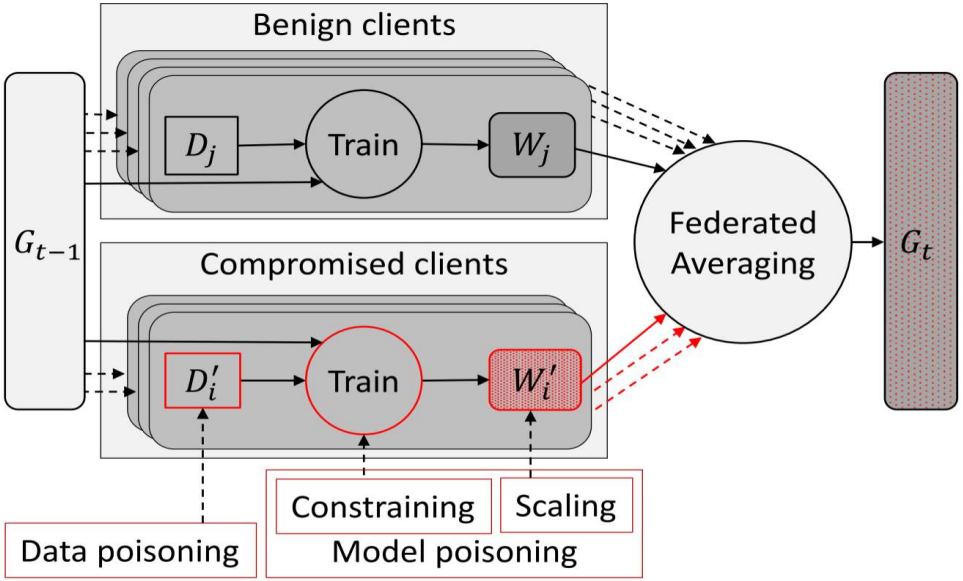
FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping

Xiaoyu Cao*[1], Minghong Fang*[2], Jia Liu[2], Neil Zhenqiang Gong[1]
[1] Duke University, {xiaoyu.cao, neil.gong}@duke.edu
[2] The Ohio State University, {fang.841, liu.1736}@osu.edu

NDSS 2021

NDSS / USENIX Security / S&P / CCS

Poison attacks in FL

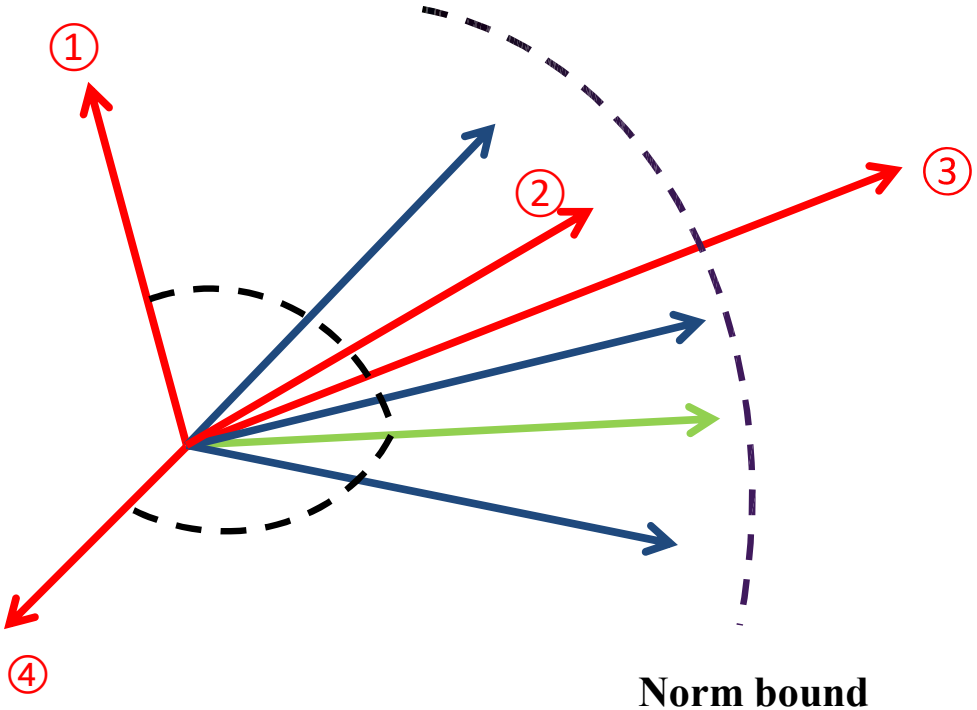


投毒攻击

- 非目标攻击 → 非特定类别样本攻击。目的降低全局模型性能
- 目标攻击(后门) → 特定类别的样本攻击。攻击者选定某类别部分样本添加触发器并将标签修改为指定类别，目的是在推理阶段对拥有触发器该类别样本错误预测为设定的类别。

