计 算 机 研 究 与 发 展
Journal of Computer Research and Development

https://crad.ict.ac.cn
62(9): 2259−2282, 2025

本期推荐

数据治理中的隐私审计

许婧楠　王雷霞　孟小峰
（中国人民大学信息学院　北京　100872）
（jnxu@ruc.edu.cn）

Privacy Auditing in Data Governance

Xu Jingnan, Wang Leixia, and Meng Xiaofeng
（*School of Information*, *Renmin University of China*, *Beijing* 100872）

Abstract　Privacy auditing is a crucial issue of data governance, aiming to detect whether data privacy has been protected effectively. Typically, scholars would protect personal private data to meet differential privacy guarantees by perturbing data or adding noise to them. Especially in scenarios of machine learning, an increasing number of differential privacy algorithms have emerged, claiming a relatively stringent level of privacy protection. Although rigorous mathematical proofs of privacy have been conducted before the algorithms' release, the actual effect on privacy in practice is hardly assured. Due to the complexity of the theory of differential privacy, the correctness of their proofs may not have been thoroughly examined, and imperceptible errors may occur during programming. All of these can undermine the extent of privacy protection to the claimed degree, leaking additional privacy. To tackle this issue, privacy auditing for differential privacy algorithms has emerged. This technique aims to obtain the actual degree of privacy-preserving of differential privacy algorithms, facilitating the discovery of mistakes and improving existing differential privacy algorithms. We survey the scenarios and methods of privacy auditing, summarize the methods from three aspects—data construction, data measurement, and result quantification, and evaluate them through experiments. Finally, we present the challenges of privacy auditing and its future direction.

Key words　data governance; privacy preserving; machine learning; privacy auditing; differential privacy

摘　要　隐私审计是数据治理中的关键问题，旨在判断数据的隐私是否得到了有效保护.通常，学者们通过对数据添加噪声或扰动实现差分隐私，从而保护个人隐私.特别在机器学习场景下，出现越来越多的差分隐私算法，并且这些算法均声称自己可以达到较为严格的隐私保护水平.然而，即使这些算法在发布之前会经过严格的数学证明，其实际应用中的隐私保护程度亦难以确定.鉴于差分隐私理论本身的复杂性，隐私算法中证明的错误和编程实现的错误时有发生，使得这些算法无法达到其声称的隐私保护水平，导致隐私泄露.为了解决这一问题，隐私审计应运而生.隐私审计可以获取隐私算法的真实隐私保护水平，有助于算法设计者对算法进行改进.将综述隐私审计相关算法，从数据构建、数据测算、结果量化3个维度进行总结，并对隐私审计算法进行实验说明，最终提出隐私审计面临的挑战以及未来研究方向.

关键词　数据治理；隐私保护；机器学习；隐私审计；差分隐私

中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202440530　　**CSTR:** 32373.14.issn1000-1239.202440530

收稿日期：2024-06-17；**修回日期**：2025-02-11

基金项目：国家自然科学基金项目（62172423）
This work was supported by the National Natural Science Foundation of China (62172423).

通信作者：孟小峰（xfmeng@ruc.edu.cn）

数据治理是当下炙手可热的话题之一,旨在实现对数据的质量、规范、制度等方面的优化.虽然已有诸多规章制度对数据的使用、传播等进行了约束,但这些规章制度通常要求数据所有者自证清明,并且没有对数据治理的对象、所需使用的技术及手段作出明确规定.通过对数据治理以及数据应用场景的分析可知,数据治理的核心问题为数据审计,其中包含公平审计、遗忘审计、隐私审计(privacy auditing).公平审计是指相似数据信息是否可以获得相似决策,可通过领域专家对决策结果的评判度量;遗忘审计是指数据模型对用户个人隐私数据的遗忘程度,可通过对比重训练模型对数据的记忆程度度量;隐私审计是指隐私技术对个人数据的保护程度,但隐私审计缺乏严格的定义以及统一度量标准,难以判断数据中隐私信息的保护程度.因此,本文重点对隐私审计进行综述.在广泛应用的机器学习场景下,差分隐私、同态加密和零知识证明等为常见的隐私计算技术[1],同态加密和零知识证明主要关注数据安全,而差分隐私侧重于数据隐私,是当前主流的数据隐私保护技术,因此差分隐私也是本文重点考虑的对象.

为实现对个人隐私数据的保护,学术界涌现了大量以差分隐私技术[2-4]为基础的隐私保护算法.差分隐私技术在保护了个人隐私的同时保证了数据的可用性,通过对数据加噪的方式,有效防止了对单个数据参与者信息的泄露,将删除或插入任何一条数据对查询结果产生的影响控制在了一定范围之内.

差分隐私(differential privacy, DP)常用在医疗领域以保护用户个人隐私.在该领域,不同的医疗机构或研究机构间可能会通过共享病例数据的方式进行医学研究,这些医疗数据中可能包含患者的病例信息,例如年龄、性别、疾病等.若将这些医疗数据直接进行发布,攻击者可以通过多次攻击推断出其所需要的信息.假设攻击者想要获取患者 Alice 是否患有某种特殊疾病,攻击者首先通过查询获得某种疾病的患者数量的统计信息,根据性别和年龄分别进行分组查询获得结果,并结合其自身拥有的背景知识,例如患者就诊时间等,极有可能推断出患者 Alice 是否患有特殊疾病.差分隐私通过在当前数据或查询结果中加入适量噪声,实现对隐私信息的保护.因为噪声的引入是随机的,攻击者每次查询所得的查询结果的变化也是随机的,所以攻击者无法确定当前结果中所含噪声大小,也就无法确定其所得结果的真实性.

差分隐私技术在工业界的应用越来越广泛.例如,苹果公司利用差分隐私技术对用户设备上的信息添加噪声后再收集数据,从而确保对单个用户具体信息的保护;小米公司在其操作系统中加入差分隐私技术,使定位软件只能获得用户的模糊定位信息.虽然每年有上千个差分隐私算法发布,但这些算法在实际应用中的隐私保护效果仍有待验证.例如,Backpropagation Clipping 算法[5]声称可以达到$(0.21, 10^{-5})$-差分隐私,但 Tramer 等人[6]发现,由于样本梯度计算的错误,该算法并未达到其声称的隐私保护程度.若这种算法被应用于工业界,用户的个人隐私将无法得到有效的保护.因此,检验隐私算法是否达到所声称的隐私保护程度具有很大的现实意义.

检验隐私算法目前主要可以分为隐私审计和隐私验证(privacy verification)两类技术.隐私审计关注于隐私算法的真实隐私保护程度,即经验隐私预算下界;隐私算法验证关注于隐私算法在部署或执行时的算法正确性.当前关于差分隐私的隐私算法验证已有诸多研究与综述[7-8],这类方法利用静态验证方式(static methods)[9-15]、隐私威胁检测(detecting violation)[16-17]和编程平台验证[18-20]三种方式对简单差分隐私算法进行验证.当前机器学习场景应用广泛,利用差分隐私实现隐私保护更为常见,但由于机器学习模型与隐私算法的复杂性,其审计也极具挑战性,并且隐私审计刚刚起步,目前尚无系统的总结.

本文重点针对差分隐私算法的审计进行综述,这些方法主要针对于复杂的机器学习下的差分隐私算法,例如 DP-SGD[21]算法、联邦学习下的差分隐私算法 DP-FedAvg[22]等.

当前机器学习下的差分隐私算法存在 3 个问题:1)一部分算法设计者证明隐私理论上界的能力不足,导致其发布算法时声称的理论值过于宽松,对于这种情况,利用审计可发现该上界仍有较大改进空间.2)一部分算法在证明或代码实现过程中出现了错误,导致算法真实的隐私保护程度远低于其声称的隐私保护程度.对于这种情况,利用较弱攻击者审计即可发现经验隐私预算下界大于理论隐私预算上界,可根据该现象对算法进行修改.3)随着组合定理的逐渐改进,已经发布的隐私算法的理论值上界可能会因此变得更严格,但这一现象是否会持续下去,可利用较强攻击者对算法审计并进行判断.

本文总结差分隐私算法的隐私审计的最新研究进展与研究方向.将现有的隐私审计算法划分为三大类别并分别进行介绍.最后,针对隐私审计的特性,提出了隐私审计的未来研究方向并加以具体分析.

1 差分隐私基础

本节将简单概述隐私审计中主要涉及的4种定义及算法：中心化差分隐私定义、本地化差分隐私定义、面向中心化机器学习的差分隐私算法及面向联邦学习的差分隐私算法.

1.1 中心化差分隐私与本地化差分隐私

差分隐私(DP)是Dwork[2]提出的一种隐私保护方法，通过对原始数据加入噪声进行扰动，使得攻击者无法根据所得信息推断出某条数据是否参与了训练. 差分隐私定义如下.

定义1[2]. (ε,δ)-差分隐私. 给定随机化算法$\mathcal{A}$，对于任意相邻数据集$D\simeq D'$，以及算法的输出子集$O\in Range(\mathcal{A})$，满足不等式(1)，则算法$\mathcal{A}$满足(ε,δ)-差分隐私，即(ε,δ)-DP. 其中D和D'之间仅相差1条数据，ε为隐私预算.

$$Pr(\mathcal{A}(D)\in O)\leqslant e^{\varepsilon}Pr(\mathcal{A}(D')\in O)+\delta, \tag{1}$$

ε越小，加噪越多，隐私算法在相邻数据集上的输出分布越接近，隐私保护程度越高. δ为松弛项，表示相邻数据集D和D'对应输出分布的比值超出隐私预算ε约束的概率，δ越小，隐私保护得越好. 当$\delta=0$时，隐私算法满足ε-差分隐私，即纯差分隐私定义. 若$\delta\neq 0$，则为松弛差分隐私，表示当前隐私算法以$1-\delta$的概率满足ε-差分隐私，即(ε,δ)-差分隐私.

(ε,δ)-差分隐私为隐私审计中常用的差分隐私定义，也被看作中心化差分隐私定义(简称CDP). 在中心化差分隐私中，用户的数据会传给可信的数据收集者，数据收集者对用户数据进行扰动，并将扰动后的数据进行发布.

本地化差分隐私(简称LDP)与中心化差分隐私定义略有不同. 在本地化差分隐私中，假设每个用户只有1条数据，并且数据收集者是不可信的. 用户在本地对数据进行扰动，将扰动后的数据发送给数据收集者，数据收集者对收集到的数据进行校正、统计并发布. 两者的核心区别在于是否包含可信第三方以及相邻数据集的定义的不同. 中心化差分隐私依赖于可信第三方. 用户数据收集完成后，可信第三方统一对用户数据进行处理，相邻数据集定义在整体的数据集上. 本地化差分隐私中舍弃了可信第三方的假设，每个用户独立地对其隐私数据进行扰动，且相邻数据集定义在用户数据上. 中心化差分隐私与本地化差分隐私的处理框架对比如图1所示，其中图1(a)为中心化差分隐私处理框架，图1(b)为本地化差分隐私处理框架.

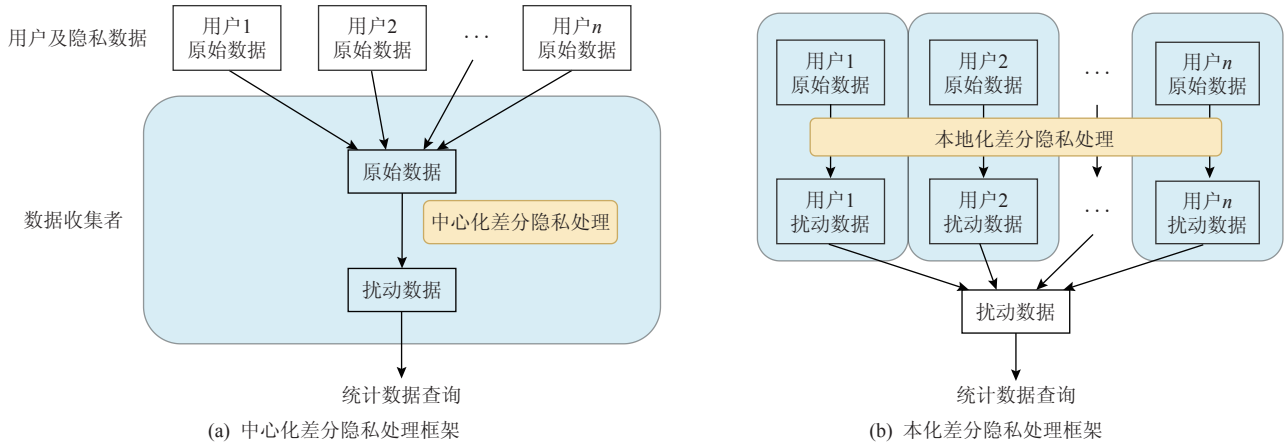
本地化差分隐私的形式化定义如下：

定义2[23]. 给定n个用户，每个用户只包含1条记录，对于$\varepsilon\geqslant 0$，若随机化算法$\mathcal{A}$满足ε-LDP，当且仅当对于任意2个输入数据x_1,x_2输出y，$y\in O$，O为算法$\mathcal{A}$的输出域：

$$Pr(\mathcal{A}(x_1)=y)\leqslant e^{\varepsilon}Pr(\mathcal{A}(x_2)=y). \tag{2}$$

1.2 面向中心化机器学习的差分隐私算法

训练机器学习模型需要大量数据以及多次迭代. 通过优化器对模型参数进行调整，最小化损失函数，以获得高精度的模型. 其中随机梯度下降法(简称SGD)为机器学习中常用的优化算法之一，其核心思想为通过迭代的方式，在训练过程中沿着函数负梯度的方向逐步更新模型参数，以最小化损失函数. 机器学习模型在训练过程中有可能会记住一些训练数据，如果攻击者攻击了某个利用隐私数据训练的模型，则有可能造成隐私的泄露. 因此研究者们通常将差分隐私算法与机器学习相结合，实现对隐私数据的保护.



(a) 中心化差分隐私处理框架 (b) 本化差分隐私处理框架

Fig. 1 Comparison of frameworks for different differential privacy data processing

图1 不同差分隐私数据处理框架对比

面向中心化机器学习的差分隐私算法有对模型输入添加扰动、对目标函数添加扰动、对模型梯度添加扰动、对模型输出添加扰动4种实现方式. 其中最常用的方法DP-SGD算法为对模型梯度添加扰动的方式, 如算法1所示.

算法 1. 基于差分隐私的随机梯度下降法(DP-SGD).

输入: 数据集大小 N, 总迭代次数 T, 采样概率 q, 梯度裁剪参数 C, 高斯噪声参数 σ;

输出: 加噪模型参数 $\boldsymbol{\theta}^T$.

① $\boldsymbol{\theta}_0 \leftarrow$随机初始化模型参数;

② for $t=0$ to $T-1$ do

③ $C^t \leftarrow$(按均匀分布对数据进行采样); /*数据抽样*/

④ foreach 数据记录 $b_i \in B^t$ do

⑤ $\boldsymbol{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\boldsymbol{\theta}_t, x_i)$; /*计算梯度*/

⑥ $\bar{\boldsymbol{g}}_t(x_i) \leftarrow \boldsymbol{g}_t(x_i)/\max\left(1, \frac{\|\boldsymbol{g}_t(x_i)\|_2}{C}\right)$; /*梯度裁剪*/

⑦ $\tilde{\boldsymbol{g}}_t \leftarrow \frac{\sum_t \boldsymbol{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)}{L}$; /*加噪*/

⑧ $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t \tilde{\boldsymbol{g}}_t$;

⑨ end foreach

⑩ end for

⑪ return $\boldsymbol{\theta}^T$.

