计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

联邦学习的隐私保护与安全防御研究综述

肖雄[1),2),3)] 唐卓[1),2)] 肖斌[3)] 李肯立[1),2)]

[1)](湖南大学信息科学与工程学院，长沙 410082)

[2)](国家超级计算长沙中心（湖南大学），长沙 410082)

[3)](香港理工大学电子计算学系，香港)

摘 要 联邦学习作为人工智能领域的新兴技术，它兼顾处理“数据孤岛”和隐私保护问题，将分散的数据方联合起来训练全局模型同时保持每一方的数据留在本地。联邦学习在很大程度上给需要将数据融合处理的数据敏感型应用带来了希望，但它仍然存在一些潜在的隐私泄露隐患和数据安全问题。为了进一步探究基于联邦学习的隐私保护和安全防御技术研究现状，本文对联邦学习的隐私和安全问题在现有最前沿的研究成果上进行了更清晰的分类，并对威胁隐私和安全的手段进行了威胁强度的划分。本文首先介绍了涉及联邦学习隐私和安全问题的威胁根源，并从多个方面罗列了其在联邦学习中的破坏手段及威胁性。其次，本文总结了关于联邦学习隐私和安全问题所面临的挑战。对于隐私保护而言，本文同时分析了包括单个恶意参与方或中央服务器的攻击和多方恶意合谋泄露隐私的场景，并探讨了相应的最先进保护技术。对于安全问题而言，本文着重分析了影响全局模型性能的多种恶意攻击手段，并系统性地阐述了先进的安全防御方案，以帮助规避构建安全的大规模分布式联邦学习计算环境中潜在的风险。同时与其他联邦学习相关综述论文相比，本文还介绍了联邦学习的多方恶意合谋问题，对比分析了现有的联邦安全聚合算法及安全开源框架，致力于为研究人员提供该领域更清晰的视野。最后，本文讨论了联邦学习技术面临的挑战和未来研究方向，以期望进一步推进联邦学习在人工智能场景下的安全应用。

关键词 联邦学习；数据隐私；数据安全；大规模分布式学习；人工智能

中图法分类号 TP393

A Survey on Privacy and Security Issues in Federated Learning

Xiao Xiong[1) 2) 3)] Tang Zhuo[1) 2)] Xiao Bin[3)] Li Ken-Li[1) 2)]

[1)](College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082)

[2)](National Supercomputing Center in Changsha (Hunan University), Changsha, 410082)

[3)](Department of Computing, The Hong Kong Polytechnic University, Hong Kong)

Abstract As an emerging technology of artificial intelligence, federated learning takes into account the issues of "isolated data islands" and data privacy protection. Federated learning can assist data fusion processing for data-sensitive applications by allowing distributed data participants to train a global model while keeping each participant's data locally. However, Federated learning encounters data privacy leak risks and various attacks. In order to explore the current research status on privacy protection and security attacks in federated learning, this paper makes a clear classification of the state-of-the-art methods. In this paper, we first introduce the threats to privacy and security in federated learning from many aspects. For privacy protection, we analyze the root causes of privacy threats from multiple scenarios, including single malicious participant attacks, central server attacks,

———————————————

本课题得到国家重点研发计划(2018YFB1701400)、湖南省自然科学基金青年项目(2021JJ40612)、国家自然科学基金(Grant Nos.61873090，L1824034，L1924056)资助. **肖 雄**，博士，是计算机学会(CCF)会员（80357G）,主要研究领域为云计算、分布式机器学习和联邦学习的隐私和安全.E-mail: xx@hnu.edu.cn. **唐 卓**（通信作者），博士，教授，博士生导师，是计算机学会(CCF)会员,主要研究领域为分布式计算系统、大数据并行处理、分布式机器学习和安全模型、以及该领域的资源调度和管理.E-mail: ztang@hnu.edu.cn. **肖 斌**（通信作者），博士，教授，博士生导师，是计算机学会(CCF)会员,主要研究领域为网络安全、区块链技术和人工智能安全.E-mail: b.xiao@polyu.edu.hk. **李 肯 立**，博士，教授，博士生导师，是计算机学会(CCF)理事，主要研究领域为并行计算、云计算和大数据计算.E-mail: lkl@hnu.edu.cn.

and multiple participants malicious collusion attacks to leak privacy. At the same time, we describe the specific attack process and the attack effect of these privacy stealing methods in detail. Moreover, we show the current defense methods as how to enhance privacy protection, including differential privacy, homomorphic encryption, secure multi-party computation, verification network frameworks and collaborative training schemes. These methods are widely used in privacy protection and have shown good effectiveness. The protection effect of the system needs to be balanced on some performance issues such as model accuracy and calculation efficiency. For security issues, we focus on analyzing a variety of malicious attack methods that affect the performance of the global federated learning model, including independent attacks where malicious participants use multiple methods to poison data or models, and malicious participants colluding to launch the attack. Meanwhile, we introduce the attack process and attack threat in detail. Subsequently, we systematically elaborate and summarize the advanced security defense solutions, which can better maintain the security of the model in solving attacks from single or multiple malicious adversaries, while also alleviating communication bottlenecks and helping the model to converge faster. Compared with the existing related state-of-the-art surveys, our work summarizes the recent progress on the multiple participants malicious collusion problem in federated learning, including multiple participants malicious parameter collusion and multiple participants malicious ownership collusion. The two methods both have strong attack destructiveness while maintaining high attack concealment. This is a novel research direction, and there is not much current research work. In addition, we also carefully analyze the existing federated learning secure aggregation algorithms and secure open-source frameworks. For federated learning secure aggregation algorithms, we list the vulnerabilities of traditional methods on privacy and security issues, and explain the advantages of the technology proposed by the existing reliable security aggregation method. For secure open-source frameworks, we analyze the work done by several mainstream security frameworks in privacy protection and security defense. We compare their applicable federated modes, federated scenarios, and defects that need to be improved, providing researchers with a clear vision for privacy and security design. Finally, we discuss the challenges and future research directions on the privacy protection and security defense of federated learning, which aims to improve the design of privacy protection and security defense methods in future research work. We expect to promote further that federated learning can be safely applied in artificial intelligence scenarios.

Key words federated learning; data privacy; data security; large-scale distributed learning; artificial intelligence

1 引言

在以传统的服务器或云为中心的模型训练模式中，通常所有参与计算的移动设备的数据都会被收集到传统服务器或上传到云中进行集中处理，以开发出一个性能更优、更具普适性的机器学习模型。然而，在这样的训练过程中，模型的精度和效率将严重依赖于中央服务器或云端的计算能力，以及训练的数据量。基于当前用户数据存在的数据量小、数据分散的特点，这让基于集中式训练模式的机器学习方法在计算时间和模型精度上存在很大的挑战。更为关键的是，用户数据的安全性和隐私性都暴露在被攻击窃取的风险之中。

当前，随着大数据、机器学习、人工智能等技术的大力发展，用户移动数据爆炸式的增长，人脸识别[1]、指纹识别[2]等个性化服务的普及，人工智能应用也越来越多地深入到人们的生产生活中。但是，在这些看似华丽的科技产品背后，却存在更多的包括用户的私人信息、医院的医疗隐私信息以及社交媒体的大量敏感数据等的暴露风险。已经提出的许多基于大规模数据的收集和处理技术帮助现有的计算平台极大地提升了数据处理性能，但也把用户隐私和安全保护问题置于一个更大的挑战之中[3]。以 2019 年发生的 Facebook 大规模用户信息泄露事件为例，该事件中超过 2.67 亿条 Facebook

用户的电话号码、姓名和用户 ID 等隐私信息能直接通过外部访问被非法下载。这也使得用户数据隐私和安全性问题得到世界范围内的广泛关注，越来越多的企业和学者在考虑兼顾处理大数据需求的同时也把保护用户的隐私和安全性问题当作当前计算系统着重需要考虑和解决的主要问题。

作为对用户隐私泄露和安全性问题的强势回应，世界各国都在积极出台相关法律法规以对用户的数据隐私和安全问题进行强力保护。在 2017 年，中国就颁布了《中华人民共和国网络安全法》，其在建立健全用户信息保密制度的同时也要求网络运营商不得泄露、篡改和损毁他人信息，此外任何个人和组织也不能非法获取他人信息。次年 5 月，欧盟也通过了《通用数据保护条例》（General Data Protection Regulation, GDPR），该条例的生效也为企业中存储的用户数据带来了更强的安全保障。企业在未经用户授权的情况下泄露或外传用户私人信息将面临巨额罚款或其他更严厉的制裁。由于这些法律法规的主导，包括 Google、腾讯、阿里巴巴等在内的众多海内外企业已纷纷向用户更新隐私政策，重新获取安全授权，旨在为所有用户提供更安全的软件服务。这些法规条例的建立与实施给当下产生的海量数据带来了安全性的保障，但也给需要进行数据交互的人工智能应用程序[4]引入了新的挑战，使得数据中潜在的巨大价值无法被充分挖掘和高效利用[5]。

在传统机器学习训练模型的过程中，用户数据往往会被统一收集并集中存储在中央服务器中，用于后续的训练和测试，以便开发出更全面的通用机器学习模型[6]。但随着数据体量的不断增大，中央服务器处理数据的能力也成为了影响模型训练效率和精度的重要性能瓶颈。面对越来越大的网络模型，为了达到更好的拟合效果，许多公司都需要给模型应用极大的数据量，而传统方法往往体现出的是巨大的时间开销和延迟，因此大规模分布式机器学习系统也应运而生。具有代表性的便是参数服务器[7]的提出，它将数据或模型分布式的存储在不同机器上进行执行并聚合最终结果，这样的计算方式极大地提升了系统的整体计算性能。当然，云计算[8]和雾计算[9]等技术的提出也都为分布式系统提供了更高效的解决方案。相比之下，边缘计算[10]将用户隐私信息存储在网络边缘设备上，不仅降低了数据在上传到云服务器的过程中潜藏的数据泄露的风险，还提升了数据的执行效率。

尽管分布式网络中设备的计算能力和存储能力在不断增强，但由于设备间存在因为数据隐私和安全性问题无法直接进行信息交互的限制，这让很多应用的数据集无法进行有效聚合。例如不同智能手机之间的历史消息记录和各地的银行存储的储蓄信息，它们几乎很难完成有效的数据共享和聚合，因此存在很严重的“数据孤岛”问题。在人工智能迅猛发展的当下，数据融合与共享的需求愈发强烈，为了同时兼顾处理数据隐私和“数据孤岛”带来的挑战，许多企业和学者都在不违反各地政府法律的前提约束下致力于构建模型进行数据利用和训练。最为典型的便是 Google 公司在 2017 年提出的安全解决方案：联邦学习[11]。在联邦学习实现过程中，多个参与方的数据无需上传到中央服务器上进行联合训练，而是在本地进行训练后将更新的本地参数上传至服务器中，进而由服务器执行聚合，迭代收敛后得到的全局模型会进一步被分发给所有参与方。联邦学习提供了一种去中心化的思路，将计算分散到各个本地端来保护数据隐私，这为一些数据敏感型的人工智能应用带来了极大的益处。同时，联邦学习也将分布式的小数据片段进行了有效的训练和聚合，在机器学习模型的安全性和隐私性上展现出了极大的优势。此外，联邦学习实现了较好的数据隔离并保证了模型质量，维护了多个参与方的对等地位，实现公平合作和独立成长。也是基于其强大的隐私保护特性，联邦学习已成为大规模机器学习和分布式优化场景下进行模型训练的主流选择。诸如联合大量智能手机进行虚拟键盘热词预测，联合建筑、行人和路况信息构建模型的自动驾驶汽车等应用都已证明联邦学习已经表现出比传统机器学习方法更为出色的性能。

虽然联邦学习已经在机器学习领域获得了很大的关注和利用，但是，因为联邦学习在执行过程中需要每一个聚合的参与方进行本地训练后上传本地模型，每一次更新的全局模型参数也都会与所有参与方共享，而参数中可能包含参与方的贡献信息和其他敏感数据，这让联邦学习在构建模型的过程中存在很严重的数据隐私泄露风险和安全隐患。尽管联邦学习缓解了机器学习领域中数据敏感型应用带来的隐私挑战，但是在执行用户本地参数上传、服务器进行模型聚合以及全局模型共享的过程中仍存在新的安全风险。参与方在每次上传的本地

参数中的隐藏特征很有可能会被不可信的服务器或其它恶意参与方窃取，服务器和用户执行通信的过程中，恶意对手的后门入侵和数据追踪也会导致敏感信息被恶意访问，另外，恶意参与方也可以上传中毒的本地模型进行聚合更新以操纵机器学习模型的输出，并最终影响全局模型的性能。为了保护数据隐私以及抵御这些恶意攻击，已经有一些研究[12-16]聚焦在联邦学习的执行过程中来设计可靠的数据加密算法和安全的防御技术。因此，本文对涉及联邦学习的数据隐私保护方法和安全防御手段进行了细致的调研，旨在展现联邦学习模型在这些策略的保护下表现出的更好的稳定性。

与现有的其他已经发表的关于联邦学习隐私和安全相关的综述文献[17-19]相比，1）本文更具体地对联邦学习涉及的隐私威胁和安全问题进行了详细的定义，同时依照其本身的攻击行为和造成的不良影响进行了新的分类。通常，在联邦学习模型构建的过程中，因参与方保有隐私数据在本地训练而较好地维护数据隐私性，但是由于恶意参与方旨在获取其他诚实参与方的私密数据，他们会在模型训练或执行聚合的任意过程发动攻击以提取敏感数据，甚至精确到用户记录信息的某一条具体数据，这无疑对参与方隐私数据的保护提出了更大的挑战。而联邦学习的安全威胁更多在于恶意参与方在模型构建的过程中发动诸如数据投毒或模型投毒等恶意攻击以严重破坏全局模型的性能。由于联邦学习系统在模型构建过程对所有参与方都是公开且透明的，同时参与方匿名执行参数传递，这使得一些不诚实的参与方有机会攻击他人，实施对本地训练数据的窃取和对模型的篡改及其他恶意操作而不会被溯源，最终引导模型往非良性的方向发展。2）本文更系统性地整理了这一领域最前沿的研究成果，并进行了整体性的总结和分类。相较于文献[17-19]对涉及联邦学习隐私和安全问题的不全面描述，本文还重点描述了联邦学习的多方恶意合谋窃取隐私和攻击全局模型的问题及其防御方案，相较于上述文章描述的传统的单个恶意参与方的独立攻击，本文总结了多个恶意参与方的合谋攻击所带来的攻击破坏性和发动攻击的隐蔽性，以及在面对基于单个恶意参与方的异常检测时所展现出来的能够较好逃避检测的能力。3）本文还对这些威胁隐私和安全的攻击手段进行了威胁强度的划分并罗列了对应的防御方案。4）为了更全面地展现出联邦学习存在的安全风险和防御策略，本文还细致地分析了现有的联邦安全聚合算法及安全开源框架，致力于为研究人员提供该领域更清晰的研究视野。我们将本文与其它联邦学习综述文献的对比展示在如下的表 1 中。

