doi :10.3969/j.issn.1671-1122.2023.12.008

联邦学习与攻防对抗综述

杨丽，朱凌波，于越明，苗银宾

（西安电子科技大学网络与信息安全学院，西安 710126）

摘　要：随着机器学习技术的不断发展，个人隐私问题被广泛重视。由于用户数据被发送至中心节点导致集中学习受到相当程度的制约，所以联邦学习作为一个数据不出本地便可以完成模型训练的框架应运而生。但联邦学习机制依旧会受到各种攻击的影响而导致安全性和隐私性降低。文章先从联邦学习的基本定义入手，再对机密性和完整性两个方面进行重点分析、总结联邦学习中的威胁和防御手段，最后结合这些问题来讨论该领域在未来的发展方向。

关键词：联邦学习；机密性；完整性；防御手段

中图分类号：TP309　**文献标志码：**A　**文章编号：**1671-1122（2023）12-0069-22

中文引用格式：杨丽，朱凌波，于越明，等 . 联邦学习与攻防对抗综述 [J]. 信息网络安全，2023，23（12）：69-90.

英文引用格式：YANG Li, ZHU Lingbo, YU Yueming, et al. Review of Federal Learning and Offensive-Defensive Confrontation[J]. Netinfo Security, 2023, 23(12): 69-90.

Review of Federal Learning and Offensive-Defensive Confrontation

YANG Li, ZHU Lingbo, YU Yueming, MIAO Yinbin

(*School of Cyber Engineering*, *Xidian University*, *Xi'an* 710126, *China*)

Abstract: With the continuous development of machine learning technology, personal privacy issues have attracted widespread attention. Centralized learning is subject to a considerable degree of constraints due to the fact that user data is sent to the central node. Therefore, federal learning as a data can be completed locally. The framework of model training came into being. However, the federated learning mechanism will still be affected by various attacks and reduce the security and privacy. This paper started with the basic definition of federal learning, and then analyzed and summarized the threats and defense means in federal learning from two aspects of confidentiality and integrity. Finally, through these problems, the future development direction of this field was discussed.

Key words: federal learning; confidentiality; integrity; defensive means

收稿日期：2023-10-24

基金项目：国家自然科学基金 [62072361]；陕西省重点研发计划 [2022GY-019]；陕西省数理基础科学研究项目 [22JSY019]

作者简介：杨丽（1994—），女，安徽，博士研究生，主要研究方向为信息安全和隐私计算；朱凌波（1999—），男，安徽，硕士研究生，主要研究方向为网络安全和机器学习；于越明（1998—），女，河北，硕士研究生，主要研究方向为网络安全和机器学习；苗银宾（1989—），男，河南，教授，博士，主要研究方向为云计算、数据安全和隐私保护。

通信作者：朱凌波 zlb326511@163.com

0 引言

近些年，机器学习（Machine Learning，ML）在人工智能（Artificial Intelligence，AI）应用领域中的迅猛发展，例如计算机视觉、自动语音设别、自然语言处理以及推荐系统等[1-3]。这些机器学习技术的成功，尤其是深度学习，建立在大量数据（亦称大数据）基础之上[1,4]。通过使用这些数据，深度学习系统能够在许多领域执行人类难以完成的任务。例如，由数百万张图像训练得到的深度学习人脸识别系统，能够达到应用领域所需级别的人脸识别准确度。这些系统的训练都需要很大的数据量才能达到一个令人满意的性能水平，例如，Facebook 公司的目标检测系统是由来自 Instagram 的 3.5 亿张图片训练得到[5]。

随着人工智能在各行各业的应用落地，人们对于用户隐私和数据安全的关注度也在不断提高。用户开始更加关注他们的隐私信息是否未经自己许可，便被他人出于商业或者政治目的而利用或滥用。在这样的法律环境下，随着时间的推移，我们在不同组织间收集和分享数据将会变得越来越困难。更加重要的是，某些金融交易数据和医疗健康数据等高度敏感数据的拥有者也会极力反对无限制地计算和使用这些数据。在这种情况下，数据拥有者只允许这些数据保存在自己手中，进而会形成各自孤立的数据。由于来自行业竞争、用户隐私、数据安全和复杂的管理规程等的约束，在同一家公司的不同部门之间，数据整合也会遇到很大的阻力。与此同时，高昂的成本也导致在不同机构之间聚合分散的数据显得十分困难[6]。

因为各方面原因造成的数据孤岛阻碍着训练人工智能模型所需的大量数据，所以一种“所有参与方的隐私数据足不出户地参与模型训练”的新思路应运而生。该方法是由每一个拥有数据源的组织训练一个模型，之后让各组织在各自的模型上彼此交流数据，最终通过模型聚合得到一个全局模型。为了确保用户隐私和数据安全，各个组织间交换的模型数据无法被其他任何组织猜测到。同时全局模型整合了所有组织的数据，这便是联邦学习的核心思想。

然而联邦学习依旧面临很多安全问题，主要包括隐私泄露和模型窃取。攻击者可以通过梯度或参数信息获取本地数据的隐私信息和模型窃取，极大威胁数据隐私安全。在攻击方面，研究人员提出了许多攻击方法，例如梯度泄露攻击、模型反转攻击和成员推理攻击等。为了防御这些攻击，研究人员提出了一些防御方法，例如差分隐私、同态加密和安全多方计算等。现有技术虽可以在保护数据隐私的同时确保模型的准确性和效率，但如何权衡效率和安全性之间的关系，依旧是中重要的研究方向。

1 基本概念

1.1 概述

由于各方面原因造成的数据孤岛正阻碍着训练人工智能模型所必需的大数据的使用，所以人们开始寻求一种不必将所有数据集中到一个中心存储点就能训练机器学习模型的方法。其中一种可行的方法是由每一个拥有数据源的组织训练一个模型，之后让各个组织在各自的模型上彼此交流沟通，最终通过模型聚合得到一个全局模型。为确保用户隐私和数据安全，各组织间交换模型信息的过程会被精心设计，使得没有组织能够猜测到其他组织的隐私数据内容。同时，当构建全局模型时，各数据源仿佛已被整合在一起，这便是联邦机器学习（Federated Machine Learning）或简称联邦学习（Federated Learning）的核心思想。

MCMAHAN[7] 等人通过使用边缘服务器架构，将联邦学习用于智能手机上的语言预测模型更新。大多智能手机都存有私人数据，为了更新谷歌 Gboard 系统的输入预测模型，即谷歌的自输入补全键盘系统，谷歌的研究人员开发了一个联邦学习系统，以便定期更新智能手机上的语言模型。谷歌 Gboard 系统能查询到用户的建议输入，以及用户是否点击了建议输入的词。谷歌 Gboard 系统的单词预测模型可以不断改进和优化，该模型不仅基于单部智能手机存储的数据，而且可通过一种联邦平均（Federated Averaging）的技术，让所

有智能手机的数据都能被利用，使该模型得以不断优化。而这一过程并不需要将智能手机上的数据传输到某个数据中心位置，也就是说，联邦平均并不需要将数据从任何边缘终端设备传输到一个中心位置。通过联邦学习，每台移动设备（智能手机或平板电脑）上的模型将会被加密并上传到云端。最终，所有加密的模型都会被聚合到一个加密的全局模型中，因此云端的服务器也不能获知每台设备的数据或者模型[8-10]。在云端聚合后的模型仍然是加密的(如使用同态加密)，之后将会被下载到所有的移动终端设备上[11,12]。在上述过程中，用户在每台设备上的个人数据并不会传给其他用户，也不会上传至云端。

1.2 联邦学习的定义

联邦学习旨在建立一个基于分布数据集的联邦学习模型。联邦学习包括两个过程，分别是模型训练和模型推理。在模型训练的过程中，模型相关的信息能够在各方之间交换或以加密形式进行交换，但数据不能。这一交换不会暴露每个站点上数据的任何受保护的隐私部分。已训练好的联邦学习模型可置于联邦学习系统的各参与方，也可以在多方之间共享。

更一般地，设有 N 位参与方 $\{F_i\}_{i=1}^{N}$ 协作通过使用各自的训练数据集 $\{D_i\}_{i=1}^{N}$ 训练机器学习模型。传统方法是将所有的数据 $\{D_i\}_{i=1}^{N}$ 收集起来并存储在地方，如存储在某台云端数据服务器上，从而在该服务器上使用集中后的数集训练得到一个机器学习模型 M_{SUM}。在传统方法的训练过程中，任何一位参与者 F_i 会将自己的数据 D_i 暴露给服务器甚至其他参与方。联邦学习是一种不需要收集各参与方所有的数据 $\{D_i\}_{i=1}^{N}$，便能协作训练一个模型 M_{FED} 的机器学习过程。V_{SUM} 和 V_{FED} 分别为集中型模型 M_{SUM} 和联邦型模型 M_{FED} 的性能量度（如准确度、召回度和$F1$分数等）。接下来，对性能保证的含义进行更准确的解释。设 $\delta > 0$，当且仅当满足公式（1）时，联邦学习模型 M_{FED} 具有 δ 的性能损失。

$$V_{SUM} - V_{FED} < \delta \tag{1}$$

上式表述了以下客观事实：如果使用安全的联邦学习在分布式数据源上构建机器学习模型，这个模型在未来数据上的性能近似于把所有数据集中到一个地方训练所得到的模型的性能。

允许联邦学习模型在性能上比集中训练的模型稍差，因为在联邦学习中，参与方 F_i 并不会将他们的数据 D_i 暴露给服务器或者任何其他的参与方，所以相比准确度的 δ 的损失，额外的安全性和隐私保护更有价值。

1.3 系统架构

具体来讲，联邦学习是一种具有以下特征的用来建立机器学习模型的算法框架，其中，机器学习模型是指将某一方的数据实例映射到预测结果输出的函数。

1）有两个或以上的联邦学习参与方协作构建一个共享的机器学习模型。各参与方都拥有若干能够用来训练模型的训练数据。

2）在联邦学习模型的训练过程中，各参与方拥有的数据都不会离开该参与方，即数据不离开数据拥有者。

3）联邦学习模型相关的信息能够以加密方式在各方之间进行传输和交换，并且需要保证任何一个参与方都不能推测出其他参与方的原始数据。

4）联邦学习模型的性能要能够充分逼近理想模型（指通过将所有训练数据集中在一起并训练获得的机器学习模型）的性能。

根据应用场景的不同，联邦学习系统可能涉及也可能不涉及中央协调方。图1中展示了一种包括协调方的联邦学习架构示例。在此场景中，协调方是一台聚合服务器，可以将初始模型发送给各参与方A～C。参与方A～C分别使用各自的数据集训练该模型，并将模型权重更新发送到聚合服务器。之后，聚合服务器将从参与方处接收到的模型更新聚合起来，并将聚合后的模型更新发回参与方。这一过程将会不断迭代，直到模型收敛或达到时间、训练轮数阈值。在这种体系结构下，参与方的原始数据永远不会离开自己。这种方法不仅保护了用户的隐私和数据安全，还减少了

发送原始数据所带来的通信开销。此外，聚合服务器和参与方还能使用加密方法来防止模型信息泄露。

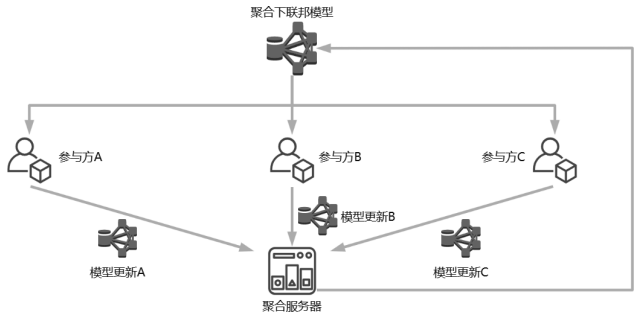

图 1 联邦学习系统示例

1.4 联邦学习的分类

1）横向联邦学习

横向联邦学习是按样本划分的联邦学习[13]，主要应用于各个参与方的数据集有相同的特征空间和不同的样本空间的场景。例如，两个地区的城市商业银行可能在各自地区拥有区别较大的客户群体，所以他们的客户交集非常小，并且数据集有不同的样本ID，然而他们的业务模型却非常相似，因此他们的数据集的特征空间相同。由此，两家银行就可联合起来进行横向联邦学习以构建更好的风控模型。

