Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

面向联邦学习的后门攻击检测与防御

博士研究生　李佳龙

2026 年 07 月 05 日

问题回溯

- 总结反思
  - PPT模板使用不够规范
  - 基础知识部分讲解不够细致
  - 部分实验结果图较模糊
- 相关内容
  - 2025.12.21 李佳龙《面向联邦基础模型的安全评测与防御方法研究》
  - 2025.10.27 陈星星《面向数据异构与通信高效的联邦大模型优化与应用研究》
  - 2026.05.31 满乐彤《基于大模型微调的后门攻击》
  - 2025.09.14 赵怡清《扩散模型的后门攻击研究》
  - 2025.03.23 赵怡清《文本生成大模型后门攻击研究》

内容提要

- 预期收获
- 内涵解析与研究目标
- 研究背景与研究意义
- 研究历史与现状
- 知识基础
- 算法原理
  - FedID
  - GBHINDER
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 熟悉联邦学习与联邦图学习场景下后门攻击的基本问题
  - 了解联邦后门攻击防御的发展历史
  - 理解两种面向联邦后门防御的新思路
  - 掌握GBHINDER由被动异常检测转向主动模型净化的防御思路

内涵解析与研究目标

- 研究目的
  - 面向联邦学习与联邦图学习，研究分布式隐私保护训练场景下的后门攻击防御问题
  - 针对联邦学习中恶意梯度与良性梯度高度相似、数据分布非独立同分布、攻击设置未知等挑战，实现对隐蔽恶意更新的动态识别
  - 针对联邦图学习中图结构触发器传播、原始图数据不可访问、中心服务器不可信等挑战，实现基于本地可信历史知识的主动模型净化
- 内涵解析
  - 联邦图学习：在保护本地图数据隐私的前提下，多个客户端利用各自的节点特征、边关系和拓扑结构协同训练图神经网络，用于处理具有复杂关联关系的图数据
  - 主动模型净化：客户端利用自身可信历史模型作为良性锚点，对下载的全局模型进行表征约束与拓扑一致性约束，实现由“被动异常检测”向“主动后门净化”的防御转变

研究背景与研究意义

- 研究背景
  - 联邦学习场景的分布式特性导致后门攻击风险更大，恶意梯度与良性梯度可能高度相似，且数据分布呈现高度Non-IID特征，现有联邦学习后门防御方法（差分隐私）难以同时兼顾防御效果与主任务性能
  - 联邦图学习场景中，图数据包含节点特征与拓扑关系，良性客户端更新天然具有较强异质性，容易模糊恶意偏离与良性差异之间的界限。图结构触发器可通过节点、边或局部子图模式灵活嵌入，使污染模型识别难度进一步提高
  - 现有联邦图学习防御方法多依赖可信服务器，既增加服务器端计算开销，也与联邦学习保护数据隐私的目标存在冲突
- 研究意义
  - 提升联邦学习中对隐蔽恶意梯度的识别能力，降低后门攻击成功率的同时保持主任务性能
  - 探索联邦图学习中无可信服务器的客户端主动防御范式，推动防御思路由“被动异常检测”转向“主动模型净化”

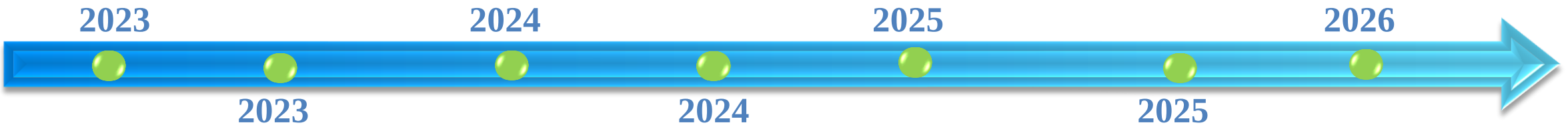
研究历史和现状

Zhang等人提出一种面向安全聚合场景的联邦后门检测框架SAFELearning，通过随机不可见分组和部分参数披露，在不泄露单个客户端模型隐私的前提下，对子组聚合模型进行异常检测，缓解安全聚合机制下后门攻击更难发现的问题。

Kasyap等人提出一种本地模型投毒攻击框架Sine，指出仅依赖欧氏距离或余弦相似度不足以识别复杂恶意更新；该方法通过构造相似外观的恶意模型更新持续降低全局模型性能，并提出FLTC防御方法，利用可信坐标的方向和幅度变化限制攻击影响。

Yang等人提出增强型模型投毒攻击ScaleSign与多策略防御方法MS Guard，通过梯度缩放和符号统计修改，使恶意梯度在余弦相似度和符号分布上更接近良性梯度，从而绕过多种鲁棒聚合规则；同时结合余弦机制、符号统计和谱方法提升防御能力。

Zhu等人提出一种面向联邦图学习的无可信服务器后门净化框架GBHINDER，通过历史知识锚点、通道注意力正则化和自适应动量信息更新，使客户端能够自主净化下载的全局模型，推动联邦图后门防御由服务器端被动检测转向客户端主动净化。



2023

Hu等人提出一种面向恶意客户端攻击的鲁棒联邦学习框架RFFL，通过在聚合前迭代过滤恶意客户端，提高联邦模型对不同类型、不同强度攻击的防御能力，并进一步扩展到数据异构场景，增强IoT环境下联邦学习的鲁棒性。

2024

Miao等人提出一种面向后门攻击的高效安全联邦学习方案ESFL，通过自适应本地差分隐私与压缩感知机制，在保护模型更新隐私的同时降低通信开销，并提升对后门攻击的抵抗能力，探索隐私保护、通信效率与安全防御之间的平衡。

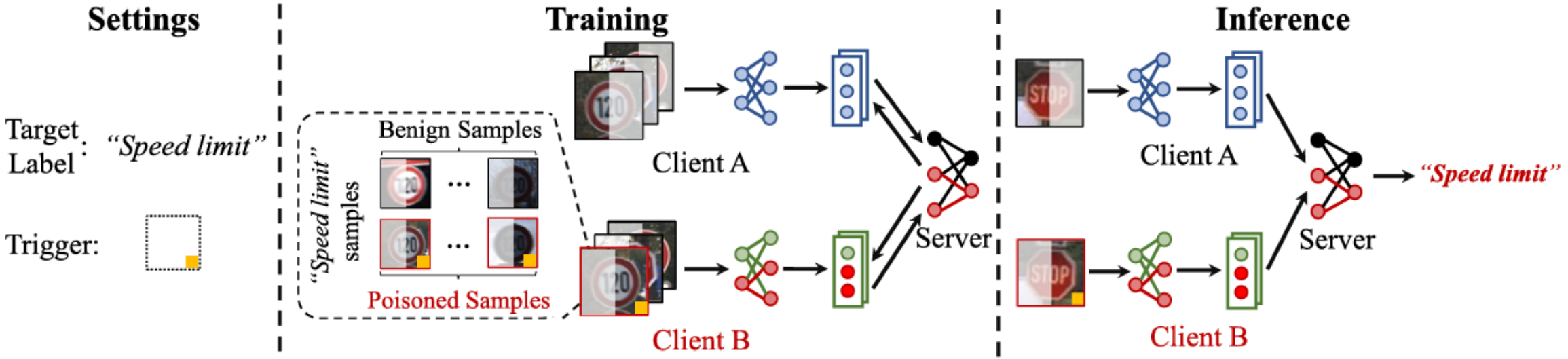
2025

Huang等人提出一种联邦学习后门动态识别方法FedID，通过曼哈顿距离、欧氏距离和余弦相似度的多度量动态加权，以及改进z-score的良性梯度选择机制，解决高维空间中欧氏距离失效和单一度量泛化能力不足的问题。

知识基础

对比维度	联邦学习（Federated Learning, FL）	联邦图学习（Federated Graph Learning, FedGL）
定义	多个客户端在不共享原始数据的前提下，协同训练一个共享全局模型	融合联邦学习与图神经网络，在不共享本地图数据的前提下协同训练图模型
数据	图像、文本、表格、时序等常规本地样本	图数据$G = (V, E, X)$，同时包含节点特征与拓扑结构
模型	CNN、RNN、Transformer、MLP等通用深度学习模型	GCN、GAT、GIN、GraphSAGE图神经网络模型
后门攻击特征	攻击者通过污染本地数据或模型更新，使全局模型在触发器输入下输出目标标签	触发器可通过节点、边、局部子图或自然拓扑模式嵌入，隐蔽性更强，也更易沿消息传递过程扩散
防御难点	恶意梯度与良性梯度可能高度相似，Non-IID数据会放大异常检测误差	图结构复杂且良性更新天然异质，容易模糊良性异质性与恶意偏离之间的边界
现有防御瓶颈	基于距离相似度的防御依赖服务器端异常检测，差分隐私方法会损害主任务性能	FedGL专用防御仍处于早期阶段，FedTGE等方法多依赖可信服务器，存在计算开销与隐私保护冲突
联系	FedGL不是简单地将FL应用于图数据，而是在图拓扑结构、Non-IID异质性、触发器隐蔽性和服务器可信性共同作用下形成的更复杂联邦安全场景	

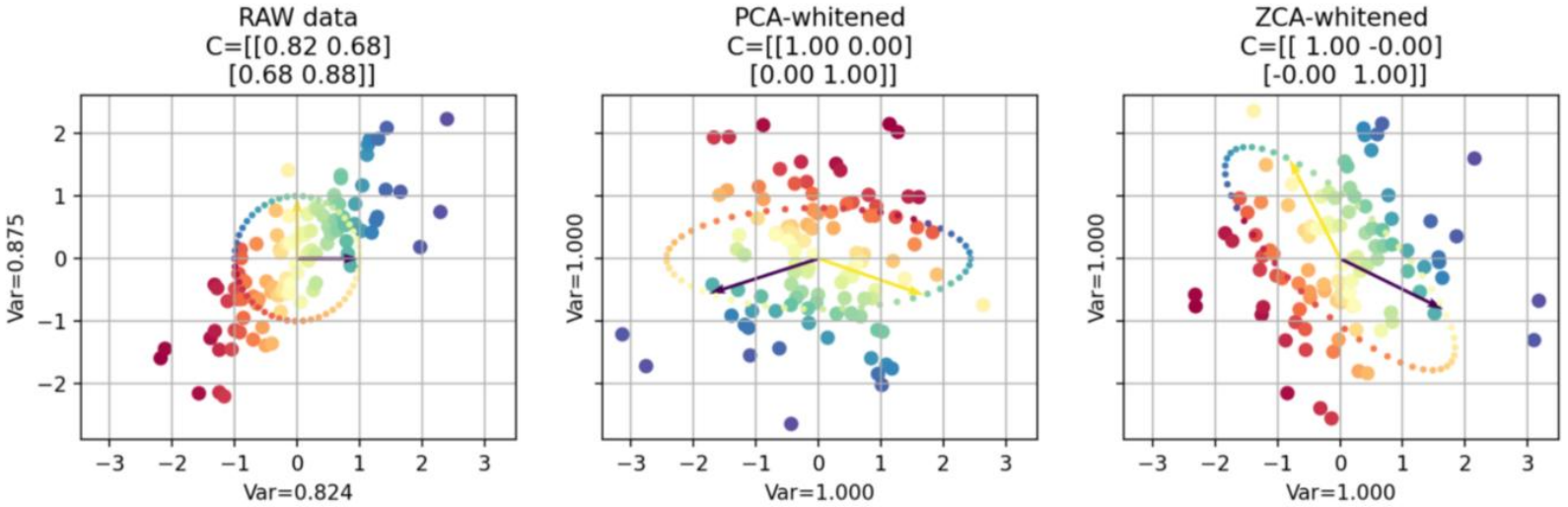
知识基础



- 基于距离（相似度）的异常检测
  - 基本思想：假设良性梯度与恶意梯度在向量空间中存在可度量差异
  - 优势：只聚合被判定为良性的客户端更新，通常对主任务性能影响较小
  - 局限：在Non-IID数据和隐蔽后门攻击下容易误判，存在额外计算开销
- 基于差分隐私的方法
  - 基本思想：通过梯度裁剪和噪声注入削弱恶意更新对全局模型的影响
  - 优势：不依赖特定数据分布假设，对隐蔽后门攻击具有较强缓解能力
  - 局限：额外噪声会降低模型主任务性能，并减缓联邦训练收敛速度

知识基础

- 数据白化处理（Whitening）
  - 一种重要的数据预处理方法，用于降低输入特征之间的冗余性
  - 标准归一化主要调整数据的均值和尺度，而白化进一步通过旋转与缩放消除特征之间的相关结构
  - 基本流程
    - 零中心化：$X_c = X - \mu$
    - 数据去相关：$\Sigma = \frac{1}{n} {X_c}^T X_c = U \Lambda U^T$
    - 方差缩放：$X_{PCA-white} = X_c U (\Lambda + \varepsilon I)^{-\frac{1}{2}}$
  - PCA白化先旋转到主成分空间，再按特征值进行方差归一，白化后数据坐标方向发生变化，更强调去相关与主成分表达
  - ZCA白化在PCA白化基础上，再旋转回原始特征空间，更强调去相关后仍接近原始数据分布