DP-SGD与SGD相比加入了梯度裁剪(clipping)和梯度加噪2个操作. 在模型每一轮的梯度更新之后, 首先对梯度进行裁剪, 实现对梯度更新值敏感度的约束, 然后对裁剪后的梯度加入高斯噪声, 使模型训练的整体过程满足 (ε,δ)-DP. 由于所加入的噪声是随机的, 每一次的梯度计算都会有略微的变化, 因此可以有效地防止攻击者根据模型的输出推断单条数据的信息. 梯度裁剪不仅可以设置敏感度, 并且会改变梯度的取值范围, 因此也可以在一定程度上隐藏当前梯度的真实值.

1.3 面向联邦学习的差分隐私算法

分布式机器学习中, 中心端的机器学习算法受限于服务器本身收集的数据集. 为了充分利用用户本地的隐私数据参与训练, 谷歌公司在2016年提出了联邦学习的模型训练框架. 该框架中参与模型训练的有两方: 服务器端(server)和若干个客户端(client). 与先收集数据后训练的中心模型相比, 该框架坚持“数据不出本地”的原则, 即各客户端之间不直接传输原始数据, 共同训练一个全局模型, 实现了对隐私信息的保护. 该框架中各客户端利用从服务器端下载的模型参数进行本地训练, 将训练后的模型参数上传至服务器端, 服务器端根据各个客户端传回的参数对全局模型进行聚合并更新后再将全局模型参数分发, 该过程重复多次直至全局模型收敛. 其框架图如图2所示. 在这一框架下, 谷歌公司提出了联邦平均算法(FedAvg)[24]. FedAvg通过2步抽样操作实现了对隐私信息的保护, 其中包括服务器端对当前迭代中参与训练的客户端的抽样以及客户端本地训练中对训练数据的抽样. 该算法的基本流程如算法2所示.

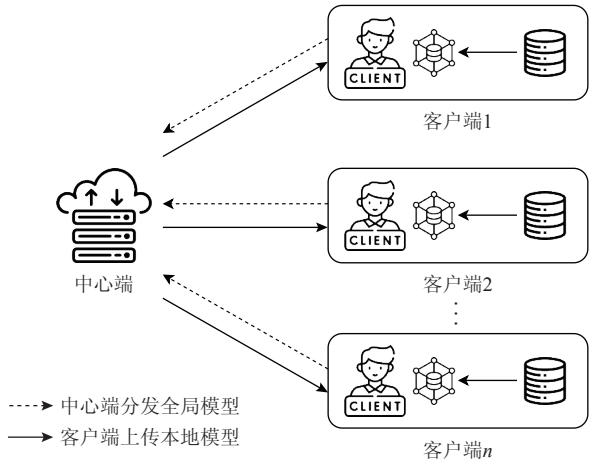


Fig. 2 Framework of federated learning

图2 联邦学习框架

算法 2. FedAvg.

输入: M 个客户端集合 C, 全局训练轮数 T, 迭代次数 E, 客户端学习率 η;

输出: 服务器模型参数 $\boldsymbol{\theta}^T$.

服务器端函数:

① $\boldsymbol{\theta}_0 \leftarrow$随机初始化模型参数

② for $t=0$ to $T-1$ do

③ $C^t \leftarrow$(从 M 个客户端中选取部分客户端参与训练); /*客户端抽样*/

④ foreach $C_m \in C^t$ do

⑤ $\Delta v \leftarrow ClientUpdate(\boldsymbol{\theta}^t, E)$;

end foreach

⑥ $\Delta\bar{\boldsymbol{\theta}}^{t+1} \leftarrow \frac{\sum_{C_m \in C^t} \Delta\boldsymbol{\theta}_m^t}{|C^t|}$; /*参数聚合*/

⑦ $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \Delta\bar{\boldsymbol{\theta}}^{t+1}$;

⑧ end for

⑨ 将 $\boldsymbol{\theta}^T$ 发送至客户端.

客户端函数:

① $\boldsymbol{\theta}^{t,0} \leftarrow \boldsymbol{\theta}^t$;

② for $e=0$ to $E-1$ do /*训练数据抽样*/

③ $B^j \leftarrow$(从本地数据集中随机采样训练数据);

④ $\boldsymbol{g}^{t,j} \leftarrow \frac{1}{|B^j|}\sum_{b\in B^j}\nabla\ell\left(\boldsymbol{\theta}^{t,j},b\right)$;

⑤ $\boldsymbol{\theta}^{e,j+1} \leftarrow \boldsymbol{\theta}^{e,j}-\eta\boldsymbol{g}^{t,j}$;

⑥ end for

⑦ $\Delta\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{t,0}-\boldsymbol{\theta}^{t,E}$;

⑧ 将$\Delta\boldsymbol{\theta}$发送至服务器端.

联邦学习框架虽然减小了原始数据被直接泄露的风险，但在模型训练过程中模型参数会记忆隐私数据，攻击者们可以利用其背景知识以及模型相关信息进行推理，导致隐私数据的泄露. 因此学者们将联邦学习与差分隐私等隐私保护技术结合，加强了对隐私数据的保护.

DP-FedAvg[22] 和 DP-FedSGD[22] 是联邦学习中常用的差分隐私算法，通过对 FedAvg 中的处理过程加入梯度裁剪和梯度加噪 2 个操作来实现隐私保护. 梯度裁剪执行在参与训练的客户端. 各客户端在本地利用数据训练好模型后，将其梯度进行裁剪再发给服务器端. 梯度加噪执行在服务器端. 服务器端利用加权求和的方式聚合各客户端传回的参数并对聚合后的参数进行加噪. DP-FedAvg 和 DP-FedSGD 的区别在于各客户端的模型训练方式以及传回的参数. DP-FedAvg 中客户端的训练方式是小批量梯度下降，各客户端将本地模型进行更新后再将参数传给服务器端，DP-FedSGD 中客户端的训练方式是批梯度下降，各客户端计算后的梯度不对本地模型进行更新，而是直接将梯度传给服务器端.

2 隐私审计定义与框架

差分隐私算法层出不穷，但其实际的隐私保护程度仍有待探究. 隐私审计方法可在一定程度上检验机器学习下差分隐私算法的真实隐私保护程度. 本节将给出隐私审计相关定义，介绍隐私审计框架，结合图 3 介绍隐私审计通用流程.

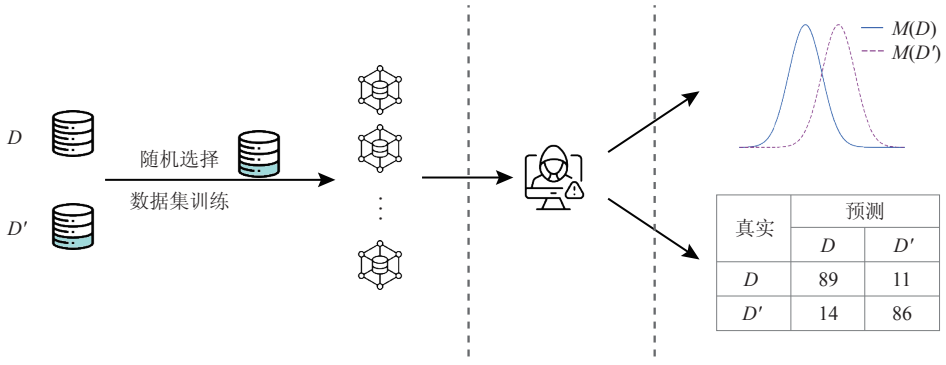


Fig. 3 General workflow of privacy auditing

图 3 隐私审计通用流程

2.1 隐私审计定义

差分隐私算法中的重要参数为隐私预算ε，用于表示当前隐私算法的隐私保护程度. 这一参数是由算法设计者利用数学证明等方式求得，考虑的是隐私算法在最差攻击者情况下的隐私保护程度，可称为理论隐私预算上界. 隐私审计关注于隐私算法的实际隐私保护程度，利用更接近现实情况的攻击者对隐私算法进行攻击并获取对应的隐私泄露程度，可称为经验隐私预算下界ε'.

隐私审计旨在获取隐私算法的经验隐私预算下界，其定义如下.

定义 3. 隐私审计. 已知满足(ε,δ)-DP 的随机算法M，其理论隐私预算上界为ε. 利用攻击者对其进行攻击并获取对应的经验隐私预算下界ε'，将ε和ε'进行对比，判断算法M的正误及其理论隐私预算上界的松弛.

2.2 隐私审计框架

隐私审计框架中包含审计场景、审计方法及审计目标. 本节首先介绍审计场景及审计目标，概括隐私审计通用流程，最后介绍审计方法.

2.2.1 审计场景

现有的隐私审计场景可划分为 3 个：中心化机器学习场景、分布式联邦学习场景及分布式数据收集场景.

中心化机器学习（centralized machine learning, CML）场景的差分隐私算法审计通常为对机器学习下差分

隐私算法 DP-SGD 的审计. 神经网络模型中网络层数及模型参数较多, 隐私预算的推导十分复杂并且很难计算严格的隐私上界, 只能假设攻击者具有强大的背景知识并利用组合定理计算得出理论上界. 但在实际情况中隐私算法并不会将隐私预算完全消耗, 因此可以利用隐私审计判断当前隐私算法的实际隐私消耗, 即经验隐私预算下界, 观察理论隐私预算上界与经验隐私预算下界之间的关系.

分布式联邦学习(distributed federated learning, DFL)场景的差分隐私算法审计涉及 DP-FedAvg 和 DP-FedSGD 这 2 个算法. 联邦学习模型中包含大量神经网络模型的训练以及参数的传输, 并且不同客户端的数据分布可能不同, 因此在联邦学习场景中严格隐私上界的计算也十分困难. 隐私审计可以观察在某一轮训练或联邦学习模型整体训练完的实际隐私消耗, 以及当前算法给出的理论隐私预算上界是否严格.

分布式数据收集(distributed data collection, DDC)场景的差分隐私算法审计涉及 LDP-SGD 算法和 8 种可用于频数估计的本地化差分隐私算法, 例如 OLH[25] 等. 根据本地化差分隐私定义与中心化差分隐私定义在算法输入上的不同, 分布式数据收集场景的差分隐私审计无需构建 2 个相邻数据集, 只需构建 2 个数据点. 但仍需训练大量模型以获得算法在 2 个数据点上的输出分布. 对该场景下的隐私算法审计可以观察可变参数对算法的真实隐私保护程度的影响.

2.2.2 审计目标

隐私审计的目标可以分为 2 个: 判断差分隐私算法的正误和检验差分隐私算法的松弛, 如图 4 所示.

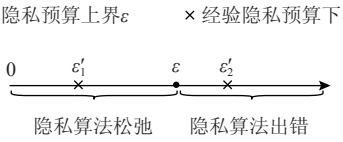


Fig. 4 Goal of privacy auditing

图 4 隐私审计目标

对于判断差分隐私算法的正误, 目前很多机器学习方法都未将代码开源, 在差分隐私领域也存在相同问题. 对于这些未开源代码的算法, 使用者并不知道其真实的隐私保护能力, 只能根据发布者自己理论证明得到的理论隐私预算上界作为算法真实隐私保护能力. 一旦这一复杂的理论证明存在失误, 用户的隐私信息将不能得到保护. 审计差分隐私算法的正误有助于解决该困境. 隐私审计中最常用的攻击为成员推理[26]. 在黑盒情况下, 该攻击者具有的背景知识与其他攻击者相比较少, 仅能获取模型最终的分类预测结果或某一数据点的损失函数值. 若在该种情况下所获得的经验隐私预算下界高于隐私算法的理论隐私预算上界, 则说明当前隐私算法无法防御这一攻击, 同样无法防御其他攻击, 因此该算法的设计中存在缺陷. 如图 4 右侧隐私算法出错部分所示, 审计所得的经验隐私下界 ε'_2 大于理论隐私预算上界 ε. 因为当前攻击者远远弱于算法设计时所考虑的最坏情况对应的攻击者, 在这种情况下隐私泄露的情况应远小于最坏情况, 即实际消耗的隐私预算应小于理论隐私预算上界.

对于检验差分隐私算法的松弛, 隐私算法在设计的时候往往假设攻击者拥有强大的背景知识, 但在实际情况中, 攻击者的背景知识和攻击能力往往是有限的. 例如 DP-SGD 算法在设计中假设攻击者可以获取所有中间模型, 但大多数情况下攻击者只能获取最终模型. 因此实际消耗的隐私预算会远小于算法设计者假设的情况. 利用不同强度的攻击者进行审计可以获得对应的隐私预算消耗, 进而判断差分隐私算法中隐私上界的松弛程度, 从而促进算法设计者对差分隐私算法的改进. 若在当前攻击情况下所获得的经验隐私预算下界小于理论隐私预算上界, 且两者十分接近, 则说明当前隐私算法的上界足够严格, 如图 4 中左侧隐私算法松弛部分所示, 审计所得的经验隐私下界 ε'_1 小于理论隐私预算上界 ε. 若经验隐私预算下界小于理论隐私预算上界且两者差值较大, 则说明当前隐私算法的上界还可以进行优化. 与隐私算法设计过程中仅从理论证明的角度相比, 隐私审计以攻击者视角利用统计方法获取经验隐私预算下界, 是从现实情况的角度度量隐私算法的真实隐私保护程度的方法.

2.2.3 隐私审计通用流程

隐私审计主要依赖于隐私攻击以及统计计算, 其中攻击者通常由数据构建者和辨别者 2 个角色组成. 整体流程可以划分为 3 步: 1)数据构建者构建相邻数据集; 2)利用相邻数据集训练模型, 辨别者观察差异点参与训练的情况; 3)对观察结果进行统计, 并将统计结果转换成概率或分布, 计算获得经验隐私预算下界. 该过程如图 3 所示.

步骤 1 中, 通常需要构建一对相邻数据集. 因为差分隐私算法设计之时需要考虑最坏情况下的攻击者, 对相邻数据集并没有特殊要求. 而隐私审计的目的是获取隐私算法在接近于真实攻击者情况下的隐私泄露程度. 因此可以构建一对具有最大差距的相

邻数据集使审计所得的经验隐私预算下界更严格.

步骤 2 中, 隐私审计最常用的辨别者的目的为判断某一数据点是否参与了当前模型的训练. 对于已经构建好的相邻数据集, 辨别者通常关注于相邻数据集之间相差的数据点的损失函数值或分类预测结果. 关注损失函数值是因为在机器学习模型中, 若 1 个数据点参与了模型的训练, 其损失函数值将明显小于其没有参与模型训练的损失函数值. 关注分类预测结果的辨别者可以根据大量模型对数据点的预测结果获取真阳性率(true positive rate, *TPR*)和假阳性率(false positive rate, *FPR*). 由差分隐私的形式化定义可知, 差分隐私算法保证的是算法在相邻数据集上的输出的概率分布的不可区分. 假设当前输出集合为S, *TPR* 对应于攻击者正确预测了来自数据集D(或D')的数据点的输出包含在集合S中, *FPR* 对应于攻击者错误预测了来自数据集D'(或D)的数据点输出包含在集合S中. 因此可以利用 *TPR* 和 *FPR* 的比值获取经验隐私预算下界.