本文更加深入地从发生隐私泄露和安全威胁的根源入手，从多个角度进行全面剖析。文章的其余部分组织如下：第 2 节对联邦学习的背景和基础知识进行了概述。第 3 节描述了关于联邦学习的隐私威胁根源。第 4 节提供了有关联邦学习隐私风险的解决方案。第 5 节介绍了涉及联邦学习的安全威胁根源。第 6 节给出了有关联邦学习安全问题的解决方案。第 7 节分析了几种现有的主流联邦学习开源框架，并对比了它们在隐私和安全层面的工作。第 8 节总结了联邦学习的隐私保护和安全防御问题，并讨论了潜在的研究方向。最后，本文在第 9 节对全文进行了总结。

2 联邦学习概述

本节主要对联邦学习的基本定义、执行流程和不同模式进行了总体的描述，同时与传统的集中式机器学习相比较，总结了联邦学习在大规模计算和分布式优化上的优势。

2.1 传统机器学习

在传统的机器学习训练过程中，所有参与训练的数据会被集中收集到一台中央服务器中进行学习，以便开发出最终的全面而准确的机器学习模型[6]。通常，它的计算方式可以表述为从已有的样本当中学习到输入和输出之间对应的映射关系，进而根据学习到的映射关系对所有新输入的样本预测其可能存在的输出结果。尽管针对这些待学习的参数求解的是一个最优化模型构建的问题，也已经提出了许多优秀的方法帮助机器学习算法和应用获得更好的性能效果，但执行参数学习的所有过程全部都堆积在一台中央服务器上进行实现。在很大程度上来说，模型的精度依赖于在中央服务器上进行训练的数据，模型的效率取决于中央服务器的计算能力以及计算方式。

表 1 本文与其它联邦学习综述文献的对比

综述文章	隐私问题定义/分类	安全问题定义/分类	文献前沿性/充分性	攻击分类/强度划分	隐私/安全的多方合谋	隐私/安全防御技术分类	聚合算法/安全框架
本文	清晰的定义及详细的攻击分类	清晰的定义及详细的攻击分类	整理了前沿且充分的文献	清晰分类并做了威胁强度划分	总结了独立和多方合谋攻击	对隐私/安全防御技术清晰的分类	清晰的总结了聚合算法/安全框架
Zhou, et al.[17]	简单的定义，未做攻击分类	简单的定义及简单的攻击分类	整理了较前沿但非充分的文献	清晰分类但未做威胁强度划分	仅有独立攻击，未介绍合谋攻击	对隐私/安全防御技术较清晰的分类	未涉及聚合算法/安全框架
Chen, et al.[18]	简单的定义，未做攻击分类	简单的定义及简单的攻击分类	整理了前沿但非充分的文献	清晰分类但未做威胁强度划分	仅有独立攻击，未介绍合谋攻击	对隐私/安全防御技术较清晰的分类	未涉及聚合算法/安全框架
Yang, et al.[19]	清晰的定义及详细的攻击分类	未做安全定义及攻击分类	整理了前沿且较充分的文献	清晰分类但未做威胁强度划分	仅有独立攻击，未介绍合谋攻击	仅对隐私/防御技术清晰的分类	未涉及聚合算法/安全框架

尤其在数据体量和网络模型越来越大的当下，许多公司和机构为了让自己精心训练的模型获得更好的拟合效果，往往会投入极大的数据量进行训练，而这些数据在传统方式中很难高效运行，同时会产生大量的开销和延迟。为了解决训练过程中存在的这一性能瓶颈，越来越多的例如并行学习、分布式学习技术[7] [20]被提出，这些技术一般将训练的过程分发到多个子处理节点上来执行并行处理。分布式学习的实际执行过程通常是通过中央处理器分别分发数据的不同子集或是模型的各个部分在每个处理节点上分散地执行并行计算，然后等待每一个独立的子节点计算完成，并接收这些计算节点返回的参数，最后进行聚合更新以生成最终的结果。尽管分布式优化技术帮助机器学习算法获得了较好的计算性能和训练效率，但它始终没有脱离中心式训练的模式，同时在训练过程中会产生较大的通信开销和同步等待开销，如果发生子计算节点故障或丢失的情况，其他所有计算节点的后续工作也会中断。此外，由于数据的集中存储和统一分发，数据的隐私泄露和安全威胁问题也越发突出。这些挑战都让当前传统的机器学习方法不得不同时兼顾考虑大规模计算、分布式优化以及隐私和数据安全的保护。

2.2 联邦学习

为了帮助克服这些挑战，Google 在 2017 年提出了一种安全的解决方案：联邦学习[11]。该方案主要通过分布式的多个设备进行本地训练后，服务器接收上传的本地参数进行聚合，在多次迭代直至收敛后生成最终稳定的全局模型。联邦学习引入了基于随机梯度下降（Stochastic Gradient Descent，SGD）优化技术的联邦平均聚合算法（Federated Averaging，FedAvg），该算法就是利用平均思想在聚合过程中计算中央服务器上的所有本地参数加权平均来生成全局模型。此外，所有用户的私有数据都保留在本地进行训练，不用上传到中央服务器中造成不必要的通信开销和潜在的隐私泄露和安全风险。

通常，联邦学习在分布式网络中实现，具体定义如下：假设存在 N 个参与方分别表示为{C_1，... C_n}，它们分别保存各自的数据{Dc_1，... Dc_n}，并且期待整合这些数据构建一个完整的机器学习模型。常规方法是先将所有数据进行统一合并，生成一个全局数据集 $D = Dc_1 \cup Dc_2 \cup ... \cup Dc_n$，然后进行训练得到模型 M_{SUM}。而联邦学习是通过所有参与方在本地训练后进行聚合，协同训练出来一个全局模型 M_{FED}。在执行训练的过程中，任何参与方 C_i 都不会向其他方公布自己私有的本地数据 Dc_i。与此同时，训练出来的全局模型 M_{FED} 的精度 V_{FED} 应该非常接近于常规方法训练的模型 M_{SUM} 的精度 V_{SUM}。具体表示如下：

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

其中，δ 为非负实数。由上式可见，联邦学习算法存在 δ 精度损失。

联邦学习的执行流程如图 1 所示。其计算环境包括中央服务器（Central Server），通常中央服务器可以由物理服务器或是云服务器组成。同时还包括多个分散的参与方（Client）和所属的本地数据

（Data），参与方和本地数据分别用缩写 C_i 和 Dc_i 表示，意为编号为 i 的参与方及其数据。

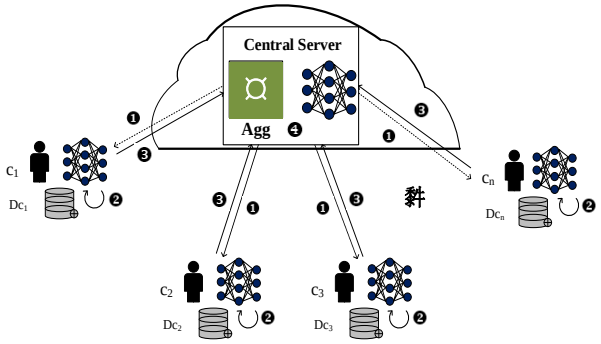

图 1 联邦学习执行流程

同时，联邦学习的执行过程也被视为一次次迭代的过程，每一次的迭代计算都在改进全局模型，进而分享给每一个参与方。如图 1 所示，其实现步骤可以概括如下：

1）初始模型分发：中央服务器将初始模型分发给计算环境中的所有参与方。

2）本地训练：所有参与方获取了中央服务器共享的全局模型，并利用本地存储的私有数据进行本地模型训练。

3）本地模型上传：所有由参与方在本地完成训练得到的本地模型会被上传到中央服务器中，以等待执行下一步聚合操作。

4）聚合：每一位参与方的本地模型在上传到中央服务器后，其更新的梯度参数会被中央服务器进行聚合并执行联邦平均，以构建全局模型。

在这个过程中，每个参与方共享的都是完整且同样的模型，它们在本地训练阶段独立进行，互相之间无沟通。直到全局迭代收敛，中央服务器会把最终构建好的全局模型分发给每一位参与方。

目前，联邦学习已经在很多不同的领域展现着其优异的性能[21]。也有学者根据训练数据特征的不同将联邦学习分为三种模式[22]，分别是横向联邦学习、纵向联邦学习以及联邦迁移学习。

1）横向联邦学习

横向联邦学习所做的事情引导的最终结果是参与方的联合，其适用于参与方的数据特征重叠较多，而参与方本身重叠较少的情况，如图 2（a）所示。例如在两个不同地区的医院，他们的病人群体交集很小，但它们的业务却十分相似。Google 最早提出的联邦学习版本[23]就属于横向联邦学习这一类。在面向大量分散的智能手机进行下一个单词预测的建模过程中，联邦学习模型通过不断更新每一个子集独立训练生成的参数而不断进行自我增强。

通常，横向联邦学习系统暴露的安全风险往往发生在好奇的服务器或是恶意的参与方之中。好奇的服务器会损害所有参与方数据的隐私，而恶意参与方则会发动攻击来破坏全局模型，实现自己的恶意目的。文献[24]就基于该恶意环境提出了一种安全的聚合手段，其旨在以更安全的方式计算来自大量移动设备的数据向量，并在该联邦学习系统的更新过程中保证数据隐私和全局执行安全性。此外，也有更多更强的攻击手段[25-26]被提出用以影响横向联邦学习的数据安全，并从根本上逃避一些常规的防御策略的检测。这也将横向联邦学习的隐私和安全问题推向了一个更高的关注点。

2）纵向联邦学习

纵向联邦学习所做的工作的本质是参与方数据特征的联合，其适用于参与方本身重叠较多，而数据特征重叠较少的情况，如图 2（b）所示。比如针对在同一个地区的医院和银行而言，它们的数据样本基本都是本地区的居民，但执行的业务却大不相同。Hardy[27]等人就根据相关公共实体的私有记录所涉及到的不同功能集合设计了纵向联邦学习方案，通过实体解析和同态加密的端到端策略进行联合逻辑回归，并将该解决方案扩展到了数百万个实体。

因为纵向联邦学习的样本所属方所产生的交集很大，但是较少的特征相似性让它们的功能空间产生很大的出入。这使得纵向联邦学习可以扩展到很多的关联领域，包括统计分析和分类等。纵向联邦学习的隐私和安全问题也备受关注，在纵向联邦学习的防御方案中，通常计算环境存在一些恶意的参与方，他们往往会破坏本地数据或串谋其他参与方以窃取数据，Nock[28]等人就提出了实体解析方案训练了一种安全性强的逻辑回归模型，不仅在学习过程中提升了整体性能，也在数据加密过程中降低了开销。为了更好地保证数据隐私和计算安全，诸如同态加密和安全多方计算等防御手段都被引入来增强系统的安全防护能力。

3）联邦迁移学习

联邦迁移学习主要是针对参与方的数据集所属样本的特征和参与方本身都重叠较少的情况，以克服数据规模小和样本标签少的情形，如图 2（c）所

示。比如处在不同地区的银行和商场，两者的用户和执行业务都不相同。它的具体实现与纵向联邦学习相似，但在参与方之间交换加密中间结果时做了一些细节上的优化和改变。Yang[29]等人通过联邦迁移学习来训练个性化的分布式模型用以实现安全的图像分析，不仅轻松地完成了大规模的安全识别，同时维持了较高的鲁棒性。

联邦迁移学习系统的隐私和安全风险与纵向联邦学习类似，计算环境潜在的恶意行为可以在迁移学习系统中发生，尤其是在医疗保健[30]和无人驾驶[31]领域更为突出。尽管作为联邦学习的重要扩展和分支，但现有的一些联邦学习算法在处理联邦迁移学习问题时会有较大的瓶颈，这也需要更多的学者深入研究和探索。

2.3 联邦学习优势

联邦学习作为兼顾隐私保护和去中心化协作的机器学习技术，与传统机器学习和分布式机器学习技术相比，其训练数据具有完全的本地性、自主性和隐私保障性，同时在模型构建的过程中具有统计异质性和系统异构性的特点，它不局限于数据设置为独立同分布的假设，同时能较好地在异构系统中完成容错。它的优势可以概括如下：

1）数据独立性。所有参与方的数据均保留在本地执行而无需上传至中央服务器，故除数据所有者外，其他参与方无法访问他人数据，这也给了数据所有者很大的空间进行本地模型的独立训练。

2）用户对等性。所有涉及模型构建的参与方在执行过程中拥有平等地位，它们执行相同迭代并上传本地模型，接收来自中央服务器聚合的相同全局模型，实现框架内的平等合作。

3）模型稳定性。联邦学习技术在兼顾分布式计算、安全聚合的同时也维持了与集中式学习相当的模型性能，全局模型在多方加密聚合的条件下实现安全交互同时保证模型不断成长。

4）统计异质性。联邦学习系统中执行的数据往往不局限于独立同分布（Independently and Identically Distributed，I.I.D.）的数据假设，同时它还聚合包含众多不同模式的数据，诸如包括中文和英文在内的文本数据进行语言建模。

5）系统异构性。联邦学习在融合数百万台移动设备进行模型构建时，往往会存在由于移动设备的不同品牌导致的 CPU 或内存异质问题、不同的网络连接问题（5G，WiFi）、不同设备的不同能耗和通信问题，系统都能完成较好的容错。

6）数据隐私性。不同于被数据收集器集中存储和处理，联邦学习分布式地将所有参与方的数据分散在本地并进行联合训练，所有参与方未经授权不得访问其他参与方的本地数据及其他敏感信息，这也很好地保障了用户的隐私和数据安全。

尽管联邦学习本身提供了多层隐私和安全保障，但在实际运行过程中仍然存在被恶意攻击者窃取隐私，执行恶意攻击或是遭受其他的安全威胁。

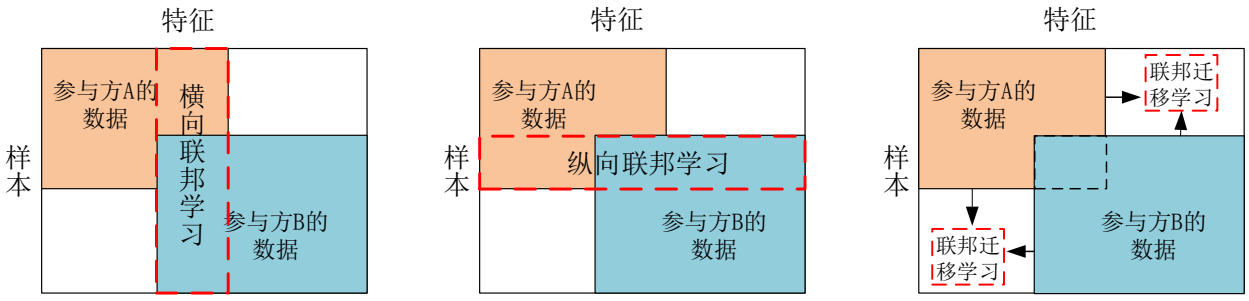

图 2 联邦学习分类

3 隐私威胁根源

联邦学习的设计和实现为全局模型构建期间的参与方数据隐私提供了更好的安全性。对于横向联邦学习来说，为了构建全面的全局模型，所有参与方都需要上传本地模型的梯度参数至服务器进行聚合，而这些参数本质是本地数据训练的映射，

