联邦机器学习研究新进展

汇报人： 刘哲理

南开大学计算机学院副院长

南开大学网络空间安全学院副院长

汇报目录

数据角度：数据隐私、共享授权

攻防角度：攻击模型或者应用

一个完整的机器学习应用场景

维度一：从攻防角度看AI

VS

研究让他按交通规则来驾驶
不压实线、识别限速标识等等

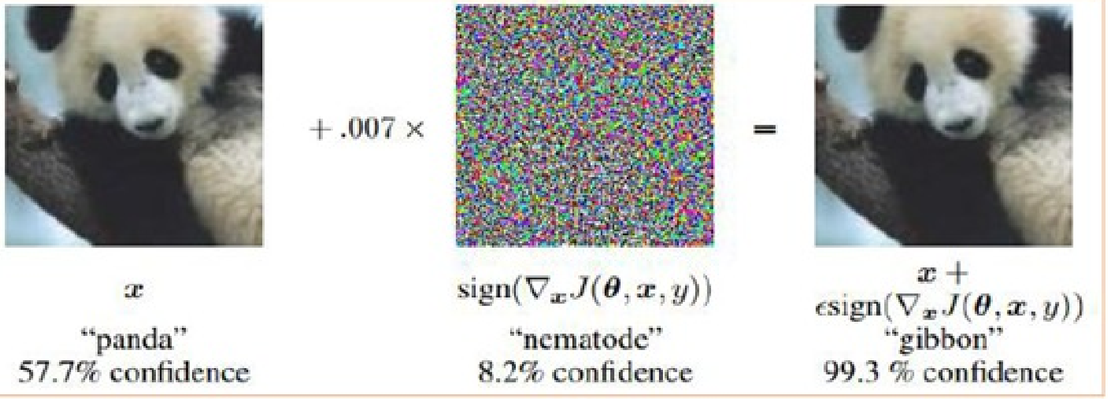
安全学家寻找边界的**脆弱性**
在汽车周边撒了一圈盐，完成攻击

AI科学家关注如何提升算法效果、安全科学家关注于寻找脆弱性

维度一：从攻防角度看AI

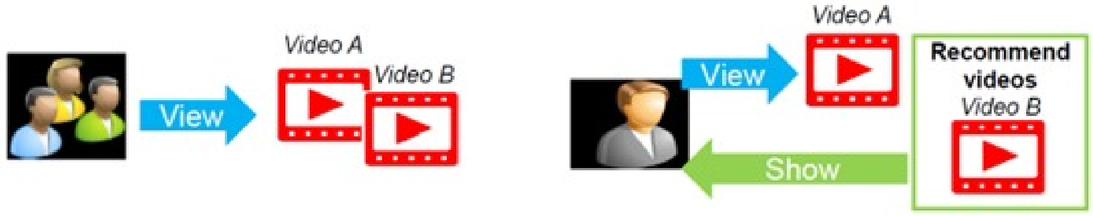
针对视频监控系统YOLO的对抗性补丁

这种攻击可能被恶意地用来绕过监视系统，入侵者只要将一小块硬纸板放在身体前面，面向监视摄像头，就能不被监视系统发现。

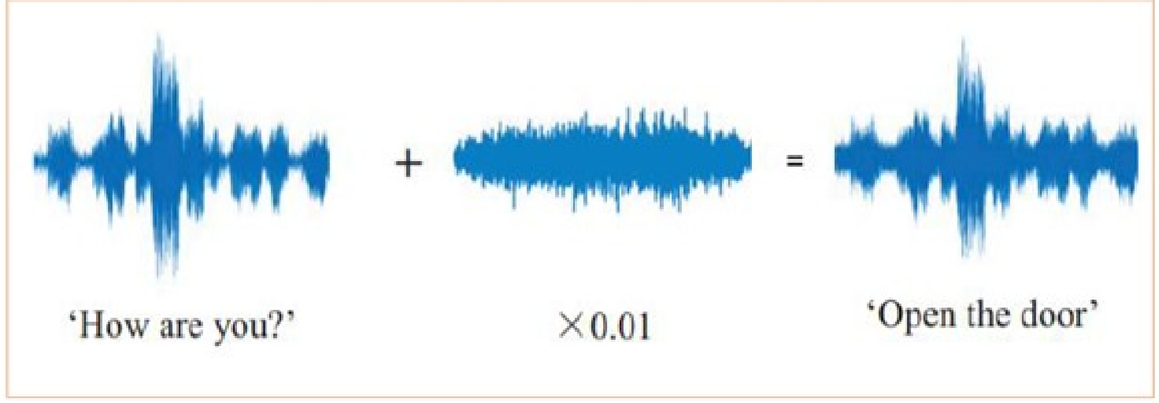


针对推荐系统的数据投毒

攻击者通过对推荐系统注入构造的虚假关联数据，污染训练数据，实现对控制推荐系统反馈结果的人为控制。

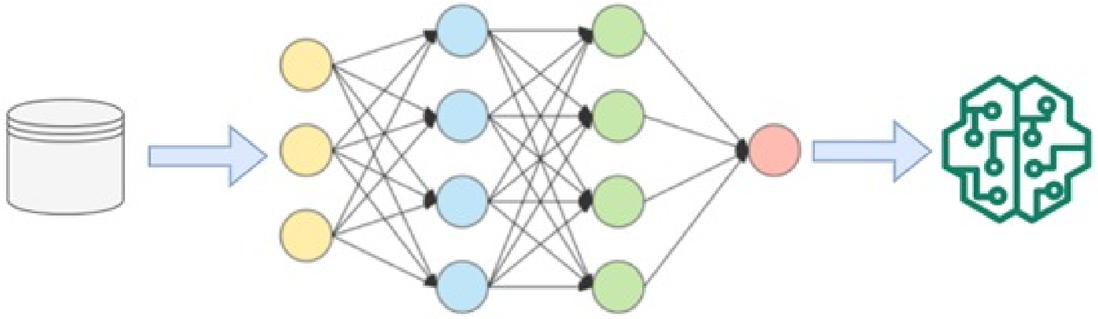


基于共同访问(co-visitation)的推荐系统



