



训练 – 推理协同的图神经网络成员推理防御框架

戚富琪^{1,2}, 全星屹^{1,2}, 高海昌^{1,2*}, 王萍^{1,2}

1. 西安电子科技大学计算机科学与技术学院, 西安 710126

2. 西安市智能系统安全重点实验室, 西安 710126

* 通信作者. E-mail: hchgao@xidian.edu.cn

收稿日期: 2025-05-30; 修回日期: 2025-08-13; 接受日期: 2025-10-13; 网络出版日期: 2025-11-06

国家重点研发计划 (批准号: 2023YFB3107505)、国家自然科学基金 (批准号: 62302371)、国家资助博士后研究人员计划 (批准号: GZC20232035) 和陕西省杰出青年科学基金 (批准号: 2023-JC-JQ-52) 资助项目

摘要 图神经网络因其在结构化数据建模中的优势, 已广泛应用于社交推荐、医疗诊断、金融分析等涉及敏感信息的场景. 然而, 图神经网络在训练过程中对局部结构及节点属性的强依赖, 使其易受到成员推理攻击. 攻击者可通过分析模型对成员与非成员样本的预测差异, 轻易推断样本是否参与训练, 造成严重的隐私泄露. 现有防御策略虽在计算机视觉任务中取得初步成效, 但直接应用于图神经网络时常面临模型性能下降、图结构扰动开销大、隐私保护效果有限等问题. 为此, 本文提出一种训练 – 推理协同的隐私保护框架, 该框架从预测熵角度出发, 提升模型输出分布的混淆性并缓解性能退化. 在训练阶段, 构建基于节点重要性的自适应高熵软标签生成机制, 抑制模型对成员节点的过度自信, 并通过熵正则项进一步惩罚低熵输出并增强隐私混淆效果. 同时, 引入节点级对比学习模块, 提高图神经网络在聚合过程中的判别性表征能力, 从而缓解隐私保护可能引起的模型性能下降. 在推理阶段, 提出基于随机图的输出扰动方法, 利用随机图生成的非隐私输入重构模型输出, 在保持类别排序不变的前提下统一输出置信度, 进一步模糊模型在成员与非成员之间的输出差异. 实验结果表明, 所提方法在多个公开图数据集和图神经网络上实现了更优的隐私 – 效用平衡, 显著提升了对成员推理攻击的防御能力.

关键词 图神经网络, 成员推理攻击, 隐私保护, 图数据, 信息熵

1 引言

图数据是一种以图结构组织的数据形式, 其中节点代表现实中的实体, 边表示实体之间的关系或交互模式. 图结构能够清晰地刻画实体之间的内在关联, 广泛应用于社交网络、学术引用网络以及分子相互作用等领域^[1,2]. 由于图数据结构复杂, 传统的机器学习方法难以有效处理复杂关联, 设计专门的图分析方法至关重要.

图神经网络 (graph neural networks, GNNs) 是一种专门处理图数据的深度学习模型, 通过逐层的信息传递机制和邻居聚合机制实现节点特征的优化^[3~5]. 具体来说, 在 GNNs 中, 节点通过聚合邻居

引用格式: 戚富琪, 全星屹, 高海昌, 等. 训练 – 推理协同的图神经网络成员推理防御框架. 中国科学: 信息科学, 2025, 55: 2852–2869, doi: 10.1360/SSI-2025-0252

Qi F Q, Quan X Y, Gao H C, et al. A training-inference collaborative defense framework against membership inference attacks for graph neural networks. Sci Sin Inform, 2025, 55: 2852–2869, doi: 10.1360/SSI-2025-0252

节点的信息更新自身特征,多层堆叠后能够捕获丰富的上下文信息.根据任务目标的不同,GNNs可用于3种不同层面的任务^[6~8],包括节点级任务(如社交网络的用户兴趣预测、生物学中的分子预测)、边级任务(如推荐系统中的好友推荐、知识图谱中的关系补全)和图级任务(如图分类).因其良好的泛化能力与结构感知能力,GNNs已成为图数据建模的主流方法.

然而,GNNs广泛应用的同时,潜在的隐私安全风险也逐渐凸显.尤其在医疗、金融及社交推荐等涉及敏感信息的领域,GNNs可能记忆或过拟合训练数据,从而致GNNs被成员推理攻击(membership inference attacks, MIAs)^[9]利用,造成隐私泄露.MIAs是一类隐私攻击方法,其核心思想是利用机器学习模型在训练数据(成员数据)和未见数据(非成员数据)上的预测输出差异判断样本是否参与模型训练.当前主流的训练方法使用硬标签监督模型的学习过程,使得模型在训练样本上的预测置信度非常高,而面对非成员时的预测则相对不确定,这种差异易被攻击者利用.与此同时,图数据的结构特性进一步增强了GNNs对此类攻击的脆弱性^[10~13],攻击者可推断节点、边甚至子图是否参与训练,引发严重的隐私问题.例如,在医疗诊断场景下,攻击者可以通过MIAs判断某位患者的病历记录是否参与了疾病预测模型的训练,从而推测患者的疾病状态,严重侵犯患者隐私;在社交推荐系统中,攻击者可以通过MIAs识别用户行为数据的使用情况,从而窃取用户的隐私行为与偏好.

鉴于隐私泄露带来的巨大安全风险,如何有效抵御MIAs成为当前的研究热点.差分隐私(differential privacy, DP)^[14]通过向模型梯度添加噪声的形式限制单个数据对模型输出的影响,但显著降低了模型的准确性,难以有效平衡隐私与模型效用.经验性的防御方法如正则化、模型蒸馏和输出扰动,更关注隐私保护与模型性能之间的权衡.正则化方法^[15]降低成员与非成员数据之间的预测分布差异,提高模型的泛化性,从而降低MIAs的成功率.模型蒸馏方法^[16]通过训练一个学生模型来模仿教师模型的行为,间接减轻模型对训练样本的依赖.输出扰动方法^[17]对模型输出的置信度向量添加精心设计的噪声,干扰攻击者的判断.然而,针对于GNNs的隐私保护缺乏足够关注.现有方法受限于节点之间的高度依赖与图结构的复杂性,没有考虑到GNNs的消息传递机制,难以保证GNNs的分类精度,权衡隐私与效用非常困难.

从信息论的角度来看,成员推理攻击之所以能够成功,其根本原因在于模型的预测输出与样本的成员身份之间存在着不可忽略的互信息(mutual information).换言之,目标模型在成员样本上的输出分布 P_{mem} 与在非成员样本上的输出分布 P_{non} 之间的统计距离与攻击成功率有直接关联,攻击者正是利用这一互信息来推断未知的成员身份.因而,任何能缩小两个分布差异的策略,都会大幅度降低攻击者的成员推理成功概率.形式化上来说,一个理想的防御策略,其核心目标是使成员样本和非成员样本在模型输出空间中的概率分布尽可能相似,即 $P_{\text{mem}} \approx P_{\text{non}}$,这样就能让攻击者的成员推理变为随机猜测.

基于上述研究现状与理论分析,本文提出了一种训练-推理协同的隐私保护框架,从预测熵的角度提升目标模型输出的混淆性,实现隐私保护与模型性能的平衡.在训练阶段,本文依据节点重要性指标、预测熵值以及梯度敏感性自适应分配熵阈值并迭代生成高熵软标签,设计了软标签拟合损失,引导模型拟合出平滑的高熵预测分布.随后,本文设计了熵正则项进一步惩罚过度自信(低熵)的预测输出,并在目标模型的特征聚合过程中引入节点级对比学习模块,增强特征向量的判别性,从而有效缓解隐私保护对模型精度的影响.在推理阶段,提出一种基于随机图的输出扰动机制,利用非隐私输入统一模型输出置信度,进一步增加辨别成员样本的难度.本文的主要贡献如下:

- 提出了自适应高熵软标签生成机制,通过节点重要性动态调整熵阈值并迭代生成软标签,设计了软标签拟合损失,引导模型生成高熵输出,避免对训练样本的过度自信预测;
- 设计了熵正则项与节点级对比学习模块,进一步惩罚低熵输出,提高GNNs的节点表征能力,减少隐私保护对性能的影响;
- 构建了基于随机图的推理阶段输出扰动模式,统一成员与非成员样本的输出分布,强化隐私保护;
- 广泛实验表明,提出的防御框架在有效保证模型性能的同时,显著提高了对MIAs的防御效果.

2 相关工作

2.1 成员推理攻击

成员推理攻击 (MIAs) 最早由 Shokri 等人^[9] 系统提出, 其核心为影子模型技术, 通过训练多个结构相似的影子模型获取成员与非成员数据的预测分布差异. Yeom 等人^[18] 深入分析过拟合与隐私泄露的关系, 指出泛化误差与攻击成功率之间存在紧密联系: 模型的泛化误差越大, 过拟合程度越严重, 启发了后续攻击方法及防御研究.

根据攻击者对目标模型的访问程度, MIAs 可分为黑盒攻击与白盒攻击. 在黑盒攻击场景下, 攻击者仅能访问目标模型的预测输出, 而无法获取其内部结构或训练细节. Salem 等人^[15] 提出了在未知模型结构和训练集分布下的攻击策略, 丰富了黑盒场景, 提高了攻击的泛化性. Song 等人^[19] 通过比较预测指标与设定阈值来区分成员样本与非成员样本. 不同于传统方法对详细数据的依赖, 标签攻击方法^[20~23] 在仅硬标签可知的条件下评估了模型的隐私风险. 这些黑盒攻击强调, 即使访问权限有限, 深度学习模型的隐私风险仍广泛存在. 相比之下, 白盒攻击场景提供了更加宽松的访问权限^[24], 攻击者可以自由访问模型结构、参数和训练过程甚至是训练数据. Nasr 等人^[25] 利用梯度范数差异实现攻击, Leino 等人^[26] 则通过模型参数差异捕捉模型对训练数据的记忆模式.

