Privacy and Robustness in Federated Learning: Attacks and Defences

Abstract

In this paper, we conduct a comprehensive survey on privacy and robustness in federated learning over the past 5 years. Through a concise introduction to the concept of FL, and a unique taxonomy covering:

- 1. threat models;
- 2. privacy attacks and defenses;
- 3. poisoning attacks and defenses, we provide an accessible review of this important topic.

We highlight the intuitions, key techniques as well as fundamental assumptions adopted by various attacks and defenses. Finally, we discuss promising future research directions towards robust and privacypreserving FL, and their interplays with multidisciplinary goals of FL.

1.Introduction

A.Categorization of Federated Learning based on Distribution

Based on the distribution of data features and data samples among participants, federated learning can be generally classified as horizontally federated learning (HFL), vertically federated learning (VFL) and federated transfer learning (FTL).

B.Categorization of Federated Learning based on Architectures

FL with Homogeneous Architectures

Sharing gradients is typically limited only to homogeneous FL architectures,i.e., the same model is shared with all participants.

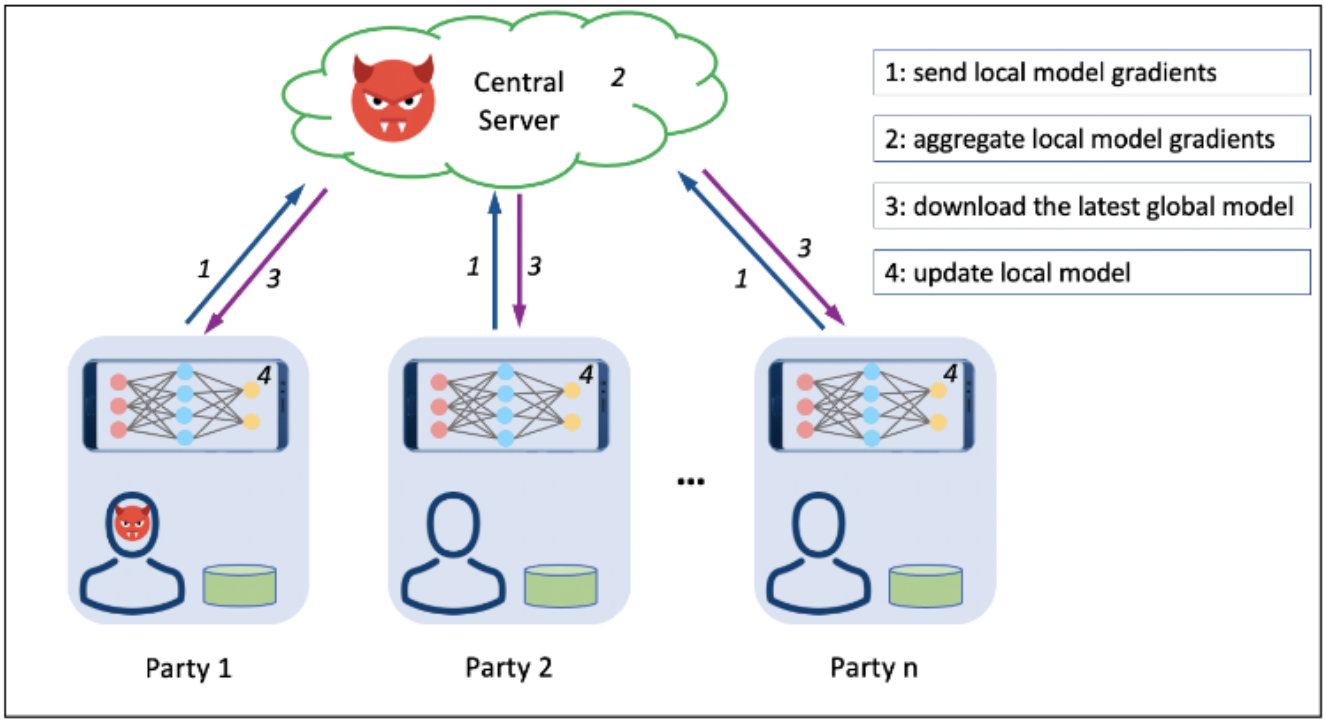


Figure 1 FL with Homogeneous Architectures

FL with Heterogeneous Architectures

Unlike the existing federated learning algorithms, Federated Model Distillation(FedMD) does not force a single global model onto local models.

C.Threats to FL

Existing FL protocol designs are vulnerable to:

- (1) a malicious server who aims to infer sensitive information from individual updates over time, tamper with the training process or control the view of the participants on the global parameters;
- (2) any adversarial participant who can infer other participants' sensitive information, tamper the global parameter aggregation or poison the global model.