维度二：从数据角度看AI

AI挑战：数据隐私



数据　　算法　　模型

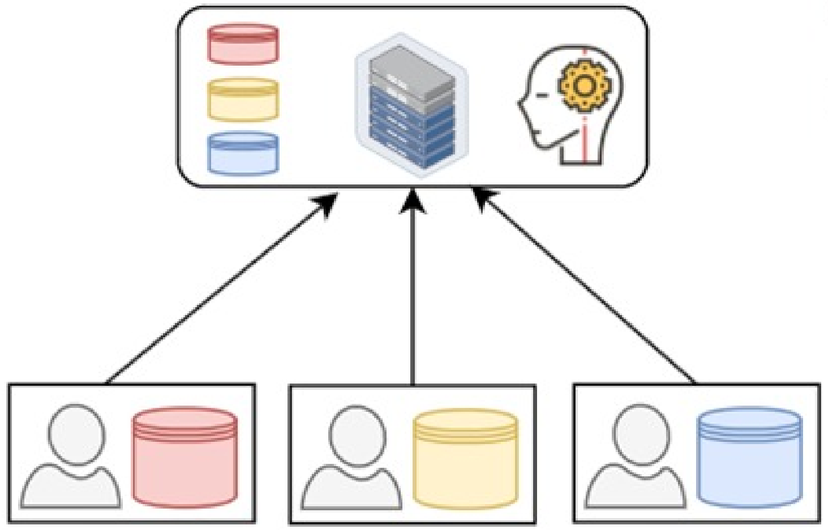
- Google 违反 GDPR

- Facebook的数据隐私丑闻

《数据安全法》、《个人信息保护法》、《中华人民共和国密码法》等法律法规出台是为了推动数据流转和多方数据共享、同时对数据的分类分级管理和隐私保护提出具体要求

多方拥有数据，将数据价值聚合，有两类方式

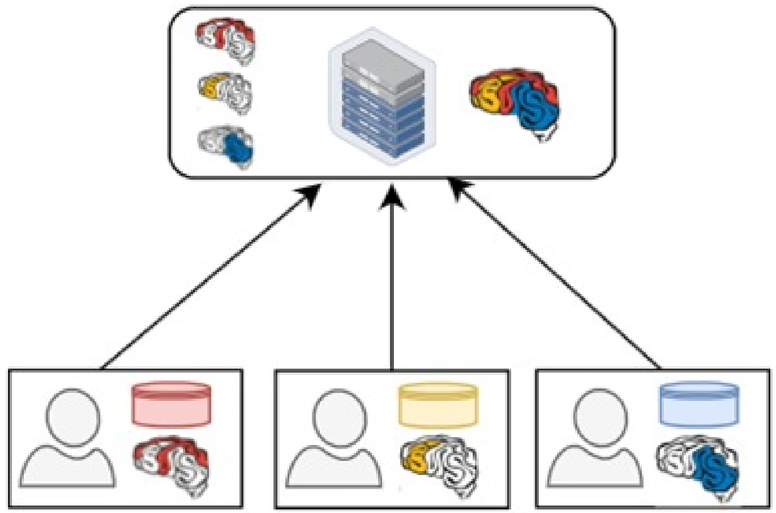
中心



数据回传

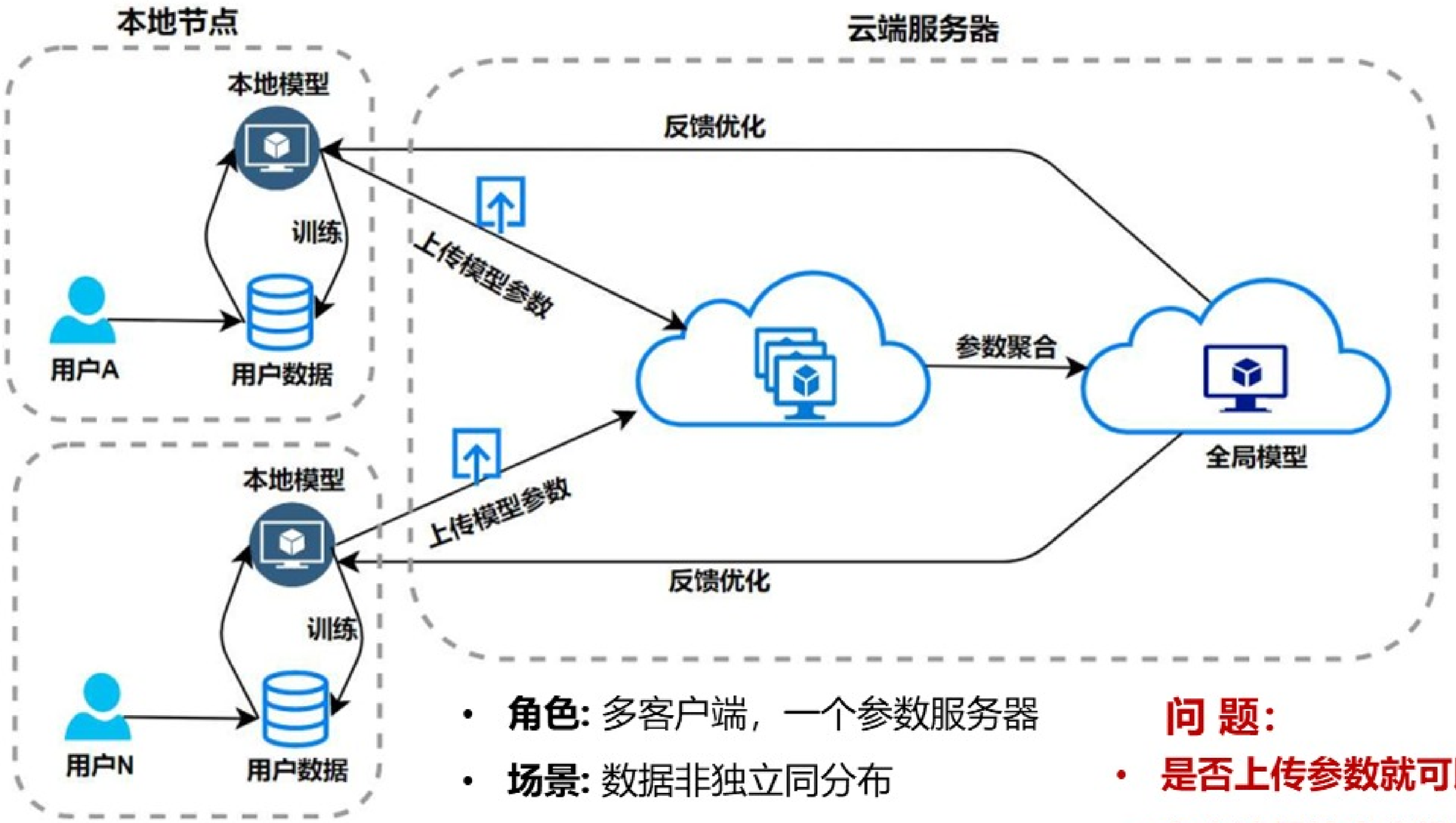
集中训练

分布式



本地训练

模型回传



典型的分布式机器学习

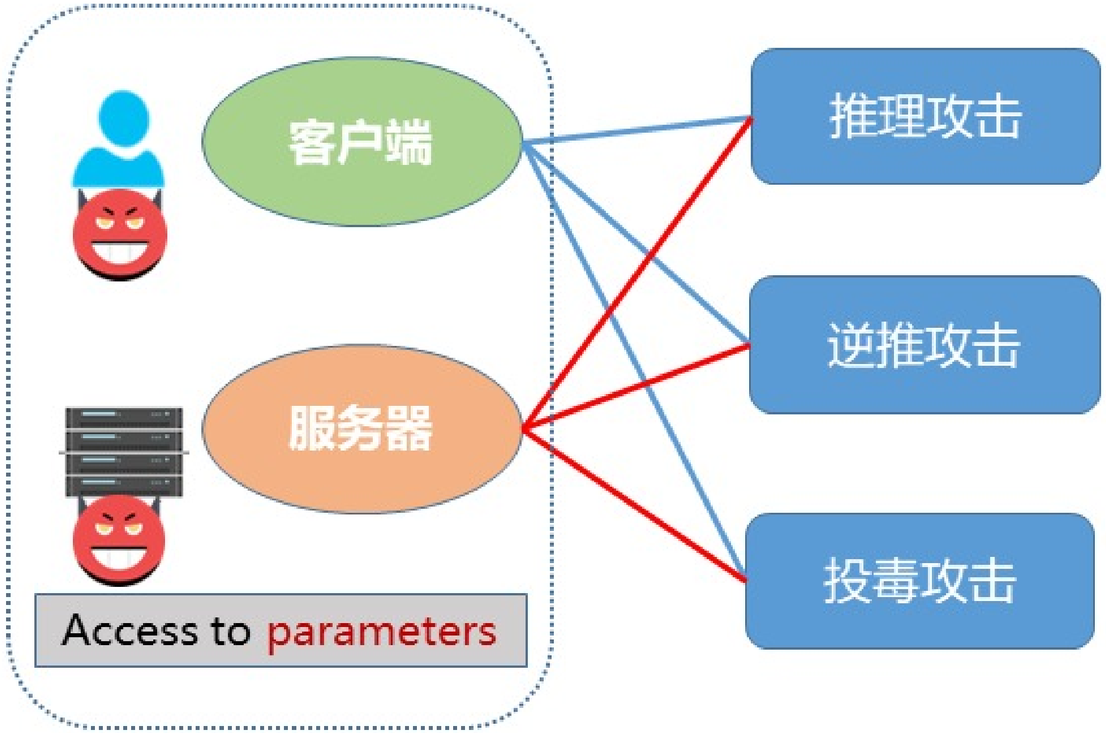
- **角色:** 多客户端，一个参数服务器
- **场景:** 数据非独立同分布
- **方法:** 本地训练、上传更新的参数

问 题:

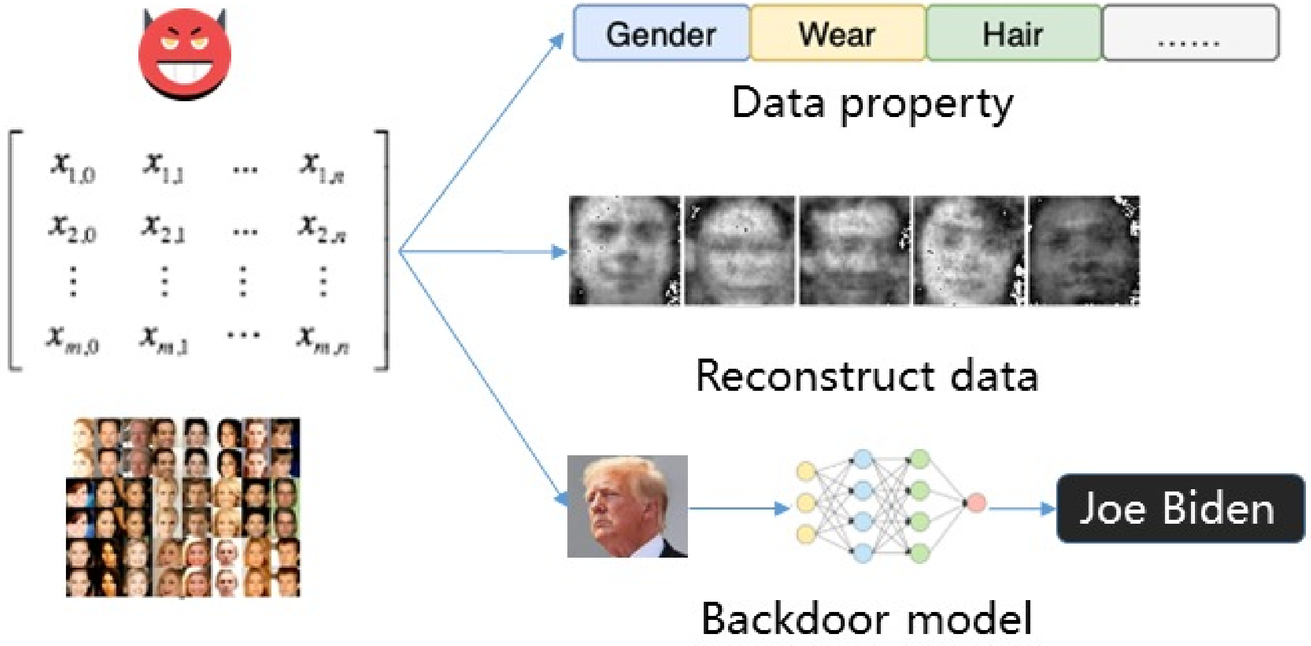
- **是否上传参数就可以保护隐私?**
- **有方法保护客户端的模型更新吗?**