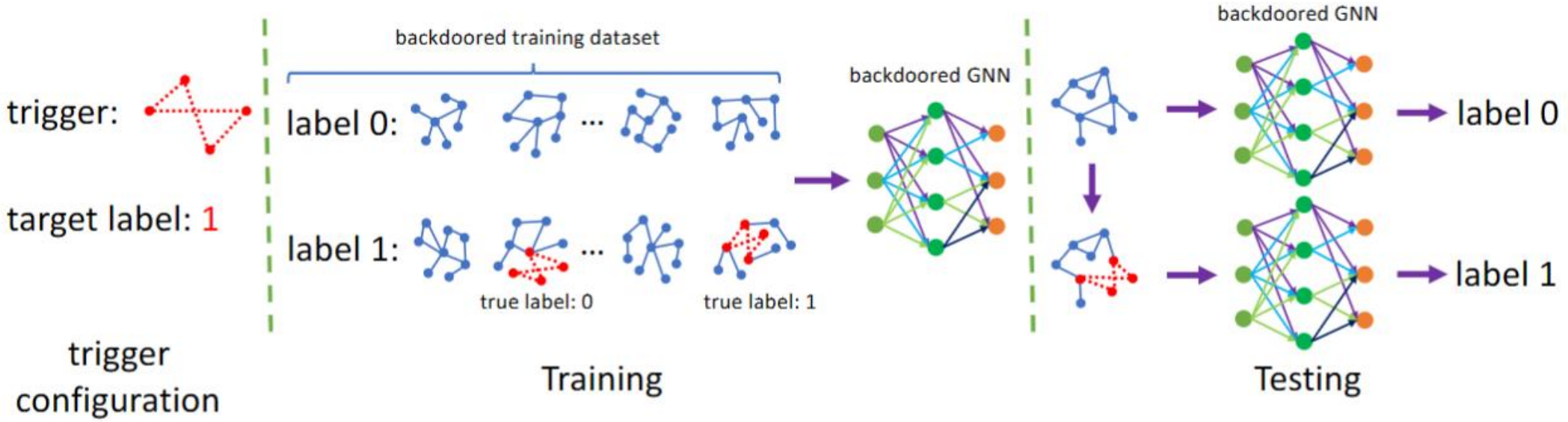


知识基础

- 图神经网络与图后门触发器
  - 图数据表示为$G = (V, E, X)$，与图像、文本不同，图数据同时包含特征信息和拓扑关系
  - GNN通过消息传递机制聚合邻居节点信息，逐层学习节点或图级表征：

$$h_v^{(l+1)} = \sigma(W^{(l)} \cdot AGG\{h_u^{(l)} : u \in \mathcal{N}(v)\})$$

  - 浅层更关注局部结构，深层逐渐融合多跳邻居信息
  - 触发器不再局限于像素块，可表现为节点、边、局部子图或自然拓扑模式
  - 由于图结构复杂，局部异常可能在消息传递过程中被逐层放大，隐蔽性更强



算法原理 FedID

Fedid: Enhancing federated learning security through dynamic identification

算法原理 TIPO

T	目标	在联邦学习场景中，防御针对全局模型的隐蔽后门攻击
I	输入	K个客户端的本地数据集与本地模型更新 各客户端在t轮训练后上传的本地模型梯度序列 $\{\Delta w_1^t, \Delta w_2^t, \dots, \Delta w_K^t\}$
P	处理	1.梯度特征计算：基于曼哈顿距离、欧氏距离、余弦相似度计算客户端更新差异 2.动态加权：基于白化处理消除量纲差异与指标相关性，自适应调整不同度量权重 3.良性聚合：基于改进z-score筛选良性梯度，并对良性更新进行FedAvg聚合
O	输出	1个能够有效防御隐蔽后门攻击并保持主任务性能的全局模型w_g^{t+1}

P	问题	1.联邦学习在保护数据隐私的同时，分布式特性导致更易遭受隐蔽后门攻击威胁 2.差分隐私方法虽能有效缓解后门攻击，但额外添加噪声会降低模型主任务性能并减缓收敛速度
C	条件	攻击者无法控制良性客户端访问其本地数据或篡改服务器聚合流程，恶性<50%
D	难点	FedID需要在Non-IID数据分布条件下区分良性异质更新与真实恶意更新，同时避免牺牲主任务性能
L	水平	IEEE Transactions on Pattern Analysis and Machine Intelligence 2025（中科院一区TOP，CCF A）

算法原理 问题定义

- 联邦聚合过程

$$w_t = w_{t-1} + \eta \cdot \frac{\sum_{i=1}^{K} n^{(i)} \Delta w_t^{(i)}}{\sum_{i=1}^{K} n^{(i)}}$$

  - 服务器根据各客户端上传的本地更新进行加权聚合
- 攻击假设
  - 采用白盒攻击设置：攻击者了解聚合算法和配置参数，可调整恶意客户端本地训练过程；但恶意客户端数量不超过总客户端数的50%，且无法控制良性客户端或篡改聚合流程
- 防御目标
  - 尽可能识别并剔除恶意更新
  - 保持全局模型的良性性能
  - 适应不同数据分布和攻击策略

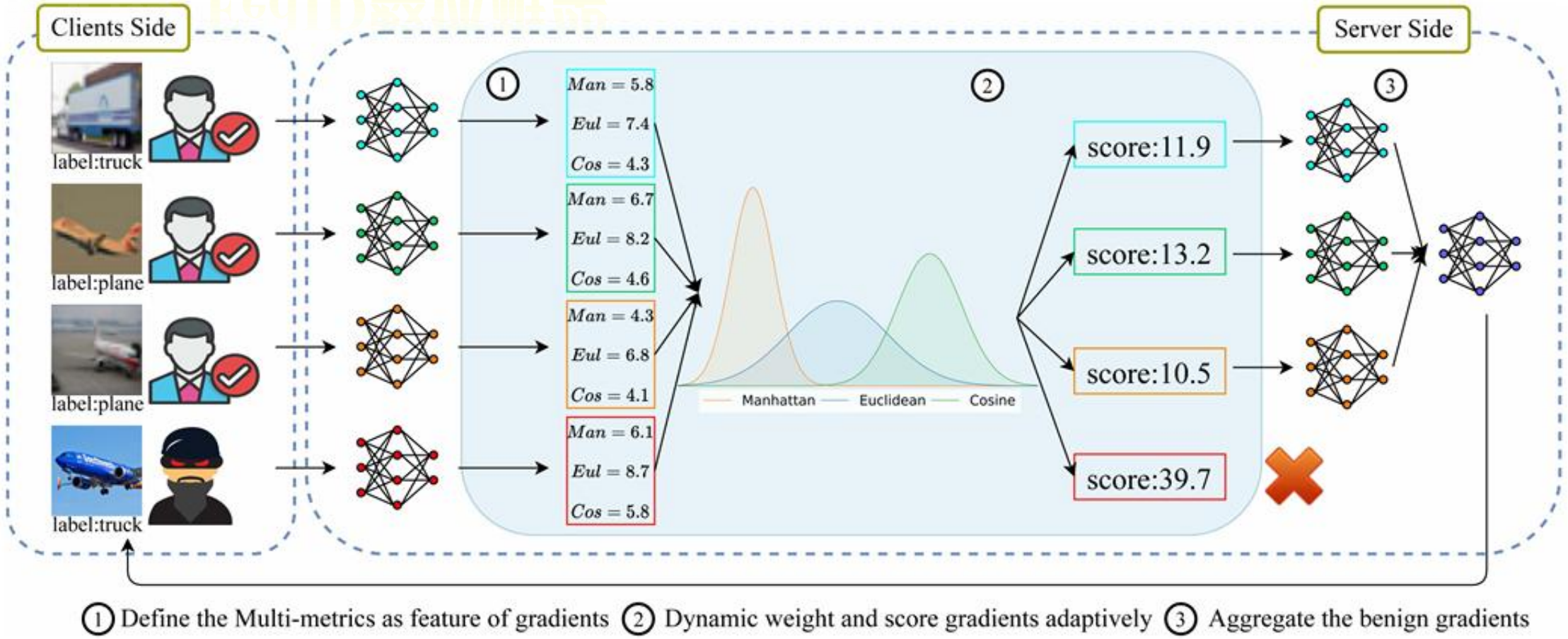
Input: Total number of clients in each round K, models of clients $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(K)}$, last global model $\boldsymbol{w}_0$, global learning rate η, size of the $i-th$ client dataset $n^{(i)}$

Output: Global model w'

算法原理 FedID

- 创新点来源：基于距离防御仍是性能友好方向，尤其是对FL来说
  - 差分隐私方法能够缓解隐蔽后门攻击，但额外噪声会降低主任务性能并减缓收敛
  - 基于距离的方法只聚合良性梯度，关键在于有效区分恶意更新与良性更新
  - 现有距离防御存在两点不足：欧氏距离在高维空间区分能力下降，单一度量难以识别具有不同特征的恶意梯度
  - 在Non-IID数据分布与攻击设置未知条件下，良性梯度本身也存在差异，进一步增加了恶意梯度识别难度
- 对应创新点：多度量动态识别与良性梯度聚合
  - 梯度特征计算：引入曼哈顿距离，并结合欧氏距离、余弦相似度（累计差异、长度差异、方向差异）
  - 基于白化处理的梯度特征动态加权：自适应调整不同度量权重
  - 基于改进z-score的良性梯度聚合：动态筛选良性梯度，避免依赖预设攻击比例

算法原理 FedID整体框架



FedID三阶段防御流程：先提取多度量梯度特征，再通过动态加权计算异常得分。最终聚合良性梯度，从而在抑制后门攻击的同时保持主任务性能

算法原理 梯度特征计算

- 命题：曼哈顿距离更适合高维空间
  - 单一欧式距离或余弦相似度难以覆盖所有攻击特征
  - 欧式距离在高维神经网络参数空间中易受维度灾难影响
    - $\frac{D^k_{max,d}-D^k_{min,d}}{D^k_{min,d}} \xrightarrow{p} 0$　　$\lim_{d\to\infty} E\left[\frac{D^k_{max,d}-D^k_{min,d}}{d^{(1/k)-(1/2)}}\right] = C_k$
    - 随着维度d的增长，最远和最近距离的相对差异趋近于0
    - 在不同k范数下，最远距离和最近距离的绝对差异增长速度不一样
    - $D^k_{max,d} - D^k_{min,d}$的增长速率与$d^{(1/k)-(1/2)}$成正比
  - 高维空间中的距离区分能力
    - 曼哈顿距离：$M_d = D^1_{max,d} - D^1_{min,d}$
    - 欧式距离：$U_d = D^2_{max,d} - D^2_{min,d}$
    - $\lim_{d\to\infty} E\left[\frac{M_d}{U_d \cdot d^{1/2}}\right] = C'$，$M_d \approx C' \cdot U_d \cdot \sqrt{d}$
    - 曼哈顿距离能更准确地捕获高维空间中的梯度细微差异

for $i \in \{1, 2, \ldots, K\}$ **do** ▷ compute the gradients features

$x^{(i)}_{Man} \leftarrow \|\boldsymbol{w}_0 - \boldsymbol{w}_i\|_1$

$x^{(i)}_{Eul} \leftarrow \|\boldsymbol{w}_0 - \boldsymbol{w}_i\|_2$

$x^{(i)}_{Cos} \leftarrow \frac{\langle \boldsymbol{w}_0, \boldsymbol{w}_i \rangle}{\|\boldsymbol{w}_0\| \cdot \|\boldsymbol{w}_i\|}$

$\boldsymbol{x}^{(i)} \leftarrow (x^{(i)}_{Man}, x^{(i)}_{Eul}, x^{(i)}_{Cos})$

end for

for $i, j \in \{1, 2, \ldots, K\}$ **do** ▷ compute the sum of the distance between each gradient

$x'^{(i)}_{Man} \leftarrow \sum_{j,j\neq i}^{K} |x^{(i)}_{Man} - x^{(j)}_{Man}|$

$x'^{(i)}_{Eul} \leftarrow \sum_{j,j\neq i}^{K} |x^{(i)}_{Eul} - x^{(j)}_{Eul}|$

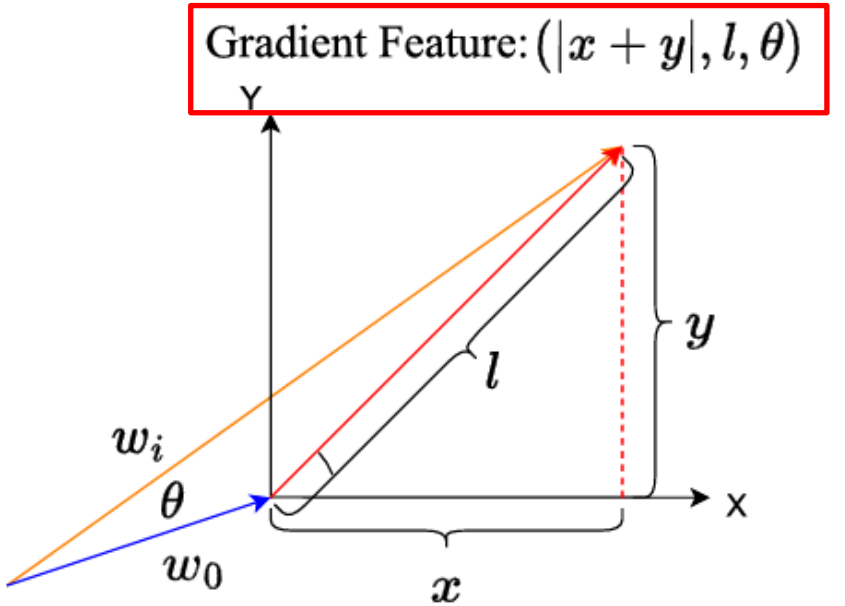
$x'^{(i)}_{Cos} \leftarrow \sum_{j,j\neq i}^{K} |x^{(i)}_{Cos} - x^{(j)}_{Cos}|$