包含了重要的私密信息。与传统的集中式机器学习不同的是，所有的数据和训练过程都被放置在了边缘，而不是像往常一样被服务器集中收集处理，这种方式会让参与方失去对数据的控制。尽管联邦学习降低了不可靠的服务器直接泄露数据隐私的风险，但也将分布式的数据隐私保护问题提升到了一个新的高度。由于需要对大量的参与方进行聚合，分散在各地的参与方上传的参数存在隐私和贡献值泄露的风险。另外，如果不诚实的参与方进行恶意隐私提取或不诚实的中央服务器端进行恶意敏感数据的泄露，这对全局模型的构建无疑会产生巨大的影响。诸如此类的关于横向联邦学习的隐私窃取方法已经从多个层面被大量地设计出来了，而涉及纵向联邦学习的隐私窃取手段是近段时间的研究热点。在纵向联邦学习中，参与训练的组织存在大量重叠但它们的数据特征不同，同时只有一个组织拥有标签。在联合训练期间，恶意对手通过梯度交换窃取其他组织样本的敏感信息，此外，由于预测的输出结果会包含其他私人数据的私密信息，这会让参与方的特征值暴露在被不诚实的组织推断窃取的风险之中。与纵向联邦学习的隐私威胁类似，联邦迁移学习在克服数据小、标签少的训练过程中，通常会进行参与方之间的梯度传递。在数据的反向传播过程中极易将隐私信息暴露出来，同时，不诚实的参与方也会在传输汇总的任何阶段通过推断提取到对方的敏感数据，这在很大程度上会将参与者的私人信息暴露在巨大的风险之中。故本文对联邦学习中包含三种模式下的隐私泄露发生的主要威胁根源，也即联邦系统存在的弱点和漏洞进行了以下总结，同时本文也将同一类型的隐私威胁手段放在同一小节中一并描述。下述隐私威胁手段主要包括恶意参与方之间进行未授权隐私访问和提取、恶意中央服务器端进行隐私泄露和攻击、以及恶意多方合谋窃取隐私，多方合谋涉及多个恶意参与方或是恶意参与方与恶意中央服务器之间的非法串通。更好的罗列以便联邦学习的开发者和使用者在之后的实现过程中规避漏洞，巩固防御。

3.1 恶意参与方获取隐私

联邦学习旨在通过所有参与方进行本地训练后聚合参数进行模型构建，而不需要集中他们的数据以维护多方隐私安全。这也在隐私保护问题上体现出了比集中式机器学习更高的优越性。但根据最近的一些研究显示，在联邦学习的实现过程中，恶意对手会在没有其他参与方任何背景信息的黑盒环境下仅通过他们的共享参数的贡献和部分数据的细化分析来发动恶意攻击以窃取隐私。如图 3 所示，展示的是恶意参与方端的威胁模型。在这个场景下，服务器是诚实的，其不会泄露其他参与方的参数信息，也不会被恶意对手妥协。恶意对手不会影响其他参与方的积极贡献，也不会破坏全局网络模型和算法，他们主要通过隐私推断、提取重构等攻击方式来窃取隐私。本文将主要手段列举如下：

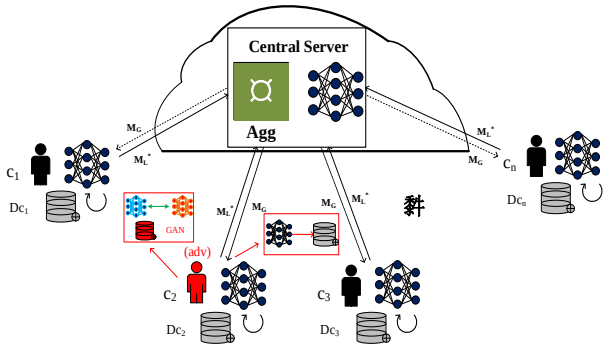

图 3 恶意参与方端威胁模型

3.1.1 隐私推断攻击

通常，隐私推断攻击通过使用推断提取技术对其他参与方的数据进行窃取。Zhang[32]等人就通过利用在黑盒环境下的隐私推断攻击对联邦学习的诚实参与方发动攻击，以获取训练数据中的敏感记录信息。同时，利用了生成对抗网络（Generative Adversarial Networks，GAN）来丰富攻击数据，在实际的攻击场景中完成了 98%的攻击准确率。Luo[33]等人研究纵向联邦学习模型预测阶段中存在的隐私泄露问题，还在不依赖任何背景信息的条件下设计了一种基于对手预测输出的通用攻击方法以进行特征值推断，并展现出了较高的攻击有效性。而这样的隐私推断窃取技术同样适用于联邦迁移学习。当参与者中存在恶意对手时，执行多方协作的联邦学习存在很大的成员推断窃取隐私的风险。

3.1.2 提取重构攻击

在参与方执行参数上传更新时，不诚实的参与方会进行参数的恶意提取并重构其他参与方的数据。Tramer[34]等人就针对当前包括逻辑回归、神经网络和决策树等流行的多类模型，进行无参数和无训练数据的黑盒访问，以实现高逼真度的目标机器学习模型提取，进而分析参数并对他人训练数据进行成功重构。Wu[35]等人也在纵向联邦学习环境下

提出了一种贪婪攻击方法，一些存在于联邦系统的攻击者利用公开的辅助信息发动攻击以重构诚实参与方的先验数据分布，从而实现隐私参数的成功提取。

3.1.3　窃取反演攻击

恶意参与方尝试通过访问训练好的模型进行训练参数和其他统计信息的未授权窃取，进而反演分析以获取参与方的隐私信息。Hayes[36]等人就通过窃取训练好的模型进行数据集基础分布的估计，然后根据该分布生成替代模型。这种方法的攻击面很广，同时对参数的敏感性很高，会导致被攻击的对象模型的性能显著下降。

3.1.4　参与方 GAN 攻击

GAN（Generative Adversarial Networks）是生成对抗网络，由两个模块组成。第一个模块 G 负责利用接收到的参数生成替代样本，另外一个模块 D 负责对生成的替代样本进行判断。Hitaj[25]等人就提出了一个由恶意参与方发起的 GAN 攻击，并根据模型学习过程中的实时性交互来获取更多与目标相关的敏感数据，进而进行生成对抗网络的构建来影响全局性能。基于 GAN 的攻击框架[37]也可以融合推理技术在获取全局模型的基础上推断其他参与方的私有信息，甚至精确到用户的具体某一条数据，并具有较高的攻击准确性。

3.2　恶意中央服务器泄露隐私

在模型的训练过程中，中央服务器负责分发初始模型、聚合所有参与方上传的参数并共享最终生成的全局模型，其理应保证诚实和安全。但往往恶意利用者会针对一些潜在的漏洞进行恶意攻击。如图 4 所示，本文展示了恶意中央服务器端的威胁模型。在这个场景中，所有参与方是诚实的，他们操纵本地数据积极参与训练和贡献，同时不会泄露自己的敏感数据和参数信息，也不会被恶意对手妥协。恶意对手不会影响所有参与方的积极贡献，也不会破坏全局网络模型和聚合算法，他们主要通过如下攻击方式窃取隐私。

3.2.1　服务器泄露隐私

协作式联邦学习技术需要聚合分散的参与方的本地模型，通过不断更新参数进行全局模型的构建。而聚合过程中来自参与方的本地贡献和敏感数据会被恶意的中央服务器泄露。Melis[38]等人就证明了在每一次迭代过程的更新阶段，不诚实的中央服务器会泄露相关参与方的训练参数。他们利用这种漏洞开发一套主动和被动兼顾的推断攻击策略来重构其他参与方的隐私数据。

3.2.2　服务器 GAN 攻击

在进行了错误标记的训练样本中引入基于服务器 GAN 攻击可以进行全局模型的破坏，同时推断参与方的样本信息。Wang[39]等人就首次提出了利用恶意服务器的攻击来推断联邦学习中的用户隐私，他们构建了一个融合 GAN 和多任务鉴别器的合并框架，旨在无形中对参与方身份和私有数据进行恢复。该策略不会干扰模型的正常训练同时隐蔽地发动恶意攻击来检索敏感信息，并完成了较好的攻击有效性。

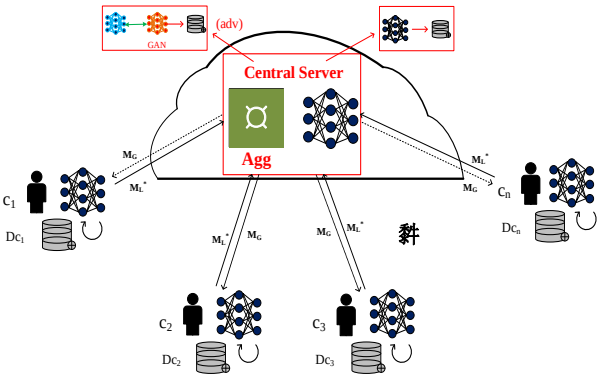

图 4　恶意中央服务器端威胁模型

3.3　恶意多方合谋获取隐私

恶意多方合谋攻击往往会展现出更强的攻击效果。如图 5 所示，本文展示了恶意多方合谋的威胁模型。在两种不同合谋模式下，诚实方都是积极参与训练和贡献，他们不会被恶意方所妥协。而恶意对手不会破坏全局体系结构和聚合算法，他们通过如下方式窃取隐私信息。

3.3.1　多参与方恶意合谋

在联邦学习计算环境中，当不可信的参与方不止一个时，全局模型的构建和诚实参与方的数据隐私保护会越发具有挑战性。这些恶意对手根据中央服务器共享的全局模型实施串谋攻击以推断其他参与方的隐私信息，并在全局模型中窃取诚实参与方的贡献值，进而达到提取敏感数据的目的。Yang[40]等人就指出多个恶意客户端合谋可以提取全局模型中的敏感信息，攻击者在未经授权的情况下窃取其他参与方的参数和贡献信息，同时利用这些隐私信息进行进一步的恶意篡改来创建中毒模型，并将中毒模型发送到中央服务器来防止全局模型收敛到理想状态。这种合谋攻击不仅窃取隐私，

还影响全局模型精度，具有较高的破坏性。

3.3.2 参与方和中央服务器的恶意合谋

不可信的参与方与不诚实的中央服务器在执行合谋攻击时，其他诚实参与方的私有信息以及贡献值在上传之后存在极大的可能性被窃取。恶意参与方利用中央服务器负责聚合所有本地模型的权限，对其他参与方的私有数据进行推断重构。如Lim[41]等人在文章中描述，不诚实的服务器会和好奇的参与者进行串通，将参数信息和其他隐私数据泄露给这些恶意用户，同时故意干扰原始参数的聚合，进而影响全局模型的构建。

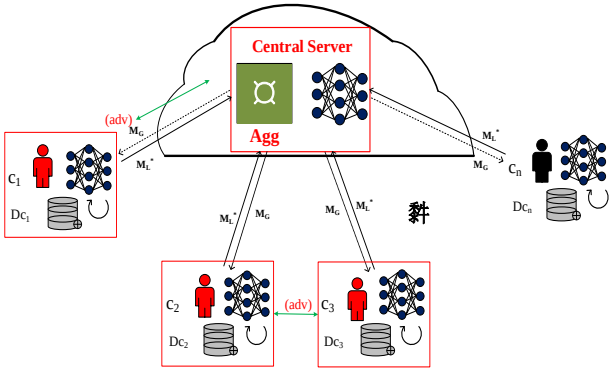

图 5 恶意多方合谋威胁模型

3.4 小结

首先，以上所提及的这些窃取隐私的攻击方法对用户的隐私数据和模型的敏感参数带来了非常大的影响。对于恶意参与方获取隐私的攻击手段而言，在全局联邦计算环境中，恶意参与方只需要在本地正常执行训练，和其他诚实参与方一样上传本地参数到服务器进行聚合或者在参与方之间进行梯度传递和汇总。其中，这些恶意对手会执行潜在的攻击进行推断或重构，发动的恶意行为最终会让诚实用户的私密数据悄无声息地被窃取。而恶意中央服务器泄露隐私的手段会相对来的更直接一些。服务器不仅控制着聚合流程，同时也执行模型共享操作，诚实参与方的敏感数据会在此过程被窃取和推断，这样的攻击往往能达到很好的效果。恶意多方合谋获取隐私的攻击手段不仅会发生在多恶意参与方中，也存在于不可信的参与方与不诚实的中央服务器之间，它们会比前两种模式的攻击效果更突出，且更具有多样性。其次，两种基于 GAN 的攻击方式在系统的执行过程中会取得很好的攻击效果，因为它们很难从本地输入样本与生成的待执行样本中被区分。在发动攻击后，恶意对手利用中毒的模型执行聚合来攻击指定用户并窃取更多的私密信息，甚至可以精确到具体的某一条数据。这样的攻击手段会比同类的其他方式更具准确性和高效性。更为典型的应用场景包括融合多个银行数据建立的金融系统，在多个银行进行信息交互时，需要列出查询的恶意多方借贷结果而不暴露诚实用户的私人信息[21]。还有涉及基因、私人疾病诊断的智能医疗系统，联邦学习技术可以较好地联合这些小数据进行训练并进一步提升模型预测效果，而不会公布这些隐私病例信息[22]。但这些数据都需要进行较强的隐私保护，故本文在下一节中进一步阐述隐私问题解决方案。

4 隐私问题解决方案

联邦学习在基于隐私保护的条件下构建模型训练体系，但就目前实现的联邦学习聚合算法而言，其计算过程针对隐私保护上还存在诸多漏洞。当然，当前越来越多的学者也已经聚焦到这一问题，并提出了许多较为可靠和稳定的解决方案。

4.1 差分隐私

在联合多参与方构建联邦学习共享模型的执行过程中，恶意参与方会对全局模型进行各式的攻击以获取其他参与方的敏感隐私信息。为了防止诚实方的隐私在模型共享时受到威胁，通常会采用差分隐私技术进行保护。差分隐私[42]最早被 DWork 等人提出，它融合了可靠的信息背景和严格的理论基础，以及较为简单的算法逻辑。假设给定的两个任意相邻数据集 D 和 D'，存在一个随机算法 A，都有

$$\Pr\{A(D)=O\} \le e^{\varepsilon} \cdot \Pr\{A(D')=O\} \quad (2)$$

也即，如果该算法作用在任何的相邻数据集上会得到一个在概率上相差无几的特定输出 O，则该算法能达到差分隐私的效果。换言之，通过观察输出结果很难察觉出数据集上微小的变化，从而达到了保护隐私的目的。如今这一思想也已经越来越广泛地被融合于各式差分隐私保护方法的变体中[43-45]。

此外，k 匿名和其他涉及为敏感参数添加噪声等多样化的方法也陆续被提出，Truex[46]等人就利用归纳方法在模型执行过程中对一些暴露的敏感数据进行掩饰，让其他恶意方无法进行参数窃取从而保护了数据隐私，但这种方法在一定程度上降低

了全局模型的精度。Geyer[47]等人为了解决分布式模型中用户端的训练参数和贡献被分析窃取的问题，提出了一种基于用户端的差分隐私联合优化算法，以在多用户参与训练时隐藏用户的贡献和数据集相关信息，进而以最小的成本在模型的性能和隐私开销上做出权衡。Agarwal[48]等人通过添加二项式噪声来掩盖敏感数据进行隐私信息的保护，但往往这种方法会对梯度进行微调，同时添加噪声也会影响模型的精度。这意味着在每次迭代期间应用差分隐私也要平衡好这种扰动对模型带来的影响。针对恶意参与方获取敏感数据的 GAN 攻击，Xin[49]等人就提出了一种差分隐私 GAN 模型，在严格的隐私权证明条件下，通过在模型学习过程中添加设计的噪声来实现差分隐私，这种方法不仅在训练过程中抵御了恶意成员的推理攻击同时获得了高质量的全局模型。Ghazi[50]等人提出了差分隐私与隐形斗篷算法相结合的安全聚合协议，保证了每一个参与方的敏感数据和训练贡献值不会在训练过程中被泄露，从而大大提升了联邦学习执行过程中的隐私安全保障。但这种方式也会让联邦服务器对参与方上传的参数产生误差评估，并可能在进一步聚合过程中损害全局性能。

