# 联邦学习安全攻防调研报告

学 生 姓 名　　童 梓 恒

高 校 名 称　　北 京 航 空 航 天 大 学

培 养 院 系　　网 络 空 间 安 全 学 院

完 成 时 间　　2024 年 10 月 1 日

# 目 录

# 第一章 引言

## 1.1 研究背景

随着计算设备的日益普及，人们在日常使用中产生了大量数据。将这些数据集中存储到中央设施既成本高昂又耗时。传统的集中式机器学习方法由于基础设施的不足，如有限的通信带宽、间歇性网络连接和严格的延迟限制，无法支持如此广泛的部署和应用。另一个关键问题是数据隐私和用户保密性，因为使用数据通常包含敏感信息。敏感数据如面部图像、基于位置的服务或健康信息可能被用于针对性的社交广告和推荐，带来直接或潜在的隐私风险。因此，在没有隐私保护的情况下，不应直接共享私人数据。随着社会对隐私保护意识的提高，诸如《通用数据保护条例》（GDPR）之类的法律限制正在出现，这使得数据聚合实践变得不太可行。

在这种背景下，联邦学习（FL）（也称为协作学习），作为一种将模型训练分布到数据来源设备上的新兴替代机器学习范式，应运而生。FL 允许众多参与者构建一个联合机器学习模型，而不会暴露他们的私有训练数据。它还能处理在现实世界中自然出现的不平衡和非独立同分布（non-I.I.D.）数据。近年来，FL 已经惠及了广泛的应用，如下一个单词预测、安全视觉目标检测、实体解析、推荐、工业物联网和基于图的分析。

## 1.2 联邦学习分类

### 1.2.1 基于数据分布的联邦学习分类

根据数据特征和数据样本在参与者之间的分布，联邦学习通常可以分为水平联邦学习（HFL）、垂直联邦学习（VFL）和联邦迁移学习（FTL）。在 HFL 中，每个参与者拥有的数据集具有相似的特征，但涉及不同的用户=。例如，几家医院可能各自存储了关于不同患者的相似类型的数据（如人口统计、临床和基因组数据）。如果他们决定使用 FL 共同构建一个机器学习模型，我们称这样的场景为 HFL。在本文中，我们进一步将 HFL 细分为跨企业的 HFL（Cross-Silo）和跨设备的 HFL（Cross-Device）。他们之间主要区别在于参与者数量、FL 训练参与程度和网络连接状态，这些因素可能影响对手试图破坏 FL 系统的方式。在跨企业的 HFL 中，通常只有少数参与者。他们可以在 FL 训练中频繁被选中。参与者往往拥有显著的计算能力和较高的通信带宽。在跨设备的 HFL 中，可能有成千上万甚至数百万潜在的参与者。在每一轮训练中，只有一部分设备被选

由于他们的数据集往往较小，参与者被反复选中进行 FL 训练的机会很低。它们通常拥有有限的计算能力和较低的通信带宽。

VFL 适用于参与者在样本空间有大量重叠但在特征空间不同的情况，即不同的参与者持有相同记录的不同属性。VFL 主要针对商业参与者。因此，VFL 参与者的特点与 Cross-Silo FL 参与者相似。

如今，FTL 在金融、医疗和保健等行业正受到越来越多的关注。FTL 处理的是 FL 参与者在样本空间和特征空间都很少有重叠的场景。在这种情况下，可以应用迁移学习技术，为联邦下的整个样本和特征空间提供解决方案。FTL 使得在数据联邦中跨域传递互补知识，从而使目标域的参与者能够利用来自源域的丰富标签构建灵活有效的模型。

### 1.2.2 基于模型架构的联邦学习分类

具有同质架构的 FL：通常只有具有同质架构的 FL 才会限制仅共享梯度，即所有参与者共享相同的模型。参与者的目标是协作学习一个更准确的模型。不同的分类器（例如，逻辑回归、深度神经网络）使用不同的目标函数。在 FL 中，每个参与者维护其本地训练数据集的本地模型。服务器通过聚合来自多个参与者的本地模型来维护全局模型。具体来说，具有同质架构的 FL 执行图 1 中的步骤。具有同质架构的 FL 通常有两种形式：(1) FedSGD，其中每个参与者将每个 SGD 更新发送给服务器；(2) FedAvg，其中参与者在将更新发送给服务器之前本地批量多次 SGD 迭代，这在通信上更有效。这些方法都基于平均聚合规则，将本地模型参数的平均值作为全局模型。然而，即使只有一个参与者被破坏，全局模型的平均值也可以被任意操纵。

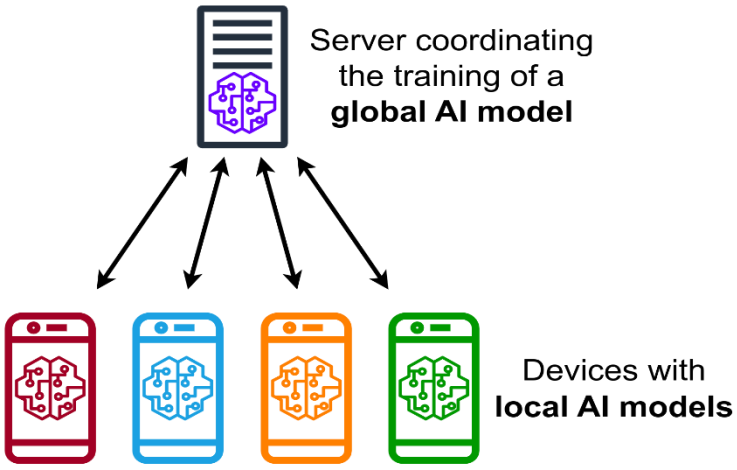


图 1 同质架构的联邦学习框架

具有异构架构的 FL：最近的工作致力于将 FL 扩展到协作训练具有异构架构的模型。传统的联邦模型训练，如果所有本地模型具有相同的模型结构，则可以直接平均模型权重。其限制了具有异构模型架构的数据所有者之间的协作。共享模型预测而不是模型参数或更新可以消除这个障碍，并消除传统联邦学习中白盒推理攻击的风险。与现有的联邦学习算法不同，联邦模型蒸馏（FedMD）不会将单一的全局模型强加给本地模型。相反，它是以简洁、黑盒和模型不可知的方式进行的。每个本地模型分别更新，参与者通过他们在一个未标记的公共集上的预测来共享他们本地模型的知识。共享 logits 的另一个明显好处是降低了通信成本，而不会显著影响效用。

## 1.3 联邦学习安全威胁

FL 提供了一种注重隐私的模型训练范式，它不需要数据共享，并允许参与者自由加入和退出。然而，最近的研究工作表明，FL 可能并不总是提供足够的隐私和鲁棒性保证。现有的 FL 协议设计容易受到以下攻击：(1) 一个恶意服务器，它旨在从随时间推移的个别更新中推断敏感信息，篡改训练过程或控制参与者对全局参数的看法；(2) 任何敌对参与者，他们可以推断其他参与者的敏感信息，篡改全局参数聚合或毒化全局模型。

在隐私泄露方面，在整个训练过程中传递梯度可能会泄露敏感信息，甚至导致深度泄露，无论是对第三方还是中央服务器。即使是梯度的一小部分也可以泄露有关本地数据的相当数量的敏感信息。最近的研究进一步表明，通过简单地观察梯度，恶意攻击者可以成功窃取训练数据。

在鲁棒性方面，FL 系统容易受到数据投毒攻击和模型投毒攻击。恶意参与者可以通过故意改变他们的本地数据（数据投毒）或他们的梯度上传（模型投毒）来攻击全局模型的收敛或在全局模型中植入后门触发器。

更广泛地说，投毒攻击可以分为：(1) 无目标攻击，例如拜占庭攻击，其中对手旨在破坏全局模型的收敛性和性能；以及(2) 有目标攻击，例如后门攻击，其中对手旨在将后门触发器植入全局模型，以便在保持主要任务良好性能的同时，诱使模型在子任务上不断预测敌对类别。

这些隐私和鲁棒性攻击对 FL 构成了重大威胁。在集中式学习中，服务器负责所有参与者的隐私和模型鲁棒性。然而，在 FL 中，任何参与者都可以攻击服务器并窥探其他参与者，有时甚至不需要涉及服务器。因此，了解这些隐私和鲁棒性攻击背后的原则