$\boldsymbol{x}'^{(i)} \leftarrow (\boldsymbol{x}'^{(i)}_{Man}, \boldsymbol{x}'^{(i)}_{Eul}, \boldsymbol{x}'^{(i)}_{Cos})$

end for

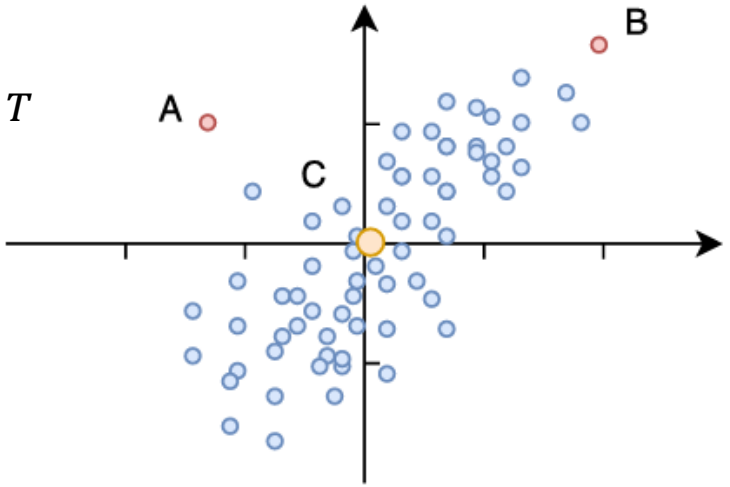
算法原理 梯度特征计算

- 多距离协同刻画梯度特征
  - 联邦学习场景中，攻击者可能在不同环境和数据分布下实施攻击，从而生成具有不同特征的恶意梯度
  - 梯度特征定义
    - $x = (x_{Man}, x_{Eul}, x_{Cosine})$
    - $x_{Man}^{(i)} = ||w_i - w_0||_1, x_{Eul}^{(i)} = ||w_i - w_0||_2, x_{Cosine}^{(i)} = \frac{w_0^T w_i}{||w_0|| ||w_i||}$
    - 曼哈顿距离刻画高维参数变化的累计，欧式距离刻画客户端更新的长度差异，余弦相似度刻画方向差异
  - 恶意梯度识别指标
    - 计算每个梯度与其他梯度间距离之和
    - $x'^{(i)} = (\sum_{j,j\neq i}^{K} \left| x_{Man}^{(i)} - x_{Man}^{(j)} \right|, \sum_{j,j\neq i}^{K} \left| x_{Eul}^{(i)} - x_{Eul}^{(j)} \right|, \sum_{j,j\neq i}^{K} \left| x_{Cosine}^{(i)} - x_{Cosine}^{(j)} \right|)$
    - 通过同时考虑曼哈顿距离、欧氏距离和余弦相似度，更全面地刻画客户端更新差异，缓解单一指标在复杂后门攻击场景下识别能力不足的问题



算法原理 基于白化处理的梯度特征动态加权

- 核心问题
  - $x_{Man}^{(i)}$，$x_{Eul}^{(i)}$，$x_{Cosine}^{(i)}$量纲和数值范围不同，不能直接简单相加
  - 不同特征之间存在相关性，简单最大值归一化难以消除冗余影响
  - 在Non-IID场景下，良性客户端更新本身存在差异，固定权重容易造成误判
- 数据白化处理
  - 构造当前轮客户端梯度特征矩阵：$X = [x'^{(1)}, x'^{(2)}, \ldots, x'^{(K)}]^T$
  - 计算特征协方差矩阵：$\Sigma = Cov(X)$
  - 利用其逆矩阵进行白化投影：$\delta^{(i)} = \sqrt{x'^{(i)T}\Sigma^{-1}x'^{(i)}}$



- 动态加权
  - Σ^{-1}由当前轮次梯度特征分布计算得到，权重会随训练轮次和攻击环境自适应变化
  - 输出梯度异常得分序列$\{\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(K)}\}$

算法原理 基于改进z-score的良性梯度聚合

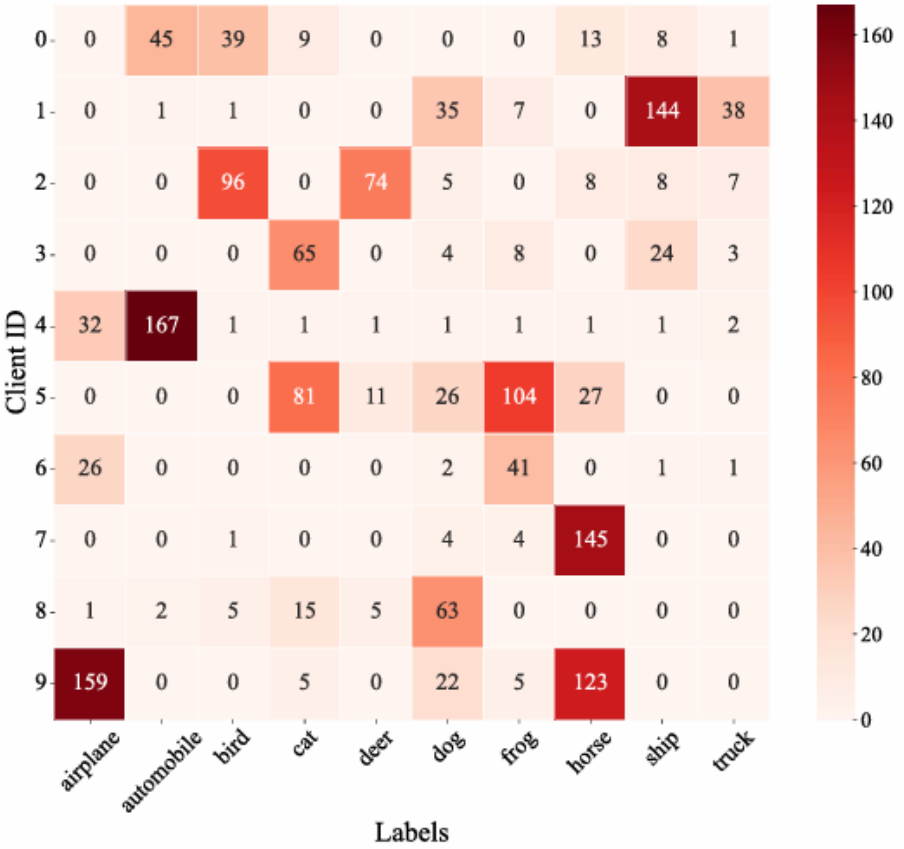
- 核心问题
  - 得到每个梯度的异常得分$\delta^{(i)}$后，选择得分较低的良性梯度进行聚合
  - 防御方并不知道每轮更新中恶意客户端所占比例
  - 聚合比例过大，可能选入后门梯度；过小，全局模型泛化能力较差
- 改进z-score动态筛选
  - 传统z-score：$\gamma = \frac{x-\mu}{\sigma}$，恶意更新存在时均值和标准差易受极端异常值影响
  - FedID采用中位数绝对偏差（MAD）代替标准差：

$$MAD(\{\delta^{(i)}\}_{i=1}^{K}) = median(|\delta^{(i)} - median(\{\delta^{(i)}\}_{i=1}^{K})|)$$

  - 同时，用$\{\delta^{(1)}, \delta^{(2)}, \dots, \delta^{(K)}\}$中的最小值代替均值：$\gamma^{(i)} = \frac{\delta^{(i)} - min_i \delta^{(i)}}{MAD(\{\delta^{(i)}\}_{i=1}^{K})}$
  - FedAvg聚合：$w_t = w_{t-1} + \eta \cdot \frac{\sum_{i \in S_t} n^{(i)} \Delta w_t^{(i)}}{\sum_{i \in S_t} n^{(i)}}, S_t = \{i | \gamma^{(i)} < 1\}, \Delta w_t^{(i)} = w_t^{(i)} - w_0$

实验流程 实验设置

- 视觉分类数据集
  - CIFAR-10：50,000张训练图像，10,000张测试图像，10个类别
  - EMNIST：280,000张真实手写数字图像，类别为0–9
  - Fashion-MNIST：10个类别，每类7,000张时尚商品图像
- 泛化验证数据集
  - CINIC10：270,000张图像，规模约为CIFAR-10的4.5倍
  - LOAN：2,260,668条贷款与信用评分样本，9个类别
  - Sentiment140：160万条英文Twitter情感分类数据
- Non-IID数据划分
  - 采用 Dirichlet分布$Dir(\alpha)$模拟联邦学习中的数据异构场景
  - α越小，客户端数据分布的 Non-IID程度越高
  - LOAN：$\alpha = 0.9$，其余数据集均为0.5



- 模型设置：CIFAR-10：VGG-9; EMNIST：LeNet-5; Sentiment140：LSTM; LOAN：三层全连接网络

实验流程 实验设置

- 攻击基线
  - 同时设置像素级后门与语义级后门
  - 为提升防御难度，采用多轮攻击而非单轮攻击
  - Model Replacement：通过放大攻击梯度替换全局模型
  - DBA：将触发器拆分后由多个恶意客户端协同上传
  - PGD：通过缩放与投影约束恶意梯度
  - Edge-case PGD：利用长尾样本与投影约束增强隐蔽性
- 防御基线
  - FedAvg、Krum、Multi-Krum、RFA、FoolsGold、Weak-DP、FLPruning、Flame
- 评估指标
  - MA(Main Accuracy)
  - BA(Backdoor Accuracy)
  - 在保持高 MA 的同时降低 BA

	Backdoor type	Semantic trigger					Artificial trigger			
$\mathcal{D}$	dataset	CIFAR10	CINIC10	EMNIST	Fashion-MNIST	Sentiment140	CIFAR10	EMNIST	Fashion-MNIST	LOAN
$\mathcal{D}_{\mathcal{A}}$	backdoor data	Southwest Airlines		Ardis		Greek Director	-			
N	#clients	200				200	300	100		
K	#clients selected in each round	10								
-	#attackers in each round	1					4			
-	#attackers local iteration	2				5	6	10		5
-	#benign local iteration	2					1			
T	#global iteration	1500			500	300	70			
-	batch size	32				20	64			
-	attack interval	10					1			
α	non-IID parameter	0.5				1	0.5			0.9
η	benign learning rate	0.02				0.05	0.1			0.001
$\eta_{\mathcal{A}}$	attackers learning rate	0.02				0.05				0.0005

