基于图神经网络模型校准的成员推理攻击

谢丽霞[①] 史镜琛[①] 杨宏宇*[①②] 胡 泽[②] 成 翔[③]
[①](中国民航大学计算机科学与技术学院 天津 300300)
[②](中国民航大学安全科学与工程学院 天津 300300)
[③](扬州大学信息工程学院 扬州 225127)

摘 要：针对图神经网络(GNN)模型在其预测中常处于欠自信状态，导致该状态下实施成员推理攻击难度大且攻击漏报率高的问题，该文提出一种基于GNN模型校准的成员推理攻击方法。首先，设计一种基于因果推断的GNN模型校准方法，通过基于注意力机制的因果图提取、因果图与非因果图解耦、后门路径调整策略和因果关联图生成过程，构建用于训练GNN模型的因果关联图。其次，使用与目标因果关联图在相同数据分布下的影子因果关联图构建影子GNN模型，模拟目标GNN模型的预测行为。最后，使用影子GNN模型的后验概率构建攻击数据集以训练攻击模型，根据目标GNN模型对目标节点的后验概率输出推断其是否属于目标GNN模型的训练数据。在4个数据集上的实验结果表明，该文方法在2种攻击模式下面对不同架构的GNN模型进行攻击时，攻击准确率最高为92.6%，性能指标优于基线攻击方法，可有效地实施成员推理攻击。

关键词：图神经网络；成员推理攻击；模型校准；因果推断；隐私风险
中图分类号：TN915.08; TP309 **文献标识码**：A **文章编号**：1009-5896(2025)03-0780-12
DOI: 10.11999/JEIT240477

1 引言

面向机器学习(Machine Learning, ML)模型的成员推理攻击(Membership Inference Attacks, MIAs)[1,2]是一种针对ML模型训练数据的隐私攻击，目的是推断攻击者感兴趣的数据样本是否被用于训练目标ML模型。由于揭示了ML模型中可能存在的隐私泄露风险，MIAs在人工智能安全领域中具有重要地位。随着人工智能技术在医疗、金融和社交网络等领域的广泛应用，保护用户隐私和确保数据安全变得尤为关键。研究MIAs有助于安全研究者评估模型在攻击下的脆弱性，规范敏感数据的使用，并设计提升模型安全性的方法。

目前，根据攻击原理的不同，面向ML模型的MIAs主要分为基于2元分类器的MIAs[3,4]和基于评估机制的MIAs[5,6]。前者攻击方法准确性高且灵活性强，但存在计算开销较大、依赖大规模数据等不足；后者攻击方法对于不同类型的目标模型需要预定义不同的评估指标，通用性较差。

近年来，面向图神经网络(Graph Neural Network, GNN)模型[7]的MIAs逐渐成为研究热点。面向GNN模型的MIAs是为了确定攻击者感兴趣的目标节点是否为GNN模型训练数据集的一部分。文献[8–11]初步探索面向GNN模型的MIAs，证明了MIAs在GNN的节点分类任务、图分类任务和链接预测任务中，都有成功实施的可能性，揭示GNN模型面对该类隐私攻击时潜在的脆弱性。

Olatunji等人[8]和He等人[9]研究了GNN节点分类任务中的MIAs，采用基于2元分类器的MIAs方法，利用GNN模型的后验概率输出进行实验，但攻击准确率均相对较低。其原因是GNN模型在其预测中常处于欠自信状态(under-confidence)[12]，即模型在输出阶段通常给出较为谨慎的概率估计。由于GNN模型的许多预测值集中在低置信度范围内，所以基于此后验概率实施的MIAs[8,9]导致攻击的准确率较低，以及攻击的漏报率(False Negative Rate, FNR)较高。

现有GNN模型校准方法中，后处理校准方法通过调整模型输出的概率分布，提高预测的置信度和可靠性，从而改善GNN模型预测的欠自信现象。Wang等人[12]提出了一种基于图卷积网络的模型校准方法，通过利用图卷积网络对预训练分类器输出层的输出结果进行处理，生成节点温度值，从而调整模型的后验概率。Hsu等人[13]提出了一种图注意

收稿日期：2024-06-12；改回日期：2025-02-17；网络出版：2025-02-26
*通信作者： 杨宏宇 yhyxlx@hotmail.com
基金项目：国家自然科学基金民航联合研究基金重点项目(U2433205)，国家自然科学基金(62201576, U1833107)，江苏省基础研究计划自然科学基金青年基金(BK20230558)
Foundation Items: The Civil Aviation Joint Research Fund Project of the National Natural Science Foundation of China (U2433205), The National Natural Science Foundation of China (62201576, U1833107), Jiangsu Provincial Basic Research Program Natural Science Foundation-Youth Fund (BK20230558)

力温度缩放方法(Graph Attention Temperature Scaling, GATS)，使用基于注意力的架构生成节点温度以调整模型的后验概率。Liu等人[14]通过分析模型容量、图密度和损失函数对校准的影响，提出了一种拓扑感知校准方法，使校准后的模型输出更准确地反映节点间的关系。上述方法聚焦于直接调整GNN模型输出概率分布，方法的有效性依赖于特定的数据集和应用场景，且对不同数据集或任务的适应性和泛化能力较差。此外，上述方法忽略了GNN模型会学习训练数据中输入特征与标签之间的所有统计相关性，缺乏对特征间因果关系的深入分析，无法有效处理非因果特征，导致校准结果存在偏差，可靠性不足。