非常重要。基于服务器的 FL 中代表性隐私和鲁棒性攻击的属性总结在表 1 中。

| 攻击类型 | | 攻击目标 | | 攻击角色 | | 攻击复杂度 | |
|---|---|---|---|---|---|---|---|
| | | 模型 | 数据 | 客户端 | 服务器 | 一轮 | 多轮 |
| **鲁棒性** | 非针对性攻击 | Y | N | Y | N | Y | Y |
| | 针对性攻击 | Y | N | Y | N | Y | Y |
| **隐私** | 类别推断 | N | Y | Y | Y | N | Y |
| | 成员推断 | N | Y | Y | Y | N | Y |
| | 属性推断 | N | Y | Y | Y | N | Y |
| | 训练数据推断 | N | Y | N | Y | Y | Y |

表 1 针对 FL 的各类攻击目标、角色和复杂度

## 1.4 安全联邦学习

对 FL 的攻击要么来自隐私角度，其中恶意参与者或中央服务器试图推断受害者参与者的私有信息时，要么来自鲁棒性角度，其中恶意参与者旨在破坏全局模型。

为了保护 FL 免受隐私攻击，已经尝试将集中式机器学习中的现有隐私保护方法应用于 FL，包括同态加密（HE）、安全多方计算（SMC）和差分隐私（DP）。然而，HE 和 SMC 可能不适用于大规模 FL，因为它们会带来大量的通信和计算开销。在基于聚合的任务中，DP 要求聚合值包含一定幅度的随机噪声，以确保，因此也不理想用于 FL。DP 所需的噪声添加在 FL 中也很难执行。在理想情况下，如果信任服务器（聚合器），服务器可以向聚合梯度添加噪声。然而，在许多现实世界的场景中，参与者可能不信任中央服务器或彼此。在这种情况下，参与者会相互竞争，都想通过向他们的本地梯度添加尽可能多的噪声来确保局部差分隐私（LDP）。这往往会在服务器端累积显著的误差。分布式差分隐私（DDP），在至少有一定比例的参与者是诚实的并且不进行这种恶意竞争时，可以在一定程度上缓解这个问题。

防御 FL 免受各种鲁棒性攻击（例如，无目标的拜占庭攻击，有针对性的后门攻击）是一个极其具有挑战性的任务。这主要有两个原因。首先，防御只能在服务器端执行，其中只有本地梯度可用。这使得许多在集中式机器学习中开发的后门防御方法无效，例如，去噪（预处理）方法，后门样本/触发器检测方法，鲁棒数据增强[53]，微调方法[53]，基于神经注意力蒸馏（NAD）的方法，以及最近的反后门学习方法。其次，防御方法必

须能够抵御数据投毒和模型投毒攻击。大多数现有的鲁棒性防御是梯度聚合方法，主要是为防御无目标的拜占庭攻击者开发的，如 Krum/Multi-Krum，AGGREGATHOR，拜占庭梯度下降（BGD），基于中位数的梯度下降，修剪均值梯度下降和 SIGNSGD。这些防御方法从未在有针对性的后门攻击上进行过测试。已经研究了专门防御数据投毒和模型投毒攻击的方法，例如范数裁剪，基于几何中位数的鲁棒联邦聚合（RFA）和鲁棒学习率。对于 Sybil 攻击的串谋，贡献相似性可以作为防御策略。

## 1.5 综述贡献与报告结构

接下来，我们针对近年来联邦学习中存在的各种攻击与防御手段进行了系统的调研，第二章对威胁模型进行了详细的定义，可以细分为内部攻击者和外部攻击者、训练阶段敌手与推理阶段敌手、半诚实敌手与恶意敌手、目标性攻击与非目标性攻击。第三章对隐私攻击与防御手段进行了系统的调研，其中隐私攻击可以划分为类别推断攻击、成员推断攻击、属性推断攻击、训练数据推断攻击，隐私防御手段可以分为同态加密、多方安全计算、差分隐私。第四章系统介绍了投毒攻击与防御，投毒攻击可以分为非针对性攻击和针对性攻击，而针对投毒攻击的手段可以划分为面向非针对性攻击的防御和面向针对性攻击的防御。最后，第 5 章，我们对联邦学习的研究热点与未来研究方向做了总结与前瞻，旨在为后续的联邦学习安全攻防研究提供指导作用。

# 第二章 威胁模型

在回顾对 FL 的攻击之前，我们首先总结一下威胁模型。一般而言，FL 中的威胁模型可以分为两种类型：(1) 内部敌手与外部敌手；(2) 训练阶段与推理阶段。这些威胁模型适用于隐私和鲁棒性。此外，隐私和鲁棒性各自也有其特定的威胁模型。

## 2.1 内部敌手与外部敌手

攻击可以由内部成员和外部成员发起。内部攻击包括由 FL 服务器和 FL 系统中的参与者发起的攻击。外部攻击包括由监听参与者和 FL 服务器之间通信渠道的窃听者发起的攻击，以及在最终的 FL 模型作为服务部署时由最终用户发起的攻击。

内部攻击通常比外部攻击更危险，因为这严格提高了对手的能力。因此，我们对 FL 攻击的讨论将主要关注内部攻击。

## 2.2 训练阶段敌手与推理阶段敌手

训练阶段。在训练阶段进行的攻击试图学习、影响或破坏 FL 模型本身。在训练阶段，攻击者可以运行数据投毒攻击来破坏训练数据集的完整性，或者运行模型投毒攻击来破坏学习过程的完整性。攻击者还可以在训练阶段对单个参与者的更新或所有参与者的聚合更新发起一系列推理攻击。

推理阶段。在推理阶段进行的攻击被称为规避或探索性攻击。它们通常不会改变目标模型，而是欺骗它产生错误的预测（针对性/非针对性），或收集有关模型特征的证据，导致隐私和鲁棒性问题。这些攻击的有效性在很大程度上取决于对手对模型的了解程度。推理阶段的攻击可以分为白盒攻击（完全访问 FL 模型）和黑盒攻击（只能查询 FL 模型）。在 FL 中，服务器维护的全局模型在目标模型作为服务部署时，遭受的规避攻击与常规机器学习（ML）环境中的相同。虽然在集中式设置中可能更自然地考虑黑盒攻击，但 FL 中的模型广播步骤使得全局模型对任何恶意参与者来说都是白盒。因此，FL 需要额外的努力来防御白盒规避攻击。

## 2.3 半诚实敌手与恶意敌手

半诚实敌手。对手被认为是被动的或诚实但好奇的。他们试图在不偏离 FL 协议的情况下了解其他参与者的私有状态。对手只能观察到接收到的信息，即全局模型的参数。

恶意敌手。一个主动的或恶意的对手试图了解诚实参与者的私有状态，并通过修改、重放或删除消息任意地偏离 FL 协议。这种设置允许对手进行特别具有破坏性的攻击。

## 2.4 目标性攻击与非目标性攻击

无目标攻击：无目标投毒攻击旨在任意破坏目标模型的完整性。拜占庭攻击是一种无目标投毒攻击，它向服务器上传任意恶意的梯度，以导致全局模型的失败。

有目标攻击：有目标投毒攻击诱导模型对特定测试样本输出由对手指定的目标标签，而其他测试样本的测试错误不受影响。

# 第三章 隐私攻击与防御

## 3.1 隐私攻击

虽然 FL 防止参与者直接共享他们的私有数据，但一系列研究已经证明，在 FL 中交换梯度也可以泄露有关参与者私有数据的敏感信息，无论是对于被动还是主动攻击者。例如，梯度或 FL 模型参数的两个连续快照可能会泄露参与者训练数据的意外特征给敌对参与者，因为深度学习模型倾向于识别并记住比主学习任务所需的更多的数据特征。图 2 说明了对手可以从梯度推断出的信息集。梯度可能导致隐私泄露的原因是梯度源自参与者的私有训练数据，学习模型可以被视为其训练数据集的高级统计特性的表示[75]。在深度学习模型中，给定层的梯度是根据该层的特征和该层之后的误差（即，反向传播）计算的。在序列全连接层的情况下，权重的梯度是当前层特征和该层之后的误差的内积。类似地，对于卷积层，权重的梯度是该层特征和该层之后的误差的卷积[28]。因此，梯度的观察可以用来推断大量的私有信息，如类别代表、训练数据子集的成员资格和属性。更糟糕的是，攻击者可以从共享的梯度中推断出标签，并在没有任何关于训练数据的先验知识的情况下恢复原始训练样本。接下来，我们根据攻击者目标的敏感信息类型详细说明 FL 的潜在隐私泄露。梯度可能导致隐私泄露的原因是梯度源自参与者的私有训练数据，学习模型可以被视为其训练数据集的高级统计特性的表示。在深度学习模型中，给定层的梯度是根据该层的特征和该层之后的误差（即，反向传播）计算的。在序列全连接层的情况下，权重的梯度是当前层特征和该层之后的误差的内积。类似地，对于卷积层，权重的梯度是该层特征和该层之后的误差的卷积。因此，梯度的观察可以用来推断大量的私有信息，如类别代表、训练数据子集的成员资格和属性。更糟糕的是，攻击者可以从共享的梯度中推断出标签，并在没有任何关于训练数据的先验知识的情况下恢复原始训练样本。接下来，我们根据攻击者目标的敏感信息类型详细说明 FL 的潜在隐私泄露。