2）纵向联邦学习

纵向联邦学习适用于样本总体具有相同的样本空间、不同的特征空间的参与方所组成的联邦学习场景，纵向联邦学习是按照特征划分的[13]。例如，假设有两家公司A和B想要协同训练一个机器学习模型，每家公司都拥有各自的数据，例如，保险公司与银行合作，根据同一用户的购买历史与消费习惯，为该用户提供定制化的服务；医院与制药公司合作，通过利用同类患者的医疗记录，从而治疗患者的慢性疾病，降低患者未来住院治疗的风险。由于用户隐私和数据安全的原因，A方和B方不能直接交换数据，为保证训练过程中数据的保密性，此时需要加入了一个第三方的协调者C。C是一个半诚实的第三方，它主要用来帮助参与方进行安全的联邦学习。C独立于各参与方，它将会收集中间结果用来计算梯度和损失值，并将结果转发给各参与方。C收到的来自参与方的信息是被加密过或被混淆处理过的，因此各方的原始数据并不会暴露给彼此，且各参与方只会收到与其拥有的特征相关的模型参数。

3）联邦迁移学习

横向联邦学习和纵向联邦学习要求所有的参与方具有相同的特征空间或样本空间，从而建立起一个有效的共享机器学习模型。然而，在更多的实际情况下，各参与方所拥有的数据集在用户和数据特征上的重叠部分都比较小。该情况下，通过迁移学习技术[13]，使其可应用于更广泛的业务范围，同时可以帮助只有少量数据（较少重叠的样本和特征）和弱监督（较少标记）应用建立有效且精确的机器学习模型，且遵守数据隐私和安全条例的规定。这种组合即称为联邦迁移学习（Federated Transfer Learning，FTL），它可以处理超出现有横向联邦学习和纵向联邦学习能力范围的问题。

一个联邦迁移学习系统一般包括两方，即源域和目标域。一个多方的联邦迁移学习系统可以被认为是多个两方联邦迁移学习系统的结合。

1.5 隐私保护联邦学习

1）降噪隐私保护：主要通过差分隐私等方法实现。

差分隐私（Difference Privacy，DP）通过给数据添加噪声，或使用归纳方法隐藏参与方的某些敏感属性，直到第三方无法通过差分攻击来区分个人为止，使数据无法还原，从而达到保护用户隐私的目的。但是DP会带来模型准确性上的损失，因此通常需要在参与方隐私与模型准确性之间进行权衡。

2）加密隐私保护：主要通过安全多方计算、同态加密等方法来实现,通常需要设计复杂的加密计算协议来隐藏真实的输入和输出。

安全多方计算（Secure Multi-Party Computation，MPC）是一种允许两个或多个参与者协同地从各方的隐私输入中计算函数的结果的方式，而不用将这些输入展示给其他方。在联邦学习中，各参与者的模型参数作为隐私输入，通过MPC，各参与者可以在不暴露自己模型实际

数据的情况下，共同完成模型的迭代和优化，可有效提高联邦学习的安全性和隐私性。但是MPC成本较高,为降低数据传输成本,参与方可能需要降低对数据安全的要求来提高训练的效率。

同态加密（Homomorphic Encryption，HE）逐渐被认是实现安全多方计算的一种可行方法。该方法允许在密文状态下进行计算，得到的结果在解密后与明文状态下的计算结果相同。在联邦学习中，使用HE可以使得参与者能够在保持数据加密的状态下进行建模和训练，这样即使攻击者获得了加密后的模型参数，也无法推理出原始的数据信息，有效保障数据层面的安全。

1.6 联邦学习安全问题

在联邦学习模型的训练过程中，可能存在恶意的参与方或恶意的中心聚合服务器，他们会利用联邦学习协议的漏洞去控制上传的参数，并恢复部分参与者的部分敏感数据，以达到毒害目标模型或窃取目标信息等效果。

1）联邦学习的安全要求

机密性是指联邦学习系统保证模型不会泄露相关敏感信息，要求系统必须保证未得到授权的用户无法接触到系统中的私密信息，包括模型的训练数据、架构和参数等信息。

完整性是指联邦学习系统在模型训练和特征推理的过程中不受到恶意的侵害，要求模型的推理结果不能过分偏离预期情况。

2）安全威胁

机密性攻击：机密性攻击的目的是窃取联邦学习系统中的敏感信息，如模型参数、模型结构、训练方式和训练数据等。攻击者通过预先设定的侵害方式窃取模型的信息或恢复与模型特征相关的部分敏感信息。

完整性攻击：完整性攻击的目的是破坏联邦学习系统的完整性，使得联邦学习系统中的数据、程序或设备受到损害。攻击者通过利用漏洞或未经授权的访问，篡改联邦学习系统中的数据、程序或设备，从而影响联邦学习系统的正常运行，或对模型训练过程和推理过程进行干扰，使得模型的推理结果不符合预期。

本文对目前联邦学习所面临的主要攻击手段进行了分类，如图2所示，按照攻击方式可分为成员推理攻击、属性推理攻击、对抗攻击、数据投毒攻击、模型投毒攻击和后门攻击6种方式。

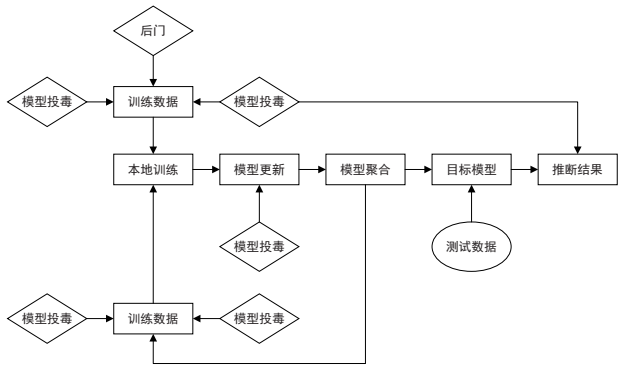

图 2 联邦学习面临的主要攻击手段

2 机密性攻击与防护

2.1 成员推理攻击

攻击者通过访问目标模型以及对应目标数据，试图推断目标数据是私有数据的成员（如推断某位患者是否在某个科室有记录）。攻击者主要利用已有信息去训练得到一个攻击模型，然后利用它分辨出目标模型对于数据点的梯度反应以推断是否是目标模型的成员。

在联邦学习中，恶意攻击者可以是参与者或中心聚合服务器。作为参与者，攻击可以通过观察模型更新以观察全局参数的更新，并可以改变其参数的上传。全局攻击者（中央聚合器）可通过收集更新的参数来被动或主动地攻击所有参与者。

常见的成员推理攻击方法有4种，分别是利用影子模型的成员推理攻击[14]、利用梯度更新特点的成员推理攻击[15]、基于模型参数记忆性的成员推理攻击[16]和针对生成网络的成员推理攻击[17,18]。

SHOKRI[14]等人为了判断某样本是否被用作训练集中的一部分，提出了影子模型。通过影子模型对攻击模型进行训练，他们实际上是将识别训练数据集的成员与模型输出之间复杂关系的问题转化为二元分类

问题。他们通过构造的训练数据训练影子模型（与目标模型类似的功能），再利用影子模型的推断对攻击模型进行训练，最后再用攻击模型来区分目标模型的输出是否在目标模型的训练集之中。

MELIS[15]等人通过批数据嵌入层的向量更新特点来进行成员推理，恶意方通过检测非零更新序列来判断特定序列是否包含在模型训练中收集到的训练序列中，以此判断目标是否是参与训练的成员。但他们的攻击方法有较大的局限性，例如实验中的参与方数量较少并且需要辅助数据进行训练，且在收集到的序列中检索目标序列的过程中未考虑一些附属因素（如序列的排序），可能会影响推理成功率。

NASR[16]等人发现参与模型训练的样本都会对应损失函数不同特点的梯度更新，因此设计了一个由卷积神经网络和全连接网络组成攻击模型，输入目标样本的特征标签、损失函数以及样本的隐藏层梯度更新和输出，并向全连接编码器发送隐藏层梯度更新和输出中对应目标样本的特征，最后得到全连接层的推理概率输出。NASR[16]等人还提出了一个主动攻击的方式：如果X是另外某一节点的训练数据，那么该节点在收到被恶意篡改的数据之后，会主动尝试降低模型损失函数在目标数据的梯度，而这个操作会被推理攻击算法探测到（因为接受了这个节点上传的数据），从而达到成员推理攻击的效果。在联邦学习的环境下重复这种主动攻击，就能高置信地完成推理攻击任务。

CHEN[17]等人利用生成对抗网络（Generative Adversarial Network，GAN）和分类模型实现成员推理攻击，但只适用于各训练模型成员具有不同的样本标签，且所有标签需被声明出来的情况。他们首先利用GAN生成其他成员带有不同标签的样本，利用这些样本训练得到一个二分类模型。然后对比二分类模型的输出结果与目标成员的标签，以推断目标数据为训练数据集的成员，但他们的攻击方式会误判参与训练但与全局模型有一定偏差的样本。ZHANG[18]等人同样利用GAN去增加训练数据的多样性，不同的是他们用GAN得到的样本及其真实标签y附近的推理值分布作为二分类模型的训练数据，以通过目标数据的标签推理值判断它是否是成员数据。最后通过实验说明，过拟合是导致模型易受到成员推理攻击的一个重要原因，过拟合程度越高的联邦学习模型往往会泄漏更多的成员信息。

2.2 属性推理攻击

联邦学习中的属性推理攻击是一种恶意攻击者试图窃取未知明确特征或与学习任务无关的样本特征方法，这些属性独立于联邦学习模型特征。例如，攻击者可以利用多任务学习来欺骗联邦学习模型，使其学习更好地分离具有和不具有目标属性的数据，从而提取更多信息。

1）相关属性推理

相关属性推理是指攻击者利用模型的输出或梯度来推断训练数据中与学习任务相关的敏感属性，如年龄、性别和种族等。这种攻击一般会利用GAN和求解最优化问题重构与其他参与方样本特征高度相似的“伪样本”，以“重现”其特征，所以这种攻击也被称为数据重构攻击。

HITAJ[19]等人提出恶意参与方利用GAN生成具有其他参与方的数据后修改成自己的标签，然后作为训练样本毒害全局模型，以至于在不断的迭代中学习到更多其他参与方的样本特征的攻击。但这种攻击不适合在多方参与和多样本标签的情况下进行，会有较大的误报率。

WANG[20]等人则是利用恶意的服务器进行推理攻击，服务器可以利用GAN生成指定受害者的训练样本。恶意服务器利用GAN学习目标参与方的样本分布，判别网络D在判别输入样本的真实性的同时判别是否是受害者，其中输入样本由生成网络G在不断学习后生成。但是上述方法仅限于联邦学习框架中每个参与者仅上传单批样本数据的更新，并不适用于多批次数据的场景。而SONG[21]等人在WANG[20]等人工作的基础上进一步深入研究，提出了在匿名条件下利用数据代表近似出参与方的样本特征的方法。他们主要是从

每个用户的更新中计算出相应的数据代表，然后用这些近似真实的数据训练GAN的鉴别器和生成器，最终GAN的生成器便可以生成某个client的私有数据。由于模型更新可以反映出参与方一定的数据分布规律且一直会保持不大的差异，因此通过用卷积孪生网络找出前后迭代中相似度之差最小的代表数据去关联它们的归属用户，在得到每个用户的多次迭代数据后便可以实现数据重构。上述攻击方式在MNIST[22]数据集（Mixed National Institute of Standards and Technology Database）上效果显著，但一旦每个用户的部分样本的特征差异较大，便会错误关联前后迭代的数据导致数据重构的真实效果大打折扣。

ZHU[23]等人则是利用求解最优化问题来实现数据重构。因为局部梯度更新会暗含样本训练信息，所以通过让全局模型产生与受害者相似梯度以重构与受害者样本相似的重构样本。随机生成一份和真实数据同样大小的假输入样本和假的标签，然后把这些假样本和假标签输入到现有的模型当中，得到假的模型梯度。该方法的目标是生成与原模型相同梯度的假梯度，这样的假样本和假标签就和真实的样本标签一致。但上述方法不适合重构较为复杂的样本，并且重构质量不足，因此有后续研究在此方面进行了精进。其中GEIPING[24]等人介绍了现有的攻击方法是基于欧几里得损失函数和通过L-BFGS优化的方法，这种方法是寻找梯度与观测梯度最相近的图像，如果把向量分解为范数和方向的话，这就是寻找范数相近的图像，但范数仅仅包含测量数据点相对于当前模型的局部最优性，相比之下，高维方向可以携带重要信息。他们的设定是恶意方是诚实且好奇的服务器，设计目标是利用Adam算法[25]求解全局梯度和局部梯度的余弦相似度的最小值。而ZHAO[26]等人则是利用从局部梯度中获取的样本标签信息，求解样本x这一最优化问题，有效地提高了攻击的效率。WEI[27]等人则在考虑目标函数的同时优化基于标签的正则化项，提高了攻击的容错率和重构数据的真实性。

