计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：联邦学习中的隐私保护技术研究综述
作者：王腾，霍峥，黄亚鑫，范艺琳
收稿日期：2021-12-09
网络首发日期：2022-04-26
引用格式：王腾，霍峥，黄亚鑫，范艺琳. 联邦学习中的隐私保护技术研究综述[J/OL]. 计算机应用. https://kns.cnki.net/kcms/detail/51.1307.TP.20220425.1937.008.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2021122072

联邦学习中的隐私保护技术研究综述

王腾[1],霍峥[2*]，黄亚鑫[2]，范艺琳[2]

(1. 中国电科网络通信研究院，石家庄，050000
2. 河北经贸大学 信息技术学院，石家庄 邮编 050061)

(*通信作者电子邮箱: huozheng@heuet.edu.cn)

摘 要: 近年来，联邦学习成为解决机器学习中数据孤岛与隐私泄露问题的新思路。联邦学习架构不需要多方共享数据资源，参与方在本地数据上训练局部模型，周期性地将参数上传至服务器更新全局模型，就可以获得在大规模全局数据上建立的机器学习模型。联邦学习架构具有数据隐私保护的特质，是未来大规模数据机器学习的新方案；然而，该架构的参数交互方式也可能导致数据隐私泄露。目前，研究联邦学习架构中的隐私保护机制已经成为新的热点。从联邦学习中存在的隐私泄露问题出发，探讨了联邦学习中的攻击模型与敏感信息泄露途径，重点综述了联邦学习中的几类隐私保护技术：以差分隐私为基础的隐私保护技术、以同态加密为基础的隐私保护技术、以安全多方计算为基础的隐私保护技术。最后，探讨了联邦学习中隐私保护中的若干关键问题，展望未来研究方向。

关键词: 联邦学习；隐私保护；差分隐私；同态加密；安全多方计算

中图分类号: TP311.5 **文献标志码:** A

Survey of privacy-preserving technologies in federated learning

WANG Teng[1], HUO Zheng[2], HUANG Yaxin[2], FAN Yilin[2]

(1. China Electronics Technology Group Corporation Network Communication Research Institute,
2. Information Technology School, Hebei University of Economics and Business)

Abstract: In recent years, federated learning has become a new way to solve the problems of data island and privacy leakage in machine learning. Federated learning architecture does not require multiple parties to share data resources, it only needs participants to train local models on local data and periodically upload parameters to the server to update the global model, and then they can obtain machine learning models built on large-scale global data. Federated learning architecture has the nature of privacy protection and is a new scheme for large-scale data machine learning in the future. However, the parameter interaction mode of federated learning may also lead to data privacy disclosure in both model training stage and model prediction stage. At present, strengthening the privacy protection mechanism in federated learning architecture has become a new research hotspot. Starting from the privacy disclosure in federated learning, the attack models and privacy information disclosure paths in federated learning was discussed, and several types of privacy protection techniques in federated learning were highlighted and reviewed, such as privacy protection technology based on differential privacy, privacy protection based on homomorphic encryption, and privacy protection mechanism based on secure multi-party computing. Finally, the key issues of privacy protection in enhanced federated learning were discussed, focusing on the problems of model convergence and performance, personalized privacy requirement, etc.

Keywords: Federated learning; privacy preserving; differential privacy; homomorphic encryption; secure multiparty computation

0 引言

机器学习算法在自动识别、智能决策等方面具备显著优势，已逐渐成为人工智能和大数据处理的技术基础。大部分机器学习算法需要庞大的训练数据集来保证训练模型的性能[1]，在这背后是大量的个人数据被采集，包括姓名、身份证件号码、联系方式、住址、账号密码、财产状况、行踪轨迹、消费状况等，甚至还有生理特征、就医记录等更敏感的信息。上述信息不但被采集、利用，甚至还会售卖给第三方获取利益，个人隐私遭到严重的泄露。随着个人用户、政府部门及数据采集方对个人数据隐私的关注，国家相继出台各种法律法规，严禁非法采集公民的个人数据：2017 年 6 月起，我国实施了《中华人民共和国网络安全法》[2]，2021 年 9 月 1 日

收稿日期: 2021-12-09; **修回日期**: 2022-01-21; **录用日期**: 2022-01-28。
基金项目: 国家自然基金项目(62002098);河北省自然科学基金项目(F2020207001,F2021207005)。
作者简介: 王腾(1980—)，男，贵州遵义人，高级工程师，博士，主要研究方向：机器学习、数字化治理；霍峥(1982—)，女，河北邯郸人，副教授，博士，CCF 会员(18239M)，主要研究方向：隐私保护技术、联邦学习；黄亚鑫(1999—)，男，河北邢台人，硕士研究生，主要研究方向：隐私保护技术；范艺琳(1998—)，女，河北石家庄人，主要研究方向：联邦学习。

正式实施了《中华人民共和国数据安全保护法》[3]；2021 年 11 月 1 日实施了《中华人民共和国个人信息保护法》[4]。

即使能合法采集个人数据，个人数据大多分散存储在不同的机构中，由于政策壁垒与存储资源的限制，很难实现数据的集中存放。近年来，联邦学习(Federated Learning, FL)[5]的出现，成为机器学习领域的新热点。联邦学习的概念最早是在 2016 年由谷歌提出的[6]，它是一种分布式的机器学习框架。分布在多个结点上的数据集协同训练，最终可获取全局数据集上的机器学习模型。联邦学习具有天然的隐私保护特质，数据不需要集中存放，仅需在数据分散存储的节点上训练模型，服务器无法获取原始数据，个人数据隐私得到有效的保护。在数据隐私与安全问题备受关注的今天，联邦学习在避免数据泄露、避免中心点数据受到攻击等方面具备显著优势。此外，传统的机器学习模型不能直接处理异构数据，利用联邦学习技术，无需处理异构数据即可建立全局数据上的机器学习模型，既保护了数据隐私，又解决了数据异构问题[7]。联邦学习可应用在涉及个人敏感数据的机器学习任务中，如个人医疗数据、可穿戴设备数据、面部特征数据、个人资产数据等等[8-10]。

目前，许多机器学习模型已扩展到联邦学习架构中，比如线性回归[11]、支持向量机[12]、神经网络[13-14]、聚类[15]、决策树[16-17]、深度学习[18-19]等。然而，研究发现，联邦学习架构的隐私保护度不足以完全防御外部隐私攻击[20]，具体来说，在模型训练和模型预测阶段都可能泄露数据隐私。在模型训练阶段，通常需要构建经验损失函数，采用随机梯度下降(Stochastic Gradient Descent, SGD)方法找到损失函数的最小值，将最小值对应的参数作为模型参数上传给服务器。不可信服务器/外部攻击者可能利用参与方的模型参数逆推数据分布特征，甚至逆推出具体的训练集数据，导致参与方的数据隐私泄露。此外，在模型预测阶段，攻击者可反复调用模型进行预测，特别是对某些泛化能力不足的模型，在预测某些训练集中出现过的数据时，模型的表现与训练集中未出现过的数据有较大差距，攻击者通过这一特征可判断某些数据是否出现在训练集之中，如果训练集包含敏感信息，则个人隐私泄露。

隐私保护技术经过多年的发展，逐渐形成了几类较为成熟的方法：以差分隐私为代表的数据扰动法[21]、以 *k*-匿名为代表的数据泛化法[22]、以安全多方计算为代表的数据加密法[23]等。隐私保护的应用场景从最初的关系型数据发布、基于位置的服务等简单场景，逐渐发展到较为复杂的社交网络、电子商务、图像识别等领域。在上述隐私保护应用场景中，数据可用性与隐私保护度是一对矛盾，研究的关键问题在于如何在保护隐私的前提下提高数据可用性。而在机器学习/联邦学习场景下，隐私保护度和模型精确度是一对矛盾，隐私保护度的提升意味着模型预测精确度的下降、模型的收敛速度变慢等问题。尤其是深度学习模型结构异常复杂，且不具备可解释性，使得隐私保护与模型可用性之间的矛盾关系无法量化。针对联邦学习中的隐私泄露问题，需要设计新的隐私保护方案。

目前，联邦学习中的隐私保护技术已经成为联邦学习领域的研究热点，研究者们发表了不少相关研究内容的综述，如表 1 所示。

表 1 联邦学习中隐私保护技术的相关综述

Tab.1 Reviews related to privacy-preserving federated learning

方法来源	环境		攻击模型		安全防御	隐私保护	评价标准
	机器学习	联邦学习	安全攻击	隐私攻击			
文献[20]	√			√		√	
文献[24]	√		√		√		
文献[25]		√	√		√		
文献[26]		√	√		√		
文献[27]		√	√		√		
文献[28]	√		√		√		
文献[29]		√		√		√	
本文		√		√		√	√

文献[20]中对机器学习中的隐私攻击和隐私保护方法进行了调研和分析，侧重机器学习中的隐私保护技术；文献[24]中对分布式深度学习中的隐私与安全攻击模型、防御措施进行了综述；文献[25-27]中对联邦学习架构中的安全攻击与防御措施进行了综述，侧重于安全攻击与防御。文献[28]中重点介绍了机器学习环境中安全攻击的类型及防御方法；文献[29]中综述了联邦学习的概念及隐私保护技术，提出了联邦学习中隐私问题的“5W”；文献[30]中综述了物联网领域中，利用联邦学习训练基于用户隐私数据的机器学习模型的研究现状，重点讨论了其中的隐私保护策略、通信代价和数据异构问题。

1 预备知识

1.1 隐私与隐私保护

隐私是指个人或实体不愿被外界知晓的信息。早在 19 世纪发表在《哈佛法律评论》上的《论隐私权》[31]中就将隐私定义为“不受打扰的权利”。随后，各国不断修整完善涉及隐私权的法律法规，直到 2018 年 5 月欧盟实施了最严格的隐私保护法——《通用数据保护条例》[32]，要求企业赋予用户“被遗忘的权利”。同年，数据隐私被纳入计算机专有名词，指数据中直接或间接蕴含的，涉及个人或组织的，不宜公开的，需要在数据收集、数据存储、数据查询和分析、数据发布等过程中加以保护的信息。敏感信息是指不当使用或未经授权被人接触或修改会不利于国家利益、联邦政府计划的实行、不利于个人依法享有的个人隐私权的所有信息。隐私保护技术通过对原始数据的变换达到保护个人敏感信息不泄露的目的，同时保证能在变换后的数据上获取信息、模型或服务。