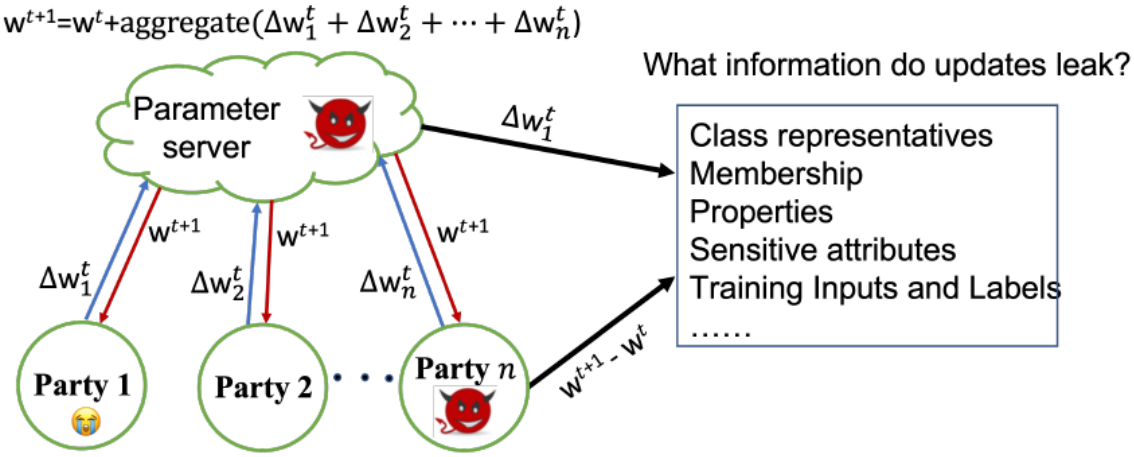


图 2 梯度能够泄露的信息集合

### 3.1.1 类别推断攻击

Hitaj 等人[1]首次设计了一种名为生成对抗网络（GAN）攻击的主动推理攻击，针对深度联邦学习（FL）模型。在这种攻击中，一个恶意参与者可以有意地破坏任何其他参与者。GAN 攻击利用了 FL 学习过程的实时性，允许敌对参与者训练一个 GAN 来生成目标私有训练数据的原型样本。生成的样本看起来像是来自与训练数据相同的分布。因此，GAN 攻击的目标不是重建确切的训练输入，而只是类别代表。值得注意的是，GAN 攻击假设给定类别的整个训练语料库来自单一参与者，这意味着 GAN 构建的代表只有在所有类别成员都相似时才与训练数据相似。这类似于集中式机器学习设置中的模型反演攻击。需要注意的是，这些假设在 FL 中可能不太实际。由于 GAN 攻击需要大量的计算资源来训练 GAN 模型，因此它不太适合跨设备场景。

### 3.1.2 成员推断攻击

给定一个确切的数据点，成员资格推理攻击（MIA）旨在确定它是否被用来训练模型[2]。例如，攻击者可以推断某个特定患者的病历是否被用来训练与某种疾病相关的分类器。FL 为此类攻击开辟了新的可能性。在 FL 中，对手可以推断一个特定样本是否属于某个特定参与者的私有训练数据（如果目标更新来自单个参与者）或任何参与者（如果目标更新是聚合的）。例如，在 FL 模型训练期间，针对文本数据训练的深度自然语言处理模型的嵌入层的非零梯度可以揭示诚实参与者的训练批次中哪些单词[3]。

FL 系统中的攻击者可以进行主动和被动的成员资格推理攻击[4],[3]。在被动情况下，攻击者观察更新的模型参数并进行推理，而不修改学习过程。在主动情况下，攻击者可

以篡改 FL 模型训练协议，并对其他参与者进行更有力的攻击。例如，攻击者可能共享恶意更新，并诱使 FL 模型暴露更多关于其他参与者本地数据的信息。一种这样的攻击是梯度上升攻击[4]，攻击者对目标数据样本运行梯度上升，并观察其增加的损失在下一轮通信中是否可以大幅降低，如果可以，该样本很可能在训练集中。这种攻击可以同时应用于一批目标数据样本[4]。

#### 3.1.3 属性推断攻击

攻击者可以发起被动和主动的属性推理攻击，以推断其他参与者训练数据的某些属性[3]。属性推理攻击假设攻击者拥有正确标记了目标属性的辅助训练数据。被动攻击者只能观察或窃听梯度，并通过训练一个二元属性分类器来进行推理。主动攻击者可以利用多任务学习来欺骗 FL 模型，使其学会更好地区分具有和不具有目标属性的数据，以提取更多信息。敌对参与者还可以推断属性何时出现在训练数据中或从训练数据中消失（例如，识别一个人首次出现在用于训练性别分类器的照片上）。属性推理攻击中对辅助训练数据的假设可能限制了其在跨设备场景中的适用性。

#### 3.1.4 训练数据推断攻击

最近的一项名为“从梯度中深度泄露（DLG）”的研究提出了一种优化算法，用于提取训练输入和标签[5]。这种攻击比以前的方法更强大，能够准确地恢复用于训练深度学习模型的原始图像和文本。在后续的研究[6]中，提出了一种名为“改进的从梯度中深度泄露（iDLG）”的分析方法，基于共享的梯度和对标签与梯度符号之间相关性的探索来提取标签。iDLG 可以应用于攻击任何使用交叉熵损失和独热标签训练的不同分模型，这是分类任务的典型设置。

总之，推理攻击通常假设对手拥有精湛的技术能力，并且拥有无限的计算资源。此外，大多数攻击假设敌对参与者可以在 FL 训练过程的许多轮次中被选中（以更新全局模型）。在 FL 中，这些假设在 H2C 场景中通常不切实际，但更有可能发生在 H2B 场景中。这些推理攻击强调了 FL 中梯度保护的必要性，可能通过各种隐私保护机制实现。

### 3.2 隐私防御

虽然隐私保护在机器学习界已经被广泛研究，但由于对电力和网络连接的间歇性访问、数据的统计异质性等因素，联邦学习中的隐私保护可能更具挑战性。现有的隐私保

护联邦学习工作大多基于众所周知的隐私保护技术，包括：(1) 同态加密（HE），如Paillier[7]、Elgamal[8]和 Brakerski-Gentry-Vaikuntanathan 密码体系[9]；(2) 安全多方计算（SMC），如混乱电路[10]和秘密共享[11]；以及(3) 差分隐私（DP）[12][13]

### 3.2.1 同态加密

同态加密方案允许直接在密文上执行算术运算，这相当于对明文进行特定的线性代数操作。现有的同态加密技术可以分为：1) 完全同态加密，2) 部分同态加密，以及 3) 有限同态加密。完全同态加密可以支持对密文的任意计算，但效率较低。另一方面，部分同态加密和有限同态加密效率更高，但只能执行有限次数的操作。有限同态加密方案在实践中更广泛使用，包括 RSA、El Gamal、Paillier 等。

同态加密被广泛使用，尤其适用于通过在加密数据上进行计算来保护学习过程。然而，在加密数字上进行算术运算需要付出内存和处理时间的代价。例如，使用 Paillier 加密方案，一个编码的浮点数（无论是单精度还是双精度）的加密长度为 2m 位，其中 m 通常至少为 1024，两个加密数字的加法比未加密的等效操作慢 2-3 个数量级。此外，需要进行多项式近似来评估机器学习算法中的非线性函数，这导致了效用和隐私之间的权衡。例如，为了保护个人梯度，Aono 等人[14]使用了加性同态加密来保护梯度的隐私并增强分布式学习系统的安全性。然而，他们的协议不仅带来了巨大的通信和计算开销，而且还导致了效用损失。此外，它无法抵御服务器和多个参与者之间的串谋。Hardy 等人[15]在垂直分割的数据上应用了同态加密的联合逻辑回归，以抵御诚实但好奇的对手。总的来说，所有这些工作都带来了额外的通信和计算开销，这限制了它们在 H2C 场景中的应用。