针对上述问题，本文研究面向GNN节点分类任务的成员推理攻击，提出一种基于GNN模型校准的成员推理攻击方法(Membership Inference Attacks based on Model Calibration, MIAs-MC)，使攻击模型在面对准确、可靠的后验概率输出时，能够更准确地判断目标节点是否来自于目标GNN模型的训练数据集。

2 成员推理攻击方法

在基于2元分类器的MIAs方法基础上，针对GNN模型预测的欠自信问题，本文提出一种基于GNN模型校准的成员推理攻击方法MIAs-MC，本方法架构如图1所示，主要包括基于因果推断的模型校准、影子模型构建、攻击模型构建与成员推理攻击3部分。各部分的主要功能和处理过程设计如下。

(1) 基于因果推断的模型校准：首先，基于节点和边的注意力机制，从初始训练子图中初步提取因果图。其次，通过解耦因果图与非因果图，得到用于区分因果特征和非因果特征的分类损失函数。然后，通过后门路径调整策略，得到后门路径调整对应的损失函数。最后，将3个损失函数加权相加作为总损失函数以更新模型参数，生成用于训练目标模型和影子模型的因果关联图。

(2) 影子模型构建：使用基于因果推断的模型校准方法中生成的影子因果关联图，构建并训练一个校准的影子模型，用于模拟已校准目标模型在图神经网络节点分类任务中的预测行为。

(3) 攻击模型构建与成员推理攻击：首先，根据校准的影子模型中影子数据集节点的后验概率及其对应的成员标签，构建攻击模型数据集。然后，使用构建的攻击模型数据集训练一个攻击模型。最后，将已校准目标模型中得到的目标节点的后验概率输入攻击模型，通过识别目标节点的攻击特征，判断该节点是否为成员节点。

3 基于因果推断的模型校准

传统的GNN模型校准方法通常聚焦于直接调整GNN模型的后验概率输出，缺乏对特征间因果关系的深入分析，无法有效处理非因果特征，导致GNN模型的预测结果存在偏差，可靠性不足。

因此，本文提出一种基于因果推断的模型校准方法，用于生成目标和影子GNN模型初始训练子图的因果关联图。目标因果关联图和影子因果关联图分别用于训练校准的目标GNN模型和影子GNN模型。首先，使用GNN编码器对初始的目标训练子图和影子训练子图进行节点嵌入表示，并通过节点和边的注意力机制初步提取因果注意力图。其次，在对因果注意力图和非因果注意力图进行解耦操作后，使用后门路径调整策略进一步降低非因果注意力图的影响。最后，通过最小化损失函数迭代更新因果注意力图，生成用于训练目标GNN模型和影子GNN模型的目标因果关联图和影子因果关联图，本方法如图2所示。