步骤 3 中, 将辨别者的攻击结果转换成概率或分布. 转换为概率即为转换成 *TPR* 和 *FPR* 的形式, 利用置信区间进行校正后, 可直接计算 2 个概率的比值获取经验隐私预算下界. 转换为分布的方式为: 若隐私算法训练了大量模型, 即使攻击者仅关注于某一个点的损失函数值, 也会获取与模型数量对应个数的损失函数值. 辨别者获取该点在利用 2 个相邻数据集训练所得的模型上的损失函数值, 通过正则化等方式将其转换成 2 个分布, 将审计问题转换为区分 2 个概率分布的问题, 与差分隐私形式化定义相同.

2.2.4 隐私审计方法分类

基于隐私审计通用流程, 考虑不同的审计场景与目标, 研究者们提出了隐私审计的数据构建方法、隐私审计的数据测算方法、隐私审计的结果量化 3 类方法. 这 3 类方法并不是完全相互独立的, 它们整体上都遵循隐私审计通用流程, 但每个类别的侧重点不同.

隐私审计的数据构建方法侧重于隐私审计通用流程中的步骤 1, 通过构建可以最大化相邻数据集之间的差异点(也可称为告密点(canary)或投毒点), 使当前隐私算法在相邻数据集上的输出分布相差最大. 数据构建的方式通常采用投毒攻击. 该类别方法可分为针对差分隐私的特殊数据点构建和针对本地化差分隐私的差异数据项构建. 差分隐私的特殊数据点构建关注于差分隐私算法, 构建的是相邻数据集之间的差异点; 本地化差分隐私的差异数据项构建关注于本地化差分隐私算法, 构建的是 2 个不同的输入数据项. 这类方法可应用于 3 个隐私审计场景, 且通常与另外 2 类审计方法相结合以提高审计所得经验隐私预算下界的严格程度.

隐私审计的数据测算方法侧重于隐私审计通用流程中的步骤 2, 观察差异点在所有模型上的训练情况. 该类别方法可划分为单点数据测算和多点数据测算. 单点数据测算中, 需利用相邻数据集训练成千上万个模型, 辨别者观察差异点在这些模型上的训练情况以判断其是否参与了训练, 辨别标准通常利用损失函数值或分类预测结果. 由于隐私审计需要利用统计结果获取经验隐私预算下界, 样本数量越大, 最终所获得的下界越准确, 因此单点数据测算方法需要训练大量模型, 但会导致审计代价较大. 多点数据测算方法只需训练单个模型, 辨别者关注数据集中若干个点在当前模型上的训练情况, 以获取较为准确的经验隐私预算下界. 虽然在该种情况下获取的经验隐私预算下界并不如训练单点数据测算方法所得的经验隐私预算下界准确, 但审计代价较小.

隐私审计的结果量化方法侧重于隐私审计通用流程中的步骤 3, 即如何对差异点的训练情况进行统计并利用统计结果计算经验隐私预算下界. 该类别方法可划分为基于差分隐私变体的量化和基于贝叶斯的量化 2 类. 基于差分隐私变体的量化利用较为严格的差分隐私定义对相对松弛的差分隐私算法审计, 包括 *f*-DP(functional-differential privacy)[27]、LiDP[28]、RKRDP[29]、Rényi 散度. 例如, 基于 *f*-DP 的量化需要将其隐私参数μ与差分隐私定义中的隐私参数ε进行转换. 基于贝叶斯的量化中包含基于差分可辨的量化和基于贝叶斯估计的量化. 其中基于差分可辨的量化需要将其重要参数与差分隐私定义中的理论隐私预算上界进行转换, 基于贝叶斯估计的量化需要将统计结果看作变量求积分计算.

3 隐私审计的数据构建方法

差分隐私通过隐私预算ε控制隐私算法在一对相邻数据集上的输出分布之间的不可区分性. 理论上, 差分隐私算法考虑的是最坏情况下的攻击者, 实际审计时, 考虑的是更接近于真实情况的攻击者, 例如构建多个具有不同能力的攻击者判断当前差分隐私算法的最大隐私泄露程度. 因此, 在审计中可以构建使算法输出差距最大的相邻数据集. 本节主要介绍隐私审计的数据构建方法. 其中 3.1 节介绍 5 种针

对差分隐私的特殊数据点构建方法, 3.2 节介绍 1 种针对本地化差分隐私的差异数据项数据构建方法.

3.1 针对差分隐私的特殊数据点构建

针对差分隐私的特殊数据点构建可以划分为构建单强度数据和构建多强度数据 2 类. 单强度数据构建的方法只需构建一个相邻数据集之间的差异点. 多强度数据构建的方法利用多种不同强度的攻击者对隐私算法进行审计, 其中不同攻击者构建的差异点不同.

3.1.1 单强度数据构建

隐私审计中常用的攻击者的攻击强度弱于定义隐私算法时所考虑的攻击者. 为最大化在这种情况下的隐私泄露程度, 可以通过构建一对具有较大差异的相邻数据集获取更严格的经验隐私预算下界.

投毒攻击是针对机器学习模型的一种攻击方式, 通过恶意更改训练数据来影响模型的预测结果. 隐私审计将投毒攻击用于相邻数据集的构建, 通过对某一数据点进行投毒, 最大化差分隐私算法在这 2 个数据集上输出分布的差异. 在投毒攻击中, 攻击者可以修改投毒点的数据梯度, 使投毒点的梯度与其他数据点梯度大小和方向之间具有明显差异, 提高攻击者攻击成功的概率. 但 DP-SGD 算法中的裁剪操作会降低这种攻击带来的影响. 因为梯度裁剪会将所有数据点的梯度都限制在一定范围内, 若投毒后的梯度值远大于裁剪的范围, 投毒的梯度也将会缩小为裁剪值, 降低了投毒带来的影响. 并且, 梯度裁剪会破坏投毒点的梯度值与投毒点个数之间的比例关系.

Jagielski 等人[30] 在中心化机器学习场景下, 对投毒攻击进行改进并提出了 ClipBKD 方法, 并基于此攻击对 DP-SGD 算法审计. ClipBKD 关注于投毒点梯度的不可区分性, 因此需使投毒点被投毒前后的梯度方差尽可能小, 以减小梯度裁剪对投毒梯度的影响. 除此之外, 还需将该数据点的标签替换为其预测概率最小的标签. 为增强投毒攻击的效果, 在审计过程中加入k个相同的投毒点. 该方法需要利用相邻数据集分别训练大量模型, 观察投毒点在所有模型上训练所得的损失函数值. 因此 ClipBKD 需要找到判断投毒点是否参与了训练的损失函数阈值. 对攻击结果进行统计后, 可获得投毒点参与了数据集D的模型训练概率p_0和数据集D'的模型训练概率p_1, 其中p_0和p_1分别对应差分隐私定义中的$P[M(D)\in S]$和$P[M(D')\in S]$. 为了减小统计不确定性, 需利用置信区间对p_0和p_1校正, 并利用校正后的概率计算比值获取经验隐私预算下界.

Lu 等人[31] 在中心化机器学习场景下提出 ML-Audit 方法, 该方法可对 3 种传统差分隐私机器学习算法以及 DP-SGD 算法审计, 算法流程图如图 5 所示. 3 种传统差分隐私学习算法包括逻辑回归(logistic regression, LR)算法[32-33]、高斯朴素贝叶斯(Gaussian Naive Bayes, NB)方法[34] 及随机森林(random forest, RF)方法[35]. ML-Audit 方法的核心思想为构建可以最大化在相邻数据集上训练所得的模型之间的参数距离投毒点, 以得到最大的经验隐私预算下界. ML-Audit 方法遵循隐私审计通用流程, 其中图 3 的步骤 1 需要扩展为构建相邻数据集, 并选取在当前数据集下后验概率较大的输出集合作为当前审计所需要的输出, 目的是利用这些输出最大差异化在 2 个相邻数据集上的输出分布. 图 3 的步骤 3 采用 Katz-log 置信区间统计攻击结果.

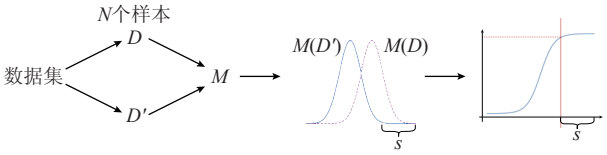


Fig. 5 Methodology flowchart of ML-Audit[31]

图 5 ML-Audit[31] 方法流程图

ML-Audit 方法构建相邻数据集是通过一种基于模型影响的(influence-based)投毒攻击, 根据算法在数据点上的预测结果或损失函数值等指标找到数据集中对所审计算法影响最大的点, 并对该点的数据及标签进行扰动, 以最大化模型参数在该数据点上扰动前后的距离. 但计算所审计算法中所有参数之间的距离并不现实, 因此 ML-Audit 选取了 4 个算法中的关键参数用于计算, 这些关键参数利用总结函数(summary function)表示, 如表 1 所示. 该方法关注于投毒点的分类结果. 与 ClipBKD 方法相同, 获得投毒点参与训练的概率后需要利用置信区间对其进行校正, 但该方法选用的置信区间为更严格的 Katz-log 置信区间. 在 Logistic Regression 算法中, 可以对算法的目标函数、梯度扰动或输出扰动实现隐私保护. ML-Audit 审计关注于对目标函数扰动[32] 和输出扰动的 Logistic Regression 算法[33]. 数据点对该模型参数的影响可以利用梯度和 Hessian 矩阵计算所得的损失函数值计算得到, 其计算形式为$\boldsymbol{I}_{\theta}(x^*)=\boldsymbol{H}_{\theta}^{-1}\nabla_{\theta}L(x^*,\widehat{\boldsymbol{\theta}})$, 因此对于 Logistic Regression 算法, 选用系数向量(coefficient vector)作为关键参数. 高斯朴素贝叶斯模型[34] 中假定属于类别i的数据的特征服从(μ_i,δ_i)的高斯分布, 并且假定每个类别具有先验概率π_i, 该方法

通过对模型参数加入噪声来实现该算法的隐私保护. 因此高斯朴素贝叶斯的关键参数为μ,σ,π. 差分隐私下的随机森林算法[35]中，利用指数机制对叶子节点中的数据标签进行采样以实现隐私保护，因此对数据的扰动是指只对随机森林中每棵树的叶子节点的标签产生影响，所以其关键参数为数据点在随机森林中每棵树的预测概率. 对于 DP-SGD 算法，ML-Audit 方法采用 ClipBKD 对数据进行投毒，在这一方法中只关注对数据点投毒前后算法预测结果的变化，因此选取预测概率作为重要参数.

Table 1 Summary Function

表 1 总结函数

方法	总结函数
逻辑回归[32-33]	协方差系数θ
高斯朴素贝叶斯[34]	$\mu_{yd},\sigma^2_{yd},\pi_d$
随机森林[35]	每棵树的 $Pr(x^*)$
DP-SGD[21]	$Pr(x^*), Pr(\mathbf{0}), Pr(x_i^{\text{test}})$

虽然 ML-Audit 方法是第 1 个可以对机器学习模型进行审计的方法，但其构建数据集计算时间较长，并且对于一些复杂模型，若想提高其审计结果，可能需要大量参数，导致审计代价提高.

3.1.2 多强度数据构建

3.1.1 节介绍的 ClipBKD 方法和 ML-Audit 方法均对利用接近于实际攻击者的攻击方式进行审计，一定程度上证明了其审计的隐私算法的隐私保护程度并没有违背其理论隐私预算上界，同时也证明了其审计方法的有效性，但并没有验证被审计的隐私算法上界的严格性. 随着组合定理的改进，一些已经提出隐私算法的理论隐私预算上界可能会随着组合定理的改进而越来越严格，但这种现象是否会一直持续下去是一个未知的问题.

Nasr 等人[36]在中心化机器学习场景下，通过设计 6 种具有不同背景知识的攻击者，对 DP-SGD 算法进行审计，以判断其理论隐私预算上界的严格性. 该审计方法的审计过程与隐私审计通用流程相同，首先需要构建相邻数据集. 该方法的流程图如图 6 所示.

这一方法中共有 6 种不同强度的攻击者，分别由不同的数据构建者、不同能力的辨别者组合得到，辨别者根据辨别信息判断数据点是否参与了模型的训练. 该审计方法共构建 6 种不同强度的攻击者，攻击者能力如表 2 所示.

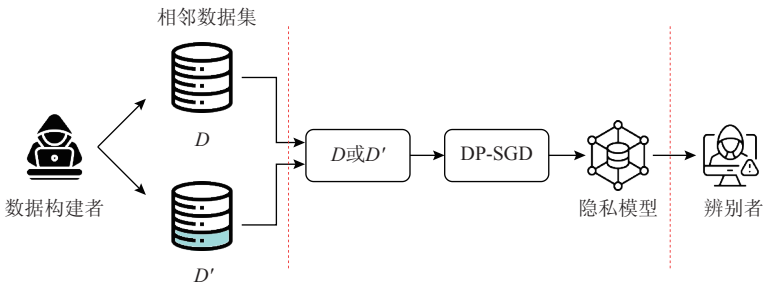


Fig. 6 Methodology flowchart of Nasr[36]

图 6 Nasr[36] 方法流程图

Table 2 Advserary Ability of Nasr[36] Method

表 2 Nasr[36] 方法攻击者能力

攻击者	数据构建者	攻击目标	辨别信息
攻击者 1	$D'=D-(x,y)$	最终模型	点 (x,y) 在最终模型上的损失函数值
攻击者 2	利用影子模型梯度生成对抗样本	最终模型	
攻击者 3	利用影子模型梯度生成对抗样本	所有中间模型	点 (x,y) 在中间模型的损失函数值
攻击者 4	构建投毒点	特殊数据点	
攻击者 5	在模型参数中加入水印信息	中间模型的所有参数	模型参数的距离
攻击者 6	构建有毒数据集	模型梯度	

对于每一种攻击者，Nasr 方法均需要训练 1 000 个模型并获取对应的统计攻击结果，利用置信区间对统计结果进行校正后，计算经验隐私预算下界. 该审计方法不仅验证了 DP-SGD 理论隐私预算上界的严格性，同时也说明该理论上界无法再利用先进的组合定理进行改进，因为在该审计工作中所假设的

最强攻击者的情况下, 经验隐私预算下界已经与理论隐私预算上界十分接近.

Matsumoto 等人[37]在分布式数据收集场景下, 同样设计 6 种具有不同背景知识的攻击者对 LDP 扰动算法 LDP-SGD[38]进行审计. 该算法的应用场景与联邦学习相同, 均包含服务器端和客户端.LDP-SGD 可划分为客户端扰动和服务器端扰动 2 部分. 客户端利用本地化差分隐私算法对本地梯度进行扰动并发送给服务器端, 服务器端将收到的客户端的梯度进行聚合并加以扰动. 该方法同样利用 6 种具有不同背景知识的攻击者对 LDP-SGD 进行审计, 每个攻击者同样包含数据构建者和辨别者 2 个角色, 攻击者的攻击能力如表 3 所示.