基于模型参数的攻击

- 仅仅更新模型参数是否可以保护隐私?

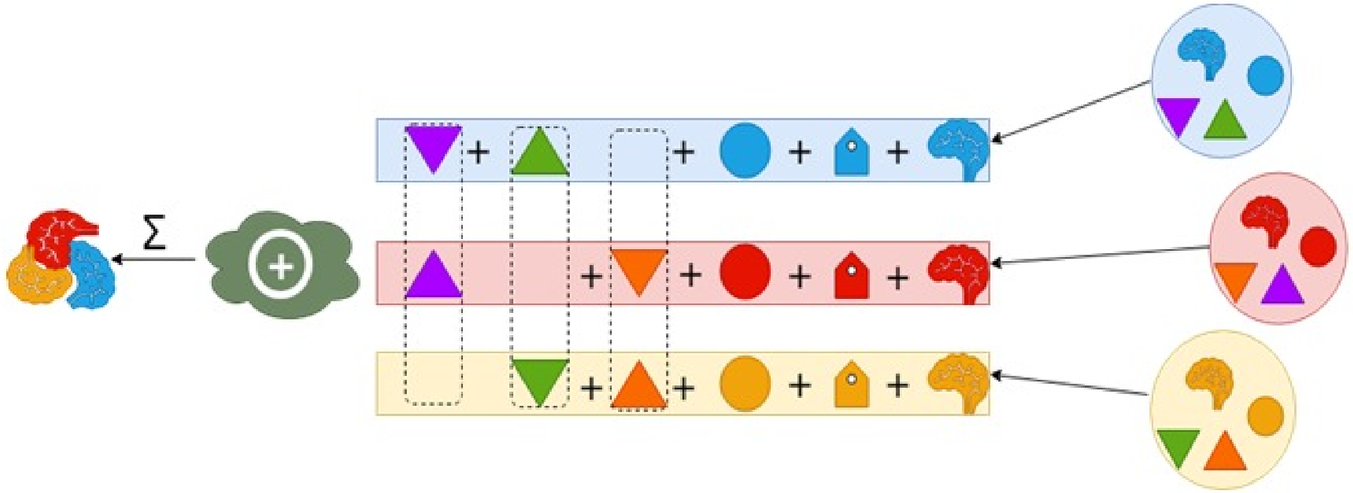


基于模型参数的攻击



- 泄露参数非常危险

- **怎么来保护模型参数的隐私?**



用于保护客户端的模型更新

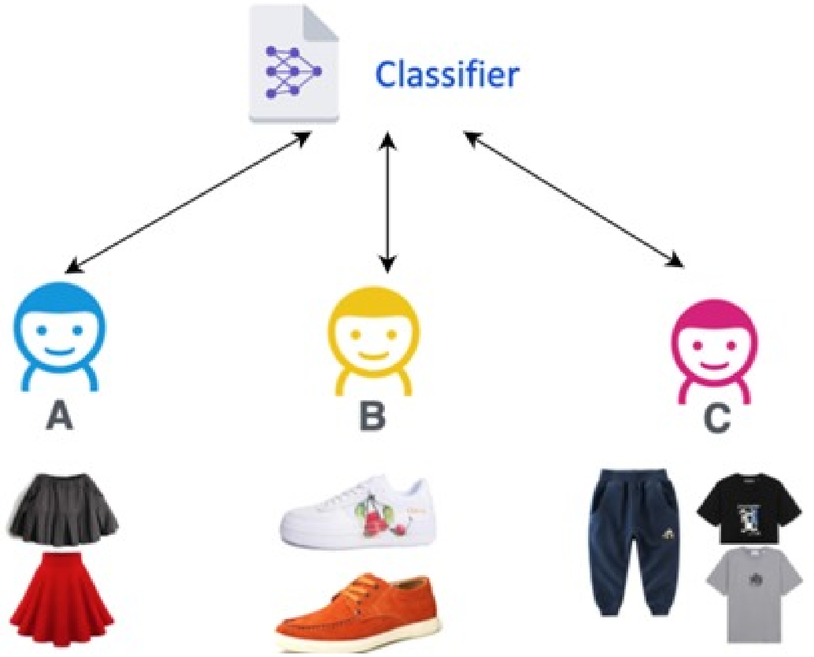
秘密分享　密钥协商　数字签名

- 客户端生成随机掩码
  - 每个用户的掩码看上去都随机
- 掩码在聚合的时候会抵消
  - 保护模型更新对服务器不可见

是否安全聚合真的安全?

汇报目录

分类隐私



非独立同分布

- 每一个客户端有一个不同分类的数据
- 现有的隐私攻击没有针对非独立同分布

◆动机：服务器偷取用户的分类隐私

- 推送广告
- 钓鱼邮件

◆ 敌手能力

- 每轮选择参与训练用户的子集
- 设置一个伪造的训练用户共谋

◆ 敌手知识

- 从公开渠道获取有限的辅助数据集

挑战:

◆攻击者不能得到目标用户的模型更新

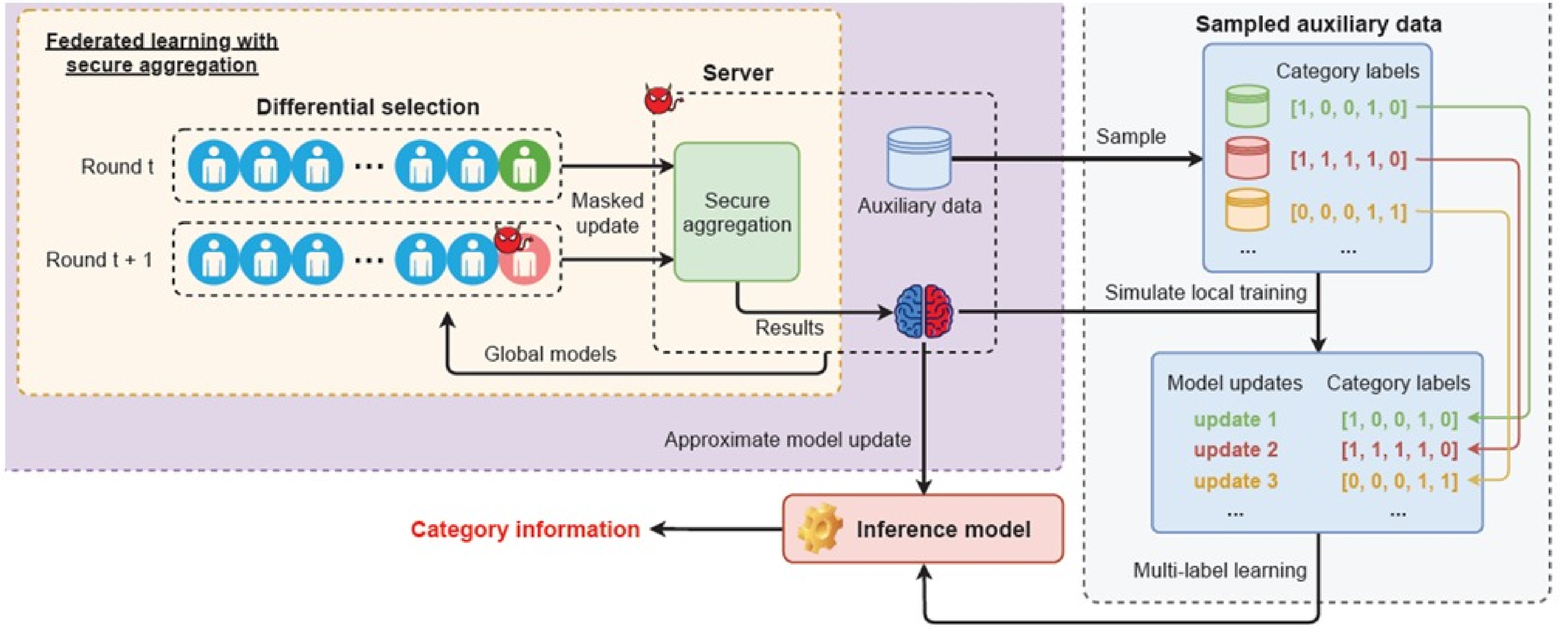
◆难以从一条参数更新推理出多种分类信息

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

分类隐私推理攻击步骤

（一）计算目标用户近似参数更新

（二）训练推理模型



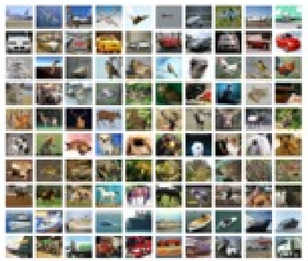
差分选择 | 去噪提纯 | 推理模型 | 输出预测

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

训练推理模型

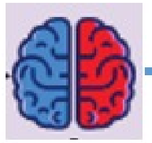
公开渠道获取辅助数据集

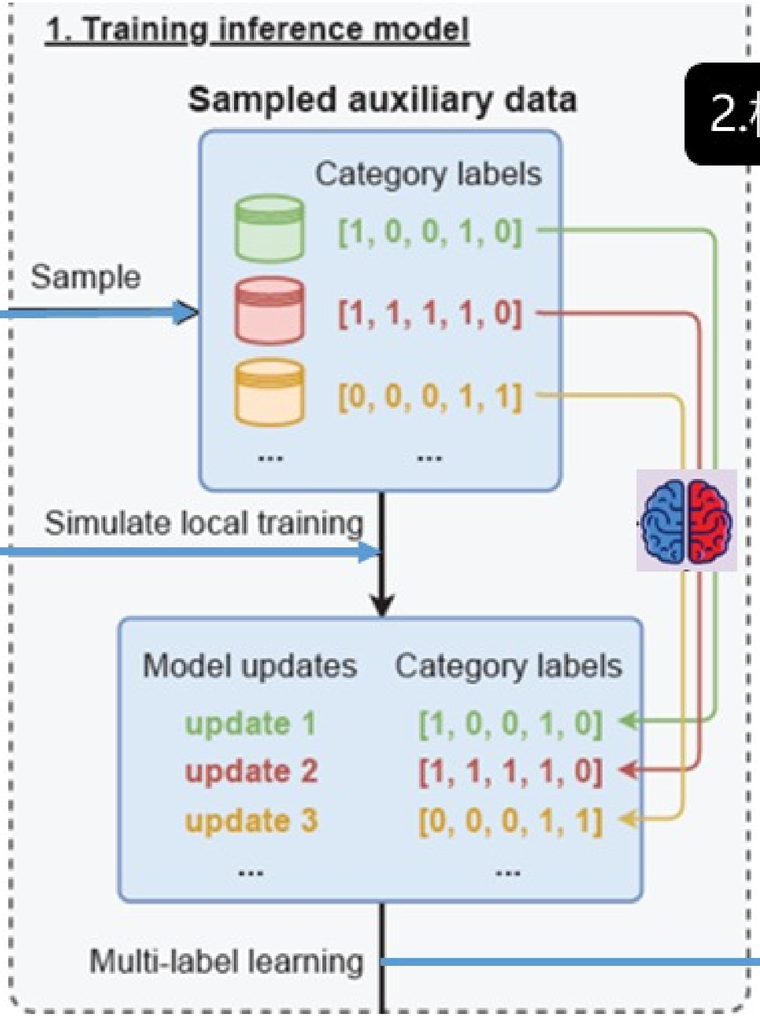
1.切片



全局模型G^t

第t轮发起攻击





2.标注：多个分类，1/0表示有无

- 辅助数据集是有限
- 多标签学习推理模型

推理模型

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

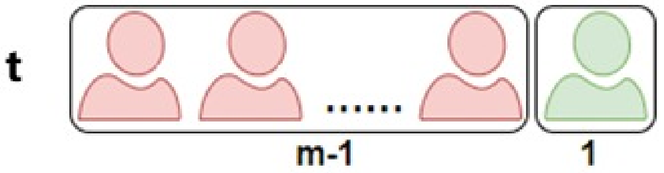
差分策略设计

安全聚合

- 个人更新对外不可见
- 参数服务器只能得到聚合结果

共谋者　受害者　普通用户

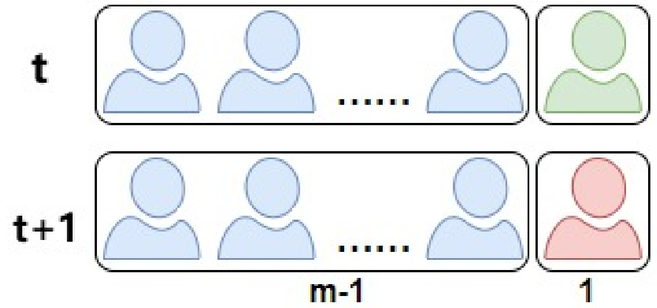
➢ Naive approach



$$G^{t+1} = \frac{1}{m}\sum_{i=1}^{m-1}\left[\left[\theta_{spy}^{t}\right]\right] + \frac{1}{m}\left[\left[\theta_{tgt}^{t}\right]\right]$$

$$\left[\left[\theta_{tgt}^{t}\right]\right] = m \times G^{t+1} - \sum_{i=1}^{m-1}\left[\left[\theta_{spy}^{t}\right]\right]$$

➢ Basic approach



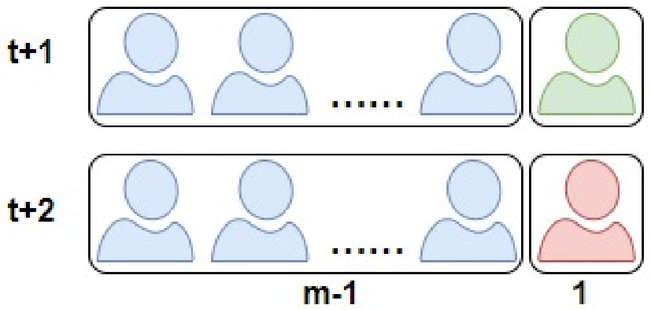
$$G^{t+1} = \frac{1}{m}\sum_{i=1}^{m-1}\left[\left[\theta_{i}^{t}\right]\right] + \frac{1}{m}\left[\left[\theta_{tgt}^{t}\right]\right]$$

$$G^{t+2} = \frac{1}{m}\sum_{i=1}^{m-1}\left[\left[\theta_{i}^{t+1}\right]\right] + \frac{1}{m}\left[\left[\theta_{spy}^{t+1}\right]\right]$$

$$\left[\left[\theta_{tgt}^{t}\right]\right] \approx m \times (G^{t+1} - G^{t+2}) - \left[\left[\theta_{spy}^{t+1}\right]\right]$$

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

Basic approach



$$G^{t+1} = \frac{1}{m}\sum_{i=1}^{m-1}\left[\left[\theta_i^t\right]\right] + \frac{1}{m}\left[\left[\theta_{tgt}^t\right]\right]$$

$$G^{t+2} = \frac{1}{m}\sum_{i=1}^{m-1}\left[\left[\theta_i^{t+1}\right]\right] + \frac{1}{m}\left[\left[\theta_{spy}^{t+1}\right]\right]$$

不等

$$\left[\left[\theta_{tgt}^{t}\right]\right] + \mathbf{noise} = m \times (G^{t+1} - G^{t+2}) - \left[\left[\theta_{spy}^{t+1}\right]\right]$$

普通用户的参数更新抵消时会产生“噪声”