随着 GNNs 的广泛应用, MIAs 研究逐步拓展至图任务上. 针对图数据进行适应性调整后, Duddu 等人^[10] 使用传统基于阈值和基于影子模型的 MIAs 揭露了 GNNs 的隐私安全问题. Olatunji 等人^[11] 探索了孤立/非孤立训练图下的隐私安全, 并分析了数据集属性和模型架构对隐私泄露程度的影响. He 等人^[12] 利用 0-hop 与 2-hop 邻居节点的最大后验概率值来查询攻击模型, 以判定获取节点的成员身份. 针对输出结果仅包含标签的情况, Conti 等人^[13] 融合节点自身特征与多邻居节点训练攻击模型, 提升了攻击效果. 在此基础上, 近期的研究更加关注攻击特征的可区分性与结构信息的利用. Xie 等人^[27] 提出了一种基于因果推断的模型校准技术, 其核心在于强化输入特征与预测标签之间真实的因果联系, 抑制 GNN 训练中产生的非因果关联干扰. Guan 等人^[28] 意识到图结构特征在成员推理攻击中的重要性, 通过显式提取节点局部子图的拓扑特征并将其与 GNN 的预测输出结合, 构造出增强的攻击特征. Guan 等人^[29] 还提出了基于注意力机制的攻击模型, 对目标节点及其邻居节点的后验概率进行自适应聚合, 通过对相邻节点动态分配权重, 将关键的局部拓扑信息有效地编码到攻击特征中. 这些工作暴露了 GNNs 面对 MIA 的脆弱性, 也推动了相关防御措施的研究.

2.2 成员推理防御

为了缓解 MIAs 带来的隐私泄露风险, 研究者提出了多类防御策略, 主要通过降低目标模型对训练样本的过拟合程度来实现隐私增强.

一类是基于差分隐私的防御方法, 通过在训练过程中引入可控噪声, 限制模型泄露关于训练数据中任何单个记录的信息, 为模型提供理论可证明的隐私保障. Rahimian 等人^[30] 通过梯度裁剪与噪声注入的方式控制隐私泄露, 缓解模型对单个样本的敏感性. Chen 等人^[14] 发现了隐私预算与模型精度的对数关系, 指出低隐私预算会造成效用急剧下降. 针对图任务, Olatunji 等人^[31] 提出的防御方法结合了泊松采样与 K 近邻子图构建, 他们发现轻微的隐私预算提升就会引发显著的模型性能下降. Sajadmanesh 等人^[32] 在 GNNs 的聚合阶段引入随机噪声以保护图结构隐私. 在文献 [31, 32] 的基础上, Xiang 等人^[33] 提出了异构泊松子图采样与多元噪声扰动机制, 提高了节点级隐私保护的效果. 需要注意的是, DP 方法普遍面临隐私保护与模型效用的权衡难题, 模型性能难以保障.

另一类经验性防御方法关注于隐私保护和模型性能的平衡. 诸如 L1/L2 正则化、Dropout 等经典正则化方法^[15], 能有效提升模型泛化性, 从而缓解过拟合带来的隐私泄露风险. Nasr 等人^[34] 进一步提出了对抗性正则化, 将防御过程建模为隐私博弈, 最小化分类误差与攻击准确率的加权目标. 其次, 基于扰动的防御方法通过向模型输出层^[35] 或中间层^[17] 中注入噪声, 扰乱攻击者的判别边界, 从而

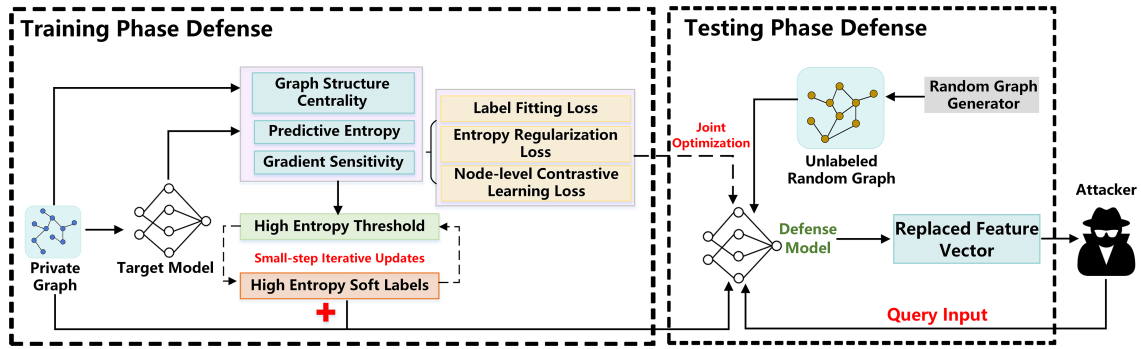


图 1 (网络版彩图) 训练 – 推理协同的隐私保护框架。

Figure 1 (Color online) A privacy-preserving framework for training-inferencing synergy.

保护隐私. 为了适用于图数据和 GNNs, 很多工作提出了针对性的扰动防御策略. He 等人^[12] 提出了随机边添加策略, 其对拓扑结构的严重破坏导致模型可用性变差. Olatunji 等人^[11] 关注于 GNNs 的聚合过程, 通过限制目标节点的邻居数量来抑制信息传播, 以此保护数据隐私. Chen 等人^[36] 使用遮蔽策略动态修改输出的置信度输出, 有效提高了输出的模糊性, 但防御效果因攻击模式和数据集的不同而存在明显差异. Guan 等人^[37] 对图的拓扑结构进行了扰动, 对节点特征进行优化实现隐私保护. Wang 等人^[38] 基于图自编码器将图结构映射到隐空间并引入高斯噪声实现防御, 但所需的高强度噪声会损害图结构信息, 导致模型性能下降. Wu 等人^[39] 提出 GraphGuard 方法, 在无需访问原始训练数据的前提下, 通过引入放射性数据检测对训练数据的过拟合, 并利用合成图进行遗忘操作以缓解过拟合的影响. 知识蒸馏提供了一种简洁的隐私保护策略, 学生模型仅模仿教师模型的预测行为, 而不需要访问教师模型的私有数据集. DMP^[16] 方法选用低熵且远离边界的非隐私数据作为学生训练集, 并利用教师模型对这些数据进行标注. Zheng 等人^[40] 通过共享预训练参数减少计算开销, 提高了蒸馏的效率. AKD^[41] 融合对抗性训练与蒸馏策略, 使学生模型在保持性能的同时减少对私有训练数据的依赖. 需要注意的是, 在蒸馏过程中, 图结构信息中的隐私仍有被泄露的风险.

3 面向 GNNs 的训练 – 推理协同隐私保护框架

为提升 GNNs 在面对 MIA 时的隐私鲁棒性, 更好地实现隐私保护与模型效用之间的权衡, 本文提出一种训练 – 推理协同的防御方法, 系统性地地在模型训练与推理两个阶段施加针对性的隐私保护. 在训练阶段, 本文设计了一种结构 – 行为联合驱动的更新策略, 自适应地为节点更新动态熵阈值与高熵软标签, 并使用软标签拟合损失引导模型输出高熵分布. 本文使用熵正则项进一步限制模型的过度自信预测, 并设计了节点级对比学习模块增强节点表征的判别能力. 在推理阶段, 为进一步统一模型对成员节点与非成员节点的输出分布, 本文引入基于随机图的输出扰动机制, 对模型最终输出进行结构性替换. 图 1 展示了完整防御框架的示意图, 以下将分别对各模块进行详细说明.

本框架的理论核心在于通过训练与推理阶段的协同作用, 系统性地降低目标模型在成员样本与非成员样本上的输出差异. 如图 1 所示, 协同框架通过两个阶段实现这一目标. 在训练阶段, 防御框架提高预测熵, 缩小分布差异. 理想情况下, 成员样本的预测熵较低 (模型较为自信), 非成员样本的预测熵较高 (模型不确定). 训练阶段的目标是通过自适应高熵软标签和熵正则化显著提升成员样本的预测熵, 直接缩小成员与非成员在模型内在预测模式上的差异. 仅靠训练阶段的防御可能仍会暴露出微弱的分布差异, 推理阶段防御通过基于随机图的输出扰动, 将所有测试样本的置信度输出分布强制统一到一个高熵、标准化的状态, 使得攻击者几乎无法从预测输出中提取有区分性的信息.

3.1 训练阶段防御

训练阶段的防御旨在保证模型效应的前提下,削弱模型对成员样本的过拟合与输出置信度的偏置. 为了实现该效果,本文设计了自适应高熵软标签生成模块,使用高熵软标签代替硬标签参与模型的训练,通过熵正则化进一步惩罚低熵(高置信度)输出,并利用节点级对比学习模块保证模型的预测精度.

3.1.1 自适应高熵软标签生成模块

在传统的输出模式下, GNNs 的预测输出为硬标签. 形式化来说,对于节点 v_i ,模型的预测输出为 one-hot 向量,即正确类别的预测值为 1, $p_i[c = c_t] = 1$,其余所有类别的预测输出值为 0,即 $p_i[c \neq c_t] = 0$. 自适应高熵软标签生成模块使用迭代更新策略,每步迭代使用小步长 η 压低正确类别的预测输出值,即 $p_{c_t} \leftarrow p_{c_t} - \eta$,并均匀提升其余 $C - 1$ 个类别的预测输出值 $p_{c_o} \leftarrow p_{c_o} + \frac{\eta}{C-1}$.

