1 introduction

FL

概念：client：local train、global:返回给client

Benefit，vulnerable to poisoning attacks, Backdoor attacks(more challenge)

**Defence**

1 Dectection:

作用：identifying and removing potentially poisoned model updates

弊端：very specific adversary models，detailed assumptions about the attack strategy

2 (DP)：

作用：individual weights of model updates are clipped to a maximum ，random noise is added to the weights

对比1好处：generic adversary model

弊端：severely deteriorates the benign，因为太多noise

提出新防御：

结合前两种defence的优点

**贡献**：

FLAME：

定义

由3模块组成：

DP-base 噪音

消除后门贡献

自动模块

identifying and removing potentially poisoned model updates

自动模块

权重剪枝

后两个模块减少DP引起的噪音

贡献：

FLAME：4.1

对比之前的后门防御：FLAME可以generic和不用underlying data distributions of benign

高斯噪音可以：4.3

remove potentially malicious model updates

clipping the weights

噪音边界可以：5

消除后门贡献

评估：7

在3个真实数据库

一种安全多方计算算法：8

preserve the privacy of individual model

updates while realizing our backdoor defense approach

3.1过滤

图显示了不同聚类方法在一组模型更新的权重向量上的行为，包含三类权重，良性模型、后门模型、随机输入模型。

(a)显示了攻击场景的基本情况，其中使用两组客户端：一组用于注入后门，而另一组提供随机模型，目的是欺骗基于集群的防御。

(b)显示了在这种情况下，K-means未能成功地分离良性模型和中毒模型，因为所有中毒模型最终与良性模型位于同一集群中。

论文采用了HDBSCAN的聚类方法来过滤，区分良性与恶意模型。FLAME使用成对余弦距离来测量所有模型更新之间的角度差异，应用动态聚类算法HDBSCAN对两两之间的余弦相似度进行聚类，超过50 % 的类为良性更新。其他类均视为离群值，将其剔除，得到剩余的L 个良性模型。

这种行为在图5d中描述，其中图5c中簇B和簇C的所有模型都被认为是离群值。

这里的优点是，即使对手扩大模型更新以增强其影响，余弦距离也不会受到影响，因为这不会改变更新的权重向量之间的角度。由于HDBSCAN算法根据模型的余弦距离分布密度对模型进行聚类，和动态确定所需的集群，因此我们将其用于动态聚类方法。

因此，据我们所知，我们的方法是第一个FL后门防御，它也能够优雅地处理动态攻击，其中注入的后门的数量可能会有所不同。