In terms of robustness, FL systems are vulnerable to both data poisoning [33], [34] and model poisoning attacks [35], [36], [37], [38].Malicious participants can attack the convergence of the global model or implant backdoor triggers into the global model by deliberately altering their local data (data poisoning) or their gradients uploads (model poisoning). More broadly, poisoning attacks can be categorized into :

- (1) untargeted attack such as Byzantine attack where the adversary aims to destroy the convergence and performance of the global model [39], [22]; and
- (2) targeted attack such as backdoor attack where the adversary aims to implant a backdoor trigger into the global model so as to trick the model to constantly predict an adversarial class on a subtask while keeping good performance on the main task [35], [36], [34].

D.Secure FL

Attacks on FL come from either the privacy perspective when a malicious participant or the central server attempts to infer the private information of a victim participant, or the robustness perspective when a

malicious participant aims to compromise the global model.

E.Motivation of this Survey and Our Contribution

Contributions:

- 本次调查对联邦学习进行了全面的分类，系统地总结了联邦学习的威胁和相应的防护措施。
- 针对现存的隐私和鲁棒性攻击进行了深入探讨，以帮助读者更好地理解联邦学习在隐私和鲁棒性领域当前进展地假设、原则和差异。
- 调查了棒性和隐私性之间的冲突，以及多个设计目标之间的冲突；总结了当前研究成果和联邦学习实际应用场景的差距。
- 未来的研究方向将帮助社区重新思考和改进他们目前的设计，以实现真正实用和有影响力的健壮和保护隐私的联邦学习系统。同时，建议在联邦学习系统设计中融入多方面目标。

本文结构：
Section 2:总结联邦学习所面临的威胁。
Section 3:回顾联邦学习中的隐私泄露风险，尤其是试图窃取具有同质结构的水平联邦学习的敏感信息。
Section 4:解决相应的隐私攻击，本节列出了FL中应用这些技术最具有代表性的隐私保护技术和当前实践情况。
Section 5:总结破坏系统鲁棒性的投毒攻击。
Section 6:应对投毒攻击的方法。
Section 7:关于可信联邦学习的研究及其未来方向。
Section 8:结束语。

2.Threat Models

联邦学习系统中的威胁可以分为两类：

- 内部威胁和外部威胁。
- 训练阶段的威胁推理阶段的威胁。

A.内部威胁和外部威胁

内部威胁：来自联邦学习系统中的服务器和参与者。
外部威胁：对服务器和客户端的通信的窃听，最终FL作为服务部署时候的应用客户。

B.训练阶段和推理阶段

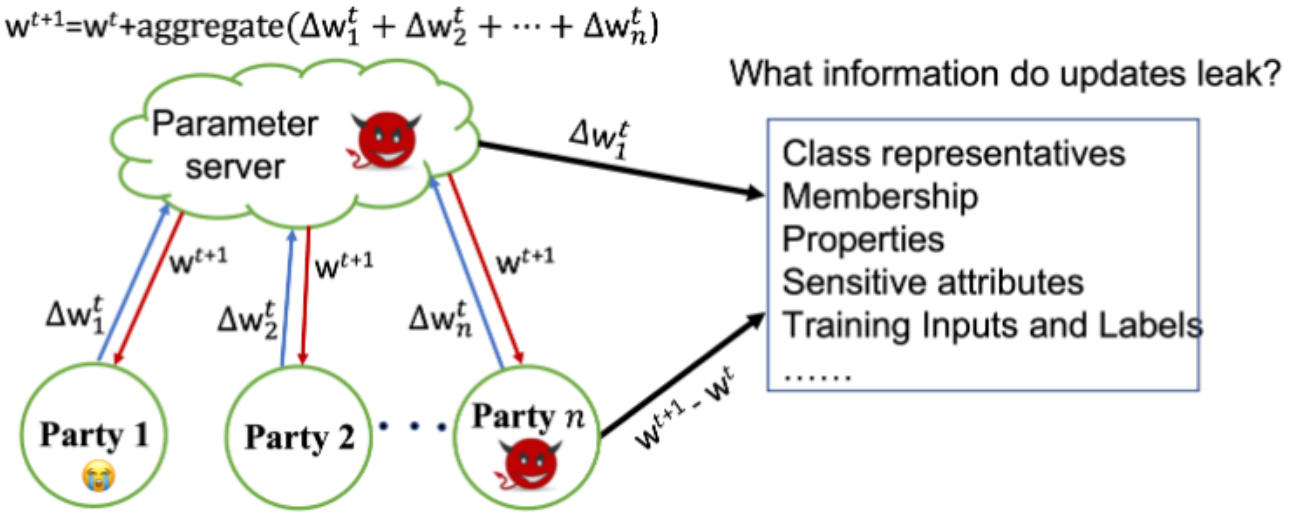
训练阶段：数据投毒，模型投毒，聚合更新攻击。
推理阶段：evasion and exploratory attack.