预测熵 $H(\cdot)$ 表示了模型预测的不确定性,熵值越大,模型预测的不确定越大. 给定目标模型 $F(\cdot)$,模型在节点 v_i 上的预测熵表示为

$$H(F(v_i)) = - \sum_{c=1}^C p_c \log p_c, \quad (1)$$

其中, p_c 表示模型在该节点上关于类别 c 的预测置信度输出. 一般情况下,低熵预测输出表示模型对预测结果非常自信,高熵输出表示模型对预测结果非常不自信. 如文献 [42] 证明,当模型在每个类别上的预测输出完全相同即每个类别的概率为 $\frac{1}{C}$ 时,预测熵为最大熵值.

$$H_{\max} = -C \cdot \left(\frac{1}{C} \cdot \log \frac{1}{C} \right) = \log C. \quad (2)$$

最大熵对应于最不确定的预测能力,也对应于最强的节点隐私性. 但与此同时,目标模型丧失了正常的分类能力. 因而,高熵软标签生成不能直接以最大熵为优化目标.

为此,本方法依据节点重要性为节点自适应分配熵阈值 κ_i ,调节高熵软标签生成的熵优化目标,以保证模型分类性能. 给定目标模型 $F(\cdot)$ 和训练集 D ,在对于训练集合中的任一样本 v_i ,其节点特征向量表示为 x_i . 本方法基于节点重要性自适应分配熵阈值,重要性评估指标包括图结构中心性 I_i (如度中心性、介数中心性、特征向量中心性)、节点的当前预测熵 $H(F(v_i))$ 和节点的梯度敏感度 $g_i^n = \|\nabla_{v_i} L^{n-1}\|$. 其中,预测熵值衡量了模型预测的不确定性;而梯度敏感性衡量了上一迭代轮次中节点对模型优化的影响程度,由梯度范数表示,为标量. 给定熵阈值取值范围 $\kappa_i \in [\kappa_{\min}, \kappa_{\max}]$,本文使用非线性映射网络 $f_\phi(\cdot)$ 为节点分配自适应熵阈值:

$$\kappa_i = f_\phi(z_i) = \text{sigmoid}(Wz_i + b) \cdot (\kappa_{\max} - \kappa_{\min}) + \kappa_{\min}, \quad (3)$$

其中, z_i 表示节点 v_i 的节点重要性向量,由图结构中心指标值 I_i 、节点熵值 $H(F(v_i))$ 和梯度敏感性直接拼接得到,拼接前对 3 部分进行了归一化. 如式 (3) 所示,映射网络由一个轻量级的单层感知机构成,包含一层线性层、Sigmoid 激活以及用于控制输出范围的线性放缩. 该网络生成得到的阈值用于接下来的高熵软标签生成.

接着,本模块按照小步长 η 迭代更新软标签 y' . 如式 (4) 所示,一旦达到调节后的最大熵则停止迭代:

$$H(y') = -p_t \log p_t - (1 - p_t) \log \frac{1 - p_t}{C - 1} \geq \kappa_i \cdot \log C. \quad (4)$$

为了在训练过程中引导目标生成类似的高熵输出,本方法设计了高熵软标签拟合损失 $L_{\text{soft-label}}$,通过 KL 散度将模型的输出分布拟合至生成的高熵输出分布:

$$L_{\text{soft-label}} = \sum_{v_i \in D} \text{KL}(F(v_i) || y'_i). \quad (5)$$

3.1.2 熵正则化模块

为了进一步避免低熵输出, 本方法设计了熵正则项 L_{reg} , 以增强高熵软标签对模型训练的监督能力.

$$L_{\text{reg}} = - \sum_{v_i \in D} H(F(v_i)) = \sum_{v_i \in D} \sum_{j=1}^C p_{ij} \log p_{ij}, \quad (6)$$

其中, p_{ij} 表示模型面对样本 v_i 时在类别 j 上的预测置信度. 该正则项通过对模型预测熵的直接约束, 惩罚在经过高熵软标签修改的节点上的低熵预测行为, 协同高熵软标签的训练目标, 引导模型输出更加平滑的预测概率分布, 破坏成员样本上稳定的置信度结构, 从而提升隐私保护效果.

3.1.3 节点级对比学习模块

如何权衡隐私保护与模型效用是困扰所有防御工作的一大难题. 传统方法通常会引入扰动或不约束来降低模型对成员样本的过拟合程度, 这不可避免地牺牲了模型在下游任务上的预测性能. 尤其是面对结构复杂且类别数量不均衡的图数据时, 这种缺陷会更为明显. 此外, 本文设计的高熵软标签会带来额外的副作用 —— GNNs 在节点聚合过程中的表征能力下降, 导致模型判别节点类别的能力下降.

为此, 本文设计了节点级对比学习模块, 通过强化隐层空间的判别性, 使模型在特征聚合过程中依然保持良好的类内聚合与类间分离效果, 从而补偿防御策略引发的模型性能下降问题. 该模块的核心思想在于: 在训练阶段构造正负样本对, 引导模型学习区分度更高的节点特征表示.

具体来说, 对于任一节点 v_i , 该节点在目标模型的第 l 层的特征向量表示为 h_i^l . 从训练集中筛选出与节点 v_i 标签一致的节点构成正样本集合 P_i , 其余异类节点构成负样本集合 N_i . 在训练过程中, 采用如下对比学习损失函数监督训练:

$$L_{\text{contrast}} = \sum_{i \in D_m} \frac{1}{|P_i|} \sum_{p \in P_i} -\log \frac{\exp(\text{sim}(h_i^l, h_p^l)/\tau)}{\sum_{a \in P_i \cup N_i} \exp(\text{sim}(h_i^l, h_a)/\tau)}, \quad (7)$$

$$\text{sim}(h_i^l, h_p^l) = \frac{(h_i^l)^T h_p^l}{\|h_i^l\| \cdot \|h_p^l\|}, \quad (8)$$

其中, $\text{sim}(h_i^l, h_j^l)$ 表示节点 v_i 与 v_j 在特征聚合层的余弦相似度, $\tau > 0$ 为温度系数, 通过调节分布的陡峭程度从而控制模型对正负样本相似性差异的敏感性. 较小的温度会放大高相似度节点对之间差异, 引导模型更加关注“真正相似”的节点组合, 从而有利于形成紧密聚合的类别簇. 相反, 较大的温度会平滑分布, 增强训练的稳定性与泛化能力. 该机制为特征表征层带来更具区分性的几何结构, 弥补高熵标签在预测层的模糊化效应.

综上所述, 目标模型的完整防御训练由联合损失函数引导:

$$L_{\text{train}} = L_{\text{soft-label}} + \lambda_1 L_{\text{reg}} + \lambda_2 L_{\text{contrast}}, \quad (9)$$

其中, λ_1 和 λ_2 为控制损失权重的超参数. 需要注意的是, 在自适应高熵软标签模块中, 梯度敏感性的计算仅包含 $L_{\text{soft-label}}$ 和 L_{reg} 两项.

3.2 推理阶段防御

尽管训练阶段已通过高熵软标签和熵正则手段削弱了模型对成员样本的过度自信输出, 但在实际部署时, 攻击者仍可能通过推理阶段的置信度分布差异实施成员推理攻击. 为进一步增强模型在部署阶段的隐私鲁棒性, 本文设计了一种基于随机图驱动的输出扰动策略, 统一模型在成员与非成员节点

表 1 数据集统计.
Table 1 Dataset statistics.

Dataset	Nodes	Edges	Features	Classes
Cora	2708	10556	1433	7
CiteSeer	3327	9104	3703	6
PubMed	19717	88648	500	3

上的输出置信度表现, 从而降低攻击者在推理过程窃取成员隐私的可能性. 推理阶段防御由两个关键步骤构成: 构造无标签的随机图输入和置信度输出替换.

首先, 为尽可能脱离真实图数据, 本文引入自定义生成的无标签随机图样本作为输出扰动的驱动源. 随机图的生成需同时包含拓扑结构与节点特征两部分. 本方法采用经典的 Erdos-Rényi 随机图模型^[43]构造图结构, 其中每对节点以独立概率 p 构建随机图的边. 通过调节 p 的取值, 可灵活控制图的稀疏性与连接模式, 以满足不同复杂度要求. 考虑图神经网络对输入特征的高度敏感性, 本文提供两种生成方式提高输入的随机性: 其一, 从标准正态分布中随机抽取, 构造与原始图形状相同的特征矩阵; 其二, 利用 Mixup 策略对真实图中多个节点的特征向量进行线性组合, 生成分布未知的新特征.

接着, 利用目标模型在这些生成样本的输出构造输出替换策略. 由于随机样本远离训练样本的分布, 模型通常会给出较低的置信度输出. 因此, 对于任一推理阶段的输入样本 v_i , 可以随机构造出一个扰动样本 v_i^{rand} , 将模型的原始预测 $F(v_i)$ 的概率预测值替换为 $F(\tilde{v}_i)$ 的预测值. 如式 (10) 所示, 替换过程保持类别的预测排序顺序不变

$$\arg \text{sort}(F(\tilde{v}_i)) = \arg \text{sort}(F(v_i)), F(\tilde{v}_i) \leftarrow \text{sort_values}(F(v_i^{\text{rand}})). \quad (10)$$

该模块在确保模型分类攻击的前提下, 统一成员样本与非成员样本的置信度分布水平, 使攻击者难以通过推理阶段的置信度输出差异进行成员样本识别, 从而有效提升模型的推理阶段的隐私鲁棒性.

整体而言, 训练阶段防御侧重于通过高熵软标签与熵正则项有效降低模型对训练数据的过拟合程度, 增强模型在训练样本上的不确定性, 推理阶段防御则着眼于消除模型输出的可区分性, 模糊成员与非成员样本在模型输出上的差异. 这种训练与推理阶段的协同防御策略, 确保了模型在整体流程中始终保持较强的隐私保护能力. 防御方法的完整流程如算法 1 所示.

4 实验与分析

4.1 实验设置

数据集. 本文在 3 个常见的图数据集 Cora^[44], CiteSeer^[45], PubMed^[46] 上进行防御测试, 并对数据集进行了划分, 一部分作为目标模型的私有数据集参与训练, 另一部分作为测试集验证防御效果. 数据集的详细信息如表 1 所示.