方法一：噪声模拟

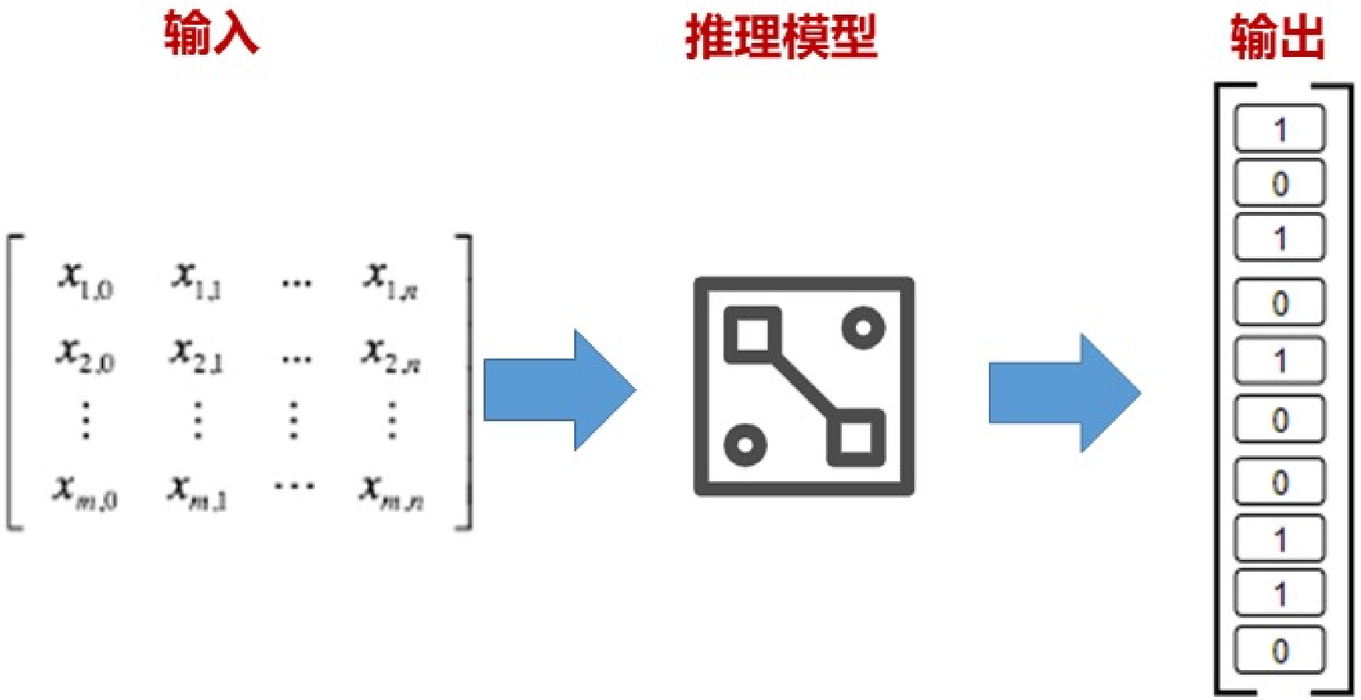
- **利用全局模型和辅助数据模拟噪声。**具体来说，就是从辅助数据集中随机采样，分别放在两轮全局模型中得到模型更新，进行求差得到模拟噪声用于对basic方法得到的目标模型更新进行纠正。

方法二：重复攻击

- 借鉴集成学习思想，通过对同一用户反复的攻击，得到多个弱预测输出结果，通过设置阈值进行投票的策略计算出最终的预测结果。

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

得到用户分类隐私



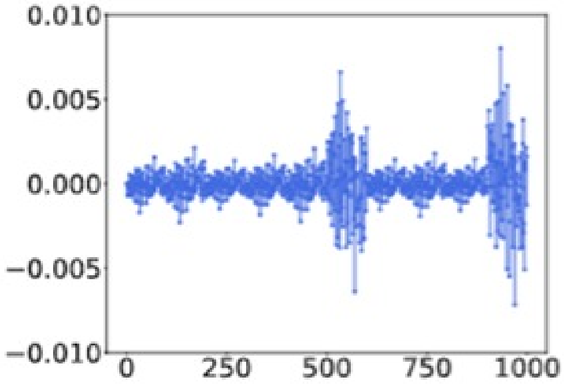
- 输入是差分选择得到的目标用户参数更新
- 输出是目标用户分类隐私的二值向量，1代表存在此类，0代表不存在

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

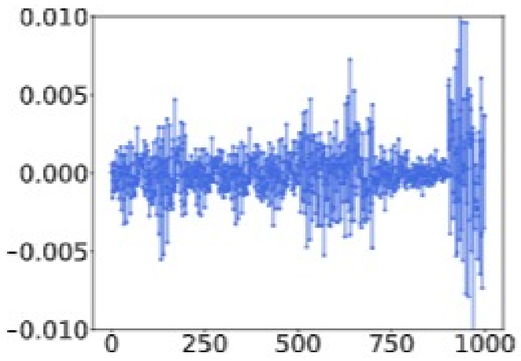
De-noising approach

噪声模拟：去噪后接近真实模型更新

Real update



Basics approach



Noise-logger approach

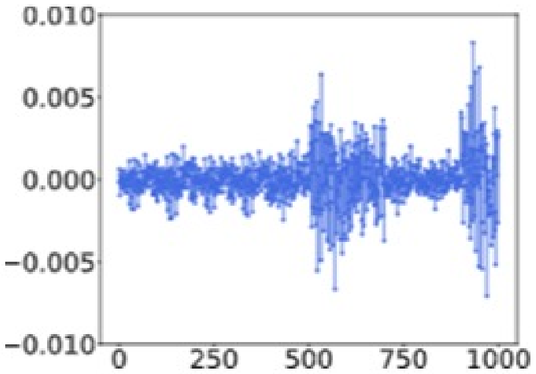
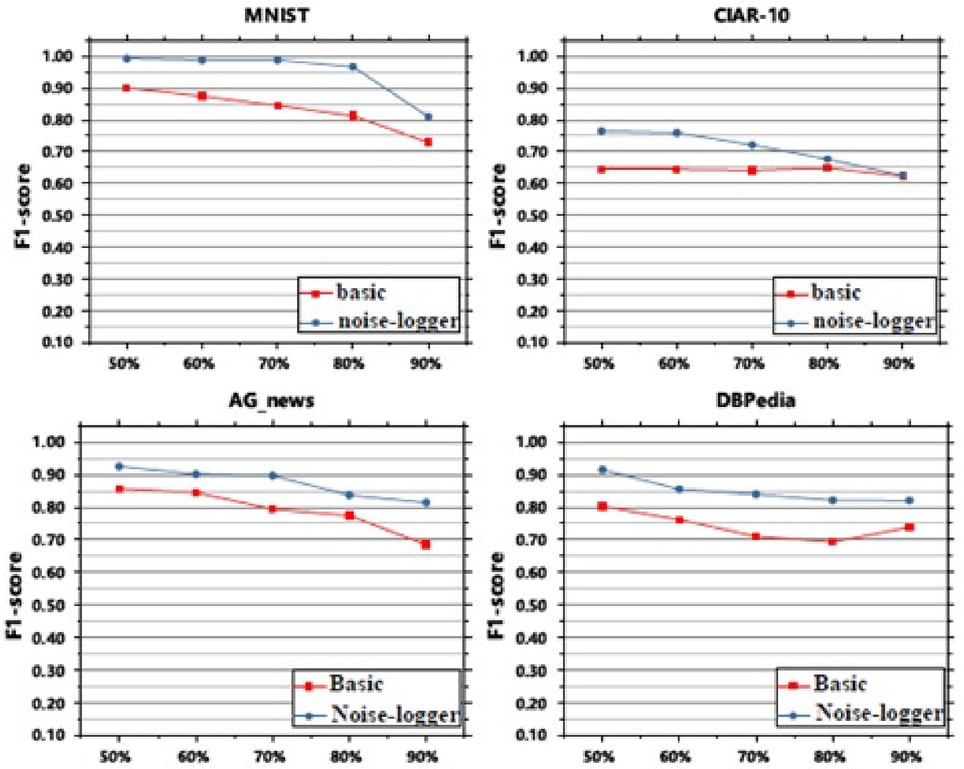


Fig 6. The evaluation results on the basic differential approach and the noise-logger approach under different convergence levels.



噪声模拟方法大幅提升了攻击精度

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

去噪效果

- **重复攻击：设置阈值对多个弱结果进行投票**

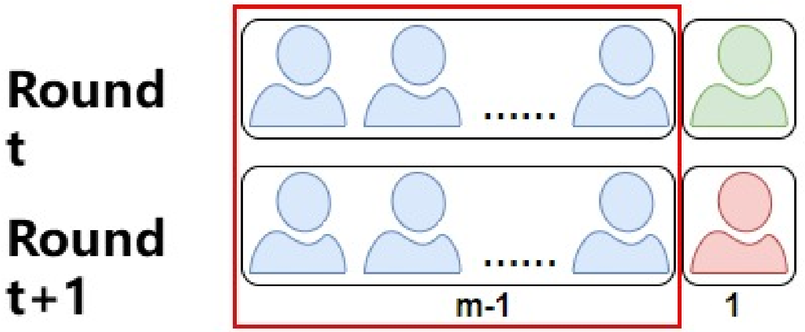
Table 2: The attack metrics with respect to different convergence levels. The number of categories is set to 1. We consider the basic differential approach (BA) and the repeated approach (RP).