Types of backdoor attacks

Backdoor update Distance and Direction attacks



Backdoor update

Benign update

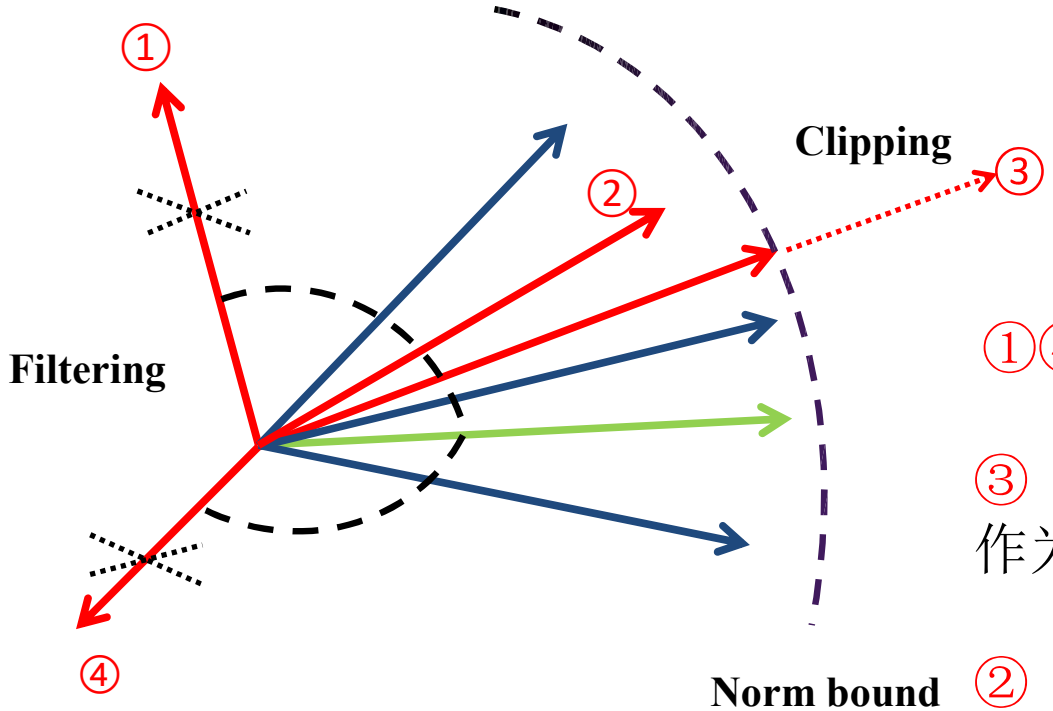
Global update

①④ Direction Attack

③ Distance Attack

Types of backdoor attacks

Distance and Direction defenses



Backdoor update

Benign update

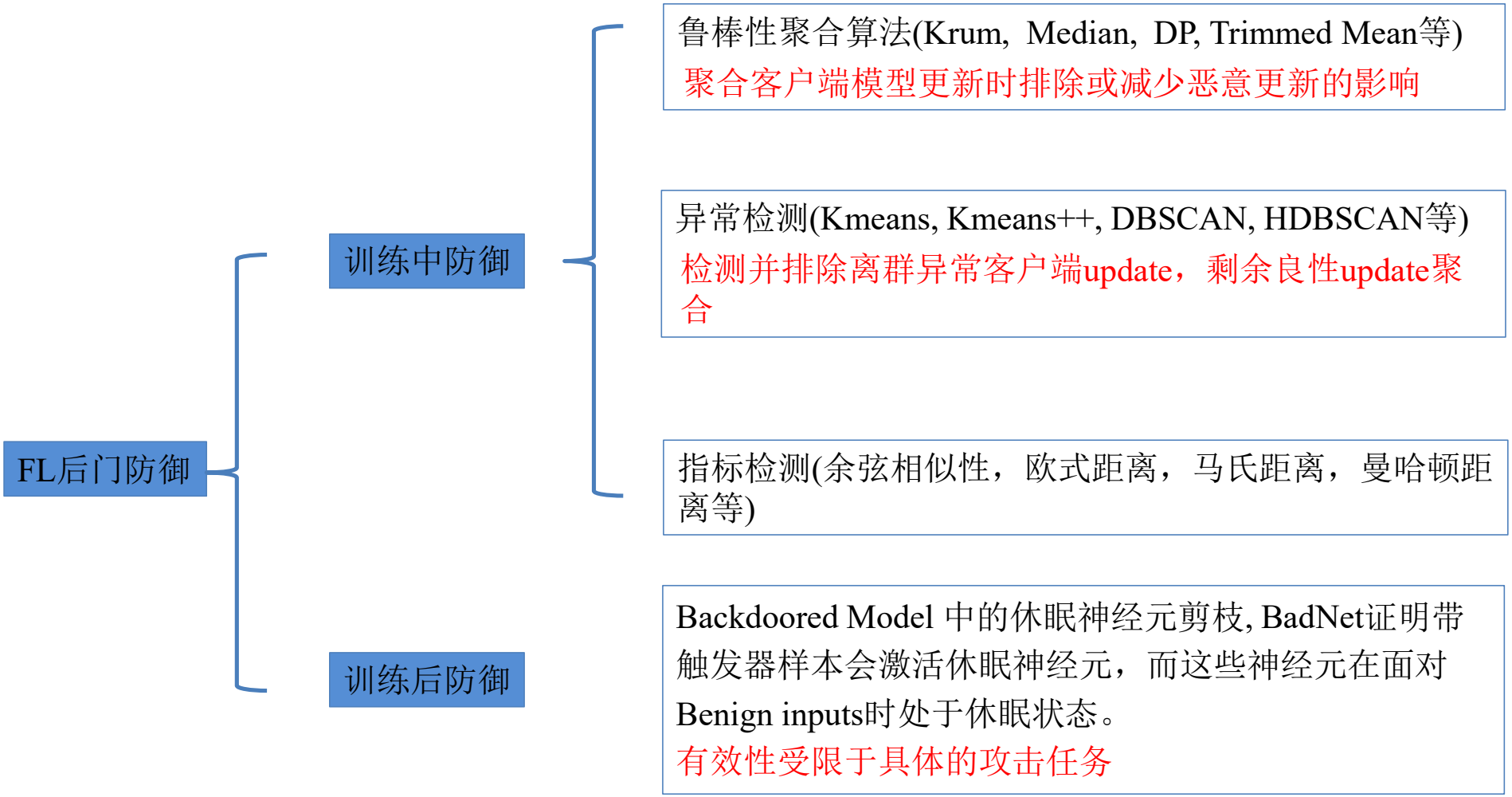
Global update

①④ 计算与剩余update余弦值总得分

③ 选择所有update Norm 中位数作为Norm bound裁剪

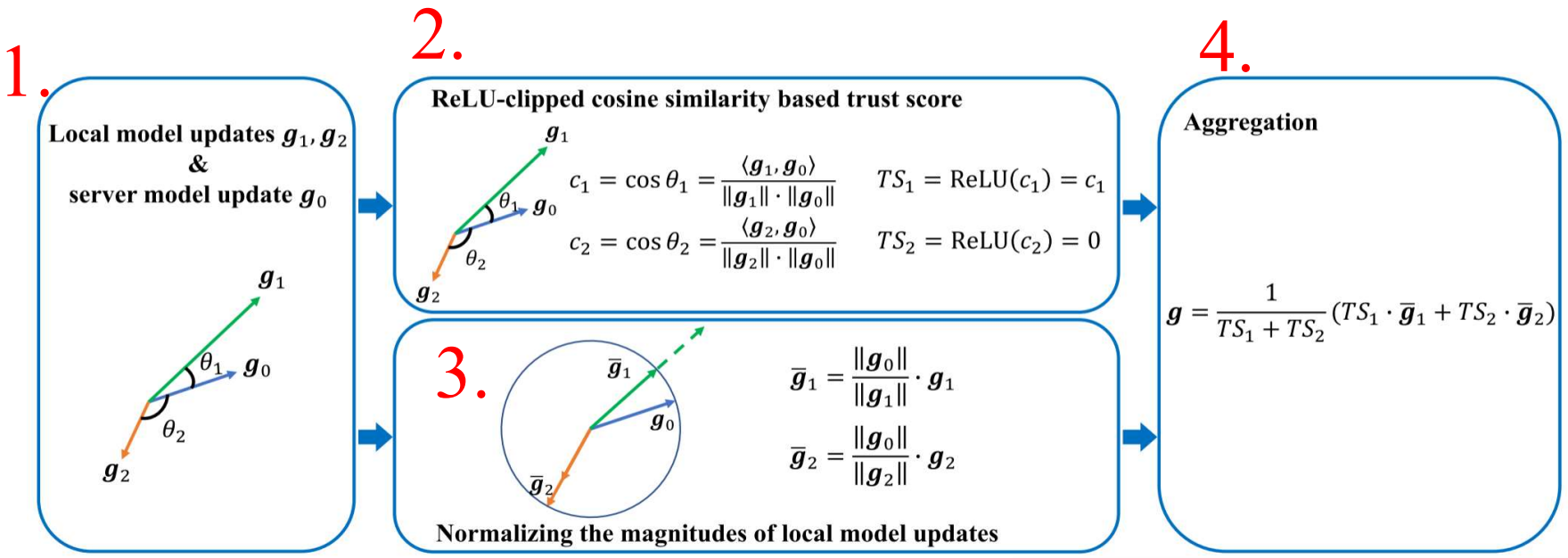
② 聚合之后加Noise扰动

Types of backdoor defenses



FLTrust (Robust Aggregation)

Server拥有root dataset(100个样本，label balance), Server与agents同步训练全局模型w_0，获得指导梯度g_0



Limitation

① 后门梯度与指导梯度方向有较大偏差时，防御效果越好。但是实际上一部分后门梯度方向不会产生大偏差，这类梯度通过训练累积依旧能训练出后门模型。

② 训练中后期性能难以提升。中后期良性梯度与指导梯度的偏差会增大， 对良性梯度的加权越来越小，导致模型性能难以提升。

FLAME (HDBSCAN)

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

FLAME: Taming Backdoors in Federated Learning

Thien Duc Nguyen[1], Phillip Rieger[1], Huili Chen[2], Hossein Yalame[1], Helen Möllering[1], Hossein Fereidooni[1], Samuel Marchal[3], Markus Miettinen[1], Azalia Mirhoseini[4], Shaza Zeitouni[1], Farinaz Koushanfar[2], Ahmad-Reza Sadeghi[1], and Thomas Schneider [1]

[1]*Technical University of Darmstadt, Germany*; [2]*University of California San Diego, USA*; [3]*Aalto University and F-Secure, Finland*; [4]*Google, USA*