1.2 联邦学习

联邦学习是一种分布式机器学习架构，由中心服务器、参与方 P_i ($1 \le i \le n$)及用户构成。其中，参与方各自持有本地数据集 D_i，无需进行数据共享，通过协作的方式训练在全局数据集 $D = \bigcup_{i=1}^{n} D_i$ 上的模型[33]。与传统的分布式系统不同，联邦学习的各参与方可以是“异质”的，即参与方软硬件配置、持有的数据格式、数据分布、模型结构等都可不同，依据不同角度可对联邦学习进行如下分类：

(1) 根据参与方数量的多寡与算力的强弱，联邦学习可分为 cross-device 和 cross-silo 两类[34]。Cross-silo 中参与方往往为大型组织（如医疗、金融等相关机构），数量较少但算力较强；cross-device 中参与方为个人设备，数量庞大且算力较弱，在该场景下，不是每个参与方都有机会参与每一轮训练，通常利用采样的方式确定哪些用户可以参与训练过程。

(2) 根据联邦学习架构中是否存在中心服务器，联邦学习架构可以分为中心化架构与去中心化架构，如图 1 所示。去中心化架构[35]不需要可信服务器，在每次迭代中，参与方在本地数据上更新梯度，将梯度发送到选定的一方，选定方使用其本地数据和梯度值再度更新模型，直到所有参与方都更新了模型，最后将模型广播给所有参与方。为了保证模型的公平性，充分利用各方数据，参与方事先约定迭代相同的轮数。

(3) 根据不同参与方之间的数据特征分割方式，联邦学习又可分为横向联邦学习(Horizontal federated learning)、纵向联邦学习(Vertical federated learning)和联邦迁移学习(Transfer Federated Learning, TFL)[4]。数据水平分割指数据持有方存储了不同用户的具有相同属性的数据；数据垂直分割指数据持有方存储了相同用户的不同属性的数据；联邦迁移学习指数据持有方持有的数据中用户和属性重叠都较少的情况，如图 1 所示。

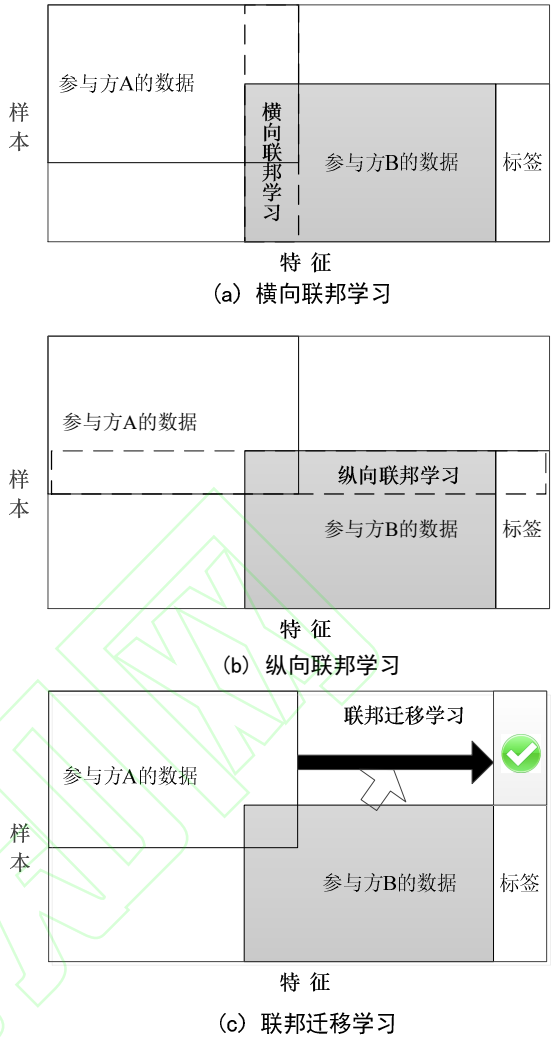


图1 基于数据分割方式的联邦学习分类[5]

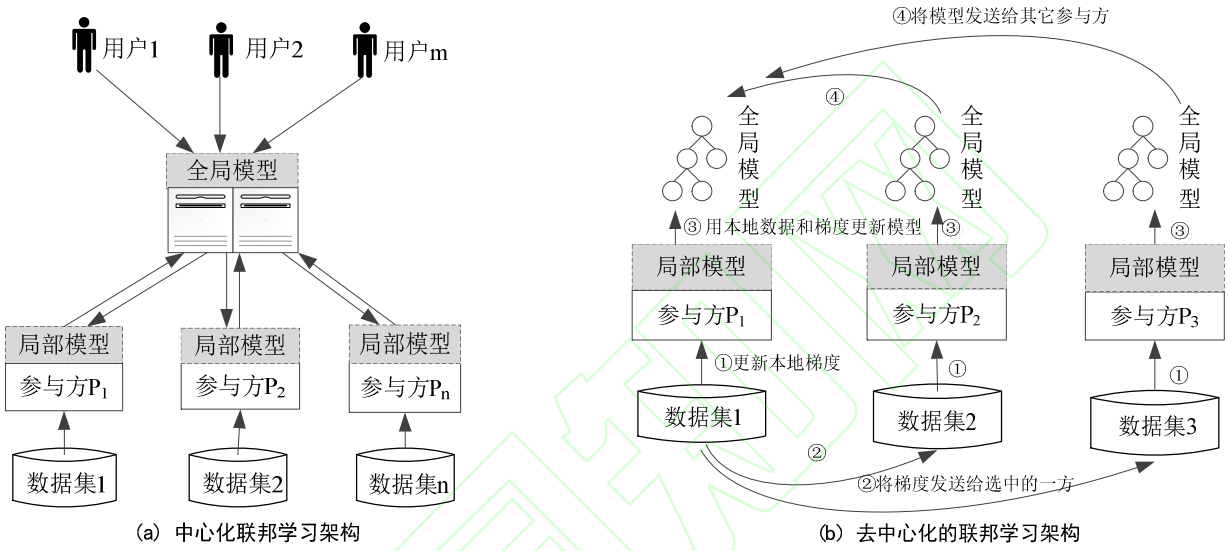
Fig. 1 Federated learning classification based on data segmentation[5]

联邦学习中参与方的参数更新方式可分为两类：一类是基于随机梯度的更新方法(SGD-based)，另一类为基于模型特征的更新方法(Model specialized)[36]。FedSGD 和 FedAVG[37]是基于随机梯度更新的典型方法。FedSGD 指参与方将每轮机器学习的梯度值传给服务器，服务器聚合后返回给参与方；FedAVG 方法允许参与方在服务器聚合参数之前多次迭代计算梯度值，服务器不必每次计算中间结果的均值，减少了通信轮数。FedSVRG [38]、FedProx[39]、FedNova [40]等方法对 FedAVG 的参数聚合进行了改进。FedSVRG 向服务器发送的不是简单的梯度值，而是随机方差缩减梯度，与 FedSGD 相比，FedSVRG 方法在相同迭代轮数下模型精度更高；FedProx 和 FedNova 考虑到参与方数据异构的问题，以限制本地更新的次数与全局聚合的方式提高模型精确度。常用的 FedAVG 聚合方式为加权平均，即：$\boldsymbol{w} = \sum_{i=1}^{N} p_i \boldsymbol{w}_i$，其中 $\boldsymbol{w}_i$ 表示第 i 个参与方的模型参数，p_i 表示第 i 个参与方的数据量占全局数据量的比例，$\boldsymbol{w}$ 表示经服务器聚合后的模型参数。通过

$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} p_i F_i(D, \boldsymbol{w})$ 计算可获取全局数据上的机器学习模型参数。经验风险最小化是常用的求解最优参数 $\boldsymbol{w}^*$ 的算法。$F_i(.)$ 表示第 i 个参与方的经验损失函数，通常采用随机梯度下降求解。

另一类参数更新方式为基于模型(Model specialized)的方法，指参与方与服务器交互参数时，不直接更新梯度值，而是依据模型特征设计更新参数，已应于梯度增强决策树 [16]、联邦森林[41]、线性/逻辑回归等模型[42]。Zhao 等[16]提出了联邦学习梯度增强决策树的模型，参与方无需通信在本地数据上训练决策树，将训练好的决策树模型送到下一个参与方，连续训练若干个树模型。文献[43]利用图像中的相似信息通过使用位置敏感哈希建立联邦梯度提升决策树(Gradient Boosting Decision Tree, GBDT)，通过聚集类似实例的梯度值来利用参与方本地数据。

无论上述哪种参数交互方式，参与方之间或参与方与服务器之间的模型参数的交互都必不可少，数据传输也会导致数据隐私的泄露。



(a) 中心化联邦学习架构　　(b) 去中心化的联邦学习架构

图2 联邦学习架构

Fig. 2 Framework of Federated Learning

1.3 联邦学习中的隐私泄露问题

Papernot 等[44]提出了机器学习中的 CIA 安全模型，即机密性(Confidentiality)、完整性(Integrity)及可用性(Availability)。机密性是指未经授权的用户无法获取训练数据、模型参数等信息；完整性指模型的预测结果不能偏离预期；可用性指模型在异常甚至恶意输入的情况下，仍然可以正常使用。本文主要关注 CIA 模型中的机密性。联邦学习中数据无需集中存放，不会产生由大规模数据采集带来的直接数据隐私泄露问题，但在联邦学习中，模型训练阶段及预测阶段可能产生数据隐私泄露的问题，具体表现为：

(1) 在模型训练阶段，不可信服务器可利用参与方上传的参数进行攻击，获取训练数据的敏感信息[45]，或利用接收到的中间参数进行成员推断攻击，推测某条记录是否出现在参与方的敏感训练集中[46]，或获取参与方数据的分布特征后，利用生成模型重构参与方的训练集。

(2) 在模型预测阶段，由于训练模型的泛化能力不足、训练模型简单导致的参与方数据泄露攻击[19]。为了避免向模型训练服务缴费，攻击者通过部分模型结构信息和标签信息，试图获取完整的模型参数信息的模型参数提取攻击[47]。在模型预测阶段，若模型预测结果较敏感，如患某种疾病的概率等，预测结果也可能泄露数据隐私。

2 联邦学习中的隐私攻击

本节从联邦学习模型的机密性保护入手，从敌手能力、攻击目标、攻击方式三方面对联邦学习中的隐私攻击模型进行归纳总结。

2.1 攻击方式

联邦学习中的攻击者包括内部攻击者和外部攻击者。内部攻击者指不可信的服务器或参与方；外部攻击者指模型用户或外部窃听者。从攻击能力来看，可分为黑盒攻击(black-box attack)、白盒攻击(white-box attack)[33]。黑盒攻击指攻击者无法直接获取模型内部参数，但可通过模型的使用观测到输入数据与输出结果，依据获取的“输入-输出”发起推理攻击，通常情况下，模型用户可以发起黑盒攻击。白盒攻击指攻击者能获取训练过程中任一轮的模型中间参数，不可信的服务器和参与方在训练过程中持续交互参数，可发起白盒攻击。外部窃听者通过监听，非法获取服务器与参与方

