提出了有效的后门攻击方法，并揭示了这些攻击对联邦学习系统安全性的危害。

因此，在将联邦学习技术应用于实际场景之前，设计和实施一套全面的防御机制显得尤为重要。首先，必须有效识别和过滤后门攻击，以减少恶意客户端对全局模型的负面影响。其次，需要开发有效的后门去除策略，确保在攻击成功注入后能够彻底消除后门威胁，同时保持模型性能。此外，为了处理稀疏数据、保护数据隐私和提升服务质量，通常会采用例如同态加密、差分隐私、数据压缩等其他技术手段。这些技术手段与防御策略之间的兼容性及潜在冲突也需要进行深入分析和验证。

7　未来研究方向

联邦学习作为一项在数据隐私保护领域取得广泛关注的技术，实现了多个客户端在不共享数据的前提下进行分布式模型训练。但随着技术的迅猛发展，联邦学习也面临着安全威胁，其中日益隐蔽和复杂的后门攻击对系统的安全性和完整性构成了严重挑战。正如第四章和第五章所讨论的，后门攻击及其防御方法正处于一场持续的攻防博弈之中。为了增强联邦学习系统的安全性和完整性，需要同步推进联邦学习中的后门攻击与防御方法的研究，以此来指导开发更为先进的检测与防御机制。

技术增强的后门攻击。研究者可以充分利用当前先进的生成模型技术，如大语言模型[181]和扩散模型[182]，以进一步增强后门攻击的效果。这些强大的生成模型不仅能够生成高质量的文本和数据，还能在模拟环境中进行复杂的策略推演，从而为后门攻击和防御提供新的思路。攻击者可能会利用这些生成模型，设计出更加隐蔽、持久且难以检测的后门攻击。这些攻击方式可能通过生成多样化的虚假数据或精心构建的攻击样本，使得后门行为更加隐秘，从而增加检测难度。此外，利用基于大模型的智能体进行红蓝对抗模拟，能够深入分析模型训练过程，挖掘潜藏的后门模式。这种方法不仅可以帮助研究者识别和检测后门行为，还能基于对抗过程提出针对性的防御。通过这种方式，研究者能够更全面地理

解后门攻击的机制，从而在防御策略的设计上做到更加精准和有效。这种深度分析与防御策略相结合的研究策略，将有助于在不断演化的攻击背景下，提升模型的安全性和鲁棒性。

多层次综合性防御。构建更加鲁棒的联邦后门防御机制，对于提升联邦学习系统的整体安全性至关重要。正如第五章的分析所指出，传统防御方法（如鲁棒聚合[86,91]和随机扰动[97,141]）在面对隐蔽性极强的后门攻击时，往往效果有限。因此，研究者应从多层次、多维度的角度构建综合性防御体系，以最大程度降低后门攻击的风险。首先，通过深入分析模型内部行为，可以更好地理解其决策过程和工作机制。这有助于设计高效的攻击检测方法，揭示潜在后门触发器，提高模型对异常行为的感知能力，进而增强其安全性。其次，应探讨联邦学习中的多任务关系，挖掘任务间的协同效应。通过利用多任务学习的鲁棒性，可以构建更稳健的联邦训练架构，使模型具备更强的抗干扰能力，提升整体可靠性。最后，现有研究通常采用多种防御机制的组合策略，但其相互作用仍缺乏系统性分析。研究者应进一步探索不同机制的协同效应，优化组合方式，以最大化防御效果，并增强系统稳定性。

鲁棒性、隐私性以及可用性的平衡。在联邦学习框架下，鲁棒性、隐私性和可用性三者的平衡是防御机制设计中的核心挑战。鲁棒性要求模型能够抵御如后门攻击和拜占庭攻击等多种潜在威胁，确保稳定性和可靠性；隐私性则强调保护客户端数据的安全，通常依赖加密技术和隐私保护机制；可用性关注模型的性能与效率，确保防御措施不会显著降低模型的准确性或训练速度。例如，差分隐私（DP）技术能够有效保护梯度信息的隐私，并增强模型的鲁棒性，但它会显著增加训练时间，且难以确定最优的梯度裁剪和噪声添加阈值，从而影响模型性能[141]。同样，加密技术（如同态加密）提升隐私保护水平的同时，也会带来较大的计算开销，降低训练效率。因此，在设计联邦后门防御机制时，研究者需要在保障安全性的同时，平衡隐私保护与计算效率。防御策略不能仅追求鲁棒性和隐私性，而忽视可用性。研究者应全面评估防御措施的安全性、隐私保护能力，以及对模型准确性和训练时间等的影响。

大语言模型在联邦训练框架下的后门脆弱性。联邦学习作为一种分布式训练方法，能够有效整合分散的算力资源，用于训练大语言模型（LLMs）。近年来，随着大语言模型的迅猛发展，联邦学习已广泛应用于这些模型的训练与优化工作[183-185]。然而，尽管已有研究关注联邦大语言模型的隐私泄露问题[186]，针对后门攻击的系统性分析仍较为匮乏。联邦学习的分布式特性和大语言模型的多任务处理能力使得后门攻击风险不可忽视[187]。攻击者可能通过恶意客户端的局部模型更新，在全局模型中植入隐蔽的后门，导致推理阶段的异常行为，严重影响模型的安全性与可靠性。未来的研究应聚焦于联邦学习环境下客户端模型带来的安全挑战，尤其是攻击者如何利用客户端模型作为攻击载体危害整个联邦大语言模型的安全。为此，需要深入分析客户端模型的安全性，系统探讨后门攻击的传播机制，并设计更为鲁棒的联邦学习框架，优化模型更新流程和防御策略，以确保训练过程的安全性与鲁棒性，进而提升大语言模型的抗攻击能力。

