联邦学习中的拜占庭攻防研究综述

赵晓洁[1]，时金桥[1]，黄梅[1]，柯镇涵[1]，申立艳[2]
（1. 北京邮电大学网络空间安全学院，北京 100088；2. 北京信息科技大学计算机学院，北京 100192）

摘　要：联邦学习作为新兴的分布式机器学习解决了数据孤岛问题。然而，由于大规模、分布式特性以及本地客户端的强自主性，使得联邦学习极易遭受拜占庭攻击且攻击不易发现，这严重破坏了模型的完整性和可用性等。首先，以拜占庭攻击为研究对象，对攻击原理进行细化分类与剖析。其次，以经典的网络安全防御模型为指导，从防御机制的角度针对联邦学习防御方法进行分类和分析。最后，提出了联邦学习抗拜占庭攻击需要解决的关键问题和研究挑战，为未来相关研究者提供了新的参考。

关键词：联邦学习；拜占庭攻击；防御方法；攻防策略
中图分类号：TP181；TP309
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024208

Survey on Byzantine attacks and defenses in federated learning

ZHAO Xiaojie[1], SHI Jinqiao[1], HUANG Mei[1], KE Zhenhan[1], SHEN Liyan[2]
1. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100088, China
2. School of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China

Abstract: Federated learning as an emerging distributed machine learning, can solve the problem of data islands. However, due to the large-scale, distributed nature and strong autonomy of local clients, federated learning is extremely vulnerable to Byzantine attacks and the attacks are not easy to detect, which seriously damages the integrity and availability of the model. First, taking Byzantine attacks as the research object, a detailed classification and analysis of the attack principles were conducted. Secondly, guided by the classic network security defense model, federated learning defense methods were classified and analyzed from the perspective of defense mechanisms. Finally, the key issues and research challenges that need to be solved in federated learning to resist Byzantine attacks were proposed, providing new references for future relevant researchers.

Keywords: federated learning, Byzantine attack, defense method, attack and defense strategy

0　引言

2016年，谷歌提出了联邦学习（FL, federated learning）[1]，通过分布式训练的方式，在不共享用户原始数据的情况下进行模型的训练和优化，解决了数据孤岛问题，并有效保护了用户隐私。近年来，联邦学习已被广泛应用于用户行为分析[2-3]、智慧医疗[4]、无线通信[5-6]、信号识别[7]以及安全检测[8-9]等领域，并提出了工业级联邦学习框架（FATE, fede-rated AI technology enabler)[10]、隐私保护通用框架（Pysyft）[11]、共同协作的开源框架（PaddleFL, paddle federated learning) [12]、开源联邦学习框架（TFF, tensorflow federated）[13]等。虽然联邦学习已经在机器学习领域得到广泛的应用，但

收稿日期：2024–05–27；修回日期：2024–11–15
通信作者：时金桥，shijinqiao@bupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62302057）；国家重点研发计划基金资助项目（No.2023YFB3106400）
Foundation Items：The National Natural Science Foundation of China (No.62302057), The National Key Research and Development Program of China (No.2023YFB3106400)

由于其大规模、分布式的特性以及参与训练客户端的强自主性，联邦学习在训练过程易受到拜占庭攻击的影响[14-27]。攻击者可以通过修改本地训练数据或者上传的梯度，来达到降低模型准确率，影响特定样本预测结果等目的。拜占庭攻击极易实施且不易发现，是联邦学习面临的最主要攻击威胁之一。例如，Chen 等[28]研究发现仅单个拜占庭攻击者，也可能导致全局模型训练失效。

目前，国内外已有许多联邦学习中的攻击与防御相关的研究。例如，He 等[29]分析了深度学习中安全威胁的 4 种攻击，但在联邦学习的模型训练中难以达到相同的攻击效果。杨丽等[30]从数据的机密性和模型的完整性 2 个方面进行重点研究，其侧重于隐私泄露和模型窃取攻击。高莹等[31]设计了一种独特的防御分类方法，将防御手段划分为鲁棒性提升方法和隐私性增强技术两类，但他们未进一步考虑整个联邦学习训练模型全流程的安全防御。文献[32-34]侧重于联邦学习中的隐私攻防，仅涉及少数拜占庭攻防方法的研究。Wan 等[35]侧重研究无线联邦学习中的攻击和防御机制。Roszel 等[36]分析了经典的拜占庭容忍聚合方法的有效性和局限性。孙钰等[37]针对联邦学习拜占庭攻防展开详细描述，将拜占庭防御策略分为基于梯度统计特性、基于梯度间距离、基于额外验证数据、基于优化算法补偿和基于差分隐私 5 类，强调了拜占庭攻防博弈是联邦学习安全研究的重要问题，但也未进一步考虑整个联邦学习攻防框架。

对比以上这些综述，本文以拜占庭攻击为研究对象，分析了拜占庭攻击的威胁模型，对攻击原理进行细化分类与剖析，增加了联邦学习中拜占庭攻击和防御的最新研究成果的描述。在实际联邦学习应用场景中，仅仅针对一种具体攻击方法进行防御远远无法保障联邦学习的安全性，亟须开展系统性的攻防对抗研究。为此，本文从攻防对抗的视角，以经典的网络安全防御模型，包括风险分析、安全策略、系统防护、攻击检测、实时响应和灾难恢复（APPDRR）为指导，进一步细化了联邦学习训练模型全周期防护，包括安全评估、安全策略、安全防护、恶意检测、异常响应和模型恢复等各个环节。从拜占庭攻防机制的角度，进一步扩展了联邦学习中的防御分类。具体来说，在安全防护方面，划分为客户端自身防护、过滤和参数空间稀疏化；在恶意检测方面，增加了基于参数方向间、组合特征和相对分布相似性的防御分类方法；在异常响应环节，分为直接剔除恶意更新、基于信誉规避异常和基于投票规避异常。此外，并详细分析了现有防御方法的优势与局限性。基于上述防御方法分析，尽管目前已有许多关于联邦学习攻防的研究，但主要集中在恶意检测和异常响应环节。而联邦学习模型训练事前部署和事后恢复，如安全评估、安全策略、安全防护和模型恢复研究较少。因此，在实际联邦学习应用中，需要考虑事前部署、事中防护和事后恢复更全面的安全防御系统。此外，从攻防对抗视角，当前提假设条件变动时，研究鲁棒性强、隐蔽性强和更持久的盲攻击将为未来相关研究者提供新的参考，为基于联邦学习中的拜占庭攻防提供更加全面的理论基础与研究框架。

1 拜占庭攻击

拜占庭攻击[38]是指攻击者主动或被动通过控制或操纵一个或者多个本地客户端篡改客户端持有的数据集或者在训练过程中上传的梯度参数等方式对全局模型进行攻击以实现降低模型性能和注入后门等目标，如图 1 所示。在攻击过程中，攻击者可以结合攻击时机[39]、攻击强度[22]和协同方式[33]等制定复杂的攻击策略。在此声明，本文考虑的拜占庭攻击主要是以投毒方式[32]造成的攻击，不考虑通过攻击硬件设备和通信网络造成的故障等[40]。

1.1 威胁模型

在联邦学习的实际场景中，拜占庭攻击的威胁模型包括攻击者的目标、攻击者的背景知识和能力假设。

拜占庭攻击者的目标是降低全局模型的性能，根据攻击是否针对特定的类别标签，分为定向攻击[14]和非定向攻击[23]。定向攻击是指损坏全局模型在特定类别预测时的准确率且其余类别预测任务几乎不受影响。攻击者可以通过翻转数据标签、植入特定的触发器等方式来实现攻击，其中对数据集以一种隐蔽的方式植入触发器方式称为后门攻击[22, 41]。非定向攻击是指损坏全局模型对所有目标的准确率，使其无法收敛或模型整体性能下降。相比于定向攻击，非定向攻击不针对任何特定的类别或数据样本。

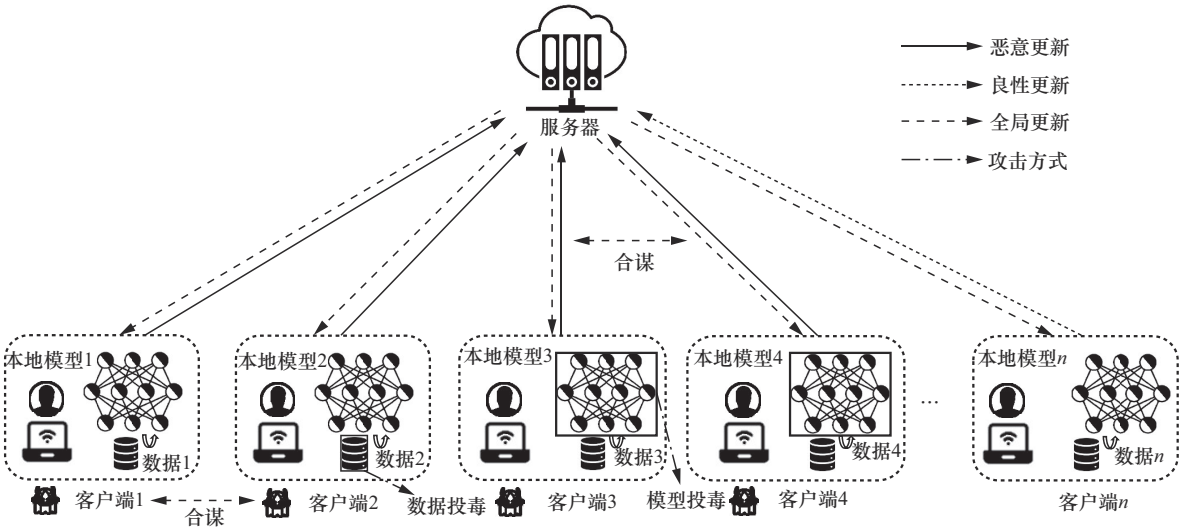

图1　拜占庭攻击

攻击者的背景知识是指攻击者是否了解服务器采用的安全聚合算法和本地训练数据集和模型[18]。攻击者的能力假设是指对其受损客户端的控制权以及对发送到服务器上的模型参数的操纵权限[18]，主要包括操纵恶意客户端的数量占比[42]、操纵恶意客户端实施攻击的能力。攻击者可以控制恶意客户端的数量直接影响攻击的规模和强度。攻击者可以操纵恶意客户端，通过修改本地训练数据、篡改本地模型参数、选择恶意的训练算法等方式，使其产生具有破坏性的梯度更新，还可以调整恶意客户端的攻击强度、选择恶意更新的类型和分布、调整攻击的时机[43]等，以适应不同的攻击目标和场景。

1.2　攻击方式

联邦学习的拜占庭攻击根据攻击方式通常可分为数据投毒[44]和模型投毒[18]。数据投毒和模型投毒都会对联邦学习中的全局模型训练产生影响，主要原因为数据投毒本质上与模型投毒都会修改局部模型的更新权重[37]。然而，不同之处在于，前者专注于对本地客户端数据集进行攻击，涉及对数据集标签翻转、添加触发器等；后者专注于对机器学习模型本身进行攻击，涉及局部模型梯度、局部模型参数等。