3.1 因果图提取

生成因果关联图的第1步是从初始训练子图中

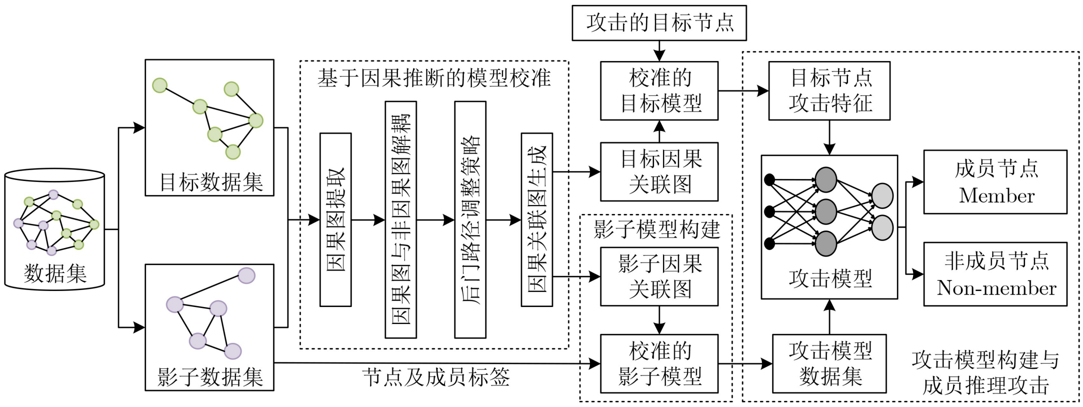

图 1 MIAs-MC攻击方法架构

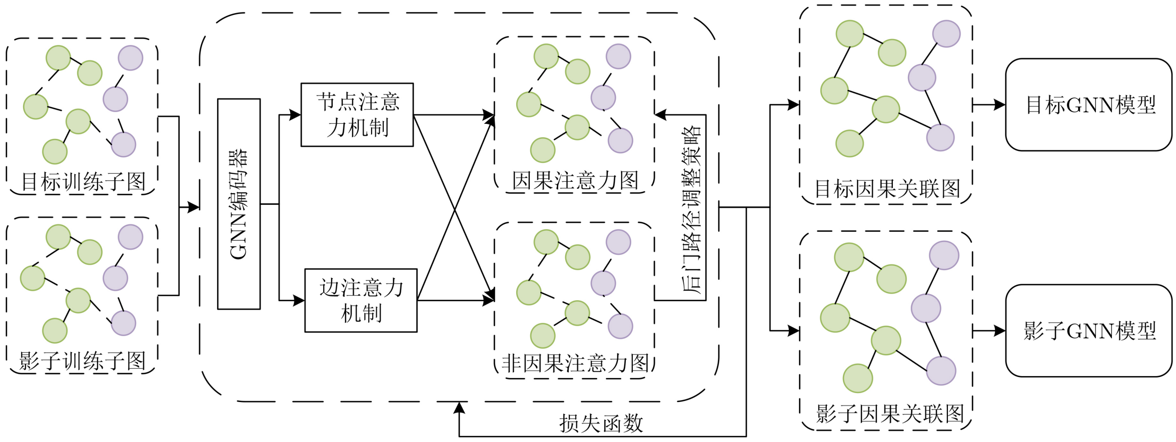

图 2 基于因果推断的模型校准方法

提取因果注意力图(即因果图)。本文将用于训练GNN模型的子图形式化记为$G=(\boldsymbol{A},\boldsymbol{X},V,E)$。其中，$\boldsymbol{A}$表示训练子图的邻接矩阵，$\boldsymbol{X}$表示训练子图上所有节点的特征矩阵，$V$表示节点集合，$E$表示边的集合。

针对初始的目标训练子图$G_{\rm t}$和影子训练子图$G_{\rm s}$，对其节点特征和边特征分别使用软掩码$\boldsymbol{M}_{\rm N}\in R^{|V|\times|V|}$和$\boldsymbol{M}_{\rm E}\in R^{|V|\times|V|}$。其中，软掩码$\boldsymbol{M}_{\rm N}$和$\boldsymbol{M}_{\rm E}$中每个值代表GNN模型感兴趣特征的相关注意力分数，值范围为(0,1)。对于给定软掩码$\boldsymbol{M}$，定义其互补的软掩码为$\bar{\boldsymbol{M}}=\boldsymbol{1}-\boldsymbol{M}$。本文对因果图的初步提取过程如下。

首先，使用一个GNN模型作为编码器对初始训练子图中的节点进行嵌入表示，如式(1)所示

$$\boldsymbol{h}_i^{(l)}=\gamma^{(l)}(\boldsymbol{h}_i^{(l-1)},\mathop{||}_{j\in N(i)}\phi^{(l)}(\boldsymbol{h}_i^{(l-1)},\boldsymbol{h}_j^{(l-1)},\boldsymbol{e}_{(i,j)}))\tag{1}$$

其中，γ表示用于节点特征映射的全连接层，$\boldsymbol{h}_i^{(l-1)}$和$\boldsymbol{h}_j^{(l-1)}$分别表示节点i和节点j在第l–1层的特征嵌入向量，$N(i)$表示节点i邻居节点的集合，$||$表示特征的拼接操作，ϕ表示置换不变性操作，$\boldsymbol{e}_{(i,j)}$表示节点i和j之间边的特征向量。

其次，使用2个全连接层分别计算节点和边的注意力分数，计算过程为

$$\alpha_{c_i},\alpha_{u_i}=\sigma(\boldsymbol{W}_{\rm node}(\boldsymbol{h}_i))\tag{2}$$

$$\beta_{c_{ij}},\beta_{u_{ij}}=\sigma((\boldsymbol{W}_{\rm edge}\boldsymbol{h}_i)^{\rm T}(\boldsymbol{W}_{\rm edge}\boldsymbol{h}_j))\tag{3}$$

其中，$\sigma(\cdot)$表示激活函数，$\boldsymbol{W}_{\rm node}$和$\boldsymbol{W}_{\rm edge}$分别表示节点和边的线性变换矩阵，$\alpha_{\rm c}$和$\beta_{\rm c}$分别表示因果注意力图中节点和边的注意力分数，$\alpha_{\rm u}$和$\beta_{\rm u}$分别表示非因果注意力图中节点和边的注意力分数。由于因果注意力图与非因果注意力图互补，所以因果注意力图与非因果注意力图中节点和边的注意力分数之和都为1。

最后，根据计算得到的节点和边的注意力分数$\alpha_{\rm c}$, $\alpha_{\rm u}$, $\beta_{\rm u}$, $\beta_{\rm c}$构建软掩码$\boldsymbol{M}_{\rm N}$, $\bar{\boldsymbol{M}}_{\rm N}$, $\boldsymbol{M}_{\rm E}$, $\bar{\boldsymbol{M}}_{\rm E}$。将初始的训练子图经过软掩码后分离出因果注意力图$G_{\rm c}$和非因果注意力图$G_{\rm u}$，计算过程为

$$G_{\rm c}=\{\boldsymbol{A}\odot\boldsymbol{M}_{\rm E},\boldsymbol{X}\odot\boldsymbol{M}_{\rm N}\}\tag{4}$$

$$G_{\rm u}=\{\boldsymbol{A}\odot\bar{\boldsymbol{M}}_{\rm E},\boldsymbol{X}\odot\bar{\boldsymbol{M}}_{\rm N}\}\tag{5}$$