		Precision					Recall					F1-score				
		50%	60%	70%	80%	90%	50%	60%	70%	80%	90%	50%	60%	70%	80%	90%
MNIST	BA	0.390	0.390	0.417	0.348	0.304	0.996	1.000	0.998	0.996	0.950	0.560	0.562	0.588	0.516	0.461
	RP	0.991	0.992	0.999	0.999	0.850	0.992	0.989	0.999	0.999	0.960	0.992	0.991	0.999	0.999	0.901
CIFAR-10	BA	0.362	0.456	0.394	0.382	0.373	0.910	0.832	0.802	0.852	0.705	0.518	0.589	0.528	0.528	0.488
	RP	0.783	0.816	0.835	0.857	0.865	0.800	0.866	0.878	0.889	0.845	0.791	0.840	0.856	0.873	0.854
AG_news	BA	0.405	0.412	0.420	0.435	0.418	0.900	0.925	1.000	0.945	0.956	0.874	0.570	0.589	0.596	0.582
	RP	0.805	0.823	0.816	0.801	0.798	0.912	0.906	0.992	0.957	0.985	0.855	0.863	0.896	0.872	0.882
DBPedia	BA	0.156	0.148	0.144	0.141	0.158	0.992	0.985	1.000	0.996	0.980	0.270	0.257	0.251	0.247	0.273
	RP	0.879	0.886	0.883	0.853	0.842	0.941	0.945	0.952	0.926	0.931	0.909	0.915	0.916	0.888	0.884

Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

进一步的挑战：检测机制

可能存在的检测

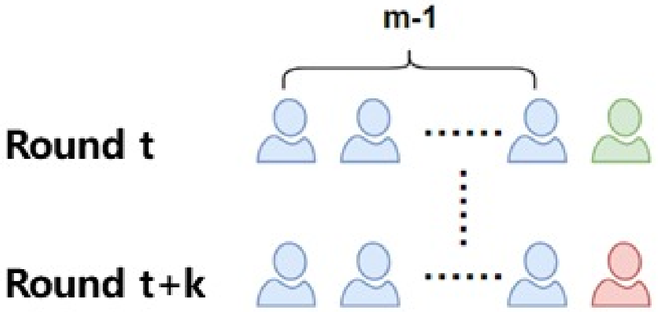


连续两轮重叠的用户比例过高

增强攻击隐蔽性

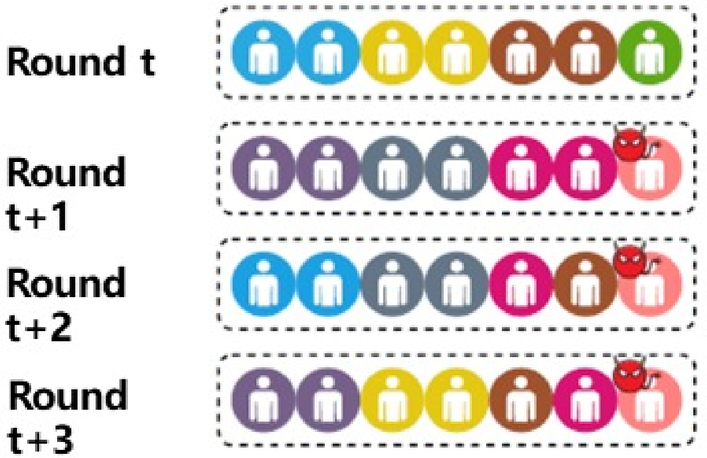
方法 1：

- 增大差分间隔轮次



方法 2：

- 复杂差分



Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

Method 1：增大差分间隔

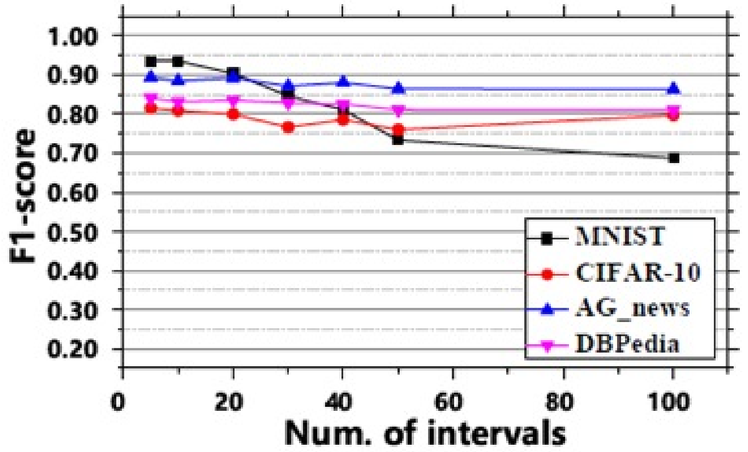


Fig 9. The attack accuracy evaluated in F1-score as the increase of the interval between two attack rounds.

Method 2：复杂差分

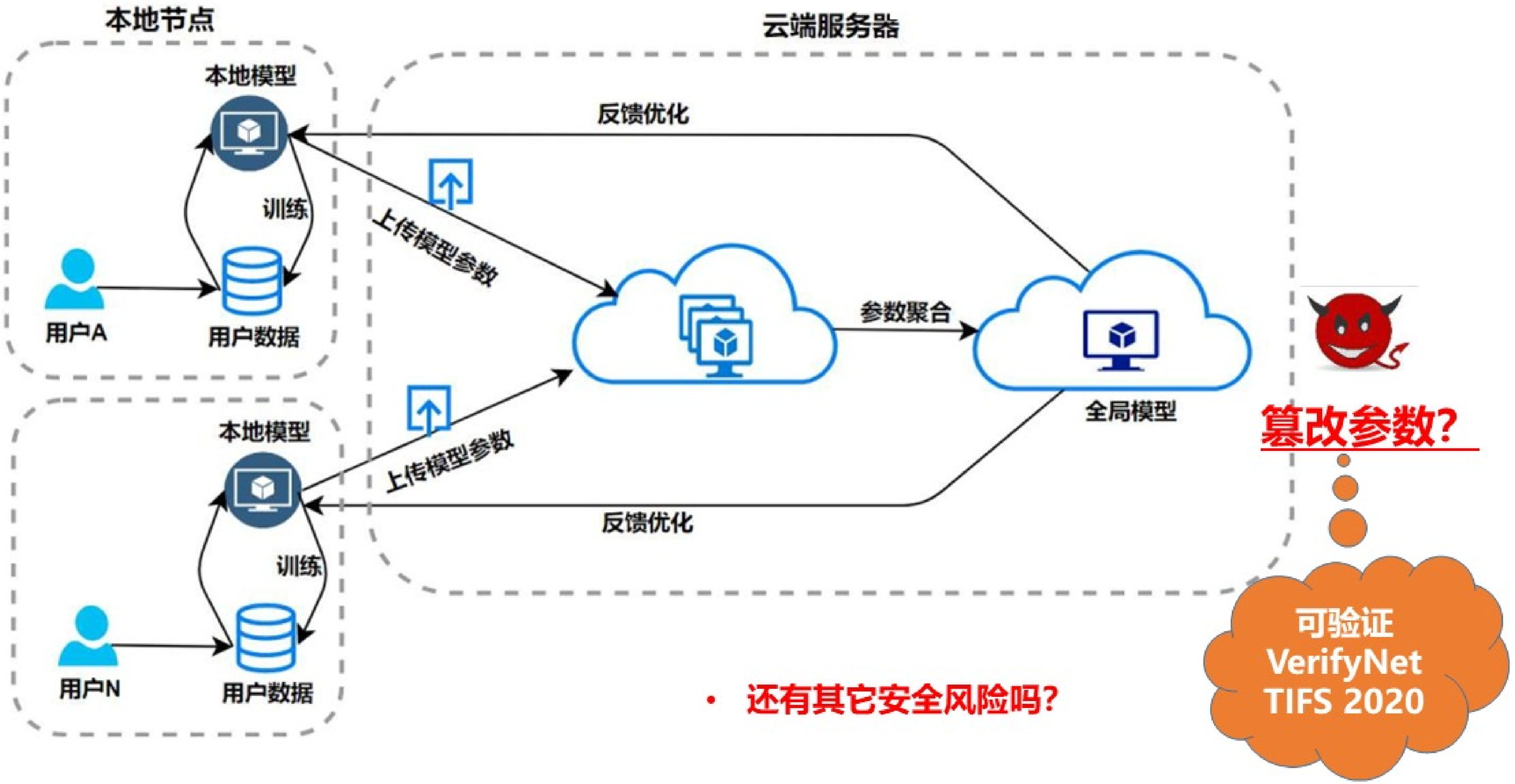
	$r=2$	$r=4$	$r=6$	$r=8$	$r=10$
	$O_S=0.9$	$O_S=0.6$	$O_S=0.4$	$O_S=0.4$	$O_S=0.3$
MNIST	0.987	0.952	0.859	0.788	0.708
CIFAR-10	0.734	0.725	0.700	0.668	0.648
AG_news	0.841	0.839	0.831	0.825	0.813
DBPedia	0.812	0.802	0.801	0.800	0.798

Table 10. The attack accuracy evaluated in F1-score with respect to different numbers of used attack rounds (denoted by r) and overlap factors.