之间的交互的参数或非法获取模型结果，根据获取数据不同可发起两种类型的攻击。

2.2 攻击目标

破坏机密性的攻击目标主要包括：(1) 获取参与方数据的分布特征或敏感信息[48,49]，利用生成模型重构参与方训练集数据，从训练数据方面破坏了模型的机密性。文献[49]训练了多个参与方联合训练人脸识别的分类模型，参与方的训练集图像是参与方本人的照片，利用模型反演攻击，采用生成模型可以重构该参与者的面部图像。(2) 推测机器学习模型的参数或功能，复制出一个功能相似甚至完全相同的机器学习模型[50]，从模型参数方面破坏模型的机密性。

2.3 攻击模型

联邦学习中的攻击模型总结在表 2 中。推理攻击包括数据泄露攻击、属性推理攻击、成员推断攻击、模型反演攻击。数据泄露攻击易发生在简单线性模型的训练中。机器学习的训练过程通常需要构建经验损失函数，采用随机梯度下降方法找到损失函数的最小值，将最小值对应的参数作为模型参数。在联邦学习中，梯度值一般由学习率(learning rate)和函数微分的乘积构成，如果损失函数过于简单，则发送梯度值大致等同于发送原始数据。此外，若机器学习模型的泛化能力较弱，则也易遭受数据泄露攻击，如递归神经网络(Recursive Neural Network, RNN)具有记忆并暴露训练数据中敏感、特殊模式的缺点。文献[48]中指出，谷歌键盘 Gboard 基于用户的历史文本数据联合学习文本预测模型，从而实现联想词智能提示功能。如果用户的键盘上曾经输入过信用卡号码、身份证号码等具有特殊模式的敏感信息，模型中会以某种方式包含该值，导致数据隐私泄露。

表 2 隐私攻击模型分类

Tab. 2 Types of Privacy Attack Models

攻击名称	描述	攻击者类型	攻击阶段	文献
数据泄露攻击 (Data leakage attack)	数据泄露攻击是指恶意服务器采取训练简单或易攻击模型的手段，通过参与方上传的参数获取参与方的敏感数据或数据特征。	HbC 攻击	预测阶段	Liu 2019 [51]
属性推理攻击 (Attribute inference attack)	攻击者结合辅助数据训练用来攻击模型的分类器，以判断观察到的更新是否为基于“包含目标属性”数据集的训练结果[52]，由此判断目标属性在训练集中的存在状况，获取数据属性信息。	HbC 攻击		Melis 2019 [52]
模型反演攻击 (Model inversion attack)	攻击者试图重构训练数据中特定个体的敏感信息。攻击者反复发送数据并查看预测结果，推测机器学习模型的参数或功能，从而复制出一个功能相似甚至完全相同的机器学习模型。	HbC 攻击	训练阶段	Tramèr 2016 [47] Hitaj 2017 [53], Wang 2019 [50]
成员推断攻击 (Membership inference attack)	对指定的模型和数据样本，攻击者可推断某个样本是否为训练样本。若训练样本均为敏感数据，成员隶属推断攻击将直接给个体带来隐私威胁。	HbC 攻击	训练阶段	Shokri 2017 [46], Hayes 2019[54], Nasr 2019 [55]

成员推断攻击和模型反演攻击在机器学习隐私保护技术中已有研究。Shokri 等[46]首次提出了成员推断攻击，利用训练目标模型影子模型的方式，推断某些数据是否属于训练集。Hayes 等[54]提出了针对生成模型的成员推断攻击。在目标模型生成的样本上训练了生成对抗性网络(Generative adversary network, GAN)，依靠 GAN 对真实记录和合成记录进行分类，可区分样本是否是基于训练集的输入。在联邦学习架构下，不可信服务器通过成员隶属攻击可获取参与方数据的敏感信息。在训练过程中，攻击者通过白盒攻击获取目标模型的多个版本，对多个版本的模型分别进行成员隶属攻击提高攻击成功概率。联邦学习中模型反演攻击威胁更大，个人设备作为参与方，其数据敏感且相似(如，同一个手机端的数据)，经模型反演攻击后得到的数据完全暴露了参与方的敏感信息。文献[49]研究了多个参与方联合训练人脸识别的分类器的问题，每个参与方的训练图像都是参与方本人的照片，利用模型反演攻击与生成模型可以重构该参与者的面部图像。

3 联邦学习中的隐私保护技术

本文依据机器学习/分布式机器学习中的隐私保护技术分类，将联邦学习中的隐私保护分为基于差分隐私的隐私保护技术、基于同态加密的隐私保护技术、基于安全多方计算的隐私保护技术及其他技术，本节对此进行阐述。

3.1 基于差分隐私的隐私保护技术

基于差分隐私的隐私保护技术指向数据中添加噪声达到扰动数据、保护隐私的目的，实现技术主要包括差分隐私

(Differential privacy, DP)[21]、本地化差分隐私(Local differential privacy, LDP)[56]、混洗(shuffle)差分隐私[57]等。

3.1.1 基本概念

差分隐私是建立在严格的数学理论基础之上的强隐私保护模型，能保证攻击者即便在具有最大背景知识的前提下，即已知数据库中除目标记录以外其他所有记录的信息，也无法推测出目标记录的敏感信息。

定义 1：(ε, δ)-差分隐私. 给定任意相邻数据集 D 和 D’，对随机算法 M 及任意输出结果 S，有不等式 $\Pr[\mathrm{M}(\mathrm{D})\in \mathrm{S}]\le \exp(\varepsilon)\times \Pr[\mathrm{M}(\mathrm{D}')\in \mathrm{S}]+\delta$ 成立，则称算法 M 满足(ε, δ)-差分隐私。

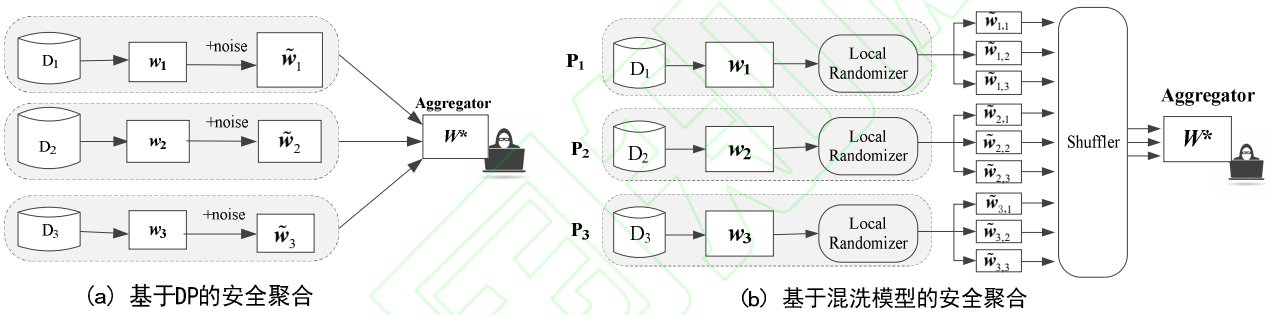
实现差分隐私的机制包括拉普拉斯机制、指数机制[58]、高斯机制[59]等。差分隐私需要有可信的第三方数据收集者，保证所收集的数据不会被窃取和泄露。在实际应用中，第三方数据收集者是否真正可信很难保证。本地化差分隐私将数据隐私化的工作转移到用户端，在数据发出用户设备之前先进行扰动，避免了不可信第三方造成的数据泄露。

定义 2：ε-本地化差分隐私. n 个用户分别持有一条记录，若算法 M 在任意两条记录 t 和 t'上的输出结果满足不等式：$\Pr[\mathrm{M}(t)=t^*]\le \exp(\varepsilon)\times \Pr[\mathrm{M}(t')=t^*]$，则称算法 M 满足 ε-本地化差分隐私。

实现本地化差分隐私的机制主要是随机响应技术、混洗模型(Shuffled model)[57]。混洗模型在本地差分隐私的基础上，增加了一个可信的 shuffler 部件，将用户端发来的数据随机打散后再发给服务器，达到匿名的效果。

3.1.2 实现原理

差分隐私技术在 FL 中应用的原理是：在发布的模型参数中引入一定程度的不确定性噪声，掩盖任何个体用户对训练结果的贡献。在集中式机器学习中，可通过输入扰动、输出扰动、目标扰动及梯度扰动四种方式保护训练数据及模型参数不被泄露；在联邦学习中，数据不集中存放，原始数据无需扰动，隐私保护主要实施在模型训练阶段及模型发布阶段，保护参与方输出的本地模型参数或全局模型参数。



(a) 基于DP的安全聚合　　(b) 基于混洗模型的安全聚合

图3 基于差分隐私的参数安全聚合

Fig. 3 Secure aggregation based on differential privacy

(1) 模型训练阶段的隐私保护

模型训练阶段的隐私保护目的：使攻击者无法获知参与方的本地模型参数，聚合服务器可在扰动后的参数上计算出全局模型参数。在模型训练阶段，其采用的方法大多基于图 3 中展示的两种架构：基于差分隐私的安全聚合及基于混洗差分隐私的安全聚合。