与上述基于横向联邦学习的差分隐私保护技术不同的是，Feng[51]等人提出了一种基于多参与方多类的纵向联邦学习隐私保护框架 MMVFL，该框架允许以隐私保护的方式将参与方的本地标签信息有效地共享给其他纵向联邦学习参与方并匹配现有方法的多类分类性能，同时保证在执行多轮通信下的原始数据不会被推导出来。Liu[52]等人也提出了一种非对称纵向联邦学习的虚拟差分隐私保护方法，并以一个联邦逻辑回归算法作为实例来展示这种方法在样本 ID 对齐阶段的隐私保护效果。Li[53]等人也提出了一种融合差分保护技术的联邦迁移学习隐私保护方案，他们通过对每一个样本的图像数据进行个性化保护以在后续执行梯度参数聚合时不会泄露个人信息。

还有一些其他的相关隐私定义方法[54]被提出运用在大规模分布式计算系统和联邦学习环境中用以增加敏感数据的保护性。同时，大量的学者也聚焦在设计基于联邦学习的差分隐私保护框架[55-58]，在维护模型准确性的同时还为数据隐私泄露问题提供更强大的安全保障。尽管陆续有学者在关注如何维护数据隐私，同时不降低模型准确性的问题，但是现有的绝大多数方法都是在模型的性能损失和差分隐私的保护效果上做权衡。

4.2 同态加密

作为不需要直接使用明文的技术手段，同态加密通过对敏感数据进行加密传递，诚实方解密获取结果的方式来针对恶意中央服务器泄露隐私问题进行保护。同态加密作为一种常用且高效的加密手段，其包括 ElGamal 乘法和 Paillier 加法两种典型的运算方式。前者是一种随机的乘加密方法，后者是一种基于合数剩余类问题的加法加密技术。以乘法同态加密的典型方案 Elgamal 加密方案为例，其加密方案的密文形式表述如下：

$$\mathrm{Enc} = (C_1, C_2) = (g^r, k^r \cdot s) \tag{3}$$

其中 C_1 和 C_2 是密文，r 是执行加密过程中选取的一个随机数，g 是生成元，k 是公钥，s 是共享秘密。假定存在两个密文：

$$\mathrm{Enc}_1 = (g^{r_1}, k^{r_1} \cdot s_1) \tag{4}$$

$$\mathrm{Enc}_2 = (g^{r_2}, k^{r_2} \cdot s_2) \tag{5}$$

将两个密文进行乘同态加密，得到：

$$\begin{aligned}\mathrm{Enc} &= (g^{r_1} \cdot g^{r_2}, k^{r_1} \cdot s_1 \cdot k^{r_2} \cdot s_2) \\ &= (g^{r_1+r_2}, k^{r_1+r_2} \cdot s_1 s_2)\end{aligned} \tag{6}$$

乘同态加密的密文正好是 $s_1 s_2$ 所对应的密文，参与方进行解密之后得到的信息恰好是明文对应的运算值。这与差分隐私方法不同，参数本身不会进行传递也不会被恶意方获取。

Fang[59]等人就提出了一种基于部分同态加密联邦学习的多方隐私保护框架 PFMLP，其核心思想在于让所有的参与方只通过同态加密的方式传输加密的梯度，尽管暴露在恶意或是好奇的中央服务器中，但基于同态加密的安全系统让原始数据和敏感信息泄露的可能性降到了 1%以下。Hall[60]等人也利用同态加密方法来构建安全分析协议以保护隐私，他们根据安全性定义对数据统计计算并进行合并而无需传递原始数据源。Phong[61]等人针对好奇的服务器设计了一个基于加性同态加密与异步随机梯度下降技术相融合的深度学习系统，与普通深度学习系统相比，他们保障了数据的隐私性同时还维持了模型相当的准确性。在分布式协作训练场景下设计基于同态加密的隐私保护技术还需要考虑密钥持有者的诚实性，以及不可信的参与方之间通

信产生的巨大开销。

Zhang[62]等人基于多方协作建模的纵向联邦学习的隐私问题设计了一种安全的同态加密聚合策略，在聚合权重中增加了一个随机值来掩盖原始参数以增强聚合过程的安全性。该策略也在大量的基准测试中展现出了有效的扩展性和收敛性。文献[63-64]也都提出了不同的融合同态加密和秘密共享技术的隐私保护方案，用来构建一个具有高安全性和可扩展性的联邦迁移学习系统。

还有一些融合同态加密的隐私保护技术[65]在多种结构数据上展现出来较高的扩展性，同时也降低了数据泄露的可能性，但是如果构建了过于复杂的同态密码结构反而会给联邦学习系统增加过多的计算开销，因为联邦学习的训练涉及大量的迭代过程，过长的通讯时间会导致本地模型在聚合阶段产生延迟和参数过期，从而影响最终的全局模型性能。

4.3 安全多方计算

顾名思义，安全多方计算使用了加密方法对多个参与方协作计算传输的参数和生成的模型进行保护，同时它提供了确切的安全证明来保证其是零知识的。安全多方计算也融合了加密方法来保护多参与方之间的通信安全，同时在每一方数据都保留在本地的前提条件下完成数据安全交互，这也解决了在互不信任的多个参与方之间进行隐私保护，同时执行协作计算的问题。

Hao[66]等人就针对中央服务器和多个参与方合谋的情况，提出了一种融合加法同态加密和差分隐私的联合安全多方深度学习协议，使用这种方法在支持大规模用户场景的联邦学习中能展现出较好的安全性、准确性和效率。Mohassel[67]等人就基于两个服务器之间的联合数据进行安全计算来训练对应的模型，并且与当前先进的隐私保护技术相比具有更高的计算准确性同时更快达到收敛。之后，他们在文献[68]中将协议扩展到三台服务器上并使用三方安全计算对联合数据进行训练，同时设计了一个用于机器学习隐私保护的通用框架，并将它用在多个网络模型中，所有服务器之间执行秘密数据共享，构建的框架可以防御来自多个恶意对手的联合隐私窃取。

Wu[69]等人研究了一种用于在纵向联邦学习上进行隐私保护的纵向决策树训练和预测的新颖解决方案 Pivot，它不依赖于第三方，通过增强的多参与方安全协议来保护所有参与方的隐私，让他们同意公开预测结果而不会披露任何中间信息，大量的理论和实验分析也证明了 Pivot 在隐私保护方面是有效的。Gu[70]等人也提出了一种基于纵向联邦学习的异步联邦随机梯度下降算法及其在强凸条件下具有理论保证的变体，以维护在多参与方之间的安全计算，保证隐私不被泄露的同时还能减少通信开销，提升计算效率。

此外，Sharma[71]等人考虑在联邦迁移学习的半诚实模型中提出一套涉及多方计算的安全协议，通过合并秘密共享来提高现有模型在联合数据环境下进行协作训练的效率和安全性。文献[72-74]等都在存在恶意中央服务器或恶意参与方的计算环境中融合了安全多方计算技术来保证多方敏感数据的秘密共享和安全计算，只是在保护隐私和执行计算的效率上需要进行权衡。此外，花费大量的时间开销在执行安全多方计算，这也会在较长的训练时间中对大规模联邦用户的贡献造成损失。

4.4 验证网框架和协作训练方案

Xu[75]等人第一个提出保护隐私和可验证的联邦学习验证网框架，它基于一种双重屏蔽协议，以在模型的训练过程中保证多参与方本地参数的隐私性。同时，在参数聚合阶段，需要中央服务器向每一位参与方提供聚合结果正确性的证明，让计算环境中的包括中央服务器和参与方在内的恶意对手无法窃取隐私，从而维护联邦学习计算过程中的私有数据安全。但是在更大规模的联邦学习计算环境中，服务器与所有参与方的频繁的验证和通信会产生一个较高的计算开销。Liu[76]等人提出了一种支持强制聚合的极端梯度增强联邦学习隐私保护方案，它在协同训练过程中将安全聚合方案扩展到分布式机器学习模型上，较好地支撑了敏感数据在执行传递时的安全性同时保证了效率。文献[77-78]等都提出了基于纵向联邦学习隐私保护的验证算法和框架，并在大量的基准测试中证明了这样的防御方案不仅可以减轻隐私泄露的风险，同时对联邦学习的性能影响也可以忽略不计。

4.5 小结

结合上述，本文对联邦学习包括三种模式下的隐私威胁及隐私保护方案进行了总结，细节如表 2 所示。尽管上述的这些防御方案都在隐私保护上取得了较好的效果，但细化来说，差分隐私技术对敏感数据带来了较强的保护性，但它需要对应用在模型上引起的性能损失和数据保护效果做出权衡。同态加密使用密文形式对隐私数据进行加密传递，用

来对恶意隐私泄露问题做出保护，虽然它在保障数据的隐私性和维护模型的性能上起到了较好的效果，但它却增加了更多的计算开销。安全多方计算让来自于多方的敏感信息得以秘密共享并执行安全计算，在联邦环境中表现出了较好的安全性，只是在隐私保护和系统计算效率之间需要进行平衡。验证网框架和协作训练方案引入了可验证的思想，并建立了双重屏蔽协议以保证全局计算的数据隐私问题，但在大规模计算环境中，这样频繁的验证和通信会引发较大的计算开销和较低的系统执行效率。

表 2　联邦学习的隐私威胁及隐私保护方案

威胁类型	恶意对手	威胁根源	威胁描述	保护方案
隐私	参与方	隐私推断攻击[32-33] 提取重构攻击[34-35] 窃取反演攻击[36] 参与方 GAN 攻击[25] [37]	恶意参与方通过多种手段对其他诚实方的私有数据进行贡献值和敏感信息的窃取和推断	差分隐私[42-58] 同态加密[59-65] 安全多方计算[66-74] 验证网框架和协作训练方案[75-78]
	中央服务器	服务器泄露隐私[38] 服务器 GAN 攻击[39]	恶意中央服务器进行隐私泄露或窃取敏感数据	
	多方合谋	多参与方恶意合谋[40] 参与方和服务器恶意合谋[41]	涉及包括多参与方和参与方与中央服务器之间合谋窃取隐私	

5 安全威胁根源

在联邦学习中，不诚实或是被妥协的参与方往往会利用系统的潜在弱点对模型发起攻击。因为联邦学习系统的执行对所有的参与方都是公开且透明的，同时执行参数传递的过程中都是匿名的，参与方无须告知自己的身份信息，中央服务器也不会对参数上传的所有者进行溯源，这让一些不诚实或是被妥协的参与方有机会对本地模型投毒及篡改或是攻击他人实施未授权的本地数据、参数的入侵访问，并将中毒后的参数或篡改的本地模型进行上传，最终影响全局模型的性能。故本文对安全威胁的产生根源进行了以下分类并做了详细的讨论。

5.1　数据投毒攻击

在模型训练过程中，中央服务器无法检查参与方上传参数的真实性，恶意参与方便可以通过训练中毒数据来生成有毒的本地模型，进而上传到中央服务器上进行全局模型的聚合。如图 6 所示，编号为 c_2 的恶意参与方利用本地的中毒数据进行训练以毒害全局模型。Gonzalez[26]等人就提出了一种新的投毒方法，通过进行中毒样本的训练，在梯度更新的过程中中毒全局模型，并在全局环境中大大降

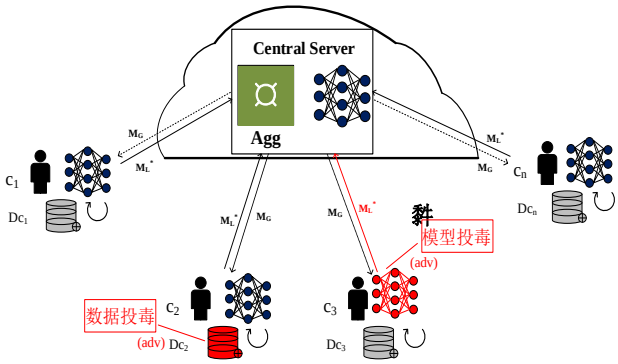

图 6　数据投毒攻击和模型投毒攻击

低了数据投毒攻击的复杂性。Jiang[79]等人设计了一种更彻底的中毒攻击模型，不仅展示了完整的攻击效果，同时在攻击者的目标函数中添加了折中参数来增强隐藏攻击的灵活性，并且保持了较高的攻击有效性。Tolpegin[80]等人研究了联邦学习中恶意参与方执行的标签翻转攻击，攻击者旨在通过发送错误标签数据训练的模型更新来中毒全局模型，即使只有小部分的恶意参与方也具有很好的攻击效果。

5.2　模型投毒攻击

与数据投毒攻击通过制造脏数据来影响全局模型不同的是，模型投毒攻击利用参与方中毒的本地模型与其他各方干净的本地模型进行聚合来影响全局模型的性能，如图 6 所示，编号为 c_3 的恶意参与方上传中毒模型到中央服务器进行聚合，以影

响全局模型的性能。这样的攻击方式往往具有较高的有效性。文献[81-82]已经证明模型投毒攻击的效果比数据投毒攻击的效果要更为明显，因为对于模型投毒攻击来说，恶意参与方可以对本地模型进行恶意修改后上传，让中央服务器在聚合所有本地模型时生成中毒的全局模型，尤其对于涉及多参与方的大规模联邦学习，模型投毒攻击也展现出了更高的成功率和更强的攻击性。

5.3 后门投毒攻击

后门投毒攻击如图 7 所示，通过在联邦学习系统的训练过程中对选定样本创建后门触发器，以便在模型中注入后门任务同时保持隐身效果来完成高效的攻击目的。后门攻击通常会误导训练模型在测试过程中将攻击样本分类为预先设定的目标标签类别。Chen[83]等人就研究了后门投毒策略，来误导受害者学习系统将对手样本分类为指定的目标类别，该策略也在一定基数的样本上获得了 90%以上的成功率。Sun[84]等人在文章中提出的后门攻击方法允许目标任务中选定样本被标记，同时在联邦学习的 TensorFlow 框架 TFF（TensorFlow Federated）中实现了高效攻击。后门攻击对后门样本的误导性很强，同时在攻击过程中对非后门样本产生的影响可以忽略不计。

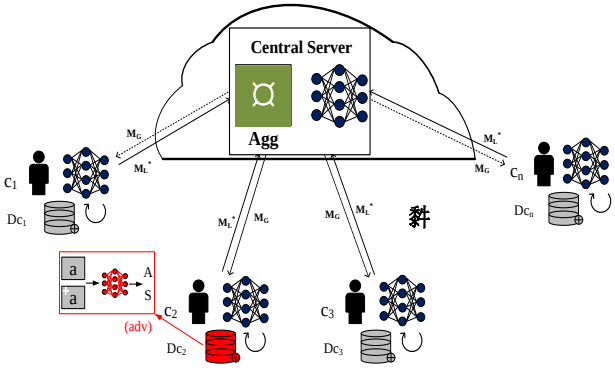

图 7 后门投毒攻击

5.4 搭便车攻击

搭便车攻击不同于其他攻击方法，攻击者只是为了利用全局模型的益处和优势而本身对训练模型不提供任何帮助。正如 Yao[85]等人在文中所述，当贡献极少的参与方想从全局模型中受益，就会发生恶意参与方搭便车的情况，它们在计算环境中只训练少量的数据，以极少的资源消耗来从全局模型中获取更大的利益。Lin[86]等人就在联邦环境下探索了这种搭便车攻击，让恶意攻击者不需要任何本地训练数据就能够构造梯度更新,同时扩大了攻击范围以考虑多方搭便车或是窃取其它除全局模型外的有利隐私信息。搭便车攻击的破坏性较小，同时对全局模型的精度造成的影响比较轻微。