模型. 本文使用了 3 种常用的 GNN 模型作为目标模型, 包括图卷积网络 (GCN)^[3]、GraphSAGE^[4]、图注意力网络 (GAT)^[5]. 每个目标模型和影子模型的层数设置为 2 层, GCN 和 GraphSAGE 的隐藏层维度为 16, GAT 的隐藏层维度为 64. 每个模型训练 200 epoch. 在防御的训练阶段中, 3 个模型采取了不同的学习率, 其中 GCN 的学习率为 0.01, GraphSAGE 的学习率为 0.05, GAT 的学习率为 0.005.

对比方法. 为了评估所提方法的有效性, 论文选择了 4 种 SOTA 的防御方法, 包括 LBP^[11], DMP^[16], AdjMask^[36], GraphGuard^[39]. LBP 在后验概率上添加拉普拉斯 (Laplace) 噪声, 先进行分区, 对同一区域添加相同噪声. 实验中分区数设为 2. DMP 采用了经典的知识蒸馏策略, 使用无保护模型生成软标签, 通过 KL 散度将知识迁移到保护模型. AdjMask 通过邻居节点的置信度加权重构目

算法 1 训练 – 推理协同隐私保护框架.

输入: 训练图数据集 D , GNN 模型 $\mathcal{F}$, 高熵软标签修改步长 δ , 温度系数 τ , 随机图边生成概率 p .

主训练过程:

```

1: 初始化模型参数  $\theta$ ;
2: while 未达到收敛条件 do
3:   计算每个节点的重要性向量  $z_i$ , 由结构中心性、预测熵、梯度敏感度拼接组成;
4:   利用非线性映射网络  $f_\phi$  生成自适应熵阈值  $\kappa_i = f_\phi(z_i)$ ;
5:   for 每个训练节点  $v_i \in D$  do
6:     初始化软标签  $\mathbf{y}'_i \leftarrow \text{one-hot}(\text{label}(v_i))$ ;
7:     while  $H(\mathbf{y}'_i) < \kappa_i \cdot \log C$  do
8:        $\mathbf{y}'_i[c_t] \leftarrow \mathbf{y}'_i[c_t] - \delta$ ;
9:       其他类别分量均匀增加  $\frac{\delta}{C-1}$ , 保持归一性;
10:    end while
11:  end for
12:  计算软标签拟合损失:  $\mathcal{L}_{\text{soft}} = \sum_{v_i \in D} \text{KL}(F(v_i) || \mathbf{y}'_i)$ ;
13:  计算熵正则项:  $\mathcal{L}_{\text{reg}} = - \sum_{v_i \in D} \sum_{j=1}^C p_{ij} \log p_{ij}$ ;
14:  构造正负样本对, 计算节点级对比损失 (余弦相似度):
15:   $\mathcal{L}_{\text{contrast}} = \sum_i \frac{1}{|P_i|} \sum_{p \in P_i} -\log \frac{\exp(\text{sim}(h_i^l, h_p^l)/\tau)}{\sum_{a \in P_i \cup N_i} \exp(\text{sim}(h_i^l, h_a^l)/\tau)}$ ;
16:  总损失为  $\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{soft}} + \lambda_1 \mathcal{L}_{\text{reg}} + \lambda_2 \mathcal{L}_{\text{contrast}}$ ;
17:  更新模型参数  $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\text{train}}$ ;
18: end while
    
```

输出: 训练完成的隐私保护模型参数 θ ;

19: **推理阶段防御**

```

20: 构造随机图  $\tilde{G}$ : 结构来自 ER 模型 (边生成概率  $p$ ), 节点特征来自正态分布或 Mixup 合成;
21: 用  $\mathcal{M}$  推理得到随机输出  $\tilde{\mathbf{y}} = \mathcal{M}(\tilde{G})$ ;
22: for 每个推理节点  $x$  do
23:   获取  $\mathcal{M}(x)$  的预测排序  $\pi = \text{argsort}(\mathcal{M}(x))$ ;
24:   用  $\tilde{\mathbf{y}}$  中的置信度值, 按照  $\pi$  重排序后替换原始输出;
25: end for
    
```

输出: 扰动后的最终推理结果.

标节点输出, 掩盖单节点特征. GraphGuard 利用基于节点属性生成的合成图微调模型, 最大化遗忘样本的损失, 无需原始训练图. 对于本文提出的协同防御框架, $\lambda_1 = 0.9$, $\lambda_2 = 0.01$, 对比学习的温度系数 $\tau = 1$, 推理阶段的构造随机图的边概率为 $p = 0.001$. 与此同时, 防御框架中的 mixup 策略每次随机从训练集与验证集的合集 (与测试集完全正交) 中随机采样两个节点, 并按照 0.5 的比例进行混合得到新节点.

攻击测试. 本文主要针对于黑盒场景下的成员推理攻击, 并在黑盒场景下测试方法的防御效果. 本文使用基于影子模型的 SOTA 黑盒攻击方法^[15] 测试各种方法的防御效果. 攻击者使用与目标模型结构相似的模型作为影子模型, 并精心构建影子模型的训练集, 使其特征分布近似于目标模型的私有数据集. 需要注意的是, 攻击者为上述 3 种目标模型分别构造影子模型, 影子模型与其对应的目标模型具有相同的模型架构. 此外, 攻击模型包含了 3 层 MLP 网络. 在攻击测试过程中, 攻击者为测试集的每个类别单独训练对应的攻击模型, 并对每个类别的数据分别进行推理攻击.

评价指标. 为了评估各种方法的防御效果, 本文采用目标模型的分类准确率 Acc 和成员推理攻击测试的攻击成功率 (ASR) 作为评估防御性能的指标. Acc 量化了防御措施对分类准确率的损害程度, 用于评估防御机制对模型性能的影响. ASR 量化了 MIAs 成功确定目标模型成员状态的可能性, ASR 越高表示攻击者区分成员数据与非成员数据的能力越强, 模型的隐私泄露风险越大. 防御方法的目标是使攻击模型的成功率趋近于 50% (随机猜测水平), 此时攻击模型无法区分成员与非成员, 从而实现

表 2 受保护模型在面对黑盒 MIAs 时的平均攻击成功率与最高攻击成功率 (%). 加粗数值表示在对应数据集 – 模型组合下的最佳防御效果 (包括平均攻击成功率与最高攻击成功率).

Table 2 Avg-ASR and Max-ASR of protected models facing blackbox MIAs (%). Bold values indicate the best defense performance under the corresponding dataset-model combination, including both Avg-ASR and Max-ASR.

Model	Dataset	Method					
		No-defense	DMP	LBP	AdjMask	GraphGuard	Ours
GCN	Cora	71.07/80.56	50.00 /51.25	65.54/76.19	59.59/68.57	53.98/57.92	50.00/50.00
	CiteSeer	74.65/90.38	49.58/50.00	58.33/66.25	68.81/90.38	50.68/54.43	50.00/ 50.00
	PubMed	74.17/80.00	51.67/55.00	63.33/76.25	61.67/65.00	55.17/58.94	50.00/50.00
GraphSAGE	Cora	76.76/83.33	51.74/58.75	58.00/69.44	65.07/75.50	51.45/55.37	50.00/50.00
	CiteSeer	79.98/96.15	49.58/50.00	69.26/86.25	67.45/88.46	50.68/53.90	50.00/ 50.00
	PubMed	69.17/77.75	52.92/58.75	69.17/75.00	62.08/63.75	60.68/63.31	50.00/50.00
GAT	Cora	71.92/80.56	50.00/ 50.00	62.06/83.33	59.09/74.29	48.19 /51.68	50.00/ 50.00
	CiteSeer	80.69/90.38	50.00/50.00	53.46/63.75	70.79/86.47	50.68/53.67	50.00/50.00
	PubMed	66.67/70.00	50.00/50.00	54.58/60.00	61.67/66.25	45.91/47.86	50.00/50.00

最优的隐私保护效果. 由于攻击者针对数据集中的每一个类别训练攻击模型, 本文统计了平均攻击成功率 (Avg-ASR) 和最高攻击成功率 (Max-ASR), Avg-ASR 计算了所有类别上的平均攻击成功率, 评估了攻击与防御的泛化性, 而 Max-ASR 反应了隐私风险最高的类别的攻击成功率, 评估了极端情况下的防御性能.

4.2 实验结果与分析

为了全面评估防御方法的有效性, 论文在 3 个模型和 3 个数据集上进行了防御测试, 并与 4 种现有先进的防御方法进行了对比. 表 2 统计了目标模型在防御前和防御后受到成员推理攻击的成功率, 攻击成功率包含了平均攻击成功率与最高攻击成功率, 分别反映了防御的泛化能力与应对极端攻击情况的能力, 该表反映了防御的稳定性与鲁棒性. 表 3 统计了目标模型在防御前后的分类精度, 反映了防御方法在隐私与效用之间的权衡能力.

实验结果表明, 在没有任何保护措施的情况下, 3 种模型均面临严重的隐私风险, Avg-ASR 普遍超过 70%, Max-ASR 甚至达到 96.15%, 验证了 GNNs 对成员推理攻击的高度脆弱性. 此外, 模型类型与数据集特征对攻击成功率具有显著影响. 例如, 在 CiteSeer 数据集上, GCN 模型的 Max-ASR 最高, 主要由于该数据集稀疏且边结构不稳定, 导致模型更依赖节点本身特征, 从而加剧过拟合.

现有防御方法虽然在部分场景下能有效降低攻击成功率, 但普遍存在防御效果不稳定的缺陷, 在面对敏感性强的极端类别时防御效果非常有限. LBP 在 Cora 数据集上仍有超过 65% 的 Avg-ASR, 表明其在稀疏图中的有效性有限. AdjMask 在 CiteSeer 上出现高达 90.38% 的 Max-ASR, 其在低连通图结构下对重要节点的隐私保护非常有限. 类似地, GraphGuard 虽然能够显著降低大部分节点的隐私泄露风险, 但对重要节点的平滑效果有限.