USENIX Security 2022

FLAME

3-5 客户端训练获得更新模型w_i

6 w_i与w_j两两之间计算余弦相似性，获得相似性matrix

7 HDBSCAN聚类(-1和0)，保留0类(benign)

8 计算所有客户端Update L2-Norm

9 计算L2-Norm中位数S_t

10 S_t对0类中的更新Update 大小进行修正

11 聚合0类模型参数作为全局模型参数

13-14 对全局模型参数添加自适应高斯噪声

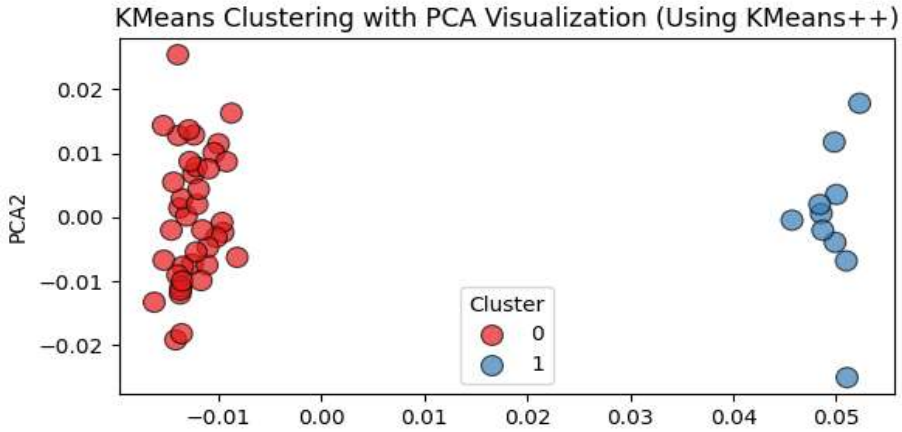
Algorithm 1 FLAME

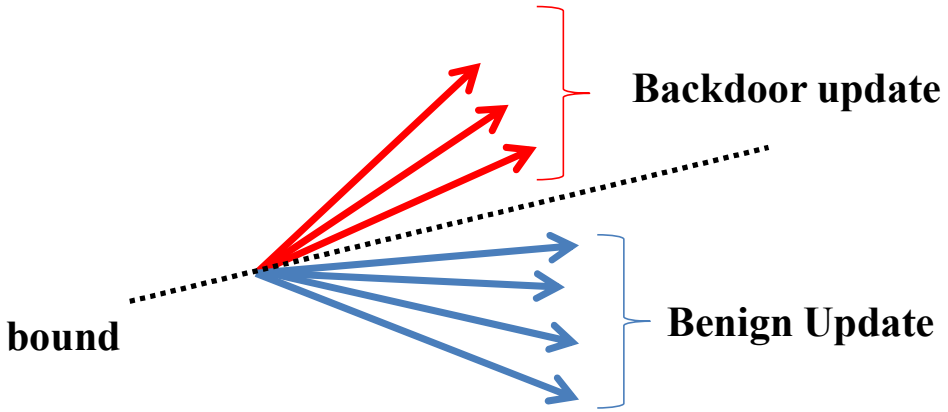
1: **Input:** n, G_0, T ▷ n is the number of clients, G_0 is the initial global model, T is the number of training iterations
2: **Output:** G_T^* ▷ G_T^* is the updated global model after T iterations
3: **for** each training iteration t in $[1,T]$ **do**
4: **for** each client i in $[1,n]$ **do**
5: $W_i \leftarrow \text{CLIENTUPDATE}(G_{t-1}^*)$ ▷ The aggregator sends G_{t-1}^* to Client i who trains G_{t-1}^* using its data D_i locally to achieve local modal W_i and sends W_i back to the aggregator.
6: $(c_{11},\ldots,c_{nn}) \leftarrow \text{COSINEDISTANCE}(W_1,\ldots,W_n)$ ▷ $\forall i,j \in (1,\ldots,n)$, c_{ij} is the cosine distance between W_i and W_j
7: $(b_1,\ldots,b_L) \leftarrow \text{CLUSTERING}(c_{11},\ldots,c_{nn})$ ▷ L is the number of admitted models, b_l is the index of the l^{th} model
8: $(e_1,\ldots,e_n) \leftarrow \text{EUCLIDEANDISTANCES}(G_{t-1}^*,(W_1,\ldots,W_n))$ ▷ e_i is the Euclidean distance between G_{t-1}^* and W_i
9: $S_t \leftarrow \text{MEDIAN}(e_1,\ldots,e_n)$ ▷ S_t is the adaptive clipping bound at round t
10: **for** each client l in $[1,L]$ **do**
11: $W_{b_l}^c \leftarrow G_{t-1} + (W_{b_l} - G_{t-1}) \cdot \text{MIN}(1,\gamma)$ ▷ Where $\gamma (= S_t/e_{b_l})$ is the clipping parameter, $W_{b_l}^c$ is the admitted model after clipped by the adaptive clipping bound S_t
12: $G_t \leftarrow \sum_{l=1}^{L} W_{b_l}^c / L$ ▷ Aggregating, G_t is the plain global model before adding noise
13: $\sigma \leftarrow \lambda \cdot S_t$ where $\lambda = \frac{1}{\varepsilon} \cdot \sqrt{2ln\frac{1.25}{\delta}}$ ▷ Adaptive noising level
14: $G_t^* \leftarrow G_t + N(0,\sigma^2)$ ▷ Adaptive noising

IID

K-means || K-means++ || DBSCAN || HDBSCAN || LOF || COF || GMM || Isolation Forest