实验结果 整体防御性能

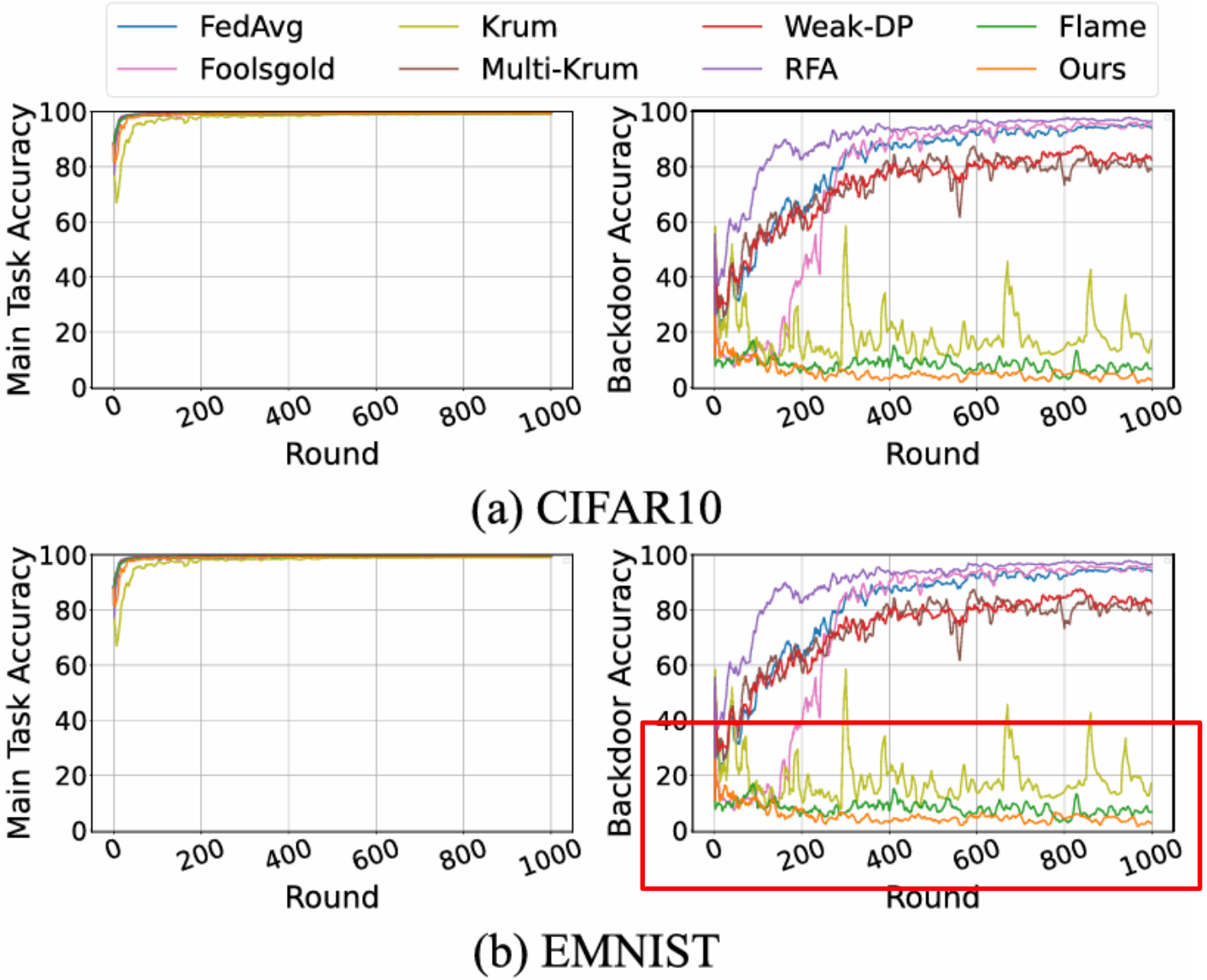
Dataset	Defense	Model Replacement		DBA		PGD		Edge-case PGD		Rank Score
		MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓	
CIFAR10	FedAvg	**86.95**	64.80	79.23	90.44	87.04	14.44	87.14	55.10	+0.00
	RFA	86.69(+0.00)	25.56(-0.61)	**79.6(+0.00)**	57.69(-0.36)	**87.10(+0.00)**	52.56(+2.64)	86.47(-0.01)	65.31(+0.19)	-1.86
	Foolsgold	85.71(-0.01)	6.67(-0.90)	77.56(-0.02)	**3.43(-0.96)**	84.92(-0.02)	14.44(+0.00)	84.76(-0.03)	51.53(-0.06)	+1.84
	Krum	82.17(-0.05)	6.11(-0.91)	78.18(-0.01)	6.01(-0.93)	82.32(-0.05)	66.67(+3.62)	81.23(-0.07)	59.18(+0.07)	-2.04
	Multi-Krum	86.55(+0.00)	1.67(-0.97)	79.33(+0.00)	91.39(+0.01)	86.52(-0.01)	17.78(+0.23)	**87.4(+0.00)**	60.20(+0.09)	+0.63
	Weak-DP	74.41(-0.14)	46.11(-0.29)	N/A	N/A	74.43(-0.14)	22.22(+0.54)	73.84(-0.15)	53.06(-0.04)	-0.66
	FLpruning	85.72(-0.01)	6.67(-0.90)	78.23(-0.01)	8.57(-0.91)	85.77(-0.01)	9.44(-0.35)	85.74(-0.02)	48.47(-0.12)	+2.23
	Flame	80.58(-0.07)	**0.56(-0.99)**	76.78(-0.03)	37.24(-0.59)	81.24(-0.07)	**0.56(-0.96)**	81.41(-0.07)	5.12(-0.91)	+3.21
	FedID	86.66(+0.00)	**0.56(-0.99)**	79.46(+0.00)	7.24(-0.92)	86.47(-0.01)	**0.56(-0.96)**	86.58(-0.01)	**2.04(-0.96)**	**+3.82**
EMNIST	FedAvg	99.54	96.00	97.68	94.13	99.55	10.00	99.37	96.00	+0.00
	RFA	99.57(+0.00)	6.00(-0.94)	**97.87(+0.00)**	1.39(-0.99)	99.32(+0.00)	4.00(-0.60)	99.29(+0.00)	97.00(+0.01)	+2.51
	Foolsgold	96.42(-0.03)	98.00(+0.02)	97.24(+0.00)	0.64(-0.99)	99.07(+0.00)	94.00(+8.40)	99.13(+0.00)	98.00(+0.02)	-7.49
	Krum	99.22(+0.00)	**0.00(-1.00)**	97.7(+0.00)	**0.56(-0.99)**	99.12(+0.00)	1.00(-0.90)	99.14(+0.00)	12.00(-0.88)	+3.76
	Multi-Krum	**99.58(+0.00)**	**0.00(-1.00)**	97.85(+0.00)	47.43(-0.50)	99.54(+0.00)	**0.00(-1.00)**	**99.57(+0.00)**	84.00(-0.13)	+2.63
	Weak-DP	99.37(+0.00)	86.00(-0.10)	N/A	N/A	99.41(+0.00)	14.00(+0.40)	99.39(+0.00)	89.00(-0.07)	-0.23
	FLpruning	99.41(+0.00)	24.00(-0.75)	97.24(+0.00)	4.79(-0.95)	99.48(+0.00)	14.00(+0.40)	99.46(+0.00)	92.00(-0.04)	+2.14
	Flame	99.39(+0.00)	**0.00(-1.00)**	97.12(-0.01)	17.38(-0.82)	99.39(+0.00)	**0.00(-1.00)**	99.44(+0.00)	13.00(-0.86)	+3.67
	FedID	99.49(+0.00)	**0.00(-1.00)**	97.22(+0.00)	3.76(-0.96)	**99.57(+0.00)**	**0.00(-1.00)**	99.46(+0.00)	**0.00(-1.00)**	**+3.96**
Fashion-MNIST	FedAvg	**91.50**	99.00	**87.59**	99.69	**91.60**	99.00	**91.95**	99.00	+0.00
	RFA	89.76(-0.02)	99.00(+0.00)	87.41(-0.06)	44.21(-0.56)	91.20(+0.00)	99.00(+0.00)	91.04(-0.01)	99.00(+0.00)	+0.52
	Foolsgold	87.62(-0.04)	98.00(-0.01)	82.58(-0.05)	92.78(-0.07)	87.11(-0.05)	35.00(-0.65)	90.74(-0.01)	99.00(-1.00)	+0.56
	Krum	87.21(-0.05)	**0.00(-1.00)**	82.87(-0.05)	8.57(-0.91)	86.27(-0.06)	**0.00(-1.00)**	85.05(-0.08)	**0.00(-1.00)**	+3.68
	Multi-Krum	90.92(-0.01)	**0.00(-1.00)**	87.19(+0.00)	3.08(-0.97)	90.98(-0.01)	1.00(-0.99)	90.08(-0.02)	97.00(-0.02)	+2.94
	Weak-DP	90.51(-0.01)	**0.00(-1.00)**	87.11(-0.01)	99.81(+0.00)	89.92(-0.02)	17.00(-0.83)	89.34(-0.03)	99.00(+0.00)	+1.89
	FLpruning	N/A	N/A	85.63(-0.03)	9.81(-0.90)	88.38(-0.04)	15.00(-0.85)	88.68(-0.04)	87.00(-0.12)	+1.66
	Flame	89.12(-0.03)	**0.00(-1.00)**	86.26(-0.02)	3.26(-0.97)	88.24(-0.04)	**0.00(-1.00)**	87.16(-0.05)	**0.00(-1.00)**	+3.84
	FedID	90.32(-0.01)	**0.00(-1.00)**	86.33(-0.01)	**2.81(-0.97)**	90.52(-0.01)	**0.00(-1.00)**	89.80(-0.02)	4.00(-0.96)	**+3.87**

结果分析

- FedID在三类数据集上均取得最高Rank Score
- 在Model Replacement、PGD、Edge-case PGD等攻击下保持较低BA
- 与差分隐私方法相比，FedID避免额外噪声对MA的明显影响

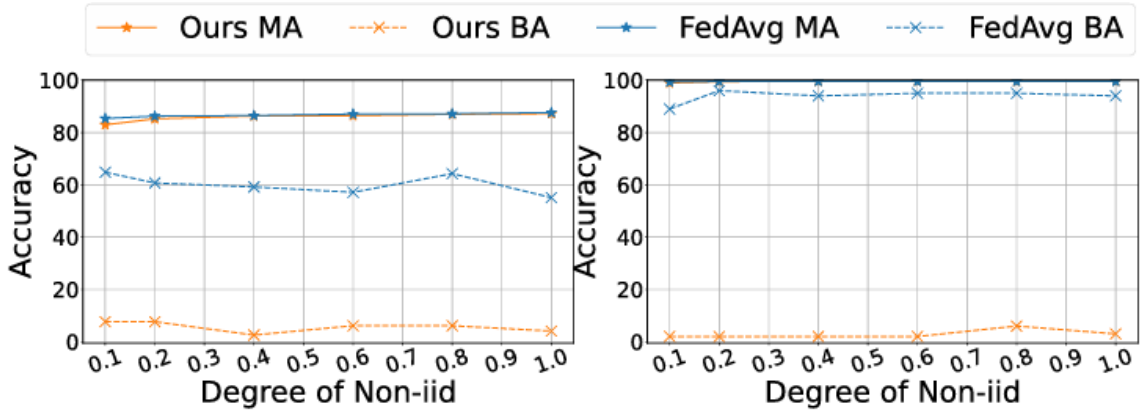
结论：FedID能够在降低后门攻击成功率的同时，基本保持全局模型主任务性能

实验结果 Edge-case PGD场景下的防御性能

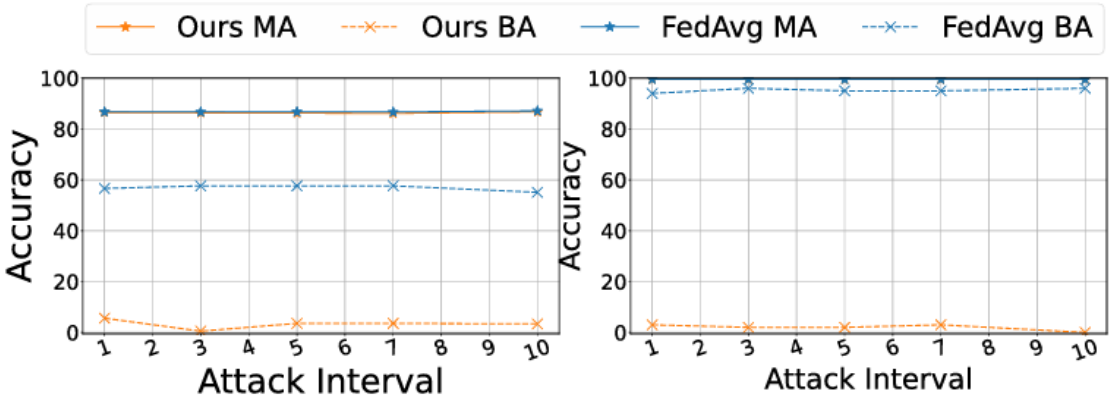


- 结果分析
  - FedID与Flame能抵御Edge-case PGD，但Flame会牺牲MA
  - Krum、FoolsGold等方法在更隐蔽攻击下容易失效
- 结论
  - FedID对隐蔽性更强的Edge-case PGD场景有更稳定的防御能力

实验结果 Non-IID与攻击频率影响

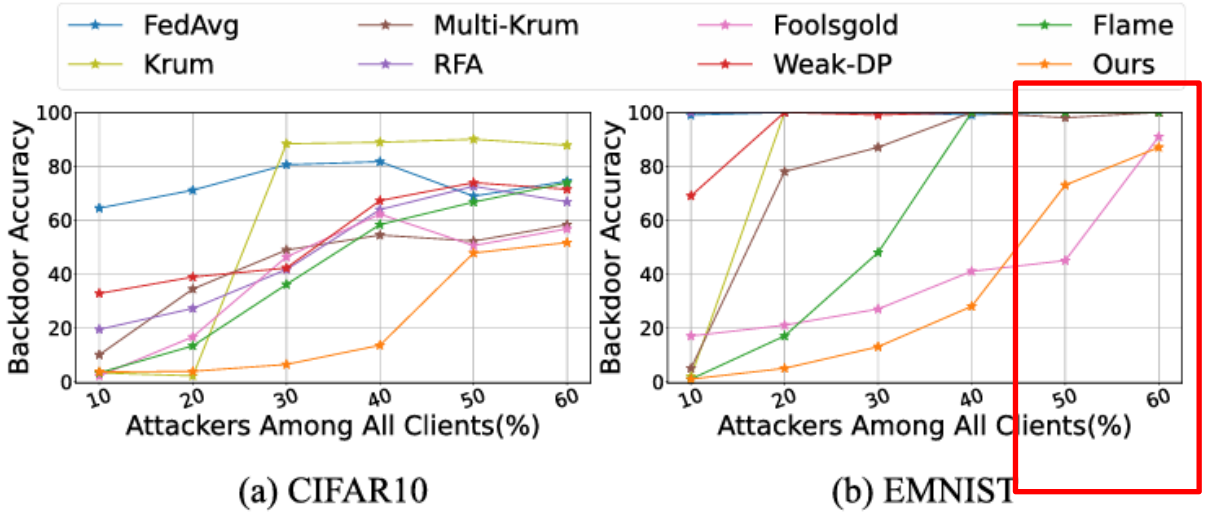


(a) CIFAR10 (b) EMNIST



(a) CIFAR10 (b) EMNIST

- 结果分析
  - 不同α下BA均维持低水平，MA轻微波动
  - 攻击频率变化时，FedID仍保持稳定低BA
  - 当恶意客户端占比高于50%时，各方法防御性能均明显下降



(a) CIFAR10 (b) EMNIST

结论：FedID对Non-IID程度、攻击频率和恶意客户端占比变化具有较强适应性

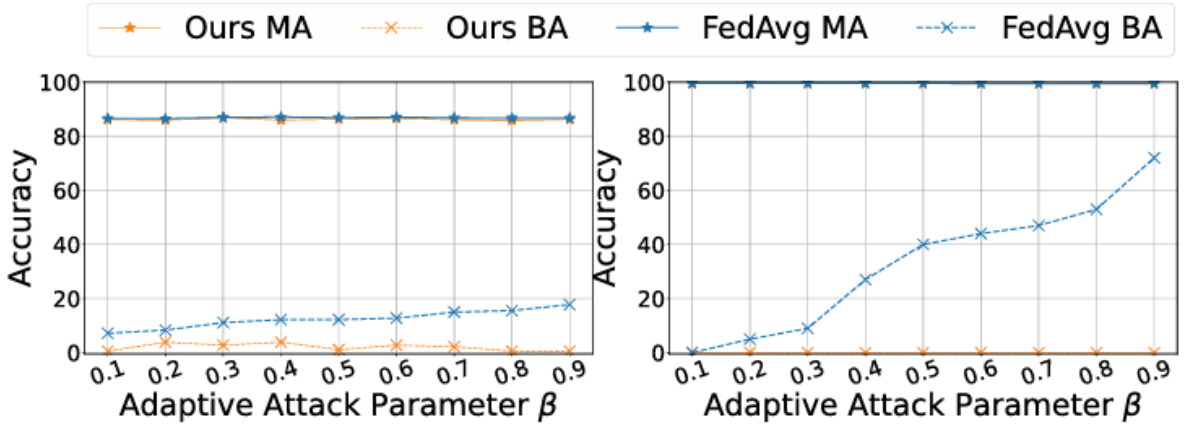
实验结果 自适应攻击防御性能

- 根据凯克霍夫原则，假设攻击者全面了解FedID防御机制
- 攻击者通过异常检测损失调整恶意梯度特征

$$\mathcal{L}_{model} = \beta\mathcal{L}_{class} + (1-\beta)\mathcal{L}_{ano}$$

$$\mathcal{L}_{ano} = \mathcal{L}_{Man} + \mathcal{L}_{Eul} + \mathcal{L}_{Cos}$$