C.Privacy: Semi-honest vs Malicious

D.Robustness: Untargeted vs Targeted

3.隐私攻击

攻击者可以从接收到的梯度或者联邦学习模型参数的快照中推断出受害者参与者的各种私人信息。



A.Inferring Class Representatives

列举了FL中的GAN攻击，目标不是重建精确的训练输入，而只是重建类代表。
但是这种攻击在FL中不太实际，因为需要大量资源，不适用于H2C场景。

B.Inferring Membership

推定特定样本是否属于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）。
FL系统中的攻击者可以进行主动和被动的成员推理攻击[73]，[28]。在被动情况下，攻击者在不修改学习过程的情况下观察更新的模型参数并进行推理。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执行更强大的攻击。

C.Inferring Properties

推理目标属性。

D.Inferring Trainging Inputs and Labels

使用IDLG来提取标签和数据样本。

4.Defenses Against Privacy Attacks

A.通过同态加密(Homomorphic Encryption)保护隐私

同态加密技术允许直接对密文进行算术计算，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分为：

- 完全同态加密。
- 部分同态加密。
- somewhat同态加密。

总的来说，同态加密技术会产生额外的计算和通信开销，使得这一项技术不是很适用于H2C场景。

B.通过SMC来保护隐私

安全多方计算 (SMC) [93] 使具有不同具有私有输入的参与者能够在其输入上执行联合计算，而不会相互揭示。SecureML，它通过SMC进行隐私保护学习，数据所有者需要在初始设置阶段的两个非串通服务器之间处理、加密和/或秘密共享他们的数据。SecureML 允许数据所有者在他们的联合数据上训练各种模型，而不会揭示结果之外的任何信息。

一般来说，SMC 技术确保了高水平的隐私和准确性，但代价是计算和通信开销很高，从而使服务无法吸引参与。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同时协调所有参与者的要求。 同时，SMC不能完全保证信息泄露的保护，这需要在多方协议中加入额外的差分隐私技术来解决这些问题。

总之，同态加密或基于SMC的方法可能不适用于大规模FL场景，因为它们会产生大量的额外通信和计算成本。

C.通过差分隐私来保护隐私

与基于加密的方法相比，差分隐私通过以 (i) 计算效率的方式扰动数据来权衡隐私和准确性，(ii) 不允许攻击者恢复原始数据，(iii) 不会严重影响效用。

5.投毒攻击

投毒攻击的目的是破坏系统的鲁棒性，根据攻击者的目标不同，投毒攻击可以分为两类：

- 非目标投毒攻击。
- 有针对性的投毒攻击。

训练阶段的时候，目标和非目标投毒攻击可以部署在数据或者模型上。
中毒更新可以来自两种投毒攻击:(1)本地数据收集过程中的数据投毒攻击;(2)局部模型训练过程中的模型中毒攻击。在高层次上，两种中毒攻击都试图以某种不希望的方式修改目标模型的行为。然而，由于FL具有同构架构的模型共享特性，数据中毒攻击通常不如模型中毒攻击有效[35]，[36]，[37]，[38]。事实上，模型中毒在FL设置中包含了数据中毒，因为数据中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。这在功能上与集中中毒攻击相同。

A.Untarfeted Attacks

无目标中毒攻击的目的是任意破坏目标模型的完整性。拜占庭攻击是一种无目标的投毒攻击，它通过向服务器任意上传恶意梯度，从而导致全局模型失效。
Byzantine Attack
诚实的用户上传$\Delta w_i = \Delta F_i(w_i)$
恶意用户上传$\Delta w_i = *$
其中“*”表示任意值，Fi表示参与者i的局部模型目标函数。

Blanchard等人[22]表明，如果FL中没有防御，则FL的聚集可以完全由单个Byzantine参与者控制。 假设服务器通过$\Delta w' = \frac{1}{n} \sum_{i=1}^n \Delta w_i$来聚合梯度，假设第n个客户端是Byzantine，它可以通过上传以下梯度使聚合梯度变成任意向量：
$\Delta w_n = n \mu - \sum_{i=1} ^{n-1} \Delta w_i$

B.Targeted Attacks

在有针对性的中毒攻击中，学习到的模型为特定的测试示例输出攻击者指定的目标标签，例如，将垃圾邮件预测为非垃圾邮件，并为具有特定特洛伊木马触发器(后门/特洛伊木马攻击)的测试示例预测攻击者期望的标签。但是，其他测试示例的测试错误不受影响。一般来说，有针对性的攻击比无针对性的攻击更难以实施，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[124]，[37]。在保持数据特征不变的情况下，将一类诚实训练样例的标签翻转到另一类。