DARIO[28]等人提出了两种针对联邦学习中安全聚合协议的攻击：梯度退化攻击和金丝雀梯度攻击。通过向非目标用户发送恶意全局模型，间接控制安全聚合的输入以实现梯度退化，得到目标用户的本地模型更新。然后在对非目标用户进行梯度退化攻击的同时，将“金丝雀梯度的注入问题”转化成“对网络末端的残差块的分类训练”，以通过损失函数对特定参数的偏导数结果去判断当前训练样本中是否存在目标属性。

2）无关属性推理

无关属性推理攻击针对的是与目标模型输出无关的属性，故又被称为无意识的特征泄露，如图3所示。例如，如果目标模型是一个人脸识别模型，那么无关属性可能是人脸图片中的背景、服装和发型等特征。实现的主要方法是利用目标模型的预测结果和部分模型参数来构建一个辅助分类器，该分类器可以根据输入数据判断是否具有某个无关属性。

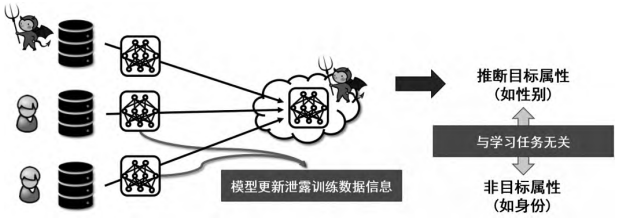

图 3 联邦学习中的无关属性推理攻击

MELIS[15]等人利用全局模型在辅助数据的推理结果去确定受害者是否有自己的若干目标属性信息，然后以“目标属性的标签”和“辅助数据上的局部更新”作为输入数据训练一个二分类模型，可以利用它去判断受害者的训练集中是否有自己的目标属性。为了提高攻击的效率，在此基础上开发了多任务学习的机制，使得目标模型学习更多目标属性。SHEN[29]等人是在恶意的中心聚合服务器一端利用辅助数据训练分类模型，该模型输入一批参与方聚合后的梯度更新，输出目标属性出现的概率，在遍历所有批后以出现最高推理概率的该批作为推理结果。

ZHOU[30]等人提出了一种新型的隐私推理攻击——偏好分析攻击（Preference Profiling Attack，PPA），可以准

确地分析本地用户的隐私偏好。通过构造模型灵敏度提取算法来确定每个类的梯度变化，然后设计了元分类器，PPA可以推断出FL中用户本地数据的偏好类别。然后验证PPA可在4个数据集（MNIST、CIFAR10、RAF-DB和Products-10K）上展现出较高的攻击成功率。实验表明，PPA对FL的本地用户数和一个用户使用的本地训练时间不敏感。对于现有的安全防护策略，如丢包、差分隐私保护等，虽在一定程度上降低了PPA的精度，但不可避免地会对全局模型产生显著的影响。

2.3 隐私保护方法

面对联邦学习存在的各种机密性攻击的威胁，诸多学者不断地研究出众多隐私保护方案，以保护模型训练参与方的敏感信息不被泄露。根据技术方案的分类，主要有利用安全多方计算技术、差分隐私技术、加密技术和安全聚合和部分共享等4种方案。

1）安全多方计算

安全多方计算（Secure Multi-Party Computation，SMC）是一种密码学技术，1982年由YAO[31]提出和推广。它可以让多个参与方在不泄露各自的私密输入的情况下，协同计算一个公共函数。安全多方计算涉及多个研究领域和技术，如秘密共享、不经意传输、混淆电路和同态加密等。这些技术可以构建通用或专用的安全多方计算协议，以适应不同的计算需求和效率要求。安全多方计算也可以利用云计算等新兴技术来提高性能和可扩展性。

安全多方计算是一种密码学原语，可以让多个参与方在不泄露各自的隐私输入的情况下，合作计算一个公共函数。用数学语言来描述，安全多方计算可定义如下。

定义1 安全多方计算 给定一个函数$f(x_1,x_2,\cdots,x_n)=(y_1,y_2,\cdots,y_n)$，其中$x_i$是参与方$P_i$的私密输入，$y_i$是参与方$P_i$的期望输出。一个安全多方计算协议可以让每个参与方$P_i$仅通过与其他参与方交换信息，就能得到正确的输出$y_i$。

XU[32]等人利用函数加密系统[33]来实现安全多方计算，该系统与普通公钥加密系统的区别是私钥可以为函数f确定一个密钥，而其他参与方可以利用该密钥和密文做运算，但不会泄露明文。因此XU[32]等人提出的隐私保护框架引入一个可信的第三方（Third Party Administrator，TPA），参与方在上传加密后的局部更新后，TPA生成各个参与方对应的密钥发送给服务器，服务器利用该密钥和各个参与方加密的局部更新计算出全局模型，从而保证在聚合过程中所有参与方的局部更新信息不会被任何一方所窃取。

KHAZBAK[34]等人提出的ML Guard是一个轻量级的分布式协作学习系统，利用秘密共享技术实现安全多方计算。该方案中参与方在每一轮都会拆分局部模型更新为两个部分，分别发给两个服务器，每个服务器在收到所有共享后再聚合，最后两个服务器一起计算最终的全局模型，所以在此过程中任何参与方或者聚合服务器都无法获得别的参与方的本地更新信息。

LI[35]等人的方案没有拆分模型更新，而是将所有的参与方进行拆分分组，然后不同的组安排到不同的链上，参与方的每轮更新都与所在链上的上一个节点的更新进行聚合，一直迭代下去最后将每条链的最终聚合结果发送给聚合服务器，再全局更新。这样有力保证了聚合服务器无法知道每一个参与方的梯度更新信息。

2）差分隐私

差分隐私（Differential Privacy）最早于2008年由DWORK[36]提出，使用随机应答（Randomized Response）方法确保数据集在输出信息时受单条记录的影响始终低于某个阈值，从而使第三方无法根据输出的变化判断单条记录的更改或增删，该方法被认为是目前基于扰动的隐私保护方法中安全级别最高的方法。

基于差分隐私的特性，可以将聚合算法作为M，通过在参与方的模型更新D上添加噪声成为D'，使聚合的全局模型与真实的全局模型尽可能接近，同时也可以防止攻击者从D'中推断出参与方的隐私信息。差分隐私也可以在全局模型上应用，以保护模型隐私。

GEYER[37] 等人为了保护全局模型，只在服务器聚合所有模型更新，确保一个学习模型不会显示客户是否参与了训练，这意味着客户机的整个数据集受到了保护，不受来自其他客户机的差分攻击。而 JAYARAMAN[38] 等人是通过添加拉普拉斯噪声来实现差分隐私的保护。

文献 [39-44] 的方案是在上传本地更新时就在更新中加入噪声，以保护梯度信息不被聚合服务器窃取。其中 BHOWMICK[39] 等人提出的差分隐私保护方法是一种新的隐私保护机制，可以在保护个人隐私的同时提高模型的准确性，这种方法通过对数据集进行加密和混淆来保护个人信息，并使用差分隐私技术来保证个人信息的隐私性；此外，这种方法还可通过使用正则化技术来提高模型的准确性。TRIASTCYN[40] 等人采用贝叶斯差分隐私（Bayesian Differential Privacy，BDP）将噪声校准到数据分布，即根据数据分布的实际情况来添加噪声。由于同一个区域的用户的样本可能拥有相似的特征，所以他们对不同区域的用户进行了个性化处理，采用贝叶斯差分隐私可添加满足阈值条件的较小噪声。因此，对于从相同或任意分布中抽取的任何两个数据集，给定具有相同噪声量的相同隐私机制，BDP 提供比 DP 更严格的保证。HUANG[41] 等人提出了一种新的差分隐私框架，首先判断梯度下降的方向，然后据此不断动态设置需要添加的噪声，使得添加的差分隐私更加有效地减少了敏感信息的泄露。而 WU[42] 等人提供了一种新思路，在训练轮次中多次加入高斯噪声，有效降低了差分隐私损失模型性能的代价，提高了模型的健壮性。

在梯度中添加差分噪声并不会有过多的额外开销产生，但是会降低模型的准确性。因此，在考虑隐私保护和模型性能之间的平衡时，需要在隐私预算和模型性能之间进行权衡。

3）同态加密

加密可以用来保护模型参数和训练数据的安全性，并防止未经授权的计算机访问联邦学习系统。其主要是利用同态加密（Homomorphic Encryption，HE）算法实现。作为一种不需要对密文进行解密的密文计算解决方案，同态加密的概念首先由 RIVEST[43] 等人在 1978 年提出。

设 $Enc_{pk}(\cdot)$ 表示使用 pk 作为加密密钥的加密函数。设 M 表示明文空间，且 C 表示密文空间。一个安全密码系统若满足条件 $\forall m_1, m_2 \in M, Enc_{pk}(m_1 \odot_M m_2) \leftarrow Enc_{pk}(m_1) \odot_C Enc_{pk}(m_2)$，则可称为同态的。

其中，$\odot_M$ 为 M 中的运算符，$\odot_C$ 为 C 中的运算符，$\leftarrow$ 符号表示左边项等于或可直接由右边项计算出来，而不需要任何中间解密。在此，设 $Dec_{sk}(\cdot)$ 表示使用 sk 作为解密密钥的解密函数，将同态加密运算符设为 $[[\cdot]]$，如 $[[u]]$、$[[v]]$ 为两个同态加密运算符，并且对密文的加法操作和乘法操作按如下方式重载。

（1）加法如公式（2）所示。在 PAILLIER[44] 的方案中，$\odot_C$ 表示密文的乘法。

$$Dec_{sk}([[u]] \odot_C [[v]]) = Dec_{sk}([[u+v]]) \quad (2)$$

（2）标量乘法如公式（3）所示。在 PAILLIER[44] 的方案中，$\odot_C$ 表示取密文的 n 次方。

$$Dec_{sk}([[u]] \odot_C n) = Dec_{sk}([[u \cdot n]]) \quad (3)$$

PHONG[45] 等人提出了一种改进的同态加密算法，以提高加密深度学习模型的性能。它的目标是在保护用户隐私的同时，尽可能保持模型的准确性。他们比较了几种不同的加密算法，并提出了一种新的方法，该方法在保护隐私的同时具有较高的计算效率。HAO[46] 等人的目标是提高联合学习中同态加密的效率和隐私保护能力。作者提出了一种基于密码学的联合学习框架，该框架可以在保护用户隐私的同时，提高计算效率。他们还比较了几种不同的加密算法，并提出了一种新的方法，该方法在保护隐私的同时具有较高的计算效率。CHAI[47] 等人提出了一种基于同态加密的联合矩阵分解算法，以保护用户隐私。作者比较了几种不同的加密算法，并提出了一种新的方法，该方法在保护隐私的同时具有较高的计算效率。但这些方案无法防御存在恶意参与方勾结聚合服务器的联邦学习场景。

为应对恶意参与方勾结聚合服务器的情况，文献[48,49]在前人的基础上提出了共享密钥的方案。参与方先使用自己的公钥对本地更新进行加密，再用共享的公钥再一次进行加密。FANG[50]等人不是直接共享公钥，而是拆分自己的私钥并共享给其他参与方。但是以上方法无法抵御恶意服务器勾结多个参与方的情况。

随后，FROELICHER[51]等人提出了一种新的隐私保护联邦学习框架，该框架利用同态加密技术对用户的数据进行加密，同时利用数据共享和计算，以减少通信开销和提高计算效率。SAV[52]等人结合混淆矩阵和同态加密算法，提出了一种新的隐私保护框架，在保护用户隐私的同时，有效提高联邦学习的效率。

BONAWITZ[53]等人将实体分为了两类：一个单独做聚合操作的服务器S，以及包括n个客户端的集合U。集合U中的每个用户u都有一个私有的m维向量$\boldsymbol{x}_u$，且$\sum_{u\in U}\boldsymbol{x}_u$在域$Z_R$中。协议的目标为安全地计算$\sum_{u\in U}\boldsymbol{x}_u$：保证服务器只学习到最终的求和结果，而用户什么也学习不到。在用户之间有顺序的前提下，且每对用户(u,v)共有一个随机向量$\boldsymbol{S}_{u,v}$，对$\boldsymbol{x}_u$做如公式（4）的盲化处理。然后服务器收到$\boldsymbol{y}_u$后进行如公式（5）的计算。

$$\boldsymbol{y}_u=\boldsymbol{x}_u+\sum_{v\in u,u<v}\boldsymbol{S}_{u,v}-\sum_{v\in u,u>v}\boldsymbol{S}_{v,u}\left(\bmod R\right) \tag{4}$$

$$\begin{aligned}z&=\sum_{u\in \mathrm{u}}\boldsymbol{y}_u=\sum_{u\in \mathrm{u}}\left(\boldsymbol{x}_u+\sum_{v\in u,u<v}\boldsymbol{S}_{u,v}-\sum_{v\in u,u>v}\boldsymbol{S}_{v,u}\right)\\&=\sum_{u\in \mathrm{u}}\boldsymbol{x}_u\left(\bmod R\right)\end{aligned} \tag{5}$$