攻击者 1 中的数据构建者不对数据做任何处理, 黑盒辨别者 1 通过计算数据点在模型 2 轮训练前后的梯度差值进行判断, 其中 $\delta(x)=|f(x;\boldsymbol{\theta}_{t+1})-f(x;\boldsymbol{\theta}_t)|$, 若 $\delta(x_1)\geqslant\delta(x_2)$则训练数据为 x_1. 攻击者 2 利用点 x_1 的梯度采用 FSGM(fast gradient sign method)[39]构建一个可以在当前模型产生最大损失函数值的新图像 x_2. 攻击者 3 利用点 x_1 在模型上一轮的梯度构建新梯度 $\boldsymbol{g}_2$. 攻击者 4 中的数据构建者仅需计算点 x_1 的梯度并对其进行翻转即可构成新梯度 $\boldsymbol{g}_2$. 攻击者 5 中服务器端与端之间相互勾结, 构建一个具有较大损失的数据点 x_1 并计算其梯度, 对其进行翻转即可构成新梯度 $\boldsymbol{g}_2$. 攻击者 2~5 的黑盒辨别者均为黑盒辨别者 2, 通过区分数据点在 2 次相邻的模型训练过程中的梯度来判断该点是否参与了训练, 若数据点的梯度值满足 $f(x;\boldsymbol{\theta}_{t+1})\leqslant f(x;\boldsymbol{\theta}_t)$, 则说明其参与了训练. 攻击者 6 中的数据构建者随机构建点 x_1 的梯度, 其每一位取值均为 λ, 点 x_2 的梯度为 $-\lambda$. 所有攻击者中均可采用白盒辨别方式进行辨别, 根据数据点扰动前后的梯度之间的余弦相似度判断其是否参与了训练.

LDP-SGD 方法总体遵循隐私审计通用流程, 但图 3 的步骤 1 和步骤 2 的具体实现方式与其他审计方式略有不同. 步骤 1 中, 由于是对本地化差分隐私算法进行审计, 因此数据构建者只需构建 2 个不同的数据点. 每个数据点可以看作联邦学习中的一个恶意客户端, 因为在本地化差分隐私中假定每个用户只有 1 条数据. 步骤 2 中, 该审计方法中的黑盒鉴别者判断某一数据参与了训练的方法是否与其他审计方法不同. 其他审计方法关注某一数据点在当前训练或当前模型上的损失函数值等指标, 而 Matsumoto 等人[37]利用当前数据点在连续 2 轮中梯度的损失函数值是否减小来判断该点是否参与了训练. 白盒鉴别者与其他审计方法类似, 通过比较扰动前后梯度的余弦值来判断数据点是否参与了训练. 该审计方法流程图如图 7 所示.

数据构建者只需构建 2 个数据点 x_1,x_2 并获取其梯度, 随机选择其中一个梯度将其扰动为 $\widetilde{\boldsymbol{g}}$, 并发送给服务器端, 服务器端利用其接收的梯度对全局模型更新. 在这一过程中, 鉴别者将会接收到数据点的扰动梯度 $\widetilde{\boldsymbol{g}}$ 以及服务器端模型更新后的梯度. 鉴别者利用这 2 个已知信息推断 x_1, x_2 中哪个数据点参与了训练.

该工作是唯一一篇从联邦学习角度审计本地化差分隐私算法的工作, 验证了 LDP-SGD 算法上界的严格性. 但其与 Nasr 等人[36]所提出的工作在整体架构上十分相似, 同样需要训练大量模型以得到统计结果, 因此审计代价较大.

2 个方法之间的对比如表 4 所示. 其中 Nasr 等人[36]对中心化差分隐私算法审计和 Matsumoto 等人[37]对本地化差分隐私算法审计的攻击者均为 6 种, 攻击方式均分为白盒和黑盒 2 种. 由于所审计方法的场景不同, 检验中的假设关注于不同的情况. 在数据构建方面, 中心化场景关注于构建相邻数据集, 本地化场景关注于构建数据点.

Table 3 Advserversary Ability in Auditing LDP-SGD[37] Method

表 3 审计 LDP-SGD 方法中的攻击者能力[37]

攻击者	数据构建者	辨别者	
		黑盒	白盒
攻击者 1	无处理	黑盒辨别者 1	白盒辨别方法
攻击者 2	利用 FSGM[39] 图像梯度	黑盒辨别者 2	白盒辨别方法
攻击者 3	生成大梯度值 $\boldsymbol{g}_1=\nabla f(x_1,\boldsymbol{\theta}_t)$; $\boldsymbol{g}_2=\nabla f(x_1,\boldsymbol{\theta}_t')$, $\boldsymbol{\theta}_t'=\boldsymbol{\theta}_t+\alpha\nabla f(x_1;\boldsymbol{\theta}_t)$	黑盒辨别者 2	白盒辨别方法
攻击者 4	$\boldsymbol{g}_1=\nabla f(x_1,\boldsymbol{\theta}_t);\boldsymbol{g}_2=-\boldsymbol{g}_1$	黑盒辨别者 2	白盒辨别方法
攻击者 5	$\boldsymbol{g}_1=\nabla f\left(x_1,\tilde{\boldsymbol{\theta}}_t\right)$; $\boldsymbol{g}_2=-\boldsymbol{g}_1$	黑盒辨别者 2	白盒辨别方法
攻击者 6	$\boldsymbol{g}_1=(\lambda,\lambda,\cdots,\lambda);\boldsymbol{g}_2=-\boldsymbol{g}_1$	黑盒辨别者 3	白盒辨别方法

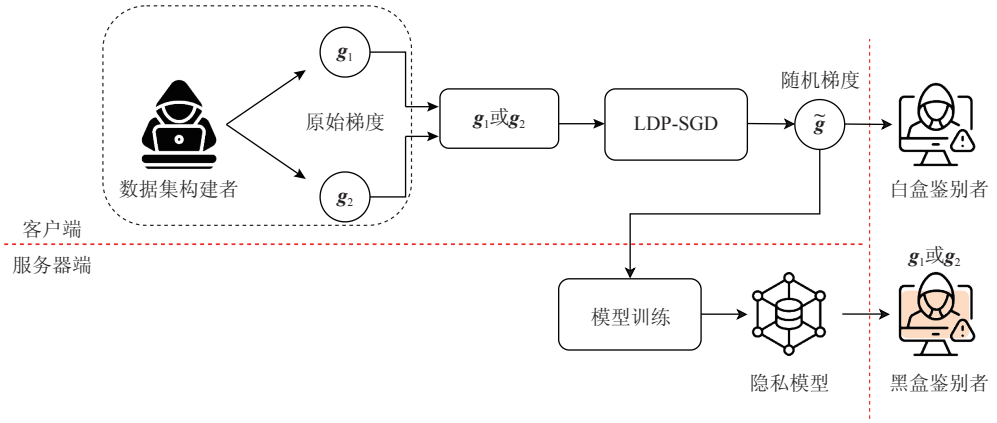


Fig. 7 Methodology flowchart of auditing LDP-SGD[37]

图 7 审计 LDP-SGD[37] 的方法流程图

Table 4 Comparison of Nasr[36] and Matsumoto[37] Methods

表 4 Nasr[36] 和 Matsumoto[37] 方法对比

对比要素	Nasr[36]	Matsumoto[37]
审计对象	DP-SGD	LDP-SGD
隐私算法类别	CDP	FL+LDP
攻击者数量	6 种	6 种
数据构建者数量	5 种	4 种
辨别者数量	3 种	2 种
假设检验	数据集	数据点

3.2 针对 LDP 的差异数据项构建

Arcolezi 等人[40] 提出针对分布式数据收集场景下的 8 种可用于频数估计任务的 LDP 算法审计方法 LDP-Auditor[40]. 由本地化差分隐私的定义可知, 本地化差分隐私关注于 2 个不同数据的输出概率分布. 因此对本地化差分隐私算法的审计需要考虑的并不是相邻数据集, 而是对不同输入点的审计. 与前文所介绍的审计本地化差分隐私算法 LDP-SGD 不同, Arcolezi 等人[40] 对应用于数据库查询的频数估计算法审计, 其中包括: GRR(generalized randomized response)[41]、SS(subset selection)[42-43]、UE(unary encoding-UE)[44] 中的 OUE(optimal nary encoding)[25] 和 SUE(symmetric unary encoding)[25], LH(local Hashing)[45] 中的 OLH(optimal local hashing)[25] 和 SLH(symmetric local hashing)[25], HE(histogram encoding)[25] 中的 THE(threshold histogram encoding)[25] 和 SHE(symmetric histogram encoding)[25].

这一审计方法遵循隐私审计通用流程, 步骤 1 仅需构建 2 个不同的输入数据点, 步骤 2 中所用辨别者为可辨性攻击 (distinguishability attack). 对于步骤 1, 审计过程中只需构建 2 个不同的输入数据点即可. 由之前综述的众多方法可知, 审计需要利用辨别者的大量预测结果构建 2 个不同的分布, 因此可以利用本地化差分隐私算法对输入数据点进行多次扰动并统计结果. 因为差分隐私算法的扰动具有随机性, 即使每次扰动的是相同的输入数据, 其扰动结果也并不相同. 对于步骤 2, 可辨性攻击不关注于本地化差分隐私算法的统计输出, 只关注于扰动后的数据取值情况. 对于算法 OLH, SLH, THE, OUE, SUE, SS, 该攻击者从支撑集(support set)中均匀采样一个数据作为其对当前算法扰动的预测. 对于 GRR 算法, 攻击者的预测与输入相同. 若攻击者的预测与算法对数据点的扰动结果相同, 则攻击者预测正确. 最终的结果统计过程与其他方法相同, 统计获得攻击者预测正确的次数后利用 Clopper Pearson 置信区间对其校正, 利用校正所得的概率计算获取经验隐私预算下界. Arcolezi 等人[40] 使用上述方法对开源隐私库 pure-LDP①中 UE 算法进行审计, 并发现该算法在实现中的错误.

Arcolezi 等人[40] 证明了 8 种用于分布式数据收集场景的频数估计算法的上界足够严格, 但本地化差分隐私算法在设计过程中经过了严格的数学证明, 并且算法中并没有用到神经网络等比较复杂的结构, 因此其理论隐私预算上界与机器学习下的差分隐私算法相比相对严格, 审计所得的经验隐私预算下界与理论隐私预算上界之间差距不大.

3.3 小 结

本节主要介绍了 2 类隐私审计的数据构建方法,

① https://pypi.org/project/pure-ldp/

Table 5 Data Constructing Methods for Privacy Auditing

表 5 隐私审计的数据构建方法

类别	方法	审计对象	优点	缺点
针对差分隐私的特殊数据点构建	ClipBKD[30]	DP-SGD	构建的投毒点受梯度裁剪影响小	所需训练的模型个数较多
	ML-Audit[31]	LR, NB, RF, DP-SGD	可以审计多个差分隐私算法	
	Nasr 等人[36]	DP-SGD	验证了隐私预算上界的严格性	
	Matsumoto 等人[37]	LDP-SGD	验证了隐私预算上界的严格性	
针对本地化差分隐私的差异数据项构建	LDP-Auditor[40]	8 种本地化差分隐私频数估计方法	可以审计多个本地化差分隐私算法	审计过程迭代次数多

表 5 为本节方法总结. 针对差分隐私的特殊数据点构建与针对本地差分隐私的差异数据项构建不同, 前者更关注于如何增大相邻数据集之间的差距, 后者只关注 2 个不同的输入点. 但关注于数据构建的审计方法均需要训练大量模型才会得到比较准确的经验隐私预算下界, 因此整体开销较大.

4 隐私审计的数据测算方法

隐私审计的数据测算方法可以划分为单点数据测算和多点数据测算 2 种. 单点数据测算为隐私审计中的常用思想, 需利用相邻数据集训练大量模型并观察差异点在所有模型上参与训练的情况. 虽然该审计思想所获得的经验隐私预算下界更加准确, 但审计代价较大. 多点数据测算只需训练 1 个模型即可. 这类方法的优点是审计代价相对较小, 缺点是审计结果与单点数据测算方法相比并不准确. 本节将对这 2 种审计方法进行介绍.

4.1 单点数据测算

单点数据测算方法利用相邻数据集训练大量模型, 观察差异点在所有模型上的训练情况以获取经验隐私预算下界. 该类方法可划分为中心化场景的单点数据测算以及分布式场景的单点数据测算.

4.1.1 中心化场景的单点数据测算

差分隐私算法层出不穷, 对 DP-SGD 进行改进的算法也有很多. 在这些改进算法中, 有一部分算法是合理有效的, 但有一部分算法可能会由于证明出错或代码实现出错而导致无法达到它所声称的保护程度. BpC(backpropagation cipping)[5] 方法则为一个真实的错误例子, 它是对 DP-SGD 改进的一种算法.DP-SGD 只会对每一个数据点的梯度裁剪 1 次, BpC 方法提出了在每层双向裁剪的方法, 即数据输入到当前层的时候进行 2 次裁剪, 继续向后传播时再进行 1 次裁剪, 最终每个数据在当前这一层的裁剪范围是 $C_1 \times C_2$. 该算法声称自己可以达到 $(0.21, 10^{-5})$-DP 的隐私保护程度.

Tramer 等人[6] 在中心化机器学习场景下对 BpC 方法进行审计, 并发现了该算法中的错误, 其实际的隐私保护程度几乎与它理论证明的保护程度相差近 10 倍. 该审计案例的总体过程遵循隐私审计通用流程. 其中步骤 1 中需要利用投毒攻击构建相邻数据集, 步骤 3 中需要找到用于统计攻击结果的损失函数阈值 τ. 对于步骤 1, 构建相邻数据集需要找到最佳投毒点. 在该审计案例中利用 MNIST 数据集进行审计, 选取 MNIST 数据集中的前 25 个原始数据, 分别训练 1 000 个模型, 同时利用完整数据集训练 1 000 个模型. 在这 25 个数据中找到可以使得当前隐私算法获取最大经验隐私预算下界的图像数据作为最佳投毒点. 利用投毒攻击对其进行投毒, 即可获得相邻数据集 D 和 D'. 相邻数据集构建完成后, 训练 2 000 个当前所需审计的 BpC 算法模型, 其中一半模型利用数据集 D 训练, 另一半利用数据集 D' 训练. 攻击者根据损失函数值判断投毒点是否参与了训练. 因为在机器学习模型中, 参与了模型训练的数据点的损失函数值远小于未参与模型训练的数据点的损失函数值. 因此通过统计最佳投毒点在这 2 000 个模型上的损失函数值, 找到可以最大化经验隐私预算下界的损失函数值作为阈值 τ. 最佳投毒点以及损失函数阈值 τ 都找到后, 即可利用构建好的相邻数据集训练大量模型并进行审计. 在该审计案例中共需训练 100 000 个模型. 审计中的攻击者只关注最佳投毒点的损失函数值. 对攻击者的攻击结果统计后即可获得 TPR 及 FPR, 并利用 Clopper Pearson 置信区间对其进行校正, 计算校正后 2 个概率的比值即可获取最终的经验隐私预算下界.

4.1.2 分布式场景的单点数据测算

中心化场景的单点数据测算通过训练大量模型即可根据差异点是否参与训练获取经验隐私预算下界. 若将这种思想直接应用于分布式联邦学习下的差分隐私算法, 训练大量模型的时间将会更长, 因为通常一个联邦学习模型的训练需要几天甚至几周的

时间. 因此, 对联邦学习下差分隐私算法的审计可以考虑在模型训练过程中进行同步审计. 在该场景下, 审计主要关注于联邦学习下差分隐私的特有算法 DP-FedSGD 的实际隐私保护程度.

Maddock 等人[46] 提出分布式联邦学习场景下对差分隐私算法 DP-FedSGD 审计的方法 CANIFE. 该方法是对联邦学习下的差分隐私算法 DP-FedSGD 审计的工作. CANIFE 可以审计在某一轮服务器端模型更新的经验隐私预算下界, 其流程图如图 8 所示.