### 3.2.2 多方安全计算

安全多方计算（SMC）[10]允许具有私有输入的不同参与者在不向彼此透露的情况下，对他们的输入进行联合计算。Mohassel[16]等人提出了 SecureML，它通过 SMC 进行隐私保护学习，其中数据所有者需要在初始设置阶段处理、加密和/或在两个不串通的服务器之间秘密共享他们的数据。SecureML 允许数据所有者在不透露任何超出结果的信息的情况下，对他们的联合数据进行各种模型的训练。然而，这需要付出高昂的计算和通信开销，这可能会阻碍参与者合作的兴趣。Bonawitz 等人[17]提出了一种基于 SMC 的安全、通信高效且抗故障的协议，用于安全聚合个人梯度。它确保服务器唯一了解的关于个人

用户的信息是从聚合结果中推断出来的。他们的协议在诚实但好奇和恶意设置下都保持安全，即使服务器和一部分用户恶意行事——串通并随意偏离协议。也就是说，没有任何一方能了解到比一大群诚实用户的输入总和更多的信息[17]。

通常，SMC 技术确保了高度的隐私和准确性，但代价是高昂的计算和通信开销，从而不利于吸引参与。基于 SMC 的方案面临的另一个主要挑战是在整个训练过程中需要所有参与者的同时协调。这种多方互动模型在实际设置中可能不可取，特别是在 FL 设置中通常考虑的参与者-服务器架构下。此外，基于 SMC 的协议可以使多个参与者协作计算一个商定的函数，而不会泄露任何参与者的输入信息，除了可以从计算结果中推断出的信息。也就是说，SMC 无法完全保证防止信息泄露，这需要将额外的差分隐私技术纳入多方协议中以解决此类问题。

总之，基于同态加密或 SMC 的方法可能不适用于大规模 FL 场景，因为它们会带来大量的额外通信和计算成本。此外，基于加密的技术需要为目标学习算法中的每个操作精心设计和实现。最后，所有基于密码学的协议都阻止任何人审计参与者对联合模型的更新，这为恶意参与者留下了攻击的空间。例如，恶意参与者可以在不被检测到的情况下向全局模型引入隐蔽的后门功能。

### 3.2.3 差分隐私

差分隐私（DP）最初是为单一数据库场景设计的，在这个场景中，对于每个查询，数据库服务器都会以一种保护隐私的方式回答查询，并且通过量身定制的随机化处理。与基于加密的方法相比，差分隐私通过以一种方式扰动数据来权衡隐私和准确性，这种方式（i）计算效率高，（ii）不允许攻击者恢复原始数据，以及（iii）不会严重影响效用。

差分隐私的概念是，单个记录的存在或缺失对输出可能性的影响被一个小因子所限制。正如如下所定义的，$(\epsilon, \delta)$-近似差分隐私[18]通过一个 $\delta$ 附加项放宽了纯粹的 $\epsilon$ -差分隐私，这意味着不太可能的响应不需要满足纯粹的差分隐私标准。

$$\Pr\{\mathcal{M}(D) \in S\} \quad \leq \quad \exp(\epsilon) \cdot \Pr\{\mathcal{M}(D') \in S\} + \delta \quad .$$

隐私社区通常根据不同的信任假设和噪声源将差分隐私（DP）分为以下三类：集中式差分隐私（CDP）、本地差分隐私（LDP）和分布式差分隐私（DDP）。

**集中式差分隐私：**集中式差分隐私（CDP）。CDP 最初是为集中式场景设计的，在这个场景中，一个受信任的数据库服务器有权查看所有参与者的明文数据，希望以保护

隐私的方式回答查询或发布统计数据，通过随机化查询结果来实现。当CDP遇到FL时，CDP假设有一个受信任的聚合器，负责向聚合的本地梯度添加噪声，以确保所有参与者数据的记录级隐私。然而，CDP旨在应对成千上万的用户进行训练以实现收敛，并在隐私和准确性之间取得可接受的权衡，导致参与者数量较少时的收敛问题。此外，CDP只有在大量参与者的情况下才能达到可接受的准确性，因此不适用于参与者相对较少的H2B场景。

同时，CDP中对受信任服务器的假设在许多应用中并不适用，因为它构成了数据泄露的单点故障，并且使受信任的管理员承担了保护用户数据安全的法律和道德义务。当聚合器不可信时，这在分布式场景中经常发生，需要本地差分隐私（LDP）或分布式差分隐私（DDP）来保护个人隐私。

**本地差分隐私：**本地差分隐私（LDP）。本地差分隐私（LDP）[19]提供了更强的隐私保证，数据所有者在将私有信息报告给不受信任的数据管理者之前，会先对其进行扰动以满足局部的差分隐私。尽管随机响应及其变体已被广泛用于在个人披露个人信息时提供LDP，我们注意到所有用于CDP的随机化机制，如拉普拉斯机制和高斯机制[96]，每个参与者都可以单独使用，以确保孤立的LDP。然而，在分布式场景中，没有密码学技术的帮助，每个参与者必须添加足够的校准噪声以确保LDP。因此，LDP的吸引力伴随着巨大的效用降低，特别是在有数十亿个体的情况下。

**分布式差分隐私：**分布式差分隐私（DDP）。分布式差分隐私（DDP）在确保每个人的隐私的同时，通过结合密码学协议，填补了LDP和CDP之间的差距。因此，DDP避免了对任何服务器的信任，并且比LDP提供了更好的效用。从理论上讲，DDP提供了与CDP相同的效用，因为总的噪声量是相同的。

DDP 的概念反映了目标统计数据所需的噪声来自多个参与者。实现整体加性噪声机制的方法，通过在每个参与者处运行相同的机制（通常噪声较小）进行求和，需要具有稳定分布的机制——以保证已知的端到端响应分布的正确校准——以及用于隐藏所有但最终结果的密码学。稳定分布包括高斯分布、二项分布等，即，高斯随机变量的和仍然遵循高斯分布，二项随机变量的和仍然遵循二项分布。DDP利用这种良好的稳定性，允许每个参与者对其本地统计数据进行较小程度的随机化，而不是LDP

# 第四章 投毒攻击与防御

## 4.1 投毒攻击

与针对数据隐私的隐私攻击不同，投毒攻击旨在破坏系统的鲁棒性。根据攻击者的目标，投毒攻击可以分为两大类：1)针对性投毒攻击 2) 无针对性投毒攻击。

在训练阶段的无针对性和有针对性投毒攻击可以同时针对数据和模型进行。图 3 显示，被污染的更新可能来自两种投毒攻击：(1) 在本地数据收集期间的数据投毒攻击；以及(2) 在本地模型训练过程中的模型投毒攻击。从高层次上看，这两种投毒攻击都试图以某种不良方式改变目标模型的行为。然而，由于同质架构的 FL 的模型共享特性，数据投毒攻击通常不如模型投毒攻击有效。实际上，在 FL 设置中，模型投毒包括数据投毒，因为数据投毒攻击最终会改变在任何给定迭代中发送到模型的更新子集。这在功能上等同于集中式投毒攻击，其中训练数据的子集被污染。

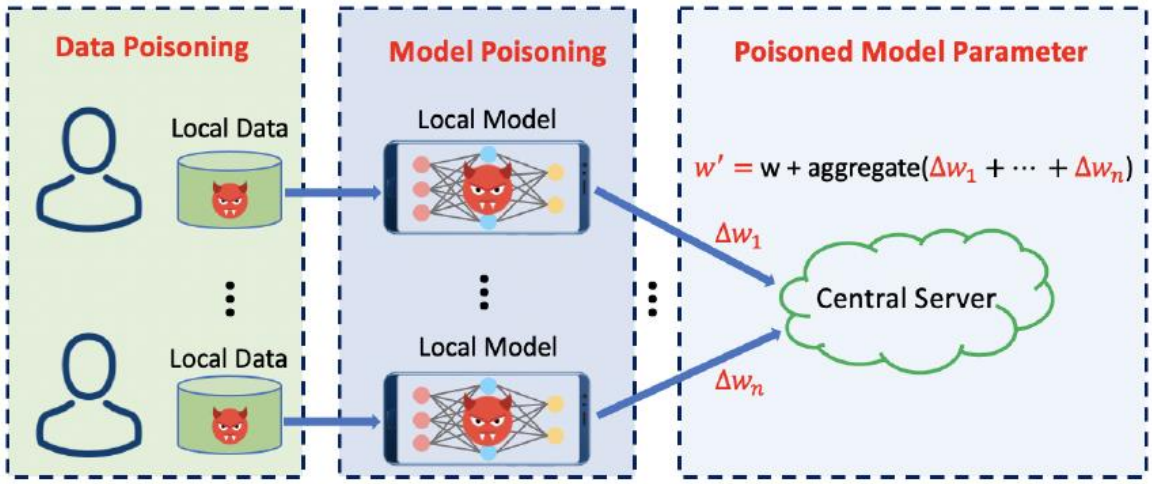


图 3 联邦学习数据与模型投毒攻击

### 4.1.1 非针对性攻击

非针对性投毒攻击旨在任意破坏目标模型的完整性。拜占庭攻击是一种无目标投毒攻击，它向服务器上传任意恶意的梯度，以导致全局模型的失败。拜占庭攻击指的是良性客户端上传本地最优梯度，而拜占庭节点则上传任意梯度