为了防止服务器因用户发送信息太慢误判“用户离线”而导致的安全问题，他们还提出双盲，在用向量$\boldsymbol{S}_{u,v}$进行盲化时加入自己生成的随机向量$\boldsymbol{b}_u$，如公式（6）所示。

$$\begin{aligned}\boldsymbol{y}_u=\boldsymbol{x}_u+PRG\left(\boldsymbol{b}_u\right)+\sum_{v\in u,u<v}PRG\left(\boldsymbol{S}_{u,v}\right)-\\\sum_{v\in u,u>v}PRG(\boldsymbol{S}_{v,u})\left(\bmod R\right)\end{aligned} \tag{6}$$

然而传统的安全聚合协议只考虑在一次训练中的隐私保证。虽然安全聚合在任何一轮都有可证明的隐私保证，即在每一轮聚合模型之外都没有信息泄露，但无法有效抵抗跨多个训练轮的攻击。具体地说，通过使用聚合模型和跨多轮的参与信息，可以从聚合模型重构单个模型。SO[54]等人提出了一种新的安全聚合框架，该框架具有多轮隐私保证。他们还介绍了一种新的度量标准，用于量化联邦学习在多个训练轮次上的隐私保证，并开发了一种结构化用户选择策略，以保证每个用户在任意数量训练轮次上的长期隐私。

在联邦学习场景中使用同态加密技术提供隐私保护，主要是为了防范不可信的服务器，所以需要参与训练的用户之间相互通信、相互验证，以保证各个用户的信息不被窃取。但是如果存在部分恶意的参与用户，还需要第三方认证机构的介入，以保证用户之间通信的可靠性。

4）部分共享

由于参与训练的用户在上传梯度更新时会受到数据重构等推理攻击的侵害，所以文献[55,56]提出只上传部分具有代表性的数据，具体来说，他们考虑到上传的参数绝对值大小对全局模型的影响程度不一样，所以让每个参与模型训练的用户选择绝对值较大的梯度更新数据上传，并且最小化其他参数，如设置为0。针对带有批标准化层[57]的联邦学习场景，ANDREUX[58]等人提出模型训练的用户只需上传批标准化层的更新参数，在减少梯度信息泄漏的同时，提高了模型在数据异构场景中的鲁棒性。

3 完整性攻击与保护

模型在推理阶段和训练阶段最容易受到完整性攻击，模型的完整性一旦被破坏，模型的预测结果就会发生偏离。在推理阶段可能会受到对抗攻击，训练阶段则会受到模型投毒攻击和数据投毒攻击。

3.1 对抗攻击

利用不同类各实例之间的边界生成使模型错误分类的输入样本，利用深度学习的缺点破坏识别系统的方法，即对输入样本故意添加一些人为无法察觉的细微的干扰，导致模型以高置信度给出一个错误的输出，

在模型的推理和预测阶段，通过在原始数据中加入精心设计的微扰，攻击者可以获得对抗样本，从而欺骗深度学习模型使其给出高信度的误判。模型在预测阶段做出错误的判断，错误分类虽不会直接侵犯联邦学习参与者的隐私数据，但会导致模型的准确性和可用性受到影响。

SZEGEDY[59]等人首次提出了对抗样本的概念。他们指出了深度神经网络学习的输入-输出映射在很大程度是不连续的，对测试图像应用一个不可察觉的非随机扰动，可以任意改变网络的预测。

对抗样本攻击在机器学习领域已有广泛研究，在联邦学习场景下的模型部署与机器学习类似，传统的对抗样本攻击可以拓展到联邦学习中，但目前针对联邦学习的对抗样本攻击研究尚少。

WANG[60]等人研究了对抗样本攻击与后门攻击之间的联系。他们提出的边缘案例后门（Edge-Case Backdoor）在数据集的边缘部分插入后门更容易实现有效的对抗样本攻击。这表明模型对后门的鲁棒性在通常情况下意味着对于对抗样本攻击的鲁棒性。PANG[61]等人则针对纵向联邦学习中用户的特征差异，提出了对抗性主导输入攻击。与传统的对抗性样本控制整个特征空间不同，对抗性主导输入攻击仅仅控制部分特征输入，以削弱其他用户输入的影响，从而影响了激励用户贡献的奖励，实现对特定的输入进行错误分类的目的。

SHI[62]等人将一种在迭代过程中通过沿损失函数的梯度方向积累速度向量来增加扰动产生对抗样本的有效方法MIFGSM扩展成联邦学习框架下的FED-MIFGSM，用对抗样本来破坏训练模型的准确性。他们假设攻击者知道服务器使用的聚合算法，但不知道良性客户端进行的更新。恶意客户端在训练阶段使用对抗样本作为输入，训练恶意更新，然后将恶意更新上传到服务器，目的是使全局模型错误分类。他们针对Krum、Bulyan、Trimmed Mean和Median4种聚合算法对FED-MIFGSM进行了不同的优化，有效地对全局模型进行了毒害。但是这种攻击对于恶意客户端比例较小的部分聚合算法是无效的。

3.2 数据投毒攻击

数据投毒攻击指攻击者污染或恶意修改了训练集中的部分数据，不改变目标联邦学习系统情况下，构造特定的输入数据来欺骗系统完成攻击，具体来说，通过恶意修改目标用户的训练数据来实现数据投毒，如图4所示。

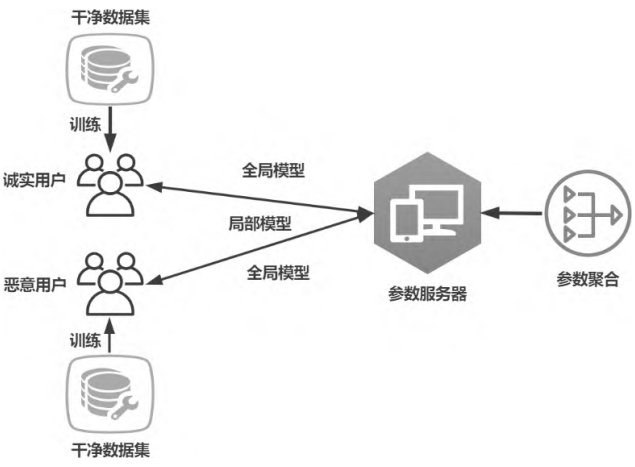

图 4 联邦学习中的数据投毒

BIGGIO[63]等人提出通过修改目标用户的训练数据来实现数据投毒，最常见的方法有伪造训练数据和改变原有数据的标签，这些方法可以使模型训练的结果对应的是错误的推理逻辑，并且可以对支持向量机进行攻击，可以有效地降低模型的准确性。

TOLPEGIN[64]等人提出的标签反转攻击考虑到了参与方中的恶意方的个数、迭代轮数和参与模型聚合的概率，从这3个维度证明他们攻击的有效性，不仅会降低全局模型在推理阶段的准确性，且对全局模型的负面影响与恶意参与者比例正相关，同时，发现适当增加靠后的训练轮次中恶意参与者参与训练的概率，可以有效提高攻击的效果。最后，他们根据恶意参与者发送的参数更新具有的独特特征，利用PCA降维思想（Principal Component Analysis）提出了一种防御策略，证明了该防御能够识别恶意参与者，并且对梯度漂移具有鲁棒性。

ZHANG[65]等人提出利用GAN[66]生成具体其他参与方特征的样本。其中，利用上一轮的全局模型作为鉴别网络，再对这些“伪样本”进行标签反转攻击。实验证明，通过更改历元数和学习速率等训练配置，可以使中毒的全局模型在主任务和目标任务上都满足预期。在利用生成对抗网络进行数据投毒攻击的基础上，ZHANG[67]等人在前人利用GAN进行投毒攻击的基础上，进一步提出了Generative Poisoning Attacks投毒方案，通过恶意修改本地训练的超参数并对局部模型进行缩放，从而更有效地降低全局模型的性能。

综上，主流数据投毒方式分为4类，如表1所示。

表 1 数据投毒方法主流方法总结

投毒方式	投毒思路
基于标签反转 [63,64,67]	通过直接修改目标类别的训练数据的标签信息，而数据的特征保持不变
基于目标优化 [64,67]	将要解决的目标转化为一系列最优解问题。在数据投毒中，目标问题通常是制作最有效的中毒样本，既可以用于计算标签中毒的最佳数据点集，也可用于找到最有效的数据修改方案
基于梯度优化 [62]	使中毒样本朝着对抗目标函数L对中毒样本x'_c的梯度方向$\frac{\partial L}{\partial x'_c}$移动，直到实现最大的投毒效果
基于干净标签 [59,60,62]	干净标签数据投毒攻击中，中毒图像的标签与视觉感官是一致的，但测试图像会被错误分类

3.3 模型投毒攻击

模型投毒攻击主要指攻击者在全局聚合过程中通过发送错误的参数或破坏模型来扰乱联邦学习过程的方式。通过控制学习参与者传递给服务器的更新参数，影响整个学习过程模型参数走向和降低模型的收敛速度，甚至破坏训练模型的正确性，影响联邦学习模型的性能。

联邦学习中，数据投毒攻击没有将数据发送到服务器，模型投毒攻击则因将数据发送到服务器而需要复杂的技术和较高的计算资源，其综合效果比数据投毒攻击更有效。

目前联邦学习中最典型的算法FedAvg[6]、FedProx[68]等都是基于梯度更新进行线性组合的安全聚合算法，恶意参与方可以按照自己的目标去篡改上传的梯度更新以影响全局模型，因此全局模型更容易受到模型投毒攻击的侵害。现有的部分方案，如Krum[69]、Trimmed-mean[70]等算法，虽可有效地抵抗恶意节点，降低恶意节点带来的侵害，但依旧可以被绕过而无法抵抗模型投毒攻击。

MHAMDI[71]等人的方案比较直接，他们直接勾结多个恶意参与方，每个参与方虽然仅仅修改梯度更新的单个参数，但在服务器聚合梯度后依旧对全局模型造成了较大的毒害，有效降低了模型的准确率，因此成功绕过了Krum算法的防御。与MHAMDI[71]等人不同的是，BARUCH[72]等人的投毒攻击发生在多个训练轮次中，他们在每一轮中的梯度毒害更小，毒害更难以被发现，对全局模型收敛的影响更小。然而XIE[73]等人提供了新的投毒方向，他们构造的恶意梯度更新在与其他参与方的梯度更新聚合后，会与未被投毒的聚合梯度方向夹角大于90度，这种利用梯度方向的投毒方式同样绕过了各种安全聚合算法，更加有效地降低了模型的性能。

BHAGOJI[74]等人也提供了更为新颖的定向投毒框架，他们根据投毒目标求出全局模型交叉熵损失最小的阈值δ，然后在全局模型上增加本地梯度更新δ。为了防止聚合服务器可能会检测到恶意梯度，他们还将投毒者的局部训练损失设置到目标函数内，可有效权衡模型投毒的效果和投毒的成功率，这样的攻击更加隐蔽。他们还尝试使用多种用于防御恶意训练方的算法，如Krum算法和基于坐标中位数的聚合算法，用来防御上述攻击算法，但实验结果证明这两个防御算法都无法抵抗该攻击算法。最后为了更进一步提高攻击算法的隐蔽性，提出了一种交替最小化方法，该方法既考虑了模型中毒中攻击目标的最小化，又考虑到了隐身目标的最小化，并使恶意权重更新在绝大部分回合中都能避免被发现。

FANG[75]等人针对参与者攻击的进一步细化，将训练集与训练区分开，将其看作是独立的两个过程，这对解决抗拜占庭攻击的FL比较有针对性。他们同样是利用求解最优化问题的思路去求出最大的偏离向量λ，具体来说，他们在上一轮全局模型参数变

化的反方向添加恶意向量λ，有效地让全局模型的性能受到了侵害。不仅如此，他们为了绕过Krum[69]、TrimmedMean[70]和Median[70]算法对投毒攻击的限制，将这些算法的检测条件加入到了模型训练的目标函数中，有效提高了模型投毒攻击的成功率。