8 结 论

联邦学习作为一种隐私保护的分布式机器学习范式，由于其分布式特性，易受到外部攻击者实施的后门攻击。本文深入探讨了联邦学习系统中后门攻击的机制，并对相关防御方法进行了详细分析。在后门攻击方面，本文全面调研并分析了不同类型的联邦学习系统（如横向联邦学习、纵向联邦学习、联邦迁移学习及异构联邦学习）所面临的后门攻击威胁。在防御方法方面，本文根据所使用信息的不同，进行了分类比较和归纳总结。针对实际应用场景，本文分析并讨论了具体的后门攻击实例及其相应的防御措施。此外，本文还展望了未来可能的研究方向。综上所述，本文为理解和应对联邦学习中的后门攻击提供了全面的视角，并为构建有效的防御方法提供了深入的分析和广泛的讨论，有助于推动联邦学习安全领域的进一步研究。

致 谢 我们向对本文的工作给予支持和宝贵建议的编辑和评审老师表示衷心的感谢！

参 考 文 献

[1] Brendan Mcmahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera Y Arcas. Communication-efficient learning of deep networks from decentralized data// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Lauderdale, USA, 2017: 1273-1282

[2] Peter Kairouz, H Brendan Mcmahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210

[3] Jakub Konečný, H Brendan Mcmahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, BaconDave. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016

[4] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, NaiboWang, Yuan Li, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 3347-3366

[5] Mica Goldblumh, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1563-1580

[6] Yiming Li, Yong Jiang, Zhifeng Li, Shu-Tao Xia. Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5-22

[7] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Kai Chen, et al. Coba: Collusive backdoor attacks with optimized trigger to federated learning. IEEE Transactions on Dependable and Secure Computing, 2024, 22(2): 1506-1518

[8] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning// Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 9020-9028

[9] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Yongsheng Zhu, Guangquan Xu, et al. Lurking in the shadows: Unveiling stealthy backdoor attacks against personalized federated learning// Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24. Philadelphia, USA, 2024: 4157-4174

[10] Sawsan Abdulrahman, Hanine Tout, Hakima Ould-Slimane, Azzam Mourad, Chamseddine Talhi, Mohsen Guizani. A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal, 2020, 8(7): 5476-5497

[11] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, Gautam Srivastava. A survey on security and privacy of federated learning. Future Generation Computer Systems, 2021, 1(115): 619-640

[12] Xuefei Yin, Yanming Zhu, Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR), 2021, 54(6): 1-36

[13] Chen Bing, Cheng Xiang, Zhang Jiale, Xie Yuanyuan. Overview of federated learning security and privacy protection. Journal of Nanjing University of Aeronautics and Astronautics, 2020, 52(5): 675-684
(陈兵，成翔，张佳乐，谢袁源. 联邦学习安全与隐私保护综述. 南京航空航天大学学报，2020，52(5): 675-684)

[14] Gao Ying, Chen Xiaofeng, Zhang Yiyu, Wang Wei, Deng Huanghao, Duan Pei, Chen Peixuan. A review of research on attack and defense techniques for federated learning systems. Journal of Computer Science, 2023, 46(9): 1781-1805
(高莹，陈晓峰，张一余，王玮，邓煌昊，段培，陈培炫. 联邦学习系统攻击与防御技术研究综述. 计算机学报，2023，46(9): 1781-1805)

[15] Wu Wentai, Wu Yingliang, Lin Weiwei, Zuo Wenming. Horizontal federated learning: Research status, system applications, and challenges. Journal of Computer Sciences, 2025, 48(1): 35-67
(吴文泰，吴应良，林伟伟，左文明. 横向联邦学习:研究现状、系统应用与挑战. 计算机学报，2025，48(1): 35-67)

[16] Pengrui Liu, Xiangrui Xu, Wei Wang. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. Cybersecurity, 2022, 5(1): 4

[17] Thuy Dung Nguyen, Tuan Nguyen, Phi Le Nguyen, Hieu H Pham, Khoa D Doan, Kok-Seng Wong. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. Engineering Applications of Artificial Intelligence, 2024, 127: 107166

[18] Xueluan Gong, Yanjiao Chen, Qian Wang, Weihan Kong. Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. IEEE Wireless Communications, 2022, 30(2): 114-121

[19] Liu Jialang, Guo Yanming, Lao Mingrui, et al. A review of backdoor attack and defense algorithms based on federated learning. Journal of Computer Research and Development, 2024, 61(10): 2607-2626
(刘嘉浪，郭延明，老明瑞，于天元，武与伦，冯云浩，吴嘉壮. 基于联邦学习的后门攻击与防御算法综述. 计算机研究与发展，2024，61(10): 2607-2626)

[20] Ranwa Al Mallah, David Lopez, Godwin Badu-Marfo, Bilal Farooq. Untargeted poisoning attack detection in federated learning via behavior attestational. IEEE Access, 2023, 11: 125064-125079

[21] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, Ling Liu. Data poisoning attacks against federated learning systems// Computer Security-ESORICs 2020: 25th European Symposium on Research in Computer Security. Guildford, UK, 2020: 480-501

[22] Yang Liu, Zhihao Yi, Tianjian Chen. Backdoor attacks and defenses in feature-partitioned collaborative learning. arXiv preprint arXiv:2007.03608, 2020

[23] Peng Chen, Xin Du, Zhihui Lu, Hongfeng Chai. Universal adversarial backdoor attacks to fool vertical federated learning. Computers & Security, 2024, 137: 103601