相比之下, 本文提出的协同防御机制在所有模型 – 数据集组合下均显著优于现有方法. 协同防御框架具有稳定性强的优势, Avg-ASR 与 Max-ASR 被稳定控制在 50% 左右的水平, 使成员推理接近随机猜测. 在节点规模较大、信息聚合能力强的 PubMed 数据集中, 本文方法在 3 种模型上实现了一致的防御效果, 充分体现了泛化能力强的优势. 这种防御优势源自于本文的协同设计. 本文在训练阶段引入自适应高熵软标签与熵正则项有效抑制模型对成员样本的过度置信输出, 尤其是在推理阶段, 提出构造无标签随机图并进行排序保持的概率重构, 统一成员与非成员节点的输出置信分布, 从而从源头破坏攻击者基于预测置信度差异实施推断的可行性.

另一方面, 不同防御方法对模型性能的影响差异显著. DMP 与 LBP 方法在大多数设置下大幅度

表 3 防御保护下的模型分类准确率 (%).
Table 3 Classification accuracy of protected models (%).

GNN	Dataset	Method					
		No-defense	DMP	LBP	AdjMask	GraphGuard	Ours
GCN	Cora	73.99	70.38	78.39	79.52	78.81	77.66
	CiteSeer	71.34	68.03	71.34	72.43	73.03	74.03
	PubMed	77.70	79.06	76.94	75.62	80.45	78.67
GraphSAGE	Cora	74.36	60.83	75.00	79.15	83.95	71.43
	CiteSeer	70.45	59.24	69.85	71.59	76.02	69.25
	PubMed	75.77	67.71	76.01	75.97	85.63	79.69
GAT	Cora	76.92	51.29	76.56	79.88	81.01	78.39
	CiteSeer	70.75	67.13	71.04	71.63	89.25	70.45
	PubMed	79.42	74.56	77.85	76.94	79.28	78.61

降低了模型准确率, 表明防御策略对模型学习造成较大干扰. AdjMask 与 GraphGuard 在多个场景下实现了略高甚至提升明显提升的分类准确率. 特别是 GraphGuard, 在多个场景下实现了分类准确率的明显提升, 例如在 GAT-CiteSeer 设置下, 其准确率相较于无防御模型提升了 18.5%, 表现出极强的精度保持和促进能力. GraphGuard 通过放射性图构造增强成员与非成员数据的预测分布差异, 并通过无需原始训练图的遗忘学习机制实现节点级隐私保护, 一方面消除了模型对未授权数据的过拟合, 另一方面保留对授权数据的学习效果. 然而, 这种对提升模型精度的过度追求限制了隐私保护的效果. 相反, 在多数设置下, 得益于节点级对比学习的引入, 本文提出的防御框架在提供极强隐私保护效果的同时, 保证了模型的效用. 尽管提高目标模型的精度的能力弱于 GraphGuard, 但是本文的防御框架在绝大多数情况下实现了与原始模型近似的分类精度, 且在部分设置下小幅度提升了分类精度, 显著优于 DMP. 本文认为, 提出防御框架在保证模型性能上的表现是可以接受的.

此外, 本文的协同防御方法通过较小的计算资源代价获得了更优的防御效果与模型性能. 标准 GCN 模型单个 epoch 训练的时间复杂度记为 T_{GNN} , 其主要有消息传递与梯度计算构成, 约为 $O(L \cdot E \cdot D)$, 其中 L , E 和 D 分别表示模型层数、图结构的边数与节点特征维度. LBP 方法对后验概率添加拉普拉斯噪声, 在每个 epoch 中引入的开销仅为对输出向量添加噪声扰动, 时间复杂度为 $O(N \cdot C)$, 其中 N 为节点数, C 为类别数. AdjMask 需要为每个节点聚合其邻居节点的置信度来对输出进行重构, 该过程类似于一次在 C 维向量上的消息传递. AdjMask 每个 epoch 的额外的开销受到图的稀疏性影响, 为 $O(E \cdot C)$. DMP 是一种基于知识蒸馏的防御框架, 需要完整的两阶段训练: 先训练教师模型, 再用其输出软标签来训练学生模型. 因而其需要额外的一倍时间开销, 计算开销较高. GraphGuard 基于节点属性生成合成图并利用该图对模型进行微调, 其中, 合成图的构建非常耗时, 可达 $O(N^2 \cdot D)$, 而微调训练又需要额外的 $O(L \cdot E \cdot D)$ 开销, 计算开销最高. 本文提出的协同框架的每个 epoch 的额外开销来自于训练阶段的 3 个模块. 自适应高熵软标签生成模块涉及节点重要性、熵和梯度范数的计算, 主要开销由节点重要性的计算决定, 约为 $O(N + E)$. 熵正则项的计算复杂度为 $O(N \cdot C)$, 可忽略不计. 节点级对比学习是最主要的开销来源, 大小为 B 的 mini-batch 需要进行约为 $O(B^2 \cdot D)$ 的相似度计算, 因而每个 epoch 引入的对比学习计算开销为 $O((N/B) \cdot B^2 \cdot D) = O(N \cdot B \cdot D)$. 对比来说, 本文的协同防御框架的复杂度高于 LBP 与 AdjMask 这两种轻量级的防御方法, 但低于需要额外一轮完整训练的 DMP 和需要复杂图生成的 GraphGuard. 本文认为, 协同防御框架的中等级别时间开销是合理且具有价值的, 通过额外可控的计算开销换取了更加有效的防御效果与模型性能.

综上所述, 本文提出的协同防御方法在多个模型和数据集上均能稳定地让成员推理攻击退化为随机猜测, 在达到理论最优水平的同时, 保持了与无防御模型相近的分类准确率. 相比现有方法模型性

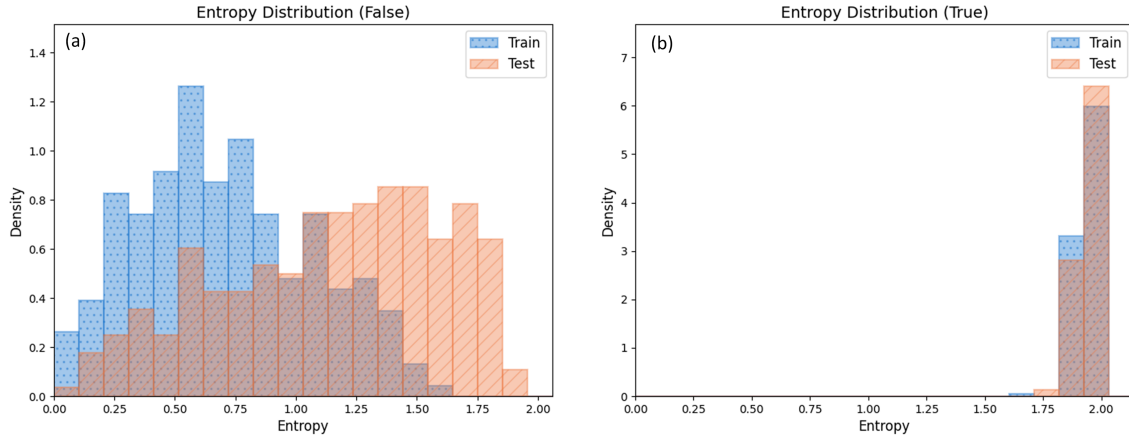


图 2 (网络版彩图) 推理阶段的熵值概率密度分布. (a) 无防御; (b) 协同防御后.

Figure 2 (Color online) Probability density distribution of the entropy value in the inference stage. (a) Without defense; (b) after the collaborative defense.

能与防御效果无法兼顾的缺陷, 本文的防御方法通过训练阶段的高熵软标签引导生成与推理阶段的输出扰动协同设计, 在隐私保护与模型效能之间实现了最优的平衡, 展现出良好的稳定性与实用性.

为进一步验证所提防御框架在输出空间中对成员推理攻击的抑制作用, 并从信息熵的角度量化和验证框架带来的增益, 本文选用 GCN 模型和 Cora 数据集进行了推理阶段的输出测试, 对模型成员数据与非成员数据上的预测熵的概率密度分布进行了可视化, 概率密度分布如图 2 所示. 通过观察两类数据在预测熵维度上的统计特征, 可进一步理解本文防御方法对模型输出可区分性削弱的实际效果.

如图 2(a) 所示, 在未采用任何防御机制时, 成员节点与非成员节点在最大预测熵上的分布差异显著, 可以轻松察觉出隐私泄露的程度. 成员节点 (蓝色区域) 的熵分布主要集中在 $[0.2, 1.0]$ 的低熵区间, 这表明 GCN 对训练集中的样本的预测非常自信. 相反, GCN 在非成员节点 (橙色区域) 的预测熵则广泛分布在 $[1.5, 2.0]$ 的高熵区域, 表明模型在未知样本上的预测更为模糊, 不确定性更高. 显然, 这种预测熵之间的分布差异为攻击者提供了明确的可区分信号, 是成员推理攻击成功的直观证据. 图 2(b) 展示了应用本文提出的训练-推理协同防御框架后, 模型预测熵的分布变化. 可以观察到, 输出熵的分布发生了根本性改变, 成员节点的预测熵整体上移, 密度分布与非成员节点显著趋同. 两者在整个熵轴上的重叠度显著提升, 原先集中于低熵区域的成员节点不再表现出极端置信度输出, 其最大熵密度向 2.0 附近集中, 与非成员几乎完全重合. 同时, 非成员样本的熵分布也集中于相同的区间内, 呈现更集中的统计特征, 反映出模型在整体输出层面进行了系统性调整. 这种几乎重叠的分布使得攻击者很难从预测输出中挖掘出有用的信息, 降低了成员推理的风险.