FANG[75]等人是对全局模型投毒，SHEJWALKAR[76]等人是针对给定数据集个性化地设置扰动向量，使得定向攻击的效果更好。他们提出了一种模型投毒攻击可有效绕过多种AGR聚合规则（Aggregation Rules）。具体来说，他们首先构造出基本的扰动向量∇^p，将∇^p设置为正常梯度更新的均值向量、正常梯度更新的标准差向量和正常梯度更新均值的符号函数；然后证明了这3种构造方式可以应对多种安全聚合算法；最后根据目的去求解最优的比例系数λ，将添加的恶意梯度方向设置为∇^p的相反方向上的$\nabla^p \times \lambda$。

3.4 后门攻击

后门攻击是通过故意改变决策边界使某些输入被错误分类，攻击者能够在模型中插入隐藏的后门，并在预测阶段通过触发简单的后门触发器完成恶意攻击，其危害性较大，如图5所示。在联邦学习中，攻击者可在训练集中插入隐蔽的后门以实现定向投毒攻击，还可根据自己的恶意目的去上传精心设计的梯度更新来插入后门，然后以模型投毒的形式干扰全局模型在某些场景下的推理结果。

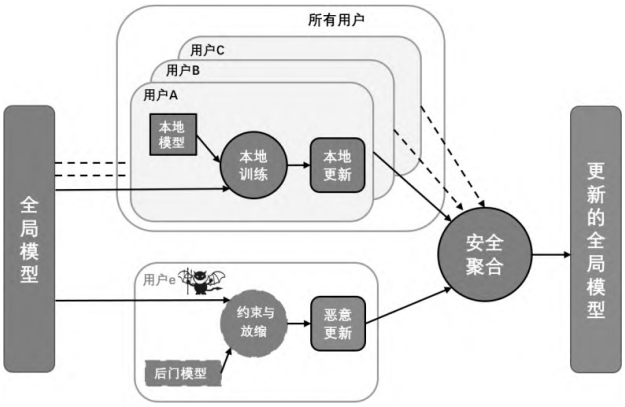

图 5 联邦学习中的后门攻击

BAGDASARYAN[77]等人通过实现模型投毒来完成后门攻击，其思路是恶意方在本地训练生成的恶意模型会与上一轮的全局模型组合，然后上传至服务器进行模型聚合，聚合的结果是自己精心设计、被毒害后的全局模型。由于恶意方上传的恶意模型可能会被各种安全检测发现，所以他们还提出恶意方在本地需要训练恶意模型至收敛，然后适当提高该恶意模型的权重以应对聚合服务器对单个模型效果的削弱。BARUCH[78]等人则是在文献[77]的基础上提出了新的植入后门的方式，他们利用Krum算法[69]流程的输出只有一个被选择的节点，而且最优的节点也会有偏离均值的参数这一缺点，生成一组偏差浮动范围的后门模型参数，以实现绕过Krum算法的防御。

SUN[79]等人提出了一种带有限制的攻击方式，可有效防止因恶意梯度过大而被聚合服务器检测出来且被裁剪的情况，具体来说，与正常参与方更新梯度相比，恶意梯度不会有较大的增加，此时既不容易被检测出来，也可以有效实现攻击目的。

WANG[60]等人考虑在数据集的边缘部分插入后门以更容易实现攻击的目的。边缘案例后门迫使模型对看似简单的输入进行错误分类，但这些输入不太可能是训练或测试数据的一部分，也就是说，它们分布在输入分布的末端。通过实验发现，在边缘案例后门攻击形式下，目标预测子任务非常隐蔽，不太可能被发现在训练或测试数据集。XIE[80]等人则不是将后门直接插入到恶意模型中，而是将需要插入的后门的触发器进行拆解，然后由多个恶意参与方进行插入，如此便可降低插入的触发器对单个上传的模型的侵害，更容易绕过安全聚合算法的防御，提高后门的植入成功率，而且模型更容易收敛，更适用于动态的、多次的攻击场景。

但在以上技术中，对手选择的触发器通常独立于学习模型和学习过程，如标志、贴纸或像素扰动而产生。因此，这种后门攻击在训练阶段并没有充分利用多个恶意用户之间的协作。为了解决这一不足，GONG[81]等人引入了Coordinated Trigger后门攻击，在这种攻击中，敌

手利用依赖于模型的触发器来更有效地注入后门。模型相关触发器是每个恶意参与者的最佳触发器配置。这是通过一个子训练过程来完成的，该过程在形状、大小和位置方面寻求触发器的理想值分配。在为每个对抗方生成本地触发器后，本地训练数据集将根据触发器中毒。在推理步骤中，通过结合局部触发器构造全局触发器。因此，模型相关触发器比以往的标准随机触发器更有效。

综上，具体的完整性对抗关系如图6所示。

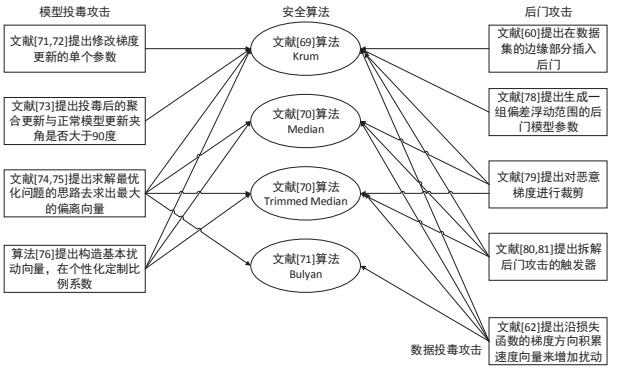

图 6 完整性对抗关系

3.5 隐私保护方法

面对联邦学习存在的各种完整性攻击的威胁，广大学者研究出了多种隐私保护方案以应对各种威胁。而目前最主要的完整性防御手段主要有安全聚合、区块链技术、可信执行环境和模型加固这4种。针对各种完整性攻击，学者们提出了与之对应的防御方式以保证隐私信息的完整性不会受到侵害。具体攻防对抗的关系如图7所示。特别的是，由于后门攻击中的触发器主要是通过投毒攻击（涵盖数据投毒攻击和模型投毒攻击）嵌入的，故可以防御投毒攻击的隐私保护方法同样可以保护隐私信息免于后门攻击的侵害，故图7没有将后门攻击单独列出，而是默认后门攻击于数据投毒攻击和模型投毒攻击中。

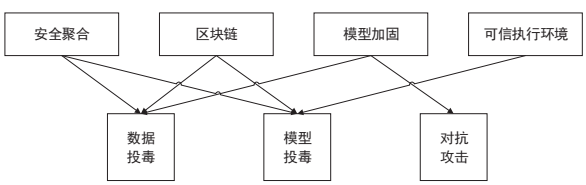

图 7 联邦学习的完整性攻击和防护

1）安全聚合

由于联邦学习的分布式学习特性，有效的安全聚合算法可以相当程度地降低恶意参与节点的侵害，也可以使得全局模型正常收敛。目前安全聚合算法主要分为以下两类：基于模型更新特征差异的聚合算法和基于验证数据集的聚合算法。

为了降低全局模型的精确性与完整性，攻击者会训练并上传恶意模型更新，但是通常情况下，该恶意更新会跟普通用户的更新有较大差异，因此聚合服务器可以通过衡量不同更新之间的过大差异性来判断某些梯度更新是否是恶意的，从而可以有效保证全局模型不会被毒害。

最简单的方法就是根据梯度的 l_p 距离来衡量两个梯度的差异，如基于梯度平均值或中位数的聚合算法。因此BLANCHARD[69]等人提出对于每一个梯度，求出它与其他梯度的最近的 $n-m-2$ 个距离，对这些距离求和，然后选择一个求和距离最小的梯度作为聚合后的全局梯度，这便是Krum算法。在此基础上，他们构思可以选择 k 个求和梯度较小的梯度以提高模型健壮性而提出Multi-Krum算法，也就是最后取这 k 个梯度的平均梯度作为安全聚合的结果。MHAMDI[71]等人提出的Bulyan安全聚合算法同样是选择 k 个求和梯度较小的梯度，但最后的聚合结果取的是 k 个被选择梯度的截尾平均数。而CHEN[82]等人提出先将所有参与方划分进 k 个组，然后计算每个组的所有梯度的平均值，最终的聚合结果就是所有均值的几何中位数。文献[70,83]提出对所有上传的梯度的每一个维度都取中位数，然后将由所有维度的中位数构成的梯度作为最终的聚合梯度。

另外一个更为行之有效的安全聚合方案是利用余弦相似度来凸显出正常更新与恶意更新之间的差异。MUÑOZ-GONZÁLEZ[84]等人提出用余弦相似度这个新的思路去筛选出恶意梯度，具体来说，他们计算所有参与方的梯度更新汇总后的加权平均值，然后求该加权平均值与每个用户上传的梯度更新的余弦相似度，如果超出一定阈值则判定该用户的上传梯度为恶意梯

度并舍弃，则最终的聚合输入为所有良性梯度的聚合结果。KHAZBAK[34]等人同样利用余弦相似度，不让每个参与方的梯度更新与所有梯度更新的加权平均值相比，而是计算与其他所有参与方梯度的余弦相似度，然后根据这些相似度和权重计算相似程度k，而聚合输入为相似程度k较高的前$n-m$个梯度更新。但文献[34,84]都在限制恶意参与方的数量，因为其方案都是在寻找与多数参与方的梯度有较大不同的恶意梯度，所以其防御手段不适用于恶意用户比较多的联邦学习场景，文献[85]做出了进一步的研究。FUNG[85]等人没有对恶意参与方的数量做约束，只要求至少有一个正常参与方，并提出了新的防御方案FoolsGold，在受到定向攻击的联邦学习场景中，恶意用户更新的特征比较单调。具体来说，如果某个参与方的历史聚合更新与其他所有参与方的梯度更新的余弦相似度比更高，那么说明该参与方是恶意参与方的可能性比较大，因此就降低该用户的梯度更新在聚合时的权值。

除了利用余弦相似度来区分恶意用户，还有众多学者利用有效的聚类算法削弱恶意梯度更新的影响。YU[86]等人利用的聚类算法是K-means[87]，其提出在使用一般聚合方案前先用聚类算法将所有梯度更新进行聚类分组，这样可以更加高效地区分出恶意和非恶意的梯度更新。而SINGH[88]等人提出的聚类方案更加灵活，其根据参与方的公开属性（如地区、偏好、性别等）进行分组，如此一来，在恶意方较少的一组中，恶意参与方的影响会因有更多的良性用户而被削弱，同时在恶意方较多的一组会因非正常属性的叠加导致该组的聚合权重被大幅降低。因此该聚类算法有效地提高了联邦学习对于投毒攻击的健壮性。

综上，基于特征更新差异的安全聚合的研究方法如表2所示。

除了可以跟其他参与方的更新进行对比来验证更新的真实性之外，还可以根据待验证的模型更新在验证数据集上的表现去判断该更新的正确性。

WANG[89]等人率先提出该类验证性算法，其让

表 2 基于特征更新差异的安全聚合的研究方法

研究	特征依据	具体方案
BLANCHARD[69]等人	欧氏距离	Krum、Multi-Krum
MHAMDI[71]等人		Bulyan
CHEN[82]等人	中位数	取分组后均值的中位数
XIE[83]等人		每个维度都取中位数
YIN[70]等人		
KHAZBAK[34]等人	余弦相似度	“用户梯度更新”与“其他所有参与方的梯度”
MUÑOZ-GONZÁLEZ[84]等人		“用户梯度更新”与“所有梯度更新的加权平均值”
FUNG[85]等人		“用户的历史聚合更新”与“其他所有参与方的梯度更新”
YU[86]等人	聚类算法	K-means
SINGH[88]等人		按照公开属性（如地区、偏好、性别等）聚类分组

中心服务器计算所有收到的梯度更新在验证数据集上的推理准确率，并判定准确率低于给定值的梯度更新的所有者为恶意用户，而聚合的输入为除去恶意梯度之后的所有良性梯度更新。CHEN[90]等人首先利用K-means算法对参与方模型更新的l_2距离进行聚类分组，然后计算每个聚类分组在验证数据集上的推理准确性，丢弃准确性低于一定阈值的组，再对剩下的组进行聚合运算。

XIE[91]等人利用验证数据集计算每一轮中的损失函数下降Δl和模型更新Δm，然后其取Δl较大且Δm较小的k个参与方的局部更新作为聚合的输入，同时丢弃剩下的$n-k$个局部更新。

FANG[75]等人在收到每个参与训练的参与方的本地更新Δm后，在验证数据集上分别计算聚合Δm和不聚合Δm所得到的全局模型的正确率，两个正确率之差越大说明该本地更新Δm对全局模型的影响越大，所以其主张丢弃正确率之差较大的k个本地更新，剩下的$n-k$个本地更新作为安全聚合的输入。这种方法虽排除了较大偏离度的本地更新的侵害，却可能会降低模型的收敛速度。