[24] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526, 2017

[25] Mauro Conti, Nicola Dragoni, Viktor Lesyk. A survey of man in the middle attacks. IEEE Communications Surveys & Tutorials, 2016, 18(3): 2027-2051

[26] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly

Shmatikov. Membership inference attacks against machine learning models//2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017: 3-18

[27] A Suri, P Kanani, Vj Marathe, Dw Peterson. Subject membership inference attacks in federated learning. arXiv 2022. arXiv preprint arXiv:2206.03317

[28] Hongsheng Hu, Zoran Salcic, Sun Lichao, Gillian Dobbie, Xuyun Zhang. Source inference attacks in federated learning// 2021 IEEE International Conference on Data Mining (ICDM). Auckland, New Zealand, 2021: 1102-1107

[29] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, Yang Zhang. Updates-Leak: Data set inference and reconstruction attacks in online learning//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020: 1291-1308

[30] Haomiao Yang, Mengyu Ge, Kunlan Xiang, Jingwei Li. Using highly compressed gradients in federated learning for data reconstruction attacks. IEEE Transactions on Information Forensics and Security, 2022, 18: 818-830

[31] Hongfu Liu, BinLi, Changlong Gao, Pei Xie, Chenglin Zhao. Privacy-encoded federated learning against gradient-based data reconstruction attacks. IEEE Transactions on Information Forensics and Security, 2023, 18: 5860-5875

[32] Xiangrui Xu, Pengrui Liu, Wei Wang, Hong-Liang Ma, Bin Wang, Zhen Han, Yufei Han. CGIR: Conditional generative instance reconstruction attacks against federated learning. IEEE Transactions on Dependable and Secure Computing, 2022, 20(6): 4551-4563

[33] Zhenqiang Neil Gong, Bin Liu. Attribute inference attacks in online social networks. ACM Transactions on Privacy and Security (TOPS), 2018, 21(1): 1-30

[34] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, et al. On the (in) feasibility of attribute inference attacks on machine learning models//Proceedings of the 2021 IEEE European Symposium on Security and Privacy (EuroS&P). San Francisco, USA, 2021: 232-251

[35] Shijie Zhang, Wei Yuan, Hongzhi Yin. Comprehensive privacy analysis on federated recommender system against attribute inference attacks. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(3): 987-999

[36] Tiantian Feng, Hanieh Hashemi, Rajat Hebbar, Murali Annavaram, Shrikanth Narayanan S. Attribute inference attack of speech emotion recognition in federated learning settings. arXiv preprint arXiv:2112.13416, 2021

[37] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, et al. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604, 2018

[38] Li Li, Yuxi Fan, Mike Tse, Kuo-Yi Lin. A review of applications in federated learning. Computers & Industrial Engineering, 2020, 149: 106854

[39] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, Wensheng Zhang. A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics, 2023, 14(2): 513-535

[40] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, Brendan H Mcmahan. Can you really backdoor federated learning? arXiv preprint arXiv:1911.07963, 2019

[41] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, Vitaly Shmatikov. How to backdoor federated learning// International Conference on Artificial Iintelligence and Statistics. Online, 2020: 2938-2948

[42] Chulin Xie, Keli Huang, Pin-Yu Chen, Bo Li. Dba: Distributed backdoor attacks against federated learning//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[43] Xueluan Gong, Yanjiao Chen, Huayang Huang, Yuqing Liao, Shuai Wang, Qian Wang. Coordinated backdoor attacks against federated learning with model-dependent triggers. IEEE Network, 2022, 36(1): 84-90

[44] Qixian Ren, Yu Zheng, Chao Yang, Yue Li, Jianfeng Ma. Shadow backdoor attack: Multi-intensity backdoor attack against federated learning. Computers & Security, 2024, 139: 103740

[45] Yingrui Tong, Jun Feng, Gaolei Li, Xi Lin, Chengcheng Zhao, Xiaoyu Yi, Jianhua Li. SBA Sem: Semantic-perturbed stealthy backdoor attack on federated semi-supervised learning// Proceedings of the 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). Ocean Flower Island, China, 2023: 1569-1576

[46] Haomin Zhuang, Mingxian Yu, Hao Wang, Yang Hua, Jian Li, Xu Yuan. Backdoor federated learning by poisoning backdoor-critical layers. arXiv preprint arXiv:2308.04466, 2023

[47] Pei Fang, Jinghui Chen. On the vulnerability of backdoor defenses for federated learning// Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 11800-11808

[48] Yanbo Dai, Songze Li. Chameleon: Adapting to peer images for planting durable backdoors in federated learning// International Conference on Machine Learning. Honolulu, USA, 2023: 6712-6725

[49] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, et al. Neurotoxin: Durable backdoors in federated learning//International Conference on Machine Learning. Baltimore, USA, 2022: 26429-26446

[50] Chenghui Shi, Shouling Ji, Xudong Pan, Xuhong Zhang, Mi Zhang, Min Yang, et al. Towards practical backdoor attacks on federated learning systems. IEEE Transactions on Dependable and Secure Computing, 2024

[51] Thuy Dung Nguyen, Tuan A Nguyen, Anh Tran, Khoa D Doan, Kok-Seng Wong. Iba: Towards irreversible backdoor attacks in federated learning. Advances in Neural Information Processing Systems, 2023, 36: 66364-66376

[52] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. Advances in Neural Information Processing Systems,

2023, 36: 61213-61233

[53] Wenqiang Sun, Sen Li, Yuchang Sun, Jun Zhang. DABS: Data-Agnostic Backdoor attack at the Server in Federated Learning. arXiv preprint arXiv:2305.01267, 2023

[54] Pu Paul Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, et al. Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523, 2020

[55] Alireza Fallah, Aryan Mokhtari, Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. arXiv preprint arXiv:2002.07948, 2020