根据式(1)–式(5)所示步骤，通过利用节点和边的注意力机制挖掘潜在的因果关系路径，从而对目标和影子GNN模型的初始训练子图实现因果图与非因果图的分离，提取包含因果特征的因果图。

3.2 因果图与非因果图解耦

初始训练子图经3.1节所示处理过程，被分离为因果图与非因果图。接下来，为了获取更丰富的节点表示，采用2层GNN逐步聚合图中节点的邻居信息，并使用一个多层感知器(MultiLayer Perceptron, MLP)作为分类器，对提取的信息进行分类预测，计算过程为

$$\boldsymbol{h}_{G_{\rm c}}=f({\rm GNN_c}(\boldsymbol{A}\odot\boldsymbol{M}_{\rm E},\boldsymbol{X}\odot\boldsymbol{M}_{\rm N}))\tag{6}$$

$$\boldsymbol{h}_{G_{\rm u}}=f({\rm GNN_u}(\boldsymbol{A}\odot\bar{\boldsymbol{M}}_{\rm E},\boldsymbol{X}\odot\bar{\boldsymbol{M}}_{\rm N}))\tag{7}$$

其中，$\boldsymbol{h}_{G_{\rm c}}$和$\boldsymbol{h}_{G_{\rm u}}$分别表示经过最后一层全连接层映射后因果特征与非因果特征的节点级向量表示，f表示图神经网络前向传播的聚合过程。

生成因果图的重点在于生成因果特征路径，即需要最大化因果特征与真实值之间的关联程度，因此，本文将其作为总损失函数计算的一部分，因果特征的分类损失函数计算过程为

$$L_{\rm c}=-\frac{1}{|N|}\sum_{G\in N}^{|N|}\boldsymbol{y}_G^{\rm T}\ln(\boldsymbol{h}_{G_{\rm c}})\tag{8}$$

其中，$L_{\rm c}$表示交叉熵损失值，N表示训练子图中节点的个数。

此外，在用于训练GNN模型的子图中，同时存在着非因果特征。非因果图可被视为对GNN节点分类任务不重要的普通范式。也就是说，将模型对非因果特征的预测结果随机分配给所有类别，不会对分类任务的结果产生实质性影响。非因果特征的分类损失函数计算过程为

$$L_{\mathrm{u}}=-\frac{1}{|N|}\sum_{G\in N}^{|N|}\mathrm{KL}(\boldsymbol{y}_{\mathrm{U}},\boldsymbol{h}_{G_{\mathrm{u}}}) \tag{9}$$

其中，$\mathrm{KL}(\cdot)$表示KL散度，为2个分布间的度量或距离函数，$\boldsymbol{y}_{\mathrm{U}}$表示均匀分布。

通过优化式(8)、式(9)这2个目标损失函数，实现因果注意力图与非因果注意力图的解耦操作。

3.3 后门路径调整策略

在GNN的结构因果模型(Structural Causal Model, SCM)中，后门路径是指从处理变量到结果变量之间的非因果路径($C\leftarrow G\rightarrow U\rightarrow R\rightarrow Y$)，如图3所示。这些路径通过共同的混淆变量连接起来，从而导致结果变量的混杂偏差。由于后门路径的存在，使得仅通过观察数据无法准确估计因果效应。因此，后门路径调整策略被用于消除这些非因果的关联，从而获得准确的因果关系估计。

为了避免非因果特征U对预测Y的干扰，后门路径调整策略通过对因果特征C进行干预运算(do-calculus)，从而消除后门路径的影响。

上文对因果图的提取过程专注于捕捉有用的因果特征，同时尽可能忽略混淆的非因果特征。然而，仅依靠解耦操作生成的因果图仍然无法完全收敛到完整的训练子图。本文进一步根据后门路径调整策略处理由非因果特征引起的因果关系估计问题。由于在初始的训练子图中很难直接观察得到非因果特征集，同时图数据又具有离散性。因此，首先对图中隐藏层的嵌入表示进行隐式操作，计算过程为

$$\boldsymbol{Z}_{G'}=\beta(\boldsymbol{h}_{G_C}+\boldsymbol{h}_{G_{\mathrm{u}}}) \tag{10}$$

其中，$\boldsymbol{Z}_{G'}$表示经过分类器β进行预测之后的隐层向量表示。得到该表示之后，根据后门路径调整策略，计算得到一个后门路径调整损失函数，计算过程为

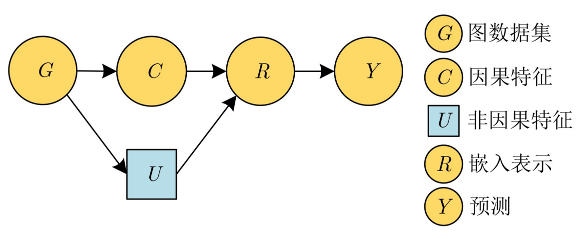

图 3 GNN中的结构因果模型

$$L_{\mathrm{cau}}=-\frac{1}{|\hat{S}|}\sum_{d'\in\hat{S}}\boldsymbol{y}_G^{\mathrm{T}}\ln(\boldsymbol{Z}_{G'}) \tag{11}$$