K-means++



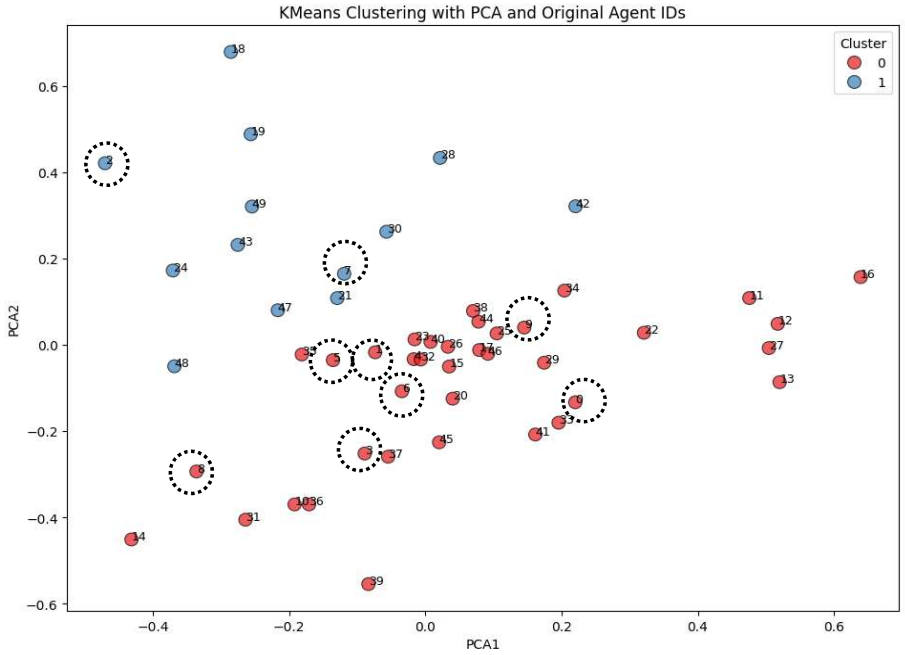


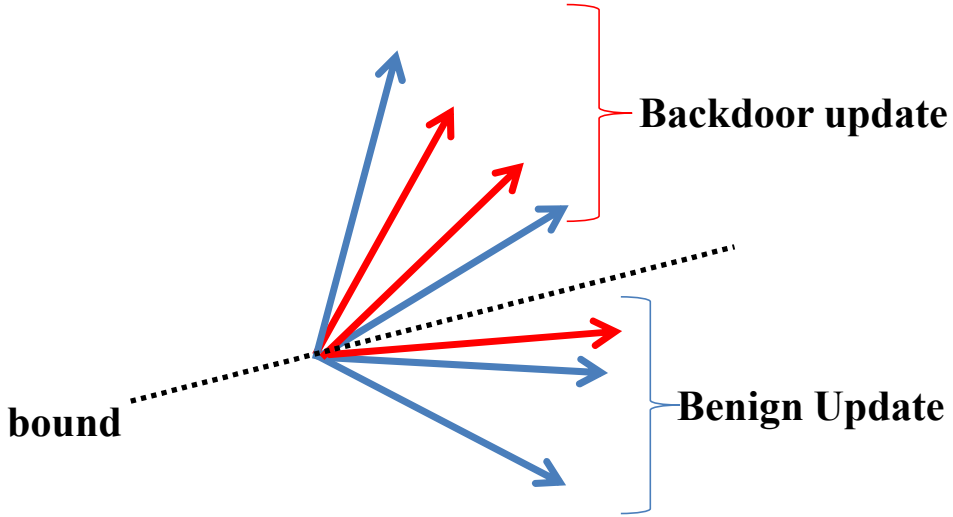
后门客户端要对部分样本进行投毒，后门客户端之间的数据分布更加相似；

良性客户端不参与投毒过程，数据依然是同分布；

后门客户端的update具有相似性，良性客户端的update同时也具有相似性。

Non-IID





Non-IID导致良性客户端之间的update差异增加，这种差异为后门update隐藏在良性update中创造了条件，清除后门update时很容易误判良性update, 同时剩余的客户端中依旧包含少量的后门update，持续攻击下每轮少量的后门update被聚合，多轮之后依旧可以训练出后门模型，Non-IID比IID难防御。

THANKS