Blanchard 等人[20] 展示了如果没有防御措施，FL 的聚合可以被单个拜占庭参与者完全控制。特别是，假设有 n-1 个良性参与者和一个拜占庭参与者，服务器可以通过以下方式聚合梯度$\Delta w_0 = \frac{1}{n}\sum_{i=1}^{n} \Delta w_i$，其中$\Delta w_0$是聚合梯度。假设第 n 个参与者是拜占庭参

与者，它总是通过上传或下载梯度使聚合梯度变成任何向量。这种简单的攻击揭示了联邦学习对拜占庭攻击的不可抵御性。Chen 等人[21]讨论了基于 Adam 优化器的 FL 中的拜占庭攻击，并提出了一种伪装攻击，可以伪装模型更新并发动有效的攻击。他们提出的攻击也适用于其他众所周知的优化器，如 AdaGrad 和 RMSProp。Baruch 等人[22]展示了梯度下降算法的核心部分是下降的方向。具体来说，对于梯度下降算法，为了确保损失的下降，真实梯度和鲁棒聚合梯度之间的内积必须是非负的。为了使聚合失败，他们提出了一种“内积操纵攻击”，可以使真实梯度和鲁棒聚合梯度之间的内积为负。为此，每个拜占庭参与者上传了平均良性梯度的负数。Xie 等人[23]声称，通过一致地对许多参数应用小的更改，拜占庭参与者可以扰乱模型的收敛。首先，他们使用拜占庭参与者的本地数据来估计分布的均值和标准差。然后，他们分析了参数更改不会被防御机制检测到的范围，并在选择这个范围的最大值时，可以阻碍收敛。

#### 4.1.2 针对性攻击

在有针对性的投毒攻击中，学习到的模型对特定的测试样本输出由对手指定的目标标签，例如，将垃圾邮件预测为非垃圾邮件，以及对具有特定特洛伊木马触发器（后门/特洛伊木马攻击）的测试样本预测攻击者期望的标签。然而，其他测试样本的测试错误不受影响。通常，有目标攻击比无目标攻击更难进行，因为攻击者有特定的目标要实现。

有针对性的投毒攻击的一个常见例子是标签翻转攻击[24]。一个类别的诚实训练样本的标签被翻转到另一个类别，而数据的特征保持不变。例如，系统中的恶意参与者可以通过将所有的 1 翻转成 7 来污染他们的数据集。一次成功的攻击会产生一个无法正确分类 1 并将它们错误预测为 7 的模型。

另一种现实的有目标投毒攻击是后门投毒攻击，在这种攻击中，对手可以修改原始训练数据集的个别特征或小区域，以在模型中植入后门触发器。模型在干净数据上将表现正常，但一旦触发器（例如，图像上的一个标记）出现，就会持续预测目标类别。例如，后门攻击可以使 FL 模型在后门任务上达到 100%的准确率，例如，控制一个图像分类器，将攻击者选择的标签分配给具有某些特征的图像，在图像分类任务中，或者在单词预测任务中，一个下一个单词预测器用攻击者选择的单词完成某些句子。

后门攻击可以进一步细分为两类： 脏标签攻击和干净标签攻击。干净标签攻击假设对手不能改变任何训练数据的标签，因为有一个过程可以认证数据属于正确的类别，

并且数据样本的污染必须是不可察觉的。相比之下，在脏标签污染中，对手可以引入一些数据样本，这些样本预计会被模型错误分类为所需的目标标签，并将这些样本放入训练数据中。干净标签攻击可以说是更隐蔽的，因为它们不改变标签。

在联邦学习中，针对性的污染攻击可以由任何 FL 参与者执行，或者通过在数据或梯度上的勾结来实现。Bhagoji 等人[25]展示了一个单独的、不勾结的恶意参与者可以使模型对一组选定的输入以高置信度进行错误分类。Bagdasaryan 等人[26]指出，污染的更新可以通过在后门本地训练数据上训练本地模型来生成，甚至一次性攻击就足以将后门注入到全局模型中。 Xie 等人[27]展示了全局触发模式可以被分解成独立的本地模式，并分别嵌入到勾结的敌对参与者的训练集中。对 FL 模型的影响取决于后门参与者参与攻击的程度，以及被污染的训练数据的数量。最近的一项研究表明，污染边缘案例（低概率）训练样本更为有效[28]。

最后，我们注意到，大多数关于污染攻击的研究都集中在拜占庭或后门攻击者上。一个允许参与者加入和离开的系统容易受到女巫攻击[29]的影响，在这种攻击中，攻击者通过加入系统来注入 c 个假参与者到 FL 系统中，或者破坏 c 个良性参与者[30]来获得影响力。女巫攻击可以以无目标和有目标的方式发起。例如，有目标的污染可以通过女巫克隆体进行，它们贡献更新以实现特定的污染目标[30]。具体来说，[30]考虑了女巫克隆体进行的两种类型的有目标攻击：标签翻转攻击和后门攻击。

## 4.2 投毒防御

在联邦学习（FL）中，对污染攻击的鲁棒性是一个理想的属性。为了应对污染攻击，文献中提出了许多鲁棒的聚合方案。在集中式环境中已知的对污染攻击的防御措施，如鲁棒损失和异常检测，假设对参与者有控制权或明确观察到训练数据。这些假设都不适用于 FL，其中服务器只观察到作为迭代机器学习算法一部分发送的模型参数/更新。我们总结了针对针对性攻击和非针对性攻击的以鲁棒性为重点的 FL 防御措施如下：

### 4.1.1 面向非针对性攻击的防御

对于抵御拜占庭攻击的聚合算法，如果即使大部分参与者是敌对的，算法的收敛性仍然稳健，那么这个算法就是拜占庭容错的。Shen 等人[31]引入了一种名为 AUROR 的统计机制，在生成准确模型的同时检测恶意用户。AUROR 基于这样的观察：大多数诚实用户的指示性特征（最重要的模型特征）将显示出类似的分布，而恶意用户的则将显

示出异常的分布。然后它使用 k-means 对参与者的更新进行聚类，并丢弃异常值，即，超出阈值距离的小聚类的贡献被移除。即使 30%的所有用户都是敌对的，使用 AUROR 训练的模型的准确性仅下降 3%。

Blanchard 等人提出了 Krum[32]，在该机制中，从模型中移除距离平均参与者贡献最远的前 f 个贡献。Krum 使用欧几里得距离来确定应该移除哪些梯度贡献，理论上可以抵御多达 33%的敌对参与者的污染攻击，即，给定 n 个代理，其中 f 个是拜占庭的，Krum 要求 $n \geq 2f + 3$。在任意时间步$t$，更新$\{\delta_1^t, \dots, \delta_n^t\}$全部发送到服务器。对于某个梯度$\delta_i^t$，$n - f - 2$最邻近的其他梯度形成一个集合$C_i$，他们之间的距离累加得出每个梯度的分数$S(\delta_i^t) = \sum_{\delta \in C_i} ||\delta_i^t - \delta||$, Krum 算法选出分数最低的梯度来进行更新。Krum 能够抵御全知敌对者的攻击——他们了解梯度的良好估计——这些敌对者发送的是乘以一个大因子的相反向量。它也能够抵御敌对者发送的从高斯分布中随机抽取的向量的攻击（分布的方差越大，攻击越强）。Multi-Krum 是 Krum 的一个变体，直观上在 Krum 和平均值之间进行插值，从而结合了 Krum 的鲁棒性属性和平均值的收敛速度。本质上，Krum 基于整个更新向量过滤异常值，但不会过滤逐坐标的异常值。