其中，$\hat{S}$表示非因果注意力图的近似估计集合。本文将式(11)得到的后门路径调整损失函数作为总的损失函数的一部分。

根据式(10)、式(11)的计算操作，可实现后门路径调整策略的效果，解决因果关系估计问题的影响。

3.4 因果关联图生成

本文将因果图提取、因果图与非因果图解耦以及后门路径调整策略这3个任务作为一个整体模型进行训练，并将子任务损失函数的加权相加作为该模型的多任务损失函数，计算过程为

$$L=L_{\mathrm{c}}+\lambda_1 L_{\mathrm{u}}+\lambda_2 L_{\mathrm{cau}} \tag{12}$$

其中，L表示该模型的总损失函数，λ_1和λ_2表示超参数，分别决定解耦和因果干预的强度，L_{c}, L_{u}和L_{cau}分别表示因果特征的分类损失函数、非因果特征的分类损失函数以及后门路径调整损失函数。

在计算出总的损失函数之后，GNN模型需要进行参数更新。通过随机梯度下降(Stochastic Gradient Descent, SGD)算法，求出损失函数对模型参数的梯度，同时根据链式法则更新参数并进行前向传播。GNN模型反复计算总的损失函数，并使用SGD算法进行反向传播，直至达到预设迭代次数或函数收敛。

综上，本文提出一种因果关联图生成算法(如算法1所示)。本算法通过生成目标因果关联图和影子因果关联图，从而用于构建校准的目标GNN模型和影子GNN模型。

算法1中，因果关联图的更新体现在每一轮迭

算法 1　因果关联图生成算法

输入：GNN模型初始的训练子图$G(G_{\mathrm{t}},\ G_{\mathrm{s}}\in G)$，迭代次数$T$

输出：目标因果关联图G_{target}，影子因果关联图G_{shadow}

(1) for $t=1$ to T

(2) 　G_{c}, $G_{\mathrm{u}}\leftarrow$ Attention(G) //因果图提取

(3) 　L_{c}, $L_{\mathrm{u}}\leftarrow$ Decouple(G) //因果图与非因果图解耦，生成对应损失函数

(4) 　$L_{\mathrm{cau}}\leftarrow$ Backdoor Adjustment(G_c, G_u) //后门路径调整，生成后门路径调整损失函数

(5) 　$L\leftarrow L_{\mathrm{c}}$, L_{u}, L_{cau} //计算模型总损失函数

(6) 　$\theta_{t+1}\leftarrow$ Update(θ_t) //更新模型参数

(7) 　$G_{t+1}\leftarrow G_t$ //迭代更新因果注意力图

(8) endfor

(9) G_{target}, $G_{\mathrm{shadow}}\leftarrow \mathrm{G}_T$ //生成目标因果关联图和影子因果关联图

(10) 结束算法返回目标因果关联图G_{target}，影子因果关联图G_{shadow}

代中，通过前向传播和反向传播对因果注意力图进行更新。每次迭代时，均在训练子图G上进行完整的前向传播和反向传播。

在前向传播过程中，利用注意力机制计算因果图与非因果图中节点和边的注意力分数，节点和边的注意力分数反映了训练子图中节点和边之间的关联程度。在反向传播过程中，通过优化总损失函数L调整模型参数，并迭代更新因果注意力图。最终，通过迭代更新因果注意力图生成因果关联图。

持续更新目的是确保模型在训练过程中不断提高其因果推断的能力。通过迭代更新因果注意力图生成因果关联图的过程，增强了模型区分因果特征和非因果特征的能力，同时降低非因果特征对模型预测输出的影响。

3.5 算法的复杂度分析

因果关联图生成算法的输入包括GNN模型初始的训练子图$G(G_{\rm t}, G_{\rm s}\in G)$和迭代次数$T$，主要步骤包括因果图提取、因果图与非因果图解耦、后门路径调整。

对于GNN模型初始的训练子图G，设图中节点的数量为N，N个节点最多可能有$N(N–1)$条有向边。在因果图提取过程中，每个节点需要与所有其他节点进行相关性计算，其时间复杂度为$O(N^2)$。在因果图与非因果图解耦过程中，需要遍历图中所有的节点及其邻居，其时间复杂度为$O(N^2)$。后门路径调整过程涉及对非因果特征的消除和后门路径的处理，复杂度取决于图结构，其时间复杂度为$O(N^2)$。

在上述算法执行过程中，空间消耗主要来自于节点邻接矩阵的存储、注意力分数的存储和节点嵌入表示，空间复杂度为$O(N^2)$。

4 影子模型构建

攻击者将用于训练影子GNN模型的影子训练子图经过算法1处理后，可以获得影子因果关联图$G_{\rm shadow}$。由于$G_{\rm shadow}$与用于训练目标GNN模型的目标因果关联图$G_{\rm target}$来自于同一数据集G，所以$G_{\rm shadow}$与$G_{\rm target}$具有相同的数据分布。于是，攻击者可以使用$G_{\rm shadow}$构建并训练一个校准的影子GNN模型(简称影子模型)，用于模拟已校准目标GNN模型(简称目标模型)在节点分类任务上的预测行为。