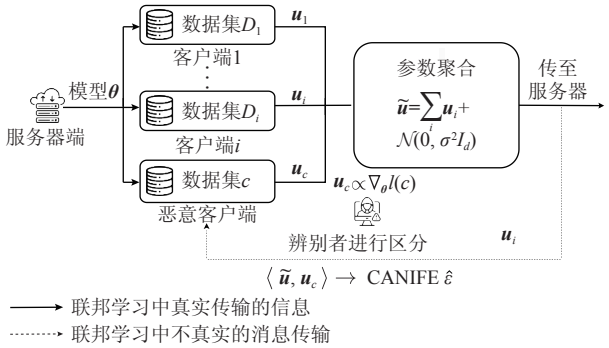


Fig. 8 Methodology flowchart of CANIFE[46]

图 8 CANIFE[46] 方法流程图

CANIFE 方法加入一个只有 1 个告密点的恶意客户端, 并且告密点的梯度尽可能地与其他所有客户端的梯度正交. 该恶意客户端参与模型训练的概率与其他客户端相同. 攻击者对告密点的攻击目的与对中心化差分隐私算法审计的攻击目的相同, 仍为判断告密点是否参与了训练. 由于每轮参与训练的客户端数量很大, 因此恶意客户端与其他所有客户端的梯度乘积即可构成一个分布. 恶意客户端的梯度与其他客户端梯度之间的乘积分布表示为

$$\langle \tilde{\boldsymbol{u}}, \boldsymbol{u}_c \rangle \propto \sum_i \langle \boldsymbol{u}_i, \nabla_{\boldsymbol{\theta}} \ell(z) \rangle + \left\langle \mathcal{N}\left(0, \sigma^2 I_d\right), \nabla_{\boldsymbol{\theta}} \ell(z) \right\rangle. \quad (3)$$

若告密点参与了该轮训练, 则梯度乘积分布的均值不为 0, 否则为 0. 因此在 CANIFE 中, DP-FedSGD 算法每一轮的经验隐私预算下界都是通过区分 2 个均值不同的正态分布获得. 其中, 告密点的加入不会影响整体模型的效果, 但会增加整体的训练时间. 因为在这一审计过程中, 服务器端只计算平均梯度, 而不用该梯度进行更新, 相当于暂停模型的训练进行审计. 但 CANIFE 方法具有高维梯度向量性质失效和采样方式不现实 2 个缺陷. 对于高维梯度向量性质失效问题, CANIFE 中构建高维梯度向量的方式会随着客户端增多而失效, 因为在高维空间中, 一旦模型的维度足够高, 那么构建的告密点将与每个模型的最终梯度正交, 导致无法根据梯度之间的乘积判断告密点是否参与了训练. 对于采样方式不现实问题, 在联邦学习中每轮训练中对客户端进行独立同分布的泊松采样(poisson sampling)并不可行.

4.2 多点数据测算

单点数据测算需要训练大量模型或暂停模型以获取差异点的训练情况, 导致审计代价较大. 为进一步降低审计代价和减少隐私审计所需的时间, 研究者们分别针对中心化场景和分布式场景对多点数据测算方法进行研究.

4.2.1 中心化场景的多点数据测算

审计方法通常利用频率学派的思想, 即攻击者将其对模型的攻击结果进行统计并转换为经验隐私预算下界. 除此之外, 隐私审计同样可以看作是一个统计泛化问题, 经验隐私预算下界为其估计的参数. 其中, 攻击者攻击后所得的大量结果可以看作是样本数据, 经验隐私预算下界即可看作是一个总体参数值, 因此可以将统计估计用于对经验隐私预算下界的估计.

Steinke 等人[47] 在中心化机器学习场景下利用统计估计的方法对算法 DP-SGD 审计. 这一审计方法包含白盒和黑盒 2 种审计策略. 白盒审计下, 攻击者可以获取数据在每一轮训练中的梯度信息; 黑盒审计下, 攻击者只能获取数据在第 1 轮和最后 1 轮的损失函数值. 以黑盒情况为例, 可计算数据点在第 1 轮训练中的损失函数值和最后 1 轮训练中的损失函数值的差值, 获取数据点的得分, 并利用这一得分判断数据点是否参与了训练. 审计过程中, 所有参与了模型训练的数据都会有一个得分, 将得分最高的 k_+ 个分数赋值为+1, 表示预测为参与了训练; 最小的 k_- 个赋值为−1, 表示预测为未参与训练; 其余位置赋为 0. 利用攻击者的预测结果与数据真实的参与训练的情况计算 p-value, 并找到最优 p-value 对应的最优经验隐私预算下界. 这一方法虽然大大减小了审计过程中的代价, 但缺陷是该审计方法所获得的经验隐私预算下界不如训练大量模型的审计方法所获得的准确.

4.2.2 分布式场景的多点数据测算

在分布式场景下, 多点数据测算的审计方法关注于联邦学习下差分隐私的特有算法 DP-FedAvg 的实际隐私保护程度. Andrew 等人[48] 在分布式联邦学习场景下针对 CANIFE 方法中的采样失效问题, 提出对 DP-FedAvg 算法审计的方法 one_shot_eps.该审计方法中加入多个告密点, 只需训练 1 个模型, 并且审计与模型的训练同步, 避免了需要获取大量模型导致的训练代价大的问题.

one_shot_eps 方法中考虑了 2 种攻击者, 可以获取模型中间状态的攻击者和无法获取模型中间状态的攻击者, 其中攻击者无法获取模型的中间状态更符合实际的情况. 这一审计方法需要在模型的训练过程中加入若干个告密点, 计算告密点与告密点在该模型上的输出之间的余弦值并判断其是否参与了训练. 若余弦值大于阈值, 则认为该告密点参与了训练. 因此该方法计算经验隐私预算下界的方式与利用多次模型攻击中需要结合阈值进行判断的方法相同, 利用 *FNR* 和 *FPR* 计算得到经验隐私预算下界. 该审计方法同时也可以用于攻击者可以观察到模型训练的中间状态的情况, 在这一情况中, 告密点将分为参与模型的训练和不参与模型的训练 2 部分. 所有告密点均与当前模型的梯度信息进行点乘, 因此可以获得参与训练的告密点获得的乘积的分布和未参与训练的告密点获得的乘积的分布 2 个高斯分布, 最终的经验隐私预算下界利用式(4)即可计算获得, 其中 F 为累积分布函数.

$$\varepsilon(X_0\|X_1;\delta,\alpha)=\max\left(\lg\frac{F_0(\alpha)-\delta}{F_1(\alpha)},\lg\frac{1-\delta-F_1(\alpha)}{1-F_0(\alpha)}\right). \tag{4}$$

4.3 小 结

本节介绍了隐私审计的数据测算方法, 表 6 列出了本节方法的总结. 当前大多数审计方法采用单点数据测算的思想获得经验隐私预算下界, 但其审计代价很大, 通常需要训练大量模型才能保证当前所获得的经验隐私预算下界是可靠的. 多点数据测算需要构建足够多的告密点或获取若干个模型训练过程中的中间状态, 其本质仍是获取大量的经验数据. 分布式联邦学习场景下的审计与其他场景相比审计时间较长, 这与联邦学习模型自身训练时间较长相关. 中心化机器学习场景下的审计方法中, 多点数据测算方法所获得的经验隐私预算下界不如单点数据测算获取的经验隐私预算下界严格.

Table 6 Data Measurement Methods for Privacy Auditing

表 6 隐私审计的数据测算方法

类别	方法	审计场景	审计对象	优点	缺点
单点数据测算	Tramer 等人[6]	中心化机器学习	BpC	找到已发布算法中的错误	训练模型个数过多
	CANIFE[46]	分布式联邦学习	DP-FedSGD	审计可以随模型训练进行	审计时间较长
多点数据测算	Steinke 等人[47]	中心化机器学习	DP-SGD	审计只需训练 1 个模型即可	经验隐私预算下界不严格
	one_shot_eps[48]	分布式联邦学习	DP-FedAvg	审计只需训练 1 个模型即可	审计时间较长

5 隐私审计的结果量化方法

隐私审计的结果量化方法可分为基于差分隐私变体的量化和基于贝叶斯的量化 2 类. 基于差分隐私变体的量化中利用多种差分隐私变体定义或与攻击者攻击成功率相关的度量指标对隐私算法进行审计. 部分变体定义针对于特定应用场景的差分隐私算法, 例如高斯机制等. 这些隐私变体中的隐私参数上界严格程度不同, 并且可以相互转换, 由此, 可利用严格的差分隐私定义对相对松弛的差分隐私算法审计. 本节将详细介绍 2 类针对结果量化的审计方法, 其中 5.1 节中介绍的是基于差分隐私变体量化的审计方法, 其中包括高斯差分隐私、Lifted DP、RKRDP 及 Rényi 散度的 2-cut 形式, 5.2 节中介绍的是基于贝叶斯量化的审计方法, 其中包括基于差分可辨的量化和基于贝叶斯变量的量化.

5.1 基于差分隐私变体的量化

本文例举的差分隐私变体定义通常指比(ε,δ)-DP定义更严格的差分隐私定义, 包括 f-DP、LiDP、RKRDP、Rényi 散度的 2-cut 形式.

5.1.1 基于 f-DP 的量化

差分隐私定义中利用ε表示隐私上界, 当前常用的序列组合定理、高级组合定理, 以及 moment accountant[21] 所获得的理论隐私预算上界并不是十分严格, 因此产生了很多变体差分隐私定义, f-DP即为其中一种. 该隐私定义是 2022 年由 Dong 等人[27] 提出的一种假设检验角度的差分隐私定义. 在这一定义中所用的假设检验的原假设为H_0: 模型的输出来自$M(D)$; 备择假设为H_1: 模型的输出来自$M(D')$. 在假设检验中, 显著性水平α表示错误地拒绝原假设的概率, 显著性水平β表示错误地接受原假设的概率. f-DP 中利用权衡函数(trade-off function)表示显著性水平α和β之间的关系. 权衡函数可以用于表示区分 2 个分布的难度. 如果$T(P,Q)\geqslant T(P',Q')$, 则说明区分P, Q这 2 个分布之间的难度大于区分P',Q'这 2 个分布之间的难度. 差分隐私关注算法在相邻数据集上的输出分布不可区分性, 因此将权衡函数与这 2 个输

出分布相结合，可得f-DP的定义.

定义 4. f-DP. 现有相邻数据集D和D'、权衡函数f, 如果算法$\mathcal{A}$满足式(5)则满足f-差分隐私.

$$T(\mathcal{A}(D),\mathcal{A}(D'))\geqslant f. \tag{5}$$

若权衡函数f中的2个分布为高斯分布，则f-DP转换为其特殊情况，如高斯差分隐私GDP，如定义5所示. GDP对应于利用高斯机制进行加噪的差分隐私定义，可以与(ε,δ)-DP进行转换. 其转换关系为：

$$\delta(\varepsilon)=\phi\left(-\frac{\varepsilon}{\mu}+\frac{\mu}{2}\right)-\mathrm{e}^{\varepsilon}\phi\left(-\frac{\varepsilon}{\mu}-\frac{\mu}{2}\right). \tag{6}$$

定义 5. μ-高斯差分隐私(μ-GDP). 现有相邻数据集D和D', ϕ是标准正态分布的累积分布函数，如果算法$\mathcal{A}$满足式(7)，则满足μ-GDP.

$$T(\mathcal{A}(D),\mathcal{A}(D'))(\alpha)\geqslant\phi\left(\phi^{-1}(1-\alpha)-\mu\right),\quad\forall\alpha\in[0,1]. \tag{7}$$

Nasr等人[49]在中心化机器学习场景下利用μ-GDP对DP-SGD算法审计. 在黑盒情况下，需构建一对差异点为(x,y)的相邻数据集，并利用这对相邻数据集训练大量模型，获取点(x,y)在2个模型上损失函数值的分布. 在白盒情况下，只需利用数据集D训练一个模型，此时相邻数据集利用模型每轮训练的批数据进行构建. 模型的每一轮迭代会选取一批数据B并利用这批数据对模型参数进行更新. 同时以一定概率向批数据B中加入点(x,y)以构建相邻数据集B'，并计算批B'的梯度. 其中，点(x,y)的梯度尽量与其他数据点梯度正交. 在整体训练中，模型共训练T轮，每轮包含τ次迭代，每一次迭代中计算点(x,y)与批B的梯度乘积以及与批B'的梯度乘积，若点(x,y)参与了训练则其梯度与批梯度的乘积不为0，否则为0.模型整体训练完可得2个由$T\times\tau$个数据构成的分布. 由于攻击者已知模型训练时的梯度裁剪大小和批数据大小，因此可将这2个分布转化成2个具有不同均值的高斯分布并进行区分.

5.1.2 基于LiDP的量化

Pillutla等人[28]在中心化机器学习场景下利用LiDP(lifted differential privacy)定义对DP-SGD算法进行审计. 现有隐私审计方法均需要构建一对相邻数据集D和D'，构建后所得的相邻数据集之间相差的数据是确定的，因此需要利用相邻数据集训练大量模型，以得到模型在相邻数据集相差的数据点上的2个输出分布. 但利用这种审计方式获得的置信区间将会受到$\frac{1}{\sqrt{n}}$的限制，其中n为训练所得的模型个数. 为打破之前的审计方法中置信区间的这一影响，Pillutla等人[28]提出了随机的相邻数据集下的差分隐私定义LiDP及更严格的置信区间计算方法，避免了需要训练大量模型才可以获得经验隐私预算下界. LiDP中仍关注隐私算法在2个相邻数据集上的输出分布，但这2个相邻数据集是随机数据集，即数据集中的数据满足于某一分布，LiDP的定义如下：

定义 6. LiDP. 当前有一对随机的相邻数据集D和D'，一个随机算法$\mathcal{A}:\mathcal{Z}^*\to\mathcal{R}$，以及对应假设检验的拒绝域$R$，其中$\mathcal{R}$为任意一个$R$的子集，$\mathcal{P}$为独立于随机算法$\mathcal{A}$的有关于$D$, D', $\mathcal{R}$的联合概率分布. 若随机算法$\mathcal{A}$满足不等式(8)，则随机算法$\mathcal{A}$满足(ε,δ)-LiDP.

$$Pr_{\mathcal{A},\mathcal{P}}(\mathcal{A}(D_1)\in R)\leqslant\mathrm{e}^{\varepsilon}Pr_{\mathcal{A},\mathcal{P}}(\mathcal{A}(D_0)\in R)+\delta. \tag{8}$$

在LiDP中，设当前数据集为D，审计过程中可以加入K个相互独立且随机的告密点，构建的相邻数据集为$D=D\cup K$, $D'=D\cup(K-1)$，即2个相邻数据集之间相差的是第k个告密点. 由于K个告密点之间都是相互独立的，训练单个模型即可获得K个不同的告密点是否分别参与了训练. 为得到更严格的审计结果，该方法提出高阶可交换伯努利置信区间，其2阶置信区间中区间大小与样本数量的关系即被限制在$1/n^{3/4}$，ℓ阶置信区间与样本个数之间的关系为$1/n^{(2\ell-1)/2\ell}$，收紧了根据统计结果所得的经验隐私预算下界.

与其他审计方法相比，利用LiDP定义进行审计所需训练的模型个数大大减少. 利用其他差分隐私定义进行审计时，每个模型只能用于判断某个数据点是否参与了训练，因此若想获取K个不同数据点是否参与了训练需要训练大量模型. LiDP中考虑的是随机数据集的差分隐私定义，因此利用LiDP进行审计时，可以加入K个独立同分布的告密点，只训练1个模型即可获取每一个数据点是否参与了训练，可以减小隐私审计的代价.