为了解决这个问题，Yin 等人[33]提出了两种鲁棒的分布式梯度下降算法，一种基于逐坐标的中位数，另一种基于逐坐标的截断均值。不幸的是，基于中位数的规则在大规模设置中可能会带来巨大的计算开销。Mhamdi 等人[34]提出了一种名为 Bulyan 的元聚合规则，这是一个基于 Krum 和截断中位数的两步元聚合算法，先过滤恶意更新，然后计算剩余更新的截断中位数。。Pillutla 等人[35]提出了一种名为 RFA 的鲁棒聚合方法，通过用近似几何中位数替换加权算术平均值，以减少被污染更新的影响。不幸的是，RFA 只能处理少数类型的污染攻击者，而不适用于拜占庭攻击。

尽管它们提供了鲁棒性保证，但最近的研究揭示了以前的拜占庭鲁棒 FL 机制也相当脆弱，容易被轻易绕过。Bhagoji 等人[36]展示了即使针对拜占庭鲁棒聚合规则（如 Krum 和逐坐标中位数）的深度神经网络的有目标模型污染也是有效的。虽然拜占庭鲁棒聚合规则可能确保拜占庭工作者在任何一轮中的影响是有限的，但攻击者可以在各轮之间联合他们的攻击，显著地将权重从期望的方向移开，从而实现降低模型质量的目标。Xu 等人[37]证明了 Multi-Krum 对无目标污染不鲁棒。这是因为 Multi-Krum 基于每个梯度向量和平均向量之间的距离，而平均向量对无目标污染不鲁棒。Fang 等人[38]展示了聚合规

则，这些规则声称对拜占庭故障鲁棒，在实践中对优化的本地模型污染攻击并不有效，这些攻击精心制作了被破坏参与者上的本地模型，使得聚合的全局模型最大程度地偏离了没有攻击时全局模型变化的相反方向。所有这些都强调了在 FL 中对拜占庭攻击者需要更有效的防御。

其他工作从不同的角度研究拜占庭鲁棒性。Chen 等人[39]提出了 DRACO，这是一个通过算法冗余实现鲁棒分布式训练的框架。DRACO 对任意恶意计算节点都是鲁棒的，同时比最先进的鲁棒分布式系统快几个数量级。然而，DRACO 假设每个参与者都可以访问其他参与者的数据，这限制了它在 FL 中的实用性。Su 等人[40]提出了基于 Steinhardt 等人[41][40]提出的过滤过程，对拜占庭参与者计算的梯度进行鲁棒聚合。Bernstein 等人[42]提出了 SIGNSGD，它与多数投票相结合，使参与者能够上传他们梯度的逐元素符号，以防御三种类型的半“盲”拜占庭对手：(i) 任意重新缩放他们随机梯度估计的对手；(ii) 随机化随机梯度每个坐标符号的对手；(iii) 反转他们随机梯度估计的对手。

### 4.1.2 面向针对性攻击的防御

针对有目标的后门攻击的现有防御措施可以分为两类：检测方法和擦除方法。检测方法利用激活统计或模型属性来确定一个模型是否被后门植入，或者一个训练/测试示例是否是一个后门示例。

有许多检测算法被设计用来检测哪些输入包含后门，以及模型的哪些部分（特别是其激活函数）负责触发模型的对抗性行为，以便移除后门。这些算法依赖于被污染模型中后门启用和干净（良性）输入的潜在表示之间的统计差异。然而，这些后门检测算法可以通过最大化后门启用的对抗性输入和干净输入的潜在不可区分性来绕过。

虽然检测可以帮助识别潜在风险，但被后门植入的模型仍然需要被净化，因为后门触发器的潜在影响在被后门植入的模型中仍然没有被清除。擦除方法更进一步，旨在净化后门触发器对模型造成的不利影响。目前最先进的擦除方法是模式连接性修复（MCR）[43]和神经注意力蒸馏（NAD）[44]。MCR 通过在损失景观的路径中选择一个鲁棒模型来减轻后门，而 NAD 利用知识蒸馏来擦除触发器。其他先前的方法，包括微调、去噪和精细修剪，已被证明对最新攻击不足。另一项较新的工作称为反后门学习（ABL）[55]旨在给定后门污染数据的情况下训练干净模型。他们将整个学习过程框架化为学习数据的干净部分和后门部分的双重任务。从这个角度来看，他们识别了后门攻击的两个固有

特征作为它们的弱点：1) 模型学习后门数据比学习干净数据要快得多，攻击越强，模型在后门数据上的收敛就越快；2) 后门任务与特定类别（后门目标类别）相关联。基于这两个弱点，ABL 引入了一个两阶段的梯度上升机制，用于标准训练，以 1) 在早期训练阶段帮助隔离后门示例，以及 2) 在后期训练阶段打破后门示例与目标类别之间的相关性。在多个基准数据集上对 10 种最先进的攻击进行的广泛实验经验表明，ABL 可以在训练期间自动防止后门攻击，而不会降低主要性能。

尽管在集中式环境中取得了有希望的后门防御结果，但目前尚不清楚这些防御措施是否可以顺利适应 FL 环境，特别是在非独立同分布（non-iid）设置中。对于 FL 中的后门防御，Sun 等人表明，剪裁模型更新的范数并添加高斯噪声可以减轻基于模型替换范式的后门攻击。Andreina 等人在每一轮 FL 中增加了一个额外的验证阶段以检测后门。然而，这些都没有提供经过认证的鲁棒性保证。对 FL 的后门攻击的认证鲁棒性在很大程度上仍未被探索。Xie 等人提供了第一个名为 Certifiably Robust Federated Learning (CRFL)的通用框架，用于训练对后门具有认证鲁棒性的 FL 模型。

为了防御女巫克隆节点进行的有目标污染攻击，Fung 等人利用了女巫克隆节点之间的相似性比诚实客户端之间的相似性更高的特征行为，并提出了 FoolsGold：一种通过根据贡献相似性调整参与者的学习率来防御 FL 赛比尔攻击的新防御方案。请注意，FoolsGol 不通过假设攻击者可以生成大量女巫节点来限制预期的攻击者数量，这使得关于诚实参与者比例的假设变得不切实际。此外，FoolsGold 不需要学习过程之外的任何辅助信息，并且对参与者及其数据的假设更少。FoolsGold 的鲁棒性适用于参与者数据的不同分布、不同的污染目标和各种赛比尔策略，并且可以成功应用于 FedSGD 和 FedAvg。

# 第五章　研究热点与未来研究方向

我们对近五年来的联邦学习安全攻防做了调研，主要研究工作如下：

联邦学习攻击：

现有的针对联邦学习的模型投毒攻击假设攻击者能够接触到大量被破坏的真实客户端。然而，这种假设在涉及数百万客户端的生产联邦学习系统中并不现实。在这项工作中，文献[47]提出了第一种基于假客户端的模型投毒攻击，称为 MPAF。具体来说，假设攻击者将假客户端注入到联邦学习系统中，并在训练期间向云服务器发送精心制作的假本地模型更新，使得学习到的全局模型对许多不加选择的测试输入具有低准确性。为了实现这一目标，我们的攻击将全局模型拖向攻击者选择的具有低准确性的基础模型。具体来说，在联邦学习的每轮中，假客户端制作指向基础模型的假本地模型更新，并在发送给云服务器之前将它们放大以增强它们的影响。我们的实验表明，即使采用了传统的防御措施和范数裁剪，MPAF 也能显著降低全局模型的测试准确性，这突出了需要更先进的防御措施的必要性。

现有的 FL 攻击和防御方法通常关注整个模型。它们都没有认识到后门关键（BC）层的存在——这是一小部分主导模型漏洞的层。攻击 BC 层相当于攻击整个模型，但被最先进的（SOTA）防御机制检测到的机会要小得多。文献[49]提出了一种通用的原位方法，从攻击者的角度识别和验证 BC 层。基于识别的 BC 层，精心设计了一种新的后门攻击方法，该方法在各种防御策略下自适应地寻求攻击效果和隐蔽性之间的基本平衡。广泛的实验表明，该文献中 BC 层感知后门攻击可以在只有 10%的恶意客户端的情况下成功地对七种 SOTA 防御下的 FL 进行后门攻击，并超越了最新的后门攻击方法。