5.5 GAN

许多学者已经对联邦学习中的对抗攻击进行了很多细致的分析和研究，并指出了 GAN 强大的攻击能力和对系统安全带来的威胁性影响。Wang[39]等人就在联邦学习环境中利用恶意服务器进行用户数据攻击，他们构建了一个融合 GAN 和多个鉴别器的框架来使恶意服务器能在无形中发动攻击并取得较好的攻击效果。Zhang[87]等人提出了一种基于 GAN 的投毒方法，他们秘密训练 GAN 来模仿其他诚实参与方的数据样本，进而控制这些样本进行中毒更新以破坏全局模型的构建。因为 GAN 具有极强的攻击隐蔽性，还能在联邦学习中发动无形的攻击，其他参与方无法提前预判 GAN 攻击的时机和方式，所以这种方式产生的攻击威胁性较大。

5.6 数据篡改和推断

在联合多参与方执行联邦学习训练的过程中，中央服务器扮演着十分重要的角色，它负责聚合来自所有参与方上传的本地模型以及共享最终的全局模型。存在恶意的中央服务器会在模型聚合阶段对参与方上传的数据进行污染操作或恶意篡改，以影响最终全局模型的精度。Nasr[88]等人就提出了多个针对不同白盒攻击背景的联邦学习推断算法，对已知的参数信息进行推断攻击，同时也证明了通用化的模型也极易受到对手的白盒推断攻击。这种攻击产生的破坏性很直接，但是也很容易被检测出来，其攻击隐蔽性和成功率并不高。

5.7 多方合谋攻击

当计算环境中的多个恶意对手串通发起联合攻击时，如图 8 所示，c_1 为诚实参与方，他接收来自中央服务器共享的全局模型 M_G，并用本地数据训练后生成本地模型，进而上传更新参数 M^*_L 至中央服务器进行聚合。而编号为 c_2 和 c_3 等多个参与方为恶意对手，他们进行合谋一同向中央服务器上传恶意参数 M^*_{Coll}，并在之后的迭代中循环这些恶意操作，旨在影响全局模型远离期望的良性方向发展。中央服务器在迭代过程中聚合这些参数，这会对全局模型的性能造成巨大的影响。举例而言，众多移动设备构建下一个单词预测的模型时，每一个参与方会根据本地数据进行本地模型构建，倘若多个恶意参

与方合谋上传同样的恶意参数，且恶意参数与模型聚合不发生明显的冲突，这会严重影响聚合器进行参数整合时对客户端的异常识别，生成非预期模型，并做出错误决策。

此外，合谋攻击也会发生在联邦学习所生成的最终模型的所有权上。在联邦学习中，每一个全局模型都有所有权的声明，在客户端-服务器的联合训练体系下，只有一个模型所有者，同时每一个模型实现了较高的准确率都离不开大量数据的学习和花费大量的时间进行训练。然而，模型分发之后，恶意参与方以未经授权的方式恶意使用全局模型，更为严重的是，恶意对手妄想独有模型，他们会伙同多个合谋者进行模型所有权的篡改和非法占有。基于此，本文将多方合谋攻击分为多方恶意参数合谋和多方恶意所有权合谋。以便为学者提供更清晰的合谋攻击分类和攻击手段的划分。

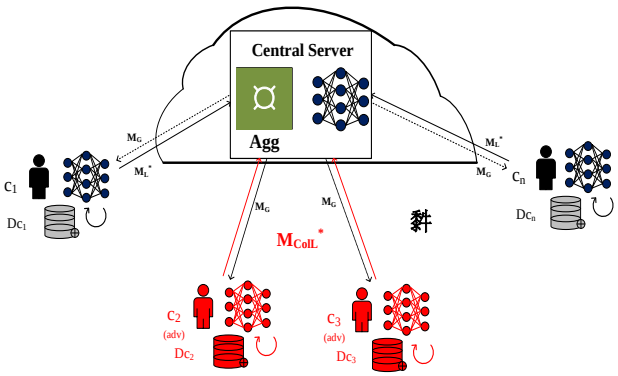

图 8 多参与方恶意合谋攻击

5. 7. 1 多方恶意参数合谋

联邦学习模型在分布式设备上进行全局训练，它们在聚合过程中选取部分参与方进行聚合，这样的聚合方式允许参与方被选择性的加入和离开。然而，这样的系统极容易受到 sybil 攻击[89]，这种攻击方式更多地出现在允许客户加入和离开的区块链系统中。恶意攻击者可能会创建多个虚假客户端身份，或者在 sybil 攻击中合拢受感染的可妥协设备，以进行更直接的模型更新操作。Fung[90]等人就基于聚合过程中的漏洞引入了 sybil 合谋攻击，它们不需要额外的辅助信息或假设条件，就可以让全局模型的性能大大降低。Jiang [91]等人针对模型投毒攻击的新研究提出了 sybil 合谋攻击用以扰乱参与方的模型更新。他们发动 sybil 攻击，欺骗诚实的参与方，操纵学习到的参数，执行合谋攻击，使得最终的全局模型具有更差的分类效果。在这样的联邦学习环境中，恶意对手不需要操作其他数据，也不会损害系统算法以及模型体系结构，他们发动 sybil 攻击可以欺骗诚实用户，执行恶意合谋，让模型远离预期方向发展。Bagdasaryan[92]等人也设计了多个恶意参与方对模型的合谋攻击，合谋攻击不仅可以混淆检测，还会放慢全局模型的收敛速度。

5.7.2 多方恶意所有权合谋

机器学习模型的最新应用已经涉及到自动驾驶，核工程以及智慧医疗等各个重要领域。而训练这样的高精度模型是一项开销巨大的过程，因为它需要处理大量的数据，同时需要使用大量的计算资源。考虑到设计和训练的昂贵流程，训练出来的最终模型会被视为模型所有者的知识产权，需要加以保护，以防止被恶意用户随意的侵权和滥用。然而，多个恶意用户会进行串谋发动合谋攻击，企图对模型进行非法篡改和占用，这让模型的所有权问题面对着极大的挑战。Li[93]等人就指出了恶意用户合谋对全局模型进行非法复制和重新分发的问题，同时恶意用户会串谋进行潜在的算法攻击和模型修改，以在窃取的模型上植入自己的标记并占用该模型。另外，合谋对手也会对一些白盒设置下的较为简单的防御方法进行解析和重构，以实现窃取最终模型的目的。多方合谋攻击涉及对联邦学习的多个安全层面发动攻击，具有较广的攻击性和较大的威慑力。

5.8 小结

上述的涉及安全威胁的攻击方法在一定程度都具有相当的攻击性。而攻击的影响程度是由大量的学者对当前前沿的工作直接研究得出的结论。其中，数据投毒攻击、模型投毒攻击和后门投毒攻击都具有很强的攻击威胁性。很大的原因来自于，全局模型在执行聚合的过程中，投毒攻击将错误标签训练的参数或者直接利用恶意参数进行上传，这些中毒参数越多，中毒效果越严重，对联邦学习的影响也会越大。它们利用聚合的漏洞，同时兼具较难检测的特性，以这种影响来操纵模型的输出，极大地降低全局模型的精度。基于 GAN 的攻击也具有很强的攻击性，因为其在联邦环境下具有不可预测的特点，同时它对所有参与方数据的隐私性和安全性具有较大的威胁，并且具有可以发动无形攻击的能力，这让它屡屡取得较好的攻击效果。数据篡改和推断的攻击威胁性中等，尽管它在白盒设置下的攻击效果很直接，但也很容易被检测到，攻击成功

率并不高。搭便车攻击所带来的影响和威慑力最小，首先，这样的攻击方式发生在少有的应用场景上，其次，它追求于获取全局模型的益处而不参与贡献，本质上并没有毒害全局模型，同时，它随着模型一起学习，并不干扰其他参与方的训练，所以这种攻击方式带来的负面影响微乎其微。多方合谋攻击的攻击威胁性很强，相较于单个恶意参与方只进行单一的独立攻击相比，多方合谋攻击可以融合多种攻击方式，执行串谋攻击，即使存在少量的恶意用户也具有一定的攻击效果。合谋攻击还可以混淆检测，篡改参数，影响全局模型的收敛速度。在实际的联邦学习应用场景中，包括涉及数百万台智能设备联合训练热词预测模型，恶意对手发动潜在攻击以使模型输出种族歧视言论[3]。还有需要联合周边路况、行人、建筑、交通标志等信息进行协同训练的无人驾驶汽车，倘使恶意对手篡改交通标志数据，这无疑会导致严重的交通事故[10]。这些应用都存在潜在的安全隐患，故本文在下一节中进一步阐述安全问题解决方案。

6 安全问题解决方案

基于联邦学习生成的全局模型，恶意对手可以进行多种攻击。这些攻击手段不仅可以影响全局计算系统的聚合过程，降低全局模型的精度，同时也可以伪造本地训练样本或是篡改更新的本地模型来实现他们的恶意目的。针对当前联邦学习系统所面临的种种安全威胁，本文在下述内容对这些威胁的最新防御技术进行详细的阐述并对它们的应用场景和效果进行了总结。

6.1 数据投毒防御

通常，数据投毒防御技术应当在训练参数进行聚合之前对参数的真实性进行安全检测，或是在训练过程中，对参数进行筛查和验证，以防止恶意对手使用不完整或是被污染的数据进行本地模型的训练，从而降低全局模型的性能。Cao[94]等人就针对恶意参与方的数据投毒问题，从参数更新方向的角度入手，提出了一种用于聚合过程中进行参数检测的新颖的 Sniper 方案，该方案能识别出诚实用户并最大程度降低样本中毒的攻击成功率，即使在面对分布式环境下多个恶意参与方的中毒样本也体现出了较高的检测率。

6.2 模型投毒防御

针对模型投毒攻击，通常诚实的服务器作为鉴别器需要检查参与方更新的本地模型是否对全局模型的性能有益，每次的筛选过程会把迭代中对模型无益的参与方累加标记以确定潜在攻击者，或是将所有参与方上传的本地模型进行比较，观察和标记存在异常的本地模型，以此来降低全局模型聚合中毒模型之后受到的性能损失影响。Bhagoli[81]等人就基于这个思想提出了恶意更新检测技术和交替最小化策略来防御恶意用户执行的模型投毒攻击，并在小规模用户上实现了较好的检测效果和收敛性。但当参与方扩大到数百万个时，这个方法会产生巨大的计算开销，同时检测成功率也会降低。

6.3 异常检测

可靠的异常检测方法通常通过一系列的技术统计和分析来计算正常行为和异常事件，对于本地数据被污染进而上传的中毒参数具有较好的检测效果。Li[95]等人就在服务器端进行用户异常检测，同时使用生成模型的权重向量作为低维替代物在 MNIST 数据集上进行分类训练并取得了较好的检测效果。Fan[96]等人也基于在联邦迁移学习模型中的恶意入侵提出了异常检测框架，该框架对未知的攻击具有出色的检测泛化能力，使所有基于该模型的物联网网络能够进行信息共享的同时还有力的维护数据安全。文献[97-98]都提出了相应的检测方法来识别客户端的负面更新，对模型的安全带来了有力的保证。

6.4 模型蒸馏

模型蒸馏通过压缩技术从原始模型中提取出另外一个模型，这样的防御手段能较好地降低对抗攻击的成功率，同时还能减少目标模型的复杂度，降低系统的计算开销。文献[99-100]就通过迁移学习和蒸馏技术开发了一个通用安全框架，让每个客户端不仅保持自己的私有数据不被泄露，而且本地还拥有独特设计的模型以进行联合训练，模型蒸馏使得测试数据集的对抗攻击成功率大大降低，同时还减轻了全局模型的复杂度，帮助系统更快达到收敛。Papernot[101]等人也引入了一种融合蒸馏技术的防御机制，用来降低训练过程中对抗攻击的有效性，并提升了训练速率。

6.5 搭便车防御

为了应对不诚实的参与方对全局模型益处的窃取而无益于模型构建的搭便车攻击，Kim[102]等人设计了一个针对该场景的区块链联邦学习架构 BlockFL，它利用区块链交换和验证用户端的本地参数更新机制来对联邦学习系统中的用户进行关

联，并给予与贡献值成正比的奖励来阻止恶意用户执行不诚实的搭便车行为，同时奖励机制也会激励所有用户一同参与贡献。Weng[103]等人也设计了一种基于区块链的价值驱动激励机制，以鼓励所有参与者积极贡献。但使用区块链技术的交换和关联操作会引发计算网络中较大的计算时延，这可能就不适用于大规模用户构建联邦学习模型的应用场景。

6.6 GAN 防御

恶意对手发动基于 GAN 的攻击往往会在训练过程中伪装成诚实的参与方，然后训练一个用来对其他参与方的训练数据进行模拟的 GAN 框架，进而通过不断训练伪造的数据样本来影响全局模型的性能，同时获得诚实参与方的信任并攻击目标用户窃取更多私有的敏感信息。Zhang[104]等人就针对数据集样本的对抗扰动和推理攻击提出了一种重要的防御对抗攻击框架 Anti-GAN，该框架能将真实样本与扰动样本进行区分，是一种有效的基于对抗网络的防御方法，同时设计了一个新的损失函数来将相似特征进行分类，并在多个数据集上展示出了可靠的防御能力。

6.7 模型剪枝

模型剪枝技术通常将对整体模型贡献很小或是受到污染的参数进行删除，然后将剩余的权重进行聚合重构，其往往被利用来对恶意参数进行防御以抵御中毒攻击，同时还能最小化全局模型以降低模型的复杂度，加速联邦学习计算且保持稳定的模型准确性。Yu[105]等人就基于恶意参与方实施的后门攻击提出了剪枝技术，将中毒参数进行清除并有效地降低了对全局模型的攻击成功率，同时还保证了边缘设备上模型的质量。Jiang[106]等人也提出了一种新颖的基于联邦学习的自适应分布式参数剪枝方法，该方法对中毒参数进行了有效削减，同时还保证了与原始模型相当的精度，减少了大量的训练时间，优化了通信瓶颈。针对大规模分布式联邦学习的深度神经网络模型训练过程中出现的包括数据投毒和后门投毒等攻击手段，可靠的剪枝技术可以有效地缓解这些恶意攻击，同时还能优化计算效率不高、通信延迟较强等性能问题。

6.8 恶意合谋防御

恶意合谋问题往往发生在多个恶意参与方或是恶意中央服务器和参与方之间，恶意参与方一同上传恶意参数以影响最终模型的精度，同时他们具有很高的隐蔽性，存在较高的检测复杂度和难以针对性捕获等问题。此外，针对花费昂贵的训练开销训练出来的最终模型会被视为模型所有者的知识产权，基于所有权的恶意合谋侵权和滥用问题也被越来越多的人关注到。也有更多的学者针对这些问题提出了不同的防御方案以对全局模型进行保护。