- β平衡攻击有效性与规避检测能力

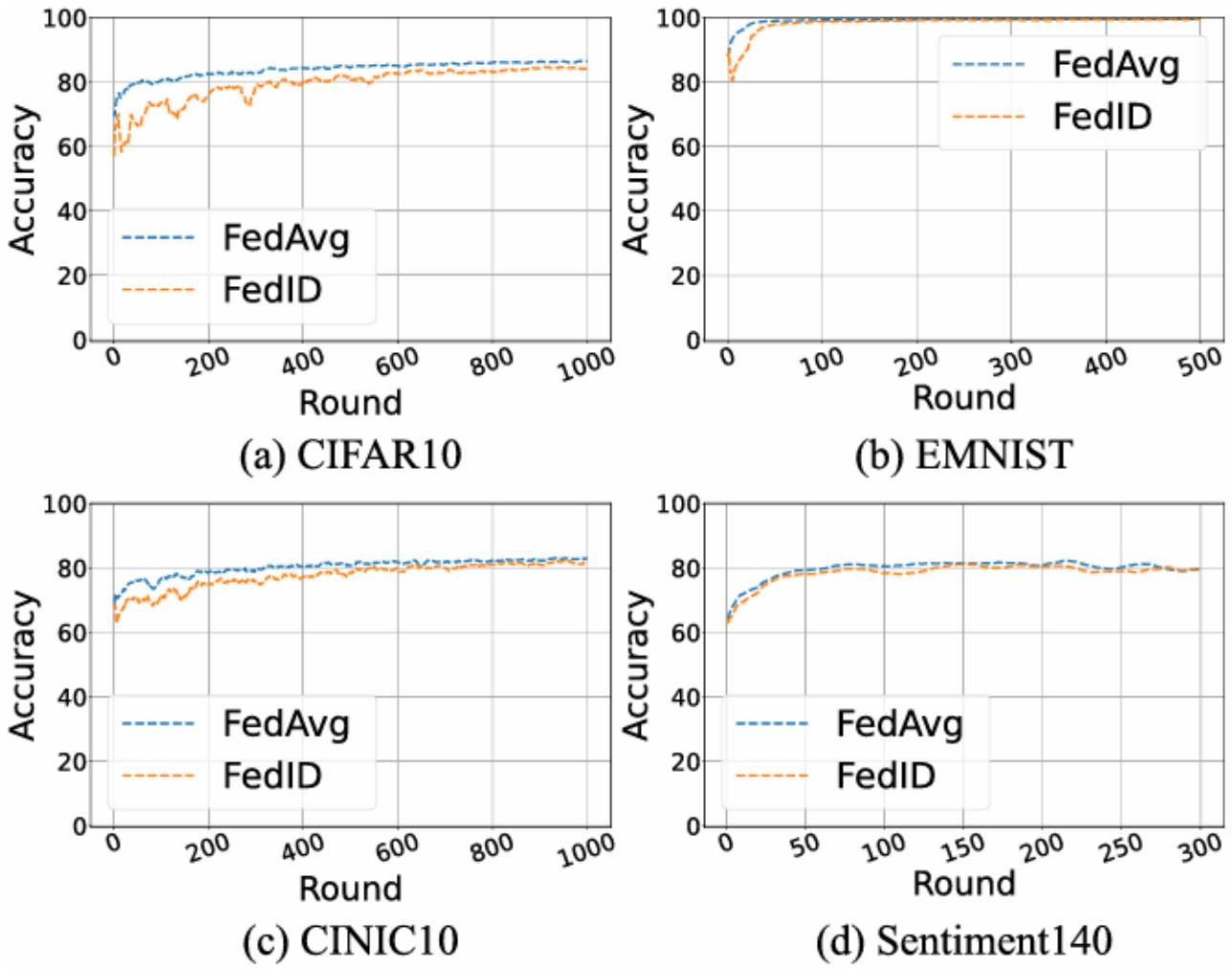


(a) CIFAR10 (b) EMNIST

Datasets	Defenses	LITTLE IS ENOUGH		RLBackdoor		Adaptive	
		MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓
CIFAR10	FedAvg	**86.91**	59.18	**85.04**	98.90	**86.89**	43.37
	RFA	86.85	33.67	84.98	99.10	86.78	51.02
	Foolsgold	85.09	75.51	84.02	4.50	85.06	26.02
	Krum	81.53	4.08	83.37	5.60	81.96	56.12
	Multi-Krum	86.75	7.14	84.64	5.50	86.53	43.37
	Weak-DP	72.58	47.96	74.64	98.50	73.64	44.39
	Flame	83.04	7.14	79.29	8.50	81.41	51.53
	FedID	86.16	**2.55**	84.78	**3.70**	86.40	**9.69**
EMNIST	FedAvg	**99.54**	80.00	**95.38**	97.75	99.51	92.00
	RFA	99.50	72.00	95.31	97.45	99.54	95.00
	Foolsgold	99.44	91.00	94.43	1.90	99.53	67.00
	Krum	98.64	11.00	94.23	1.80	99.11	12.00
	Multi-Krum	99.53	**4.00**	95.26	1.50	**99.57**	45.00
	Weak-DP	97.56	68.00	92.20	95.65	99.00	89.00
	Flame	99.42	42.00	93.47	1.45	99.43	30.00
	FedID	99.45	**4.00**	95.06	**1.10**	**99.57**	**6.00**
Fashion-MNIST	FedAvg	**90.83**	99.00	**85.28**	99.20	**92.11**	99.00
	RFA	90.24	99.00	84.81	98.80	91.80	97.00
	Foolsgold	90.32	99.00	83.87	**0.00**	91.34	48.00
	Krum	87.08	**0.00**	82.38	**0.00**	87.13	**0.00**
	Multi-Krum	90.75	**0.00**	85.11	**0.00**	90.20	98.00
	Weak-DP	88.77	98.00	81.40	97.70	90.32	99.00
	Flame	89.13	**0.00**	83.09	**0.00**	89.83	95.00
	FedID	89.88	**0.00**	84.95	**0.00**	89.49	8.00

- 当β超过0.3时，FedAvg的BA快速上升；FedID仍保持BA接近0，EMNIST更明显
- β=0.5时FedID 仍保持稳定防御效果，说明其不依赖单一攻击假设

实验结果 不同数据集的泛化性能



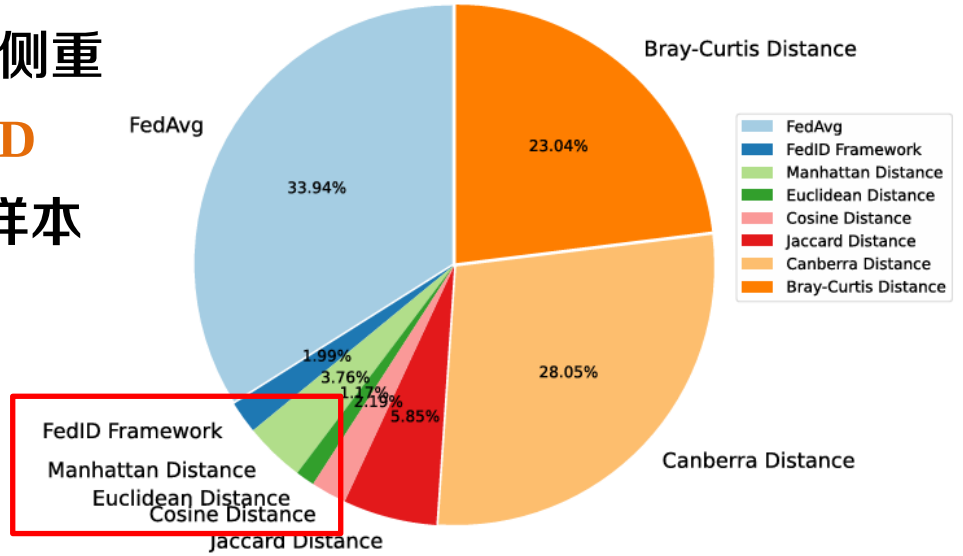
Defense	CINIC10		LOAN		Sentiment140	
	MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓
FedAvg	**84.37**	36.22	**89.05**	61.36	**82.59**	89.17
FedID	83.02	**4.59**	88.52	**0.00**	81.67	**5.83**

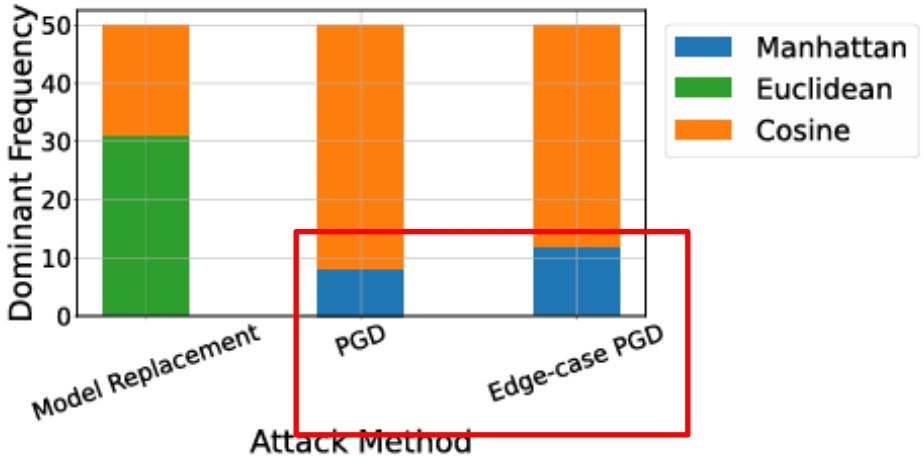
- 结果分析
  - 在特征较为简单的数据集EMNIST和Sentiment140上，FedID在整个训练过程中的收敛速度与FedAvg基本相当
  - FedID在CINIC10、LOAN和Sentiment140上均实现较高MA与低BA，说明该方法对不同任务具有一定适用性

实验结果 消融实验与计算开销

- 消融实验
  - 多指标组合优于单一指标，不同度量在识别恶意梯度时各有侧重
  - 曼哈顿距离更有助于防御隐蔽性较强的PGD与Edge-case PGD
  - 白化处理相比最大值归一化能够更好识别Non-IID下的恶意样本

Weighting	Model Replacement		PGD		Edge-case PGD	
	MA ↑	BA ↓	MA↑	BA ↓	MA ↑	BA ↓
Max Norm	83.86	**0.56**	83.74	25.56	84.08	62.24
Whitening	**86.34**	**0.56**	**86.66**	**0.56**	**86.58**	**2.04**





Defense	Model Replacement	PGD	Edge-case PGD
	MA/BA	MA/BA	MA/BA
Man	83.86/**0.56**	83.74/25.56	85.3/64.80
Eul	86.24/**0.56**	85.52/17.78	**87.12**/54.08
Cos	84.22/2.22	83.84/30.00	85.38/66.84
Man+Eul	84.09/1.11	84.17/28.63	84.3/67.35
Man+Cosine	85.74/1.67	85.16/23.68	85.86/6.63
Cosine+Eul	86.31/**0.56**	85.44/16.11	85.14/63.78
Man+Cosine+Eul	**86.66/0.56**	**86.47/0.56**	86.58/**2.04**

- 计算开销
  - 基础FedID框架仅占总计算成本的9.11%

Fend for Yourself! Backdoor Purification in Federated Graph Learning with an Evolving Knowledge Anchor

算法原理 TIPO

T	目标	在无可信服务器的联邦图学习场景中，防御针对全局图模型的后门攻击
I	输入	M个客户端的本地图数据集$D_k = \{(G_{k,i}, y_{k,i})\}$ 服务器下发的全局模型w^t以及良性客户端自身历史模型锚点h_k^t
P	处理	1.利用客户端可信历史模型作为良性知识锚点 2.历史通道注意力正则化约束全局模型表征，结合拓扑一致性损失抑制后门传播 3.基于自适应动量信息更新机制选择性融合稳健全局知识，动态更新历史锚点
O	输出	1个无需可信服务器、能够抵御图后门攻击并保持主任务性能的全局图模型w^{t+1}

P	问题	1. 联邦图学习中图结构复杂且数据呈现更高程度的Non-IID，任务难度更高 2.现有联邦图学习防御依赖可信中央服务器，带来额外计算开销且与隐私保护目标存在冲突
C	条件	攻击假设：恶意客户端可完全操控本地训练过程；每个良性客户端均作为防御者，在无可信服务器、无公开干净数据集条件下自主防御
D	难点	GBHINDER需要在本地信息有限条件下区分良性全局知识与恶意模式，同时避免静态本地锚点导致模型分化和知识孤立
L	水平	USENIX Security 2026（CCF A）

算法原理 问题定义与威胁模型

- 联邦图学习
  - 核心任务为图分类，即为每个图G分配一个来自标签空间Y的标签y

$$w = arg\min_{w} \sum_{k=1}^{|\mathcal{M}|} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \ell_k(w, \mathcal{D}_k)$$

$$\ell_k(w, \mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} f(w, (G_{k,i}, y_{k,i}))$$