隐蔽型攻击依然有效

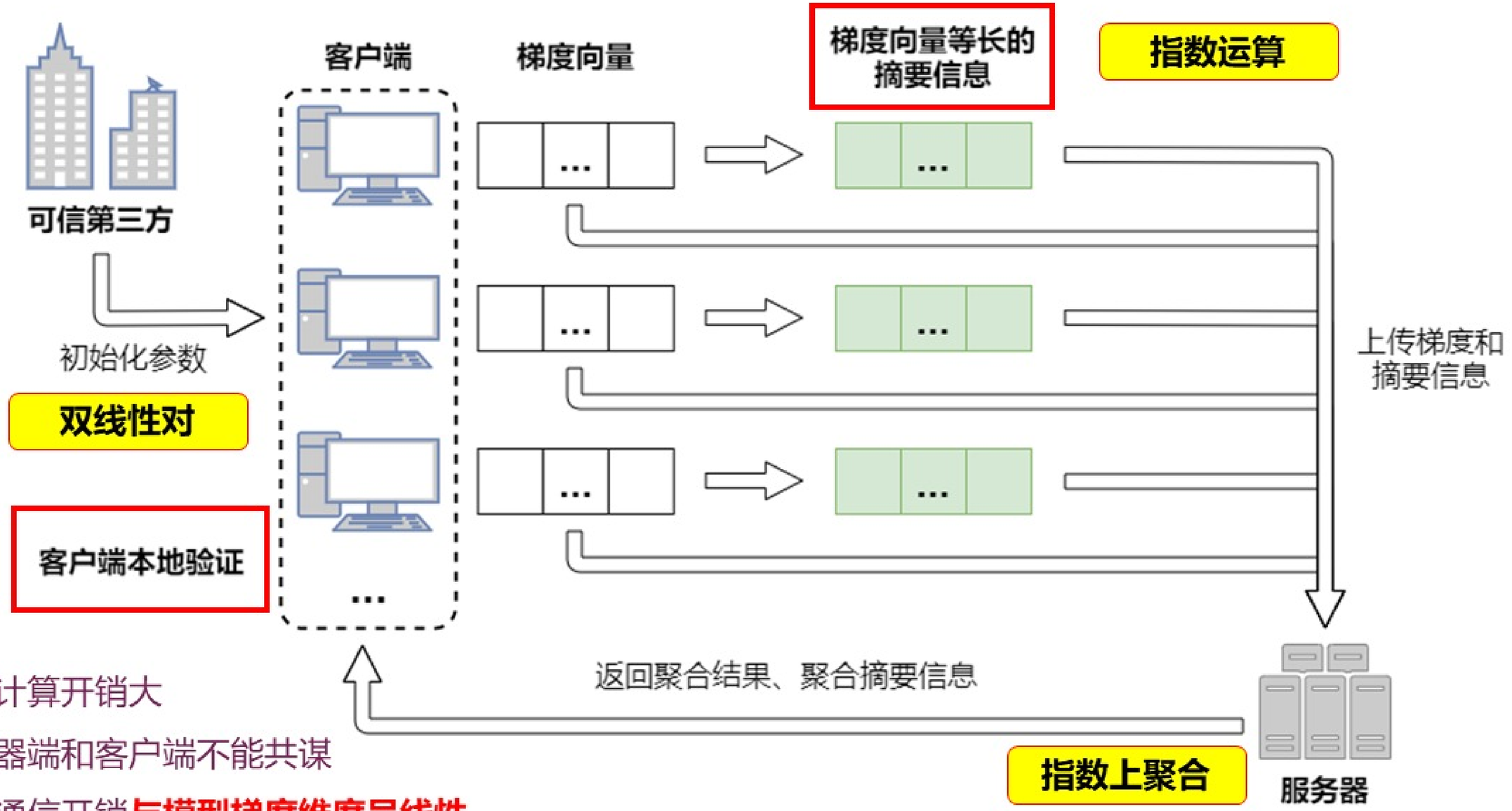
Jiqiang Gao, Baoyu Hou, Xiaojie Guo, Zheli Liu*, et al. Secure Aggregation is Insecure: Category Inference Attack on Federated Learning. **IEEE TDSC 2021**

汇报目录


本地节点
云端服务器
本地模型
反馈优化
训练
上传模型参数
用户A
用户数据
参数聚合
全局模型
篡改参数?
本地模型
上传模型参数
反馈优化
训练
用户N
用户数据
可验证
VerifyNet
TIFS 2020

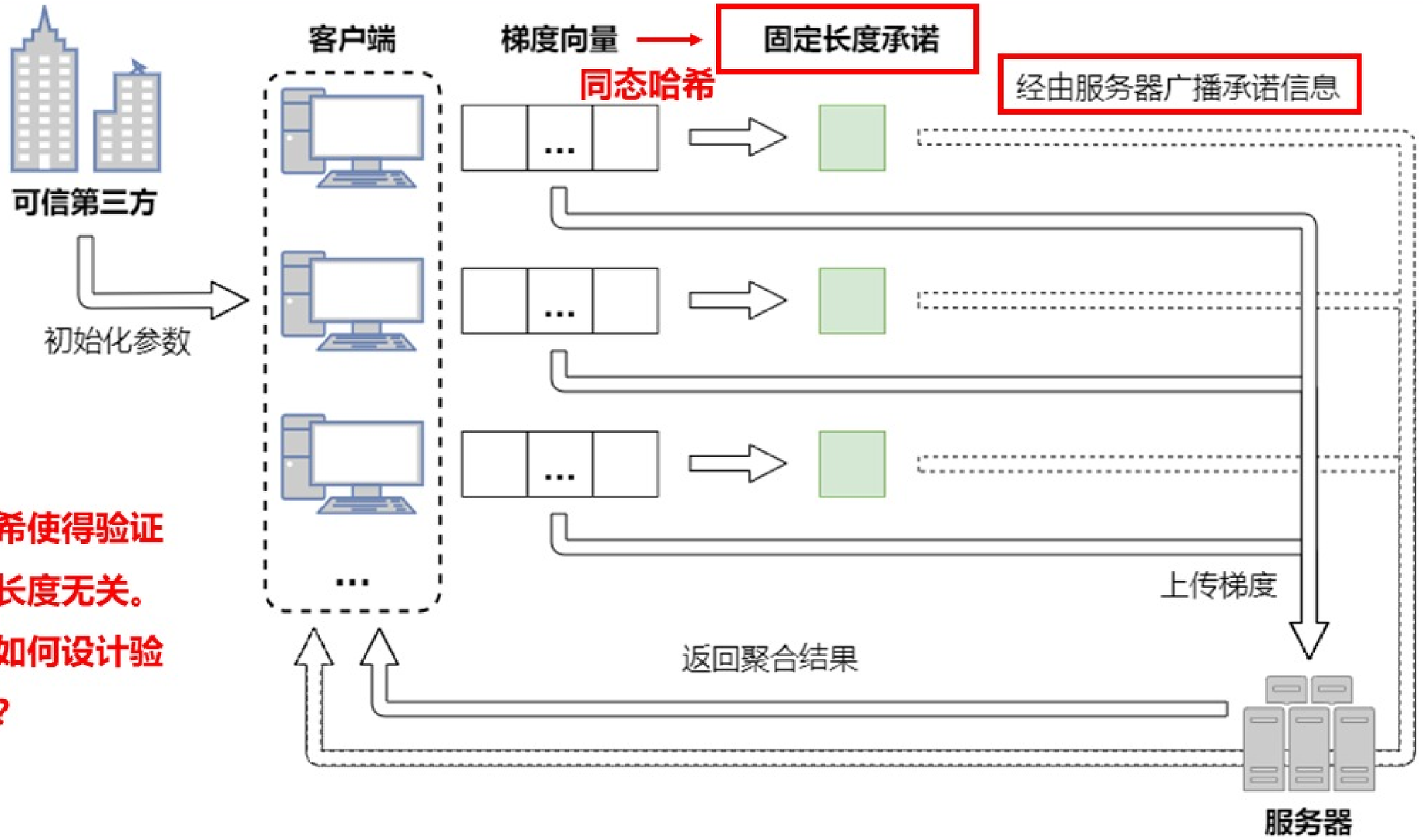
- 还有其它安全风险吗?