5.1.3 基于RKRDP的量化

Domingo-Enrich等人[29]在中心化机器学习场景下，提出RKRDP(regularized kernel Rényi differential privacy)定义并对标准差分隐私定义中的加噪机制高斯机制审计. 隐私审计方法为得到较为严格的经验隐私预算下界，需要训练上千个或更多模型. 若审计利用高维数据训练的隐私模型则将进一步加大审计的代价. 因为随着维度的增加，获取经验隐私预算下界所需要的数据量将会指数倍增长. 由此Domingo-Enrich等人[29]提出了一种高维空间中的松弛差分隐私的定义RKRDP，该定义可用于审计利用高维数据

训练的隐私模型. RKRDP 的定义如下:

定义 7. RKRDP. 一个随机算法 M, 对于任意的 $\varepsilon,\delta>0$ 的取值, 如果有 $\alpha\in[1/2)\cup(1,+\infty)$、任意的核函数 k、正则化系数 λ, 以及一对相邻数据集 D 和 D' 满足式(9), 则随机算法 M 满足 RKRDP. 其中 $\sum L(M(D))$ 和 $\sum L(M(D'))$ 为算法 M 在相邻数据集上输出分布的协方差算子, α 为 Rényi 熵的矩, ε 为理论上界.

$$D_{\alpha,\lambda}\sum L(M(D))\|\sum L(M(D'))\leqslant\varepsilon. \tag{9}$$

该差分隐私定义与差分隐私的形式化定义之间的关系为 (ε,δ)-DP 包含 $(k,\alpha,\delta e^{-\varepsilon},\varepsilon)$-RKRDP.

RKRDP 的定义形式与概率空间中常用于区分 2 个概率分布的 KL 散度(Kullback-Leibler divergence)形式相似. 该方法在高维希尔伯特(Hilbert)空间将核函数与量子态(quantum)结合, 将 KL 散度中的概率用高维 Hilbert 空间中的协方差算子代替, 实现对 2 个高维的概率分布的区分. 审计过程中, 首先从 2 个高斯分布中分别采样大量独立同分布的数据点, 并近似得到它们的协方差矩阵, 即可获取其在 RKRDP 下的经验隐私预算下界, 并转换为在差分隐私定义中隐私参数 ε 的对应值. 但 RKRDP 只能用于高斯机制的审计, 无法应用于其他差分隐私算法的审计.

5.1.4 基于 Rényi 散度的量化

Chadha 等人[50] 在中心化机器学习场景下, 利用 Rényi 散度的 2-cut 形式对利用了隐私预测接口框架的 4 个方法进行审计, 包括 PATE[51-52], CaPC[53], Prompt-PATE[54], Private-kNN[55]. 对于分类问题, 除了利用上述在训练过程中进行加噪的 DP-SGD 算法训练模型, 还可以利用以隐私预测接口 PDI(private prediction interface)[56] 为框架构建仅在模型推断或预测阶段加入噪声的分类模型. 该接口整体可分为训练阶段和预测阶段. 在训练阶段, 首先将数据集随机划分为 m 个不相交的子集, 每一个子集都需分别训练一个对应的教师模型. 在预测阶段, 假设当前查询点为 q, 从训练好的 m 个教师模型中随机抽取 k 个对查询点进行预测, 将 k 个教师模型的预测结果统计为直方图, 利用 noisy arg max 方法对预测直方图加入高斯噪声, 并选取概率最大的类别作为预测结果进行发布, 形式为 $Predictions=\arg\max_{y\in Y}H_S+\mathcal{N}\left(0,\sigma^2\right)$. 隐私预测接口满足 Rényi 差分隐私[57](简称 RDP), 如定义 8 所示.

定义 8. Rényi 差分隐私. 对于任意一对相邻数据集 D 和 D', 若随机算法 $f:\mathcal{D}\mapsto R$ 满足不等式(10), 则说明该算法满足 (α,ε)-Rényi 差分隐私(简称 (α,ε)-RDP). 其中 D_α 为 Rényi-散度(Rényi divergence)且 $\alpha\in(1,\infty)$,

$$D_\alpha(f(D)\|f(D'))\leqslant\varepsilon. \tag{10}$$

定义 9. 隐私预测接口. 对于预测接口 $A(S)$, 若在该接口上的每一个查询序列 Q 输出 $(A(S)\leftrightarrow Q)$ 满足 $(\alpha,\varepsilon_\alpha)$-RDP, 则预测接口满足 $(\alpha,\varepsilon_\alpha)$-RDP.

这类算法的审计方法的整体过程遵循审计通用流程. 利用相邻数据集训练获得 2 个接口模型后, 攻击者分别对这 2 个接口模型进行成员推理, 判断某一数据点是否参与了训练, 并获得对应的 *FNR* 和 *FPR*. 但目前没有可以用于获得严格的 RDP 隐私下界的成员推理攻击, 因此利用 Rényi 散度的 2-cut 形式获取其上界, 表达形式如式(11)所示. 将统计所得的 *FNR* 和 *FPR* 分别代入式(11)后, 即可获得对应的 RDP 隐私预算参数, 并转换为标准差分隐私的经验隐私预算下界. 这一审计方法大大减少了审计过程中所需的模型训练个数以及审计代价, 同时使得对该类隐私算法审计所得的经验隐私预算下界更加严格.

$$\overline{D_\alpha}^2(\mu_1|\mu_2):=\sup_{O\subseteq O}\frac{1}{\alpha-1}\lg\left(p_1^\alpha p_2^{1-\alpha}+(1-p_1)^\alpha(1-p_2)^{1-\alpha}\right). \tag{11}$$

5.2 基于贝叶斯的量化

大多数审计方法均为频率学派方法, 为获得相对严格的经验隐私预算下界, 审计人员需要训练成千上万个模型, 以获得某一数据点是否参与了训练的统计信息, 例如 *FPR*, *FNR* 等, 并利用 Clopper Pearson 等置信区间对所得的统计信息进行校正. 若审计过程中所获样本不足, 将导致设计所得的经验隐私下界不严格. 因此基于频率学派的审计方法的审计代价往往较高. 为降低审计代价并提高审计所得的经验隐私预算下界的严格性, 研究者们在审计方法中结合贝叶斯流派的思想, 利用贝叶斯后验概率或构建贝叶斯变量进行审计.

5.2.1 基于差分可辨的量化

Bernau 等人[58] 在中心化机器学习场景下, 利用差分可辨性质对 DP-SGD 算法审计. 差分隐私算法将单条数据对模型输出的影响限制在一定范围内, 使攻击者无法根据查询结果之间的差异推测出任意一条参与了模型训练的数据, 即将正确预测数据参与训练的概率限制在了一定范围之内. 隐私审计方法中, 辨别者常利用识别数据是否参与了训练的思想进行审计. 基于贝叶斯理论, Bernau 等人[58] 将 (ε,δ)-DP 的隐私参数转换为攻击者判断任意一条数据是否参与了训练的贝叶斯后验概率边界, 并引入攻击者优势作为互补边界, 对差分隐私算法进行审计. 在这一

审计方法中，所用的辨别者为差分可辨攻击者 DI（differential identifiability）[59]，DI 强于其他审计方法中所用的辨别者. 差分可辨性定义了该场景下推断某一数据点是否存在于某个数据集的概率上界. 具体而言，假设全局数据集为$\mathcal{U}$，数据集D为$\mathcal{U}$的子集，D与其相邻数据集D'之间只相差数据点i. 若攻击者已知算法输出以及D的相邻数据集D'，正确推断出数据点i在数据集D中的概率不超过ρ. 差分可辨性定义如下：

定义 10. ρ-差分可辨性（ρ-differential identifiability）. 给定任一查询 q，对于在完整数据集$\mathcal{U}$中的任意一对相邻数据集 D 和 D'，其中 $D'=D-i^*$，$\forall i\in\mathcal{U}-D'$，若随机算法$\mathcal{A}$满足式（12）则满足$\rho$-差分可辨性.

$$Pr\left[I(i)\in I_D|\mathcal{A}_f(D)=R,D'\right]\leqslant\rho. \tag{12}$$

基于该定义，差分可辨攻击者对攻击的准确率进行量化. 其中，对攻击者的正确预测利用后验概率进行量化，如定义 11 所示. 第k轮的后验概率与前一轮的后验概率相关，该后验概率量化了攻击者对某一数据参与了训练的置信程度，其上边界为ρ_β.

定义 11. 假设给定的算法$\mathcal{A}$需要迭代 k 次，训练所用数据集为D，利用差分可辨攻击者进行攻击预测以获得结果矩阵$\boldsymbol{R}_k=(r_0,r_1,\cdots,r_k)$，当模型迭代到第 k 轮，攻击者利用其观察到的所有信息，正确预测当前训练数据集为D的后验概率$Pr(D|\boldsymbol{R}_k)$为

$$\beta_k:=r(D|\boldsymbol{R}_k)=\frac{Pr(D,\boldsymbol{R}_k)}{Pr(D,\boldsymbol{R}_k)+Pr(D',\boldsymbol{R}_k)}. \tag{13}$$

除此之外，攻击者的攻击能力还可以用表示攻击者预测正确的频率的攻击者优势[60]进行量化，可通过计算预测成功概率与预测失败概率之差获得. 攻击者优势定义如下：

定义 12. 攻击者优势（advantage）. 将攻击者推断某一模型的训练数据用符号 Exp 表示，做出正确预测，则 Exp=1，反之 Exp=0，则其攻击者优势为

$$\begin{aligned}adv=&Pr(Exp=1)-Pr(Exp=0)=\\&2Pr(Exp=1)-1.\end{aligned} \tag{14}$$

通过计算可得攻击者优势的边界为

$$Adv^{\mathrm{DI}}\leqslant(\mathrm{e}^{\varepsilon}-1)Pr(\mathcal{A}_{\mathrm{DI}}=1|b=0). \tag{15}$$

基于上述量化的指标，可以对 DP-SGD 算法审计. 攻击者进行预测后，可获得其对应的后验概率以及攻击者优势. 后验概率与隐私预算进行转换，即可获得对应的经验隐私下界，转换关系如式（16）所示.

$$\varepsilon'=\ln\left(\frac{\beta_k}{1-\beta_k}\right). \tag{16}$$

攻击者优势通过统计后验概率获得，同样可以与隐私预算进行转换. 攻击者优势在高斯机制下的边界与经验隐私下界之间的转换关系为：

$$\varepsilon'=\sqrt{2\ln(1.25/\delta)}\phi^{-1}\left(\frac{Adv^{\mathrm{DI}}+1}{2}\right).$$

在该审计方法中，所用的攻击者 DI 是比常用的成员推理更强的攻击者，因为 DI 可以获取当前神经网络模型每一步训练的参数以及当前数据集的几乎所有数据. 通过计算攻击得到的后验概率和攻击者优势，并利用这两者与隐私预算之间的关系，可得出更加严格的经验隐私预算下界. 虽然利用该方法审计所得的经验隐私预算下界会相对准确，但需要预先由检验算法的理论隐私预算上界求出后验概率再进行审计. 与其他审计方法相比，该方法求出的经验隐私预算下界相对较为准确，但其设定的攻击者的背景知识足够多，实用性较差.

5.2.2 基于贝叶斯变量的量化

Zanella-Béguelin 等人[61]在中心化机器学习场景下基于贝叶斯估计对 DP-SGD 算法审计. 该方法是第 1 个从贝叶斯估计的角度审计的工作. 这一方法可以大大减少审计过程中所需要训练的数据点个数或所需训练的模型个数. 与其他需要利用对攻击者结果进行统计才能获得的 *FNR* 和 *FPR* 方法不同，Zanella-Béguelin 等人[61]将 *FNR* 和 *FPR* 建模为具有无信息杰弗里先验（non-informative Jeffrey prior）变量的独立 2 项式变量，该变量表达为：

$$\begin{aligned}p&\sim Beta(1/2,1/2),\\k|p&\sim Bin(N,p),\\p|k&\sim Beta(1/2+k,1/2+N-k).\end{aligned} \tag{17}$$

在这一审计方法中，并不需要获得 *FNR* 和 *FPR* 的真实值，也不用对 *FNR* 和 *FPR* 利用 Clopper Pearson 置信区间校正再进行比值计算获取经验隐私预算下界. 只需要利用 *FNR* 和 *FPR* 联合密度函数积分即可求出隐私算法的经验隐私消耗：

$$F_\varepsilon(\varepsilon)=Pr[(FNR,FPR)\in R(\varepsilon,\delta)]\iint_{R(\varepsilon,\delta)}f_{(FNR,FPR)}\mathrm{d}x\mathrm{d}y. \tag{18}$$

计算获取经验隐私消耗后，利用贝叶斯分析中常用的且更严格的杰弗里可信区间（Jeffrey credible interval）进行修正，即可获得经验隐私预算下界的取值区间.

$$\hat{\varepsilon}_-=\arg\max_{\varepsilon}F_{\hat{\varepsilon}}(\varepsilon)\leqslant\alpha/2,\ \hat{\varepsilon}_+=\arg\max_{\varepsilon}F_{\hat{\varepsilon}}(\varepsilon)\geqslant1-\alpha/2. \tag{19}$$

该方法在审计中所需的告密点数量或模型数量远小于频率学派方法中所需的数量. 若利用相同数量的告密点，该方法可以得到更紧密的置信区间.

5.3 小 结

本节主要介绍隐私审计的结果量化方法,这些方法根据所使用的量化方法可细分为基于差分隐私变体的量化和基于贝叶斯的量化,表 7 列出了该类别中各个方法之间的优缺点.

基于 f-DP 等差分隐私定义的量化可以获得更严格的经验隐私预算下界,但其构建过程相对复杂. 基于 RKRDP 定义仅能审计差分隐私中的高斯机制. 基于 Rényi 散度的审计针对在预测阶段加入噪声的隐私算法. 基于差分可辨的量化所得的经验隐私预算下界为所有审计方法中最为严格的,但其计算数据集敏感度时所需生成的中间文件超过 500 GB,计算代价较大. 基于贝叶斯变量的量化方法所需加入的告密点最少,审计代价相对较小,但需要对告密点进行选择.

总体来看,基于差分隐私变体量化的方法(除 RKRDP 外)以及基于差分可辨量化的方法最为严格,基于贝叶斯变量的量化方法效率最高.

6 实 验

隐私审计方法旨在判断当前隐私算法的真实隐私保护程度与其理论上的隐私保护程度之间的差距. 当前隐私审计方法数量较少,且涉及不同场景,本文选取中心化差分隐私算法的审计方法 ML-Audit,利用不同数据集对 DP-SGD 算法进行审计,以说明经验隐私预算下界与理论隐私预算上界之间的关系.

6.1 常用数据集及模型

隐私审计中常用的数据集共有 6 个: Purchase-100, CIFAR10[62], CelebA[63], Sent140[64], MNIST[65], FMNIST.

1)Purchase-100 数据集(简称 P100). Purchase-100 数据集为文本数据集,包含了数千位用户的购物记录. 每个购物记录为 600 个二进制特征,表示用户是否购买了该产品.

2)CIFAR10 数据集. CIFAR10 数据集是常用于机器学习的图像数据集. 该数据集主要由交通工具和动物的图像组成,包括飞机(airplane)、汽车(automobile)、鸟类(bird)、猫(cat)、鹿(deer)、狗(dog)、青蛙(frop)、马(horse)、船(ship)和卡车(truck)共 10 类,每类均包含 6 000 张三通道图像数据.

3)CelebA 数据集. CelebA 数据集是广泛应用于机器学习模型的图像数据集,其中包含超过 20 万张名人的人脸图像数据,每张图像都标注了 40 个属性,例如性别、年龄、发型等.