[56] Tian Li, Shengyuan Hu, Ahmad Beirami, Virginia Smith. Ditto: Fair and robust federated learning through personalization//Proceedings of the International Conference on Machine Learning, Virtual, 2021: 6357-6368

[57] Liam Collins, Hamed Hassani, Aryan Mokhtari, Sanjay Shakkottai. Exploiting shared representations for personalized federated learning// Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 2089-2099

[58] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623, 2021

[59] Zeyu Qin, Liuyi Yao, Daoyuan Chen, Yaliang Li, Bolin Ding, Minhao Cheng. Revisiting personalized federated learning: Robustness against backdoor attacks//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA, 2023: 4743-4755

[60] Tiandi Ye, Cen Chen, Yinggui Wang, Xiang Li, Ming Gao. BapFL: You can backdoor personalized federated learning. ACM Transactions on Knowledge Discovery from Data, 2024, 18(7): 1-17

[61] Haochen Mei, Gaolei Li, Jun Wu, Longfei Zheng. Privacy inference-empowered stealthy backdoor attack on federated learning under non-iid scenarios//Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN). Queensland, Australia, 2023: 1-10

[62] Aviv Shamsian, Aviv Navon, Ethan Fetaya, Gal Chechik. Personalized federated learning using hypernetworks// Proceedings of the International Conference on Machine Learning, Virtual, 2021: 9489-9502

[63] Phung Lai, Nhathai Phan, Abdallah Khreishah, Issa Khalil, Xintao Wu. Model transferring attacks to backdoor hypernetwork in personalized federated learning. CoRR, abs/2201.07063, 2022

[64] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, et al. Vertical federated learning: Concepts, advances, and challenges. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3615-3634

[65] Siwei Feng, Han Yu. Multi-participant multi-class vertical federated learning. arXiv preprint arXiv:2001.11154, 2020

[66] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, Beng Chin Ooi. Privacy preserving vertical federated learning for tree-based models. arXiv preprint arXiv:2008.06170, 2020

[67] Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, Chen. Tianyi Cafe: Catastrophic data leakage in vertical federated learning. Advances in Neural Information Processing Systems, 2021, 34: 994-1006

[68] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning//Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece, 2021: 181-192

[69] Otkrist Gupta, Ramesh Raskar. Distributed learning of deep neural network over multiple agents. Journal of Network and Computer Applications, 2018, 116: 1-8

[70] Iker Ceballos, Vivek Sharma, Eduardo Mugica, Abhishek Singh, Alberto Roman, Praneeth Vepakomma, Ramesh Raskar. Splitnn-driven vertical partitioning. arXiv preprint arXiv:2008.04137, 2020

[71] Daniele Romanini, Adam James Hall, Pavlos Papadopoulos, Tom Titcombe, Abbas Ismail, Tudor Cebere, et al. Pyvertical: A vertical federated learning framework for multi-headed splitnn. arXiv preprint arXiv:2104.00489, 2021

[72] Ying He, Zhili Shen, Jingyu Hua, Qixuan Dong, Jiacheng Niu, Wei Tong, et al. Backdoor attack against split neural network-based vertical federated learning. IEEE Transactions on Information Forensics and Security, 2023, 19: 748-763

[73] Mohammad Naseri, Yufei Han, Emiliano De Cristofaro. Badvfl: Backdoor attacks in vertical federated learning// Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2024: 2013-2028

[74] Yuhao Gu, Yuebin Bai. LR-BA: Backdoor attack against vertical federated learning using local latent representations. Computers & Security, 2023, 129: 103193

[75] Peng Chen, Xin Du, Zhihui Lu, Hongfeng Chai. Universal adversarial backdoor attacks to fool vertical federated learning in cloud-edge collaboration. arXiv preprint arXiv: 2304.11432, 2023

[76] Yang Liu, Tianyuan Zou, Yan Kang, Wenhan Liu, Yuanqin He, Zhihao Yi, Qiang Yang. Batch label inference and replacement attacks in black-boxed vertical federated learning. arXiv preprint arXiv:2112.05409, 2021

[77] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, Qiang Yang. A secure federated transfer learning framework. IEEE Intelligent Systems, 2020, 35(4): 70-82

[78] Sudipan Saha, Tahir Ahmad. Federated transfer learning: Concept and applications. Intelligenza Artificiale, 2021, 15(1): 35-44

[79] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, Qiang Yang. Privacy-preserving heterogeneous federated transfer learning//Proceedings of the 2019 IEEE International Conference on Big Data. Los Angeles, USA, 2019: 2552-2559

[80] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems, 2020, 35(4): 83-93

[81] Shreya Sharma, Chaoping Xing, Yang Liu, Yan Kang. Secure and efficient federated transfer learning//Proceedings of the

2019 IEEE international conference on big data (Big Data). Los Angeles, USA, 2019: 2569-2576

[82] Marco Arazzi, Stefanos Koffas, Antonino Nocera, Stjepan Picek. Let's Focus: Focused backdoor attack against federated transfer learning. arXiv preprint arXiv:2404.19420, 2024

[83] Ramprasaath Selvaraju R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 618-626

[84] Phung Lai, Nhathai Phan, Issa Khalil, Abdallah Khreishah, Xintao Wu. How to backdoor network hyper in personalized federated learning? arXiv preprint arXiv:2201.07063, 2022

[85] Xi Li, Chen Wu, Jiaqi Wang. Unveiling backdoor risks brought by foundation models in heterogeneous federated learning// Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei, China, 2024: 168-181