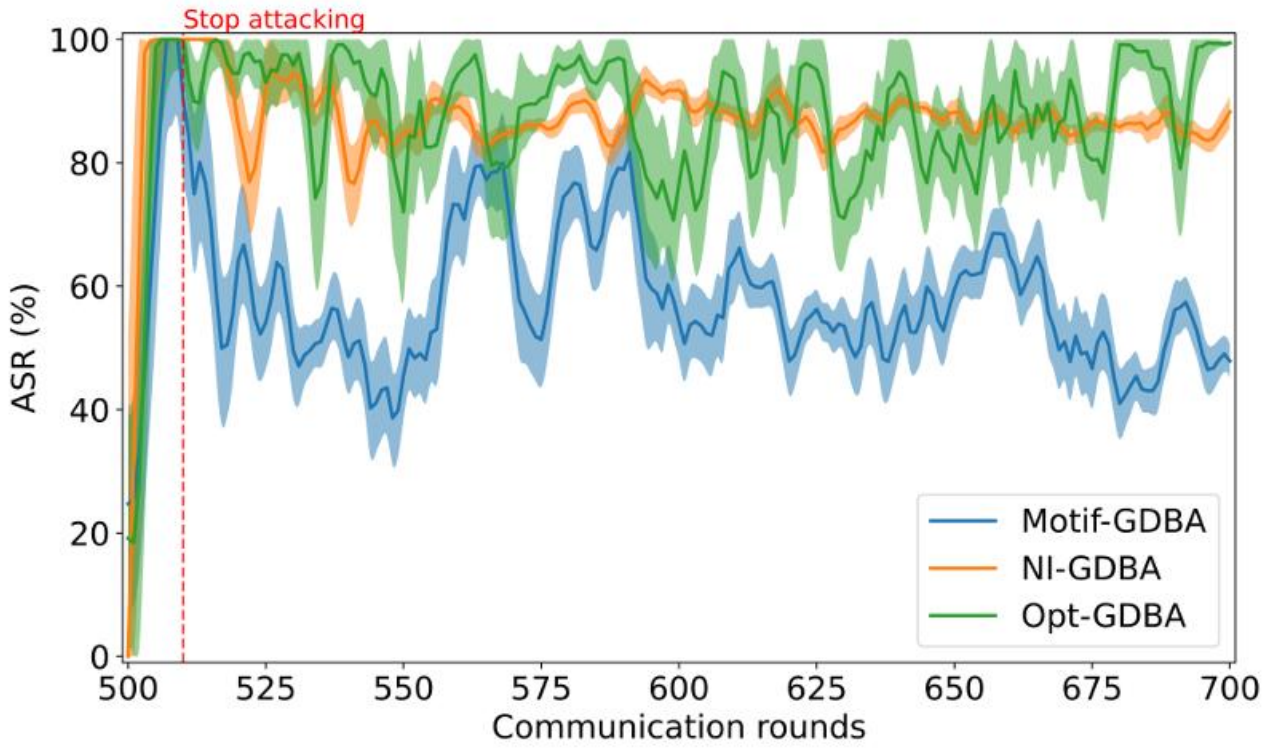
  - 第t轮训练中，服务器选择客户端集合S_t，并下发全局模型w^t
  - 客户端利用本地数据训练：$w_k^{t+1} \leftarrow w_k^t - \eta \nabla_{w_k^t} \ell(w_k^t; b)$
  - 服务器通过聚合函数更新全局模型：$w^{t+1} \leftarrow AGG(\{w_k^{t+1} | k \in S_t\})$

- 后门攻击目标
  - 攻击者试图向全局图模型注入神经木马，干净输入上保持正常预测，触发器出现时输出目标标签

$$\begin{cases} F_w(G) = \widetilde{F_w}(G) \\ \widetilde{F_w}(G \oplus \delta) = \hat{y} \end{cases}$$

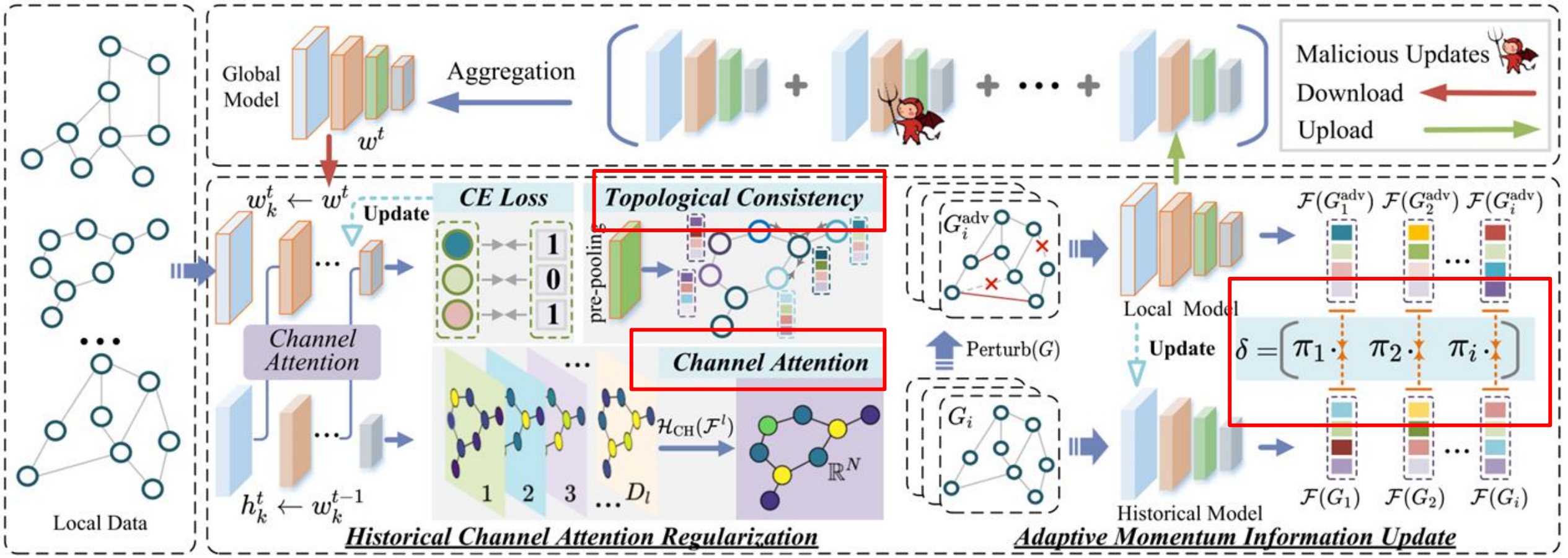
算法原理 创新点

- 创新点来源：单纯本地微调难以防御深层后门
  - 经过10轮后门攻击后，即使继续进行200轮无恶意参与者训练，单纯微调仍难以彻底消除后门
  - 后门隐蔽性
  - 孤立客户端本地数据不足



- 对应创新点：从被动微调（服务器）转向主动模型净化（客户端）
  - 提出GBHINDER：一种无可信服务器的FedGL防御范式，利用客户端可信历史知识缓解后门攻击
  - 历史通道注意力正则化：结合图拓扑特性的通道注意力机制，放大良性与恶意神经元激活差异；引入拓扑一致性损失，使局部净化后的全局模型深层表征向历史模型中的良性浅层特征靠拢
  - 自适应动量更新：根据扰动敏感性动态调整锚点更新强度，选择性吸收稳健全局知识

算法原理 GBHINDER整体框架



在无需可信服务器的条件下净化全局图模型并抑制后门传播

算法原理 历史通道注意力正则化

- 基于历史锚点的显式正则
  - 良性客户端利用上一轮本地模型作为可信历史锚点：$h_k^t \leftarrow w_k^{t-1}$
  - 良性模型与后门模型的整体表征可能高度相似，简单表征匹配难以有效净化后门
  - 如何放大良性与恶意神经元激活值差异？结合图拓扑信息与通道重要性
- 拓扑通道注意力构建
  - 对GNN第l层激活张量：$\mathcal{F}^l \in \mathbb{R}^{N \times D_l}$，定义通道注意力函数$\mathcal{H}: \mathbb{R}^{N \times D_l} \rightarrow \mathbb{R}^N$
  - 引入归一化节点度作为拓扑权重，突出结构重要节点：$D[i] = \frac{\deg(v_i)}{max_{v \in V} \deg(v)}$
  - 第d个通道的注意力图：$[\mathcal{H}_d(\mathcal{F}^l)]_i = D[i] \cdot |\mathcal{F}^{l(i,d)}|^2$
  - 考虑到后门神经元可能集中于特定的通道，引入通道权重：$\omega_d = \frac{||\mathcal{F}^{l(:,d)}||_1}{\sum_{j=1}^{D_l} ||\mathcal{F}^{l(:,d)}||_1}$
  - $||F^{l(:,d)}||_1 = \sum_{i=1}^{N} |\mathcal{F}^{l(i,d)}|$，表示第$d$个通道激活值的$L_1$范数
  - 最终通道注意力：$\mathcal{H}_{CH}(\mathcal{F}^l) = \sum_{d=1}^{D_l} \omega_d \cdot \mathcal{H}_d(\mathcal{F}^l) \in \mathbb{R}^N$

算法原理 历史通道注意力正则化

- 深浅层注意力对齐
  - GNN浅层捕获局部拓扑模式，深层融合全局信息
  - 后门特征通常通过多跳消息传播在深层逐步累积
  - 因此利用历史锚点的浅层注意力约束当前模型的深层注意力，避免后门传播

$$\mathcal{L}_{align}(h_k^t, w_k^t)$$

$$= \sum_{l=1}^{L-1} ||\psi\left(\mathcal{H}_{CH}\left(\mathcal{F}_{h_k^t}^{l}\right)\right) - \psi\left(\mathcal{H}_{CH}\left(\mathcal{F}_{w_k^t}^{l+1}\right)\right)||_2^2$$

  - 其中，$\psi\left(\mathcal{H}_{CH}(\mathcal{F}^l)\right) = \frac{\mathcal{H}_{CH}(\mathcal{F}^l)}{||\mathcal{H}_{CH}(\mathcal{F}^l)||_2}$

- 拓扑一致性约束
  - 后门攻击可能通过异常边或节点特征破坏图结构一致性
  - 利用本地干净数据约束相邻节点特征对齐：

$$\mathcal{L}_{topo}(w_k^t) =$$

$$\frac{1}{\sum_{(i,j)\in E} A_{i,j}} \sum_{(i,j)\in E} ||z_{w_k^t}^{(i)} - z_{w_k^t}^{(j)}||_2^2 \; A_{i,j}$$

- 整体优化目标
  - 对下载的全局模型进行主动净化

$$\mathcal{L}_{total} = \frac{1}{|b|}(\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{topo})$$

算法原理 自适应动量信息更新

- 静态历史锚点的局限
  - 历史通道注意力正则化依赖历史模型质量
  - FedGL随机选择客户端，良性客户端可能多轮未被选中，导致历史锚点偏差或过时
  - 静态锚点可能导致知识孤立，削弱全局模型泛化能力
- 基于动量的历史锚点更新
  - 在每个训练epoch中动态更新历史模型：

$$h_k^t \leftarrow a \cdot h_k^t + (1-a) \cdot w_k^t$$

  - a为动量系数，控制历史知识与净化后全局知识的融合比例
  - a越大，越倾向于保留历史锚点知识
  - 不直接固定a，而是根据净化模型的稳定性与可信度动态调整

算法原理 自适应动量信息更新

- 基于扰动敏感性的可信度评估
  - 构造拓扑扰动图，模拟后门触发器对图结构的影响：$G^{adv} = Perturb(G)$
  - 扰动方式：随机添加或删除一定比例的边
- 扰动敏感性得分
  - 计算净化后模型与历史模型在扰动图上的表征差异：

$$\delta = \sum_{G_i \in D_k} \pi_i \cdot ||\psi\left(\mathcal{F}_{w_k^t}(G_i^{adv})\right) - \psi\left(\mathcal{F}_{h_k^t}(G_i)\right)||_2^2$$

  - δ越大，说明净化模型对拓扑扰动越敏感，可能仍包含后门污染

- 图规模权重
  - 考虑不同图规模对整体表征贡献不同，为每个图分配权重：

$$\pi_i = \frac{|V_i|}{\sum_{G_i \in D_k} |V_j|}$$

- 动量系数自适应映射

$$a = a_0 \cdot (1 - \exp(-\gamma \cdot \delta))$$

  - 高δ：提高a，更多保留历史锚点以减少潜在恶意知识吸收
  - 低δ：降低a，更多融合全局知识，避免锚点过时

实验流程 实验设置

- 数据集
  - 图分类任务数据集
  - 节点分类任务数据集
- 基础模型
  - 默认骨干模型：图同构网络 (Graph Isomorphism Network, GIN)
  - 泛化验证模型：GCN、GAT、GraphSAGE
- 攻击基线
  - Rand-GDBA：随机图触发器后门攻击
  - Motif-GDBA：基于结构模体的图后门攻击
  - Opt-GDBA：优化触发器结构与注入位置
  - NI-GDBA：非侵入式图后门攻击，不直接修改原始图数据
  - 节点分类任务额外引入GTA-GDBA、UGBA-GDBA

Datasets	NCI1	PROTEINS	DD	AIDS
#Graphs	4,110	1,113	1,178	2,000
Avg. #Nodes	29.87	39.06	284.32	15.69
Avg. #Edges	32.3	715.66	72.82	32.3
#Classes	2	2	2	2
#Target label	1	1	0	1

Datasets	CiteSeer	Pubmed	Coauthor-CS	Amazon-Photo
#Nodes	3,327	19,717	18,333	7,650
#Edges	4,732	44,338	81,894	119,081
#Feature	3,703	500	6,805	745
#Classes	6	3	15	8
#Target label	1	1	1	0