文献[49]提出了一种新颖的 FL 后门攻击框架，即不可逆后门攻击（IBA），它联合学习最优和视觉上隐蔽的触发器，然后逐渐将后门植入全局模型。这种方法允许敌手执行可以逃避人类和机器检查的后门攻击。此外，其通过选择性地毒害模型参数，这些参数最不可能被主任务的学习过程更新，以及将被毒害的模型更新限制在全局模型的附近，从而提高了所提出攻击的效率和持久性。最后，算法在包括 MNIST、CIFAR-10 和 Tiny ImageNet 在内的几个基准数据集上评估了所提出的攻击框架，并在同时绕过现有的后门防御措施的同时，实现了比其他后门攻击更持久的后门效果，并取得了高成功率。总的来说，IBA 为 FL 中的后门攻击提供了一种更有效、隐蔽和持久的方法。

文献[50]提出了 Neurotoxin，这是一种简单的一行后门攻击，它通过攻击在训练期间变化幅度较小的参数来起作用。在十个自然语言处理和计算机视觉任务上进行了全面的评估，并发现文献可以通过添加 Neurotoxin 的单行代码将最先进的后门持久性提高一倍。

联邦学习防御：

文献[46]设计了一个算法框架，包括一个通用的联邦 LCB 算法和灵活的隐私协议。然后，利用所提出的框架，我们在两种不同的隐私约束下研究联邦 LCB。我们首先在数据孤岛级别的本地差分隐私下建立隐私和遗憾保证，这修正了现有算法中存在的问题。为了进一步提高遗憾性能，我们接下来考虑差分隐私的洗牌模型，在这个模型下，我们展示了我们的算法可以在没有可信服务器的情况下实现几乎“最优”的遗憾。我们通过两种不同的方案实现这一点——一种依赖于洗牌对 DP 机制隐私放大的新结果，另一种利用将洗牌协议整合到基于树的机制中的向量求和，这两种方案都可能具有独立的兴趣。最后，我们通过在合成和现实生活数据生成的上下文老虎机实例上的数值评估来支持我们的理论结果。

文献[51]提出了一个有效且可适应的联邦框架 FedP3，代表联邦个性化和隐私友好型网络修剪，专为模型异质性场景量身定制。文献提出的方法可以将其特定实例很好地整合和适应到已建立的技术中。该算法提供了 FedP3 及其本地差分隐私变体 DP-FedP3 的理论解释，并从理论上验证了它们的效率。

文献[52]识别了这种配置中的一个重要安全漏洞，并设计了一种能够欺骗联邦学习最先进防御的攻击。提出的攻击包括两种操作模式，第一种专注于收敛抑制（对抗模式），第二种旨在构建对全局联邦模型的欺骗性评分注入（后门模式）。实验结果表明，文献的攻击在两种模式下都有效，平均在对抗模式的所有测试中返回了 60%的性能损失，并在后门模式的测试中 93%的情况下完全有效的后门。

文献[53] 揭示了尽管 DP 噪声扰动可以提高学习鲁棒性，但 DP-FL 框架本身并不鲁棒，容易受到精心设计的攻击方法的攻击。此外，文章发现现有的鲁棒 FL 方法很难防御对 DP-FL 的攻击。这可以归因于 DP-FL 的本地梯度被随机噪声扰动，所选的中心梯度不可避免地包含了比传统 FL 更高比例的中毒梯度。为了解决这个问题，文献进一步提出了一种新的 DP-FL 防御方法（命名为 Robust-DPFL），它可以有效区分 DP-FL 中的

中毒和清洁本地梯度，并鲁棒地更新全局模型。在三个基准数据集上的实验表明，基线方法不能同时确保任务准确性、数据隐私和鲁棒性，而 Robust-DPFL 可以有效增强联邦学习的隐私保护和鲁棒性，同时保持任务性能。

在文献[54]中，作者提出以下问题：差分隐私与 FL 对抗投毒攻击的认证鲁棒性之间存在哪些内在联系？能否利用 DPFL 的固有隐私属性为 FL 提供认证鲁棒性？能否进一步提高 FL 的隐私性以改善这种鲁棒性认证？作者首先研究了 FL 的用户级和实例级隐私，并提供了正式的隐私分析以实现改进的实例级隐私。然后，作者为用户提供级和实例级 DPFL 提供了两个鲁棒性认证标准：认证预测和认证攻击无效性。从理论上讲，作者根据给定的敌对用户或实例的有界数量，基于这两个标准提供了 DPFL 的认证鲁棒性。从实证上看，作者进行了广泛的实验，以验证在不同数据集上的一系列投毒攻击下文章的理论。作者发现，在 DPFL 中提高隐私保护水平可以导致更强的认证攻击无效性；然而，这并不一定导致更强的认证预测。因此，实现最优的认证预测需要在隐私和效用损失之间取得适当的平衡。

文献[55]引入了强大的自适应对手的概念，这些对手能够同时适应多个目标。广泛的实证测试揭示了现有防御在这种对手模型中的脆弱性。并且提出了 MESAS（多指标级联），这是一种针对更现实的场景和对手模型量身定制的新型防御方法。MESAS 同时采用多个检测指标来对抗中毒的模型更新，为自适应攻击者提出了一个复杂的多目标问题。在九个后门和三个数据集的全面评估中，MESAS 在区分客户端内外与数据分布相关的后门和扭曲方面，超越了现有防御。MESAS 在现实世界的数据设置中为强大的自适应对手提供了强大的防御，平均开销仅为 24.37 秒。

文献[56]提出了 FLPurifier，这是一种联邦学习中的新型后门防御方法，能够在联邦聚合之前有效净化可能的后门属性。具体来说，FLPurifier 将一个完整的模型分为一个特征提取器和分类器，其中提取器以解耦的对比方式进行训练，以打破触发特征和目标标签之间的强相关性。与现有的后门缓解方法相比，FLPurifier 不依赖于不切实际的假设，因为它可以在训练过程中有效净化后门效应，而不是已经训练好的模型。此外，为了减少被植入后门的分类器的负面影响并提高全局模型的准确性，我们进一步设计了一种自适应分类器聚合策略，以动态调整权重系数。在六个基准数据集上的广泛实验评估表明，FLPurifier 在联邦学习中对抗已知后门攻击是有效的，性能下降可以忽略不计，并

且优于最先进的防御方法

文献[57]假设这类攻击成功的关键因素之一是在随机优化过程中，批量内每个数据的梯度之间的低纠缠。这造成了敌手可以利用的漏洞来重建敏感数据。基于这一洞见，作者提出了一种简单但有效的防御策略，用隐藏的样本混淆敏感数据的梯度。为了实现这一点，文献提出合成隐藏样本以在梯度层面模仿敏感数据，同时确保它们与实际敏感数据在视觉上的差异。与先前的技术相比，其的实证评估表明，所提出的技术在同时保持 FL 性能的同时提供了最强的保护。

文献[58]提出了 Snowball，一个新颖的反后门 FL 框架，通过双向选举从个体视角出发，受到我们推导出的一个原则和 FL 及深度学习中的两个原则的启发。它的特点是 a) 自下而上的选举，每个候选模型更新对几个同行进行投票，以便一些模型更新被选为聚合的选择者；以及 b) 自上而下的选举，选择者通过从候选人中挑选来逐步扩大自己。文献将 Snowball 与 FL 中针对后门攻击的最先进防御措施在五个真实世界数据集上进行比较，展示了其对后门攻击的优越抵抗力和对全局模型准确性的轻微影响。

文献[59]提出了 FLTracer，这是一个首个 FL 攻击溯源框架，能够准确检测各种攻击并追踪攻击的时间、目标、类型以及更新中的中毒位置。与仅依赖于跨客户端异常检测的现有方法不同，作者提出了一种基于卡尔曼滤波器的跨轮次检测，通过寻找攻击前后的行为变化来识别对手。因此，它能够适应数据异质性，即使在 non-IID 设置中也有效。

文献[60]引入了一种基于 Huber 损失最小化的新型聚合器，并提供了全面的理论研究分析。在独立同分布（i.i.d）的假设下，我们的方法与现有方法相比具有几个优点。首先，它对代表受攻击客户端比例的 epsilon 有最优的依赖性。其次，作者的方法不需要精确知道 epsilon 的值。第三，它允许不同的客户端拥有不等的数据大小。然后，我们扩展了我们的分析，包括非 i.i.d 数据，使得客户端有略微不同的分布。

# 参考文献


[1] B. Hitaj, G. Ateniese, and F. P´erez-Cruz, “Deep models under the GAN: information leakage from collaborative deep learning,” in CSS, 2017, pp. 603–618.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in SP, 2017, pp. 3–18.