图 3.a 展示了基于差分隐私的安全聚合结构。参与方在本地模型的参数上添加噪声，聚合服务器无法获取参与方的精确参数，研究的关键在于：如何降低噪声添加量，保护隐私的同时保证本地参数的可用性。Wei 等[60]对经验风险最小化后的参数添加高斯噪声。数据扰动的公式表示为：$\tilde{\boldsymbol{w}}_{\mathrm{i}}=\arg\min_{\boldsymbol{w}_{\mathrm{i}}} \mathrm{F}_{\mathrm{i}}(\boldsymbol{w}_{\mathrm{i}},\mathrm{D}_{\mathrm{i}})+\boldsymbol{\eta}_{\mathrm{U}}$，$\boldsymbol{\eta}_{\mathrm{U}}$ 是参与方上传参数时添加的噪声，当满足 $\rho(\boldsymbol{\eta})\propto e^{-\alpha\|\boldsymbol{\eta}\|}$ 时，经验风险最小化的过程满足差分隐私。α 是与隐私预算 ε 及经验风险最小化函数敏感度相关的参数。参与方从服务器端下载参数也需添加噪声。添加噪声的大小取决于函数 $\mathrm{F}_i(.)$ 的敏感度，由于各参与方的函数敏感度不同，取各个参与方函数敏感度的最大值以保证安全。Geyer 等[61]同样使用高斯机制产生噪声数据，与[60]不同之处在于，该方法可防止攻击者识别某个参与方是否参与了训练，而不是只保护参与方中的某条数据。作者提出一种随机化的参数聚合方法，该方法部署在服务器端。在每一轮迭代中，服务器随机选择若干个参与方加入集合 Z_t，模型参数仅发送给 Z_t 中的参与方。Z_t 中的参与方在本地数据上重新训练之后，将参数传给服务器。事先计算出参数更新之和操作的敏感度，再采用高斯机制扰动。Liu 等[62]提出一种分层相关传播算法，在训练神经网络模型时计算每个属性对模型输出的贡献度，针对贡献度确定隐私预算，添加自适应的噪声满足差分隐私。在确定输出层的贡献等于模型输出之后，依次计算其余神经元的贡献 $C_{a_{\mathrm{i}}}^{l_{\mathrm{k}}}(x_{\mathrm{i}})=\sum_{a_{\mathrm{j}}\in l_{\mathrm{k}+1}} C_{a_{\mathrm{i}}\leftarrow a_{\mathrm{j}}}^{l_{\mathrm{k}}\leftarrow l_{\mathrm{k}+1}}(x_{\mathrm{i}})$，通过从数据元组中提取同一属性的贡献，可计算出每个属性类对输出的平均贡献度 $C_{\mathrm{j}}(x_{\mathrm{i}})=\frac{1}{n}\sum_{i=1}^{n} C_{x_{\mathrm{i}},\mathrm{j}}(x_{\mathrm{i}})$，向属性类的贡献度中添加拉普拉斯噪声以保护数据隐私。Hu 等[63]利用差分隐私技术解决参与方计算能力各异，数据结构异质情况下

的隐私保护，提出了个性化联邦学习中的隐私保护问题。同样是在参与方的中间参数中添加高斯噪声，其设置了两个关键参数 $\boldsymbol{W}$ 和 $\boldsymbol{\Omega}$，$\boldsymbol{W}$ 是 m 个参与方的参数向量构成的矩阵，$\boldsymbol{\Omega}$ 为表示各参与方之间参数关系的协方差矩阵，则目标函数可表示为：$\min_{\boldsymbol{W},\boldsymbol{\Omega}} \mathrm{P}(\boldsymbol{W},\boldsymbol{\Omega}) := \sum_{t=1}^{m}\sum_{i=1}^{n_t} l_t(w_t^{\mathrm{T}} x_t^i, y_t^i) + \lambda \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Omega}^{-1}\boldsymbol{W}^{\mathrm{T}})$，求解时迭代多轮直到收敛后可求得最优模型参数。

降低添加的噪声量是基于 DP 的隐私保护方法的研究要点。Liu 等[51]提出了一种基于概要(sketch)数据结构的联邦学习隐私保护方法。Sketch 用少量数据描述全体数据的特征，牺牲了数据描述的准确性，但降低了数据存储及处理代价。sketch 仅描述数据的部分特征，达到同样的 ε-差分隐私在 sketch 上添加噪声量明显小于在原始参数上添加的噪声量。Liu 等[51]利用 Count sketch/Count-Min sketch 结构[64]实现 cross-device 场景下参与方模型更新参数的隐私保护，提出并证明了一个重要的规则：Count-sketch 和 Count-Min 在模型空间明显较大时能实现差分隐私，因此将应用场景放在 cross-device 联邦学习场景下。利用“参与采样+传送参数 sketch”的方法实现了参数的隐私保护，并在线性回归、多层感知模型、循环神经网络模型上进行了实验，结果表明在达到 ε-差分隐私的情况下，通信代价下降到传送原始参数通信代价的 10%。差分隐私还可结合安全多方计算技术减少噪声添加量。经典差分隐私方法需添加方差为 $C^2\sigma^2$ 的高斯噪声以实现隐私保护。假设联邦学习架构中可信成员数为 t，Truex 等[65]采用 SMC 技术将添加的噪声量从 $N(0, C^2\sigma^2)$ 减少到 $N(0, \frac{C^2\sigma^2}{t-1})$。

上述基于 DP 的安全聚集对隐私预算 ε 要求颇高，每一轮迭代所使用的隐私预算满足顺序合成定理(Sequential composition)，所有迭代轮次所用隐私预算为 ε，在迭代轮数不能确定的联邦学习过程中，事先为每一轮迭代分配多少隐私预算难以估计。基于混洗模型的安全聚集可在一定程度上避免上述问题。混洗模型是用来实现本地化差分隐私的一种模型，是 Encode-Shuffle-Analyze(ESA)模型[57]的核心思想。Shuffle 是一个介于客户端和服务器之间的可信部件。Ghazi 等[66]使用混洗模型架构实现了联邦学习中的安全的多方聚合，确保通过添加随机噪声项传递给聚合服务器的单个数字是完全随机的，而总和是一个固定值，通常情况下可为 0。零和噪声的加入不需用户之间的协调。每个本地混淆器(local randomizer)的输出接近于完全随机，对于所有可能输入与真实输入相同的和，可计算出与该输入一致的多种分解形式，从而无法逆推本地混淆器的输出。Shuffle 模型可以“放大”隐私保护度，即使用较小的本地隐私预算，实现全局数据模型上更大的隐私保护度[67]。

(2) 模型发布阶段的隐私保护

模型发布阶段的隐私保护包括模型参数隐私保护与预测结果隐私保护。Hamm 等[68]利用差分隐私技术对联邦学习全局模型参数进行扰动。针对分类模型，采用多数投票的方式确定全局模型的分类结果，在输出全局模型参数上添加符合 $\rho(\boldsymbol{\eta}) \propto e^{-\alpha\|\boldsymbol{\eta}\|}$ 分布的噪声数据，其中 $\alpha = \frac{\lambda\varepsilon}{2}$。Jayaraman 等[69]在联邦学习下对模型训练阶段的扰动和模型发布阶段的扰动进行了对比。提出参与方在安全计算中聚合本地模型，在发布模型之前添加拉普拉斯噪声的隐私保护方法，并证明了该方法的隐私放大效果。实验证明该方法能够实现与未采用隐私保护的模型十分相近的模型可用性。

Triastcyn 等[70]提出利用贝叶斯差分隐私实现模型训练及模型发布时的隐私保护。贝叶斯差分隐私与传统差分隐私的不同之处在于，两个相邻数据集相差一条符合 $\mathrm{p}(x)$ 分布的随机变量记录，而不是一条确定的数据记录。添加符合高斯分布的噪声达到贝叶斯差分隐私，但需要计算每轮迭代的隐私代价，累加各轮隐私代价后计算参数 ε 和 δ 的界限值。在参与方数据分布较相近的情况下，BDP 与传统差分隐私相比，具有显著的优势。

3.1.3 总结与分析

基于差分隐私的隐私保护技术通过添加随机噪声或采用随机应答机制就可实现隐私保护，不会带来额外的计算开销。研究的关键问题主要在于：(1) 依据添加噪声后的数据需进行何种聚集运算，计算运算函数敏感度，量化噪声添加量；(2) 在确保隐私度的前提下设法减少噪声数据的添加量，如结合 SMC 技术、使用特殊的数据结构或引入混洗机制。

基于差分隐私的方法虽然有效，但噪声数据的引入会给模型可用性带来影响，如增加模型收敛的迭代次数、影响运行时间和通讯代价、降低模型预测的精确度等。此外，由于隐私预算的限制，差分隐私处理高维数据后的可用性有待于进一步提高；基于混洗模型的方法需要可信第三方，若参与方中存在恶意用户，混洗模型就无法达到其宣称的隐私保护度。

3.2 基于加密的隐私保护技术

用于联邦学习中的加密技术主要是同态加密技术。

3.2.1 基本概念

同态加密是一种允许用户直接在密文上进行运算的加密形式，得到的结果仍是密文，解密结果与对明文运算的结果一致。即：给定明文数据 x_1 和 x_2，使用同态加密之后的密文分别表示为：$[x_1]$和$[x_2]$，则其同态性可表示为：

(1) 加法：$[x_1]\oplus[x_2]=[x_1\oplus x_2]$；

(2) 乘法：$[x_1]\otimes[x_2]=[x_1\otimes x_2]$。

根据同态加密支持的运算种类和次数，又可分为全同态加密(Fully homomorphic encryption, FHE)[71]、部分同态加密(Partially homomorphic encryption, PHE)及类同态加密(Somewhat homomorphic encryption, SHE)[72]。其中，FHE 支持密文上任意计算的同态性，且不限制计算次数。FHE 虽然足够安全可靠但计算开销太大。PHE 仅支持加法或乘法运算的同态性；SHE 介于上述两者之间，是一种支持有限次加法和乘法运算的加密方法。AHE(Additive Homomorphic Encryption)则仅支持加法运算的同态性。由于同态加密的良好性质，可委托第三方对数据进行处理而不泄露信息。常用的同态加密算法有 Paillier 加密[73]、RSA 加密[74]等。

3.2.2 实现原理

利用同态加密对本地模型参数、数据加密，服务器无法获知参与方的模型参数，也无法获知参与方的原始数据或预测结果，保护了训练阶段及预测阶段的数据隐私。图 4 展示模型训练阶段基于同态加密的参数安全聚合过程。下面分别介绍同态加密用于模型训练阶段及模型预测阶段技术。