[86] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 2017, 30

[87] Thien Duc Nguyen, Phillip Rieger, Roberta De Viti, Huili Chen, Björn Brandenburg B, Hossein Yalame, et al. {FLAME}: Taming backdoors in federated learning// Proceedings of the 31st USENIX Security Symposium. Boston, USA, 2022: 1415-1432

[88] Krishna Pillutla, Sham Kakade M, Zaid Harchaoui. Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 2022, 70: 1142-1154

[89] Clement Fung, Chris Jm Yoon, Ivan Beschastnikh. The limitations of federated learning in sybil settings//Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses. 2020: 301-316

[90] Rachid Guerraoui, Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3521-3530

[91] Dong Yin, Yudong Chen, Ramchandran Kannan, Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates//Proceedings of the International Conference on Machine Learning. Stockholmsmässan, Stockholm, Sweden, 2018: 5650-5659

[92] Xiaoyu Cao, Minghong Fang, Jia Liu, Zhenqiang Neil Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint arXiv:2012.13995, 2020

[93] Virat Shejwalkar, Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning//Proceedings of the Network and Distributed System Security Symposium (NDSS), 2021

[94] Chulin Xie, Minghao Chen, Pin-Yu Chen, Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks// Proceedings of the International Conference on Machine Learning, Virtual, 2021: 11372-11382

[95] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, Vincent PoorH. User-level privacy-preserving federated learning: Analysis and performance optimization. IEEE Transactions on Mobile Computing, 2021, 21(9): 3388-3401

[96] Robin Geyer C, Tassilo Klein, Moin Nabi. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557, 2017

[97] Jingwei Sun, Ang Li, Louis Divalentin, Amin Hassanzadeh, Yiran Chen, Hai Li. Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. Advances in Neural Information Processing Systems, 2021, 34: 12613-12624

[98] Tao Lin, Lingjing Kong, Sebastian Stich U, Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. Advances in Neural Information Processing Systems, 2020, 33: 2351-2363

[99] Stefán Páll Sturluson, Samuel Trew, Luis Muñoz-González, Matei Grama, Jonathan Passerat-Palmbach, Daniel Rueckert, Amir Alansary. Fedrad: Federated robust adaptive distillation. arXiv preprint arXiv:2112.01405, 2021

[100] Chen Wu, Xian Yang, Sencun Zhu, Prasenjit Mitra. Mitigating backdoor attacks in federated learning. arXiv preprint arXiv: 2011.01767, 2020

[101] Mustafa Safa Ozdayi, Murat Kantarcioglu, Yulia Gel R. Defending against backdoors in federated learning with robust learning rate//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 9268-9276

[102] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. arXiv preprint arXiv:2201.00763, 2022

[103] Cong Xie, Sanmi Koyejo, Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 6893-6901

[104] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer. Byzantine-tolerant machine learning. arXiv preprint arXiv:1703.02757, 2017

[105] Kang Liu, Brendan Dolan-Gavitt, Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks//Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses. Heraklion, Greece, 2018: 273-294

[106] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, Zhenqiang Neil Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2022: 2545-2555

[107] Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, Siyuan Cheng, Shengwei An, Yingqi Liu, et al. Flip: A provable defense framework for backdoor mitigation in federated learning. arXiv preprint arXiv:2210.12873, 2022

[108] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal. Sparsefed: Mitigating

model poisoning attacks in federated learning with sparsification//Proceedings of the International Conference on Artificial Intelligence and Statistics. Valencia, Spain, 2022: 7587-7624

[109] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. arXiv preprint arXiv: 2101.05930, 2021

[110] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, et al. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728, 2018

[111] Tianyuan Zou, Yang Liu, Yan Kang, Wenhan Liu, Yuanqin He, Zhihao Yi, et al. Defending batch-level label inference and replacement attacks in vertical federated learning. IEEE Transactions on Big Data, 2022,12(5):234-246

[112] Yujun Lin, Song Han, Huizi Mao, Yu Wang, William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. arXiv preprint arXiv: 1712.01887, 2017

[113] Oscar Li, Jiankai Sun, Xin Yang, Weihao Gao, Hongyi Zhang, Junyuan Xie, et al. Label leakage and protection in two-party split learning. arXiv preprint arXiv:2102.08504, 2021

[114] Jing Liu, Chulin Xie, Sanmi Koyejo, Bo Li. CoPur: certifiably robust collaborative inference via feature purification. Advances in Neural Information Processing Systems, 2022, 35: 26645-26657

[115] Yan Kang, Jiahuan Luo, Yuanqin He, Xiaojin Zhang, Lixin Fan, Qiang Yang. A framework for evaluating privacy-utility trade-off in vertical federated learning. arXiv preprint arXiv: 2209.03885, 2022

[116] Marco Arazzi, Serena Nicolazzo, Antonino Nocera. A defense mechanism against label inference attacks in vertical federated learning. Neurocomputing, 2025: 129476

[117] Jingkai Liu, Xiaoting Lyu, Li Duan, Yongzhong He, Jiqiang Liu, Hongliang Ma, et al. Pna: Robust aggregation against poisoning attacks to federated learning for edge intelligence. ACM Transactions on Sensor Networks,to appear

[118] H Brendan Mcmahan, Daniel Ramage, Kunal Talwar, Li Zhang. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963, 2017

[119] Waris Gill, Ali Anwar, Muhammad Ali Gulzar. FedDefender: Backdoor attack defense in federated learning// Proceedings of the 1st International Workshop on Dependability and Trustworthiness of Safety-Critical Systems with Machine Learned Components. San Francisco , USA, 2023: 6-9

[120] Jinyuan Jia, Zhuowen Yuan, Dinuka Sahabandu, Luyao Niu, Arezoo Rajabi, Bhaskar Ramasubramanian, et al. Fedgame: A game-theoretic defense against backdoor attacks in federated learning. Advances in Neural Information Processing Systems, 2023, 36: 53090-53111