6.8.1 多方恶意合谋防御

面对联邦学习系统中中央服务器聚合参数，全局迭代运行聚合算法来训练共享模型的过程中所受到的 sybil 合谋攻击问题，Fung[90][107]等人针对这种特性的攻击设计了一套可靠的防御方法，它可以根据分布式学习过程中参与方上传更新的多样性来识别中毒的 sybil 合谋。同时，该防御方法不限制预期的合谋攻击者数量，在基于 sybil 的标签翻转和后门中毒攻击中都展现出了极强的防御性。Jiang[91]等人针对这种合拢多个可妥协设备的合谋攻击提出了一种可以监控每个回合中所有参与者的平均损失，然后进行收敛异常检测的防御方法，并为每个参与方设计了预测成本报告来防御 sybil 攻击，该方法也有效地减轻了攻击对模型收敛带来的影响。文献[108-109]也都为这一类型的攻击手段提供了新的防御思路，并展现了较好的防御效果。

6.8.2 多方恶意所有权保护

针对恶意对手串谋，企图发动合谋攻击对模型所有权进行侵权或滥用的破坏行为，最近已经有许多学者提出了很多不同的防御技术，Tekgul[110]等人设计了 WAFFLE，他们在每次本地模型聚合生成全局模型的过程中引入再训练过程以在模型中添加水印，防止多个恶意参与方合谋使用其他标记数据进行训练覆盖本地水印，同时在全局模型被恶意窃取和替换时也能进行水印再验证，水印方法的嵌入不会影响模型的精度也不会引起任何多余的通信开销。文献[93] [111]也都分别在模型训练过程中利用指纹技术嵌入所有参与方具有唯一性的指纹，以便系统检测出错误参数同时根据指纹索引到执行了合谋的恶意用户。文献[112-113]也都在设计的防御框架中融入数字签名技术和具有独特性的数字水印技术来验证用户进而保护模型的安全。

6.9 安全聚合方案

聚合算法在中央服务器执行本地参数的聚合过程中发挥着十分重要的作用。一个安全性强的联邦学习聚合算法应该能在本地模型的迭代更新中自适应地对恶意客户端上传的异常参数进行检测，并对可疑参数进行标记删除，或是自主调节聚合方式、融合加密机制，来保证所有隐私数据和模型的安全。

现阶段的一些传统的聚合方法在安全聚合和隐私保护上存在严重的不足。举例而言，联邦平均算法 FedAvg 在执行全局聚合过程中利用参数平均得到全局模型。它在系统建模的过程中不会对参与方的任何本地数据进行上传或迁移操作，能较好地维护数据的隐私性，同时在中央服务器中只负责聚合更新的本地模型参数而无需进行其他操作。尽管在进行联邦计算过程中数据没有产生迁移，但在聚合计算期间却存在诸多漏洞和一些可进行恶意操控的安全隐患。相较于执行缓慢的联邦平均算法，作为改进，Xie[114]等人提出了 FedAsync 聚合算法，他们利用异步联合优化技术提升了中央服务器在聚合过程中的灵活性和可扩展性，较好地缓解了来自恶意用户无效聚合引发的巨大延迟带来的性能低下影响，同时也在各种应用中表现出更快的收敛速度，但它却没有在安全层面做出任何提升。Wang[115]等人也提出了一种自适应全局控制聚合方案。他们从理论的角度分析了分布梯度下降的收敛边界，并确定了本地更新和全局参数聚合之间的最佳折衷。该方案动态调整全局模型的聚合频率以确保获取期望的模型性能，与固定聚合方案相比，自适应控制聚合方案保证联邦学习模型准确性的同时，还减少了大量训练能耗。尽管这些方法在某些层面都分别表现出各自独特的优势，但它们在安全聚合方面的劣势和漏洞让很多恶意对手乘虚而入，发动不同的攻击来完成自己的恶意目的。

为了防止恶意对手利用聚合算法的弱点在本地更新期间推断联邦学习参与方的隐私信息同时利用污染参数影响全局模型的性能，大量学者设计了一系列可靠的聚合算法来维护全局模型的安全，本文也在下述内容对联邦学习的聚合算法进行详细的阐述和讨论。

6.9.1 联邦随机控制平均聚合

基于联邦学习在聚合过程中被恶意参与方篡改梯度和参与方聚合期间产生的不稳定性问题，Karimireddy[116]等人为了缓解这一类的本地梯度参数不相似和客户端更新漂移问题提出了联邦随机控制平均（SCAFFOLD）聚合算法，他们旨在利用该算法来维护参与方的梯度参数变量同时控制聚合流程，从而确保所有客户端朝着全局安全聚合的方向进行更新。

6.9.2 安全多方计算聚合协议

Bonawitz[117]等人设计了一种基于安全多方计算的安全聚合协议，用于联邦学习中移动设备的训练和模型构建。该协议对互不信任的多个参与方的参数进行聚合而不会泄露他们的隐私信息。同时该聚合方案具有较强的容错能力，即使存在部分恶意用户不参与聚合的情况，它也能出色的完成安全聚合并实现预期效果。尤其针对联合大量移动设备进行建模的联邦计算环境，该方案能轻松应对掉线挑战并实现较好的聚合性能。

6.9.3 个性化联邦聚合

因为存在不集中汇总数据的异质性问题，跨用户设备的数据统计异质性源都需要有鲁棒性的算法来保证性能。Arivazhagan[118]等人就提出了一种融合了基础和个性化层的安全聚合方法（FedPer），用于深度前馈神经网络的联合训练，以消除统计异质性对性能带来的负面影响。同时，针对异质性的数据，采用个性化的方法进行对齐聚合，而不会泄露任何一方的敏感信息。研究还表明，如果最大程度地提高全局模型的性能，就需要很好的限制本地模型进行个性化的能力，因为两者在训练过程中一直在互相权衡。故 Deng[119]等人提出了一种自适应个性化联邦学习算法（APFL），在对本地模型和全局模型混合的广义边界进行推导的同时，找出最优混合参数，来使得两者都表现出局部最佳效果。他们还提出了一种高效的通信方式来维护模型的高收敛性。

上述的这些聚合方法都提出了不同的安全聚合手段，旨在维护全局模型聚合过程的安全性和模型准确性，同时在模型收敛性、梯度漂移和优化通信问题上分别做出了重要贡献。

6.10 小结

结合上述，针对不同的基于联邦学习的攻击方法，本文分别对各类攻击方法相应的防御方案进行了罗列。当恶意的攻击者对诚实方进行攻击并实施未授权的本地数据、参数或本地模型的入侵窃取、投毒及篡改等行为来破坏模型构建时，相应的防御方案对这些针对性的攻击都具有较好的防御效果。此外，数据投毒防御可以对多个恶意参与方进行检测，并具有高检测率；模型投毒防御在保证了较好的检测效果同时实现了更好的收敛性；异常检测对中毒参数的负面更新具有较好的甄别作用；搭便车防御为企图不劳而获的不诚实参与方设计了很好的激励策略，以鼓励所有参与方积极做出正面贡献；模型剪枝和模型蒸馏技术可以降低恶意样本的对抗性，同时减少了训练时间，提高了通信效率；GAN 防御对伪造样本具有很高的防御可靠性；多方

合谋攻击的防御涉及对模型精度以及模型所有权的保护，同时，防御策略可以维持模型更好的收敛。另外，不同的联邦安全聚合手段分别给联邦学习系统带来了不同的性能效益，它们在维护全局模型聚合安全性和模型精度的同时，在一些场景下对模型收敛性、梯度漂移和优化通信问题上也分别做出了重要贡献。本文把联邦学习的安全威胁及安全防御方法总结如表 3 所示。

表 3　联邦学习的安全威胁及安全防御方法

威胁类型	威胁根源	威胁描述	威胁性	防御方法
安全	数据投毒攻击[79-80] 模型投毒攻击[81-82] 后门投毒攻击[83-84] 搭便车攻击[85-86] GAN[39] [87] 数据篡改和推断[88] 多方合谋攻击[89-93]	恶意的攻击者会对联邦学习系统实施攻击，包括对自身本地数据、参数的投毒及篡改或对其他参与方实施未授权的入侵窃取及推理，并将攻击后的中毒参数或中毒模型进行上传更新，最终影响全局模型的性能。	较强 较强 较强 较弱 较强 中等 较强	数据投毒防御[94] 模型投毒防御[81] 异常检测[95-98] 模型蒸馏[99-101] 搭便车防御[102-103] GAN 防御[104] 模型剪枝[105-106] 恶意合谋防御[90-91,93,107-111] 安全聚合方案[116-119]

7 联邦学习安全框架

针对联邦学习设计的一些安全开源框架已经开发如下，它们旨在为科研人员和领域开发者提供一个平台以探索联邦学习的安全可行解决方案。

7.1　TensorFlow Federated

TensorFlow Federated[120]简称 TFF，是 Google 基于 TensorFlow 开发的用于执行分布式机器学习的开放框架，在解决分散的数据计算和跨设备的联邦学习建模上具有较高的灵活性。它主要运用于横向联邦学习模式，现阶段大多数关于隐私保护和安全防御的研究主要是基于 TFF 框架实现的。同时它包括一些高级接口，允许在现有的 TensorFlow 模型上实施联邦学习计算，同时集成了 Kubernetes 集群，以执行在众多参与方和中央服务器之间的交互。此外，借助 TFF，可以使用 LEAF[121]生成的特定数据集来尝试使用用户设计的联邦学习算法。该技术典型的应用是支持用于移动设备上的下一个单词预测。尽管 TFF 使用广泛，其支持本地模式的分布式计算，虽然在一定程度上具有安全的执行保障，但在数据隐私和多方聚合上缺乏全面的保护技术。

7.2　FATE

作为由微众银行发起的开源项目，FATE[122]旨在为联邦 AI 生态提供可靠和安全的计算框架。其内在架构分为多个模块，基本集成了联邦学习横向、纵向和迁移学习三种模式下的所有功能，包括数据训练建模、特征值预处理、联邦聚合、加密共享等流程。它可以从主机中手动或使用 Docker 镜像安装，支持单机独立和多节点集群模式。更为重要的是，它内在融合了同态加密和安全多方计算等方法来应对恶意对手的攻击。Mothukuri[123]等人也分析了该框架在抵御恶意用户从数据集层面对联邦学习模型发动攻击所产生的影响。

7.3　PySyft

PySyft[124]涉及在非可信环境中进行安全和隐私保护，它是一个基于 PyTorch 的安全框架。作为一个由 OpenMined 牵头的项目，它通过给定的接口连接用户端并构建了一个维系通信的标准化协议，同时设计了一个基于张量的抽象模型，覆盖了服务器端的聚合和客户端的本地训练和信息共享流程。此外，该框架还集成了先进的差分隐私、同态加密和安全多方计算技术来保证横向联邦学习计算过程中的数据隐私安全。PySyft 能较好地执行检测并抵御破坏数据或模型的恶意攻击，但是，该框架在执行联邦学习训练过程所花费的训练时间过长，整体训练效率并不高，在扩展性和迁移性上也存在很大的改进空间。

7.4　PaddleFL

PaddleFL[125]开源联邦学习框架，作为百度牵头的研究项目，它可以较为容易地部署在大规模的分布式集群当中。同时，PaddleFL 提供了许多联邦学习策略，它让学者可以较为轻松地使用它复制和比

较不同的联邦学习方法。其集成了包括横向联邦学习、纵向联邦学习和联邦迁移学习三种模式，同时融合了差分隐私和安全聚合等隐私保护方法，并在计算机视觉、推荐式系统以及自然语言处理等多个领域得到了广泛应用。该框架搭建简易，同时在对用户的数据隐私和安全性的保障上提供了一些可靠的防御手段，但是，它在一些给用户提供的安全的个性化服务和解决用户需求方面还需要提升，并且缺乏全面的安全协议。

7.5 小结

结合上述，当前主流的联邦学习框架都在横向联邦学习模式下得到了广泛应用，但是，诸如 TFF 和 PySyft 目前还是局限于解决横向联邦学习问题。TFF 仅支持本地模式下的分布式计算，缺乏全面的隐私保护技术，也正是因为 TFF 本地实现分布式计算的能力，现阶段大多数关于隐私保护和安全防御的研究是基于 TFF 框架实现的，它主要面向学术界并为众多学者提供一个灵活性高的联邦学习平台。而 PySyft 同时在产业界和学术界上得到了众多青睐，它集成了差分隐私、同态加密和安全多方计算等技术来维护本地数据和敏感参数的传递，同时 PySyft 在对恶意参数的异常检测上表现出较好的性能，不过在整体训练效率、扩展性和迁移性上还是有很大的提升空间。FATE 和 PaddleFL 也都被广泛地应用在产业界和学术界中，它们都基本集成了横向联邦学习、纵向联邦学习和联邦迁移学习三种模式下的所有功能，同时融合了包括差分隐私、同态加密和安全多方计算等多种隐私保护方法，并广泛应用在计算机视觉、推荐式系统和物联网智能家居中。尽管 PaddleFL 安装简便，在一些数据隐私和安全问题上也提供了一系列可靠的保障，但它在当前给用户提供的一些安全服务上还需要做出一些改进。FATE 也针对恶意对手可能发动的攻击部署了一些可靠的防御措施，它是当前联邦 AI 生态下一个最为可靠和安全的计算框架。虽然这四个不同的联邦学习安全框架已经在一些场景下得到了应用，它们也在一定程度上对隐私数据和模型安全起到保护作用，但他们在全局执行效率、对模型受到攻击的防御性和收敛性以及系统训练的扩展性和迁移性上仍然存在一定的提升空间。本文把联邦学习安全框架及其属性对比总结如表 4 所示。

表 4 联邦学习安全框架及其属性对比

联邦学习安全框架	所属机构/应用领域	涉及参与方数	模式分类	保护技术	优势及不足	提出时间
TensorFlow Federated	Google (学术)	多个(≥2)	横向	差分隐私	灵活性高，使用广泛，但缺乏数据隐私和聚合的保护技术	2017
FATE	微众银行 (产业/学术)	多个(≥2)	横向，纵向，迁移	同态加密，安全多方计算	可靠性高，并集成了三种联邦模式，融合了多种加密手段	2019
PySyft	OpenMined (产业/学术)	多个(≥2)	横向	差分隐私，同态加密，安全聚合	融合多种保护技术维护数据安全，但执行训练时间过长，效率不高，扩展性有待提升	2018
PaddleFL	百度 (产业/学术)	多个(≥2)	横向，纵向，迁移	差分隐私，安全聚合	搭建简易，集成三种模式，融合隐私保护技术，个性化安全服务有待提升	2020

8 研究挑战及未来方向

为了兼顾处理数据隐私保护和“数据孤岛”带来的挑战，联邦学习给当前数据融合、共享需求愈发强烈的 AI 带来了希望。它保护本地数据的同时为多方构建联合模型，并在数据敏感的应用当中发挥着极大的作用。本文介绍了联邦学习的基本概念、并总结了联邦学习隐私和安全的威胁根源以及先进的隐私保护和安全防御方法，以期望大家规避风险，构建一个安全的联邦学习计算环境。目前，联邦学习的安全性问题仍处于研究的初步阶段，现有的方法和技术手段只能在一定的条件下提升全局模型的鲁棒性。在实际计算的过程中，仍存在一些亟待解决的挑战值得大家仔细思考。