实验流程 实验设置

- 防御基线
  - FoolsGold、Flame、RLR：通用联邦学习防御方法
  - GNNCert、FedTGE：联邦图学习专用防御方法
  - 节点分类任务中，GNNCert因面向图级预测而不参与对比
- GBHINDER超参数
  - 历史通道注意力正则：$\lambda_1 = 1$
  - 拓扑一致性约束：$\lambda_2 = 1$
  - 动量系数上界：$a_0 = 0.9$
  - 扰动敏感性缩放因子：$\gamma = 10$
  - 默认采样50%边计算$\mathcal{L}_{topo}$，采样50%本地数据计算δ
- 评估指标
  - ACC（主任务准确率）：衡量全局模型在干净样本上的分类性能
  - ASR（攻击成功率）：后门样本被误分类为目标标签的比例
  - 在保持高ACC的同时降低ASR

实验结果 整体防御性能

Datasets (ACC *w/o* attack)	Attack	No-Defense		FoolsGold		Flame		RLR		GNNCert		FedTGE		GBHINDER	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
NCI1 (ACC=80.42)	Rand-GDBA	75.72	28.66	78.39	12.37	77.26	13.13	78.17	9.54	70.58	2.03	75.97	39.58	80.56	6.73
	Motif-GDBA	77.36	54.65	78.33	53.81	78.98	44.99	79.06	50.63	71.61	22.74	76.32	37.13	79.35	5.26
	NI-GDBA	80.42	86.94	76.66	88.53	76.66	85.43	77.12	84.27	69.92	57.92	75.09	80.94	80.09	8.64
	Opt-GDBA	76.66	87.11	72.84	86.68	69.08	58.15	71.61	50.74	68.00	73.99	75.97	35.23	79.88	6.51
PROTEINS (ACC=75.95)	Rand-GDBA	66.47	29.89	65.41	26.54	67.11	24.09	70.32	33.78	61.75	5.95	73.28	30.26	76.53	7.54
	Motif-GDBA	67.30	63.78	69.76	22.02	63.39	57.48	67.58	57.48	59.77	27.60	72.27	42.07	74.01	6.72
	NI-GDBA	75.95	88.11	71.09	78.71	74.68	71.61	72.76	80.02	66.95	61.49	73.66	72.91	75.16	9.75
	Opt-GDBA	72.49	92.83	70.74	47.10	69.29	95.25	68.77	90.61	62.46	35.47	72.52	34.7	73.13	6.12
DD (ACC=71.46)	Rand-GDBA	68.82	30.19	69.49	30.68	70.09	28.35	67.86	40.63	62.49	8.47	67.34	10.31	70.13	5.24
	Motif-GDBA	69.86	40.63	68.67	47.92	68.87	40.24	65.87	40.24	62.25	29.08	64.26	21.7	69.72	6.31
	NI-GDBA	71.46	84.86	68.52	78.43	68.52	77.84	70.67	80.32	53.03	73.86	64.96	81.22	70.95	8.39
	Opt-GDBA	71.34	78.53	70.64	75.63	70.81	77.64	62.25	29.08	60.43	61.98	65.53	72.64	70.17	4.32
AIDS (ACC=99.74)	Rand-GDBA	99.56	24.56	99.76	11.56	99.76	4.16	97.12	18.45	95.24	14.05	91.96	11.98	99.72	8.34
	Motif-GDBA	99.88	38.45	98.07	9.05	99.76	7.84	98.45	35.84	92.26	8.90	93.97	5.79	99.17	6.05
	NI-GDBA	99.74	87.33	97.47	88.81	97.50	87.59	98.07	84.05	94.62	84.93	93.41	86.63	99.71	8.78
	Opt-GDBA	99.11	84.62	98.05	50.34	98.21	43.24	92.26	86.90	95.06	37.20	91.85	17.03	99.52	2.35

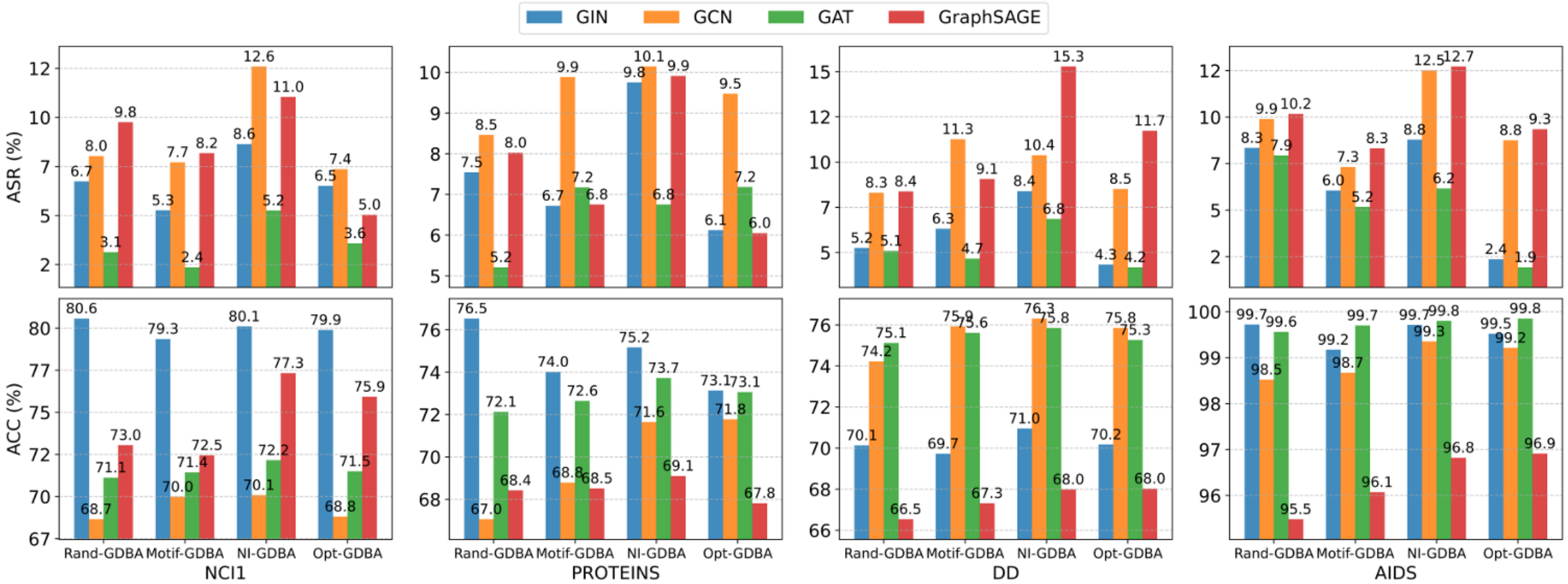
- 对于图分类任务，四种攻击场景下GEHINDER均能将ASR稳定至10%以下，同时ACC基本接近无攻击时的主任务性能，最高差2%左右
- FedTGE在隐蔽性最强的NI-GDBA攻击下明显失效，在AIDS上ASR高达越87%

实验结果 整体防御性能

Datasets (ACC w/o attack)	Attack	No-Defense		FoolsGold		Flame		RLR		FedTGE		GBHINDER	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CiteSeer (ACC=76.05)	Rand-GDBA	74.36	94.12	73.54	95.26	72.25	33.46	72.17	46.54	73.97	8.58	75.08	7.53
	GTA-GDBA	71.79	96.51	70.32	78.62	69.76	85.51	69.06	61.63	72.32	13.13	72.18	11.58
	UGBA-GDBA	75.26	86.71	67.98	6.78	73.76	26.86	74.12	84.27	74.09	18.14	74.86	5.67
Pubmed (ACC=88.89)	Rand-GDBA	84.87	74.87	83.23	63.71	84.35	41.48	80.59	80.76	85.75	22.26	86.53	13.74
	GTA-GDBA	83.58	92.08	82.85	70.01	85.45	59.49	81.58	82.48	87.34	12.07	87.19	8.92
	UGBA-GDBA	85.58	87.51	79.32	8.57	84.05	1.87	82.76	80.02	83.66	32.91	86.76	9.51
Coauthor-CS (ACC=91.25)	Rand-GDBA	86.89	93.95	85.72	89.41	86.01	32.91	84.07	84..58	85.63	27.95	88.52	15.24
	GTA-GDBA	91.39	71.29	89.59	52.56	88.92	21.50	91.35	24..54	90.53	12.18	91.13	10.54
	UGBA-GDBA	90.16	83.56	87.95	26.57	89.76	85.51	81.67	39.32	88.96	8.22	90.7	6.32
Amazon-Photo (ACC=84.08)	Rand-GDBA	84.25	95.33	84.82	94.98	81.08	31.08	78.26	92.44	80.81	2.08	83.18	7.24
	GTA-GDBA	81.84	72.93	77.18	57.73	78.68	8.82	76.87	48.24	80.26	3.17	82.05	2.11
	UGBA-GDBA	83.54	95.43	72.80	5.42	75.01	19.72	80.57	80.32	82.96	15.23	83.92	8.39

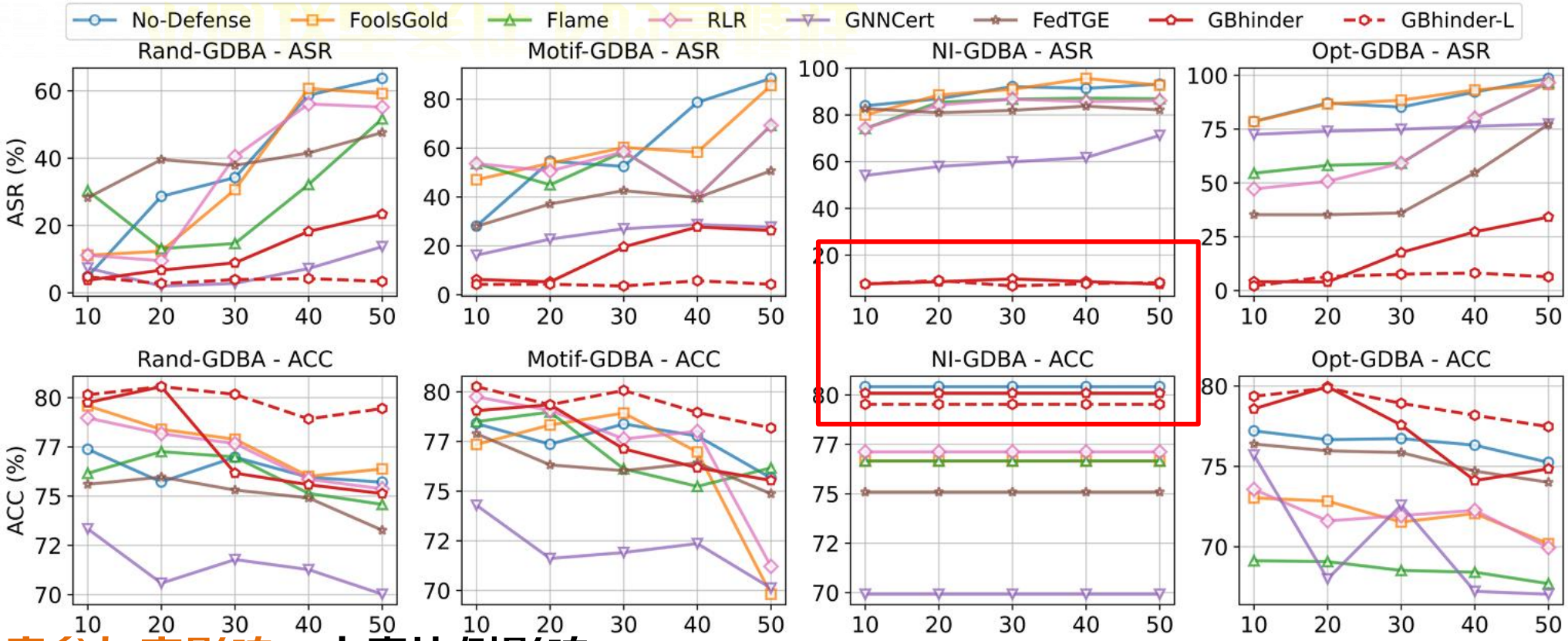
- 对于节点分类任务，GEHINDER在三种攻击场景下的ASR最高仅为15.24%
- 在PubMed+UGBA-GDBA场景下，FedTGE的ASR为32.91%，而GBHINDER降至9.51%，且ACC从83.66%提升到86.76%。说明GEHINDER有较好的任务泛化性能

实验结果 不同GNN骨干下的防御性能



- 在GIN、GCN、GAT、GraphSAGE四类骨干模型上均能有效降低ASR，同时保持较高ACC，防御性能较稳定
- 在默认GIN下GEHINDER的ASR整体略高于GAT，但整体ACC远高于GAT，说明历史通道注意力正则与注意力型GNN结构具有一定协同作用

实验结果 不同攻击条件下的鲁棒性



– 恶意参与率影响、中毒比例影响

- GBHINDER对于不同攻击场景在不同恶意客户端比例下整体保持较低ASR和较稳定ACC；其中GBHINDER-L（聚合前本地模型）的ASR始终最低

– 结论：GBHINDER的优势不只体现在全局模型防御效果，更体现在良性客户端具备持续自我保护能力，从而缓解高恶意参与率下的全局污染风险

实验结果 不同联邦设置下的适用性

Datasets	Setting	No-Defense		GBHINDER	
		ACC	ASR	ACC	ASR
NCI1	0.5	68.32	72.56	73.21 (↑4.89)	9.13 (↓63.43)
	1	75.26	79.63	78.56 (↑3.30)	7.32 (↓72.31)
	5	76.05	84.17	79.33 (↑3.28)	6.98 (↓77.19)
	10	77.12	87.21	80.12 (↑3.00)	6.01 (↓81.20)
	1000	76.71	87.52	79.96 (↑3.25)	6.44 (↓81.08)
PROTEINS	0.5	62.86	85.74	64.12 (↑1.26)	7.93 (↓77.81)
	1	70.93	90.17	72.05 (↑1.12)	8.06 (↓82.11)
	5	71.01	91.72	73.41 (↑2.40)	7.79 (↓83.93)
	10	71.98	93.56	74.92 (↑2.94)	7.71 (↓85.85)
	1000	72.49	92.83	75.05 (↑2.56)	7.68 (↓85.15)
DD	0.5	57.03	77.63	62.52 (↑5.49)	6.26 (↓71.37)
	1	67.78	76.73	69.71 (↑1.93)	5.98 (↓70.75)
	5	68.46	80.12	70.04 (↑1.58)	6.01 (↓74.11)
	10	68.82	80.64	71.19 (↑2.37)	5.16 (↓75.48)
	1000	71.02	78.53	70.78 (↓0.24)	5.29 (↓73.24)
AIDS	0.5	92.22	80.42	93.78 (↑1.56)	11.15 (↓69.27)
	1	96.12	82.77	98.05 (↑1.93)	9.78 (↓72.99)
	5	99.27	84.56	99.61 (↑0.34)	9.65 (↓74.91)
	10	98.58	86.39	99.45 (↑0.87)	7.98 (↓78.41)
	1000	99.11	84.62	99.58 (↑0.47)	8.16 (↓76.46)