[121] Jun Li, Jian Lin. A probability distribution detection based hybrid ensemble QoS prediction approach. Information Sciences, 2020, 519: 289-305

[122] Kavita Kumari, Phillip Rieger, Hossein Fereidooni, Murtuza Jadliwala, Ahmad-Reza Sadeghi. Baybfed: Bayesian backdoor defense for federated learning//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2023: 737-754

[123] Zhen Qin, Feiyi Chen, Chen Zhi, Xueqiang Yan, Shuiguang Deng. Resisting backdoor attacks in federated learning via bidirectional elections and individual perspective// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 14677-14685

[124] Zihan Ma, Tianchong Gao. Federated learning backdoor attack detection with persistence diagram. Computers & Security, 2024, 136: 103557

[125] Xiaoyu Cao, Zaixi Zhang, Jinyuan Jia, Zhenqiang Neil Gong. Flcert: Provably secure federated learning against poisoning attacks. IEEE Transactions on Information Forensics and Security, 2022, 17: 3691-3705

[126] Xiaoyun Gan, Shanyu Gan, Taizhi Su, Peng Liu. GANcrop: A contrastive defense against backdoor attacks in federated learning//Proceedings of the 2024 5th International Conference on Computing, Networks and Internet of Things. Tokyo, Japan, 2024: 606-612

[127] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[128] Songze Li, Yanbo Dai. {BackdoorIndicator}: Leveraging {OOD} data for proactive backdoor detection in federated learning// Proceedings of the 33rd USENIX Security Symposium. Philadelphia,USA, 2024: 4193-4210

[129] Yungi Cho, Woorim Han, Miseon Yu, Younghan Lee, Ho Bae, Yunheung Paek. VFLIP: A backdoor defense for vertical federated learning via identification and purification// Proceedings of the European Symposium on Research in Computer Security. Bydgoszcz, Poland, 2024: 291-312

[130] Zhou Tan, Jianping Cai, Puwei Lian, Ximeng Liu, Yan Che. FedPD: Defending federated prototype learning against backdoor attacks. Neural Networks, 2025, 184: 107016

[131] Zekai Chen, Shengxing Yu, Mingyuan Fan, Ximeng Liu, H Robert Deng. Privacy-enhancing and robust backdoor defense for federated learning on heterogeneous data. IEEE Transactions on Information Forensics and Security, 2023, 19: 693-707

[132] Chunhua Tang, Han Wang, Zhiwen Wang, Xiangkun Zeng, Huaran Yan, Yingjie Xiao. An improved OPTICS clustering algorithm for discovering clusters with uneven densities. Intelligent Data Analysis, 2021, 25(6): 1453-1471

[133] He Yang, Wei Xi, Yuhao Shen, Canhui Wu, Jizhong Zhao. Roseagg: Robust defense against targeted collusion attacks in federated learning. IEEE Transactions on Information Forensics and Security, 2024, 19: 2951-2966

[134] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the Second International

Conference on Knowledge Discovery and Data Mining(KDD-96). Portland, USA, 1996: 226-231

[135] Torsten Krauß, Alexandra Dmitrienko. Mesas: Poisoning defense for federated learning resilient against adaptive attackers// Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen, Denmark, 2023: 1526-1540

[136] Tiansheng Huang, Sihao Hu, Ka-Ho Chow, Fatih Ilhan, Selim Tekin, Ling Liu. Lockdown: backdoor defense for federated learning with isolated subspace training. Advances in Neural Information Processing Systems, 2023, 36: 10876-10896

[137] Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi. Crowdguard: Federated backdoor detection in federated learning. arXiv preprint arXiv: 2210.07714, 2022

[138] Jiale Zhang, Chengcheng Zhu, Xiaobing Sun, Chunpeng Ge, Bing Chen, Willy Susilo, Shui Yu. Flpurifier: backdoor defense in federated learning via decoupled contrastive training. IEEE Transactions on Information Forensics and Security, 2024, 19: 4752-4766

[139] Vasyl Pihur, Aleksandra Korolova, Frederick Liu, Subhash Sankuratripati, Moti Yung, Dachuan Huang, Ruogu Zeng. Differentially-private" draw and discard" machine learning. arXiv preprint arXiv:1807.04369, 2018

[140] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan Mcmahan, Ilya Mironov, Kunal Talwar, Li Zhang. Deep learning with differential privacy// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria, 2016: 308-318

[141] Mohammad Naseri, Jamie Hayes, Emiliano De Cristofaro. Local and central differential privacy for robustness and privacy in federated learning. arXiv preprint arXiv:2009.03561, 2020

[142] Chaoyi Zhu, Stefanie Roos, Lydia Y Chen. LeadFL: Client self-defense against model poisoning in federated learning// International Conference on Machine Learning. Honolulu, USA, 2023: 43158-43180

[143] Kane Walter, Meisam Mohammady, Surya Nepal, Salil S Kanhere. Mitigating distributed backdoor attack in federated learning through mode connectivity// Proceedings of the 19th ACM Asia Conference on Computer and Communications Security. Singapore, 2024: 1287-1298

[144] Gilad Baruch, Moran Baruch, Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. Advances in Neural Information Processing Systems, 2019, 32

[145] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-Yong Sohn, et al. Attack of the tails: Yes, you really can backdoor federated learning. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084

[146] Fatima Elhattab. Towards mitigation of edge-case backdoor attacks in federated learning//Proceedings of the 16th EuroSys Doctoral Workshop. Rennes, France, 2022:1-3