4)Sent140 数据集. Sent140 数据集为文本数据集,常被用于训练情感分析模型. 该数据集中包含了大量推文文本,这些推文被标记为具有正面情绪、负面情绪和中性情绪. 可用于社交媒体监测、舆情分析和个性化推荐等任务.

5)MNIST 数据集. MNIST 数据集是由美国国家标准与技术研究所收集的包含手写数字的灰度图像的数据集,每张图片为一个 0~9 的手写数字. 其中包含 70 000 张图片,包括 60 000 张训练数据和 10 000 张测试数据. 常被用于训练手写数字识别、邮政编码识别等任务模型的训练.

6)FMNIST 数据集. FMNIST 数据集由 10 类服饰的正面照片组成,包括 T 恤(T-shirt)、裤子(trouser)、套头衫(pullover)、连衣裙(dress)、外套(coat)、凉鞋(sandal)、衬衫(shirt)、运动鞋(sneaker)、包(bag)和踝靴(ankle boot)共 10 类. 包括 70 000 张 28×28 的灰度图像.

隐私审计中常用的神经网络包括 ResNet18[66]、WideResNet[67]、CNN、多层感知机(MLP)等.

ResNet 是 2016 年由 He 等人[66]提出的深度神经网络模型,在此之前所有的神经网络都是由卷积层和池化层进行堆叠组成的. 但仅由这 2 个网络层组成的模型具有梯度消失或梯度爆炸的问题,并且随着神经网络深度的增加,神经网络模型的训练效果会出现变差的现象. 为解决这些问题,ResNet 引入了残差块(residual block)结构,提高了神经网络的训练性能.

Table 7 Methods for Results Quantification of Privacy Auditing

表 7 隐私审计的结果量化方法

类别	方法	审计对象	优点	缺点
基于差分隐私变体的量化	基于 f-DP 的量化	DP-SGD	审计所得经验隐私下界严格	审计构建过程复杂
	基于 LiDP 的量化	DP-SGD	所需样本数量少	加入的告密点较多
	基于 RKRDP 的量化	高斯机制	适用于高维数据	仅能审计高斯机制
	基于 Renyi 散度的量化	PID 框架方法	审计所需训练模型数量少	仅针对于隐私预测接口算法
基于贝叶斯的量化	基于差分可辨的量化	DP-SGD	审计所得经验隐私下界最严格	中间结果空间代价大
	基于贝叶斯变量的量化	DP-SGD	所需样本数量最少	需要选择数据点进行审计

WideResNet 是 ResNet 网络的变体，引入了宽度因子（width factor）的概念，通过增大残差块中的通道数提高神经网络模型的宽度，使模型可以学习更复杂的特征表示.

除此之外，还有 SmallCNN 等结构相对简单的神经网络模型，SmallCNN 中包含 3 个卷积层、1 个池化层以及 2 个全连接层.

6.2 审计方法分析

对差分隐私算法进行隐私审计的工作当前共 15 个，方法汇总如表 8 所示. 这些审计方法通常比较复杂，并且所使用的理论工具差异较大，例如利用 f-DP 审计、利用贝叶斯估计审计等. 目前将方法代码开源的工作比较少，部分开源代码的方法中缺少一些函数或文件，或无法完全得到其论文中展示的结果，一些未开源的审计方法所用理论工具较为复杂，复现具有很大难度. 当前开源并且可完整运行的方法仅有 ML-Audit 方法，但该方法对 DP-SGD 的审计时间较长. 若不采用并行方式运行，较为简单的神经网络 SmallCNN 需 7 天时间才可完整运行，若采用其他更复杂的神经网络训练，例如 ResNet 等，即使采用并行的方式其审计时间也至少为 7 天. 因此当前隐私审计工作难度较大、实验代价较高.

现有的隐私审计方法中，所用的神经网络基本不同，例如 ResNet18[66]、WideResNet[67]、CNN 网络、LSTM 等. 这些神经网络之间参数量相差较大，并且对数据集的训练精度不完全相同. 虽然现有的审计方法均利用实验证明自己的审计方法得到的经验隐私预算下界与之前的方法相比更为严格，但并未排除审计下界与其所用的神经网络之间的关联性.

6.3 审计结果分析

本文所用实验数据集为 FMNIST 和 CIFAR10.为了提高隐私审计过程的训练速度，本文对数据集的采样与其他论文相同，FMNIST 数据集仅采用前 2 个类别的数据，T-shirt 和 trouser. CIFAR10 数据集同样仅采用前 2 个类别的数据，airplane 和 automobile.数据集的信息表如表 9 所示. 实验所用神经网络为 SmallCNN 网络，其中包含 3 个卷积层、1 个池化层以及 2 个全连接层. 隐私审计方法主体为 ML-Audit 方法，其中对数据点的投毒方法为 ClipBKD 方法. 本节的隐私审计实验遵循隐私审计通用流程. 首先利用 ClipBKD 方法构建相邻数据集.

实验中对比 4 种 ε 取值情况，即理论隐私预算上界，分别为 1, 2, 4, 8. 图 9 为隐私审计实验结果图，通过观察可得出以下 3 个结论：1）随着理论隐私预算上界的增加，对模型加噪逐渐变小，模型泄露的信息会逐渐变多，因此经验隐私预算下界也会随之增加. 2）随着隐私预算上界的增加，经验隐私预算下界与理论隐预算上界之间的差距逐渐增大，表明当前算

Table 8 Analysis of Privacy Auditing Methods

表 8 隐私审计方法分析

方法	文献来源	应用场景	审计对象	数据集	训练模型	代码	
						开源	可用性
Jagielski 等人[30]	2020NeurIPS	CML	DP-SGD	FMNIST，P100	LR，FNN	是	可运行
Lu 等人[31]	2022NeurIPS	CML	LR，NB，RF，DP-SGD	Iris，FMNIST 等	SmallCNN	是	可完整运行
Nasr 等人[36]	2021SP	CML	DP-SGD	MNIST，P100 等	CNN	否	
Matsumoto 等人[37]	2022Arxiv	DDC	LDP-SGD	MNIST，SVHN 等	CNN	否	
Arcolezi 等人[40]	2023TPDP	DDC	8 种频数估计算法			否	
Tramer 等人[6]	2022Arxiv	CML	DP-SGD	MNIST	CNN	否	
CANIFE[46]	2023ICLR	DFL	DP-FedAvg	CelebA Sent140 等	ResNet18 2-layerLSTM	是	需手动调大量参数文件
Steinke 等人[47]	2023NeurIPS	CML	DP-SGD	CIFAR-10	WideResNet	否	
one_shot_eps[48]	2023Arxiv	DFL	DP-FedSGD	word prediction data	LSTM	是	需调整代码
Nasr 等人[49]	2023USENIX	CML	DP-SGD	CIFAR-10，P100	WRN-16，ConvNet	否	
Pillutla 等人[28]	2023NeurIPS	CML	DP-SGD	FMNIST，P100	MLP 等	否	
Domingo-Enrich 等人[29]	2023Arxiv	CML	DP-SGD	生成数据		否	
Chadha 等人[50]	2024Arxiv	CML	PDI 框架方法	MNIST，SST2 等	ViT-L/16 等	否	
Bernau 等人[58]	2021PVLDB	CML	DP-SGD	MNIST，Adult 等	NN	是	缺少函数
Zanella-Béguelin 等人[61]	2023PLMR	CML	DP-SGD	CIFAR-10，SST2	CNN	否	

Table 9 Datasets of Experiments

表 9 实验数据集

数据集	维度	训练集数	测试集数	标签数
FMNIST	28×28	60 000	10 000	10
CIFAR10	32×32×3	50 000	10 000	10

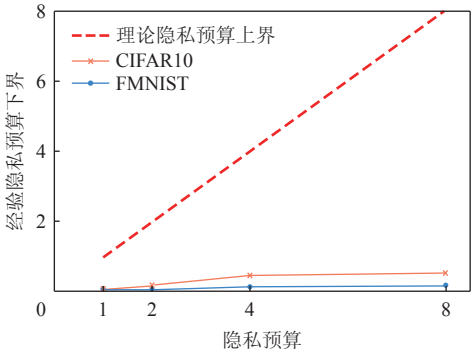


Fig. 9 Results of privacy auditing

图 9 隐私审计结果

法的理论隐私预算上界仍可进行改进. 3)FMNIST 数据集中数据为灰度图像, 与 CIFAR10 数据集中的三通道图像相比, 包含的数据信息相对少. 在加噪相同的情况下, CIFAR10 数据集将会导致更多信息泄露, 所得的经验隐私预算下界更高. 因此在训练模型几乎相同的情况下, 所用数据集越复杂, 消耗的经验隐私预算下界越多.

7 挑战问题

当前对差分隐私算法的审计, 以审计方法的侧重点为依据进行划分, 可划分为隐私审计的数据构建方法、隐私审计的数据测算方法、隐私审计的结果量化方法; 以审计对象为依据, 可划分为对中心化差分隐私算法 DP-SGD 等算法的审计、对联邦学习下差分隐私算法的审计、对本地化差分隐私算法中频数估计算法的审计. 虽然目前的审计工作已涉及大多数隐私定义, 但通常局限于几种常被审计的隐私算法, 并且通常需要训练成千上万个模型, 审计代价较高. 若只训练少量模型进行审计, 审计所得经验隐私预算下界将会不准. 由此, 本节提出现有隐私审计方法面临的主要问题.

7.1 审计结果的正确性

差分隐私算法的理论隐私预算上界在证明过程中关注于在最坏情况下的隐私泄露情况, 而隐私审计方法更关注接近真实攻击者情况下的隐私泄露情况, 因此隐私审计中获得的经验隐私预算下界通常小于理论隐私预算上界. 若经验隐私预算下界高于理论隐私预算上界则认为当前差分隐私算法出错. 在一些隐私审计方法中, 审计人员会选取具有较强背景知识的白盒攻击者作为辨别者以获取更加严格的经验隐私预算下界, 但利用置信区间对统计结果进行校正后其置信区间可能会超出被审计方法的理论隐私预算上界. 虽然这一现象较少, 但目前没有工作对其进行解释说明, 因此无法从这一审计结果中准确判定当前被审计算法的正确与否以及隐私审计结果的正确与否. 对于出现这一现象的隐私审计方法, 需提供理论保证以确保其所得经验隐私预算下界的正确性.

7.2 审计方法的有效性

当前隐私审计方法可以根据其方法侧重点不同划分为 3 类, 这 3 类方法对隐私审计通用流程中的不同步骤分别改进或多个步骤同时进行改进, 例如利用数据构建者构建投毒点或选取具有不同背景知识的辨别者等. 虽然现有审计方法均表明其审计方法与前人方法相比可以获得更加严格的经验隐私预算下界, 但这一结果是基于其隐私审计方法整体获得的, 并未根据审计通用流程分别进行对比, 因此无法判定对隐私审计通用流程中的哪一步进行改进可以获得更加严格的经验隐私预算下界. 例如, 若审计方法对步骤 1 进行改进, 可设置其审计方法中步骤 2 与步骤 3 中所用的策略与其他审计方法完全相同并进行对比. 根据审计通用流程进行分段对比的结果, 可得出改进的方法中最有效的步骤.

7.3 审计方法的统一评估

当前的隐私审计方法的审计场景可划分为 3 类, 由于不同审计场景下对应的隐私算法以及算法训练模式不同, 因此无需进行整体对比. 但同一审计场景下的审计方法大多相同, 例如中心化机器学习场景下, 审计方法普遍关注于 DP-SGD 算法. 但这些审计方法在审计过程中所选用的神经网络并不相同, 例如 CNN, WideResnet 等. 这些神经网络的网络结构以及网络参数数量相差较大, 但在审计过程中并未考虑这些因素对审计结果带来的影响. 除此之外, 现有审计方法所用的数据集并不完全统一. 大多数审计方法关注 MNIST 数据集或 CIFAR10 数据集, 一部分方法选取其他数据集进行审计. 现有审计方法并未在相同数据集下与其他审计方法进行对比. 因此需要设计一个 Benchmark 对审计方法进行统一评估, 以获取审计方法的对比.

7.4 联邦学习隐私审计的性能问题

目前对于联邦学习审计的算法仅有 CANIFE 和

one_shot_eps 这 2 种, 分别对 DP-FedAvg 和 DP-FedSGD 进行审计. 这两者都是从高维空间的角度构建相邻数据集上的分布, 并加以区分获得经验隐私预算下界. 在高维空间中, 若维度足够高则空间中的向量就会相互正交, 联邦学习下的差分隐私审计方法基本都是基于此思想去构建数据点的向量. 从这一角度进行审计通常需要构建较多的客户端进行审计, 但联邦学习方法的训练时间代价往往较大, 因此审计代价通常大于对中心化差分隐私或者本地化差分隐私算法的审计代价. CANIFE 方法虽然可以做到对联邦学习算法暂停并对每一轮训练都进行审计, 但其总体审计时间较长. one_shot_eps 方法针对 CANIFE 中的高维向量问题进行改进, 但该方法在审计过程中需要加入较多的告密点. 因此对于联邦学习下的差分隐私算法, 如何降低审计代价同时提高审计方法性能是需要解决的问题.

7.5 复杂场景差分隐私算法的审计

部分本地化差分隐私算法中用户在本地端对自己的隐私信息进行加噪后再将数据传给数据收集者, 但这类隐私算法并不涉及神经网络等较为复杂的数据处理方式, 并且理论隐私预算上界是通过严谨的数学证明所得, 已经是比较严格的证明, 因此对这类隐私算法审计所得的经验隐私预算下界与理论隐私预算上界十分接近. 但有一部分本地化差分隐私算法针对复杂数据类型[68-69], 例如键值对数据、流数据等, 这类算法的理论隐私预算上界的计算较为简单, 因此并不十分严格. 以键值(key-value)对数据为例, 对这一数据类型加噪的算法中, value 数据的扰动常与键值相关. 若利用现有隐私审计思想, 审计所得的经验隐私下界将大于算法的理论隐私预算上界. 因此隐私审计通用流程无法直接应用于这类差分隐私算法. 如何减小复杂场景差分隐私算法中数据之间的相关性以获取较为严格的经验隐私预算下界, 是对这一类复杂场景差分隐私算法审计的挑战问题.

7.6 差分隐私变体的审计问题

差分隐私技术的实际应用促使研究人员提出了越来越多的差分隐私的变体定义, 例如混洗差分隐私、标签差分隐私[70-77]、量子差分隐私等. 混洗差分隐私通过加入混洗器加强了对隐私的保护, 加入少于本地化差分隐私算法的噪声即可达到本地化差分隐私的保护程度, 模型可用性大大提高. 标签差分隐私只对数据集中的标签信息进行加噪, 将特征视为公开信息, 无需进行加噪处理. 这一隐私定义大大提高了数据的可用性, 并被应用于推荐系统中.

混洗差分隐私和标签差分隐私这 2 种隐私定义虽然已被广泛应用, 但是并没有针对于这 2 类隐私定义的隐私审计方法对其理论隐私预算上界的松弛或算法正误进行验证. 对于标签差分隐私来说, 当前已有工作证明, 从攻击者预测的准确率角度, 其隐私保护能力远低于其声称的隐私保护程度. 但这一结论为理论证明, 与隐私审计思想并不相同. 因此对于差分隐私变体的隐私审计当前仍为空白.