- 结果分析
  - No-Defense：在不同α设置下ASR始终较高，在NCI1上最高达87.52%，PROTEINS最高93.56%
  - GBHINDER：在强Non-IID场景下仍能显著降低ASR，PROTEINS上α=0.5时ASR由85.74%降至7.93%
- 结论
  - 服务器端防御容易将良性客户端间的统计漂移误判为异常
  - GBHINDER不比较不同客户端更新，而是基于客户端自身可信历史知识净化全局模型，因此对数据异构性更稳定

实验结果 消融实验

$\mathcal{L}_{ce}$	$\mathcal{L}_{align}$	$\mathcal{L}_{topo}$	ACC	ASR
✓	✗	✗	76.66 (↓3.21)	87.11 (↑80.60)
✗	✓	✗	68.42 (↓11.45)	35.74 (↑29.23)
✗	✗	✓	65.39 (↓14.48)	48.75 (↑42.24)
✗	✓	✓	62.52 (↓17.35)	11.73 (↑5.22)
✓	✗	✓	75.94 (↓3.93)	40.53 (↑34.02)
✓	✓	✗	77.53 (↓2.34)	28.79 (↑22.28)
✓	✓	✓	**79.87**	**6.51**

- **损失函数消融**
  - 仅使用$\mathcal{L}_{ce}$，模型保持基本分类能力，但ASR高达87.11%
  - 去除$\mathcal{L}_{align}$后，ASR升至40.53%；去除$\mathcal{L}_{topo}$后，ASR升至28.79%
  - 三者均存在，ACC高达79.87%，ASR低至6.51%

Setting	*w/o* Both		*w/o* HCAR		*w/o* AMIU		GBHINDER	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
$\alpha = 0.5$	68.32 (↓4.89)	72.56 (↑63.43)	68.79 (↓4.42)	51.37 (↑42.24)	62.50 (↓10.71)	41.72 (↑32.59)	73.21	9.13
$\alpha = 1$	75.26 (↓3.30)	79.63 (↑72.31)	72.57 (↓5.99)	46.71 (↑39.39)	69.75 (↓8.81)	25.71 (↑18.39)	78.56	7.32
$\alpha = 5$	76.05 (↓3.28)	84.17 (↑77.19)	75.86 (↓3.47)	42.24 (↑35.26)	74.61 (↓4.72)	30.27 (↑23.29)	79.33	6.98
$\alpha = 10$	77.12 (↓3.00)	87.21 (↑81.20)	78.28 (↓1.84)	29.93 (↑23.92)	77.76 (↓2.36)	14.06 (↑8.05)	80.12	6.01
$\alpha = 1000$	76.71 (↓3.25)	87.52 (↑81.08)	81.09 (↑1.13)	33.19 (↑26.75)	78.15 (↓1.81)	12.72 (↑6.28)	79.96	6.44

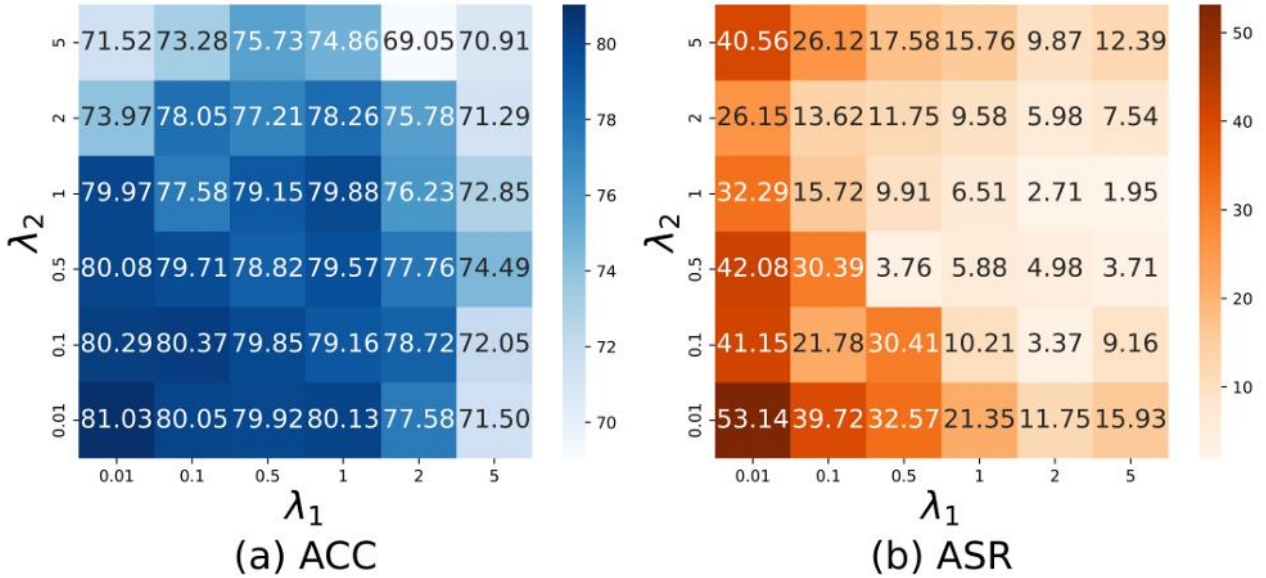
- **核心模块消融**
  - 去除HCAR后，α=0.5时ASR升至51.73%，说明历史通道注意力正则是后门净化的核心
  - 去除AMIU后，α=0.5时ACC降至62.50%、ASR升至41.72%，说明强Non-IID下静态历史锚点容易产生偏差导致主任务泛化能力骤降

实验结果 超参数实验

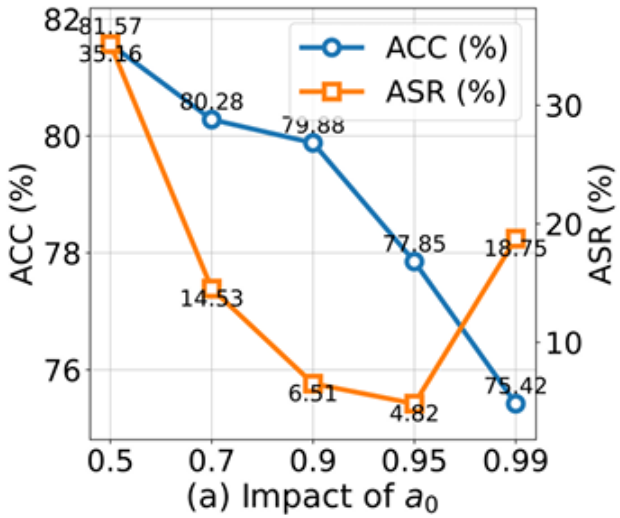
	ACC	ASR
No-Defense	76.66	87.11
$<l,l>$	80.53 (↑3.87)	17.39 (↓69.72)
$<l,l+1>$	79.88 (↑3.22)	6.51 (↓80.60)
$<l,l+2>$	72.34 (↓4.32)	15.24 (↓71.87)
$<l,l+3>$	57.56 (↓19.10)	47.71 (↓39.40)

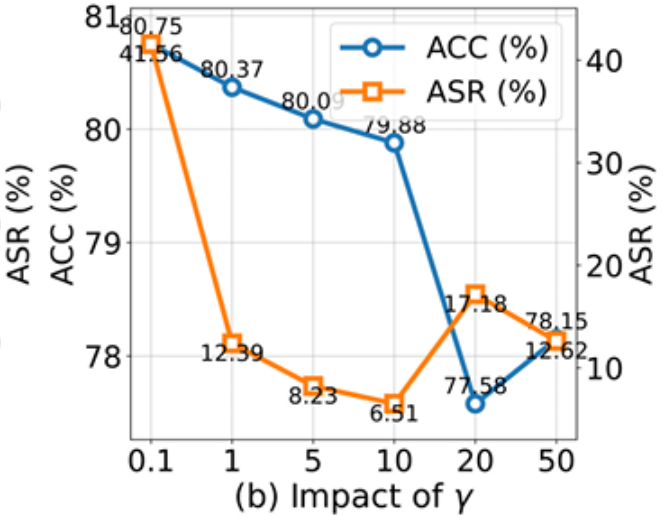
- 层对齐策略超参
  - $<l,l+1>$ 在低ASR与高ACC之间取得最佳平衡
  - $<l,l+1>$ 利用历史模型第l层的干净浅层特征约束全局模型第$l+1$层，可主动打断后门信号的放大链
  - $<l,l>$偏向被动修正，$<l,l+2>$ 等跨度过大，容易导致训练不稳定并降低ACC

后门触发器最初表现为局部异常模式，并在GNN消息传递过程中逐层放大，最终在深层形成稳定恶意表征



(a) ACC (b) ASR





实验结果 计算开销

DataSets	Methods	ASR (%)	Client (s)	Server (s)
NCI1	Fedavg	92.67	0.5012	–
	FedTGE	18.96 (↓73.71)	2.0965 (↑1.5953)	+0.019
	CNNCert	13.75 (↓78.92)	1.7474 (↑1.2462)	+0.521
	GBHINDER (%5)	21.80 (↓70.87)	1.8686 (↑1.3674)	–
	GBHINDER (%25)	17.57 (↓75.10)	2.2751 (↑1.7739)	–
	GBHINDER (%50)	12.39 (↓80.28)	3.8260 (↑3.3248)	–
AIDS	Fedavg	86.96	0.2067	–
	FedTGE	12.97 (↓73.99)	1.1423 (↑0.9356)	+0.021
	CNNCert	31.47 (↓55.49)	0.8064 (↑0.5997)	+0.298
	GBHINDER (%5)	18.13 (↓68.83)	0.6932 (↑0.4865)	–
	GBHINDER (%25)	11.59 (↓75.37)	1.6507 (↑1.4440)	–
	GBHINDER (%50)	6.96 (↓80.00)	2.0160 (↑1.8093)	–
PROTEINS	Fedavg	90.62	0.1340	–
	FedTGE	30.18 (↓60.44)	1.0432 (↑0.9092)	+0.032
	CNNCert	20.35 (↓70.27)	0.4071 (↑0.2731)	+0.309
	GBHINDER (%5)	18.75 (↓71.87)	0.5146 (↑0.3806)	–
	GBHINDER (%25)	9.31 (↓81.31)	1.1395 (↑1.0055)	–
	GBHINDER (%50)	11.39 (↓79.23)	2.7264 (↑2.5924)	–
DD	Fedavg	79.97	0.1224	–
	FedTGE	18.97 (↓61.00)	0.7658 (↑0.6434)	+0.026
	CNNCert	35.47 (↓44.50)	0.6957 (↑0.5733)	+0.698
	GBHINDER (%5)	19.13 (↓60.84)	0.4932 (↑0.3708)	–
	GBHINDER (%25)	9.72 (↓70.25)	0.9507 (↑0.8283)	–
	GBHINDER (%50)	4.66 (↓75.31)	2.2160 (↑2.0936)	–

- 结果分析
  - GBHINDER的计算开销主要来源于$\mathcal{L}_{topo}$的图边集遍历和敏感性分数δ的计算
  - 在NCI1上，FedAvg客户端耗时为 0.5012s，GBHINDER在50%采样率下增至 3.8260s，但ASR由92.67%降至12.39%
  - 降低采样率可显著减少开销：GBHINDER在AIDS数据集上5%采样率下客户端耗时0.6932s，ASR仍可降至18.13%
  - 相比FedTGE和GNNCert，GBHINDER额外计算主要发生在客户端侧，不会增加服务器端计算负担
- 结论
  - GBHINDER的计算开销可控，可通过简单采样策略实现防御效果与计算效率的灵活平衡

特点总结与未来展望

特点总结与未来展望

- 算法创新
  - FedID：在联邦学习场景中引入曼哈顿距离、欧氏距离和余弦相似度的多度量动态识别机制，通过白化加权与改进z-score实现隐蔽恶意梯度筛选
  - GBHINDER：提出无可信服务器的联邦图学习客户端主动净化范式，利用历史知识锚点、通道注意力正则化和自适应动量更新防御图后门攻击
- 算法优势
  - FedID：在多种后门攻击和Non-IID场景下保持较低BA与较高MA
  - GBHINDER：减少对可信服务器端异常检测和公开干净数据集的依赖
- 未来展望
  - 锚点轻量化维护： 将完整历史模型锚点通过低秩增量上传、模型量化压缩等技术降低客户端存储开销
  - 根据扰动敏感性δ动态决定是否启用$\mathcal{L}_{topo}$、AMIU或高频锚点更新，在防御性能与通信效率之间实现更灵活的平衡

参考文献

[1] Huang S, Li Y, Chen C, et al. Fedid: Enhancing federated learning security through dynamic identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[2] Zhu C, Mao Y, Zhang J, et al. Fend for Yourself! Backdoor Purification in Federated Graph Learning with an Evolving Knowledge Anchor[J].

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！