[147] Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017

[148] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, Shutao Xia. Rethinking the trigger of backdoor attack. arXiv preprint arXiv:2004.04692, 2020

[149] Haoyang Li, Qingqing Ye, Haibo Hu, Jin Li, Leixia Wang, Chengfang Fang, Jie Shi. 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning// Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2023: 1893-1907

[150] Yijie Bai, Yanjiao Chen, Hanlei Zhang, Wenyuan Xu, Haiqin Weng, Dou Goodman. {VILLAIN}: Backdoor attacks against vertical split learning//Proceedings of the 32nd USENIX Security Symposium. Anaheim, USA, 2023: 2743-2760

[151] Mauro Barni, Kassem Kallas, Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning//Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, 2019: 101-105

[152] Alexander Turner, Dimitris Tsipras, Aleksander Madry. Label-consistent backdoor attacks. arXiv preprint arXiv: 1912.02771, 2019

[153] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, Seraphin Calo. Analyzing federated learning through an adversarial lens// Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 634-643

[154] Eugene Bagdasaryan, Vitaly Shmatikov. Blind backdoors in deep learning models//Proceedings of the 30th USENIX Security Symposium. Virtual, 2021: 1505-1521

[155] Amir Jalalirad, Marco Scavuzzo, Catalin Capota, Michael Sprague. A simple and efficient federated recommender system// Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. Auckland New Zealand, 2019: 53-58

[156] Ziming Ye, Xiao Zhang, Xu Chen, Hui Xiong, Dongxiao Yu. Adaptive clustering based personalized federated learning framework for next poi recommendation with location noise. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(5): 1843-1856

[157] Yue Wu, Lei Su, Liping Wu, Weinan Xiong. FedDeepFM: a factorization machine-based neural network for recommendation in federated learning. IEEE Access, 2023, 11: 74182-74190

[158] Shahriar Badsha, Xun Yi, Ibrahim Khalil. A practical privacy-preserving recommender system. Data Science and Engineering, 2016, 1: 161-177

[159] Jiangcheng Qin, Baisong Liu, Jiangbo Qian. A novel privacy-preserved recommender system framework based on federated learning// Proceedings of the 2021 4th International Conference on Software Engineering and Information Management. Yokohama, Japan, 2021: 82-88

[160] Dazhong Rong, Shuai Ye, Ruoyan Zhao, Hon Ning Yuen, Jianhai Chen, Qinming He. Fedrecattack: Model poisoning attack to federated recommendation//2022 IEEE 38th International Conference on Data Engineering (ICDE). Kuala Lumpur, Malaysia, 2022: 2643-2655

[161] Wei Yuan, Quoc Viet Hung Nguyen, Tieke He, Liang Chen, Hongzhi Yin. Manipulating federated recommender systems: Poisoning with synthetic users and its countermeasures// Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China, 2023: 1690-1699

[162] Guancheng Wan, Wenke Huang, Mang Ye. Federated graph learning under domain shift with generalizable prototypes// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 15429-15437

[163] Zhenzhen Xie, Yan Huang, Dongxiao Yu, Reza Parizi M, Yanwei Zheng, Junjie Pang. FedEE: A federated graph learning solution for extended enterprise collaboration. IEEE Transactions on Industrial Informatics, 2022, 19(7): 8061-8071

[164] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, et al. Graph neural networks: A review of methods and applications. AI open, 2020, 1: 57-81

[165] Zhiwei Liu, Liangwei Yang, Ziwei Fan, Hao Peng, Philip S Yu. Federated social recommendation with graph neural network. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(4): 1-24

[166] Rui Liu, Pengwei Xing, Zichao Deng, Anran Li, Cuntai Guan, Han Yu. Federated graph neural networks: Overview, techniques, and challenges. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(3): 4279-4295

[167] Xingjie Zeng, Tao Zhou, Zhicheng Bao, Hongwei Zhao, Leiming Chen, Xiao Wang, Feiyue Wang. Feature-contrastive graph federated learning: Responsible ai in graph information analysis. IEEE Transactions on Computational Social Systems, 2022, 10(6): 2938-2948

[168] Chaoyang He, Emir Ceyani, Keshav Balasubramanian, Murali Annavaram, Salman Avestimehr. Spreadgnn: Decentralized multi-task federated learning for graph neural networks on molecular data// Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, Pennyslvania, 2022: 6865-6873

[169] Chuizheng Meng, Sirisha Rambhatla, Yan Liu. Cross-node federated graph neural network for spatio-temporal data modeling// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore, 2021: 1202-1211

[170] Pengwei Xing, Songtao Lu, Lingfei Wu, Han Yu. Big-fed: Bilevel optimization enhanced graph-aided federated learning. IEEE Transactions on Big Data, 2022, 10(6): 903-914

[171] Jing Xu, Rui Wang, Stefanos Koffas, Kaitai Liang, Stjepan Picek. More is better (mostly): On the backdoor attacks in federated graph neural networks// Proceedings of the 38th Annual Computer Security Applications Conference. Austin, USA, 2022: 684-698

[172] Yuxin Yang, Qiang Li, Jinyuan Jia, Yuan Hong, Binghui Wang. Distributed backdoor attacks on federated graph learning and certified defenses// Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 2829-2843

[173] Wei Li, Cheng Zhang, Yoshiaki Tanaka. Pseudo label-driven federated learning-based decentralized indoor localization via mobile crowdsourcing. IEEE Sensors Journal, 2020, 20(19): 11556-11565

[174] Abdullatif Albaseer, Bekir Sait Ciftler, Mohamed Abdallah, Ala Al-Fuqaha. Exploiting unlabeled data in smart cities using federated edge learning//Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC). Limassol, Cyprus, 2020: 1666-1671

