《Understanding distributed poisoning attack in federated learning》

文章的主要思想是：

进行了分布式投毒攻击的实验，并提出了应对分布式投毒的解决方案-Sniper。
在 Sniper 中，服务器构建一个图，图中的顶点是指在更新过程中从参与者收集的局部模型，如果两个局部模型足够接近（即欧氏距离相对较小），则它们之间存在一条边。然后，服务器通过解决图中的“最大团”问题来识别诚实的本地模型。仅通过汇总所得集团所包含的那些局部模型来获得全局模型。实验结果表明，在 Sniper 的保护下，即使有三分之一的参与者是攻击者，攻击成功率也下降到 2%。

本文对分布式投毒的假设：

1. 所有攻击者合谋，有着相同的攻击目标。攻击类别为 Target error poisoning attack（标签反转攻击、后门攻击）。
2. 攻击者无法观察其他诚实客户端的训练数据，即攻击者无法推断任何诚实参与者的参数。
3. 攻击者的数量不超过 N/3。

文章对分布式投毒的实验设置：

1. 对 MNIST 数据集进行标签反转攻击。
2. 客户端总数 N=10。
3. 攻击者数目 P=1/2/3。
4. 每个客户端选取的样本训练数：500。
5. 中毒样本数：300~500（间隔 50），平均分配给所有攻击者。
6. 20 轮训练。

（攻击者的最大数量应小于参与者总数的 1/3 [17]，否则，联合平均算法的损失将不再减少。）

分析攻击实验：

1. 攻击成功率随中毒样本数量**线性**增加。
2. 中毒样本数量不变的情况下，增加攻击者数量可以提高攻击的成功率。
3. 攻击成功率的增速随攻击次数增加。（攻击次数越多，攻击成功率增长得越显著）

防御设计（Sniper）：

Algorithm 2 Identify Honest Participansts

$\gamma \Leftarrow 0.5$
$\epsilon \Leftarrow 0.05$ //the length to decrease γ
$Honest \Leftarrow \phi$ //$Honest$ is the set of honest models
$V \Leftarrow \{M_1, M_2, ..., M_N\}$ //V contains all the models
while $Honest == \phi$ **do**
  //if distance between two models is less than γ, they are neighbors G
  for i *in* $range(N-1)$ **do**
    $G(M_i) \Leftarrow \phi$
    for j *in* $range(i+1, N)$ **do**
      if $Dis(M_i, M_j) < \gamma$ **then**
        $G(M_i) \Leftarrow G(M_i) \cup M_j$
      end if
    end for
  end for
  $Cliques \Leftarrow BronKerbosch(V, G)$ //find all cliques
  //sort $Cliques$ by size and find the largest one
  $MaxClique \Leftarrow FindLargestClique(Cliques)$
  if $|MaxClique| > \frac{N}{2}$ **then**
    $Honest \Leftarrow MaxClique$
    return $Honest$
  end if
  $\gamma \Leftarrow \gamma + \epsilon$
end while

1. 服务器 S 从 N 个参与方收集其局部模型M_i，然后计算每两个模型之间的欧式距离
2. S 设置初始阈值γ（实验中设为 0.5），然后 S 构建一个图，图中的顶点对应每个局部模型，如果任两个局部模型的欧式距离小于γ，则这两个顶点之间存在一条边$e_{i,j}$。
3. 找到图中的最大团，并检查团中的顶点数是否大于 N / 2。如果是，则使用下式汇总团中的顶点（局部模型），得到全局模型，否则用增大 γ 转到第 2 步。

$$M_G^{(t)} = \frac{1}{N}\sum_{k=1}^{N} M_k^{(t)}$$

最大团问题为 NP 完全问题，文章使 BronKerbosch 算法找到途中所有的团，进而识别最大团作为诚实候选集。

文章特点：

1. 用欧式距离考察更新之间的相似度。
2. 构建图，将分类问题转化为寻找图中“最大团”问题，用图论算法筛选诚实客户端。