影子模型的构建通常假设攻击者对目标模型的架构和训练过程有一定的了解，影子模型构建的过程包括以下3个步骤：

(1) 从$G_{\rm shadow}$中选择训练子集：攻击者首先需要从影子因果关联图$G_{\rm shadow}$中选择一部分数据$G_{\rm shadow}^{\rm train}$用于训练影子模型。在选择影子训练子集的方案中，需要确保影子模型的训练数据集与目标模型的训练数据集在特征和模式的覆盖上拥有足够的相似性。

(2) 训练影子模型：攻击者使用选出的影子训练数据集$G_{\rm shadow}^{\rm train}$训练影子模型，根据攻击者拥有的外部知识，影子模型的架构选定为与目标模型相同或者相似。在影子模型的训练过程中，目标是使其尽可能地模拟目标模型的预测行为。为了实现这一目标，攻击者需要将影子模型训练到与目标模型具有相似的性能水平。

(3) 使用目标模型对$G_{\rm shadow}^{\rm train}$的后验概率输出作为训练影子模型的真实值标签：攻击者首先利用目标模型对$G_{\rm shadow}^{\rm train}$中的每个节点进行预测，得到每个节点的后验概率分布。然后，将这些后验概率作为标签，用于训练影子模型。这样做目的是使影子模型更好地模拟目标模型对相同节点的预测行为，从而为后续攻击模型训练数据集的构建过程提供更加准确的内容。

通过以上3步，影子模型被训练为一个能够模仿目标模型预测行为的代理模型。该模型将被用于生成攻击模型的训练数据集，并作为推断目标模型是否使用特定数据节点作为训练数据集的依据。

5 攻击模型构建与成员推理攻击

本文攻击方法是一种基于2元分类器的成员推理攻击方法，在此类攻击方法中，攻击者需要构建一个攻击模型A，以对目标模型实施成员推理攻击，推断目标节点是否来自于目标模型的训练数据集。

在构建攻击模型的过程中，首先需要构建用于训练攻击模型的攻击模型数据集。攻击模型数据集的构建过程如下：

在影子模型构建过程中，影子模型的训练数据集被选定后，攻击者将$G_{\rm shadow}$分为$G_{\rm shadow}^{\rm train}$和$G_{\rm shadow}^{\rm test}$(即$G_{\rm shadow}$中未用于训练影子模型的那部分数据节点)。然后，攻击者将$G_{\rm shadow}^{\rm train}$和$G_{\rm shadow}^{\rm test}$分别输入影子模型，并获取全部的后验概率。这些后验概率构成了攻击模型的训练数据集，用于训练攻击模型A。其中，$G_{\rm shadow}^{\rm train}$中节点对应的后验概率被标记为‘1’(代表成员节点)，而$G_{\rm shadow}^{\rm test}$中节点对应的后验概率被标记为‘0’(代表非成员节点)。这些标记帮助攻击模型A学习如何根据后验概率分布，判断目标节点是否为目标模型训练数据集的一部分。

接下来，攻击者使用攻击模型数据集构建并训练一个MLP作为攻击模型A。训练好的攻击模型A将被攻击者用于实施对目标模型的成员推理攻击，攻击的具体过程如下：

(1) 攻击者从目标数据集中选择一个节点v及其已知的邻居节点，将其输入到目标模型，获取节点

v及其邻居节点在目标模型中输出的后验概率(即目标节点攻击特征)。

(2) 攻击者将获取到的后验概率作为攻击模型A的输入，攻击模型A根据目标节点的攻击特征判断该节点是否为成员节点。

(3) 攻击模型A输出一个“1”或“0”的预测值，其中，“1”表示目标节点v是成员节点，“0”表示目标节点v不是成员节点。

6 实验与分析

6.1 实验设置

考虑到在实际应用中，特别是涉及敏感数据(如，医疗记录)的场景，研究人员通常只在敏感数据的一个子图上训练GNN模型，以保护患者的个人隐私不被泄露。这导致在研究面向GNN模型的MIAs时通常遇到2种不同的数据处理和模型评估设置。因此，本文设置2种攻击模式，以开展本文方法的成员推理攻击实验，分别是：

(1) 攻击模式1：GNN模型在子图上进行训练，在完整图上进行测试。

(2) 攻击模式2：GNN模型在子图上进行训练，同时也在子图上进行测试。

为验证本文方法的有效性，在验证实验中采用4种评估GNN模型性能的基准数据集：Cora数据集、CiteSeer数据集、PubMed数据集[15]和Flickr数据集[16]。各数据集的统计信息如表1所示。

本文的目标模型和影子模型分别使用由目标因果关联图和影子因果关联图训练的2层图卷积网络(Graph Convolutional Network, GCN)[17]、图注意力网络(Graph ATtention network, GAT)[18]、图采样与聚合(Graph SAmple and aggreGatE, GraphSAGE)[19]和简化图卷积(Simplified Graph Convolution, SGC)[20]模型架构，模型的训练轮次为300轮，训练学习率设置为0.000 1；攻击模型使用3层的MLP模型，模型的训练轮次为100轮，训练学习率设置为0.01。此外，使用由初始训练子图训练的影子模型的后验概率输出，构建相对应的基准攻击模型[8]，作为基线攻击方法，与本文方法MIAs-MC进行对比分析，以评估本文方法的攻击性能。

在实验中，为了确保模型均衡处理因果特征和非因果特征，将损失函数的超参数λ_1和λ_2均设置为0.5。上述两个参数的设置，既要考虑模型在解耦因果特征与非因果特征过程中分类损失函数的权重平衡，也要确保模型在因果干预过程中的调整力度恰当。