[175] Yunlong Lu, Xiaohong Huang, Xiaohong Daix, Sabita Maharjan, Yan Zhang. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186

[176] Linghe Kong, Xiao-Yang Liu, Hao Sheng, Peng Zeng, Guihai Chen. Federated tensor mining for secure industrial internet of things. IEEE Transactions on Industrial Informatics, 2019, 16(3): 2144-2153

[177] Yunlong Lu, Xiaohong Huang, Ke Zhang, Sabita Maharjan, Yan Zhang. Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4298-4311

[178] Haoye Chai, Supeng Leng, Yijin Chen, Ke Zhang. A hierarchical blockchain-enabled federated learning algorithm for knowledge sharing in internet of vehicles. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(7): 3975-3986

[179] Yunfei Song, Tian Liu, Tongquan Wei, Xiangfeng Wang, Zhe Tao, Mingsong Chen. FDA^3: Federated defense against adversarial attacks for cloud-based IIoT applications. IEEE Transactions on Industrial Informatics, 2020, 17(11): 7830-7838

[180] Yongsheng Zhu, Chong Liu, Chunlei Chen, Xiaoting Lyu, Zheng Chen, Bin Wang, et al. Privacy-preserving large-scale AI models for intelligent railway transportation systems: hierarchical poisoning attacks and defenses in federated learning. CMES-Computer Modeling in Engineering & Sciences, 2024, 141(2): 1305-1325

[181] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 2023, 103: 102274

[182] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, et al. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2023, 56(4): 1-39

[183] Tao Fan, Guoqiang Ma, Yan Kang, Hanlin Gu, Yuanfeng Song, Lixin Fan, et al. Fedmkt: Federated mutual knowledge transfer for large and small language models. arXiv preprint arXiv:2406.02224, 2024

[184] Jingang Jiang, Haiqi Jiang, Yuhan Ma, Xiangyang Liu, Chenyou Fan. Low-parameter federated learning with large language models// Proceedings of the International Conference on Web Information Systems and Applications. Yinchuan, China, 2024: 319-330

[185] Jiaying Zheng, Hainan Zhang, Lingxiang Wang, Wangjie Qiu, Hongwei Zheng, Zhiming Zheng. Safely learning with private data: A federated learning framework for large language model. arXiv preprint arXiv:2406.14898, 2024

[186] Minh Vu, Truc Nguyen, My T Thai. Analysis of privacy leakage in federated large language models//Proceedings of the International Conference on Artificial Intelligence and Statistics. Valencia, Spain, 2024: 1423-1431

[187] Lorenzo Sani, Alex Iacob, Zeyu Cao, Bill Marino, Yan Gao, Tomas Paulik, et al. The future of large language model pre-training is federated. arXiv preprint arXiv:2405.10853, 2024

LYU Xiao-Ting, Ph. D., assistant professor. Her main research interests include federated learning and backdoor attacks.

LIU Jing-Kai, Ph. D. candidate. His research interests mainly include LLM security

LIU Zhi-Chen, M. S. candidate. His research interests mainly include federated learning.

CHEN Zheng, Ph. D. candidate. His research interests mainly include federated learning.

XU Guang-Quan, Ph. D., professor. His research interests mainly include AI security.

LUO Wen-Jian, Ph. D., professor. His research interests mainly include AI security.

SHEN Meng, Ph. D., professor. His research interests mainly include AI security.

WANG Bin, Ph. D., professor. His research interests mainly include IoT security.

JI Shou-Ling, Ph. D., professor. His research interests mainly include AI security.

CHEN Kai, Ph. D., professor. His research interests mainly include AI security.

WANG Wei, Ph. D., professor. His research interests mainly include AI security and blockchain security.

Background

Federated learning (FL) is a privacy-preserving distributed machine learning paradigm that enables multiple participants to collaboratively train models while keeping their private data localized. However, as FL continues to grow in application, it faces significant security challenges, particularly from backdoor attacks. These attacks involve embedding backdoors during model training, resulting in models that produce incorrect outputs when triggered by specific inputs. Such vulnerabilities severely compromise the integrity, security, and reliability of the models. Once deployed and activated, backdoor models can have devastating consequences, especially in high-security environments.

In the realm of machine learning, backdoor attacks entail the attacker secretly embedding backdoors into models during the training phase, which leads the model to generate predetermined incorrect outputs under specific conditions after deployment. In FL systems, these attacks are particularly insidious, as attackers can create fake clients or manipulate existing ones to participate in the training process, subtly integrating backdoors into the global model through malicious local updates that affect all clients. Researchers have developed various defense strategies, including model regularization and anomaly detection, to identify and mitigate the risks associated with backdoor attacks. However, the stealthy, diverse, and evolving nature of these attacks presents significant challenges for detection and defense.

In this context, this paper offers an in-depth discussion and analysis of backdoor attacks and defense strategies in FL. We thoroughly review and analyze backdoor attacks applicable to different FL systems from the perspective of the target system. Subsequently, we summarize existing defense methods, categorizing them based on the type of information they utilize. Additionally, we explore the threats posed by backdoor attacks in practical applications, revealing the security challenges that FL may encounter in real-world contexts. Finally, the paper provides a forward-looking discussion on future research directions concerning backdoor attacks and defenses in FL. Our aim is to furnish researchers and practitioners in the FL field with a comprehensive perspective, enabling them to better understand and address security issues while advancing the research on FL security.

This work was supported by the Beijing Natural Science Foundation (L221014), Systematic Major Project of China State Railway GroupCorporation Limited (P2023W002, P2024S003, P2024W001-4), and Hangzhou Qianjiang Distinguished Experts programme (2024).