上述分布变化充分说明, 本文提出的协同防御策略能够有效重构成员样本的预测输出特征. 其一, 自适应高熵软标签生成机制在训练阶段主动提升模型在关键节点上的预测不确定性, 缓解了模型在训练集上的过拟合现象. 其二, 熵正则项进一步抑制低熵输出, 驱动模型学习平滑输出分布. 其三, 在推理阶段, 基于随机图生成的输出扰动策略对模型输出进行置信度重构, 从根本上消除成员与非成员之间的输出差异. 三者协同作用显著削弱了熵空间中的可区分性, 从而极大降低了攻击者通过熵差判别成员身份的可能性.

综上所述, 本文方法不仅在分类准确率保持的基础上, 有效控制了模型输出空间中的熵结构差异, 更通过熵分布统计分析验证了方法在打破成员身份信息残留方面的可解释性与实效性.

4.3 消融实验

为了系统评估本文所提出的训练-推理协同防御框架中各个模块和策略的作用, 本文设置了相应

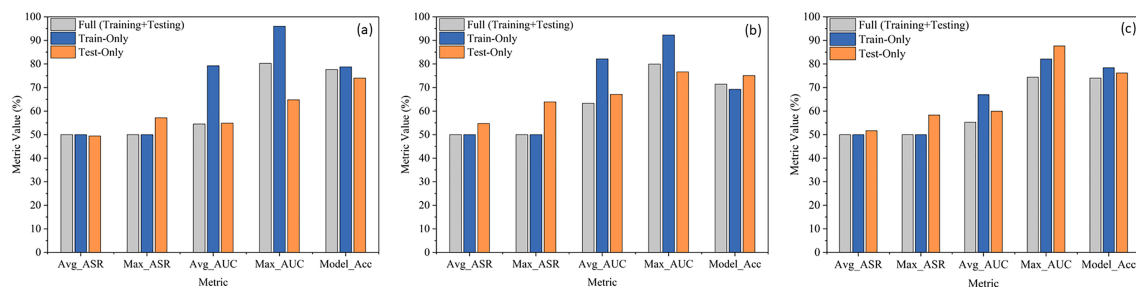


图 3 (网络版彩图) 针对防御阶段的消融实验结果. (a) GCN; (b) GraphSAGE; (c) GAT.

Figure 3 (Color online) Results of ablation experiments targeting the defense phase. (a) GCN; (b) GraphSAGE; (c) GAT.

的消融实验进行详细验证.

4.3.1 针对防御阶段的消融实验

首先, 为了验证两个阶段防御策略的独立贡献, 本节讨论了第一组消融实验, 构造了 3 组对照防御框架: 完整的防御策略、仅包含训练阶段的防御策略和仅包含推理阶段的防御策略.

此外, 为了进一步评估防御方法的泛化性, 除了 4.1 小节中设置的评价指标外, 本节引入了 AUC (area under the ROC curve) 作为额外的评估指标. 该指标衡量了模型在不同判别阈值下的整体区分能力, 该阈值用于判断输入是否为成员样本. 相比之下, 在直接以 ASR 为评估指标的情况下, 阈值默认为 0.5. 引入 AUC 指标, 能够覆盖更广的攻击设定, 从而更加鲁棒地反映攻击者识别成员样本的潜力. 类似地, 本文计算了所有类别下攻击模型的 AUC 平均值 Avg-AUC, 以及最容易被攻击的类别对应的 AUC 最大值 Max-AUC. 本节选用 Cora 数据集在 3 个不同目标模型上进行了实验, 实验结果如图 3 所示.

单独的训练阶段防御在攻击成功率与模型精度指标的表现与完整防御框架近似, 但是在 AUC 指标上表现不佳. 以 GCN 为例, 其 Avg-AUC 与 Max-AUC 均比推理阶段防御高 30%. 单独的推理阶段防御在模型精度指标上优于完整的防御框架, 但是在攻击成功率指标上的表现弱于训练阶段防御. 以 GraphSAGE 为例, 推理阶段防御的 Avg-ASR 比完整框架高 5%, Max-ASR 比完整框架高 15%, 这说明, 单一阶段的防御并不能完全消除模型对训练样本的过度自信, 成员样本与非成员样本之间的分布差异仍然存在. 相比之下, 完整防御框架在 3 个评价维度上均表现出最均衡的性能.

上述实验表明, 训练阶段防御策略在缓解模型过拟合、提升输出熵值方面具有积极作用, 而推理阶段防御通过统一置信度输出有效打破输出结构中的可分性边界. 若缺乏推理阶段的结构性输出扰动补充, 成员与非成员在输出空间中的差异仍难以完全消除, 防御性能存在显著缺陷. 类似地, 若缺乏训练阶段的不确定性高熵引导, 模型在训练样本上仍易产生记忆性偏差. 两者协同作用可在模型精度保持的基础上, 显著降低攻击成功率并提高了防御的泛化性, 实现更为全面且稳健的隐私保护.

4.3.2 针对高熵阈值的消融实验

与传统的手动设置阈值策略^[42]不同, 本文基于节点重要性指标自适应分配高熵阈值. 为了验证自适应策略的有效性, 本节设计了与固定高熵阈值对比的消融实验. 该组消融实验选用了 5 个连续的固定阈值 (0.5, 0.6, 0.7, 0.8, 0.9), 在 3 种不同的数据和模型上与本文的自适应策略进行了对比, 实验结果如图 4 所示.

从攻击成功率来看, 固定高熵阈值策略在大多场景下未能有效压制成员推理攻击, 且在不同的数据集和模型上有较大波动. 这说明固定阈值方法难以适应不同数据结构和模型特性, 在面对复杂图结构和异质节点分布时, 不能有效识别并重点保护高风险节点, 导致整体攻击成功率维持在较高水平, 防

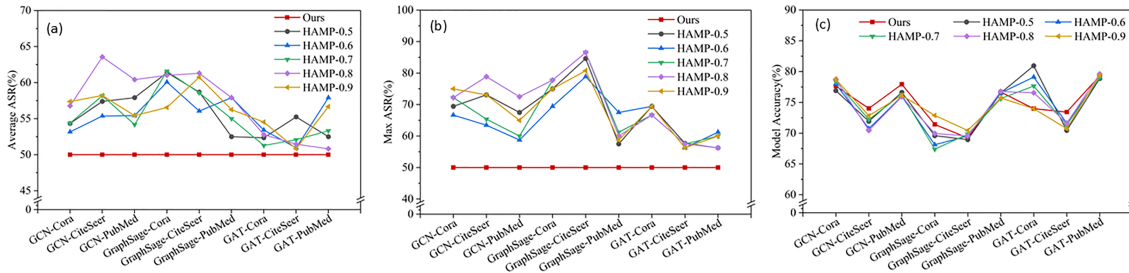


图4 (网络版彩图) 针对高熵阈值的消融实验结果. (a) Avg-ASR; (b) Max-ASR; (c) Acc.

Figure 4 (Color online) Results of ablation experiments for high entropy thresholds. (a) Avg-ASR; (b) Max-ASR; (c) Acc.

御的泛化性有限. 与之相比, 本文提出的自适应高熵阈值策略在所有设置中均实现了最优的攻击抑制效果, Avg-ASR 和 Max-ASR 保持在随机猜测的水平, 显著优于任意固定阈值设定. 这一优势得益于该策略融合结构中心性、熵值水平与梯度敏感性等多个指标对节点重要性进行综合建模, 动态分配熵扰动强度, 有效增强了对隐私易泄露节点的重点防护能力, 从而在不引入额外性能损失的情况下, 实现更细粒度的隐私控制.

从分类性能来看, 高熵阈值 (如 0.8 或 0.9) 可带来略微的精度提升, 但同时攻击成功率也显著上升, 导致隐私保护效果显著下降. 这表明, 固定的高熵阈值策略不能实现隐私与效用的有效平衡. 相比之下, 自适应策略在分类精度上的波动幅度最小, 与最佳固定阈值结果基本持平, 但能显著降低攻击成功率, 体现出更优的隐私-效用协同优化能力.

综上所述, 固定高熵阈值策略在不同模型和数据集上表现波动较大, 缺乏对结构异质性和节点风险差异的适应能力, 隐私保护性能缺乏一致性和鲁棒性. 而本文提出的自适应阈值策略则能动态调节扰动目标与强度, 在不同攻击场景下均表现出稳定而优越的隐私泄露抑制效果.

4.3.3 针对节点更新和节点选择的消融实验

为进一步评估节点选择策略对高熵软标签训练机制效果的影响, 本文设计了两组消融实验, 对比不同节点参与方式下成员推理防御性能的差异. 本文提出的防御框架默认所有训练节点均参与高熵软标签的构造与参数迭代, 两组对照组分别设定为: (1) 每个 epoch 随机选择 20% 的节点参与高熵标签扰动; (2) 每 5 个 epoch 固定一批随机选择的 20% 节点参与高熵标签扰动. 本节按照 3 种设置在不同的模型和数据集上进行了测试, 实验结果如图 5 所示.

原始全节点参与的策略在所有模型与数据集上均将 ASR 稳定压制至 50.00%, 相比之下, 两种稀疏节点参与策略的 ASR 明显上升, 这表明高熵软标签作用范围受限时, 部分成员节点暴露风险显著增加. 在模型精度方面, 两种稀疏参与策略优于全节点方案, 但提升幅度有限, 且伴随一定的隐私损失.

结合上述实验结果, 本文认为扰动节点范围的覆盖度与更新频率对防御效果影响显著. 仅使用少量节点参与高熵扰动虽可提供最佳的模型性能, 但牺牲了大部分节点的隐私保护. 相较而言, 所有节点均参与生成与更新机制能确保扰动均匀覆盖, 构建稳定可靠的成员推理防御边界. 需要注意的是, 由于在高熵软标签生成过程中, 防御框架基于节点重要性自适应分配高熵阈值, 隐私-效用的权衡不会因为全节点参与而被破坏, 相反, 这种自适应策略使隐私-效用的权衡能够在全节点上达到最优.