可验证安全聚合：VerifyNet的构造



- 验证计算开销大
- 服务器端和客户端不能共谋
- 验证通信开销**与模型梯度维度呈线性**

可验证安全聚合：VerifyNet的构造



利用哈希使得验证与向量长度无关。

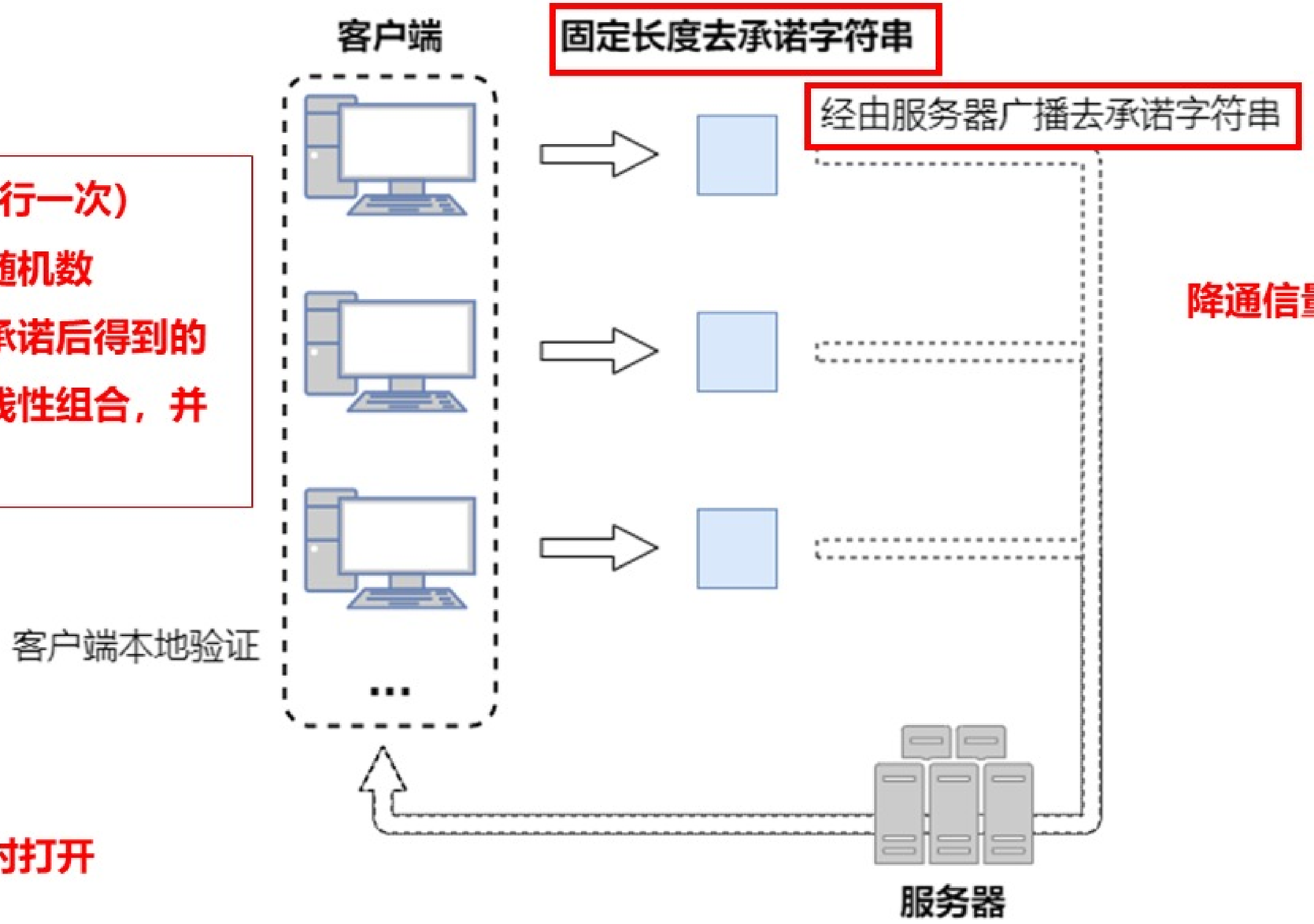
挑战：如何设计验证方案?

Xiaojie Guo , Zheli Liu , Jin Li , Jiqiang Gao, et al. VeriFL: Communication-Efficient and Fast Verifiable Aggregation for Federated Learning. **IEEE TIFS 2020**

可验证安全聚合：VerifyNet的构造

验证阶段（可多轮进行一次）

1. **客户端本地选择随机数**
2. **使用随机数对去承诺后得到的同态哈希值进行线性组合，并验证。**

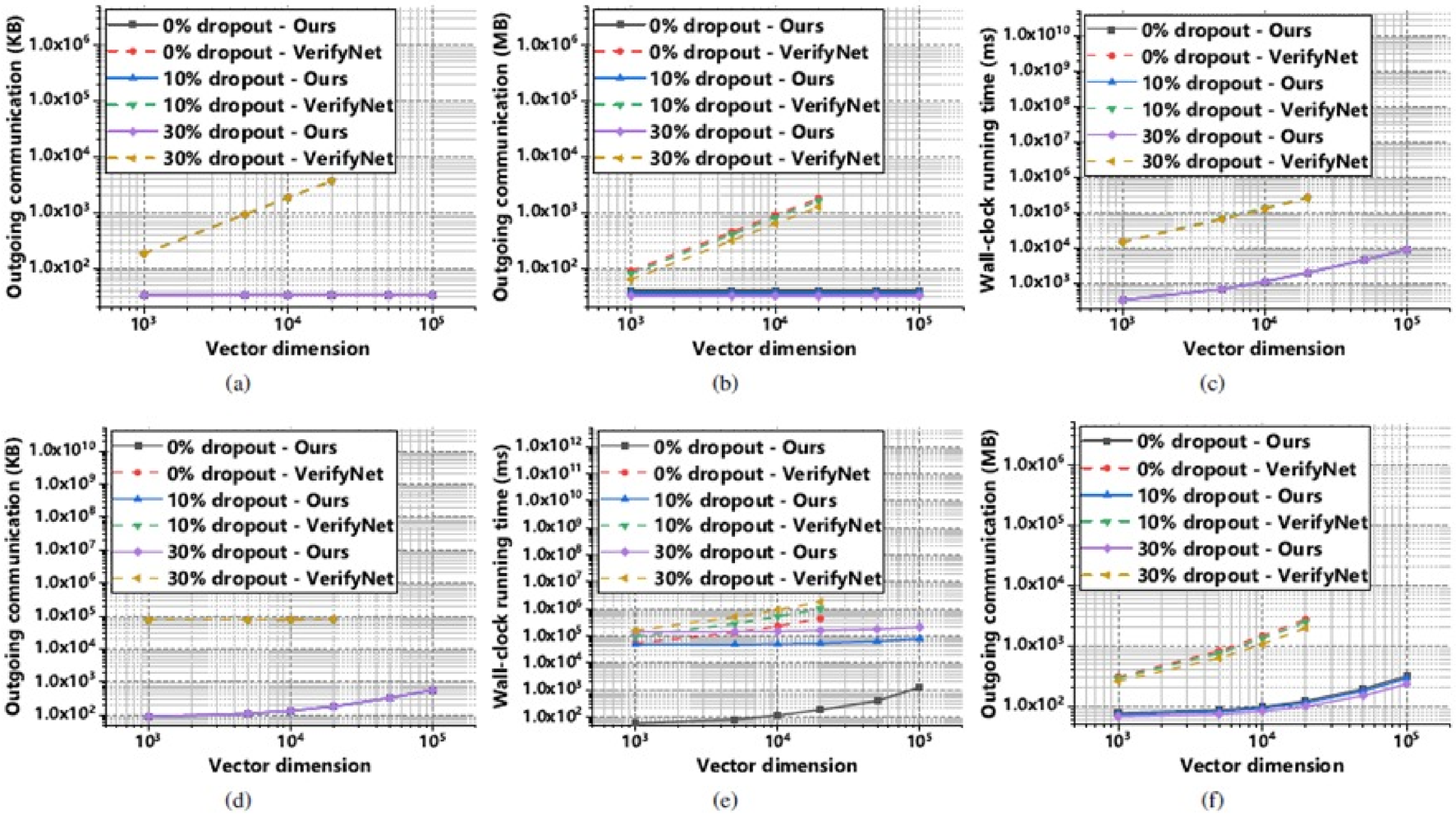


降通信量：服务器广播

验证：承诺-需要时打开

Xiaojie Guo , Zheli Liu , Jin Li , Jiqiang Gao, et al. VeriFL: Communication-Efficient and Fast Verifiable Aggregation for Federated Learning. **IEEE TIFS 2020**

可验证安全聚合：性能对比



与VerifyNet的性能对比

Xiaojie Guo , Zheli Liu , Jin Li , Jiqiang Gao, et al. VeriFL: Communication-Efficient and Fast Verifiable Aggregation for Federated Learning. **IEEE TIFS 2020**

谢 谢