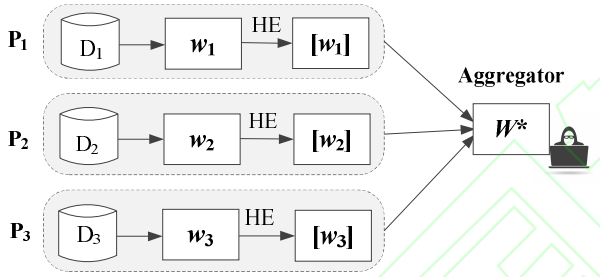


图4 基于同态加密的参数安全聚合

Fig. 4 Secure parameter aggregations on HE

(1) 模型训练阶段的隐私保护

模型训练阶段的隐私保护任务主要是保证训练过程中的中间参数不泄露。Phong 等[75]基于加法同态加密方法(Additive Homomorphic Encryption, AHE)实现了一个保护隐私的深度学习算法 PPDL(Privacy-preserving Deep Learning)。算法分别部署在参与方和聚合服务器上。每个参与方从服务器下载全局加密参数 $\mathrm{E}(\boldsymbol{w}_{\mathrm{global}}^{(\mathrm{j})})$，并用私钥 *sk* 解密后得到权重参数 $\boldsymbol{w}_{\mathrm{global}}^{(\mathrm{j})}$，进而可得权重向量 $\boldsymbol{w}_{\mathrm{global}}$。在本地数据上训练模型，利用 AHE 加密方法将参数加密为 $\mathrm{E}(-\alpha\cdot\boldsymbol{G}^{(\mathrm{i})})$ 后传送给聚合服务器，服务器收到参与方发来的参数后无需解密，计算 $\mathrm{E}(\boldsymbol{w}_{\mathrm{global}}^{(\mathrm{j})})+\mathrm{E}(-\alpha\cdot\boldsymbol{G}^{(\mathrm{i})})$ 更新参数值。该方法通过理论分析与实验，验证了该方法牺牲了效率但不损失模型的精确度。Zhang 等[33]以同态加密和中国余数定理(Chinese Reminder Theorem, CRT)为基础，研究了联邦学习神经网络训练中的隐私保护问题与可验证问题。在数据处理过程中，参与方 $\mathrm{P_i}$ 将神经网络每一层的梯度值 $\boldsymbol{w}_{\mathrm{i}}$ 分成 *r* 份，联合各个分值做线性同余运算，利用 CRT 原理可得到唯一的解，表示为 $\overline{\boldsymbol{w}}_{\mathrm{i}}$，随后利用 $\mathrm{P_i}$ 的私钥对其加密，得到 $[\overline{\boldsymbol{w}}_{\mathrm{i}}]_{\mathrm{pk}}$，$\mathrm{P_i}$ 利用同态哈希函数 *h* 和双线性聚合签名 *x* 计算签名值 $\sigma_{\mathrm{i}}=(\mathrm{h}(\overline{\boldsymbol{w}}_{\mathrm{i}}))^{\mathrm{x}}$，并将加密值和签名同时发送给服务器。服务器收到加密梯度值和签名后，直接在密文上聚合各个参与方上传的参数，得出聚合结果。验证阶段，参与方需要检验服务器是否诚实地聚合了上传的参数，先将参数解密获得 $\overline{\boldsymbol{w}}$，如果公式 $\mathrm{e}(g_1,\sigma)=\mathrm{e}(g_1^{\mathrm{x}},\mathrm{h}(\overline{\boldsymbol{w}}))$ 成立则可验证服务器诚实，其中，e 为双线性映射，g_1 是一个随机生成数。随后，计算 $\overline{\boldsymbol{w}}\bmod m_{\mathrm{i}}$ 得到每一层的梯度值。反复执行上述参数“上传-聚合-下载”过程，直到模型收敛为止。

(2) 预测阶段的隐私保护

预测阶段的隐私保护最早出现在“机器学习即服务(Machine learning as a service, MLaaS)”场景中。数据持有方将数据上传给 MLaaS 服务器，服务器将预测结果返回给数据持有方。在该交互过程中，数据持有者的数据及预测结果都泄露给了 MLaaS 服务器。由于同态加密算法仅对加法及乘法运算有效，非线性运算仍由数据持有方完成，将中间结果加密后发送给云服务器，云服务器将计算结果返回给数据持有方，直到训练完成[76]。显然，这种方法把中间结果暴露给了服务器。Rahulamathavan 等[77]利用 Paillier 加密技术将支持向量机(Support vector machine, SVM)模型的函数及分类样本转换为密文的形式，客户端以加密格式将数据样本发送到服务器。服务器利用同态加密属性直接在加密数据上分类样本。若部分运算不能由同态性质处理，则客户和服务器之间基于安全两方计算协议进行有限次交互。Xie 等[78]提出一种保护隐私预测方法 Crypto-nets，将加密后的数据传送给神经网络模型进行预测，预测结果也同样用加密的方式传给用户，可以保证在模型预测阶段不泄露隐私。由于神经网络模型的函数不是多项式函数，故一个关键问题是如何在密文上利用神经网络模型进行预测。Xie 等提出可以根据 Stone-Weierstrass 定理构造一个逼近神经网络函数的多项式函数，从而可使用同态加密进行预测和输出。

3.2.3 总结与分析

基于加密的隐私强化技术可以达到较高的隐私保护度，既可以保护训练阶段中间参数的隐私不泄露，也可以保证预测阶段的预测结果隐私不泄露。同时，不需多项式逼近的同态加密方法不牺牲模型可用性，但是同态加密需要价高的计算花费及通信代价，且其不支持机器学习中 sigmoid 函数、softmax 函数等非线性运算，需要利用多项式近似表示这些函数，因此在一定程度上造成模型精度的下降。基于同态加密的隐私保护技术计算代价较高，不适于参与方计算能力较差的场景。但在要求较高隐私保护度的场景下，同态加密依然不失为一个最佳选择。

3.3 基于 SMC 的隐私保护技术

安全多方计算(Secure Multiparty Computation, SMC)[23]可使多个参与方以一种安全的方式正确执行分布计算任务，任何一方不能获取其他参与方的额外信息。

3.3.1 基本概念

安全多方计算的原理可描述为：有 n 个参与 $P_1, P_2, \ldots P_n$，每个参与方 P_i 持有一个秘密输入 m_i，在不泄露 m_i 的情况下，n 个参与方可协作计算出函数 $f(m_i)$的值。参与方 P_i 可能是诚实参与方、半诚实参与方或恶意参与方。

多方安全计算的协议众多，在联邦学习中常用的协议有安全两方计算协议与秘密共享协议[79]。Yao 使用混淆电路(Garbled Circuits, GC)技术将计算函数表示为布尔电路[80]，实现了安全两方计算，保证在半诚实模型下的计算安全性。秘密共享协议(Secret Share, SS)包括(t, n)门限秘密共享协议[80]、Blakley 秘密共享协议[81]和中国余数定理。(t, n)门限秘密共享协议是指，用户将某个秘密信息 s 分成 n 份，任意 t ($t \leqslant n$) 份可以重构 s，而任何 t-1 份均无法重构 s。

3.3.2 实现原理

基于 SMC 的隐私保护技术能保护联邦学习模型训练阶段的隐私，但无法保护预测阶段的隐私。Kanagavelu 等[82]提出了一种基于 SMC 的两阶段联邦学习架构。重点保护参与方生成的本地参数 $\boldsymbol{w}_i$，参与方将 $\boldsymbol{w}_i$ 分解为 n 个无意义的值：前 n-1 个值是随机数，第 n 个值通过公式 $\mathbf{V}(i,n)=(\mathbf{V}(i)-\sum_{j=1}^{n-1}\mathbf{V}(i,j)) \bmod Q$ 计算得出。参与方之间互相秘密交换份额，每个参与方持有参数向量的一部分。参与者对秘密份额进行局部聚合，再做全局聚合得到 $\boldsymbol{w}^*$。两轮秘密份额的交换和相加之后，可以消除份额拆分的随机性，即 $\boldsymbol{w}^*=\sum_{i=1}^{n}\mathrm{V}(i)$，该方法的参数聚合机制如图 5 所示。为了解决互相交互秘密份额导致通讯代价过高的问题，采用两阶段联邦学习架构，通过投票方式产生参与方委员会，委员会成员之间进行秘密份额的交换与聚合，产生聚合后的参数。

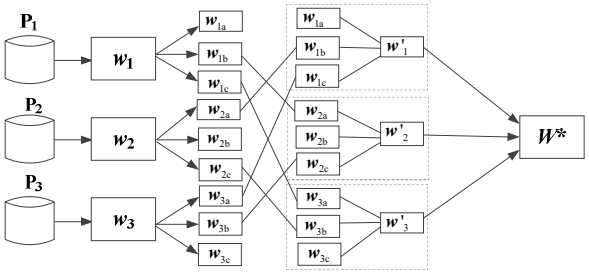


图5 基于 SMC 的参数安全聚合[82]

Fig. 5 Secure parameter aggregation on SMC[82]

Bonawitz 等[83]使用一次性掩码对本地模型参数加密。将 n 个参与方做全序排列，任意一对参与方(u, v)用某个随机向量 $\boldsymbol{s}_{u,v}$ 作为加密参数， 参与方 u 的参数 $\boldsymbol{w}_u$ 与该向量求和，参与方 v 的参数 $\boldsymbol{w}_v$ 就与该向量求差，保证服务器收到的每一对参与方的参数总和不变。但这种方法通信代价太大且容错度较低。为了降低通信代价，Wu 等[84]提出了一种名为 Pivot 的方法。该方法使用 TPHE 和 SMC 的混合框架训练垂直分割数据的树模型。每个客户端在 TPHE 的帮助下执行尽可能多的本地计算以降低通信代价。与之前的结构不同，Pivot 方法需要一个超级参与方协调训练过程。在初始化阶段，参与方确定协作训练某种树模型，并对齐关联样本、确定参数，如密钥、修剪阈值等等。参与方共同生成门限同态加密密钥，接收公钥 pk 和私钥 sk_i。在模型训练阶段，超级参与方广播加密参数协助其他参与方计算加密统计信息。然后，参与方联合将上述加密统计信息转换为 SMC 兼容的输入信息，也就是若干份秘密分享的值。计算当前树节点的最佳分裂方式，并以加密形式表示。整个过程中不会向参与方披露中间信息。获取树模型后，整个树以明文形式发布。内部节点的分割阈值和叶节点上的预测标签以秘密共享的形式出现，参与方不可见，保证不会泄露除预测标签外的任何信息。

3.3.3 总结与分析

基于安全多方计算的联邦学习隐私保护的方法能保证较高的隐私保护度，不需要可信聚合服务器即可完成学习任务，但安全多方计算并非解决联邦学习中隐私问题的杀手锏，这是由于：(1) 基于 SMC 的隐私保护方法的计算代价大、通信轮数多。参与方之间的信息交互造成的通信代价可能成为整个训练过程的瓶颈，基于 SMC 的隐私保护技术的研究目标在于降低系统通信代价。(2) 服务器无法评估通过秘密共享产生的聚合参数是否可用。Bonawitz 等[85]指出有恶意参与方存在的情况下，此类方法无法保证联邦学习模型的可用性。(3) 基于 SMC 的方法仅能对训练过程中的参数进行隐私保护，无法对预测结果进行隐私保护。

3.4 其他方法

近年来，区块链技术的出现也为隐私保护技术提供了新的研究思路。区块链是一个分布式的共享账本和数据库，由于其具备去中心化、不可篡改、全程留痕等优点。联邦学习中参与方众多，分布式记账方式不仅能保证本地模型参数不泄露，还能保证参数聚合过程是可审计的，亦可通过调整激励策略，保证参与方对模型的贡献/收益比是公平的。