4.3.4 针对训练阶段防御模块的消融实验

为了系统性地评估训练阶段防御策略中各个核心模块的独立贡献, 本文进行了一系列的消融实验, 包含 3 种消融场景, 并在 Cora 数据集上对 3 种模型进行了测试. 仅软标签 (基线) 对照组移除了完整防御框架的正则项和对比学习模块, 无对比学习对照组移除了对比学习模块, 无熵正则项对照组移除了熵正则化模块.

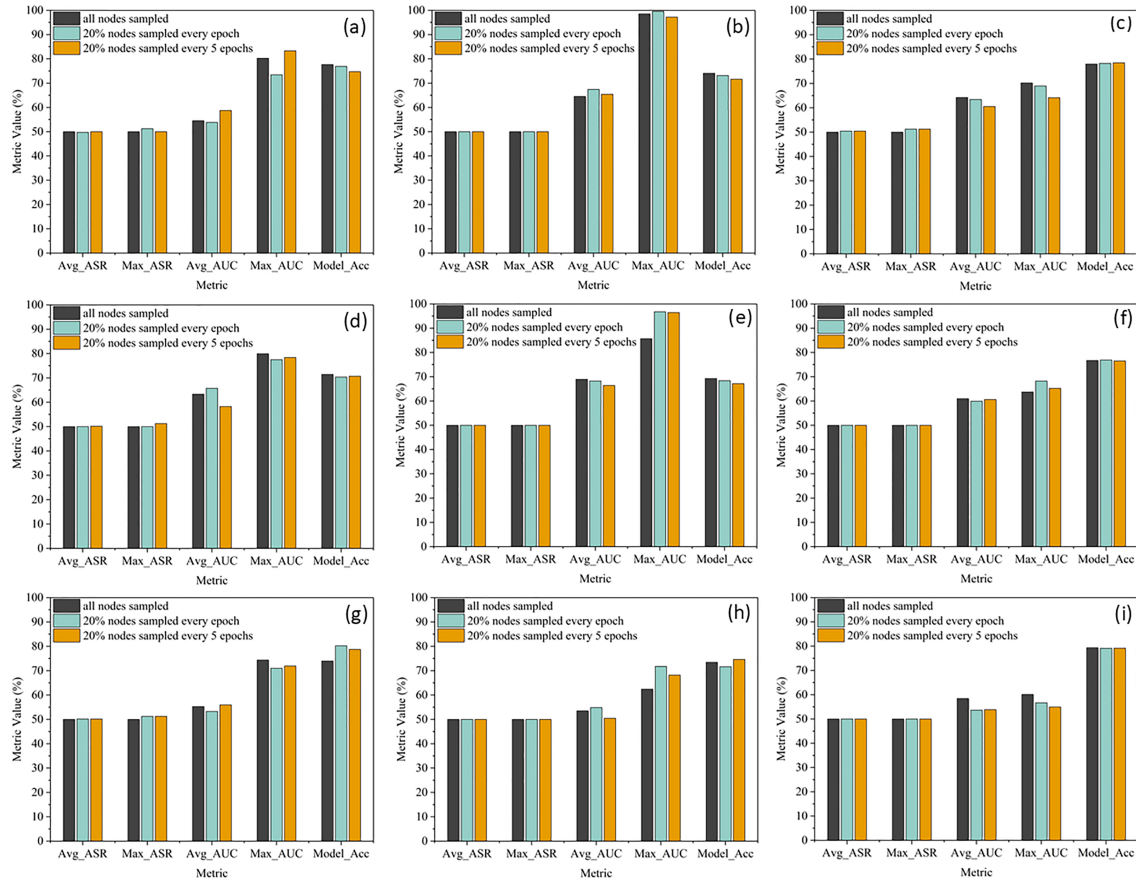


图 5 (网络版彩图) 针对节点更新和节点选择的消融实验. (a) GCN-Cora; (b) GCN-CiteSeer; (c) GCN-PubMed; (d) GraphSAGE-Cora; (e) GraphSAGE-CiteSeer; (f) GraphSAGE-PubMed; (g) GAT-Cora; (h) GAT-CiteSeer; (i) GAT-PubMed.

Figure 5 (Color online) Ablation experiments for node updating and node selection. (a) GCN-Cora; (b) GCN-CiteSeer; (c) GCN-PubMed; (d) GraphSAGE-Cora; (e) GraphSAGE-CiteSeer; (f) GraphSAGE-PubMed; (g) GAT-Cora; (h) GAT-CiteSeer; (i) GAT-PubMed.

图 6 揭示了各模块的贡献程度与协作. 首先, 若仅使用高熵软标签作为唯一的防御手段, 虽然提供了一定程度的隐私保护, 但其效果远非理想. 例如, 在 GraphSAGE 模型上, 其 Avg-ASR 和 Max-ASR 分别高达 71.15% 和 86.11%, Max-AUC 值也达到了 90.74%, 这表明模型输出仍有冗余信息供攻击者区分成员与非成员. 在此基础上引入熵正则项后 (即无对比学习), 模型的隐私保护能力得到了大幅度提升. 对于 3 个目标模型, 加入熵正则项后 ASR 指标均出现了大幅下降, 这一趋势在 GCN 和 GraphSAGE 上尤为显著. 这有力地证明了, 熵正则项通过对模型输出分布的强力约束, 是本防御框架中降低攻击成功率、提升隐私安全的核心驱动模块. 然而, 隐私保护并非没有代价, 模型准确率有小幅下降. 这恰恰凸显了节点级对比学习模块不可或缺的价值. 在引入了对比学习的“无熵正则化”场景中, GCN 的精度恢复到了基线水平. 这表明, 节点级对比学习模块具有模型精度的代偿作用, 它通过在特征表示层强制模型学习更具区分度的信息, 有效对冲了由高熵标签和熵正则化在输出层造成的预测模糊化效应, 从而确保了模型在获得强大隐私保护的同时, 其核心任务性能不被牺牲.

综上所述, 本文设计的训练阶段防御框架通过 3 个模块的协同机制, 实现了隐私保护 – 模型效用之间的平衡. 其中, 高熵软标签模块奠定了防御的基础, 熵正则模块提供了强力的隐私保护效果, 对比学习模块抵消了防御策略带来的模型性能下降.

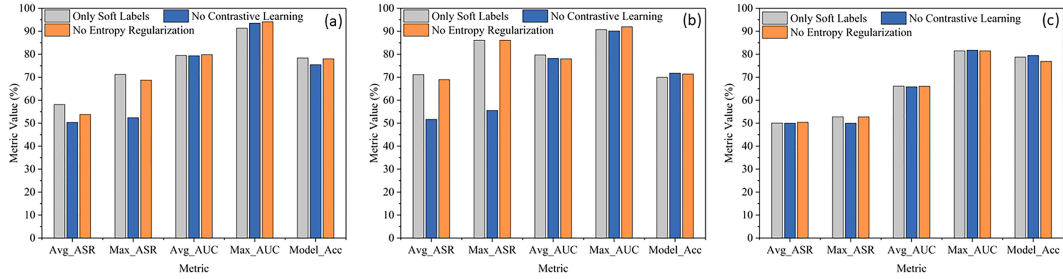


图 6 (网络版彩图) 针对节点更新和节点选择的消融实验. (a) GCN; (b) GraphSAGE; (c) GAT.

Figure 6 (Color online) Ablation experiments for node updating and node selection. (a) GCN; (b) GraphSAGE; (c) GAT.

表 4 受保护模型在白盒 MIA 下的平均攻击成功率与模型准确率. 加粗数值表示在对应数据集 – 模型组合下的最佳防御效果 (包括平均攻击成功率与最高攻击成功率).

Table 4 Avg-ASR and Model-Acc of protected models under white-box MIAs. Bold values indicate the best defense performance under the corresponding dataset-model combination, including both Avg-ASR and Max-ASR.

Model	Dataset	Method (Avg-ASR / Model-Acc)					
		No-defense	Ours	LBP	AdjMask	GraphGuard	DMP
GCN	Cora	0.678/0.816	0.540/0.831	0.685/0.731	0.710/0.804	0.572/0.813	0.566/0.692
	CiteSeer	0.708/0.707	0.650/ 0.716	0.690/0.699	0.708/0.712	0.552/0.625	0.606/0.697
	PubMed	0.799/0.784	0.649/0.750	0.792/0.788	0.768/0.766	0.550/0.819	0.732/0.804
GraphSAGE	Cora	0.715/0.817	0.595/0.805	0.669/0.814	0.693/0.822	0.681/ 0.919	0.600/0.647
	CiteSeer	0.687/0.687	0.569/0.693	0.659/0.694	0.695/0.699	0.627/ 0.805	0.522/0.653
	PubMed	0.785/0.783	0.671/0.756	0.772/0.778	0.784/0.774	0.509/0.841	0.733/0.758
GAT	Cora	0.768/0.797	0.515/0.827	0.669/0.717	0.689/0.807	0.646/0.610	0.548/0.606
	CiteSeer	0.700/0.721	0.552/0.731	0.721/0.635	0.693/0.690	0.638/0.723	0.638/0.620
	PubMed	0.787/0.757	0.547/ 0.776	0.781/0.754	0.695/0.756	0.527/0.615	0.722/0.752

4.4 针对白盒攻击场景的迁移防御效果

为了进一步验证协同防御框架的鲁棒性和泛化性, 本文在白盒 (white-box) 攻击场景下评估了其防御性能. 本文选用了经典梯度攻击方法^[10], 该攻击假设攻击者能够访问目标模型的内部梯度信息, 从而进行更具侵略性的隐私窃取. 该部分实验对比了本文提出方法与 4 种主流的防御方法.