2）区块链

联邦学习与分布式机器学习不同的是联邦学习需

要一种去中心化分布系统来保证用户的隐私安全，在保障数据安全和交换、训练效率前提下进行有效的机器学习，去中心化联邦学习如图8所示。区块链作为一个去中心化、数据加密和不可篡改的分布式共享数据库，可为联邦学习的数据交换提供数据保密性来对用户隐私进行保障，保证各参与方之间的数据安全，也可保证多参与方提供数据进行模型训练的数据一致性，区块链的价值驱动激励机制也能增加各参与方之间提供数据、更新网络模型参数的积极性。目前已经有一些应用区块链的联邦学习系统提出了基于区块链的应用优点。

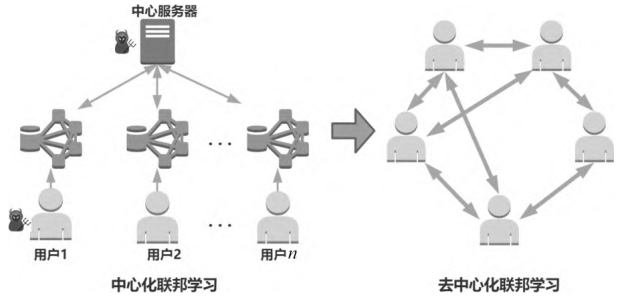

图 8 去中心化联邦学习

区块链具有去中心化的特征，不需要中心服务器对多个参与方的数据进行统一化计算、处理、更新，而是可以由部分被选择、被认可的节点代替原本的聚合服务器进行操作，并且将数据的运转都记录在链上，这不仅实现了数据的可验证性，而且有效防御了恶意服务器的各种完整性攻击。基于区块链的防御手段如表3所示，具体包括各方案的共识算法以及涉及的聚合节点。

表 3 基于区块链的防御手段

研究者	节点分类	共识机制	聚合节点
BAO[92] 等人	所有参与方	基于可信度	可信度最高的结点
LI[93] 等人	委员会、其他参与方	选举委员会	委员会成员
PENG[94] 等人	委员会、其他参与方	优先级	委员会
SHAYAN[95] 等人	验证、聚合委员会、其他参与方	PoF	聚合委员会
QU[96] 等人	区块链矿工	工作量证明	PoW 赢家
LIU[97] 等人	区块链矿工	PoS	聚合服务器

BAO[92] 等人为了增加每个参与方的数据可信度，提出将参与方的个人数据、每一轮的训练数据（包括本地梯度更新，上一轮下发的全局模型等）保存在事先设计好的区块链FLChain中，参与方可自发在每一轮中从FLChain中下载其他参与方的模型或上传自己的本地更新等数据。将参与方的各数据都记录在区块链上的另外一个好处是，参与方可以相互验证数据的真实性，可以及时检测出恶意的梯度更新，有效提高抵抗完整性攻击的能力。

LI[93] 等人提出将所有参与方分成委员会和普通参与方两个部分，委员会由达成共识的部分参与方组成，负责接收普通参与方记录在普通链上的本地模型更新。委员会在收到普通参与方的本地模型更新后，在自己的本地数据上进行验证，只有验证准确率较高的本地模型才能被认可并保存在委员会的区块链上。只有当委员会保存了足够的模型更新才会进行安全聚合，然后将新的聚合结果记录在区块链上，在每一轮的最后会选择本轮验证准确率较高的参与方作为委员会委员。这种利用验证集的准确率决定数据是否有效、决定委员会委员的方式，有效防御了完整性攻击，提高了联邦学习系统的鲁棒性。PENG[94] 等人利用区块链的共识机制，按照优先级得到一个委员会来集体聚合模型，并将可验证的证明记录在区块链中，从而提供了可验证性。SHAYAN[95] 等人利用区块链的共识机制，通过可验证随机函数（VRF）和证明联邦（PoF）来选择验证委员会和聚合委员会，从而提供了去中心化和可信任性。委员会也分成两个部分，委员会的职责分别是用Multi-Krum算法检验客户端上传的本地模型更新是否真实有效和聚合有效模型更新，这可以有效提高系统的安全性和鲁棒性。

QU[96] 等人提出了一种基于区块链系统大数据驱动认知计算的去中心化范式，结合了联邦学习和区块链技术，具体来说，其用多个矿工组成的区块链代替聚合服务器收集设备上传的局部更新再生成区块，然后通过工作量证明（PoW）选出赢家，再由赢家聚合所有局部更新后记录在区块链上，实现了去中心化和可信任的模型训练与更新。

LIU[97] 等人提出了一个基于区块链的联邦学习框

架来创建智能合约，并结合以太坊的PoS[98]共识算法。具体来说，其没有舍弃中心聚合服务器，而是让聚合服务器自动执行智能合约以测试收到的梯度更新是否真实有效。以太坊的矿工将验证通过的梯度更新发送给聚合服务器。

3）可信安全计算

上述各种攻击都是发生在训练阶段，一旦模型投毒攻击绕过训练阶段，直接上传恶意的梯度更新至聚合服务器，那么多数完整性防御手段将无法有效应对，因此部分研究者考虑利用可信执行环境（TEE）来保护本地训练不受到外界攻击的侵扰。典型的为用户提供可信执行环境的软件保护扩展SGX[99]是由因特尔公司推出的一套指令集扩展，SGX禁止所有其他软件访问或修改enclave内部的代码和数据。

CHEN[100]等人提出在可信执行环境的enclave中进行参与方的本地训练和中心服务器的聚合操作，有效地保证了本地训练的完整性不会被破坏，同时防止恶意参与方绕过训练阶段并直接上传恶意的梯度更新。不仅如此，参与方与中心服务器之间上传梯度更新、下发全局模型的过程也在enclave中的安全通道中进行，有效地保证了数据不会被恶意篡改、替换。但SGX存在效率低下的问题，因为它不支持GPU加速，所以不太适用于训练深度学习的场景，因此ZHANG[101]等人提出在用GPU训练深度模型时依旧将部分训练轮次转移到可信执行环境中进行，该方案可在考虑到效率的同时兼顾训练的安全性。

因此利用可信执行环境可以保护联邦学习系统不受模型投毒的攻击，但是无法防御数据投毒攻击，因为如果参与方的训练数据存在问题，会导致在可信执行环境中训练被投毒的数据，不可避免地降低模型的准确性。

4）模型加固

针对可信执行环境无法抵御的数据投毒攻击，后续研究者认为可以对模型进行加固，以有效降低由数据投毒产生的有毒数据造成的消极影响，同时提高全局模型的鲁棒性。模型加固的优点是可以提高模型的安全性和可信度，防止数据泄露或模型被盗用，同时可以增强模型的鲁棒性和稳定性，抵抗噪声、异常或敌对的数据影响，但是模型加固的手段与特点，不可避免地会引入额外的误差或噪声，影响模型的精度；同时也会存在技术上的难度或局限性，难以适应复杂和动态的场景。

ZHAO[102]等人首次在边缘计算领域对后门攻击的影响进行了讨论，并提出了多种稳定性诱导操作，常见的有Dropout、Weight decay等，可以有效地提高本地训练时的模型健壮性，同时提高模型的泛化能力，最后通过大量智能边缘计算场景下的实验证明Dropout等操作可有效削弱后门攻击，提高全局模型的鲁棒性。ZHANG[103]等人提出了一种基于对抗训练技术的防御方法，利用两个模型（f_1和f_2）的交叉优化来提高全局模型的鲁棒性，具体来说，f_1是全局模型，其目标是最小化全局损失函数；f_2是一个分类模型，负责判断收到的梯度的归属方。为了降低分类模型f_2的分类准确率，需要模糊恶意梯度更新和良性梯度更新之间的界限，并将f_2的损失函数加入到目标函数中一起训练直至收敛，以提高全局模型的鲁棒性和防御受到毒害的样本。

IBITOYE[104]等人提出不需要修改联邦学习的通信协议和优化算法，只需要在每个客户端的本地模型中添加差分隐私和自归一化的层。具体来说，首先，其利用差分隐私噪声来增加模型的随机性和不确定性，从而降低对抗样本[105]（Adversarial Samples）的影响；然后，利用自归一化技术来保持模型的稳定性和收敛性，从而提高模型的泛化能力。最终实现简单且可扩展的防御方案，可以有效地提高模型的鲁棒性。

4 总结与展望

4.1 机密性相关的挑战与展望

现有几乎所有的成员推理攻击都非常依赖于辅助数据，而恶意成员拥有的数据难以推断出额外的信息，所以现有的研究几乎都是通过一定的方式获得足够多

的辅助数据来增加数据的多样性，以更好地拟合目标模型，从而推理出相关成员。因此可知，过拟合是导致模型易收到成员推理攻击的重要原因，且样本会在函数梯度上留下明显的痕迹，所以有必要让服务器无法记录任何有关本地模型更新的任何信息，所以能否可以找到一种方式去减弱此种痕迹，是否可以增加训练过程将这种痕迹被另外一种无关痛痒的痕迹所替代，需要进一步研究。

4.2 完整性相关的挑战和展望

现有的大多数FL投毒攻击和后门攻击依赖于不同的假设，包括对被攻击客户的百分比、FL客户的总数量和训练数据的全局分布的假设。例如，最新的攻击[60,77]使用从全局分布中抽取的良性样本来操纵中毒数据集。其他攻击[60,67]需要被攻击客户端的持续参与，或需要大量恶意客户端的参与。这一假设受到了文献[106]的挑战，并被证明是不切实际的。因此，探索设计攻击策略的可能性将是有趣的，这种策略需要有限的假设，并可以应用于各种场景，如在对系统操作的知识有限的大规模FL系统中。为了实现这一目的，对手可以通过共享全局模型[67]利用泄漏信息，模拟与全局数据分布一致的辅助训练数据集，以加强能力有限的对手的后门影响。此外，当对手只控制了一小部分参与者（即小于0.1%）时,其可以考虑设计单枪（One Shot）攻击，延长后门的耐久度。

由于模型投毒对于攻击者的要求比较高，需要有至少一个或多个恶意方相互勾结，随着联邦学习部署应用的延伸，越来越少的敏感且脆弱的参与方使得模型投毒的门槛日益提高。与此同时，与模型投毒相比，数据投毒对攻击者能力要求较低，因而有更多的实施空间，且在联邦学习的规模比较大时更不容易被察觉。然而现有数据投毒攻击的研究不够深入，主要是在验证各种投毒手段是否可行，缺乏对投毒效果的优化。由于数据投毒贯穿于参与方的本地训练阶段，所以其产生的恶意更新需要与良性更新有较大相似性才能保证不被检测出。是否可以根据受害者的数据特征对被数据投毒后的本地更新再进一步优化，以绕过现有异常检测聚合算法的防御；如何防止数据投毒对全局模型的毒害效果被各种安全聚合算法削弱等问题都亟待深入研究。

目前来看，关于联邦学习的机密性和完整性保护都具有相当程度的独立性，但在实际应用中，所有威胁都是同时存在的，无法对以上的各安全手段进行简单的叠加，所以需要更加深入的研究去解决不同安全算法叠加可能造成的精度降低和部分安全算法失效等隐患。

5 结束语

随着人工智能与联邦学习的发展与应用，随着GPT 4.0的爆火，联邦学习模型的安全和隐私问题吸引了许多学者的关注，关于训练数据可能会造成的隐私泄露引发多个国家政界和学者的担忧，目前尚有很多安全问题急待解决。本文在深度调研其他学者研究成果和学习的基础上，对联邦学习在攻防对抗领域的研究成果进行综述，系统地总结归纳了联邦学习在机密性与完整性两个方面所面临的威胁，并对现有的隐私保护方式进行了分析与总结。联邦学习依旧存在诸多问题有待解决，未来需要更多的深入研究。

参考文献：

[1] POUYANFAR S, SADIQ S, YAN Yilin, et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications[J]. ACM Computing Surveys, 2019, 51(5): 1−36.

[2] HATCHER W G, YU Wei. A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends[J]. IEEE Access, 2018(6), 24411−24432.

[3] GOODFELLOW I, COURVILLE A, BENGIOB Y. Deep Learning[M]. Cambridge: MIT Press, 2016.

[4] TRASK A W. Grokking Deep Learning[M]. Greenwich: Manning Publications, 2019.

[5] YANG Qiang, LIU Yang, CHEN Tianjian, et al. Federated Machine Learning: Concept and Applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1−19.

[6] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication−Efficient Learning of Deep Networks from Decentralized Data[EB/OL]. (2023−01−26)[2023−10−10]. https://arxiv.org/abs/1602.05629.