表 1 数据集的统计信息

数据集	类别数	节点数	边数	节点特征维度	使用节点数
Cora	7	2 708	5 429	1 433	2 520
CiteSeer	6	3 327	4 732	3 703	2 400
PubMed	3	19 717	44 338	500	18 000
Flickr	7	89 250	449 878	500	42 000

通过赋予λ_1和λ_2相同的数值，使模型在训练初期均衡处理解耦因果特征与非因果特征以及因果干预这两个关键过程，从而能有效区分因果特征与非因果特征，通过后门路径调整抑制非因果特征的影响。最终，在生成因果关联图时，既能确保因果特征的精确度，也能减少非因果特征对模型预测的干扰。

本文采用攻击准确率(Accuracy)、攻击精确率(Precision)、曲线下面积(Area Under the Curve, AUC)、攻击召回率(Recall)、攻击F1分数(F1-score)和攻击漏报率(False Negative Rate, FNR)6个常见的指标衡量本文方法的攻击性能。

6.2 实验结果与分析

为验证本文方法的攻击性能，在Cora数据集、CiteSeer数据集、PubMed数据集和Flickr数据集上分别训练GCN, GAT, GraphSAGE和SGC 4种不同架构的GNN节点分类模型作为本实验的目标模型和影子模型，从而通过影子模型实现对相应目标模型的成员推理攻击实验。

为验证本文方法的可行性和有效性，分别在设置的2种攻击模式下进行成员推理攻击实验。在攻击模式1下，在4个数据集上进行本文所提攻击方法MIAs-MC和基线攻击方法的攻击实验，使用攻击准确率、攻击精确率、AUC、攻击召回率和攻击F1分数5个指标分析实验结果。在攻击模式2下，在4个数据集上进行本文所提攻击方法MIAs-MC和基线攻击方法的攻击实验，使用攻击准确率、攻击精确率和攻击漏报率3个指标分析实验结果。

6.2.1 攻击模式1下的成员推理攻击实验

在攻击模式1下，尽管目标GNN模型仅限制在图数据的一个子集上进行训练，但模型的性能和泛化能力将在完整的图数据上进行测试。这样可以评估模型对未见过数据的处理能力，但同时也增加了通过分析模型输出推断训练数据(即，成员推理攻击)的可能性。在攻击模式1下，本文所提攻击方法MIAs-MC具体实验结果如表2所示，基线攻击方法的攻击实验结果如表3所示。

由表2和表3可见，本文攻击方法针对在4种数据集上分别训练出的4种不同架构的GNN模型均取得了较好的攻击效果，攻击的关键指标均优于基线攻击方法。

表 2 攻击模式1下MIAs-MC的攻击结果

数据集	GNN架构	Accuracy	Precision	AUC	Recall	F1-score
Cora	GCN	0.926	0.920	0.912	0.913	0.912
	GAT	0.911	0.914	0.910	0.911	0.911
	GraphSAGE	0.905	0.908	0.904	0.905	0.905
	SGC	0.914	0.923	0.915	0.914	0.914
CiteSeer	GCN	0.918	0.912	0.917	0.918	0.918
	GAT	0.857	0.879	0.857	0.857	0.855
	GraphSAGE	0.933	0.936	0.931	0.933	0.933
	SGC	0.930	0.938	0.929	0.930	0.930
PubMed	GCN	0.750	0.784	0.750	0.751	0.743
	GAT	0.642	0.686	0.643	0.642	0.621
	GraphSAGE	0.748	0.754	0.747	0.748	0.748
	SGC	0.690	0.702	0.691	0.690	0.690
Flickr	GCN	0.841	0.846	0.841	0.841	0.841
	GAT	0.786	0.801	0.787	0.786	0.785
	GraphSAGE	0.732	0.764	0.732	0.732	0.725
	SGC	0.907	0.916	0.908	0.907	0.907

表 3 攻击模式1下基线攻击方法的攻击结果

数据集	GNN架构	Accuracy	Precision	AUC	Recall	F1-score
Cora	GCN	0.763	0.770	0.764	0.763	0.763
	GAT	0.721	0.728	0.718	0.721	0.720
	GraphSAGE	0.825	0.837	0.825	0.825	0.824
	SGC	0.806	0.812	0.808	0.806	0.807
CiteSeer	GCN	0.860	0.865	0.859	0.860	0.860
	GAT	0.772	0.775	0.769	0.772	0.771
	GraphSAGE	0.858	0.875	0.859	0.858	0.827
	SGC	0.863	0.868	0.862	0.863	0.863
PubMed	GCN	0.647	0.655	0.647	0.647	0.647
	GAT	0.593	0.612	0.593	0.593	0.580
	GraphSAGE	0.554	0.560	0.553	0.554	0.553
	SGC	0.664	0.685	0.665	0.664	0.658
Flickr	GCN	0.774	0.805	0.775	0.774	0.769
	GAT	0.601	0.613	0.602	0.601	0.598
	GraphSAGE	0.689	0.755	0.688	0.689	0.668
	SGC	0.877	0.893	0.878	0.877	0.876