由表 4 可知, 与黑盒场景相比, 白盒攻击的针对性更强, 所有防御方法的 Avg-ASR 都有大幅度上升, 无法被控制在随机猜测水平. 根本原因在于信息获取维度的增加. 黑盒场景下, 攻击者只能观察到模型的最终预测输出, 而白盒攻击者能够获取模型内部的梯度信息与完整预测输出. 梯度是模型对输入最敏感、最细粒度的响应, 能更直观地反映出模型决策的依据. LBP 和 AdjMask 这类主要作用于模型输出层的防御方法, 虽然能够混淆最终的预测置信度, 但它们几乎没有改变模型内部的梯度计算过程, 导致被基于梯度的白盒攻击者轻易绕过. 相比之下, 本文的协同防御框架在面对白盒攻击时仍然展现了相对良好的防御效果, 在绝大多数情况下都要优于其他 4 种防御方法. 不同于仅在模型输出添加防御机制的方法, 本文提出的协同框架在训练阶段就通过自适应高熵软标签和熵正则化, 主动且深刻地改变了模型的学习范式. 通过抑制模型对成员样本的过拟合, 本文方法从根本上缩小了成员与非成员样本在模型空间中的差异, 这使得基于梯度的攻击者难以轻松地挖掘出判别成员样本的决策边界. 而推理阶段的输出扰动防御进一步巩固了防御效果. 尽管防御效果弱于黑盒场景, 白盒攻击实验充分证明了本文方法的迁移性与泛化性. 未来, 我们的工作将进一步针对白盒攻击场景提升协同防御机制的泛化性.

5 总结

本文针对图神经网络面临的成员推理攻击风险,提出了一种融合训练与推理双阶段的协同防御框架.该框架设计了基于节点重要性的高熵软标签生成机制,强化对高风险节点的保护,通过标签拟合损失引导平滑模型的输出分布,并引入了节点级的对比学习模块提高模型的判别能力.与此同时,推理阶段的输出扰动策略有效隐藏了成员信息在模型输出中的显性表达.实验结果表明,本文提出的协同防御框架在隐私保护与模型性能保持之间实现了更优的平衡,显著优于现有主流防御方法,具有良好的鲁棒性与泛化性.

参考文献

- 1 Hu B D, Wang X G, Wang X Y, et al. Survey on hypergraph learning: algorithm classification and application analysis. *J Soft*, 2022, 33: 498–523 [胡秉德, 王新根, 王新宇, 等. 超图学习综述: 算法分类与应用分析. *软件学报*, 2022, 33: 498–523]
- 2 Cao W M, Zheng C T, Yan Z Y, et al. Geometric deep learning: progress, applications and challenges. *Sci China Inf Sci*, 2022, 65: 126101
- 3 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv:1609.02907
- 4 Hamilton W, Ying R, Leskovec J. et al. Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems (NIPS)*, 2017
- 5 Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018
- 6 Velickovic P. Everything is connected: graph neural networks. 2023. ArXiv:2301.08210
- 7 Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications, and evaluations. 2019. ArXiv:1906.05017
- 8 Dai E, Zhao T, Zhu H, et al. A comprehensive survey on trustworthy graph neural networks: privacy, robustness, fairness, and explainability. *Mach Intell Res*, 2024, 21: 1011–1061
- 9 Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models. In: *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2017. 3–18
- 10 Duddu V, Boutet A, Shejwalkar V. Quantifying privacy leakage in graph embedding. In: *Proceedings of the 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, 2020. 76–85
- 11 Olatunji I E, Nejd W, Khosla M. Membership inference attack on graph neural networks. In: *Proceedings of the 3rd IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2021. 11–20
- 12 He X, Wen R, Wu Y, et al. Node-level membership inference attacks against graph neural networks. 2021. ArXiv:2102.05429
- 13 Conti M, Li J, Picek S, et al. Label-only membership inference attack against node-level graph neural networks. In: *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, 2022. 1–12
- 14 Chen J, Wang W H, Shi X. Differential privacy protection against membership inference attack on machine learning for genomic data. In: *Proceedings of the Pacific Symposium*, 2020. 26–37
- 15 Salem A, Zhang Y, Humbert M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models. 2018. ArXiv:1806.01246
- 16 Shejwalkar V, Houmansadr A. Membership privacy for machine learning models through knowledge transfer. *AAAI*, 2021, 35: 9549–9557
- 17 Xue M, Yuan C, He C, et al. Use the spear as a shield: an adversarial example based privacy-preserving technique against membership inference attacks. *IEEE Trans Emerg Top Comput*, 2022, 11: 153–169
- 18 Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: analyzing the connection to overfitting. In: *Proceedings of the 31st Computer Security Foundations Symposium (CSF)*, 2018. 268–282
- 19 Song L, Mittal P. Systematic evaluation of privacy risks of machine learning models. In: *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, 2021. 2615–2632
- 20 Wu Y, Qiu H, Guo S, et al. You only query once: an efficient label-only membership inference attack. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024

- 21 Li Z, Zhang Y. Membership leakage in label-only exposures. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), 2021. 880–895
- 22 Choquette-Choo C A, Tramer F, Carlini N, et al. Label-only membership inference attacks. In: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021. 1964–1974
- 23 Peng Y, Roh J, Maji S, et al. OSLO: one-shot label-only membership inference attacks. 2024. ArXiv:2405.16978
- 24 Rezaei S, Liu X. On the difficulty of membership inference attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 7892–7900
- 25 Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2019. 739–753
- 26 Leino K, Fredrikson M. Stolen memories: leveraging model memorization for calibrated white-box membership inference. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020. 1605–1622
- 27 Xie L, Shi J, Yang H, et al. Membership inference attacks based on graph neural network model calibration. J Electron Inform Technol, 2025, 47: 780–791
- 28 Guan F, Zhu T, Zhou W, et al. Topology-based node-level membership inference attacks on graph neural networks. IEEE Trans Big Data, 2025, 11: 2809–2826
- 29 Guan F, Zhu T, Tong H, et al. Attention-based membership inference attacks on graph neural network through topological features. IEEE Trans Dependable Secure Comput, 2025, doi: 10.1109/TDSC.2025.3586251
- 30 Rahimian S, Orekondy T, Fritz M. Sampling attacks: amplification of membership inference attacks by repeated queries. 2020. ArXiv:2009.00395
- 31 Olatunji I E, Funke T, Khosla M. Releasing graph neural networks with differential privacy guarantees. 2021. ArXiv:2109.08907
- 32 Sajadmanesh S, Shamsabadi A S, Bellet A, et al. GAP: differentially private graph neural networks with aggregation perturbation. In: Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023. 3223–3240
- 33 Xiang Z, Wang T, Wang D. Preserving node-level privacy in graph neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2024. 4714–4732
- 34 Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), 2018. 634–646
- 35 Jia J, Salem A, Backes M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), 2019. 259–274
- 36 Chen C, Zhang X, Qiu H, et al. MaskArmor: confidence masking-based defense mechanism for GNN against MIA. Inf Sci, 2024, 669: 120579
- 37 Guan F, Zhu T, Tong H, et al. Topology modification against membership inference attack in graph neural networks. Knowl-Based Syst, 2024, 305: 112642
- 38 Wang K, Wu J, Zhu T, et al. Defense against membership inference attack in graph neural networks through graph perturbation. Int J Inf Secur, 2023, 22: 497–509
- 39 Wu B, Zhang H, Yang X, et al. GraphGuard: detecting and counteracting training data misuse in graph neural networks. In: Proceedings of Network and Distributed System Security Symposium (NDSS), 2024
- 40 Zheng J, Cao Y, Wang H. Resisting membership inference attacks through knowledge distillation. Neurocomputing, 2021, 452: 114–126
- 41 Alvar S R, Wang L, Pei J, et al. Membership privacy protection for image translation models via adversarial knowledge distillation. 2022. ArXiv:2203.05212
- 42 Chen Z, Pattabiraman K. Overconfidence is a dangerous thing: mitigating membership inference attacks by enforcing less confident prediction. In: Proceedings of the Network and Distributed System Security Symposium (NDSS), 2024
- 43 Erdos P, Rényi A. On random graphs. I. Publ Math Debrecen, 1959, 6: 290–297
- 44 McCallum A K, Nigam K, Rennie J, et al. Automating the construction of internet portals with machine learning. Inf Retrieval, 2000, 3: 127–163
- 45 Giles C L, Bollacker K D, Lawrence S. CiteSeer: an automatic citation indexing system. In: Proceedings of the 3rd ACM Conference on Digital Libraries, 1998. 89–98
- 46 Sen P, Namata G, Bilgic M, et al. Collective classification in network data. AI Mag, 2008, 29: 93–106

A training-inference collaborative defense framework against membership inference attacks for graph neural networks

Fuqi QI^{1,2}, Xingyi QUAN^{1,2}, Haichang GAO^{1,2*} & Ping WANG^{1,2}

1. *School of Computer Science and Technology, Xidian University, Xi'an 710126, China*

2. *Xi'an Key Laboratory of Intelligence System Security, Xi'an 710126, China*

* Corresponding author. E-mail: hchgao@xidian.edu.cn

Abstract Graph neural networks (GNNs) have been widely adopted in sensitive domains such as social recommendation, medical diagnosis, and financial analysis due to their powerful capability in modeling structured data. However, their strong reliance on local structures and node attributes during training exposes them to membership inference attacks (MIAs), where attackers can infer whether a node was part of the training set by analyzing prediction discrepancies. Existing defense strategies often suffer from limited effectiveness, high graph perturbation costs, or significant performance degradation when applied to GNNs. To address this, we propose a training-inference collaborative defense framework that enhances output obfuscation from the perspective of predictive entropy. In the training phase, we design an adaptive high-entropy soft labeling mechanism based on node importance, combined with entropy regularization to suppress overconfident predictions. A node-level contrastive learning module is further introduced to mitigate the accuracy drop caused by high-entropy supervision. In the inference phase, we propose a random graph-driven output perturbation strategy that reconstructs the model's output using non-private synthetic inputs while preserving the prediction order. Experimental results on multiple benchmark datasets and GNN models demonstrate that our method achieves a superior trade-off between privacy and utility, significantly enhancing resistance against MIAs.

Keywords graph neural networks, membership inference attacks, privacy protection, graph data, information entropy