[7] MCMAHAN H B, MOORE E, RAMAGE D, et al.Federated Learning of Deep Networks Using Model Averaging[EB/OL]. (2017−02−28)[2023−

10−10]. https://arxiv.org/abs/1602.05629v3.
[8] LIU Yang, YANG Qiang, CHEN Tianjian, et al. Federated Learning and Transfer Learning for Privacy, Security and Confidentiality[EB/OL]. (2019−02−21)[2023−10−10]. https://aisp−1251170195. cos.ap−hongkong. myqcloud.com/fedweb/1552916850679. pdf.
[9] YANG T, ANDREW G, EICHNER H, et al. Applied Federated Learning: Improving Google Keyboard Query Suggestions[EB/OL]. (2018−12−07)[2023−10−10]. https://arxiv.org/abs/1812.02903.
[10] HARD A, RAO K, MATHEWS R, et al. Federated Learning for Mobile Keyboard Prediction[EB/OL]. (2019−02−28)[2023−10−10]. https://arxiv.org/abs/1811.03604.
[11] CRAMER R, DAMGARD I, NIELSEN J B. Multiparty Computation from Threshold Homomorphic Encryption[C]//IACR.Proceedings of the International Conference on the Theoryand Application of Cryptographic Techniques: Advances in Cryptology. Berlin:Springer,2001: 280−299.
[12] DAMGARD I, NIELSEN J B. Universally Composable Efficient Multiparty Computation from Threshold Homomorphic Encryption[C]//IACR. Proceedings of Advances in Cryptology. Berlin: Springer,2003: 247−264.
[13] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Adances and Open Problems in Federated Learning[EB/OL]. (2019−12−10)[2023−10−10]. https://arxiv.org/abs/1912.04977.
[14] SHOKRI R, STRONATI M, SONG Congzheng, et al. Membership Inference Attacks against Machine Learning Models[EB/OL]. (2017−03−31)[2023−10−10]. https://arxiv.org/abs/1610.05820.
[15] MELIS L, SONG Congzheng, DE C E,et al. Exploiting Unintended Feature Leakage in Collaborative Learning[C]//IEEE.Proceedings of 40th IEEE Symposium on Security and Privacy. Berlin: IEEE, 2019: 691–706.
[16] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White−Box Inference Attacks against Centralized and Federated Learning[C]//IEEE. Proceedings of the IEEE Symposium on Security and Privacy. Berlin: IEEE, 2019: 739−753.
[17] CHEN Jiale, ZHANG Jiale, ZHAO Yanchao, et al. Beyond Model−Level Membership Privacy Leakage: An Adversarial Approach in Federated Learning[C]//IEEE. International Conference on Computer Communications and Networks. New York: IEEE, 2020: 1–9.
[18] ZHANG Jingwen, ZHANG Jiale, CHEN Junjun, et al. GAN Enhanced Membership Inference: A Passive Local Attack in Federated Learning[C]//IEEE.ICC 2020–2020 IEEE International Conference on Communications. New York: IEEE, 2020: 1–6.
[19] HITAJ B, ATENIESE G, PEREZ−CRUZ F.Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning[C]// ACM. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017: 603–618.
[20] WANG Zhibo, SONG Mengkai, ZHANG Zhifei,et al. Beyond Inferring Class Representatives: User−Level Privacy Leakage from Federated Learning. [C]//IEEE. IEEE Conference on Computer Communications. New York: IEEE, 2019: 2512–2520.
[21] SONG Mengkai, WANG Zhibo, ZHANG Zhifei, et al. Analyzing User–Level Privacy Attack against Federated Learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430−2444.
[22] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient–Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.
[23] ZHU Ligeng, LIU Zhijian, HAN Song.Deep Leakage from Gradients[C]//NIPS. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: NIPS, 2019: 14747−14756.
[24] GEIPING J, BAUERMEISTER H, DRÖGE H, et al. Inverting Gradients–How Easy is It to Break Privacy in Federated Learning?[C]//NIPS. Proceedings of the 34rd International Conference on Neural Information Processing Systems.Vancouver: ACM, 2020: 16937–16947.
[25] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[EB/OL]. (2017−01−30)[2023−10−10]. https://arxiv.org/abs/1412.6980.
[26] ZHAO Bo, MOPURI K R, BILEN H.IDLG: Improved Deep Leakage from Gradients[EB/OL]. (2020−01−08)[2023−10−10]. https://arxiv.org/abs/2001.02610.
[27] WEI Wenqi, LIU Ling, LOPER M, et al. A Framework for Evaluating Client Privacy Leakages in Federated Learning[C]//Springer. Computer Security−ESORICS 2020. Berlin: Springer, 2020: 545–566.
[28] DARIO P, DANILO F, GIUSEPPE A. Eluding Secure Aggregation in Federated Learning via Model Inconsistency[C]//ACM. Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2022: 2429–2443.
[29] SHEN Meng, WANG Huan, ZHANG Bin, et al. Exploiting Unintended Property Leakage in Blockchain–Assisted Federated Learning for Intelligent Edge Computing[J]. IEEE Internet of Things Journal, 2021, 8(4): 2265–2275.
[30] ZHOU Chunyi, GAO Yansong, FU Anmin, et al. PPA: Preference Profiling Attack against Federated Learning[EB/OL]. (2022−08−09)[2023−10−10]. https://arxiv.org/abs/2202.04856.
[31] YAO A C. Protocols for Secure Computations[EB/OL]. (2008−07−18)[2023−10−10]. https://ieeexplore.ieee.org/document/4568388.
[32] XU Runhua, BARACALDO N, ZHOU Yi, et al. HybridAlpha: An Efficient Approach for Privacy−Preserving Federated Learning[C]//ACM. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 13–23.
[33] BONEH D, SAHAI A, WATERS B. Functional Encryption: Definitions and Challenges[C]//Springer. Proceedings of the 8th Theory of Cryptography Conference IACR, Berlin: Springer. 2011: 253–273.
[34] KHAZBAK Y, TAN Tianxiang, CAO Guohong.MLGuard: Mitigating Poisoning Attacks in Privacy Preserving Distributed Collaborative Learning[C]//IEEE.Proceedings of the 29th International Conference on Computer Communications and Networks. New York: IEEE, 2020: 1–9.
[35] LI Yong, ZHOU Yipeng, JOLFAEI A, et al. Privacy−Preserving Federated Learning Framework Based on Chained Secure Multiparty Computing[J]. IEEE Internet of Things Journal, 2021, 8(8): 6178–6186.
[36] DWORK C. Differential Privacy[EB/OL]. (2006−07−10)[2023−10−10]. https://doi.org/10.1007/978−1−4419−5906−5_752.

[37] GEYER R C, KLEIN T, NABI M. Differentially Private Federated Learning: A Client Level Perspective[EB/OL]. (2018−03−01)[2023−10−10]. https://arxiv.org/abs/1712.07557.

[38] JAYARAMAN B, WANG Lingxiao, EVANS D, et al. Distributed Learning without Distress: Privacy–Preserving Empirical Risk Minimization[C]//ACM. Proceedings of the 32nd Internationall Conference on Neural Information Processing Systems. New York: ACM, 2018: 6346–6357.

[39] BHOWMICK A, DUCHI J, FREUDIGER J, et al. Protection against Reconstruction and Its Applications in Private Federated Learning[EB/OL]. (2019−06−03)[2023−10−10]. https://arxiv.org/abs/1812.00984.

[40] TRIASTCYN A, FALTINGS B. Federated Learning with Bayesian Differential Privacy[C]//IEEE.Proceedings of the IEEE International Conference on Big Data. New York: IEEE, 2019: 2587–2596.

[41] HUANG Xixi, DING Ye, JIANG Z L, et al. DP−Fl: A Novel Differentially Private Federated Learning Framework for the Unbalanced Data[J]. World Wide Web, 2020, 23(4): 2529–2545.

[42] WU Maoqiang, YE Dongdong, DING Jiahao, et al. Incentivizing Differentially Private Federated Learning: A Multidimensional Contract Approach[J]. IEEE Internet of Things Journal, 2021, 8(13): 10639–10651.

[43] RIVEST R L, ADLEMAN L, DERTOUZOS M L, et al. On Data Banks and Privacy Homomorphisms[J]. Foundations of Secure Computation, 1978, 4 (11): 169–180.

[44] PAILLIER P. Public–Key Cryptosystems Based on Composite Degree Residuosity Classe[C]//IEEE. Proceedings of International Conference on the Theory and Applications of Cryptographic Techniques. New York: IEEE, 1999: 223–238.

[45] PHONG L T, AONO Y, HAYASHI T, et al. Privacy–Preserving Deep Learning: Revisited and Enhanced[C]//ATIS.Proceedings of the 8th Internaitonal Conference on Applications and Techniques in Information Security. Berlin: Springer, 2017: 100–110.

[46] HAO Meng, LI Hongwei, XU Guowen, et al. Towards Efficient and Privacy–Preserving Federated Deep Learning[C]//IEEE. Proceedings of the 2019 IEEE International Conference on Communications. New York: IEEE, 2019: 1–6.

[47] CHAI Di, WANG Leye, CHEN Kai,et al. Secure Federated Matrix Factorization[J]. IEEE Intelligent Systems, 2021, 36(5): 11–20.

[48] FANG Chen, GUO Yuanbo, WANG Na, et al. Highly Efficient Federated Learning with Strong Privacy Preservation in Cloud Computing[EB/OL]. (2020−09−01)[2023−10−10]. https://doi.org/10.1016/j.cose.2020.101889.

[49] HAO Meng, LI Hongwei, LUO Xizhao, et al. Efficient and Privacy–Enhanced Federated Learning for Industrial Artificial Intelligence[J]. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6532–6542.

[50] FANG Chen, GUO Yuanbo, HU Yongjin, et al. Privacy–Preserving and Communication–Efficient Federated Learning in Internet of Things[EB/OL]. (2021−04−01)[2023−10−10]. https://doi.org/10.1016/j.cose.2021.102199.

[51] FROELICHER D, TRONCOSO–PASTORIZA J R, PYRGELIS A, et al. Scalable Privacy–Preserving Distributed Learning[J]. Proceedings on Privacy Enhancing Technologies, 2021, 2021(2): 323–347.

[52] SAV S, PYRGELIS A, TRONCOSO–PASTORIZA J R, et al. POSEIDON: Privacy–Preserving Federated Neural Network Learning[EB/OL]. (2021−01−08)[2023−10−10]. https://arxiv.org/abs/2009.00349.

[53] BONAWITZ K, IVANOV V, KREUTER B,et al. Practical Secure Aggregation for Privacy−Preserving Machine Learning[C]//ACM. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175–1191.

[54] SO J, ALI R E, GULER B, et al. Securing Secure Aggregation: Mitigating Multi–Round Privacy Leakage in Federated Learning[EB/OL]. (2023−07−27)[2023−10−10]. https://arxiv.org/abs/2106.03328.

[55] LI Wenqi, MILLETARÌ F, XU Daguang, et al. Privacy–Preserving Federated Brain Tumour Segmentation[EB/OL]. (2023−10−10)[2023−10−10]. https://link.springer.com/chapter/10.1007/978−3−030−32692−0_16.

[56] ZHAO Bin, FAN Kai, YANG Kan, et al. Anonymous and Privacy–Preserving Federated Learning with Industrial Big Data[J]. IEEE Transactions on Industrial Informatics, 2021, 17: 6314–6323.

[57] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]//ACM. Proceedings of the 32nd International Conference on Machine Learning. New York: ACM, 2015: 448–456.

[58] ANDREUX M, DU T J O, BEGUIER C, et al. Siloed Federated Learning for Multi–Centric Histopathology Datasets[EB/OL]. (2022−08−17)[2023−10−10]. https://arxiv.org/abs/2008.07424.

[59] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing Properties of Neural Networks[EB/OL]. (2013−12−21)[2023−10−10]. https://arxiv.org/abs/1312.6199.

[60] WANG Hongyi, SREENIVASAN K, RAJPUT S, et al. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning[EB/OL]. (2020−07−09)[2023−10−10]. https://arxiv.org/abs/2007.05084.

[61] PANG Qi, YUAN Yuanyuan, WANG Shuai. Attacking Vertical Collaborative Learning System Using Adversarial Dominating Inputs[EB/OL]. (2023−04−11)[2023−10−10]. https://arxiv.org/abs/2201.02775v1.

[62] SHI Lei, CHEN Zhen, SHI Yucheng, et al. Data Poisoning Attacks on Federated Learning by Using Adversarial Samples[EB/OL]. (2022−07−01)[2023−10−10]. https://ieeexplore.ieee.org/document/9853326.