1.2.1　数据投毒

数据投毒[44]是指攻击者部分或者完全破坏用户的训练数据，通过恶意修改数据类别、添加触发器等攻击手段来降低模型的性能。

数据投毒中最常见的攻击为标签翻转攻击[39,42]和后门攻击[24, 43]。标签翻转攻击是通过修改部分参与者的训练数据的类别标签的方式，来污染全局模型的参数。Tolpegin 等[16]通过大量实验表明，即使只有少量的恶意参与者，标签翻转攻击也能显著降低全局模型的准确率和召回率，尤其是对于被攻击的源类别。Xu 等[43]提出分布式后门攻击（DBA, distributed backdoor attack），对恶意客户端的数据都添加触发器，使得聚类中心在每次迭代过程中发生微小的偏移，从而扰乱模型的训练过程，进而影响模型把特定样本预测为目标类。由于这些攻击仅针对训练数据，它们的成功取决于足够数量的攻击回合以避免后门攻击效果减弱[45]。数据投毒在实际应用中对攻击者的能力要求较低，它只需要攻击者能够控制参与方的训练数据，具有广泛的实施场景。

1.2.2　模型投毒

模型投毒[17]是指攻击者部分或者完全控制用户的训练阶段，通过对上传的局部参数精心篡改破坏模型的训练过程，实现对全局模型的操纵。

1) 简单模型投毒

简单模型投毒[46]是指通过上传任意梯度来破坏全局模型的可用性，如非定向攻击。具体来说，攻击者可以篡改用户上传给服务器的梯度w_i，篡改后的梯度称为恶意或异常梯度，如式(1)所示。

$$w_i'=\begin{cases}\nabla f\left(w_i;\xi_i\right), i\text{是正常用户}\\ v, i\text{是拜占庭用户}\end{cases}\tag{1}$$

其中，ξ_i、∇f和v分别代表训练数据、局部损失函

数的梯度和任意梯度。

2) 目标优化的模型投毒

目标优化的模型投毒是指通常将模型的参数构造问题建模为优化问题，并通过最优化算法寻找最佳的参数值，以达到最大化攻击效果的目的。本节详细介绍了目标优化的模型投毒，如一点就足够的（LIE, little is enough）攻击、Fang攻击[18]和针对聚合规则设计的攻击（AGR-tailored）等，这类攻击方法旨在提高攻击的隐蔽性和有效性，从而更好地破坏联邦学习系统的全局模型，如表1所示。

表1 模型投毒的最优化函数

模型投毒	最优化函数
LIE攻击[15]	$\max\limits_{z}\left(\Phi(z)<\dfrac{n-\left\lfloor\frac{n}{2}+1\right\rfloor}{n-m}\right)$
Fang攻击[18]	$\max\limits_{w'_1,\cdots,w'_m} \boldsymbol{s}^{\mathrm{T}}(W-W')$
AGR-tailored攻击[19]	$\max\limits_{\gamma,\nabla^P}\lVert W-W'\rVert_2$
Min-Max攻击[19]	$\max\limits_{\gamma,i\in[m+1,n]}\lVert w'-w_i\rVert_2$
Min-Sum攻击[19]	$\max\limits_{\gamma}\sum_{i\in[m+1,n]}\lVert w'-w_i\rVert_2^2$
CMP攻击[47]	$\hat{W}'^{*}=\operatorname{agr\,min}F_A(\hat{W}')$

LIE攻击[15]的目标使得经验方差足够高，导致当时的防御机制始终选择拜占庭攻击者而丢弃良性客户端。恶意更新的表达式为$w_i=\mu_i-z\sigma_i, i\in[d]$。表1显示$z$的求解的最优函数，其中，$n$表示客户端的总数，$m$是拜占庭数量，$\Phi(z)$是累积标准正态函数。

Fang等[18]的攻击基于全局模型最大程度地偏离未攻击时全局模型更新变化方向的反方向的机理。具体的优化函数为$\max\limits_{w'_1,\cdots,w'_m}\boldsymbol{s}^{\mathrm{T}}(W-W')$，其中，$W$是全局模型更新，$W'$是带有攻击的全局模型更新，$\boldsymbol{s}^{\mathrm{T}}$是所有未受到攻击的全局模型更新变化方向的列向量和，$w_1,\cdots,w_m$是恶意更新的前m个维度参数。

另外，Shejwalkar等[19]提出了针对聚合算法的攻击，旨在降低恶意全局模型更新与良性全局模型更新之间的相似性。与Fang攻击不同，该攻击的恶意更新为$w'_{i\in[m]}=W+\gamma\nabla^P$，其中，$\gamma$是缩放因子，$\nabla^P$是扰动向量。基于未知防御方法的前提条件下，他们还提出了2种变体攻击最小化最大距离（Min-Max, minimize maximum distance）攻击[19]和最小化距离和（Min-Sum, minimize sum of distance）攻击[19]。其中，Min-Max的攻击策略为攻击者必须确保恶意新靠近良性更新的群集，同时最大化与良性更新的距离或L_P范数差异。因此，攻击者的目标是最大化恶意更新与良性更新之间的距离，约束条件是恶意更新与良性更新的距离应小于良性更新距离的最大值。相应的优化函数是求解恶意更新与良性更新的距离最大化，即$\max\limits_{\gamma,i\in[m+1,n]}\lVert w'-w_i\rVert_2$，其中，$\gamma$是缩放因子，$w_i$是良性更新，$w'$是恶意更新。而Min-Sum攻击的策略为攻击者必须确保最大化恶意更新与良性模型更新之间的距离，同时确保与其他良性参数之间的最大距离小于任意2个良性更新之间的最大距离。与Min-Sum攻击类似，相应的优化目标为求解恶意更新与良性更新距离和的最大化，即$\max\limits_{\gamma}\sum_{i\in[m+1,n]}\lVert w-w_i\rVert_2^2$。

Wei等[47]提出了隐蔽的模型投毒（CMP, covert model poisoning）攻击，考虑拜占庭攻击的隐蔽性。通过最小化操纵模型与指定模型之间的欧氏距离，并受到防御性聚合方法的约束。攻击者的目标通常是找到一组m恶意更新参数，使得上传服务器时目标函数$F_A(\cdot)$最小化，如表1所示，其中$\hat{W}'$是指带有恶意客户端聚合参数，$\hat{W}'^{*}$是指最优的聚合参数。实验结果表明，所提出的攻击方法比现有的标签翻转攻击更有效。

此类攻击通过设计不同的优化函数能够在不同的攻击场景中实现不同的攻击目标[46]。例如，在Fang攻击中，攻击者利用梯度更新对全局模型训练的重要性，构造优化函数，使得恶意客户端上传的梯度更针对性地破坏全局模型的准确性。

3) 合谋的模型投毒

合谋的模型投毒是指攻击者之间可以共谋实施攻击，攻击者可以采取相同或者不同的攻击策略修改上传的梯度。依据攻击目标不同进行分类，包括攻击目标相同的合谋模型投毒，例如Yang等[48]提出的定向合谋攻击以及上述的目标优化的模型投毒攻击；攻击目标不同的合谋模型攻击，例如Xu等[44]提出新型混合攻击（ByzMean），能够使聚合梯度篡改成任意的恶意梯度。具体而言，把恶意客户端分成2个集合，其中，m_1个客户端

选择任意梯度向量$\boldsymbol{w}_{m_1} = m$，$m_2 = m - m_1$个客户端选择梯度向量$\boldsymbol{w}_{m_2}$且使所有梯度的平均值正好是$\boldsymbol{w}_{m_1}$，如式(2)所示。

$$\boldsymbol{w}_{m_1} = *, \boldsymbol{w}_{m_2} = \frac{(n - m_1)\boldsymbol{w}_{m_1} - \sum_{i=m+1}^{n} \boldsymbol{w}^i}{m_2} \tag{2}$$

目前，所提出的攻击都可以集成到ByzMean中，以实现多目标攻击。例如，可以将$\boldsymbol{w}_{m_1}$设置为随机噪声或由攻击者精心设计的梯度，此类攻击增加了攻击的灵活性和强度以及检测攻击的难度。

4) 自适应的模型投毒

自适应的模型投毒是针对所提出的防御方法动态调整攻击策略和参数，以最大程度地影响全局模型的可用性和完整性。例如，由于联邦学习中的全局模型和局部模型之间的差异使得局部优化触发器在转移到全局模型时效果明显降低。为了解决上述挑战，Zhang等[49]在联邦学习中提出了对抗自适应的后门攻击（A3FL, adversarially adaptive backdoor attack to federated learning），对抗性地调整后门触发器。Wu等[50]研究当前的防御机制是否能够在实际环境中消除联邦学习后门攻击的威胁。从攻防对抗的视角，直接修改小部分局部模型权重通过符号翻转注入后门触发器中和与客户端模型联合优化触发模式，从而更持久、更隐蔽地规避现有防御。Kasyap等[41]针对相似性防御方法提出仅有相似性是不足（Sine, similarity is not enough）的攻击，它能够将梯度平均值偏移到恶意更新使其作为全局模型更新进行迭代，并且保持攻击的持久性使得全局模型无法收敛。

目标优化的模型投毒攻击方法中，如Fang攻击和AGR-tailored攻击也是自适应攻击，分类存在一定的交叉，不同之处在于后者只针对有防御方法的联邦学习模型进行攻击，前者则侧重于建模优化方法，部分涵盖自适应攻击。从攻防对抗的角度，研究自适应攻击，甚至攻击极其隐蔽极难防御的方法，是实现联邦学习中攻击方法极具前景的研究方向。

5) 其他的模型投毒攻击

其他的模型投毒攻击是通过一小部分的坐标维度参数进行扰动或者对攻击效果进行细粒度的控制来对联邦学习模型进行攻击。例如，为了提供攻击方法的隐蔽性，Jin等[51]精心设计一小部分更新参数来创建恶意更新，从而绕过多个最先进的防御方法，如基于符号防御（SignGuard）[44]和抑制联邦学习中的后门攻击的防御方法（FLAME）[14]等。Zhang等[52]提出了灵活模型投毒攻击（FMPA, flexible model poisoning attack），恶意的FL服务提供商可以在不被注意的情况下获得优于竞争对手的优势，从而在FL中开辟了除拒绝服务攻击之外的新攻击面。此外，该攻击不仅仅保证攻击的有效性，旨在增加攻击的隐蔽性。

1.3　评估指标

针对不同的攻击类型，全面的评估指标能更好地反映攻击的隐蔽性、有效性、攻击强度等。目前许多学者提出和使用了已有多种评估指标，但仍然存在一些问题，比如难以覆盖复杂攻击场景和缺乏标准性等。本节总结了相关的评估指标。

1) 攻击成功率[53]是指如果恶意模型将一个数据X输入，输出为所期望的目标类，则认为攻击成功。模型F_p攻击成功率为

$$\mathrm{SR}_{F_P} = \frac{\left(\sum_{X \in D} S_X\right)}{|D|} \times 100\% \tag{3}$$

其中，D是所有样本，$S_X = 1$指目标攻击成功。式(3)所示指标主要针对像标签翻转攻击等定向攻击的评估，攻击成功率越高说明攻击效果越好。

2) 攻击对参数的影响[54]是指联邦学习系统对恶意设备进行攻击后，累积到第t轮的全局模型参数的变化，记为

$$\delta_t \equiv W_t(S \backslash M) - W_t \tag{4}$$

其中，$W_t\left(S_i \backslash M\right)$表示当$S_i (i \leqslant t)$中的所有恶意设在第$i$个训练轮中不执行攻击时第$t$轮中的全局模型权重。

3) 后门精度[55]是指表示模型在后门任务中的准确性，该评估指标适用于对后门攻击效果的评估，有效的防御方法将使得后门精度值降低。

4) 主任务精度[55]是指模型正确预测的良性输入的比例。在定向攻击中，攻击者通常会尽量减少对主任务精度的影响，以降低被检测的可能性。而防御方法的目标则是在保护模型免受攻击的同时，尽可能不影响主任务的精度。

5) 模型准确率[56]是指采用联邦学习全局模型准确性作为评估度量，计算方法为

$$\mathrm{ACC} = 1 - \frac{\sum_{x=1}^{D} y_x \neq \bar{y}_x}{|D|} \tag{5}$$

其中，$\overline{y_x}$为样本x的预测标签。模型准确率越低，说明联邦学习系统中遭受非定向攻击的可能性越大。

6) 缓解轮[55]是指一个模型在受到拜占庭攻击后，其精度首次降至50%，然后再次回升至50%所经历的通信轮数。该指标用于评估模型对攻击的抵抗能力，可以作为模型恢复到正常所需的时间或步骤的度量。

1.4 小结

拜占庭攻击是联邦学习中常见的安全问题，特别对于大规模的模型训练[38]。本文依据攻击目标、攻击者背景知识和能力、攻击方式等细化分类，梳理了攻击常用的评估指标，为未来应对联邦学习中的攻防对抗问题以及实际落地奠定基础。此外，数据投毒对攻击者的背景知识和能力要求低，只需要对恶意客户端的数据进行操纵，实际场景中应用广泛，但攻击效果不直接高效。模型投毒对模型可用性的影响更为显著，其可能导致模型难以收敛，或者使模型朝无效的方向收敛。该攻击表现出更好的攻击成功率和更长的持续时间[57]。但模型投毒攻击成本相对较高，攻击者需要获得相应上传的参数信息和聚合过程的完整信息。从攻防对抗的视角，基于对上述攻击方法的总结，攻击形式从初始的随机攻击到基于优化和自适应的攻击逐渐优化，也凸显出联邦学习中的防御方法日益增强，进一步增加了整个联邦学习模型训练的安全性。未来，研究隐蔽性强、难以防御且具有自适应特性的对抗盲攻击，将是联邦学习攻击领域的一个重要研究方向。

2 拜占庭防御

拜占庭防御是针对攻击者控制和操纵部分训练数据或模型参数等攻击手段，通过防护和检测部分客户端或模型参数来保护全局模型的训练过程，保障训练模型的安全性和完整性等。该防御出现在联邦学习的模型训练阶段，防御者通过在本地客户端进行防护、在训练过程中检测上传的参数、在聚合阶段及时响应以及恢复模型性能等方式对全局模型进行防护以实现提高全局模型的安全性，并且可以基于安全防护策略组合多种防御方法，从而对遭受恶意攻击的联邦学习系统进行更有效的防护。

本文参考经典的网络安全防御模型APPDRR[58-59]，面向拜占庭攻击重点分析了联邦学习训练过程全周期的安全防护需求。

APPDRR是一种经典的可适应性动态网络安全模型[59]，模型系统的设计充分考虑到风险分析、安全策略、系统防护、攻击检测、实时响应、灾难恢复等各个环节，并考虑到各部分之间的动态依赖关系使得整个系统的生存能力大大增强，最大限度地降低网络事件带来的风险和损失。本文借鉴APPDRR分析联邦学习各个环节的防护需求，表2阐述了联邦学习拜占庭防御中APPDRR环节的具体含义。

针对当前联邦学习中的拜占庭攻击研究的种种防御方法，本文在下述内容对最新防御技术从攻防框架和攻防机理2个维度进行详细的阐述并对它们的防御效果和优缺点进行了总结。具体地，从攻防框架的视角，考虑联邦学习的整体安全需求，基于安全防御模型，介绍了针对联邦学习拜占庭攻击的全流程安全防御手段。从攻防机理的视角，进一步扩展了联邦学习中的分类。例如，在安全防护环节，分为客户端自身防护、过滤和对参数空间的稀疏化；在恶意检测环节，增加了参数方向间、组合特征和相对分布相似性防御分类方法；在异常响应环节，分为直接剔除恶意更新、基于信誉规避异常和基于投票规避异常等方式。为此，本文从联邦学习攻防框架视角，介绍针对联邦学习拜占庭攻击的全流程安全防御手段，总结防御方法的前提条件如数据分布为独立同分布（IID, independent and identically distributed）和非独立同分布（Non-IID , non-independent and identically distributed）以及恶意客户端数量占比，并结合防御机制的基本原理对其分类分析，其中，K为参与本地客户端的数量，f为恶意客户端的数量，如表3所示。

2.1 安全防护

安全防护是针对攻击者破坏客户端持有数据集、本地模型参数等攻击手段，通过客户端自身防护、过滤和对参数空间的稀疏化等方法，在本地客户端部署防御方法，限制恶意更新参数，使得拜占庭攻击的效果削弱，确保全局模型更新具有一致性和稳定性。进而从源头进行拜占庭防御，使得全局模型具备鲁棒性和抗攻击性。其核心原理是通过对本地客户端的模型参数和参数空间添加正则化项、聚类、噪声注入和干扰梯度更新来防护恶意更新对

表 2　联邦学习拜占庭防御中 APPDRR 环节的具体含义

APPDRR	传统网络安全模型		联邦学习中拜占庭防御系统	
	释义	含义	释义	含义
Analysis	风险分析	确定网络资产发生安全威胁的可能性和网络受到攻击的脆弱性，并估计由此造成的损失，主要包括定性分析和定量分析	安全评估	在联邦学习模型的训练阶段对系统进行安全评估，如受到攻击的可能性以及受到的攻击类型、攻击强度以及模型抵御有目标攻击的效果、遭受攻击模型性能降低程度等
Policy	安全策略	从全局考虑出发，负责制定一系列的控制策略、通信策略和总体安全策略，根据风险分析的结果做出决策	安全策略	针对现有和潜在的攻击策略，依据风险评估结果以及实际场景中的安全需求制定相应的一系列的安全策略，从而保护联邦学习系统的安全性
Protection	系统防护	通过系统防护限制网络中的数据包，以防止外部攻击，同时阻断内部来源的未经授权访问，主要包括防火墙、加密和认证等防御	安全防护	针对破坏客户端持有数据集、本地模型参数等攻击手段，安全防护通过客户端自身防护、过滤和对参数空间的稀疏化等方法限制恶意客户端对联邦学习模型开展的攻击
Detection	攻击检测	通过对整个网络的动态性能监控、漏洞扫描、入侵检测、反病毒、网络管理，及时发现新的攻击行为	恶意检测	通过检测恶意梯度更新及时发现联邦学习系统中存在的攻击，如数据投毒攻击和模型投毒攻击等，其主要方法包括特征相似性检测和利用异常检测算法来检测恶意更新
Response	实时响应	安全事件发生后的紧急应对策略，包括阻断攻击、隔离故障或设置陷阱和进行追踪[57]	异常响应	当检测到恶意更新时及时做出反应，通过直接剔除恶意客户端更新、基于信誉参数和基于投票等来规避异常行为
Recovery	灾难恢复	当发现外部攻击和系统漏洞时，系统应采取数据备份、容灾、生存性等方法，以及利用升级系统、升级软件和打补丁等措施来恢复重要信息[57]	模型恢复	针对联邦学习模型训练过程中检测到的恶意攻击，利用历史梯度信息、梯度添加高斯噪声、差分隐私技术、模型聚合等方法尝试恢复模型的精度，以使训练过程能够继续进行及时恢复遭到攻击的模型性能

全局模型的影响；通过对参数空间进行稀疏化操作，使得全局模型能够主动容忍恶意更新，依据参数空间的特性不同，基于参数空间稀疏化算法主要分为聚合参数空间的稀疏化和本地模型参数空间的稀疏化两大类。

2.1.1　基于客户端自身防护和过滤的防御

基于客户端自身防护和过滤的方法在防御数据投毒攻击中较为常用，比如标签翻转攻击和后门攻击。通过客户端自身防护和过滤恶意攻击的方式，主动防御可能出现的攻击形式。

例如，Shen 等[53]针对 IID 提出了 Auror 防御方法。在攻击场景中，利用大多数良性客户端的指示性特征表现出类似的分布，而来自恶意客户端的指示性特征将表现出异常分布现象来过滤恶意客户端。而 Tolpegin 等[16]对最后一层模型参数采取主成分分析（PCA, principal component analysis）进行降低模型参数的维度，达到客户端自身过滤恶意更新的目的。联邦学习中的“疫苗”（RECESS）防御方法[78]通过精心构建的聚合梯度主动查询每个参与的客户端，过滤恶意客户端，该防御方法提升了全局模型的效能。

然而，上述涉及的防御方法通常基于较强的假设，例如联邦学习的数据分布为 IID、本地梯度参数需要全部参与聚合和限制恶意客户端的占比上限等强假设，这将导致在实际应用场景中难以部署且由于联邦学习特定参数的变动使得此类防御效果并不乐观。为此，Zhang 等[60]提出了新的防御方法，无需假设数据分布和恶意客户端的占比，尽可能地减少对联邦学习特定参数的限制。其基本思想为通过检查客户端的模型更新一致性来检测恶意客户端。服务器基于客户端在每一轮中的历史模型更新来预测客户端的模型更新。如果接收的模型更新在多轮中与预测的模型更新不一致，则认为是恶意更新。该防御方法为拜占庭攻击手段提供了新的防御思路，并展现了较好的防御效果，但也需要知道历史模型更新参数的前提假设。

2.1.2　基于客户端模型和参数扰动设置的防御

基于客户端模型和通过对参数空间添加正则化项、噪声和干扰梯度等扰动设置的防御方法，主动防护联邦学习中的攻击。特别地，在恶意客户端数量变化较大的突发对抗攻击下，基于客户端自身防护和过滤的方法无法保证联邦学习的安全

表 3 联邦学习中拜占庭防御的主要研究

APPDRR	防御方法	防御机制	异常响应	数据分布	恶意客户端数量
安全防护	Auror[53]	本地模型参数聚类验证	—	IID	$K \geqslant 2f+1$
	DataPoisoning_FL[16]	最后一层模型参数聚类验证	—	IID/Non-IID	$K \geqslant 5f+1$
	FLDetector[60]	预测模型更新为基准	基于信誉规避异常	IID/Non-IID	$K \geqslant 2f$
	FLAME[14]	差分隐私模型过滤攻击	—	IID/Non-IID	$K \geqslant 2f+1$
	FL-WBC[61]	识别参数空间过滤攻击	—	IID/Non-IID	$K \geqslant 2f$
	LeadFL[55]	本地模型添加正则化项	—	IID/Non-IID	—
	SparseFed[62]	更新 Top-k 个最高幅度元素	—	IID	$K \geqslant 4f+1$
	FRL[63]	参数更新空间稀疏化	基于投票规避异常	IID/Non-IID	$K \geqslant 4f+1$
恶意检测	RGDA[38]	可信数据集为基准	基于信誉规避异常	IID	$K \geqslant f+1$
	FLTrust[64]	可信数据集为基准	基于信誉规避异常	IID/Non-IID	$K \geqslant f+1$
	FLTC[50]	可信数据集为基准	基于信誉规避异常	IID/Non-IID	$K \geqslant f+1$
	RoseAgg[46]	识别可信的良性成分	基于信誉规避异常	IID/Non-IID	$K \geqslant 2f$
	Agramplifier[54]	识别最具抑制性的特征	基于信誉规避异常	IID/Non-IID	$K \geqslant 2f$
	Krum[40]	欧氏距离	直接剔除恶意更新	IID	$K \geqslant 2f+3$
	Median[65]	梯度值取中位数	直接剔除恶意更新	IID	$K \geqslant 2f+1$
	trimmed-mean[65]	梯度值取截断平均值	直接剔除恶意更新	IID	$K \geqslant 2f+1$
	Bulyan[66]	欧氏距离	直接剔除恶意更新	IID	$K \geqslant 4f+1$
	RFA[67]	梯度值的几何中值	直接剔除恶意更新	IID	$K \geqslant 2f+1$
	DeFL[68]	联邦梯度范数度量	基于信誉规避异常	IID/Non-IID	$K \geqslant 2.5f+1$
	SignGuard[47]	参数符号间的相似性	基于信誉规避异常	IID/Non-IID	$K \geqslant 2.5f+1$
	AFA[44]	余弦相似度	基于信誉规避异常	IID	$K \geqslant 2f+1$
	PEFL[70]	Pearson 相关系数	基于信誉规避异常	IID	$K \geqslant 2f+1$
	ShieldFL[71]	余弦相似度	基于信誉规避异常	IID/Non-IID	$K \geqslant 2f+1$
	FoolsGold[22]	历史梯度的余弦相似度	基于信誉规避异常	IID/Non-IID	$K \geqslant f+1$
	DnC[19]	谱方法异常检测	直接剔除恶意更新	IID/Non-IID	$K \geqslant 2f+1$
	SAD[72]	光谱异常检测	直接剔除恶意更新	IID	$K \geqslant 2f$
	ACBD[73]	自编码器模型异常检测	直接剔除恶意更新	IID	$K \geqslant 2f+3$
	MCDFL[74]	特征空间的相对分布	基于信誉规避异常	IID	$K \geqslant 2.5f+1$
	LoMar[75]	梯度值领域的相对分布	基于信誉规避异常	IID/Non-IID	$K \geqslant 2.5f+1$
模型恢复	FAST[56]	—	客户端贡献删除	—	$K \geqslant 1.25f+1$
	FedRecover[76]	—	恢复良性客户端更新	IID/Non-IID	$K \geqslant 2f$
	FedRecovery[77]	—	模型参数添加高斯噪声	IID/Non-IID	—

性和鲁棒性。

基于上述挑战，Sun 等[61]提出了基于客户端的防御方法为白细胞联邦学习（FL-WBC, white blood cell for federated learning），其核心思想是识别对参数的长期攻击影响所在的参数空间，并在局部训练期间扰动该空间。Zhu 等[55]提出了一个客户端自我防御方法（LeadFL），其核心思想是在局部模型训练中添加一个基于梯度的黑塞（Hessian）矩阵的新正则化项来扰动局部模型更新。具体的正则项表达式为

$$R_{t,i}^{\mathrm{k}} = \mathrm{clip}\left(\nabla\left(\boldsymbol{I} - \eta_t \boldsymbol{H}'^{k}_{t,i}\right), q\right) \tag{6}$$

其中，$\boldsymbol{I}$和η_t是单位矩阵和学习率，$\boldsymbol{H'}^k_{t,i}=\frac{\theta'^k_{t,i+1}-\theta^k_{t,i}-\Delta\theta^k_{t,i}}{\eta_t}$表示评估的Hessian矩阵，$\theta'^k_{t,i+1}=\theta^k_{t,i}-\eta_t\nabla L\left(\theta^k_{t,i}\right)$代表模型参数的更新，裁剪是将正则化项限制到阈值$q$内以确保模型收敛。得益于优化的正则化，从理论上证明了优化的正则化使得模型投毒对主任务准确性的低影响。该防御方法也可以和基于服务器端的防御方法相结合，从而形成对联邦学习的全局模型训练过程的多层保护作用。

另外，Nguyen 等[14]针对后门攻击提出了FLAME防御框架，可表示为

$$G^*=\frac{1}{n}\sum_{i=1}^{n}W_i^*=\frac{1}{n}\left[\sum_{i=1}^{n}W_i+N\left(0,\sigma_i^2\right)\right]=\frac{1}{n}\sum_{i=1}^{2}W_i+N\left(0,\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)=G+N\left(0,\sigma_G^2\right)\tag{7}$$

其中，W_i^*是添加噪声的局部模型，G^*是添加噪声的全局模型，σ_i是本地模型的噪声标准差，W_i是局部模型，N是指随机噪声，G是指差分隐私模型参数，σ_G^2是指加权平均值的方差，具体为

$$\sigma_i=\frac{\alpha_i e_i}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}}\tag{8}$$

其中，$\alpha_i=\frac{\Delta_i}{e_i}$是超参数，$\Delta_i$和$e_i$分别是指本地模型$W_i^*$的灵敏度和$L_2$范数，$\varepsilon$和$\delta$是控制差分隐私强度的2个参数。后门攻击是通过在数据中插入触发器在模型训练时来破坏全局模型的性能。而该防御方法通过对本地训练模型中注入噪声，使恶意更新无法影响全局模型影响，触发器失效。

此类防御方法不同于基于客户端自身防护和过滤的方法。后者是利用特征的不一致性防护过滤恶意攻击者对全局模型性能的影响，而前者是基于参数扰动、添加噪声以及在局部模型训练中添加新的正则化项等策略削弱攻击者对全局模型的影响。但该类方法由于引入噪声不可避免会导致训练模型的性能下降，使得良性用户的数据在训练过程中有用信息提取减少。

2.1.3　基于参数空间稀疏化的防御

基于参数空间稀疏化的防御方法是通过对聚合参数、本地模型参数空间稀疏化使得恶意更新梯度信息在聚合时对全局梯度的影响减少。上述防御方法的目的是类似的，其区别是稀疏化的对象不同。

具体地，攻击者为了成功地毒害模型，则不会尝试同步更新良性客户端需要更新的坐标，而会试图向与大多数良性更新不同的方向移动。基于这一事实，Pand 等[62]提出了稀疏化梯度减轻模型投毒（SparseFed），通过仅更新聚合模型的最相关权重来减轻联邦学习中的模型投毒。其核心思想为在每一轮训练中，参与的客户端计算其本地数据的更新并剪切。服务器计算聚合梯度，并且仅更新top-k个最高幅度更新，以此削弱模型投毒对全局模型训练迭代过程的影响。而基于本地模型参数空间稀疏化方法是依据现有联邦学习中的拜占庭攻击的成功关键因素与客户端梯度的值域较大有关，即（单个，少量的）客户端就可以对模型更新产生较大的影响的事实，进行稀疏化进而减小拜占庭攻击成功的参数空间。Mozaffari 等[63]提出了联邦秩学习（FRL, federated rank learning）防御方法。FRL 将客户端更新的空间从联邦学习中的浮点数的连续空间转换成整数值的离散空间。为了能够使用参数序列来训练全局模型，FRL 利用了最近的超级掩码训练机制的思想。FRL 的防御目标是寻找全局排名R_g，并将其转化为一个全局二进制掩码M。使得最终的子网络$\theta^w M$最小化所有用户的平均损失，优化公式为

$$\min_{R_g}F\left(\theta^w,R_g\right)=\min_{R_g}\sum_{i=1}^{N}\lambda_i L_i\left(\theta^w M\right)\quad \text{s.t. } M\left[R_g<k\right]=0 \text{且} M\left[R_g\geqslant k\right]=1\tag{9}$$

其中，N是FRL客户端的总数，L_i是第i个客户端的损失函数，$\lambda_i=\frac{1}{N}$是第i个客户端的权重，M是最终的二进制掩码，θ^w是全局模型参数。FRL客户端基于其本地训练数据对随机初始化的神经网络的参数进行排名，并且FRL服务器使用投票机制来聚合由客户端提交的参数排名。

此类防御策略相较于前两类防御，提出的角度比较新颖，使得联邦学习中的全局模型能够主动容忍恶意更新。并且FRL进一步保护了数据隐私，基于投票排名的方式提升了联邦学习的通信效率，进行了性能优化，有效地提升了联邦学习系统的鲁棒性，但目前的研究和实验尚少。

基于上述分类分析，尽管上述3类防御方法采取的防御策略不同，但其主要目的是相同的。基于预测可能存在的攻击，通过对模型、参数和参数空间操作，主动防护联邦学习系统的安全性。因此，

在联邦学习系统中，安全防护作为第一道防线，使得该系统具备一定的抵御攻击能力。但如第二类防御方法表示，如何平衡为了抵御拜占庭攻击进行安全防护机制导致对全局模型精度产生的消极影响是联邦学习安全防御研究面临的主要挑战之一[70]。

2.2 恶意检测

恶意检测是通过检测恶意梯度更新并及时发现联邦学习系统中存在的攻击。检测恶意更新本质上是一个二进制分类问题，目的是检测出恶意更新和良性更新。关键思想是利用特征之间的一些统计差异来检测恶意更新和良性更新。对于每个客户端，这些检测方法首先从其一轮或多轮模型更新中提取特征，然后使用分类器来预测是否存在恶意。依据防御策略不同，将防御方法分为基于参数统计特性和数值间相似性的防御[51,66-71]、基于参数方向间相似性的防御[20,72-74]、基于异常检测算法的防御[17,75-76]和基于参数相对分布间相似性的防御[77-78]。

2.2.1 基于可信基准数据的防御

此类防御方法的前提假设是服务器能够预先从所有本地客户端收集小部分纯净的数据，并基于此数据集进行训练，得到全局模型的基准梯度。如果上传的本地梯度与基准梯度大小和方向都相似，说明未受到拜占庭攻击；反之，说明受到拜占庭攻击的可能性比较大。

例如，Cao等[38]提出抗任意数量的拜占庭攻击的鲁棒聚合算法（RGDA, robust gradient descent algorithm），服务器随机选择小型的可信数据集训练得到基准梯度，与此轮参与的客户端上传的梯度对比过滤恶意梯度，最后计算所有良性的梯度和服务器训练梯度的均值更新全局模型。此方法无需知道恶意客户端的数量，即使拜占庭的数量超过50%，该聚合方法都是收敛的。另外，Cao等[64]提出基于信任根的防御方法FLTrust，该方法利用服务器训练可信数据集得到基准梯度。在每一轮训练过程中，服务器对比本地模型更新与基准模型更新的余弦相似度，分配一个信任分数（TS, trusted score），依据此分数加权平均归一化本地模型更新，得到全局模型更新。其中，服务器对每个本地模型更新值归一化目的是抵御梯度值大幅度变化的攻击。Kasyap等[41]提出可信坐标联邦学习（FLTC, FL trusted coordinate）的防御方法，与FLTrust的信任分数计算方法不同，该方法考虑了坐标级的鲁棒聚合方法。它使用方向计算余弦相似度，并在相反方向上修剪不可信的坐标。具体而言，利用信任分数进行排序选择前$(1-2\beta)n$个更新，并对这些更新的每个维度进行加权平均得到全局聚合模型。FLTC可抵抗强自适应性模型投毒，如Sine攻击。此外，Yang等[48]提出了鲁棒的防御方法（RoseAgg），它动态地从本地更新中识别可信的良性成分，并利用良性成分来限制恶意更新的影响。Gong等[54]提出了通过本地更新放大保护联邦学习免受投毒攻击（Agramplifier）的防御方法，它的核心思想是通过识别每个梯度更新中最具抑制恶意梯度更新的特征来放大局部更新的可信性。

该类方法直接依赖可信数据集的训练梯度和测试结果，检测效果更加可靠。但同时因为需要预先收集可信数据，难以在现实场景中部署。其主要原因有2个方面，一是在大多数联邦学习场景中，可信数据集难以收集；二是可信数据来自各个参与联邦学习的客户端，与联邦学习的局部模型的隐私性需求相冲突。

2.2.2 基于参数统计特性和数值间相似性的防御

此类防御方法基于参数更新对全局模型训练提升精度至关重要的事实，进而依据参数更新的统计特性和相似性检测恶意更新。并且，在联邦学习中，为了攻击提高攻击成功率，恶意更新一般会远离良性更新，越分散的更新越有可能为恶意更新。因此，本节利用该参数数值统计特性作为全局模型更新，如坐标中位数、几何中位数和截断均值等或者利用参数数值间的相似性检测恶意更新，如欧氏距离。

传统的FedAvg[79]对拜占庭攻击不具有鲁棒性。针对上述情况，许多学者提出了鲁棒聚合算法。例如，Blanchard等[40]提出基于欧氏距离的防御方法（Krum），利用欧氏距离计算得到每个局部梯度向量与其他最近的m个梯度向量；然后针对每个局部梯度向量，计算其与m个梯度向量之间的距离之和，并对该距离之和进行分数化；最后选择分数最小的梯度向量作为全局聚合向量。Yin等[65]提出基于中位数（Median）、基于截断中位数（trimmed-mean）的防御方法，其中，Median将梯度的每一维度的中位数作为全局聚合梯度向量；trimmed-mean则是把梯度每一维度的边缘值截断再进行以

平均值作为全局聚合向量。Mhamdi 等[66]提出了防御方法 Bulyan，在使用裁剪平均进行聚合之前进行 Krum 操作。Pillutla 等[67]提出了鲁棒联邦聚合（RFA，robust federated aggregation），RFA 使用基于平滑的 weiszfeld 方法计算局部更新的几何中位数，作为全局模型更新。Xia 等[80]提出抵御拜占庭攻击的快速聚合（FABA，fast aggregation against Byzantine attack）方法，其基本思想是在每次迭代中，从参数服务器收集所有参与客户端的梯度，然后检测并排除一些离当前平均梯度最远的梯度，最后用剩余的梯度计算新的平均梯度，并更新模型参数。

然而，上述的防御方法主要依据梯度值的统计特征，如中位数、平均值、欧氏距离等，计算比较简单，适用对模型参数产生很大影响的攻击。但当攻击导致的梯度幅度变化小或者统计特征和相似性特征不明显或不能很好地区分恶意梯度时，防御方法的效果会极大地降低且易出现假阳性，导致良性更新参数被误判。为此，Yan 等[68]提出了一种新的联邦学习感知防御投毒（DeFL，defense against poisoning of FL）方法，其基本思想是通过易于计算的联邦梯度范数（FGNV，federated gradient norm vector）度量来测量深度神经网络（DNN，deep neural network）模型更新之间的细粒度差异。具体定义第 i 个客户端的全局模型第 j 层参数的更新差为

$$\Delta l_i^j = l\left(w_i^j - \eta g_i\left(w_i^j;\xi\right);\xi\right) - l\left(w_i^j;\xi\right) \quad (10)$$

其中，l 是指损失函数，w_i^j 是指模型参数，η 是指学习率，g_i 是指梯度函数，ξ 是指训练数据。

使用泰勒展开通过其梯度范数来近似，相应公式为

$$\Delta l_i^j \approx -\eta \left\| g_i\left(w_i^j;\xi\right)\right\|^2 \quad (11)$$

定义 FGNV_i 为第 i 个客户端的全局模型第 j 层参数的联邦梯度范数，可表示为

$$\mathrm{FGHV}_i^j := \Delta l_i^j \quad (12)$$

将 $\mathrm{FGNV}_i = \left(\mathrm{FGNV}_i^1,\cdots,\mathrm{FGNV}_i^L\right)$ 表示为第 i 个客户端的联合梯度范数向量，其表示 DNN 的每一层上的第 i 个客户端的全局模型更新差异。然后，可以使用所选客户端上的加权平均来近似第 t 轮处的第 l 层上的全局模型更新差异，相应公式为

$$\mathrm{FGNV}^j(t) = \sum_{i\in N(t)} \frac{\left|D_i\right|}{\sum_{i\in N(t)}\left|D_i\right|}\mathrm{FGNV}_i^j(t) \quad (13)$$

基于参数统计特性的防御利用梯度的统计特性推断全局模型参数特性，而基于参数数值间的相似性则利用梯度间的欧氏距离，此类防御策略大多数计算开销少和易于部署。特别是在梯度幅度变化比较大的情况下，能够适应不同类型的恶意攻击。

通常假设联邦学习数据分布为 IID 和恶意客户端的数量占比少于 50%。但当上传的梯度之间数据差别较大，如何精确区别恶意梯度以及 Non-IID 场景下的良性梯度，将是联邦学习防御面临的一大难题。

2.2.3　基于参数方向间相似性的防御

梯度的方向在模型迭代训练中至关重要，因为它决定了损失函数收敛下降的最快方向。通过计算梯度并沿其反方向更新参数，可以优化模型性能。因此，此类防御方法利用了参数方向间的相似性来区分恶意更新和良性更新，从而排除恶意更新对全局模型的影响。具体地，常用皮尔逊（Pearson）相关系数和余弦相似度来计算参数方向的相似性。

Liu 等[70]和 Shen 等[81]计算每个局部更新与基准的 Pearson 系数，将与基准差异较大的局部更新赋予更低的权重，中央聚合器根据权重对局部更新进行加权平权，得到全局模型。Munoz-González 等[69]提出自适应联邦平均（AFA，adaptive federated averaging）防御方法，首先计算每个参与方的模型更新与全局模型的余弦相似度，设置相似度阈值以排除不良更新。Ma 等[71]提出基于双陷门同态加密的隐私保护防御策略（ShieldFL），该防御方法首先对每个局部更新进行归一化，并设置阈值，将超出范围的局部更新舍弃，然后计算剩余局部更新与上一轮全局模型的余弦相似度，选择余弦相似度最小的局部更新，计算该局部更新与其他的梯度参数的余弦相似度用于确定其他的局部更新的可信度。最后根据每个更新的可信度，加权聚合得到全局模型。Fung 等[22]提出有目标的女巫投毒防御方法（FoolsGold），依据恶意更新目标的一致性，计算历史梯度的余弦相似度检测恶意梯度。但该防御方法主要是针对女巫攻击的防御方法，可以配合其他的恶意检测的防御方法，以增强联邦学习模型的鲁棒性。Xu 等[44]提出了恶意梯度过滤的防御方法，通过利用梯度值的方向特征进行聚类以过滤异常以及恶意的梯度更新。

这类防御方法的优势在于其简单直观，并且能

够在一定程度上检测到恶意更新，适用于检测对全局模型影响比较大的拜占庭攻击。另外，此类防御方法也存在一些局限性，例如对参数更新方向的计算可能受到噪声和不确定性的影响，如Sine攻击，同时攻击者可能采取一些对抗性的策略来规避检测，考虑对抗性攻击的防御方法研究较少，目前仍存在一些空白。

2.2.4 基于异常检测算法的防御

此类防御方法利用参数值的异常检测算法及时检测异常数据、异常行为或者不符合预期的模式并做出系统响应，来提高全局模型训练过程的安全性。例如，文献[72]提出了一个鲁棒的基于光谱异常检测（SAD，spectral anomaly detection）的框架，其中光谱异常检测在服务端进行检测和删除来自异常客户端的模型更新。光谱异常检测的思想是在去除模型参数更新中的噪声和冗余特征后，在低维潜空间中区分良性模型参数更新和异常的模型参数更新的嵌入。

Li等[73]提出异常客户端行为检测（ACBD，abnormal client behavior detection）为每个客户端分配一个信用评分，即根据预训练的自编码器模型产生的异常评分计算。其中，自编码器模型基于来自客户端的模型权重更新来定义每个客户端的异常分数。Shejwalkar等[19]提出了分而治之（DnC，divide and conquer）的防御方法，其基于奇异值分解（SVD，singular value decomposition）的谱方法来识别和去除异常值。Han等[82]提出了一种为实际场景的联邦学习系统设计的异常检测方案。当攻击发生时，该方案在攻击发生时利用早期的跨轮检查，激活随后的异常检测算法，有效地删除异常客户端模型更新，确保良性客户端提交的本地模型参数不受影响。如果局部模型参数之间的相似性得分很高，这表明这些局部模型参数更有可能是良性的客户端，并表明联邦学习模型训练的“收敛”趋势，而较低的相似性可能表明在当前训练轮中发生了攻击，因为恶意客户端可能已经上传了任意或篡改的局部模型梯度。

这类防御方法将恶意用户视为异常，并利用异常算法来识别模型的训练过程和数据等，如果检测到不符合预期的行为或异常数据，联邦学习系统将及时采取反应措施。研究适用于联邦学习分布式特性的异常检测算法，并将其与鲁棒性聚合算法结合，是实现兼顾联邦学习鲁棒性和隐私性的一个具有前景的研究方向。

2.2.5 基于参数相对分布间相似性的防御

这类防御方法通过核密度估计以及分布函数来得到模型参数和潜在空间的相对分布，利用该相对分布间的相似性来检测恶意更新和良性更新。例如，Li等[75]提出了局部恶意因子（LoMar，local malicious factor）防御方法，通过使用核密度估计预测其邻域的相对分布对来自每个本地客户端的模型更新进行评分，并确定区分恶意和良性更新的最佳阈值。此防御方法无需知道先验知识且具有强鲁棒性，但可能会受到数据分布不均匀或异常数据的影响，导致对恶意用户的检测假阴性。此外，核密度估计的计算复杂度较高，特别是当客户端数量较多时，计算量会显著增加。

Jiang等[74]提出了恶意客户端检测联邦学习（MCDFL，malicious clients detection federated learning）方法来抵御标签翻转攻击。它可以通过提取潜在特征空间分布来识别恶意客户端，以检测每个客户端的数据质量。图2阐明了第i个客户端的本地数据通过生成器$G_w(\cdot|y)$来提取潜在特征空间分布DQ_i的过程，而无需观察每个客户端的本地数据。由图2可知，该潜在特征空间的定义为

$$\mathrm{DQ}_i = \frac{1}{|D_i|} E_{x\sim D_i} E_{z\sim G_w(z|y)} \Big[\mathrm{ACC}\Big(\arg\max h(z;\theta_p), \big(\arg\max h(f(x;\theta))\big)\Big)\Big] \tag{14}$$

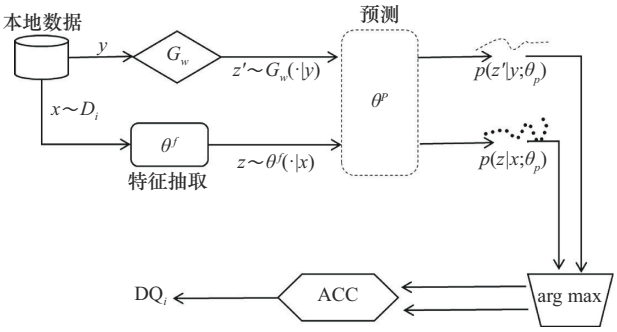

图2 特征空间的分布恢复过程

其中，$z\sim G_w(\cdot|y)$是通过生成器G_w从全局视角生成的与用户数据标签相对应的潜在特征空间分布，而y是用户i的本地训练数据的标签序列。函数$\mathrm{ACC}(a,b)$计算序列a和b中相同元素的总和。arg

$\max h\left(z;\theta_p\right)$是基于潜在特征$z$的预测，这些潜在特征对应于用户标签，并由全局模型的预测模块θ_p生成，这些预测被视为真实标签。$\arg\max h\left(f\left(x;\theta\right)\right)$是基于客户端本地数据样本的预测，这些预测被视为客户端本地样本的实际标签。

这类方法的主要优势在于可以在不直接观察每个客户端的本地数据的情况下，能够有效地识别出恶意客户端。然而，该防御方法可能受到特征空间的选取和恢复方法的影响，对于复杂的数据分布差异大或变化较快的场景可能存在一定的适用性局限性。此外，MCDFL方法相比于LoMar防御方法需要较高的计算开销和额外的通信开销，因为需要训练和使用生成器和预测模型等，对系统的性能和效率可能产生一定的影响。

2.3　异常响应

异常响应是指在检测到恶意更新或安全威胁时，能够及时采取措施来应对和处理，以减轻或消除潜在的安全风险和威胁。其主要目标是确保联邦学习的稳定性、可靠性和安全性，具体如表2所示。

根据策略方式分为直接剔除恶意更新、基于信用规避异常行为和基于投票规避异常行为。其中，直接剔除恶意更新是当检测到恶意更新时，联邦学习可以立即停止接受来自该客户端的更新，并将其剔除，防止进一步对全局模型产生影响；基于信用规避异常行为是指建立客户端信用评估机制，根据客户端的历史行为、信用等级等信息来评估其可信度，对于信用较低的客户端可以采取削弱该客户端的贡献等措施，以规避其可能的异常行为，如RoseAgg防御方法；基于投票规避异常行为是指可以采用多数投票或一致性机制来识别和规避异常行为，当多个客户端的行为出现不一致或异常时，可以根据投票结果或共识机制来做出相应的决策。通过及时的应急响应措施，可以有效地减轻或消除恶意行为对系统安全性的影响。

2.4　模型恢复

尽管上述可以检测并剔除出恶意客户端，但是恶意客户端的历史贡献也会对局部模型以及全局模型产生影响。对此，一些研究提出了模型恢复技术[83-84]，在检测到恶意客户端后排除其历史贡献，恢复全局模型的性能。

Guo等[56]提出了采用联邦学习消除服务器端恶意终端（FAST，federated learning to eliminating server malicious terminal）防御方法，基本思想是通过从服务器中减去每个恶意客户端的历史参数更新来消除恶意客户端的影响，并通过使用存储在服务器上的一小组基准样本对遗忘模型[85-87]进行额外训练来恢复模型性能偏差，相比于传统的从零开始训练迭代整个全局模型的方法，FAST节省了大量的运行时间和资源，达到了类似的模型性能，但该防御方法依赖于历史数据和基准数据选择。Cao等[76]提出了FedRecover防御方法，关键思想是在模型恢复过程中服务器估计本地客户端的梯度上传服务器端更新。服务器使用柯西中值定理来估计每个本地客户端在每一轮中的模型更新。相应的模型更新g_t^i为

$$g_t^i = \overline{g}_t^i + \boldsymbol{H}_{\mathrm{t}}^i\left(\hat{w}_t - \overline{w}_t\right) \tag{15}$$

其中，$\bar{g}_t^i$是原始模型的梯度，$\hat{w}_t$是恢复的全局模型，$\bar{w}_t$是原始的全局模型，$\boldsymbol{H}_t^i = \boldsymbol{H}\left(\overline{w}_t + z(\hat{w}_t - \overline{w}_t)\right)\mathrm{d}z$是第$t$轮中第$i$个客户端的Hessian矩阵，梯度$g$是模型参数$w$的函数。另外，文献[76]进一步优化FedRecover，使用预热、定期校正、异常修复和最终调整策略来恢复更准确的全局模型。从理论上，证明了利用FedRecover恢复的全局模型与传统的从零开始训练迭代整个全局模型的效果类似。Zhang等[77]提出了FedRecovery，其核心思想是从训练好的全局模型中去除了未参加学习客户端的影响，并采用高斯噪声来掩盖未参加学习参数和重新训练参数之间的差异。该方法实现过程相对简单，容易部署，但模型恢复的效果相较于优化FedRecover低。

拜占庭攻击检测技术可以发现恶意客户端并防止其进一步影响全局模型，但在被检测到之前，恶意客户端可能已经对联邦学习模型造成了负面影响。针对这一问题，可以利用模型恢复技术，在发现恶意客户端后，排除其已经对模型施加的影响，纠正联邦学习模型偏差并恢复模型性能，进一步提高模型准确率。就目前而言，模型恢复技术研究和实验尚少，其原因是需要了解模型机理和基于深度学习模型消除数据贡献等难题。

2.5　小结

结合上述，本节从攻防框架和攻防机理2个维度详细进行阐述。考虑到联邦学习实际场景中的安

全需求，介绍了联邦学习拜占庭防御中APPDRR环节的具体含义，对现有防御方法进行详细总结和分类。在联邦学习模型的训练阶段对系统进行安全评估；在本地客户端通过利用本地模型参数间以及客户端的历史更新和预测更新进行自我防护以及过滤恶意更新；在训练过程中利用特征的不一致性进行检测恶意更新，如考虑基准可信数据、上传参数的大小和方向、异常检测算法、两两间、基准参数和参数相对分布的相似性等；在聚合阶段及时响应，基于信用、投票或者直接删除恶意更新不参与全局模型聚合；最后，考虑恶意客户端的历史贡献也会对局部模型以及全局模型产生影响，进行模型恢复。通过上述环节形成对联邦学习系统的闭环保护，为应用联邦学习解决实际问题提供安全保障。

3 未来研究展望

本文基于APPDRR对联邦学习拜占庭攻击的原理以及相应的防御方法进行体系化的分析。结果表明，现有研究主要集中在恶意检测和异常响应等环节，但在安全评估、安全策略、安全防护和模型恢复等模块还缺少相关的研究。其原因在于，安全评估、安全策略和安全防护环节不仅需要研究大量的攻防机理，还需考虑攻击策略、前提假设和参数设定对模型的影响，以及攻击隐蔽性、强度和安全性等评估指标的标准化，这些因素使得研究相对复杂，且对服务器设备的要求高、资源消耗大。模型恢复环节则面临研究模型机理及消除恶意客户端对全局模型负面影响的挑战。此外，从攻防对抗视角，学者提出了很多关于联邦学习拜占庭攻防的研究，但对攻击和防御的强度划分、安全度量评估的研究较少。当攻防假设条件变动甚至参数未知的情况以及复杂的攻防场景下，研究鲁棒性强、隐蔽性高极难防御的盲攻击和联邦学习模型训练全过程的安全保护问题。最后，考虑了在Non-IID、联邦学习在推荐系统中应用场景和纵向联邦学习中的安全防护问题。为此，本节梳理了联邦学习拜占庭攻防未来研究的重点方向。

3.1 联邦学习模型的安全评估研究

联邦学习中的拜占庭攻击会对模型的准确率以及对特定类别的预测结果产生影响，这严重影响联邦学习的推广应用。从攻防框架的视角，亟须开展面向联邦学习拜占庭攻击的安全度量与风险评估研究，对联邦学习系统遭受攻击时可能带来的攻击成功率、主任务精度受损情况等进行量化分析，形成可度量、可比较的联邦学习安全评估方法。目前，已有部分关于联邦学习可信评估的研究[88-90]。在未来的工作中，还需要进一步加强联邦学习安全度量和评估研究，制定合适的评估联邦学习系统安全性的度量指标，例如当系统受到拜占庭攻击时模型准确性的损失值、联邦学习系统模型能够容忍的攻击类型以及能够容忍的攻击强度等，量化联邦学习系统面临的风险和威胁。此外，还需要建立攻击防御模型和安全度量指标之间的关系，以此指导安全攸关和非安全攸关场景下，联邦学习中的防御方法的研究。

3.2 联邦学习拜占庭攻防博弈策略研究

拜占庭攻击的攻防博弈策略旨在探索在联邦学习环境下，攻击者和防御者之间的博弈行为，并提出相应的策略和机制来应对拜占庭攻击，从而保护系统的安全性和隐私性。现有的防御方法在攻击者的策略和一些先验条件等方面通常做出简化，例如假设受控的多个恶意客户端同时实施攻击[22]、假设恶意客户端会在模型训练过程中一直实施攻击[16]等。事实上，攻击者可以在攻击过程中，依据联邦学习攻击需求制定复杂的攻击策略，如攻击目标的优化选择、攻击方式的组合选择、攻击强度和攻击时机的动态调整等。因此，当攻击场景呈现动态变化的特性时，仅依靠传统的单一防御方案，很难对联邦学习系统提供有效的安全保障，如何建立攻防博弈模型，针对复杂的攻击策略提出有效的防御策略成为关键问题。在复杂攻击场景下，如何根据攻防评估指标、安全度量和风险评估以及真实场景中的安全需求制定相应的安全策略，开展联邦学习拜占庭攻击的动态博弈研究将是未来的重要研究方向。

3.3 Non-IID数据的鲁棒拜占庭防御研究

先前许多研究者提出的防御方法假设数据分布为IID。基于这一假设，这类防御方法的核心思想为排除少数的梯度更新，关注大多数良性目标的梯度更新，对梯度更新参数进行操作，选取良性更新作为聚合梯度。然而，在实际联邦学习应用场景中，Non-IID [91]是不同用户协同训练必然需要面对的问题[92]。它将引发本地模型的更新参数值发散[93]，使得传统基于IID假设的鲁棒聚合防御方法

不再适用[94]。在 Non-IID 场景下，如 Krum、Median、trimmed-mean 等不仅不能防御拜占庭攻击，还会降低全局模型的精度。针对数据分布为 Non-IID 假设下，研究联邦学习的防御方法一直是个难题[95]。Zhao 等[95]为了适应数据分布的不平衡 Non-IID 的情况，设计了动态客户端分配机制，将检测任务分配给最合适的客户端，但该防御方法增加了各个客户端的通信量。因此，如何探究新的防御解决方案，结合聚合策略[96-98]、知识蒸馏[99-100]和异质任务[101-103]等技术，提升联邦学习在 Non-IID 场景下的防御性能和效果是当前亟待解决的关键问题。在这一领域中，如何区分恶意用户上传的梯度更新与由于数据异质性导致的良性用户上传的分散梯度值，成为未来提升模型应对拜占庭攻击的鲁棒性重要挑战。

3.4　联邦推荐场景中拜占庭防御研究

联邦推荐作为联邦学习在推荐系统的一个应用场景，将用户的敏感交互记录在本地，极大地保护了用户隐私信息。然而，由于大规模分布式的特性以及数据的高度个性化，使得攻击者更容易通过数据投毒等方式操纵全局模型的训练。因此，基于联邦学习框架的推荐算法较联邦学习模型更容易受到低成本的攻击方法的破坏。研究联邦推荐系统防御机制将为联邦推荐算法的安全性提供保障。近期，已有学者初步做了一些工作，如文献[69]提出了一种鲁棒学习策略来防御联邦推荐系统中的拜占庭攻击。该方法首先证明了拜占庭客户端可以发动有效攻击来伪造模型参数并逃避现有的防御方法的问题。并针对此问题，提出一种鲁棒学习策略，即服务器端不使用模型参数，而是计算并使用梯度来过滤掉拜占庭客户端。目前，关于联邦推荐中的拜占庭防御仍处于早期阶段。同时，对于存在恶意的客户端和服务器，或者存在一些数据质量较低的客户端，如何设计出提升联邦推荐性能的防御方法，将是一个值得研究的问题。

3.5　纵向联邦学习场景中拜占庭防御研究

在当前的研究背景下，尽管横向联邦学习的拜占庭攻击和防御方法已经得到了广泛的关注和研究[23-70]，但是对于纵向联邦学习而言，关于拜占庭攻防方法的研究还相对较少[96, 104]。纵向联邦学习与横向联邦学习相比最大的不同是，不同的客户端拥有不同的数据特征，其他客户端是不知道恶意客户端的数据特征的，在此条件下恶意数据或恶意梯度的检测将缺乏依据。针对此问题，Lai 等[104]提出了纵向联邦学习投毒防御（VFedAD，vertical federated learning poisoning defense）方法，通过对比学习任务和跨客户端预测任务来学习语义丰富的客户端数据表示，以识别异常。未来在此方向，还需要设计面向纵向联邦学习的普适性的拜占庭防御方法。

4　结束语

随着联邦学习技术的不断发展和广泛应用，联邦学习模型的安全问题日益严重。本文以拜占庭攻击为例，对联邦学习最新的攻击原理进行细化分类与剖析，总结了常用的评估指标。进一步地，本文以经典的网络安全防御模型 APPDRR 为指导，从防御机制的角度针对联邦学习攻击防御方法进行了分类和分析。并在此基础上，梳理了联邦学习拜占庭攻防未来研究的重点方向，为未来相关研究者提供了新的参考。

参考文献：

[1] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: distributed machine learning for on-device intelligence[J]. arXiv Preprint, arXiv: 1610.02527, 2016.

[2] HARD A, RAO K, MATHEWS R, et al. Federated learning for mobile keyboard prediction[J]. arXiv Preprint, arXiv: 1811.03604, 2018.

[3] YANG T, ANDREW G, EICHNER H, et al. Applied federated learning: improving google keyboard query suggestions[J]. arXiv Preprint, arXiv: 1812.02903, 2018.

[4] ANTUNES R S, ANDRÉ DA COSTA C, KÜDERLE A, et al. Federated learning for healthcare: systematic review and architecture proposal[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(4): 1-23.

[5] NIKNAM S, DHILLON H S, REED J H. Federated learning for wireless communications: motivation, opportunities, and challenges[J]. IEEE Communications Magazine, 2020, 58(6): 46-51.

[6] YANG Z H, CHEN M Z, WONG K K, et al. Federated learning for 6G: applications, challenges, and opportunities[J]. Engineering, 2022, 8: 33-41.

[7] SHI J B, ZHAO H J, WANG M Y, et al. Signal recognition based on federated learning[C]//Proceedings of the IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Piscataway: IEEE Press, 2020: 1105-1110.

[8] PREUVENEERS D, RIMMER V, TSINGENOPOULOS I, et al. Chained anomaly detection models for federated learning: an intrusion detection case study[J]. Applied Sciences, 2018, 8(12): 2663.

[9] KHRAMTSOVA E, HAMMERSCHMIDT C, LAGRAA S, et al. Federated learning for cyber security: SOC collaboration for malicious URL

detection[C]//Proceedings of the 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2020: 1316-1321.

[10] LIU Y, FAN T, CHEN T J, et al. FATE: an industrial grade platform for collaborative learning with data protection[J]. Journal of Machine Learning Research, 2021, 22(226): 1-6.

[11] RYFFEL T, TRASK A, DAHL M, et al. A generic framework for privacy preserving deep learning[J]. arXiv Preprint, arXiv: 1811.04017, 2018.

[12] 马艳军, 于佃海, 吴甜, 等. 飞桨: 源于产业实践的开源深度学习平台[J]. 数据与计算发展前沿, 2019, 1(1): 105-115.
MA Y J, YU D H, WU T, et al. PaddlePaddle: an open-source deep learning platform from industrial practice[J]. Frontiers of Data and Computing, 2019, 1(1): 105-115.

[13] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: system design[J]. arXiv Preprint, arXiv: 1902.01046, 2019.

[14] NGUYEN T D, RIEGER P, VITI R D, et al. {FLAME}: Taming backdoors in federated learning[C]//Proceedings of the 31st USENIX Security Symposium. Berkeley: USENIX Association, 2022: 1415-1432.

[15] BARUCH M, BARUCH G, GOLDBERG Y. A little is enough: circumventing defenses for distributed learning[C]//Proceedings of the 33rd International Conference on Neural Information. Processing Systems. New York: ACM Press, 2019: 8635-8645.

[16] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems[C]//Proceedings of the 25th European Symposium on Research in Computer Security. Berlin: Springer, 2020: 480-501.

[17] WU C H, WU F Z, QI T, et al. FedAttack: effective and covert poisoning attack on federated recommendation via hard sampling[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2022: 4164-4172.

[18] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to Byzantine-robust federated learning[C]//Proceedings of the 29th USENIX security symposium (USENIX Security 20). Berkeley: USENIX Association, 2020: 1605-1622.

[19] SHEJWALKAR V, HOUMANSADR A. Manipulating the Byzantine: optimizing model poisoning attacks and defenses for federated learning[C]//Proceeding of the 2021 Network and Distributed System Security Symposium. Reston: Internet Society, 2021: 1-18.

[20] ZHANG S J, YIN H Z, CHEN T, et al. PipAttack: poisoning federated recommender systems for manipulating item promotion[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2022: 1415-1423.

[21] RONG D Z, HE Q M, CHEN J H. Poisoning deep learning based recommender model in federated learning scenarios[J]. arXiv Preprint, arXiv: 2204.13594, 2022.

[22] FUNG C, YOON C J, BESCHASTNIKH I. The limitations of federated learning in sybil settings[C]//Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses. Berkeley: USENIX Association, 2020: 301-316.

[23] YU Y, LIU Q, WU L K, et al. Untargeted attack against federated recommendation systems via poisonous item embeddings and the defense[C]//Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM Press, 2023: 4854-4863.

[24] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learnings[J]. arXiv Preprint, arXiv: 1807.00459, 2018.

[25] WANG H Y, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: yes, you really can backdoor federated learning[J]. arXiv Preprint, arXiv: 2007.05084, 2020.

[26] XIE C, KOYEJO O, GUPTA I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation[J]. arXiv Preprint, arXiv: 1903.03936, 2019.

[27] 陈晓霖, 昝道广, 吴炳潮, 等. 面向纵向联邦学习的对抗样本生成算法[J]. 通信学报, 2023, 44(8): 1-13.
CHEN X L, ZAN D G, WU B C, et al. Adversarial sample generation algorithm for vertical federated learning[J]. Journal on Communications, 2023, 44(8): 1-13.

[28] CHEN Y D, SU L L, XU J M. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2017, 2(1):1-25.

[29] HE Y Z, MENG G Z, CHEN K, et al. Towards security threats of deep learning systems: a survey[J]. IEEE Transactions on Software Engineering, 2022, 48(5): 1743-1770.

[30] 杨丽, 朱凌波, 于越明, 等. 联邦学习与攻防对抗综述[J]. 信息网络安全, 2023, 23(12): 69-90.
YANG L, ZHU L B, YU Y M, et al. Review of federal learning and offensive-defensive confrontation[J]. Netinfo Security, 2023, 23(12): 69-90.

[31] 高莹, 陈晓峰, 张一余, 等. 联邦学习系统攻击与防御技术研究综述[J]. 计算机学报, 2023, 46(9): 1781-1805.
GAO Y, CHEN X F, ZHANG Y Y, et al. A survey of attack and defense techniques for federated learning systems[J]. Chinese Journal of Computers, 2023, 46(9): 1781-1805.

[32] GUO S W, ZHANG X, YANG F, et al. Robust and privacy-preserving collaborative learning: a comprehensive survey[J]. arXiv Preprint, arXiv: 2112.10183, 2021.

[33] 顾育豪, 白跃彬. 联邦学习模型安全与隐私研究进展[J]. 软件学报, 2023, 34(6): 2833-2864.
GU Y H, BAI Y B. Survey on security and privacy of federated learning models[J]. Journal of Software, 2023, 34(6): 2833-2864.

[34] 陈学斌, 任志强, 张宏扬. 联邦学习中的安全威胁与防御措施综述[J]. 计算机应用, 2024, 44(6): 1663-1672.
CHEN X B, REN Z Q, ZHANG H Y. Review on security threats and defense measures in federated learning[J]. Journal of Computer Applications, 2024, 44(6): 1663-1672.

[35] WAN Y C, QU Y Y, NI W, et al. Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: a comprehensive survey[J]. IEEE Communications Surveys and Tutorials, 2024, 26(3): 1861-1897.

[36] ROSZEL M, NORVILL R, STATE R. An Analysis of Byzantine-Tolerant Aggregation Mechanisms on Model Poisoning in Federated Learning[C]//Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence. Berlin: Springer, 2022: 143-155.

[37] 孙钰, 刘霏霏, 李大伟, 等. 联邦学习拜占庭攻击与防御研究综述[J]. 网络空间安全科学学报, 2023(1): 17-37.
SUN Y, LIU F F, LI D W, et al. Survey on Byzantine attacks and defenses in federated learning[J]. Journal of Cybersecurity, 2023(1): 17-37.

[38] CAO X Y, LAI L F. Distributed gradient descent algorithm robust to an arbitrary number of Byzantine attackers[J]. IEEE Transactions on Signal Processing, 2019, 67(22): 5850-5864.

[39] XIAO H, XIAO H, ECKERT C. Adversarial label flips attack on support vector machines[C]//Proceedings of the 20th European Conference on Artificial Intelligence. New York: ACM Press, 2012: 870-875.

[40] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 118-128.

[41] KASYAP H, TRIPATHY S. Sine: similarity is not enough for mitigating local model poisoning attacks in federated learning[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(5): 4481-4494.

[42] ROSENFELD E, WINSTON E, RAVIKUMAR P, et al. Certified robustness to label-flipping attacks via randomized smoothing[J]. arXiv Preprint, arXiv: 2002.03018, 2020.

[43] XU J, WANG R, KOFFAS S, et al. More is better (mostly): on the backdoor attacks in federated graph neural networks[C]//Proceedings of the 38th Annual Computer Security Applications Conference. New York: ACM Press, 2022: 684-698.

[44] XU J, HUANG S L, SONG L Q, et al. Byzantine-robust federated learning through collaborative malicious gradient filtering[C]//Proceedings of the 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2022: 1223-1235.

[45] AWAN S N, LUO B, LI F J. CONTRA: defending against poisoning attacks in federated learning[C]//Proceedings of the 26th European Symposium on Research in Computer Security. Berlin: Springer, 2021: 455-475.

[46] PRAKASH S, AVESTIMEHR A. Mitigating Byzantine attacks in federated learning[J]. arXiv Preprint, arXiv: 2010.07541, 2020.

[47] WEI K, LI J, DING M, et al. Covert model poisoning against federated learning: algorithm design and optimization[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(3): 1196-1209.

[48] YANG H, XI W, SHEN Y H, et al. RoseAgg: robust defense against targeted collusion attacks in federated learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2951-2966.

[49] ZHANG H, JIA J, CHEN J, et al. A3FL: adversarially adaptive backdoor attacks to federated learning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: ACM Press, 2024: 61213-61233.

[50] WU R H, CHEN X Y, GUO C, et al. Learning to invert: simple adaptive attacks for gradient inversion in federated learning[C]//Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence. New York: PMLR, 2023. 2293-2303.

[51] JIN T S, FU Z H, MENG D, et al. FedPerturb: covert poisoning attack on federated learning via partial perturbation[C]//Proceedings of the International Conference on Artificial Intelligence and Applications. Amsterdam: IOS Press, 2023:1172-1179.

[52] ZHANG H T, YAO Z M, ZHANG L Y, et al. Denial-of-service or fine-grained control: towards flexible model poisoning attacks on federated learning[J]. arXiv Preprint, arXiv: 2304.10783, 2023.

[53] SHEN S Q, TOPLE S, SAXENA P. Auror: defending against poisoning attacks in collaborative deep learning systems[C]//Proceedings of the 32nd Annual Conference on Computer Security Applications. New York: ACM Press, 2016: 508-519.

[54] GONG Z R, SHEN L Y, ZHANG Y J, et al. Agramplifier: defending federated learning against poisoning attacks through local update amplification[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 1241-1250.

[55] ZHU C, ROOS S, CHEN L Y. LeadFL: client self-defense against model poisoning in federated learning[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR, 2023. 43158-43180.

[56] GUO X T, WANG P F, QIU S, et al. FAST: adopting federated unlearning to eliminating malicious terminals at server side[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(2): 2289-2302.

[57] XIE C, CHEN M, CHEN P, et al. Crfl: certifiably robust federated learning against backdoor attacks[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR, 2021. 11372-11382

[58] WU L T, YUE M Q, ZHANG H B, et al. A network survivability evaluation method based on PDRR model[C]//Proceedings of the 2023 International Conference on Electronics and Devices, Computational Science (ICEDCS). Piscataway: IEEE Press, 2023: 522-526.

[59] 潘洁, 刘爱洁. 基于 APPDRR 模型的网络安全系统研究[J]. 电信工程技术与标准化, 2009, 22(7): 27-30.
PAN J, LIU A J. Study of APPDRR model-based network security system[J]. Telecom Engineering Technics and Standardization, 2009, 22(7): 27-30.

[60] ZHANG Z X, CAO X Y, JIA J Y, et al. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2022: 2545-2555.

[61] SUN J W, LI A, DIVALENTIN L, et al. FL-WBC: enhancing robustness against model poisoning attacks in federated learning from a client perspective[J]. Advances in Neural Information Processing Systems, 2021, 34: 12613-12624.

[62] PAND A, MAHLOUJIFAR S, BHAGOJI A, et al. SparseFed: mitigating model poisoning attacks in federated learning with sparsification[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2022: 7587-7624.

[63] MOZAFFARI H, SHEJWALKAR V, HOUMANSADR A. Every vote counts: ranking-based training of federated learning to resist poisoning attacks[C]//Proceedings of the 32nd USENIX Security Symposium. Berkeley: USENIX Association, 2023: 1721-1738.

[64] CAO X Y, FANG M H, LIU J, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping[J]. arXiv Preprint, arXiv: 2012.13995, 2020.

[65] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[J]. arXiv Preprint,

arXiv: 1803.01498, 2018.

[66] MHAMDI E M E, GUERRAOUI R, ROUAULT S. The hidden vulnerability of distributed learning in Byzantium[J]. arXiv Preprint, arXiv: 1802.07927, 2018.

[67] PILLUTLA K, KAKADE S M, HARCHAOUI Z. Robust aggregation for federated learning[J]. IEEE Transactions on Signal Processing, 2022, 70: 1142-1154.

[68] YAN G, WANG H, YUAN X, et al. DeFL: defending against model poisoning attacks in federated learning *via* critical learning periods awareness[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 10711-10719.

[69] MUÑOZ-GONZÁLEZ L, CO K T, LUPU E C. Byzantine-robust federated machine learning through adaptive model averaging[J]. arXiv Preprint, arXiv: 1909.05125, 2019.

[70] LIU X Y, LI H W, XU G W, et al. Privacy-enhanced federated learning against poisoning adversaries[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4574-4588.

[71] MA Z R, MA J F, MIAO Y B, et al. ShieldFL: mitigating model poisoning attacks in privacy-preserving federated learning[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 1639-1654.

[72] LI S Y, CHENG Y, WANG W, et al. Learning to detect malicious clients for robust federated learning[J]. arXiv Preprint, arXiv: 2002.00211, 2020.

[73] LI S Y, CHENG Y, LIU Y, et al. Abnormal client behavior detection in federated learning[J]. arXiv Preprint, arXiv: 1910.09933, 2019.

[74] JIANG Y F, ZHANG W W, CHEN Y X. Data quality detection mechanism against label flipping attacks in federated learning[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1625-1637.

[75] LI X Y, QU Z, ZHAO S Q, et al. LoMar: a local defense against poisoning attack on federated learning[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(1): 437-450.

[76] CAO X Y, JIA J Y, ZHANG Z X, et al. FedRecover: recovering from poisoning attacks in federated learning using historical information[C]//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2023: 1366-1383.

[77] ZHANG L F, ZHU T Q, ZHANG H B, et al. FedRecovery: differentially private machine unlearning for federated learning frameworks[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4732-4746.

[78] YAN H N, ZHANG W J, CHEN Q, et al. RECESS vaccine for federated learning: proactive defense against model poisoning attacks[J]. arXiv Preprint, arXiv: 2310.05431, 2023.

[79] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint, arXiv: 1602.05629, 2016.

[80] XIA Q, TAO Z Y, HAO Z J, et al. FABA: an algorithm for fast aggregation against Byzantine attacks in distributed neural networks[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2019: 4824-4830.

[81] SHEN L Y, KE Z H, SHI J Q, et al. SPEFL: efficient security and privacy-enhanced federated learning against poisoning attacks[J]. IEEE Internet of Things Journal, 2024, 11(8): 13437-13451.

[82] HAN S, WU W, BUYUKATES B, et al. Kick bad guys out! zero-knowledge-proof-based anomaly detection in federated learning[J]. arXiv Preprint, arXiv: 2310.04055. 2023.

[83] JIANG Y, SHEN J Y, LIU Z Y, et al. Towards efficient and certified recovery from poisoning attacks in federated learning[J]. arXiv Preprint, arXiv: 2401.08216, 2024.

[84] ZHANG X Y, LIU Q Y, BA Z J, et al. FLTracer: accurate poisoning attack provenance in federated learning[J]. arXiv Preprint, arXiv: 2310.13424, 2023.

[85] WANG J X, GUO S, XIE X, et al. Federated unlearning *via* class-discriminative pruning[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 622-632.

[86] XIA H, XU S, PEI J M, et al. FedME2: memory evaluation & erase promoting federated unlearning in DTMN[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(11): 3573-3588.

[87] WU C, ZHU S C, MITRA P. Federated unlearning with knowledge distillation[J]. arXiv Preprint, arXiv: 2201.09441, 2022.

[88] 刘晗, 李凯旋, 陈仪香 . 人工智能系统可信性度量评估研究综述[J]. 软件学报, 2023, 34(8): 3774-3792.
LIU H, LI K X, CHEN Y X. Survey on trustworthiness measurement for artificial intelligence systems[J]. Journal of Software, 2023, 34(8): 3774-3792.

[89] WEI W Q, LIU L. Trustworthy distributed AI systems: robustness, privacy, and governance[J]. ACM Computing Surveys, 2024, 152: 83-98.

[90] ZHANG Y F, ZENG D, LUO J L, et al. A survey of trustworthy federated learning with perspectives on security, robustness, and privacy[J]. arXiv Preprint, arXiv: 2302.10637, 2023.

[91] LU Z L, PAN H, DAI Y Y, et al. Federated learning with non-IID data: a survey[J]. IEEE Internet of Things Journal, 2024, 11(11): 19188-19209.

[92] PENG B, CHI M M, LIU C. Non-IID federated learning via random exchange of local feature maps for textile IIoT secure computing[J]. Science China Information Sciences, 2022, 65(7): 170302.

[93] ZHAO Y, LI M, LAI L Z, et al. Federated learning with non-IID data[J]. arXiv Preprint, arXiv:1806.00582, 2018.

[94] 马鑫迪, 李清华, 姜奇, 等 . 面向 Non-IID 数据的拜占庭鲁棒联邦学习[J]. 通信学报, 2023, 44(6): 138-153.
MA X D, LI Q H, JIANG Q, et al. Byzantine-robust federated learning over Non-IID data[J]. Journal on Communications, 2023, 44(6): 138-153.

[95] ZHAO R J, WANG Y J, XUE Z, et al. Semisupervised federated-learning-based intrusion detection method for Internet of Things[J]. IEEE Internet of Things Journal, 2023, 10(10): 8645-8657.

[96] QIU P Y, ZHANG X H, JI S L, et al. Hijack vertical federated learning models as one party[J]. IEEE Transactions on Dependable and Secure Computing, 2024(99): 1-18.

[97] ZHANG J Q, HUA Y, WANG H, et al. FedALA: adaptive local aggregation for personalized federated learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 11237-11244.

[98] MORAFAH M, VAHIDIAN S, WANG W J, et al. FLIS: clustered federated learning *via* inference similarity for non-IID data distribution[J].

IEEE Open Journal of the Computer Society, 2023, 4: 109-120.

[99] WEN H, WU Y, HU J, et al. Communication-efficient federated learning on non-IID data using two-step knowledge distillation[J]. IEEE Internet of Things Journal, 2023, 10(19): 17307-17322.

[100] ALSENANI Y, MISHRA R, AHMED K R, et al. FedSiKD: clients similarity and knowledge distillation: addressing non-i.i.d. and constraints in federated learning[J]. arXiv Preprint, arXiv: 2402.09095, 2024.

[101] CHEN H K, FRIKHA A, KROMPASS D, et al. FRAug: tackling federated learning with non-IID features via representation augmentation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 4826-4836.

[102] YANG L, HUANG J M, LIN W Y, et al. Personalized federated learning on non-IID data via group-based meta-learning[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(4): 1-20.

[103] LI Z J, SUN Y C, SHAO J W, et al. Feature matching data synthesis for non-IID federated learning[J]. IEEE Transactions on Mobile Computing, 2024, 23(10): 9352-9367.

[104] LAI J R, WANG T, CHEN C, et al. VFedAD: a defense method based on the information mechanism behind the vertical federated data poisoning attack[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2023: 1148-1157.

[作者简介]

赵晓洁（1997– ），女，山东菏泽人，北京邮电大学博士生，主要研究方向为联邦学习、攻防算法、隐私计算、人工智能安全等。

时金桥（1978– ），男，黑龙江哈尔滨人，博士，北京邮电大学教授，主要研究方向为隐私保护、人工智能安全、匿名通信技术等。

黄梅（1997– ），女，广西桂平人，北京邮电大学博士生，主要研究方向为加密流量分类、联邦学习等。

柯镇涵（2000– ），男，湖北十堰人，北京邮电大学硕士生，主要研究方向为安全多方计算、联邦学习等。

申立艳（1992– ），女，河北保定人，博士，北京信息科技大学副教授，主要研究方向为人工智能安全、隐私计算、安全多方计算、联邦学习等。