基于区块链的隐私保护以分布式事务分类账方法为基础，记录学习任务的参数、参与客户端本地及全局模型的参数更新，单独设置一个聚合器用来聚合参与方更新的参数。更新后参数包装在本地更新事务中，在矿工的协助下记入总账。Awan 等[86]提出了一个基于区块链的隐私保护联邦学习框架，利用区块链的不变性和分散信任属性来保证模型更新的安全。Weng 等[87]提出使用秘密共享协议和区块链技术实

现训练过程中的参数隐私保护。参与方对本地计算的梯度值分别加密并上传，通过秘密分享协议获得更新的参数。协同解密需要至少 t 个参与者提供其秘密分享片段。在梯度值收集过程中，参与者的事务包含加密的梯度值及正确性验证值，允许第三方审核参与方是否上传了正确加密的梯度值。另一方面，矿工通过记录在 DeepChain 中的事务来计算全局参数更新结果。参与方下载全局参数并协同验证。任何第三方都可以审计全局参数值是否正确。此外，DeepChain 提供了一种基于区块链的价值驱动激励机制，迫使参与方正确上传本地参数。

基于区块链的隐私保护技术具备可审计、无需可信结点、安全性高等优点。但区块链技术本身的局限性也限制了其在隐私保护应用领域的应用，如：吞吐量有限、可扩展性差等。因此，在大规模数据的应用场景下，基于区块链的隐私保护方法的有效性一般。

表 3　联邦学习中的隐私保护技术

Tab. 3 Privacy-preserving federated learning techniques comparison

方法类型	方法名	保护模型	防御阶段
基于差分隐私的隐私保护技术	Skeches [51], Stacey[65], Wei[60], Personalized FL[63], APFL [62], Ghazi[66]	神经网络	训练阶段
	Beyesian DP[70] PATE[88]	任意	训练阶段
	Hamm and Cao[68]	回归	预测阶段
	DPGAN[89], DPGM[90]	生成模型	训练/生成阶段
基于同态加密的隐私保护技术	Phong[75], Zhang[32]	神经网络	训练阶段
	Orlandi[91], Barni[76], Rahulamathavan[77], Crypto-nets[78]		预测阶段
基于安全多方计算的隐私保护技术	Renuga[82], Bonawitz[83], HybridAlpha[92], Shokri[93], DeepChain[87]	神经网络	训练阶段
	Pivot[84]	树模型	

4　隐私保护性能衡量标准

依据联邦学习的过程，隐私保护程度可分为计算隐私保护 (Computation Privacy) 和输出隐私保护 (Output Privacy)[65]。所谓计算隐私保护可确保在聚合参与方参数时不会泄露单个参与方的结果；所谓输出隐私保护，指敌手在反复查询模型时，防止敌手推断出训练集中的某条记录或部分数据的信息。依据联邦学习的架构，隐私保护度可分为用户数据隐私保护(Instance Level Privacy)、参与方隐私保护(Client Level Privacy)及联合隐私保护[70]。用户数据隐私保护目的是隐藏单个用户的数据，更具体地说，要限制学习结果分布上的任何单个用户暴露，模型参数的分布不能暴露单条用户数据。参与方隐私保护指参与方上传给服务器的中间参数不会泄露。参与方隐私保护可为用户数据提供额外的保护层，以防不可信服务器获取参与方的数据更新。联合隐私保护指同时达到用户数据隐私保护和参与方隐私保护。

依据联邦学习中的隐私保护方法，评价标准包括隐私保护度、模型可用性、收敛迭代次数和通信代价。其中，隐私保护度的衡量标准主要有隐私泄露率(privacy leakage)及达到的隐私模型，如(ε, δ)-差分隐私、k-匿名等。模型可用性衡量标准包括模型精度、召回率及 F1 分数(F1-Score)。收敛迭代次数指模型收敛时的迭代次数上限。通信代价的主要衡量标准包括传输数据量、算法运行时间等。

基于差分隐私的隐私保护技术采用达到的隐私模型来衡量隐私保护度，训练阶段对模型参数的隐私保护度可以达到(ε, δ)-差分隐私。文献[60]和[63]分别计算了达到(ε, δ)-差分隐私时，添加的高斯噪声参数 σ 的取值。文献[60]计算得出，噪声参数 σ 的取值是聚集次数 T，参与方个数 N 与隐私预算 ε 的某个函数。文献[63]得到类似的结论。文献[51]用泄露隐私概率衡量隐私保护度，即使服务器能完全从 sketch 中恢复参数值，参数隐私泄露的概率不超过 $1/n$，n 是模型参数的维度。在模型精确度方面，差分隐私在训练过程中引入噪声数据，影响模型精确度或训练的迭代次数。文献[63]采用模型收敛迭代次数来衡量噪声数据对模型训练的影响。其定义了一次更新质量的概念，用以衡量每次迭代的下降率，最后计算出模型收敛的迭代次数上限。

基于同态加密的隐私保护技术能达到“不泄露任何信息”的隐私保护度，线性模型的模型精度不受同态加密的影响。但涉及机器学习中 sigmoid/softmax 函数等非线性运算时，需要利用多项式近似表示这些函数，会造成模型精度的下降。基于同态加密的隐私保护技术计算量和通信代价较高。文献[32]、[75]计算出使用 Paillier 加密方法，每轮更新的通信代价是异步 SGD 方法的 2.93 倍；使用基于 LWE(Learning with

errors, LWE)加密方式，每轮通信代价是异步 SGD 方法的 2.4 倍。

基于多方安全计算的隐私保护技术在隐私保护度上可达到与同态加密相同的效果，不泄露任何隐私，但其通信代价较大。为了降低通信代价，HybridAlpha 方法[91]引入了函数加密(Functional Encryption, FE)方法和差分隐私技术降低传输的数据量。基于 SMC 的基准方法的通信量为 $2mn + n$，HybridAlpha 方法将通信代价降低为 $mn + m + n$(n 为参与方数量，m 为聚集服务器数量)。文献[82]先采用 P2P(Point to Point, P2P)的方式选举少量 FL 参与方作为模型聚合委员会成员，参与方和委员会成员交互参数后，再将所有参与方的参数发给服务器，降低了通信代价。单纯的采用 SMC 技术不会影响模型精度，但有些研究工作将 SMC 技术与差分隐私相结合，以求降低通信代价的同时减少噪声量，这种方式则会对模型精度产生影响。

5 总结与展望

联邦学习为构建跨企业、跨数据、跨领域的大数据和人工智能生态圈提供了良好的技术支持。为了进一步强化联邦学习的隐私保护特质，研究者们提出了基于加密、差分隐私、安全多方计算、区块链的隐私保护技术。本文列举了各类技术中的代表性研究工作，如表 3 所示。上述四类隐私保护技术大多是在训练阶段以保护“本地模型参数”为基本任务，防止参与方与服务器在参数交互时泄露数据隐私。多数算法都在公开的数据集(如 MINST、SVHN 等)上进行了实验，评估了模型收敛率、可验证性及通信代价等衡量标准。

随着联邦学习研究的深入与应用领域的拓展，在研究和应用领域仍有一些挑战性问题亟待解决。

(1) 隐私保护技术对联邦学习模型可用性影响的量化研究。在联邦学习中，模型的收敛性还没有理论上的证明，仅有一些研究提供了近似收敛的证明。Li 等[94]研究了 FedAvg 在非独立同分布数据上的收敛性，结果表明，收敛速度与局部迭代的总次数成反比。如果采用差分隐私方法强化联邦学习的隐私保护度，在局部模型的中间参数中加入噪声数据，亦不能保证模型的收敛性。即使模型最终收敛，可添加噪声数据后的模型性能表现不容乐观。有研究表明，在深度学习网络中加入人工噪声后，模型可以收敛，但在 MNIST 数据集上训练分类模型并进行预测时，精度下降了 40%左右[95]。因此，模型的收敛性和预测精度方面还有以下问题需要研究：第一，研究在理论上证明保护隐私的联邦学习模型收敛率的方法；目前的大多数研究在隐私保护处理之后，给出了隐私保护度与模型可用性的实验验证，但未从理论上证明隐私保护处理后的联邦学习模型的收敛问题；即使模型具备收敛性，收敛率和模型性能也需要量化的分析和研究。第二，联邦学习模型的隐私保护度与模型收敛率之间的关系需要进一步的研究；定量衡量联邦学习模型在隐私保护处理之后的精确度、通信代价、经验损失函数的变化等问题也需要深入研究。

(2) 联邦学习架构中隐私保护技术的研究，在经典的隐私保护技术，如差分隐私技术、安全多方计算及加密技术中寻求新的思路。基于差分隐私的保护技术计算量小，隐私保护度较高，但目前该研究领域仍有以下问题有待探索：第一，噪声的添加会导致全局机器学习模型的收敛速度会变慢，模型性能和隐私度是矛盾的[60]。高隐私保护度会造成较低的模型可用性和较慢的模型收敛速度，隐私保护度、模型可用性、模型收敛速度之间均衡的定量关系值得研究；第二，对一定的隐私保护级别，增加参与方数量可能会提高模型收敛速度，但缺乏理论上的证明；对一定的隐私保护级别，存在最佳的聚合时间及通信轮数，也需要定量的研究；第三，当参与方数据非独立同分布时，某一参与方对参数更新贡献较大时，需限制其对全局参数更新的贡献大小，防止其结果影响整个更新。

(3) 联邦生成模型中的隐私保护技术研究。生成模型中也存在隐私泄露的问题，主要原因是生成模型数据集中分布在训练数据点上，且训练样本很容易被记录下来。当生成模型应用到私人数据（如用户面部识别的图像）或敏感数据（如患者医疗记录）上时，会泄露个人敏感信息。目前，已有一些研究针对生成模型进行隐私保护，Xie 等[88]提出了一种满足差分隐私的 GAN 模型 DPGAN。直接发布 Wasserstein 距离相对于训练数据的梯度值会暴露训练集的数据特征，在此梯度值上添加噪声数据保护隐私。Acs 等[90]提出了一种满足差分隐私的基于 k 个神经网络的生成模型 DPGM(Differentially private generative model)，利用随机傅里叶特征将高维数据转换为低维数据，利用一种满足差分隐私的 Lloyd's 算法，将低维数据聚类。在低维数据生成的簇上训练生成模型，在训练过程中使用满足差分隐私的随机梯度下降方法, 噪声值添加到梯度更新中。可见，已有研究工作主要是在训练的梯度值上添加符合高斯分布的噪声实现的，然而，生成模型往往是多层神经网络构成的，结构非常复杂，噪声的添加会影响生成模型的精确度。另外，为了生成更复杂的数据，例如个人照片或各种序列数据，还需要对具有多个隐藏层的深度神经网络进行有效的隐私保护训练，有很多内容值得深入研究。