[63] BIGGIO B, NELSON B, LASKOV P. Poisoning Attacks against Support Vector Machines[C]//ACM. Proceedings of the 29th International Coference on Machine Learning. New York: ACM, 2012: 1467–1474.

[64] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data Poisoning Attacks against Federated Learning Systems[EB/OL]. (2020−08−11)[2023−10−10]. https://arxiv.org/abs/2007.08432.

[65] ZHANG Jiale, CHEN Junjun, WU Di, et al. Poisoning Attack in Federated Learning Using Generative Adversarial Nets[EB/OL]. (2019−10−31)[2023−10−10]. https://ieeexplore.ieee.org/document/8887357.

[66] GOODFELLOW I J, POUGET–ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]//ACM. Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2014: 2672–2680.

[67] ZHANG Jiale, CHEN Bing, CHENG Xiang, et al. PoisonGAN: Generative Poisoning Attacks against Federated Learning in Edge Computing

Systems[J]. IEEE Internet of Things Journal, 2021, 8(5): 3310–3322.
[68] LI Tian, SAHU A K, ZAHEER M, et al. Federated Optimization in Heterogeneous Networks[EB/OL]. (2020−04−21)[2023−10−10]. https://arxiv.org/abs/1812.06127.
[69] BLANCHARD P, EL M E M, GUERRAOUI R, et al. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 118–128.
[70] YIN Dong, CHEN Yudong, RAMCHANDRAN K, et al. Byzantine–Robust Distributed Learning: Towards Optimal Statistical Rates[EB/OL]. (2021−02−25)[2023−10−10]. https://arxiv.org/abs/1803.01498.
[71] MHAMDI EL M E, GUERRAOUI R, ROUAULT S.The Hidden Vulnerability of Distributed Learning in Byzantium[EB/OL]. (2018−07−17)[2023−10−10]. https://arxiv.org/abs/1802.07927.
[72] BARUCH M, BARUCH G, GOLDBERG Y. A Little is Enough: Circumventing Defenses for Distributed Learning[EB/OL]. (2019−02−16)[2023−10−10]. https://arxiv.org/abs/1902.06156.
[73] XIE Cong, KOYEJO O, GUPTA I. Fall of Empires: Breaking Byzantine–Tolerant SGD by Inner Product Manipulation[EB/OL](2019−03−10)[2023−10−10]. https://arxiv.org/abs/1903.03936.
[74] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing Federated Learning through Anadversarial Lens[C]//IMLS. Proceedings of the 36th International Conference on Machine Learning. New York: ICML, 2019: 1012–1021.
[75] FANG Minghong, CAO Xiaoyu, JIA Jinyuan, et al. Local Model Poisoning Attacks to Byzantine–Robust Federated Learning[C]//USENIX. Proceedingsof the 29th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2020: 1623–1640.
[76] SHEJWALKAR V, HOUMANSADR A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning[EB/OL]. (2021−01−01)[2023−10−10]. https://www.researchgate.net/publication/350050756_Manipulating_the_Byzantine_Optimizing_Model_Poisoning_Attacks_and_Defenses_for_Federated_Learning.
[77] BAGDASARYAN E, VEIT A, HUA Yiqing, et al. How to Backdoor Federated Learning[EB/OL]. (2019−08−06)[2023−10−10]. https://arxiv.org/abs/1807.00459.
[78] BARUCH M, BARUCH G, GOLDBERG Y. A Little is Enough: Circumventing Defenses for Distributed Learning[EB/OL]. (2019−02−16)[2023−10−10]. https://arxiv.org/abs/1902.06156.
[79] SUN Ziteng, KAIROUZ P, SURESH A T, et al. Can You Really Backdoor Federated Learning?[EB/OL]. (2019−12−02)[2023−10−10]. https://arxiv.org/abs/1911.07963.
[80] XIE Chulin, HUANG Keli, CHEN Pinyu, et al. DBA: Distributed Backdoor Attacks against Federated Learning[EB/OL]. (2023−05−06)[2023−10−10]. https://openreview.net/forum?id=rkgyS0VFvr.
[81] GONG Xueluan, CHEN Yanjiao, HUANG Huayang, et al. Coordinated Backdoor Attacks against Federated Learning with Model−Dependent Triggers[J]. IEEE Network, 2022, 36: 84–90.
[82] CHEN Yudong, SU Lili, XU Jiaming. Distributed Statistical Machine Learning in Adversarial Settings: Byzantine Gradient Descent[EB/OL]. (2017−12−19)[2023−10−10]. https://dl.acm.org/doi/10.1145/3154503.
[83] XIE Cong , KOYEJO O, GUPTA I. Generalized Byzantine–Tolerant SGD[EB/OL]. (2018−05−23)[2023−10−10]. https://arxiv.org/abs/1802.10116.
[84] MUÑOZ–GONZÁLEZ L, CO K T, LUPU E C. Byzantine–Robust Federated Machine Learning through Adaptive Model Averaging[EB/OL]. (2019−09−11)[2023−10−10]. https://arxiv.org/abs/1909.05125.
[85] FUNG C, YOON C J M, BESCHASTNIKH I. The Limitations of Federated Learning in Sybil Settings[C]//USENIX. Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses. San Sebastian: USENIX Association, 2020: 301–316.
[86] YU Lei, WU Lingfei. Towards Byzantine–Resilient Federated Learning via Group–Wise Robust Aggregation[EB/OL]. (2020−11−26)[2023−10−10]. https://doi.org/10.1007/978−3−030−63076−8_6.
[87] JAIN A K. Data Clustering: 50 Years Beyond K–Means[J]. Pattern Recognition Letters, 2010, 31 (8): 651–666.
[88] SINGH A K, BLANCO–JUSTICIA A, DOMINGO–FERRER J, et al. Fair Detection of Poisoning Attacks in Federated Learning[C]//IEEE. Proceedings of the 32nd IEEE International Conference on Tools with Artificial Intelligence. New York: IEEE, 2020: 224–229.
[89] WANG Yuao, ZHU Tianqing, CHANG Wenhan,et al. Model Poisoning Defense on Federated Learning: A Validation Based Approach[C]//Springer. Proceedings of the 14th International Conference on Network and System Security. Berlin: Springer, 2020: 207–223.
[90] CHEN Zheyi, TIAN Pu, LIAO Weixian,et al. Zero Knowledge Clustering Based Adversarial Mitigationin Heterogeneous Federated Learning[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(2): 1070–1083.
[91] XIE Cong, KOYEJO S, GUPTA I. Zeno: Distributed Stochastic Gradient Descent with Suspicion–Based Fault–Tolerance[EB/OL]. (2019−05−18)[2023−10−10]. https://arxiv.org/abs/1805.10032.
[92] BAO Xianglin, SU Cheng, XIONG Yan,et al. FLChain: A Blockchain for Auditable Federated Learning with Trust and Incentive[C]//IEEE. Proceedings of the 5th International Conference on Big Data Computing and Communications. New York: IEEE, 2019: 151–159.
[93] LI Yuzheng, CHEN Chuan, LIU Nan, et al. A Blockchain–Based Decentralized Federated Learning Framework with Committee Consensus[J]. IEEE Network, 2021, 35(1): 234–241.
[94] PENG Zhe, XU Jianliang, CHU Xiaowen, et al. VFChain: Enabling Verifiable and Auditable Federated Learning via Blockchain Systems[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(1): 173–186.
[95] SHAYAN M, FUNG C, YOON C J M, et al. Biscotti: A Blockchain System for Private and Secure Federated Learning[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(7): 1513–1525.
[96] QU Youyang, POKHREL S R, GARG S, et al. A Blockchained Federated Learning Framework for Cognitive Computing in Industry 4.0 Networks[J]. IEEE Transactions on Industrial Informatics, 2021, 17(4): 2964–2973.
[97] LIU Yi, PENG Jialiang, KANG Jiawen, et al. A Secure Federated Learning Framework for 5G Networks[J]. IEEE Wireless Communications,

2020, 27(4): 24–31.
[98] BENTOV I, LEE C, MIZRAHI A, et al. Proof of Activity: Extending Bitcoin's Proof of Work via Proof of Stake[J]. ACM SIGMETRICS Performance Evaluation Review, 2014, 42(3): 34–37.
[99] MCKEEN F, ALEXANDROVICH I, BERENZON A, et al. Innovative Instructions and Software Model for Isolated Execution[EB/OL]. (2013−06−23)[2023−10−10]. https://doi.org/10.1145/2487726.2488368.
[100] CHEN Yu, LUO Fang, LI Tong, et al. A Training–Integrity Privacy–Preserving Federated Learning Scheme with Trusted Execution Environment[J]. Information Sciences, 2020, 522: 69–79.
[101] ZHANG Xiaoli, LI Fengting, ZHANG Zeyu, et al. Enabling Execution Assuranceof Federated Learning at Untrusted Participants[C]//IEEE.Proceedings of the 2020 IEEE Conference on Computer Communications. New York: IEEE, 2020: 1877–1886.
[102] ZHAO Yi, XU Ke, WANG Haiyang, et al. Stability–Based Analysis and Defense against Backdoor Attacks on Edge Computing Services[J]. IEEE Network, 2021, 35(1): 163–169.
[103] ZHANG Jiale, WU Di, LIU Chengyong,et al. Defending Poisoning Attacks in Federated Learning via Adversarial Training Method[C]//Springer. Proceedings of the 3rd International Conference on Frontiers in Cyber Security. Berlin: Springer, 2020: 83–94.
[104] IBITOYE O, SHAFIQ M O, MATRAWY A. DiPSeN: Differentially Private Self–Normalizing Neural Networks For Adversarial Robustness in Federated Learning[EB/OL]. (2021−01−08)[2023−10−10]. https://arxiv.org/abs/2101.03218v1.
[105] GOODFELLOW I J, SHLENS J, SZEGEDY C.Explaining and Harnessing Adversarial Examples[EB/OL]. (2015−03−20)[2023−10−10]. https://arxiv.org/abs/1412.6572.
[106] SHEJWALKAR V, HOUMANSADR A, KAIROUZ P, et al. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Federated Learning[EB/OL]. (2021−12−13)[2023−10−10]. https://arxiv.org/abs/2108.10241.

西安电子科技大学闫峥教授团队研究成果被国际信息安全会议 IEEE S&P 录用

第45届国际信息安全会议IEEE Symposium on Security and Privacy（IEEE S&P 2024）将于2024年5月在美国旧金山召开。西安电子科技大学网络与信息安全学院闫峥教授团队的最新研究成果“FlowMur: A Stealthy and Practical Audio Backdoor Attack with Limited Knowledge”被该会议全文收录，并将在大会作报告。

据悉，IEEE S&P又称Oakland，与ACM CCS、USENIX Security、NDSS并列称为安全领域的四大国际学术会议，其近十年的平均录用率约为13%，发表难度在四大会议里最高，被中国计算机学会（CCF）认定为A类会议。该会议收录的论文代表着相关研究领域的较高水平，在业界具有广泛而深远的影响。

该研究成果由西安电子科技大学的闫峥教授团队和普渡大学的Elisa Bertino教授合作完成，闫峥教授的博士生兰佳禾、博士生王杰、硕士生闫保辰为论文的前三作者，通讯作者为闫峥教授。

该论文聚焦语音识别安全，提出了仅需有限敌手知识、隐蔽且实用的语音后门攻击方法FlowMur。FlowMur通过构建辅助数据集和代理模型增强敌手知识。它充分考虑了语音的动态性以及环境噪声的影响，将触发器的生成形式化为一个优化问题，提升了攻击的实用性。此外，FlowMur采用了基于信噪比的自适应数据投毒手段，实现了攻击的隐蔽性。基于多个公开数据集和经典语音识别AI模型的实验结果表明，FlowMur在数字和物理场景下均展现出强大的攻击性能，造成AI模型的失灵，并能绕过当前最先进的防御手段。用户测试进一步证实了FlowMur的不易察觉性，因此其具有极强的危害性。

近年来，语音识别快速发展，已被广泛应用于智能语音助手、语音购物、声纹认证等场景，极大地便利了人们的日常生活，但其安全性仍是一个悬而未决的问题。该论文的发表在一定程度上推动了语音识别系统安全的研究，有助于未来构建更加安全的语音识别系统。同时，这种攻击也可被用于验证语音数据集和语音识别模型的所有权，进行审计和溯源。该论文向外界展示了西安电子科技大学在人工智能安全领域研究的领先成果，标志其在该领域的研究得到了国际同行的高度认可。（来源：西安电子科技大学，https://news.xidian.edu.cn/info/2106/226520.htm）