1) 权衡隐私保护的效率和模型的精度

所构建模型的性能及其可用性是当前隐私保护方法面临的巨大挑战，在现有的隐私保护方法上，他们普遍以牺牲效率或模型精度为代价来达到增强隐私保护的目的[46-48]。但是如何构建高效的隐私保护安全协议同时保证模型的精度是当前联邦学习亟待解决的主要问题。而权衡两者也需要考虑较多的因素，包括参与方的好奇性、本地参数的敏感性、聚合模型的预期性能、服务器诚实与否以及所添加方法产生的额外开销等。如果使用的方法加密程度较弱，则参与方的隐私仍暴露在被泄露的风险之中，相反，如果加密程度过强，则会引发较大的计算开销，还可能对全局模型的性能产生负面影响。介于此，未来的研究工作可以考虑开发融合多种技术的隐私保护方案，使得在保护数据隐私的同时，不仅不会产生较大的性能损耗，同时在数据加密的保护性以及通信过程的安全性上得到互补保障，但这也需要加密方案带来更大的创新。

2) 开发适用于纵向联邦学习和联邦迁移学习的融合保护技术

在当前的研究背景下，现阶段的关于联邦学习隐私保护和抵御恶意攻击的主要工作都集中在横向联邦学习上[46-50]，在纵向联邦学习和联邦迁移学习的研究到目前来看还是十分有限[51] [53]。在这两种模式下，由于需要参与方之间进行更加密切的数据交互和协作，因此迫切需要构建能满足所有参与方安全需求的可靠安全协议，或是开发融合多种隐私保护方法的混合策略，以实现在计算中的每一个过程都能达到最优保护效果的目的。所设计的融合方法不仅需要满足参与方对通信安全的需求，同时也要考虑到在数据共享过程中如何保护每一个积极提供正面贡献的参与方的数据隐私安全。这是一个值得不断探索的新研究方向。

3) 联邦学习参数上传的溯源性

在联邦学习执行过程中，全局流程对所有参与方都是公开且透明的，执行上传参数的各参与方都进行的是匿名操作，中央服务器不会对这些参数进行参与方的溯源，这会使得很多恶意对手攻击模型并隐藏自己[79-84]。一些参数检测方法只能简单的筛选恶意参数并不能捕获这些恶意参与方，于是，在之后的多轮迭代中，恶意对手会继续上传破坏性的参数，给系统在安全检测上造成巨大的开销，影响模型收敛。一个全面的安全防御方法不仅能有效防御恶意攻击，还要具备参数溯源能力以回溯识别上传恶意参数的用户端，进行高效的联邦学习系统全局保护。因此，未来针对恶意参数的检测所提出的防御方法应当着手考虑既能高效检测攻击又能回溯恶意参数。在一般情况下，可以考虑在训练过程中进行水印和指纹编码的嵌入，在模型被窃取替换或是恶意参数检测上进行恶意对手的溯源，这样做不仅有效的检测出恶意参数，保护了模型的构建，同时捕获到了恶意对手，截断其继续利用恶意参数执行聚合的企图，极大地提升了系统计算安全性，只是使用该技术手段需要考虑添加的编码对模型的性能是否会产生不良影响。

4) 权衡联邦学习的安全交互和低效通信问题

在移动设备上执行分布式训练过程中，中央服务器需要聚合来自所有分散参与方的本地参数，但往往联邦网络由大量设备组成，例如数百万台移动智能手机和 PC 协同训练机器学习模型，网络中设备的通信速度可能会因为庞大的设备体量而降低很多个数量级，联邦学习模型也会以更慢的速度达到收敛。缓慢的通信速度会造成一些数据包的丢失和部分移动设备消耗过多能量被迫关机进而引发掉线或失活，最终退出联邦网络的情况，这也使得通信效率成为了系统的一个关键性能瓶颈[3]。同时，为了保护数据隐私和安全而对一些敏感数据添加的加密措施也会极大地加剧服务器的通信负担[69-71]。因而，未来的研究需要突破通信瓶颈和安全交互带来的权衡影响，想要降低通信开销，提升通信效率，可以考虑设计高效的聚合算法以在每一次迭代过程中选取贡献较大或是更具有代表性的样本执行参数上传，或者设计可靠的通信策略利用异步通信或是邻近通信在保护用户敏感数据的前提下尽可能减少全局交互次数、以及考虑压缩本地模型减少数据传递量等方法，只是这些方法需要在聚合收敛性、设备参与度、计算灵活性和模型准确性上做出平衡。

5) 联邦学习的通信安全问题

联邦学习在参数传递的过程中所有参与方都是匿名通信的，他们不需要向任何用户展示自己的身份信息，这也导致诚实参与方容易受到恶意对手的通信干扰攻击。就像在文献[126]所述攻击方式，恶意对手通过利用高功率射频来干扰参与方和中央服务器之间通信信号，以达到破坏模型安全聚合的目的。恶意对手让服务器聚合的模型无法共享到用户端，同时用户端的本地参数发生通信丢失无法执行上传，最终影响全局模型构建的质量。此外，信道攻击也会让本地数据的隐私和本地模型的安全

暴露在更大的风险之中。因此，未来的研究应当考虑非诚实的联邦学习网络中的通信安全问题，可以建立一套安全的通信体系，也可以设计类似抗信号干扰的方法来保证安全通信，或是设计参数备份上传方案来规避类似的通信掉线和丢失的风险，以及进行信道攻击后参数的验证。

6) 联邦学习的多用户安全贡献问题

在涉及大量用户或机构参与的联邦学习环境中，用户或机构的数据可能相差巨大，一些持着大量数据的用户或机构想从模型中获益，但却不参与贡献。这往往会在联邦学习系统中发生搭便车攻击行为，这种方式使得模型不能够聚合到所有的本地模型而引发非全面的性能效果[85-86]。虽然这样的攻击方式对模型产生的攻击效果不强，但也让全局模型的性能无法在聚合所有参与方的本地模型中获得更大的改进。当贡献极少的参与方想从全局模型中受益，它们在本地计算环境中可能只训练少量的数据，消耗少量的资源来尽可能从全局模型中获取更大的好处，此外，它们还能从全局模型中分析提取其他参与方的敏感信息。因此，未来对诸如这样搭便车攻击的防御研究需要考虑让更多的数据持有方积极参与贡献，同时也要维护参与贡献的数据隐私安全。为了让联邦学习更好的实现商业化，让更多的机构、企业或平台参与进来，可以考虑设计一套公平且安全的鼓励机制，让参与更多正面贡献的用户或机构得到更高的优先级或者更大的权限，并鼓励更多的机构加入进来。因为最终模型的高效性取决于所有参与方提供的正面贡献，同时，这套机制也必须维护所有数据的隐私，确保每一个用户或机构参与贡献而不会泄露敏感数据。这样安全的鼓励机制也会影响到更多的机构加入进来的积极性。当然，设计这样一套公平且安全的鼓励机制也需要考虑如何让所有用户达成共识，如何对用户的正面贡献进行判定，如何保证贡献更多数据的用户执行计算的数据隐私性和安全性问题。

7) 开发联邦学习安全框架

近年来，开发联邦学习相关的安全框架已经成为领域内的研究热点，已有较多的扩展版本[120-125][127]正在验证当中。由文章第 7 节可以得知，现有的联邦学习框架尽管在一定程度上对隐私数据和模型安全起到保护作用，但他们在执行效率、模型的防御性和收敛性以及系统计算能耗上仍然存在很大的提升空间。同时，现有的安全框架数量较少，而集成了三种联邦学习模式且部署了隐私保护方法和安全防御策略的全面框架更少，与最早构建的 TFF 框架相比，后续的工作应该建立更稳定的计算环境，同时面对当前的网络安全状况，开发一个具有低延迟、高普适性和强鲁棒性的联邦学习框架是保护系统安全的重大挑战。

9 总结

本文主要对近几年基于联邦学习的隐私和安全的前沿文献进行了详细梳理，从多个方面罗列了其在联邦学习中的破坏手段及威胁性，同时，对联邦学习的隐私保护和安全防御研究进行了更清晰的分类。与当前的相关综述相比，本文还总结了联邦学习的多方恶意合谋问题，细致的分析了现有的联邦安全聚合算法和安全开源框架，旨在帮助大家规避风险，构建安全的大规模分布式联邦学习计算环境。尽管这些研究起步较晚，但都在面向联邦学习的隐私和安全性问题上提供了一些新的解决思路，并成功应用在一些人工智能应用中。然而，在实现联邦学习应用的过程中，仍然存在一些亟待解决的挑战，尤其针对权衡隐私保护和模型准确性、融合隐私保护技术的开发、参与方参数溯源性、低效通信、安全通信、多用户安全贡献和联邦学习安全框架开发这七大问题值得重点考虑。

此外，在未来的研究工作中，更好的完善相关隐私保护和安全防御技术的设计，加快对联邦学习聚合方案的构建以及促进联邦学习安全框架及平台的开展，可以更多的让联邦学习的益处惠利到每一个领域。

参 考 文 献

[1] Sajjad M, Nasir M, Muhammad K, et al. Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. Future Generation Computer Systems, 2017, 108: 995-1007

[2] Cao K, Jain Anil K. Automated latent fingerprint recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 788-800

[3] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions. IEEE Signal Process Magazine, 2020, 37(3): 50–60

[4] Mohammadi M, Al-Fuqaha A. Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. IEEE Communications Magazine, 2018, 56(2): 94-101

[5] Wang S, Cao J, Philip Yu. Deep learning for spatio-temporal data mining: A survey. IEEE Transactions on Knowledge and Data Engineering, 2022 34(8): 3681-3700

[6] Roh Y, Heo G, Whang S E. A survey on data collection for machine learning: A big data - AI integration perspective. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(4): 1328-1347

[7] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks//Proceeding of the Advances in Neural Information Proceeding Systems (NIPS). Nevada, USA. 2012: 1223–1231

[8] Stergiou C, Psannis K E, Kim B G, et al. Secure integration of IoT and cloud computing. Future Generation Computer Systems, 2016, 78(3): 964-975

[9] Mukherjee M, Shu L, Wang D. Survey of fog computing: Fundamental, network applications, and research challenges. IEEE Communications Surveys & Tutorials, 2018, 20(3): 1826-1857

[10] Nguyen D C, Ding M, Pathirana P N, et al. Federated learning for industrial internet of things in future industries. IEEE Wireless communications magazine, 2021, 28(6): 192-199

[11] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA. 2017: 1273-1282

[12] Zhang J, Zhao Y, Wu J, et al. LVPDA: A lightweight and verifiable privacy-preserving data aggregation scheme for edge-enabled IoT. IEEE Internet of Things Journal, 2020, 7(5): 4016-4027

[13] Zhu T, Philip Yu. Applying differential privacy mechanism in artificial intelligence//Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). Dallas, USA. 2019: 1601-1609

[14] Goryczka S, Xiong L. A comprehensive comparison of multiparty secure additions with differential privacy. IEEE Transactions on Dependable and Secure Computing, 2017, 14(5): 463-477

[15] Phong L T, Phuong T T. Privacy-preserving deep learning via weight transmission. IEEE Transactions on Information Forensics and Security, 2019, 14(11): 3003-3015

[16] Kieu T, Yang B, Guo C, et al. Outlier detection for time series with recurrent autoencoder ensembles//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao, China. 2019: 2725-2732

[17] Zhou Jun, Fang Guo-ying, Wu Nan. Survey on security and privacy-preserving in federated learning. Journal of Xihua University (Natural Science Edition), 2020, 39(4): 9-17 (in Chinese)
(周俊, 方国英, 吴楠. 联邦学习安全与隐私保护研究综述. 西华大学学报(自然科学版), 2020, 39(4): 9-17)

[18] Chen Bing, Cheng Xiang, Zhang Jia-le, Xie Yuan-yuan. Survey of security and privacy in federated learning. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5) (in Chinese)
(陈兵, 成翔, 张佳乐，谢袁源. 联邦学习安全与隐私保护综述. 南京航空航天大学学报, 2020, 52(5))

[19] Yang Geng, Wang Zhou-sheng. Survey on privacy preservation in federated learning. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2020, 40(5): 204-214 (in Chinese)
(杨庚, 王周生. 联邦学习中的隐私保护研究进展. 南京邮电大学学报(自然科学版), 2020, 40(5): 204-214)

[20] Kozik R, Choras M, Ficco M, et al. A scalable distributed machine learning approach for attack detection in edge computing environments. Journal of Parallel and Distributed Computing, 2018, 119: 18-26

[21] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19

[22] Yang Qiang, Liu Yang, Cheng Yong, Kang Yan, Chen Tian-jian, Yu Han. Federated learning: “isolated data islands” and data protection problem solving method. Beijing: Publishing House of electronics industry, 2020 (in Chinese)
(杨强, 刘洋, 程勇, 康焱, 陈天健, 于涵. 联邦学习: 数据孤岛和数据保护难题破解之法. 北京，电子工业出版社，2020)

[23] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging. 2016, arxiv:1602.05629.

[24] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS). New York, USA. 2017: 1175–1191

[25] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS). New York, USA. 2017: 603–618

[26] Muoz-Gonzalez, L, Biggio B , Demontis A , et al. Towards poisoning of deep learning algorithms with back-gradient optimization//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York, USA. 2017: 27-38

[27] Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. 2017, arXiv:1711.10677.

[28] Nock R, Hardy S, Henecka W, et al. Entity resolution and federated learning get a federated resolution. 2018, arXiv:1803.04035.

[29] Yang H, He H, Zhang W, et al. FedSteg: A federated transfer learning framework for secure image steganalysis. IEEE Transactions on

Network Science and Engineering, 2020, 8(2): 1084-1094

[30] Sudipan S, Tahir A. Federated transfer learning: Concept and applications. Intelligenza Artificiale, 2021, 15(1): 35-44

[31] Liang X, Liu Y, Chen T, et al. Federated transfer reinforcement learning for autonomous driving. 2019, arXiv:1910.06001.

[32] Zhang J, Zhang J, Chen J, et al. GAN enhanced membership inference: A passive local attack in federated learning//Proceedings of the 2020 IEEE International Conference on Communications (ICC). Dublin, Ireland. 2020: 1-6

[33] Luo X, Wu Y, Xiao X, et al. Feature inference attack on model predictions in vertical federated learning//Proceedings of the IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece. 2021: 181-192

[34] Tramer F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIs//Proceedings of the 25th USENIX Security Symposium (USENIX Security). Austin TX, USA. 2016: 10-12

[35] Wu Z, Li Q, He B. Exploiting record similarity for practical vertical federated learning. 2021, arXiv:2106.06312.

[36] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership inference attacks against generative models. Privacy Enhancing Technologies. 2019, (1): 133-152

[37] Song M, Wang Z, Zhang Z, et al. Analyzing user-level privacy attack against federated learning. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430-2444

[38] Melis L, Song C, Cristofaro E D, et al. Exploiting unintended feature leakage in collaborative learning//Proceedings of the IEEE Symposium on Security and Privacy (SP). San Francisco, USA. 2019: 691-706

[39] Wang Z, Song M, Zhang Z, et al. Beyond inferring class Representatives: user-level privacy leakage from federated learning//Proceedings of the IEEE Conference on Computer Communications (INFOCOM). Paris, France. 2019: 2512-2520

[40] Yang R, Au M H, Lai J, et al. Collusion resistant watermarking schemes for cryptographic functionalities//Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT). Kobe, Japan. 2019: 371-398

[41] Lim W Y B, Luong N C, Hoang D T, et al. Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031-2063

[42] Dwork C, Mcsherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis//Proceedings of the Third conference on Theory of Cryptography (TCC). New York, USA. 2006: 265-284

[43] Dwork C. Differential Privacy: A Survey of Results. Theory and applications of models of computation, 2008: 1-19

[44] Li J, Khodak M, Caldas S, et al. Differentially private meta-learning. 2019. arXiv:1909.05830.