(4) 联邦学习中的参与方隐私异质性与模型可用性研究。联邦学习的一个重要优势在于可在参与方的数据格式各异、计算能力各异的情况下，协同多个参与方联合训练机器学习模型。目前的隐私保护技术可以保证结构各异的参与方达到相同的数据隐私保护度，最终获取相同的模型参数[63]。然而，各个参与方对隐私保护度与模型可用性的需求可能各不相同，有些参与方希望牺牲一些数据隐私换取更好的模型性能，而有些参与方刚好相反。目前鲜有研究考虑联邦学习系统中的“隐私异质性”，在差分隐私背景下，可以给参与方分配不同的隐私预算，初步解决隐私异质性问题。然而，

笔者认为，该问题的关键在于模型参数的聚合策略，可设计智能的模型参数聚合策略区分参与方对隐私保护度和模型性能的个性化需求。

(5) 隐私保护度、通信代价、模型精确度之间的权衡，建立统一的隐私保护度与模型可用性衡量标准。已有的研究方法在隐私保护度、通信代价、模型精确度上难以兼得。或者牺牲隐私保护度换取模型精度；或牺牲通信代价换取隐私保护度。从研究者角度来看，需要定义一个统一的衡量指标体系，综合考虑隐私保护度、模型精度、通信代价及计算开销。统一的隐私保护度与模型可用性衡量标准可为各种研究方案的对比奠定基础。

(6) 针对 cross-device 场景下离线客户端对隐私保护度的影响，研究隐私保护技术对系统稳定的鲁棒性。上述隐私保护算法假设所有参与方在每一轮参数交互过程中都可以连接到服务器，不存在无法连接的情况。当参与方数量较多时，如 cross-device 场景下，一些客户端会由于网络连接中断或其他原因暂时无法连接到服务器。若采用差分隐私添加噪声的形式实现参数的隐私保护，则客户端的退出会导致添加的噪声太少，无法达到要求差分隐私的隐私保护度。一种保守的方法是增加每个客户端的噪声量，即便存在一定比例的离线客户端，剩余客户端在进行安全参数聚合时仍能达到差分隐私的隐私保护度。但是当客户端没有掉线时，易产生大量的额外噪声，导致模型精度下降。挑战性问题在于如何处理大规模客户端参与的联邦训练模型，且能保证隐私保护技术在系统不稳定时的系统鲁棒性。

(7) 面向应用领域的联邦学习隐私保护新技术研究。由于数据隐私策略与数据孤岛问题的产生，联邦学习在未来的发展中，应用领域将越来越广泛。诸如医疗大数据、财经大数据、个人移动设备大数据等涉及敏感数据的领域都可能使用联邦学习联合训练模型。联邦学习架构不同，其隐私保护技术所采用的方法也可能不尽相同。在 cross-device 的应用场景中，更注重个性化隐私保护，研究异质性隐私；在 cross-silo 场景下，可信服务器可能难以完全可信，如何设计无可信服务器的隐私保护技术、研究在去中心化联邦学习架构中的隐私保护方案是挑战性问题。此外，基于应用领域的不同需求，一些安全领域的技术如机密计算等、可信执行环境等与隐私保护技术的结合也是一个值得研究的问题。

6 结语

作为人工智能的重要分支，机器学习和联邦学习技术已经成为处理大数据不可或缺的技术手段。而人工智能领域中的伦理问题向来受到全社会的关注。数据隐私保护是人工智能面临的重要伦理问题之一，已经成为《人工智能道德准则》[96]的组成部分。数据隐私保护技术的解决方案通常包括加密、泛化、扰动等途径，为了适应联邦学习或机器学习模型中强大的攻击能力，数据隐私保护技术可能需要更强的隐私保护模型或者结合几种隐私保护技术，设计出轻量级的隐私保护算法，在技术上探讨机器学习/联邦学习架构中的隐私保护技术，使得机器学习/联邦学习模型的机密性、完整性、可用性三个标准完美均衡；另一方面，需要制定适当的法律法规与政策引导，技术和法规的有机结合可作为解决机器学习隐私与伦理问题的新探索。除了数据隐私与安全，人工智能中的伦理问题还包括数据透明、算法的多样性、非歧视性和公平性等其他重要部分，同样需要技术上的深入研究。

参考文献

[1] ZHOU L, PAN S, WANG J, et al. Machine learning on big data: opportunities and challenges [J]. Neurocomputing, 2017, 237(10): 350-361.

[2] 中华人民共和国网络安全法[EB/OL]. [2016-11-07]. http://www.cac.gov.cn/2016-11/07/c_1119867116.htm (Network security law of the People's Republic of China [EB/OL]. [2016-11-07]. http://www.cac.gov.cn/2016-11/07/c_1119867116.htm)

[3] 中华人民共和国数据安全法[EB/OL]. [2021-6-10]. https://baijiahao.baidu.com/s?id=1702265632126727684&wfr=spider&for=pc (Data security law of the People's Republic of China [EB/OL]. [2021-6-10]. https://baijiahao.baidu.com/s?id=1702265632126727684&wfr=spider&for=pcs)

[4] 中华人民共和国个人信息保护法[EB/OL]. [2020-10-22]. http://fzzfyjy.cupl.edu.cn/info/1077/12335.htm (Law of the People's Republic of China on the protection of personal information[EB/OL]. [2020-10-22]. http://fzzfyjy.cupl.edu.cn/info/1077/12335.htm)

[5] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 12:1-12:19.

[6] KONECNY J, MCMAHAN H B, YU F X, et, al. Federated learning: strategies for improving communication efficiency [EB/OL]. [2017-10-30]. CoRR abs/1610.05492.

[7] LIU B Y, Wang L J, LIU M, et al. Federated imitation learning: a privacy considered imitation learning framework for cloud robotic systems with heterogeneous sensor data [EB/OL]. [2019-9-15]. CoRR abs/1909.00895.

[8] SHAO R L, HE H Y, LIU H, et al. Stochastic channel-based federated learning for medical data privacy preserving [EB/OL]. [2019-11-15] CoRR abs/1910.11160.

[9] HAHN S, LEE J. Privacy-preserving federated bayesian learning of a generative model for imbalanced classification of clinical data [EB/OL] [2020-8-29]. CoRR abs/1910.08489.

[10] CHOUDHURY O, GKOULALAS D A, SALONIDIS T, et al. Differential privacy-enabled federated learning for sensitive health data [EB/OL] [2020-2-27]. CoRR abs/1910.02578.

[11] ZHENG W T, POPA R A, GONZALEZ J E, et al. Maliciously secure coopetitive learning for linear models[C].// SP2019: 2019 IEEE Symposium on Security and Privacy. IEEE, 2019: 724-738.

[12] LI Q B, WEN Z Y, WU Z M, et al. Federated learning systems: vision, hype and reality for data privacy and protection [EB/OL]. [2021-12-5]. CoRR abs/1907.09693.

[13] LI Z P, SHARMA V, MOHANTY S P. Preserving data privacy via federated learning: challenges and solutions [J]. IEEE Consumer Electronics Magazine, 2020, 9(3):8-16.

[14] SHAO R L, LIU H, LIU D B. Privacy preserving stochastic channel-based federated learning with neural network pruning [EB/OL]. [2019-10-04]. CoRR abs/1910.02115.

[15] SATTLER F, MULLER K R, SAMEK W. Clustered federated learning: model-agnostic distributed multi-task optimization under privacy constraints [EB/OL]. [2019-10-04]. CoRR abs/1910.01991.

[16] ZHAO L C, NI L H, HU S S, et al. InPrivate digging: enabling tree-based distributed data mining with differential privacy [C].// INFOCOM2018: the 2018 IEEE Conference on Computer Communications. IEEE, 2018: 2087–2095.

[17] CHENG K W, FAN T, JIN Y L, et al. Secureboost: a lossless federated learning framework [EB/OL]. [2021-04-07]. arXiv:1901.08755v3.

[18] MCMAHAN H B, MOORE E, RAMAGE D, et al. federated learning of deep networks using model averaging [EB/OL]. [2017-02-28]. arXiv:1602.05629v3, 2017.

[19] HAO M, LI H W, XU G W, et al. Towards efficient and privacy-preserving federated deep learning [C]. // ICC 2019: the 2019 IEEE International Conference on Communications. IEEE, 2019: 1-6.

[20] 刘俊旭,孟小峰.机器学习的隐私保护研究综述[J].计算机研究与发展,2020,57(02):346-362. (LIU J X, MENG X F. Survey on privacy-preserving machine learning [J]. Journal of Computer Research and Development, 2020, 57(02):346-362.)

[21] DWORK C. Differential privacy [C]. // ICALP 2006: Automata, Languages and Programming, 33rd International Colloquium. Springer, 2006: 1-12.

[22] SWEENEY L. k-anonymity: A model for protecting privacy [J]. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10 (5): 557-570.

[23] BOGETOFT P, CHRISTENSEN L D, DAMGARD I, et al. Secure multiparty computation goes live [C]. // FC 2009: International Conference on Financial Cryptography and Data Security. Springer, 2009: 325-343.

[24] 周纯毅,陈大卫,王尚,付安民,高艳松.分布式深度学习隐私与安全攻击研究进展与挑战[J].计算机研究与发展, 2021, 58(05): 927-943. (ZHOU C Y, CHEN D W, SHANG W, et al. Research and challenge of distributed deep learning privacy and security attack. Journal of Computer Research and Development [J]. 2021, 58(05):927-943.)

[25] 陈兵,成翔,张佳乐,谢袁源.联邦学习安全与隐私保护综述 [J]. 南京航空航天大学学报, 2020, 52(05): 675-684. (CHEN B，CHENG X，ZHANG J L，et al. Survey of security and privacy in federated learning [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 675-684.)

[26] 周俊,方国英,吴楠.联邦学习安全与隐私保护研究综述[J].西华大学学报(自然科学版),2020,39(04): 9-17. (ZHOU J, FANG G Y, WU N. Survey on security and privacy-preserving in federated learning [J]. Journal of Xihua University (Natural Science Edition), 2020, 39(4): 9-17.)