7.7 大模型的隐私审计问题

大模型如 ChatGPT, Claude 等模型已被广泛应用, 但同时伴随着许多隐私问题, 如个人身份信息暴露. 大模型的训练过程较为复杂, 若直接利用 DP-SGD 等差分隐私算法在训练过程中加噪是不现实的. 即使在每一次迭代中都加入噪声, 但由于整体训练迭代次数过多, 最终模型训练结束的隐私预算将会很大. 因此当前已有方法利用隐私数据对公开大模型进行微调(finetune)以实现对隐私数据的保护. 但这些微调后所获得的模型是否真实达到了隐私保护仍有待验证. 若这类微调方法确实实现了隐私保护, 其真实的隐私保护程度仍有待探索. 除此之外, 这一类方法是否浪费了隐私预算也是可审计的问题. 例如在大模型训练的过程中, 某一部分数据是公开的特征, 但在微调过程的数据集中, 这一部分数据被视为隐私数据, 因此并不需要对这一部分数据进行加噪. 若可以用隐私审计思想实现对这一问题的审计, 可提高大模型的隐私保护程度并提高其可用性.

8 结束语

隐私审计是数据治理中的关键问题, 其中差分隐私为实现对用户的个人隐私数据提供了有效的保护. 不同的差分隐私算法逐渐涌现, 但这些隐私算法的正确性与隐私保护的严格程度参差不齐, 因此对差分隐私算法的审计具有重大意义. 本文首先介绍了隐私审计的场景及隐私审计的通用流程, 并将隐私审计方法划分为隐私审计的数据构建方法、隐私审计的数据测算方法、隐私审计的结果量化方法 3 类. 从实验的角度说明了不同数据集以及不同的隐私预算对审计结果的影响. 最后, 结合当前发展情况, 本文提出了 7 个隐私审计的未来研究方向.

作者贡献声明: 许婧楠负责完成实验并撰写和修改论文; 王雷霞负责论文审阅并提出修改意见; 孟小峰负责整体规划并提出指导意见.

参 考 文 献

[1] Archer D W, Pigem B B, Bogdanov D, et al. UN handbook on privacy-preserving computation techniques[J]. arXiv preprint, arXiv: 2301.06167, 2023

[2] Dwork C. Differential privacy: A survey of results [C/OL] //Proc of the 5th Int Conf on Theory and Applications of Models of Computation. Berlin: Springer, 2008[2024-01-24]. https://web.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf

[3] Dwork C, Lei J. Differential privacy and robust statistics[C]//Proc of the 41st Annual ACM Symp on Theory of Computing. New York: ACM, 2009: 371–380

[4] Smith A. Privacy-preserving statistical estimation with optimal convergence rates[C]//Proc of the 43rd Annual ACM Symp on Theory of Computing. New York: ACM, 2011: 813−822

[5] Stevens T, Ngong I C, Darais D, et al. Backpropagation Clipping for deep learning with differential privacy[J]. arXiv preprint, arXiv: 2202.05089, 2022

[6] Tramer F, Terzis A, Steinke T, et al. Debugging differential privacy: A case study for privacy auditing[J]. arXiv preprint, arXiv: 2202.12219, 2022

[7] Baldoni R, Coppa E, D'elia D C, et al. A survey of symbolic execution techniques[J]. ACM Computing Surveys, 2018, 51(3): 1−39

[8] Chen Ruoxi, Jin Haibo, Chen Jinyin, et al. Deep learning testing for reliability: A survey[J]. Journal of Cyber Security, 2024, 9(1): 33−55 (in Chinese)
(陈若曦, 金海波, 陈晋音, 等. 面向深度学习模型的可靠性测试综述[J]. 信息安全学报, 2024, 9(1): 33−55)

[9] Barthe G, Gaboardi M, Arias E J G, et al. Proving differential privacy in Hoare logic[C]//Proc of the 27th Computer Security Foundations Symp. Piscataway, NJ: IEEE, 2014: 411−424

[10] Barthe G, Danezis G, Grégoire B, et al. Verified computational differential privacy with applications to smart metering[C]// Proc of the 26th Computer Security Foundations Symp. Piscataway, NJ: IEEE, 2013: 287−301

[11] Barthe G, Fong N, Gaboardi M, et al. Advanced probabilistic couplings for differential privacy [C] //Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 55−67

[12] Barthe G, Gaboardi M, Grégoire B, et al. Proving differential privacy via probabilistic couplings[C]//Proc of the 31st Annual ACM/IEEE Symp on Logic in Computer Science. New York: ACM, 2016: 749−758

[13] Barthe G, Köpf B, Olmedo F, et al. Probabilistic relational reasoning for differential privacy[C]//Proc of the 39th Annual ACM SIGPLAN-SIGACT Symp on Principles of Programming Languages. New York: ACM, 2012: 97−110

[14] Barthe G, Olmedo F. Beyond differential privacy: Composition theorems and relational logic for *f*-divergences between probabilistic programs[C]//Proc of the 40th Int Colloquium on Automata, Languages, and Programming. Berlin: Springer, 2013: 49−60

[15] Gaboardi M, Haeberlen A, Hsu J, et al. Linear dependent types for differential privacy[C]//Proc of the 40th Annual ACM SIGPLAN-SIGACT Symp on Principles of Programming Languages. New York: ACM, 2013: 357−370

[16] Ding Zeyu, Wang Yuxin, Wang Guanhong, et al. Detecting violations of differential privacy[C]//Proc of the 25th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 475−489

[17] Bichsel B, Gehr T, Drachsler-Cohen D, et al. DP-finder: Finding differential privacy violations by sampling and optimization[C]//Proc of the 25th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 508−524

[18] McSherry F D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis[C]//Proc of the 34th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 19−30

[19] Roy I, Setty S T V, Kilzer A, et al. Airavat: Security and privacy for MapReduce[C]// Proc of the 7th USENIX Conf on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2010: 297−312

[20] Mohan P, Thakurta A, Shi E, et al. GUPT: Privacy preserving data analysis made easy[C]// Proc of the 12th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 349−360

[21] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]// Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308–318

[22] McMahan H B, Daniel R, Kunal T, et al. Learning differentially private recurrent language models[J]. arXiv preprint, arXiv: 1710.06963, 2017

[23] Ye Qingqing, Meng Xiaofeng, Zhu Minjie, et al. Survey on local differential privacy[J]. Journal of Software, 2018, 29(7): 159−183 (in Chinese)
(叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述[J]. 软件学报, 2018, 29(7): 159−183)

[24] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Proc of the 20th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273−1282

[25] Wang Tianhao, Blocki J, Li N, et al. Locally differentially private protocols for frequency estimation[C]// Proc of the 26th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2017: 729−745

[26] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]// Proc of the 37th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3−18

[27] Dong Jinshuo, Aaron R, Su W J, Gaussian differential privacy [J]Journal of the Royal Statistical Society Series B: Statistical Methodology, 2022, 84(1): 3−37

[28] Pillutla K, Andrew G, Kairouz P, et al. Unleashing the power of randomization in auditing differentially private ML[C]//Proc of the 36th Advances in Neural Information Processing Systems. New York: Curran Associates, 2023: 66201−66238

[29] Domingo-Enrich C, Mroueh Y. Auditing differential privacy in high dimensions with the kernel quantum Rényi divergence[J]. arXiv

preprint, arXiv: 2205.13941, 2022

[30] Jagielski M, Ullman J, Oprea A. Auditing differentially private machine learning: How private is private SGD?[C]//Proc of the 33rd Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2020: 22205−22216

[31] Lu F, Munoz J, Fuchs M, et al. A general framework for auditing differentially private machine learning[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 4165−4176

[32] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression[C]//Proc of the 21st Advances in Neural Information Processing Systems. New York: Curran Associates, 2008, 21: 289−296

[33] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization[J]. Journal of Machine Learning Research, 2011, 12(3): 1069−1109

[34] Vaidya J, Shafiq B, Basu A, et al. Differentially private naive Bayes classification[C]//Proc of the 10th IEEE/WIC/ACM Int Joint Conf on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Piscataway, NJ: IEEE, 2013: 571−576

[35] Fletcher S, Islam M Z. Differentially private random decision forests using smooth sensitivity[J]. Expert Systems with Applications, 2017, 78: 16−31

[36] Nasr M, Songi S, Thakurta A, et al. Adversary instantiation: Lower bounds for differentially private machine learning[C]//Proc of the 42nd IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2021: 866−882

[37] Matsumoto M, Takahashi T, Liew S P, et al. Measuring lower bounds of local differential privacy via adversary instantiations in federated learning[J].arXiv preprint, arXiv: 2206.09122, 2022

[38] Erlingsson Ú, Feldman V, Mironov I, et al. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation[J]. arXiv preprint, arXiv: 2001.03618, 2020

[39] Ian J G, Jonathon S, Christian S. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[40] Arcolezi H H, Gambs S. Revealing the true cost of local privacy: An auditing perspective[J]. arXiv preprint, arXiv: 2309.01597, 2023

[41] Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy[C]// Proc of the 33rd Int Conf on Int Conf on Machine Learning. New York: PMLR, 2016: 2436−2444

[42] Wang Shaowei, Huang Liusheng, Wang Pengzhan, et al. Mutual information optimally local private discrete distribution estimation[J]. arXiv preprint, arXiv: 1607.08025, 2016

[43] Ye Min, Barg A. Optimal schemes for discrete distribution estimation under locally differential privacy[J]. IEEE Transaction on Infomation Theory, 2018, 64(8): 5662−5676

[44] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response[C]//Proc of the 21st ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 1054−1067

[45] Bassily R, Smith A. Local, private, efficient protocols for succinct histograms[C]//Proc of the 47th Annual ACM Symp on Theory of Computing. New York: ACM, 2015: 127−135

[46] Maddock S, Sablayrolles A, Stock P. CANIFE: Crafting canaries for empirical privacy measurement in federated learning[J]. arXiv preprint, arXiv: 2210.02912, 2022

[47] Steinke T, Nasr M, Jagielski M. Privacy auditing with one (1) training run[C]//Proc of the 36th Advances in Neural Information Processing Systems. New York: Curran Associates, 2023: 49268−49280

[48] Andrew G, Kairouz P, Oh S, et al. One-shot empirical privacy estimation for federated learning[J]. arXiv preprint, arXiv: 2302.03098, 2023

[49] Nasr M, Hayes J, Steinke T, et al. Tight auditing of differentially private machine learning[C]// Proc of the 32nd USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2023: 1631−1648

[50] Chadha K, Jagielski M, Papernot N, et al. Auditing private prediction[J]. arXiv preprint, arXiv: 2402.09403, 2024

[51] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data[J]. arXiv preprint, arXiv: 1610.05755, 2016

[52] Papernot N, Song Shuang, Mironov I, et al. Scalable private learning with pate[J]. arXiv preprint, arXiv: 1802.08908, 2018

[53] Choquette-Choo C A, Dullerud N, Dziedzic A, et al. CaPC learning: Confidential and private collaborative learning[J]. arXiv preprint, arXiv: 2102.05188, 2021

[54] Duan H, Dziedzic A, Papernot N, et al. Flocks of stochastic parrots: Differentially private prompt learning for large language models[C]//Proc of the 36th Advances in Neural Information Processing Systems. New York: Curran Associates, 2024: 76852−76871

[55] Zhu Yuqing, Yu Xiang, Chandraker M, et al. Private-kNN: Practical differential privacy for computer vision[C]//Proc of the 43rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11854−11862

[56] Dwork C, Feldman V. Privacy-preserving prediction[C] //Proc of the 31st Conf on Learning Theory. New York: Curran Associates, 2018: 1693−1702

[57] Mironov I. Rényi differential privacy[C]// Proc of the 30th IEEE Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2017: 263−275

[58] Bernau D, Eibl G, Grassal P W, et al. Quantifying identifiability to choose and audit $\varepsilon $ in differentially private deep learning[J]. Proceedings of the VLDB Endowment, 2022, 14(13): 3335−3347

[59] Lee J, Clifton C. How much is enough? Choosing ε for differential privacy[C]//Proc of the 14th Int Conf on Information Security. Berlin: Springer, 2011: 325−340

[60] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]// Proc of the 31st IEEE Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2018: 268−282

[61] Zanella-Béguelin S, Wutschitz L, Tople S, et al. Bayesian estimation of differential privacy[C]// Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 40624−40636

[62] Krizhevsky A, Hinton G. Learning multiple layers of features from

tiny images[J/OL]. Handbook of Systemic Autoimmune Diseases, 2009 [2025-01-23]. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[63] Liu Ziwei, Luo Peng, Wang Xiaogang, et al. Deep learning face attributes in the wild[C]//Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 3730−3738

[64] Go A, Bhayani R, Huang Lei. Twitter sentiment classification using distant supervision[R/OL]. 2009[2025-01-23]. https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[65] Cun Y L, Boser B, Denker J, et al. Handwritten digit recognition with a backpropogation network[C]//Proc of the 2nd Advances in Neural Information Processing Systems. New York: Curran Associates, 1989: 396−404

[66] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[67] Zagoruyko S, Nikos K. Wide residual networks[J]. arXiv preprint, arXiv: 1605.07146, 2016

[68] Ye Qingqing, Hu Haibo, Meng Xiaofeng, et al. PrivKV: Key-value data collection with local differential privacy[C]// Proc of the 40th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 317−331

[69] Ye Qinging, Hu Haibo, Li Ninghui, et al. Beyond value perturbation: local differential privacy in the temporal setting[C/OL]//Proc of the IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2021[2025-01-23]. https://drive.google.com/file/d/1ODzRtejolFnKHoZr_cGjQr6E6lMBJeFL/view

[70] Ghazi B, Golowich N, Kumar R, et al. Deep learning with label differential privacy[C]//Proc of the 34th Advances in Neural Information Processing Systems. New York: Curran Associates, 2021: 27131−27145

[71] Malek Esmaeili M, Mironov I, Prasad K, et al. Antipodes of label differential privacy: Pate and alibi[C]//Proc of the 34th Advances in Neural Information Processing Systems. New York: Curran Associates, 2021: 6934−6945

[72] Ghazi B, Kamath P, Kumar R, et al. Regression with label differential privacy[J]. arXiv preprint, arXiv: 2212.06074, 2022

[73] Busa-Fekete R I, Syed U, Vassilvitskii S. On the pitfalls of label differential privacy[EB/OL]. 2021[2025-01-20]. https://openreview.net/forum?id=2sWidqliCDG

[74] Brahmbhatt A, Saket R, Havaldar S, et al. Label differential privacy via aggregation[J]. arXiv preprint, arXiv: 2310.10092, 2023

[75] Esfandiari H, Mirrokni V, Syed U, et al. Label differential privacy via clustering[C/OL]//Proc of the 25th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2022: 7055−7075

[76] Bitau A, Erlingssonú, Maniatis P, et al. Prochlo: Strong privacy for analytics in the crowd [C]//Proc of the 17th ACM Symp on Operating Systems Principles. New York: ACM, 2017: 441−459

[77] Wu Ruihan, Zhou Jinpeng, Weinberger K Q, et al. Does label differential privacy prevent label inference attacks?[C]//Proc of the 26th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2023: 4336−4347



Xu Jingnan, born in 1997. PhD candidiate. Her main research interests include differential privacy and privacy auditing.

许婧楠, 1997 年生. 博士研究生. 主要研究方向为差分隐私、隐私审计.



Wang Leixia, born in 1994. PhD. Her main research interests include secure data collection, and differential privacy and its application.

王雷霞, 1994 年生. 博士. 主要研究方向为安全数据收集、差分隐私及其应用.



Meng Xiaofeng, born in 1964. PhD, professor, PhD supervisor. Fellow of CCF. His main research interests include cloud data management, Web data management, and privacy preserving.

孟小峰, 1964 年生. 博士, 教授, 博士生导师. CCF 会士. 主要研究方向为云数据管理、网络数据管理、隐私保护.