[45] McMahan B H, Daniel R, Kunal T, et al. Learning differentially private language models without losing accuracy. 2017, arXiv:1710.06963.

[46] Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec). New York, USA. 2019: 1–11

[47] Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective. 2017, arxiv:1712.07557

[48] Agarwal N, Suresh A T, Yu F X , et al. cpSGD: Communication-efficient and differentially-private distributed SGD//Proceedings of the 32nd International Conference on Neural Information Proceeding Systems (NIPS). Montreal, Canada. 2018: 7564–7575

[49] Xin B, Yang W, Geng Y, et al. Private FL-GAN: Differential privacy synthetic data generation based on federated learning//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore. 2020: 2927-2931

[50] Ghazi B, Pagh R, Velingker A. Scalable and differentially private distributed aggregation in the shuffled model, 2019, arXiv: 1906.08320.

[51] Feng S, Yu H. Multi-participant multi-class vertical federated learning. 2020, arXiv:2001.11154.

[52] Liu Y, Zhang X, Wang L. Asymmetrical vertical federated learning. 2020, arXiv:2004.07427.

[53] Li X, Chi H, Lu W, et al. Federated transfer learning enabled smart work packaging for preserving personal image information of construction worker. Elsevier Automation in Construction, 2021, 128-103738

[54] Nergiz M E, Clifton C. δ-Presence without complete world knowledge. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(6): 868-883

[55] Jiang D, Tong Y, Song Y, et al. Industrial federated topic modeling. ACM Transactions on Intelligent Systems and Technology, 2021, 12(1):1-22

[56] Wang Y, Tong Y, Shi D. Federated latent dirichlet allocation: A local differential privacy based framework//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA. 2020, 34(04): 6283-6290

[57] Jiang D, Song Y, Tong Y, et al. Federated topic modeling//Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM). Beijing, China. 2019: 1071-1080

[58] Song T, Tong Y, Wei S. Profit allocation for federated learning//Proceedings of the IEEE International Conference on Big Data (Big Data). Los Angeles, USA. 2019: 2577-2586

[59] Fang H, Quan Q. Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet, 2021,

13(4): 94

[60] Hall R, Fienberg S E, Nardi Y. Secure multiple linear regression based on homomorphic encryption. Journal of official stats. 2011, 27(4): 669-691

[61] Phong L T, Aono Y, Hayashi T, et al. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333-1345

[62] Zhang Q, Gu B, Deng C, et al. Secure bilevel asynchronous vertical federated learning with backward updating//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021,35(12), 10896-10904.

[63] Liu Y, Kang Y, Xing C, et al. A secure federated transfer learning framework. IEEE Intelligent Systems, 2020, 35(4): 70-82

[64] Gao D, Liu Y, Huang A, et al. Privacy-preserving heterogeneous federated transfer learning//Proceedings of the IEEE International Conference on Big Data (Big Data). Los Angeles, USA. 2019: 2552-2559

[65] Zhang C, Li S, Xia J, et al. BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning//Proceedings of the USENIX Annual Technical Conference (USENIX ATC). Santa Clara, USA. 2020: 493-506

[66] Hao M, Li H, Xu G, et al. Towards efficient and privacy-preserving federated deep learning//Proceedings of the IEEE International Conference on Communications (ICC). Shanghai, China. 2019: 1-6

[67] Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning//Proceedings of the IEEE Symposium on Security & Privacy (SP). San Francisco, USA. 2017: 19-38

[68] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). New York, USA. 2018: 35–52

[69] Wu Y, Cai S, Xiao X, et al. Privacy preserving vertical federated learning for tree-based models. Proceedings of the VLDB Endowment, 2020, 13(12): 2090-2103

[70] Gu B, Xu A, Huo Z, et al. Privacy-preserving asynchronous vertical federated learning algorithms for multiparty collaborative learning. IEEE Transactions on Neural Networks and Learning Systems, 2021, 1-13

[71] Sharma S, Xing C, Liu Y, et al. Secure and efficient federated transfer learning//Proceedings of the IEEE International Conference on Big Data (Big Data). Los Angeles, USA. 2019: 2569-2576

[72] Araki T, Furukawa J, Lindell Y, et al. High-throughput semi-honest secure three-party computation with an honest majority//Proceedings of the ACM SIGAC Conference on Computer and Communications Security (CCS). Vienna, Austria. 2016: 805-817

[73] Mohassel P, Rosulek M, Zhang Y. Fast and secure three-party computation: The garbled circuit approach//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS). New York, USA. 2015: 591-602

[74] Hao M, Li H, Luo X, et al. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6532-6542

[75] Xu G, Li H, Liu S, et al. VerifyNet: Secure and verifiable federated learning. IEEE Transactions on Information Forensics and Security, 2019, (15): 911-926

[76] Liu Y, Ma Z, Liu X, et al. Boosting privately: Privacy-preserving federated extreme boosting for mobile crowdsensing. 2019, arXiv:1907.10218.

[77] Xu R, Baracaldo N, Zhou Y, et al. FedV: Privacy-preserving federated learning over vertically partitioned data/Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec). Korea. 2021: 181-192

[78] Zhang S, Xiang L, Yu X, et al. Privacy-preserving federated learning on partitioned attributes. 2021, arXiv:2104.14383.

[79] Jiang W, Li H, Liu S, et al. A flexible poisoning attack against machine learning//Proceedings of the IEEE International Conference on Communications (ICC). Shanghai, China. 2019: 1-6

[80] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems//Proceedings of the European Symposium on Research in Computer Security (ESORICS). Guildford, UK. 2020: 480-501

[81] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019: 634-643

[82] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to Byzantine-robust federated learning//Proceedings of the 29th USENIX SECURITY SYMPOSIUM (USENIX SECURITY), 2020: 1605-1622

[83] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. 2017. arXiv:1712.05526.

[84] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning?//Proceedings of the 2nd International Workshop on Federated Learning for Data Privacy and Confidentiality at NeurIPS2019. Vancouver, Canada. 2019

[85] Yao X, Huang C, Sun L. Two-stream federated learning: Reduce the communication costs//Proceedings of the IEEE Visual Communications and Image Proceedings (VCIP). Taichung, China. 2018: 1-4

[86] Lin J, Du M, Liu J. Free-riders in federated learning: Attacks and

defenses, 2019, arXiv:1911.12560.

[87] Zhang J, Chen J, Wu D, et al. Poisoning attack in federated learning using generative adversarial nets//Proceedings of the 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). Rotorua, New Zealand. 2019: 374-380

[88] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning//Proceedings of the IEEE Symposium on Security and Privacy (SP). San Francisco, USA. 2019: 739-753

[89] Xiao X, Tang Z, Li C, et al. SCA: Sybil-based collusion attacks of IIoT data poisoning in federated learning. IEEE Transactions on Industrial Informatics, 2022

[90] Fung C, Yoon C J, Beschastnikh I. Mitigating sybils in federated learning poisoning. 2018. arXiv:1808.04866.

[91] Jiang Y, Li Y, Zhou Y, et al. Mitigating sybil attacks on differential privacy based federated learning. 2020, arXiv:2010.10572.

[92] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning//Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS). California, USA. 2020: 2938–2948

[93] Li F, Wang S, Liew A W. Towards practical watermark for deep neural networks in federated learning, 2021, arXiv:2105.03167.

[94] Cao D, Chang S, Lin Z, et al. Understanding distributed poisoning attack in federated learning//Proceedings of the IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). Tianjin, China. 2019: 233-239

[95] Li S, Cheng Y, Liu Y, et al. Abnormal client behavior detection in federated learning// Proceedings of the 2nd International Workshop on Federated Learning for Data Privacy and Confidentiality, in Conjunction with NeurIPS 2019 (FL-NeurIPS 19). 2019

[96] Fan Y, Li Y, Zhan M, et al. IoTDefender: A federated transfer learning intrusion detection framework for 5G IoT//Proceedings of the IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE). Guangzhou, China. 2020: 88-95

[97] Shen S, Tople S, Saxena P. AUROR: Defending against poisoning attacks in collaborative deep learning systems//Proceedings of the 32nd Annual Conference on Computer Security Applications (ACSAC). Austin, USA. 2016: 508-519

[98] Li S, Cheng Y, Wang W, et al. Learning to detect malicious clients for robust federated learning. 2020, arXiv:2002.00211.

[99] Li D, Wang J. FedMD: Heterogenous federated learning via model distillation//Proceedings of the NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality. 2019

[100] Wang A, Zhang Y, Yan Y. Heterogeneous defect prediction based on federated transfer learning via knowledge distillation. IEEE Access, 2021, (9): 29530-29540.

[101] Papernot N, Mcdaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the IEEE Symposium on Security and Privacy (SP). San Francisco, USA. 2016: 582-597

[102] Kim H, Park J, Bennis M, et al. On-device federated learning via blockchain and its latency analysis. 2018. arXiv:1808.03949.

[103] Weng J, Weng J, Zhang J, et al. DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Transactions on Dependable and Secure Computing, 2019, 18(5): 2438-2455

[104] Zhang X, Luo X. Exploiting defenses against GAN-based feature inference attacks in federated learning, 2020, arXiv:2004.12571.

[105] Yu S, Nguyen P, Anwar A, et al. Adaptive dynamic pruning for non-IID federated learning. 2021, arXiv:2106.06921.

[106] Jiang Y, Wang S, Valls V, et al. Model pruning enables efficient federated learning on edge devices. IEEE Transactions on Neural Networks and Learning Systems, 2022:1-13

[107] Fung C, Yoon M, Beschastnikh I. The limitations of federated learning in sybil settings//Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID). San Sebastian, Spain. 2020: 301-316

[108] Blanchard P, ElMhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent//Proceedings of the in Advances in Neural Information Proceedings Systems (NIPS). Los Angeles, USA. 2017: 119-129

[109] Yin D, Chen Y, Ramchandran K, et al. Byzantine-robust distributed learning: Towards optimal statistical rates//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden. 2018: 5636–5645

[110] Tekgul B G A, Xia Y, Marchal S, et al. WAFFLE: Watermarking in federated learning//Proceedings of the 40th International Symposium on Reliable Distributed Systems (SRDS). Chicago, USA. 2021: 310-320

[111] Rouhani B D, Chen H, Koushanfar F. DeepSigns: An end-to-end watermarking framework for protecting the ownership of deep neural networks//Proceedings of the 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). New York, USA. 2019

[112] Chen Y, Luo F, Li T, et al. A training-integrity privacy-preserving federated learning scheme with trusted execution environment. Information Sciences, 2020, 522: 69-79

[113] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding Watermarks into deep neural networks//Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR). Ottawa, Canada. 2017: 269-277

[114] Xie C, Koyejo S, Gupta I. Asynchronous federated optimization. 2019. arXiv:1903.03934.

[115] Wang S, Tuor T, Salonidis T, et al. Adaptive federated learning in resource constrained edge computing systems. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1205-1221

[116] Karimireddy S P, Kale S, Mohri M, et al. SCAFFOLD: Stochastic controlled averaging for on-device federated learning//Proceedings of the 37th International Conference on Machine Learning (ICML). Kyoto, Japan. 2020: 5132-5143

[117] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for federated learning on user-held data. 2016, arXiv:1611.04482.

[118] Arivazhagan M G, Aggarwal V, Singh A K, et al. Federated learning with personalization layers. 2019, arXiv:1912.00818.

[119] Deng Y, Kamani M M, Mahdavi M. Adaptive personalized federated learning, 2020, arXiv:2003.13461.

[120] TensorFlow Federated: Machine learning on decentralized data, Google, Mountain View, USA, 2019.

[121] Caldas S, Duddu S M K, Wu P, et al. LEAF: A benchmark for federated settings. 2018. arXiv:1812.01097.

[122] Luo J, Wu X, Luo Y, et al. Real-world image datasets for federated learning, 2019, arXiv:1910.11089.

[123] Mothukuri V, Parizi R M, Pouriyeh S, et al. A survey on security and privacy of federated learning. Future Generation Computer Systems, 2020, (115): 619-640

[124] Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning. 2018. arXiv:1811.04017.

[125] He C, Li S, So J, et al. FedML: A research library and benchmark for federated machine learning, 2020, arXiv:2007.13518.

[126] Xu W, Ma K, Trappe W, et al. Jamming sensor networks: attack and defense strategies. IEEE Network, 2006, 20(3): 41-47

[127] Ulm G, Gustavsson E, Jirstrand M. Functional federated learning in erlang (ffl-erl)//Proceedings of the International Workshop on Functional and Constraint Logic Programming (WFLP). 2019: 162-178

Xiao Xiong, Ph.D. candidate. His main research interests include cloud computing, distributed machine learning and privacy and security of federated learning.

Xiao Bin, Ph.D., professor, Ph.D. supervisor. His main research interests are cyber security, and currently focuses on the network and cloud security, blockchain technology and AI security.

Tang Zhuo, Ph.D., professor, Ph.D. supervisor. His majors are distributed computing system, parallel processing for big data, including distributed machine learning, security model, and resources scheduling and management in these areas.

Li Kenli, Ph.D., professor, Ph.D. supervisor. His main research interests include parallel computing, cloud computing, and Big Data computing.

Background

As an emerging technology in the field of artificial intelligence, federated learning takes into account the issues of " isolated data islands " and privacy protection. For example, it combines millions of mobile devices scattered around for joint training without accessing their private data. Federated learning brings new solutions to many applications that cannot be used for data fusion and interaction. Although it plays an important role in the process of small data aggregation training, it still has some potential privacy leakage risks and data security issues. At present, many scholars are committed to researching the issues of privacy and security in federated learning.

In order to further explore the current research status of the privacy protection and security defense issues in federated learning. In this paper, the authors first introduce the threats to privacy and security in federated learning from many aspects. For privacy issues, the authors analyze the root causes of

privacy threats from multiple scenarios, including single malicious participant attacks, central server attacks, and multiple participants malicious collusion to leak privacy and describe the specific attack process and the attack effect of these privacy stealing methods in detail. For security issues, the authors focus on analyzing a variety of malicious attack methods that affect the performance of the global federated model, including independent attacks and collusion attacks. The authors further systematically evaluate the existing state-of-the-art researches concerning the privacy and security issues of federated learning, and make a clear classification. The authors list their destructive methods and threats from multiple aspects. Then, the authors focus on the most advanced protection methods for privacy protection and security defense, and analyze their defense capabilities to help researchers avoid risks and provide solutions to some security issues. Finally, we focus on the problem of multi-party malicious collusion in federated learning, federated security aggregation algorithms, and secure open-source frameworks, which aims to provide the researchers with a clearer security vision in the field. In addition, at the end of this paper, the authors put forward some challenges that need to be urgently solved and future directions that worth thinking about based on the current research of federated learning.

The work is supported by the National Key Research and Development Program of China (2018YFB1701400), the Hunan Provincial Natural Science Foundation of China (2021JJ40612), the National Natural Science Foundation of China (Grant Nos. 61873090, L1824034, L1924056).