[27] 王健宗,孔令炜,黄章成,陈霖捷,刘懿,卢春曦,肖京.联邦学习隐私保护研究进展[J].大数据,2021,7(03): 130-149. (WANG J Z, KONG L W, CHEN L C, et al. Research advances on privacy protection of federated learning [J]. Big Data Research, 2021,7(03): 130-149.)

[28] 纪守领,杜天宇,李进锋,沈超,李博.机器学习模型安全与隐私研究综述[J].软件学报,2021,32(01):41-67. (JI S L, DU T Y, LI J F, et al. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1):41-67.)

[29] YIN X F, ZHU Y M, HU J K. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions [J]. ACM Computing Surveys, 2021, 54(6): 131:1-131:36.

[30] BRIGGERS C, FAN Z, ANDRAS P. A review of privacy-preserving federated learning for the Internet-of-Things [EB/OL]. Federated Learning Systems, https://doi.org/10.1007/978-3-030-70604-3_2.

[31] BRANDEIS L, WARREN S. The right to privacy [M]. Harvard Law Review, 1890, 4(5): 193-220.

[32] Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46[Z]. Official Journal of the European Union (OJ), 2016, 59(1-88): 294.

[33] ZHANG X L, FU A M, WANG H Q, et al. A privacy-preserving and verifiable federated learning scheme[C]. // ICC 2020: Proceedings of the 2020 IEEE International Conference on Communications. Piscataway: IEEE, 2020: 1-6.

[34] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning [J]. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210.

[35] LI T, SANJABI M, BEIRAMI A, et al. Fair resource allocation in federated learning [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1905.10497.pdf.

[36] LI Q, WEN Z Y, WU Z M, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1907.09693.pdf.

[37] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. Lauderdale: PMLR, 2017: 1273-1282.

[38] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: distributed machine learning for on-device intelligence [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1610.02527.pdf.

[39] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks [J]. Proceedings of Machine Learning and Systems, 2020, 2: 429-450.

[40] WANG J Y, LIU Q H, LIANG H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization [EB/OL]. [2021-01-09]. CoRR abs/2007.07481.

[41] LIU Y, LIU Y T, LIU Z J, et al. Federated forest [EB/OL]. [2020-10-09]. CoRR abs/1905.10053.

[42] NIKOLAENKO V, WEINSBERG U, IOANNIDIS S, et al. Privacy-preserving ridge regression on hundreds of millions of records [C]. // S&P 2013: 2013 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2013: 334-348.

[43] LI Q B, WEN Z Y, HE B S. Practical federated gradient boosting decision trees[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2020, 34(04): 4642-4649.

[44] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1611.03814.pdf.

[45] MA C, LI J, DING M, et al. On safeguarding privacy and security in the framework of federated learning [J]. IEEE network, 2020, 34(4): 242-248.

[46] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models [C] //S&P 2017: 2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2017: 3-18.

[47] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[C]//25th USENIX Security Symposium. Berkeley: USENIX, 2016: 601-618.

[48] CARLINI N, LIU C, ERLINGSSON Ú, et al. The secret sharer: evaluating and testing unintended memorization in neural networks [C] // 28th USENIX Security Symposium. Berkeley: USENIX, 2019: 267-284.

[49] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[50] WANG Z B, SONG M K, ZHANG Z F, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C] // IEEE INFOCOM 2019: IEEE Conference on Computer Communications. Piscataway: IEEE, 2019: 2512-2520.

[51] LIU Z X, LI T, SMITH V, et al. Enhancing the privacy of federated learning with sketching [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1911.01812.pdf.

[52] MELIS L, SONG C, CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]//S&P 2019: 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2019: 691-706.

[53] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 603-618.

[54] HAYES J, MELIS L, DANEZIS G, et al. Logan: membership inference attacks against generative models[C]//Proceedings on Privacy Enhancing Technologies. Poland: De Gruyter, 2019, 2019(1): 133-152.

[55] 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]//S&P 2019: 2019 IEEE symposium on security and privacy. Piscataway: IEEE, 2019: 739-753.

[56] WANG N, XIAO X K, YANG Y, et al. Collecting and analyzing multidimensional data with local differential privacy[C]//2019 IEEE 35th International Conference on Data Engineering. Piscataway: IEEE, 2019: 638-649.

[57] BITTAU A, ERLINGSSON Ú, MANIATIS P, et al. Prochlo: strong privacy for analytics in the crowd[C]//Proceedings of the 26th Symposium on Operating Systems Principles. New York: ACM, 2017: 441-459.

[58] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//48th Annual IEEE Symposium on Foundations of Computer Science. Piscataway: IEEE, 2007: 94-103.

[59] DWORK C, ROTH A, et al. The algorithmic foundations of differential privacy [J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407.

[60] WEI K, LI J, DING M, et al. Federated learning with differential privacy: algorithms and performance analysis [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[61] GRYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1712.07557.pdf.

[62] LIU X Y, LI H W, Xu G W, et al. Adaptive privacy-preserving federated learning [J]. Peer -to-Peer Networking and Applications, 2020, 13(6): 2356-2366.

[63] HU R, GUO Y, LI H, et al. Personalized federated learning with differential privacy [J]. IEEE Internet of Things Journal, 2020, 7(10): 9530-9539.

[64] MELIS L, DANEZIS G, DE CRISTOFARO E. Efficient private statistics with succinct sketches [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1508.06110.pdf.

[65] TRUEX S, BARACALDO N, ANWAR A, et al. A hybrid approach to privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 1-11.

[66] GHAZI B, PAGH R, VELINGKER A. Scalable and differentially private distributed aggregation in the shuffled model [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1906.08320.pdf.

[67] BALLE B, BELL J, GASCÓN A, et al. The privacy blanket of the shuffle model[C]//Annual International Cryptology Conference. Cham: Springer, 2019: 638-667.

[68] HAMM J, CAO Y, BELKIN M. Learning privately from multiparty data[C]//International Conference on Machine Learning. Lauderdale: PMLR, 2016: 555-563.

[69] JAYARAMAN B, WANG L. Distributed learning without distress: privacy-preserving empirical risk minimization [J]. Advances in Neural Information Processing Systems, 2018.

[70] TRIASTCYN A, FALTINGS B. Federated learning with bayesian differential privacy[C]//2019 IEEE International Conference on Big Data. Piscataway: IEEE, 2019: 2587-2596.

[71] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//Proceedings of the forty-first annual ACM symposium on Theory of computing. New York: ACM, 2009: 169-178.

[72] ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: theory and implementation [J]. ACM Computing Surveys , 2018, 51(4): 1-35.

[73] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[C]//International conference on the theory and applications of cryptographic techniques. Berlin: Springer, 1999: 223-238.

[74] RIVEST R L, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms [J]. Foundations of Secure Computation. 1978: 169–180.

[75] PHONG T L, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345.

[76] BARNI M, ORLANDI C, PIVA A. A privacy-preserving protocol for neural-network-based computation[C]//Proceedings of the 8th workshop on Multimedia and security. New York: ACM, 2006: 146-151.

[77] RAHULAMATHAVAN Y, PHAN R C W, VELURU S, et al. Privacy-preserving multi-class support vector machine for outsourcing the data classification in cloud [J]. IEEE Transactions on Dependable and Secure Computing, 2013, 11(5): 467-479.

[78] XIE P T, BILENKO M, FINLEY T, et al. Crypto-nets: neural networks over encrypted data [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1412.6181v1.pdf.

[79] RIVEST R L, SHAMIR A, TAUMAN Y. How to share a secret[C]//Communications of the ACM. New York: ACM, 1979, 22(11): 612-613.

[80] YAO A C. Protocols for secure computations[C]//23rd annual symposium on foundations of computer science. Piscataway: IEEE, 1982: 160-164.

[81] BLAKLEY G R. Safeguarding cryptographic keys[C]//Managing Requirements Knowledge, International Workshop on. Los Alamitos: IEEE Computer Society, 1979: 313-317.

[82] KANAGAVELU R, LI Z, SAMSUDIN J, et al. Two-phase multi-party computation enabled privacy-preserving federated learning[C]//2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing. Piscataway: IEEE, 2020: 410-419.

[83] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[84] WU Y C, CAI S F, XIAO X K, et al. Privacy preserving vertical federated learning for tree-based models [EB/OL]. [2022-01-19]. http://export.arxiv.org/pdf/2008.06170.

[85] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[86] AWAN S, LI F, LUO B, et al. A reliable and accountable privacy-preserving federated learning framework using the BlockChain [C]. // CCS 2019: the 2019 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2019: 2561-2563.

[87] WENG J S, WENG J, ZHANG J L, et al. Deepchain: auditable and privacy-preserving deep learning with blockchain-based incentive [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(5): 2438-2455.
[88] PAPERNOT N, SONG S, MIRONOV I, et al. Scalable private learning with pate [EB/OL]. [2022-01-19]. http://export.arxiv.org/pdf/1802.08908.
[89] XIE L Y, LIN K X, WANG S, et al. Differentially private generative adversarial network [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1802.06739.pdf.
[90] ACS G, MELIS L, CASTELLUCCIA C, et al. Differentially private mixture of generative neural networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(6): 1109-1121.
[91] ORLANDI C, PIVA A, BARNI M. Oblivious neural network computing via homomorphic encryption [J]. EURASIP Journal on Information Security, 2007, 2007: 1-11.
[92] XU R H, BARACALDO N, ZHOU Y, et al. Hybridalpha: an efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 13-23.
[93] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. New York: ACM, 2015: 1310-1321.
[94] LI X, HUANG K X, YANG W H, et al. On the convergence of fedavg on non-iid data [EB/OL]. [2022-01-19]. https://arxiv.org/pdf/1907.02189.pdf.
[95] AÏVODJI U M, GAMBS S, MARTIN A. IOTFLA: a secured and privacy-preserving smart home architecture implementing federated learning [J]. IEEE Symposium on Security and Privacy Workshops, 2019: 175-180.
[96] Ethics Guidelines for Trustworthy AI [EB/OL]. [2021-01-03] https://www.i-programmer.info/programming/artificial-intelligence/12702-ethics-guidelines-for-trustworthy-ai-.html.

This work is partially supported by Natural Science Foundation of China (62002098), Natural Science Foundation of Hebei Province (F2020207001, F2021207005).

WANG Teng, born in 1980, Ph. D., senior engineer. His research interests include machine learning, digital governance.
HUO Zheng, born in 1982, Ph. D., associate professor. Her research interests include privacy-preserving, federated learning.
HUANG Yaxin, born in 1999, M. S. candidate. His research interests include privacy-preserving.
FAN Yilin, born in 1998. Her research interests include federated learning.