另一种现实的针对性投毒攻击是后门投毒攻击，攻击者可以修改原始训练数据集的单个特征或小区域，将后门触发器植入模型中。该模型在干净的数据上表现正常，但无论何时触发器(例如，图像上的戳)出现，它都会不断预测目标类。

后门攻击可以进一步分为两类:脏标签攻击[129]、[128]、[53]、[134]和干净标签攻击[135]、[136]、[130]、[137]、[53]、[66]。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，通过该过程，数据被证明属于正确的类别，并且数据样本的中毒必须是不可察觉的。在脏标签攻击导致中毒中，攻击者可以将许多数据样本引入训练数据中，这些样本预计会被具有期望目标标签的模型错误分类。清洁标签攻击可以说是隐形的，因为它们不改变标签。

6.针对投毒攻击的防御

对中毒攻击的鲁棒性是FL的一个理想特性。为了解决中毒攻击，文献中提出了许多鲁棒聚集方案。在集中式设置中，已知的对投毒攻击的防御，如鲁棒损失[144]和异常检测[123]，假设参与者的控制或对训练数据的显式观察。这些假设都不适用于FL，在FL中，服务器只观察作为迭代ML算法[37]的一部分发送的模型参数/更新。

A.防御非目标攻击

对于拜占庭弹性聚合，如果一个算法即使在大部分参与者是敌对的情况下也具有鲁棒性，则该算法是拜占庭容错：

- Shen等人[139]引入了一种称为AUROR的统计机制来检测恶意用户，同时生成准确的模型。AUROR基于这样的观察，即来自大多数诚实用户的指示性特征(最重要的模型特征)将呈现相似的分布，而来自恶意用户的指示性特征将呈现异常分布。然后，它使用k-means对参与者在训练回合中的更新进行聚类，并丢弃异常值，即，超过阈值距离的小聚类的贡献被删除。即使30%的用户是敌对的，使用AUROR训练的模型的准确性也只下降了3%。
- Blanchard等人提出了Krum，将离平均贡献最远的f个参与者，将它们从聚合中移除。使用欧几里得距离度量应该去除哪些参与者。本质上，Krum基于整个更新向量过滤异常值，但不过滤坐标异常值。

B.防御目标攻击

现有的针对针对性后门攻击的防御可以分为两类:检测方法和擦除方法[150]。检测方法利用激活统计或模型属性来确定模型是否为后门[151]，[152]，或者训练/测试样例是否为后门样例[47]。

有许多检测算法旨在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行为，以去除后门[48]，[128]，[153]，[42]，[47]。这些算法依赖于有毒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[154]。
虽然检测可以帮助识别潜在风险，但由于后门触发器的潜在影响在后门模型中仍然未被清除，因此后门模型仍然需要被纯化。擦除方法更进一步，旨在净化后门触发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修复(MCR)[155]和神经注意蒸馏(NAD)[54]。MCR通过在损失路径上选择一个鲁棒模型来减轻后门，而NAD利用知识蒸馏来消除触发器。

同时有人提出了ABL(Anti-Backdoor Learning)。

为了防御sybil克隆的针对性中毒攻击，Fung等人[37]利用sybil之间的相似性大于诚实客户之间的相似性这一特征行为，提出了FoolsGold:一种新的防御FL sybil攻击的方案，该方案基于贡献相似性调整参与者的学习率。

7.可信机器学习的研究方向

- 维度诅咒:具有高维参数向量的大型模型特别容易受到隐私和安全攻击[158]。大多数FL算法需要用全局模型覆盖局部模型参数。这使得它们容易受到中毒攻击，因为对手可以在不被发现的情况下对高维模型进行微小但具有破坏性的更改。几乎所有设计良好的拜占庭鲁棒聚合器[22]，[23]，[58]仍然受到维度诅咒的困扰。
- Rethinking Current Privacy Attacks.
- Rethinking Current Defenses: 出于隐私目的而使用安全聚合的FL更容易受到中毒攻击，因为无法检查单个更新。 Similarly, it is still unclear if adversarial training, one state-of-the-art defense approach against adversarial attacks in conventional ML [162], [163], [164], can be adapted to FL, as adversarial training was developed primarily for IID data and remains unclear for its performance in non-IID settings.
- Optimizing Defense Mechanism Deployment.
- Test-phase Privacy in FL.
- FL的测试阶段鲁棒性:就鲁棒性脆弱性而言，最近的研究[175]，[176]，[177]表明FL也容易受到精心制作的对抗性示例的影响。在推理期间，攻击者可以对测试数据进行非常小的扰动，使得测试数据几乎无法与自然数据区分，从而被全局模型错误地分类。