[3] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in S&P, 2019, pp. 691–706.

[4] M. Nasr, R. Shokri, and A. Houmansadr, “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning,” in S&P, 2019, pp. 739–753.

[5] L. Zhu, Z. Liu, and S. Han, “Deep leakage from gradients,” in NeurIPS, 2019, pp. 14 747–14 756

[6] B. Zhao, K. R. Mopuri, and H. Bilen, “idlg: Improved deep leakage from gradients,” CoRR, arXiv:2001.02610, 2020.

[7] P. Paillier et al., “Public-key cryptosystems based on composite degree residuosity classes,” in Eurocrypt, vol. 99, 1999, pp. 223–238.

[8] T. ElGamal, “A public key cryptosystem and a signature scheme based on discrete logarithms,” IEEE Transactions on Information Theory, vol. 31, no. 4, pp. 469–472, 1985.

[9] C. Gentry, “Fully homomorphic encryption using ideal lattices,” in STOC, 2009, pp. 169–178.

[10] A. C. Yao, “Protocols for secure computations,” in SFCS, 1982, pp. 160–164.

[11] D. Demmler, T. Schneider, and M. Zohner, “Aby-a framework for efficient mixed-protocol secure two-party computation.” in NDSS, 2015.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Theory of cryptography conference, 2006, pp. 265–284.

[13] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Foundations and Trends® in Theoretical Computer Science,vol. 9, no. 3–4, pp. 211–407, 2014.

[14] Y. Aono, T. Hayashi, L. Wang, S. Moriai et al., “Privacy-preserving deep learning via

additively homomorphic encryption," IEEE Transactions on Information Forensics and Security, vol. 13, no. 5, pp. 1333–1345, 2018.

[15] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," CoRR, arXiv:1711.10677, 2017.

[16] P. Mohassel and Y. Zhang, "Secure ml: A system for scalable privacy preserving machine learning," in S&P, 2017, pp. 19–38.

[17] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan,S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation or privacy-preserving machine learning," in CCS, 2017, pp.1175–1191.

[18] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends® in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

[19] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science, 2013, pp. 429–438.

[20] P. Blanchard, R. Guerraoui, J. Stainer et al., "Machine learning with adversaries: Byzantine tolerant gradient descent," in NeurIPS, 2017, pp. 119–129.

[21] C. Chen, J. Zhang, A. K. Tung, M. Kankanhalli, and G. Chen, "Robust federated recommendation system," arXiv preprint arXiv:2006.08259, 2020.

[22] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in Advances in Neural Information Processing Systems, 2019, pp. 8632–8642.

[23] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine tolerant sgd by inner product manipulation," in UAI. PMLR, 2020, pp. 261–270.

[24] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," CoRR, arXiv:1206.6389, 2012.

[25] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," CoRR, arXiv:1811.12470, 2018.

[26] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” CoRR, arXiv:1807.00459, 2018.

[27] C. Xie, K. Huang, P. Chen, and B. Li, “DBA: distributed backdoor attacks against federated learning,” in 8th International Conference on Learning Representations, 2020.

[28] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. y. Sohn, K. Lee, and D. Papailiopoulos, “Attack of the tails: Yes, you really can backdoor federated learning,” NeurIPS, 2020.

[29] J. R. Douceur, “The sybil attack,” in International Workshop on Peer-to-Peer Systems, 2002, pp. 251–260.

[30] C. Fung, C. J. Yoon, and I. Beschastnikh, “The limitations of federated learning in sybil settings,” in 23rd International Symposium on Research in Attacks, Intrusions and Defenses (fRAIDg 2020), 2020, pp. 301–316.

[31] S. Shen, S. Tople, and P. Saxena, “Auror: defending against poisoning attacks in collaborative deep learning systems,” in Proceedings of the 32nd Annual Conference on Computer Security Applications. ACM, 2016, pp. 508–519.

[32] P. Blanchard, R. Guerraoui, J. Stainer et al., “Machine learning with adversaries: Byzantine tolerant gradient descent,” in NeurIPS, 2017, pp. 119–129.

[33] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, “Byzantinerobust distributed learning: Towards optimal statistical rates,” CoRR, arXiv:1803.01498, 2018.

[34] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, “The hidden vulnerability of distributed learning in byzantium,” CoRR, arXiv:1802.07927, 2018.

[35] K. Pillutla, S. M. Kakade, and Z. Harchaoui, “Robust aggregation for federated learning,” arXiv preprint arXiv:1912.13445, 2019.

[36] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing federated learning through an adversarial lens,” CoRR, arXiv:1811.12470, 2018.

[37] X. Xu and L. Lyu, “Towards building a robust and fair federated learning system,” arXiv preprint arXiv:2011.10464, 2020.

[38] M. Fang, X. Cao, J. Jia, and N. Gong, “Local model poisoning attacks to byzantine-robust

federated learning," in 29th fUSENIXg Security Symposium (fUSENIXg Security 20), 2020, pp. 1605–1622.

[39] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "Draco: Byzantine-resilient distributed training via redundant gradients," CoRR,arXiv:1803.09877, 2018.

[40] L. Su and J. Xu, "Securing distributed machine learning in high dimensions," CoRR, arXiv:1804.10140, 2018.

[41] J. Steinhardt, M. Charikar, and G. Valiant, "Resilience: A criterion for learning in the presence of arbitrary outliers," arXiv preprint arXiv:1703.04940, 2017.

[42] J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar, "signSGD with majority vote is communication efficient and byzantine fault tolerant," in In Seventh International Conference on Learning Representations (ICLR), 2019.

[43] P. Zhao, P.-Y. Chen, P. Das, K. N. Ramamurthy, and X. Lin, "Bridging mode connectivity in loss landscapes and adversarial robustness," arXiv preprint arXiv:2005.00060, 2020.

[44] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," arXiv preprint arXiv:2101.05930, 2021.

[45] Li Y, Lyu X, Koren N, et al. Anti-backdoor learning: Training clean models on poisoned data[J]. Advances in Neural Information Processing Systems, 2021, 34: 14900-14912.

[46] Zhou X, Chowdhury S R. On Differentially Private Federated Linear Contextual Bandits[C]//The Twelfth International Conference on Learning Representations.

[47] Cao X, Gong N Z. Mpaf: Model poisoning attacks to federated learning based on fake clients[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3396-3404.

[48] Zhuang H, Yu M, Wang H, et al. Backdoor Federated Learning by Poisoning Backdoor-Critical Layers[C]//The Twelfth International Conference on Learning Representations.

[49] Nguyen T D, Nguyen T A, Tran A, et al. Iba: Towards irreversible backdoor attacks in federated learning[J]. Advances in Neural Information Processing Systems, 2024, 36.

[50] Zhang Z, Panda A, Song L, et al. Neurotoxin: Durable backdoors in federated learning[C]//International Conference on Machine Learning. PMLR, 2022: 26429-26446.

[51] Yi K, Gazagnadou N, Richtárik P, et al. FedP3: Federated Personalized and Privacy-friendly Network Pruning under Model Heterogeneity[C]//The Twelfth International Conference on Learning Representations.

[52] Arazzi M, Conti M, Nocera A, et al. Turning privacy-preserving mechanisms against federated learning[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023: 1482-1495.

[53] Qi T, Wang H, Huang Y. Towards the Robustness of Differentially Private Federated Learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(18): 19911-19919.

[54] Xie C, Long Y, Chen P Y, et al. Unraveling the connections between privacy and certified robustness in federated learning against poisoning attacks[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023: 1511-1525.

[55] Krauß T, Dmitrienko A. Mesas: Poisoning defense for federated learning resilient against adaptive attackers[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023: 1526-1540.

[56] Zhang J, Zhu C, Sun X, et al. FLPurifier: Backdoor Defense in Federated Learning via Decoupled Contrastive Training[J]. IEEE Transactions on Information Forensics and Security, 2024.

[57] Wu J, Hayat M, Zhou M, et al. Concealing Sensitive Samples against Gradient Leakage in Federated Learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(19): 21717-21725.

[58] Qin Z, Chen F, Zhi C, et al. Resisting Backdoor Attacks in Federated Learning via Bidirectional Elections and Individual Perspective[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(13): 14677-14685.

[59] Zhang X, Liu Q, Ba Z, et al. Fltracer: Accurate poisoning attack provenance in federated learning[J]. IEEE Transactions on Information Forensics and Security, 2024.

[60] Zhao P, Yu F, Wan Z. A huber loss minimization approach to byzantine robust federated learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence.