目录

1 论文的研究背景及主要贡献

1.1 研究背景

本文中，我们专注于联邦学习（Federated Learning, FL）环境中一个挑战性的问题——后门攻击（Backdoor Attacks）。这类攻击是一种特定的目标化投毒攻击，联邦学习旨在通过维护用户隐私的同时，集合多个客户端的数据和学习成果来构建强大的机器学习模型，这就给了攻击者可乘之机，攻击者可以将操纵的模型更新注入到联邦模型聚合过程中，以便生成的模型会对一些特定选择的输入（也就是我们说的后门）提供有针对性的错误预测。这种攻击威胁到了联邦学习模型的安全性和可靠性。后门攻击的独特之处在于攻击者通过操纵少数受损的客户端模型来影响整个全局模型，这对于如何设计有效的防御机制提出了特别的挑战。特别是在 FL 环境中，保持全局模型的良好性能至关重要，同时需要从众多潜在的模型更新中识别和消除恶意贡献，而不依赖于对数据分布或攻击策略的具体假设。因此，研究者需要探索新的方法来解决后门攻击带来的威胁，而这正是本文所致力于的。

1.2 主要贡献

- **弹性防御机制的提出**: 本研究提出了 FLAME，一个针对后门攻击的弹性聚合框架，特别适用于联邦学习（FL）环境。这个框架通过结合差分隐私（Differential Privacy, DP）和其他智能聚合策略，旨在既减少后门攻击所需的噪声量，又不显著降低聚合模型的良性性能。

- **降低所需高斯噪声量**: 通过两种方式显著减少所需的高斯噪声量：a）应用聚类方法移除可能的恶意模型更新；b）在适当水平剪裁本地模型的权重，以限制单个模型（尤其是恶意模型）对聚合模型的影响。这种方法不仅提高了模型防御后门攻击的能力，还通过减少所需噪声的量，保持了模型的良好性能。

- **基于差分隐私（DP）的噪声注入量的噪声下界证明**:FLAME 通过受 DP 启发的方式注入高斯噪声来消除后门贡献，并为所需高斯噪声的量提供了噪声下界证明。这一理论基础确保了噪声注入的策略既能有效防御后门攻击，又能尽可能减少对模型性能的影响。

- **成功抵御新型攻击**: 在本文发表后，有研究者提出新型攻击手段，声称能够绕过 FLAME 的保护。通过详尽调查，本文作者不仅找出了这些攻击策略的缺陷，还与相关作者进行了对话，使他们承认并更正了错误主张。这一过程展示了 FLAME 防御机制的有效性以及作者团队对确保其研究准确性和可靠性的承诺。

- **实际应用的验证**: 通过在真实世界数据集上的广泛评估，证明了 FLAME 在不同应用领域中减少后门攻击所需噪声的同时，保持了聚合模型良性性能的有效性。这为 FLAME 框架的实用性和适用性提供了有力的实证支持。

2 系统架构

在介绍 FLAME（一个针对后门攻击的弹性聚合框架）的系统架构时，我们可以聚焦于以下几个关键组成部分和设计挑战：

1. **动态模型过滤（Dynamic Model Filtering）**:FLAME 用于识别和过滤含有后门的模型。这些模型在权重向量的角度分布上与良性模型有较大差异。通过聚类方法，可以识别这些被投毒的模型并将它们从联邦学习（FL）聚合中移除。这种基于聚类的过滤考虑到了动态攻击场景，即注入后门的数量不定。

2. **自适应剪裁（Adaptive Clipping）和自适应加噪（Adaptive Noising）**：这些方法旨在减少所需噪声量，同时保持聚合模型的良性性能。尤其自适应加噪是关键的防御机制，用于消除剩余的后门更新，而这一点在许多针对 FLAME 的攻击设计中却没有被充分考虑。

3. **安全多方计算（Secure Multi-Party Computation，SMPC）**：这是用来保护模型更新免受诚实但好奇的聚合器侵犯隐私的方法。SMPC 旨在在不透露个别模型的详细信息的情况下，实现模型更新的隐私保护。

4. **实验验证**:FLAME 的有效性通过在三个不同应用领域的真实世界数据集上的广泛评估得到验证。实验结果表明，FLAME 在减少后门攻击所需噪声的同时，能够保持聚合模型良性性能的优势。

5. **后门和主任务准确率指标（Backdoor Accuracy, BA 和 Main-task Accuracy, MA）**：通过重复实验和使用不同的随机种子值，确保了实验的鲁棒性与可靠性。FLAME 显著减少了 BA，即使在全局模型接近收敛时，也说明 FLAME 有效地缓解了 3DFed 攻击的影响。

3 研究现状

在本文中，主要聚焦在后门攻击（backdoor attacks）的问题上 [1, 2, 3, 4]，这类攻击为有目标的毒化攻击，攻击者意图潜移默化地操纵最终的全球模型，以致于攻击者控制的输入导致的预测结果不准确。而现有的针对后门攻击的防御措施可以大致分为两类：

- **基于异常检测的方法**: 这种方法尝试识别并移除可能被毒化的模型更新。然而，异常检测的方法往往基于特定的对手模型，并且对攻击策略或数据集的分布做出了详尽的假设。例如，参考文献 [5, 6, 7, 8] 探讨了这种类别的防御。当这些特定的假设不成立时，防御可能会失败。

- **独立于数据分布和攻击策略的方法**: 理想的防御措施需要能够适应通用的敌手模型，而不需要关于后门攻击方法的先验知识，也不会假设局部客户机的特定数据分布，例如数据是独立同分布 (iid) 还是非独立同分布 (non-iid)。

文章提出了 FLAME（一个弹性聚合框架），旨在弥补现有方法的不足，并通过动态聚类、自适应剪切和噪声以及安全多方计算等技术手段有效抵御后门攻击。FLAME 的动机来源于之前的工作，如 Sun et al.[9] 使用类差分隐私的聚合模型噪声化来消除后门，这在联邦学习（FL）设置中由于无法假设聚合器能够访问训练数据，尤其是被投毒的数据集，因此存在挑战。

此外，自 FLAME 在 2022 年的安全会议上被引入 [10] 后，此方法已经引起了研究者的极大兴趣，并根据 Google Scholar[11] 获得了超过 100 次引用。论文还讨论了两次声称可以规避 FLAME 的攻击 [12, 13]，并进行了详细的调查，发现这些攻击方法存在显著的错误，并与这些作者进行了沟通，并最终获得了 Xu et al.[13] 对声称的修正。

4 先验知识

后门攻击 (Backdoor Attacks)：这是针对机器学习模型的一类攻击，其中攻击者通过在训练数据中注入恶意样本来操纵模型学习一个特定的错误行为。这种攻击在文档中被定义为目标性的毒化攻击 (targeted poisoning attacks)，旨在悄然操纵全局模型，使得攻击者控制的输入导致模型做出攻击者指定的错误预测 [1, 2, 3, 4]。

异常检测 (Anomaly Detection)：这是一类方法，用于识别并移除可能被毒化的模型更新。这些防御机制在特定的敌手模型下是有效的，它们对攻击策略和/或良性或恶意数据集的底层分布做出了详细的假设 [5, 6, 7, 8]。

聚类 (Clustering)：聚类是一种将数据点分组的技术，使得同一组内的数据点相比于其他组的数据点更为相似。文档中指出现有基于聚类的防御方法需要能够在动态攻击场景下工作，即注入的后门数量未知且可能在训练轮次间变化。为了解决这一挑战，引入了一种动态确定模型更新聚类的解决方案，使其能够适应动态攻击。

高斯噪声 (Gaussian Noise) 应用：通过在模型参数上注入高斯噪声来减少所需的噪声量的方法，在这里通过应用聚类方法移除潜在恶意的模型更新和在适当的水平剪切本地模型的权重来实现，制约了个别（特别是恶意的）模型对聚合模型的影响。

差分隐私 (Differential Privacy, DP)：为保护用户隐私，差分隐私技术限制了从发布的数据中提取个人信息的能力。文档提到，通过噪声注入所需的高斯噪声量可以根据差分隐私提供的一个噪声边界证明来显著减少，以消除后门贡献。

联邦学习 (Federated Learning)：这是一种分布式机器学习方法，允许多个客户端协作训练一个共享的模型，同时保持其训练数据的隐私。FLAME 框架就是在联邦学习的上下文中被提出，以对抗后门攻击，它不依赖于对攻击策略或底层数据分布的知识。

5 具体方案和原理

5.1 问题具体解决方案

论文的主题集中于联邦学习环境中的后门攻击问题，这些攻击旨在悄然操纵全局模型，使得攻击者能够控制输入，导致模型做出错误的预测。针对这一挑战，论文提出了一个名为 FLAME 的新型防御方法，其设计理念如下：

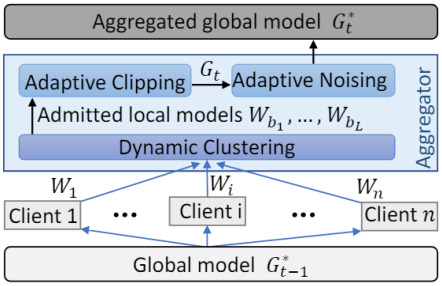

图 1: FlAME 工作流程

1. **动态估算噪声下界**:FLAME 的核心特性之一是能够在不需要访问训练数据，特别是被投毒数据集的条件下，对聚合模型进行去噪。FLAME 通过一种可靠的方法估算所需噪声的最小但充足边界，该方法不依赖于对攻击策略或数据分布的先验知识。

2. **三大组成部分—过滤、剪裁和噪声注入**:FLAME 的防御策略由三个主要组成部分构成：

   - **过滤（Filtering）**：采用聚类方法识别并排除潜在的恶意模型更新。
   - **剪裁（Clipping）**：适当剪裁本地模型的权重，限制个别模型（特别是恶意模型）对聚合模型的影响。
   - **噪声注入（Noising）**：在模型参数上注入高斯噪声，减少所需噪声量，以消除后门影响。

3. **独立于数据分布和攻击策略的设计**:FLAME 旨在适用于通用的敌手模型，而无需关于后门攻击策略的先验知识，也无需对局部客户端的特定数据分布（如 iid 或 non-iid）作出假设。

4. **基于差分隐私的噪声边界证明**:FLAME 受差分隐私（DP）技术的启发，在噪声注入部分为所需的高斯噪声提供了一个边界证明，从而确保能够有效地消除后门贡献。

5.2 FLAME 解决问题的原理

通过结合过滤、剪裁和噪声注入这三种策略，FLAME 能够有效地防御后门攻击，同时最大程度地保持了全局模型的良性性能和实用性：

- **过滤（Filtering）**：通过识别并移除含有潜在恶意更新的模型，此组件起到首要的防御线，通过对来自不同客户端的模型更新进行分析，FLAME 能够识别出可能包含恶意意图的更新，并防止它们对全局模型造成影响。
- **剪裁（Clipping）**：此步骤通过限制模型参数的更新量来减少恶意更新的潜在影响。裁剪确保了即使恶意更新没有在过滤步骤中被完全排除，其对模型的贡献也将被最小化，这有助于维持全局模型的整体性能和安全性。
- **噪声注入（Noising）**：通过在聚合过程中加入噪声，FLAME 进一步增强了模型抵御后门攻击的能力。这一策略的关键在于它能够扰乱攻击者插入的后门模式，从而防止后门被激活。通过仔细估算必要的噪声水平，FLAME 能够在不显著影响模型性能的前提下，有效中和潜在的后门攻击。

特别是，它的设计不依赖于特定攻击行为的假设，为联邦学习提供了一种更通用且鲁棒的防御策略。通过对真实世界数据集的广泛评估，FLAME 的防御框架证明了其在不同应用领域的有效性。此外，针对声称能规避 FLAME 的攻击，论文作者通过详细的调查和与攻击提出者的沟通，识别并纠正了这些声称的错误，进一步验证了 FLAME 在设计和实施上的可靠性和有效性。

6 实验分析

6.1 实验环境搭建

实验在使用 PyTorch 深度学习框架进行，并引用了 Bagdasaryan 等人 [1]、Xie 等人 [4] 以及 Wang 等人 [3] 提供的源代码来实现攻击。为了验证 FLAME 与现有防御技术的比较效果，我们重新实现了现有的防御策略。所有的实验设置都根据最新的研究成果来评估 FLAME 在三个典型应用场景下的表现，包括词预测、图像分类和物联网入侵检测。实验中除了使用标准的数据集和模型之外，还对参数进行了精心的选择和配置，重复每个实验至少 50 次以确保结果的稳健性和可靠性。此外，为引入变异性，每次实验使用不同的随机种子值。

该环境搭建确保了实验能够在控制条件下进行，从而有效地评估 FLAME 策略在对抗后门攻击（如 3DFed 攻击）中的效果。实验结果显示，在全局模型逐渐达到收敛时，后门准确率显著下降，验证了 FLAME 在处理这类攻击中的有效性。

6.2 实验数据来源

根据文档，实验评估了 FLAME 在三个典型应用场景下的性能：单词预测（Word Prediction, WP）、图像分类（Image Classification, IC）和物联网入侵检测（IoT intrusion detection）。具体的数据集包括：

- **单词预测（WP）**：使用 2017 年 11 月的 Reddit 数据集，每个作为客户端的用户拥有至少 150 篇至多 500 篇的帖子，从而形成了大小介于 298 至 32660 词之间的 80000 个客户数据集。
- **图像分类（IC）**：主要利用 CIFAR-10 数据集，该数据集由 60000 张属于 10 个不同类别的图像组成，是图像分类任务的一个标准基准测试集。
- **IoT 入侵检测**: 使用的数据源于真实世界的家庭和办公环境中采集的数据，包括由 Mirai 恶意软件感染的异常行为 IoT 设备的流量数据，以及 24 种典型 IoT 设备在三种智能家居及一个办公环境设置中的通信数据。

6.3 实验参数设置

实验中考虑了各场景下的标准差分隐私（DP）参数配置：

- 对于图像分类 (IC)，设置 $\epsilon = 3705$，$\lambda = 0.001$。
- 对于物联网入侵检测 (NIDS)，设置 $\epsilon = 395$，$\lambda = 0.01$。
- 对于单词预测 (NLP)，设置 $\epsilon = 4191$，$\lambda = 0.001$。

在 IoT-Traffic 数据集上进行的比较实验中，选取了 100 个客户端参与每轮通信，其中 40 个为恶意客户端。通过向聚合模型添加高斯噪声 ($N(0, \sigma^2)$)，以消除后门威胁。此外，实验通过改变训练回合数来观测四个预训练模型达到不同收敛水平的表现，包括后门准确率 (BA) 和主任务准确率 (MA) 两个关键指标。

BA（后门准确率）指的是模型在后门任务中的准确性，即对于被激活后门的数据集（触发集），模型给出对抗方选择的错误输出的比例。对抗方的目标是最大化 BA，而有效的防御措施则是防止对抗方提高这一指标。

MA（主任务准确率）指的是模型在其主要（良性）任务上的准确性。它表示对于良性输入，系统给出正确预测的比例。对抗方希望尽量减小对 MA 的影响，以减少被检测到的机会。防御系统不应该对 MA 造成负面影响。

6.4 实验对比结果

实验结果基于三种设置来衡量消除所有后门攻击所需的最小高斯噪声规模 (σ)：

- 无任何防御组件（相当于先前的 DP-based 防御）

- 应用动态聚类后
- 应用动态聚类和自适应裁剪（即 FLAME 设置）

实验结果显示，在单独部署聚类和裁剪之后，消除后门所需的噪声规模有所下降，这验证了理论的正确性，图表就不过多展示，接下来主要展示 FLAME 相比 No defense 的情况及其他主流防御手段对比效果。

Attack	Dataset	No Defense		FLAME	
		BA	*MA*	*BA*	*MA*
Constrain-and-scale [9]	Reddit	100	22.6	0	22.3
	CIFAR-10	81.9	89.8	0	91.9
	IoT-Traffic	100.0	100.0	0	99.8
DBA [64]	CIFAR-10	93.8	57.4	3.2	76.2
Edge-Case [62]	CIFAR-10	42.8	84.3	4.0	79.3
PGD [62]	CIFAR-10	56.1	68.8	0.5	65.1
Untargeted Poisoning [22]	CIFAR-10	-	46.72	-	91.31

图 2: FlAME 效果

FLAME 的效果：实验评估了 FLAME 对当前最先进的后门攻击方法——约束缩放（constrain-and-scale）[1]、DBA[4]、PGD 以及边缘情况 (Edge-Case)[3]，还有一种非针对性的投毒攻击 [14] 的防御效果。这些结果显示在图 2。FLAME 完全缓解了所有数据集上的约束缩放攻击 (BA = 0%)。此外，我们的防御不会影响系统的主任务准确率 (MA)，在所有实验中 MA 的减少不超过 0.4%。DBA 攻击以及边缘情况攻击 [3] 也被成功缓解 (BA = 3.2%/4.0%)。此外，FLAME 对 PGD 攻击也有效 (BA = 0.5%)。应该注意的是，提供建议词汇本身就是一个相当具有挑战性的任务，即使没有攻击，主任务准确率 (MA) 也只有 22.7%，这与之前的工作 [1] 是一致的。

Defenses	Reddit		CIFAR-10		IoT-Traffic	
	BA	*MA*	*BA*	*MA*	*BA*	*MA*
Benign Setting	-	22.7	-	92.2	-	100.0
No defense	100.0	22.6	81.9	89.8	100.0	100.0
Krum [11]	100.0	9.6	100.0	56.7	100.0	84.0
FoolsGold [25]	**0.0**	**22.5**	100.0	52.3	100.0	99.2
Auror [56]	100.0	**22.5**	100.0	26.1	100.0	96.6
AFA [46]	100.0	22.4	**0.0**	91.7	100.0	87.4
DP [20]	14.0	18.9	**0.0**	78.9	14.8	82.3
Median [67]	**0.0**	22.0	**0.0**	50.1	0.0	87.7
FLAME	**0.0**	22.3	**0.0**	**91.9**	**0.0**	**99.8**

图 3: FlAME 与其他防御手段的效果对比

FLAME 和现有防御手段对比：我们将 FLAME 与现有防御方法进行比较，包括 Krum[7]、FoolsGold[6]、Auror[15]、自适应联邦平均（Adaptive Federated Averaging, AFA）、中位数及一种通用的差分隐私（Differential Privacy, DP）方法 [1, 16]。图3显示，FLAME 在所有三个数据集上均表现有效，而以往的工作要么未能缓解后门攻击，要么降低了主任务准确度。Krum、FoolsGold、Auror 及 AFA 并未有效去除被污染的模型，后门准确率（BA）经常保持在 100%。此外，某些防御措施使得攻击比没有防御时更成功。因为它们移除了许多良性更新，但没有移除足够多的被污染更新，这些防御措施增加了被污染模型率（PMR），因此也增加了攻击的影响。例如，Krum[7]、Auror[15] 或 AFA[17] 无法处理非独立同分布（non-IID）数据场景，如 Reddit。相比之下，FoolsGold 仅在 Reddit 数据集上有效（真阳性率 TPR = 100%），因为它适应于高度非独立同分

布（non-IID）数据。类似地，AFA 只在 CIFAR-10 数据集上缓解后门，因为数据高度独立同分布（每个客户端被分配一个随机图像集），使得良性模型与全局模型的距离类似。另外，这些模型的主任务准确率（MA）受到了负面影响。基于 DP 的防御措施虽然有效，但会显著降低 MA。例如，它在 CIFAR-10 数据集上的表现最好，BA = 0，但 MA 下降到了 78.9%，而 FLAME 将 MA 提高到了 91.9%，接近于没有攻击时的正常设置（MA = 92.2%）。

从实验中不难发现，FLAME 策略通过引入动态聚类和自适应裁剪机制，显著降低了为消除后门所需加入的噪声量，这不仅有效抵抗了后门攻击，也保证了模型在主要任务上的性能几乎不受影响。此外，通过对不同数量的后门进行实验（0, 1, 2, 4, 8 个后门），展现了 FLAME 在处理单一至多重后门攻击方面的灵活性和鲁棒性。以上细致的实验参数设置和对比结果全面地展示了 FLAME 防御策略在联邦学习环境中对抗后门攻击的有效性，为进一步提高联合学习系统的安全性提供了有力证据。

7 论文阅读心得

阅读这篇关于 FLAME 防御框架的论文，深切体会到在当前快速发展的人工智能领域中，数据安全和模型的鲁棒性是维护系统健康发展的基石。特别是在联邦学习环境中，数据分布式存储和计算带来的隐私保护优势同时也伴随着新的挑战，如何有效防御恶意的后门攻击便是其中之一。FLAME 通过动态估算噪声下界，结合过滤、剪裁和噪声注入三大策略，提出了一种既考虑效率又注重效果的联邦学习防御方法，对提升联邦学习环境的安全性具有积极意义。

从实际应用的角度来看，FLAME 框架的独到之处在于它能够在不预设具体攻击模型和数据分布的情况下，通过一种相对普遍适用的方法进行防御。这种设计哲学意味着 FLAME 在应对未知或动态变化的恶意攻击方面有着较高的灵活性和鲁棒性。实验环节对各种配置的全面探索，不仅验证了 FLAME 的有效性，更为我们提供了对比实验设计和结果分析的宝贵示范。

展望未来，随着联邦学习的应用领域不断拓宽，攻击手段也会越来越多样化和复杂化。如何进一步增强 FLAME 的适应性和灵活性，以应对更复杂多变的攻击策略，将是一个值得深入探讨的问题。此外，如何平衡防御效果和系统效率，保证联邦学习在保护隐私的同时，也能高效运行，也是未来发展的重要方向之一。

总体而言，这篇论文不仅为我们提供了一种新的联邦学习环境下的后门攻击防御框架，更重要的是，它启发我们在未来的研究和实践中，持续关注和解决数据安全与模型鲁棒性问题，以确保人工智能技术的健康、可持续发展。

参考文献

[1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[2] Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, and Ahmad-Reza Sadeghi. Poisoning attacks on federated learning-based iot intrusion detection system. In *Workshop on Decentralized IoT Systems and Security*, 2020.

[3] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *NeurIPS*, 2020.

[4] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *ICLR*, 2020.

[5] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *CCS*, 2018.

[6] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *RAID*, 2020. Originally published as arXiv:1808.04866.

[7] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, 2017.

[8] Sebastien Andreina, Giorgia Azzurra Marson, Helen Möllering, and Ghassan Karame. Baffle: Backdoor detection via feedback-based federated learning. In *ICDCS*, 2021.

[9] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

[10] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. Flame: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, Boston, MA, 2022. USENIX Association.

[11] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. Flame:

Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, Boston, MA, 2022. USENIX Association.

[12] Haoyang Li, Qingqing Ye, Haibo Hu, Jin Li, Leixia Wang, Chengfang Fang, and Jie Shi. 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1893–1907. IEEE, 2023.

[13] Jing Xu, Rui Wang, Stefanos Koffas, Kaitai Liang, and Stjepan Picek. More is better (mostly): On the backdoor attacks in federated graph neural networks. In *Annual Computer Security Applications Conference (ACSAC)*. ACSAC, 2022.

[14] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Security Symposium*. USENIX Association, 2020.

[15] Shiqi Shen, Shruti Tople, and Prateek Saxena. Auror: Defending against poisoning attacks in collaborative deep learning systems. In *Annual Computer Security Applications Conference (ACSAC)*, 2016.

[16] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private language models without losing accuracy. In *International Conference on Learning Representations (ICLR)*, 2018.

[17] Luis Muñoz-González, Kenneth T. Co, and Emil C. Lupu. Byzantine-robust federated machine learning through adaptive model averaging. *arXiv preprint arXiv:1909.05125*, 2019.