在Cora数据集上，与基线攻击方法的结果相比，本文攻击方法针对GCN, GAT, GraphSAGE和SGC模型的攻击准确率的提升幅度为9.7%～26.4%。从攻击结果中可以看出，针对GCN的攻击表现最好，针对GraphSAGE的攻击表现最差，二者在攻击准确率上的差值为2.1%。

在CiteSeer数据集上，与基线攻击方法的结果相比，本文攻击方法针对GCN, GAT, GraphSAGE和SGC模型的攻击准确率的提升幅度为6.7%～11%。从攻击结果中可以看出，针对GraphSAGE的攻击表现最好，针对GAT的攻击表现最差，二者在攻击准确率上的差值为7.6%。

在PubMed数据集上，与基线攻击方法的结果相比，本文攻击方法针对GCN, GAT, GraphSAGE和SGC模型的攻击准确率的提升幅度分别为3.9%～35%。从攻击结果中可以看出，针对GCN的攻击表现最好，针对GAT的攻击表现最差，二者在攻击准确率上的差值为10.8%。

在Flickr数据集上，与基线攻击方法的结果相比，本文攻击方法针对GCN, GAT, GraphSAGE和SGC模型的攻击准确率的提升幅度分别为3.4%～30.8%。从攻击结果中可以看出，针对SGC的攻击表现最好，针对GraphSAGE的攻击表现最差，二者在攻击准确率和攻击精确率上的差值为17.5%和15.2%。

以Cora数据集和PubMed数据集为例，通过计算目标模型和影子模型在不同GNN架构下的训练准确率和测试准确率的差值，定量评估影子模型模仿目标模型预测行为的准确性，结果如表4和表5所示。

表4和表5中的差值在一定程度上可以反映模型校准与影子模型模仿目标模型预测行为准确性之间的关系，即：在不同的GNN架构下，本文提出的基于因果推断的模型校准方法，可较有效地降低影子模型与目标模型在训练和测试准确率上的差异。例如，当本文攻击方法实施攻击时，在Cora数据集上训练的GraphSAGE模型经过模型校准后，影子模型与目标模型的测试准确率差值仅为0.80%；而在未经模型校准的基线攻击下，该差值为4.92%。

上述实验结果表明，经过模型校准后，影子模型能更好地模仿目标模型的预测行为，同时有助于增加对目标模型实施MIAs的攻击成功率。

表 4 Cora数据集上影子模型与目标模型准确率差异(%)

GNN架构	基线攻击下训练准确率差值	基线攻击下测试准确率差值	模型校准后训练准确率差值	模型校准后测试准确率差值
GCN	**0.32**	3.97	0.79	**0.95**
GAT	3.65	**1.99**	**1.91**	2.22
GraphSAGE	0.32	4.92	**0.16**	**0.80**
SGC	**0.66**	**0.47**	1.11	1.70

综上所述，在攻击模式1下，本文所提攻击方法MIAs-MC在目标和影子GNN模型的训练过程中增加对因果特征的提取过程后，实现了模型校准这一目标，攻击者利用校准的影子GNN模型，可较准确地推断目标节点是否属于目标GNN模型的训练数据。这表明在减少或消除GNN模型训练数据的非因果特征后，通过训练影子GNN模型并构建攻击模型能够增加泄露目标GNN模型训练数据的隐私风险。

6.2.2 攻击模式2下的成员推理攻击实验

在攻击模式2下，目标GNN模型的训练和测试都仅限于使用同一个子图，这种模式增强了数据的隔离性，有助于保护GNN模型训练数据的隐私安全。本文攻击方法和基线攻击方法在攻击模式2下的实验结果如图4–图6所示。

实验结果表明，在攻击模式2下，本文攻击方法实施MIAs的关键评价指标皆优于基线攻击方法实施MIAs的结果，其中，攻击准确率比基线攻击方法的攻击结果提高0.5%～32.1%，攻击精确率提高0.3%～31.8%，攻击漏报率平均降低31.2%。

在Cora数据集和CiteSeer数据集上，本文攻击方法对所有目标GNN模型的攻击准确率和攻击精确率均高于0.8；在PubMed数据集和Flickr数据集上，本文攻击方法对所有目标GNN模型的攻击准确率和攻击精确率则分别在0.56～0.753和0.593～0.776。以上实验结果显示，在拓扑结构较简单的图数据集(Cora数据集和CiteSeer数据集)上训练的GNN模型相比在拓扑结构复杂的图数据集(PubMed数据集和Flickr数据集)上训练的GNN模型更容易受到MIAs的威胁。

综上，即使目标GNN模型训练和测试阶段仅使用同一个数据子图，本文所提攻击方法MIAs-MC也可以成功实施MIAs，且进一步降低攻击的漏报率，增加了MIAs对GNN模型训练数据安全的威胁。

表 5 PubMed数据集上影子模型与目标模型准确率差异(%)

GNN架构	基线攻击下训练准确率差值	基线攻击下测试准确率差值	模型校准后训练准确率差值	模型校准后测试准确率差值
GCN	1.45	0.75	**1.15**	**0.14**
GAT	**0.36**	1.15	0.82	**0.51**
GraphSAGE	0.20	5.12	**0.13**	**3.15**
SGC	1.58	0.60	**0.73**	**0.56**

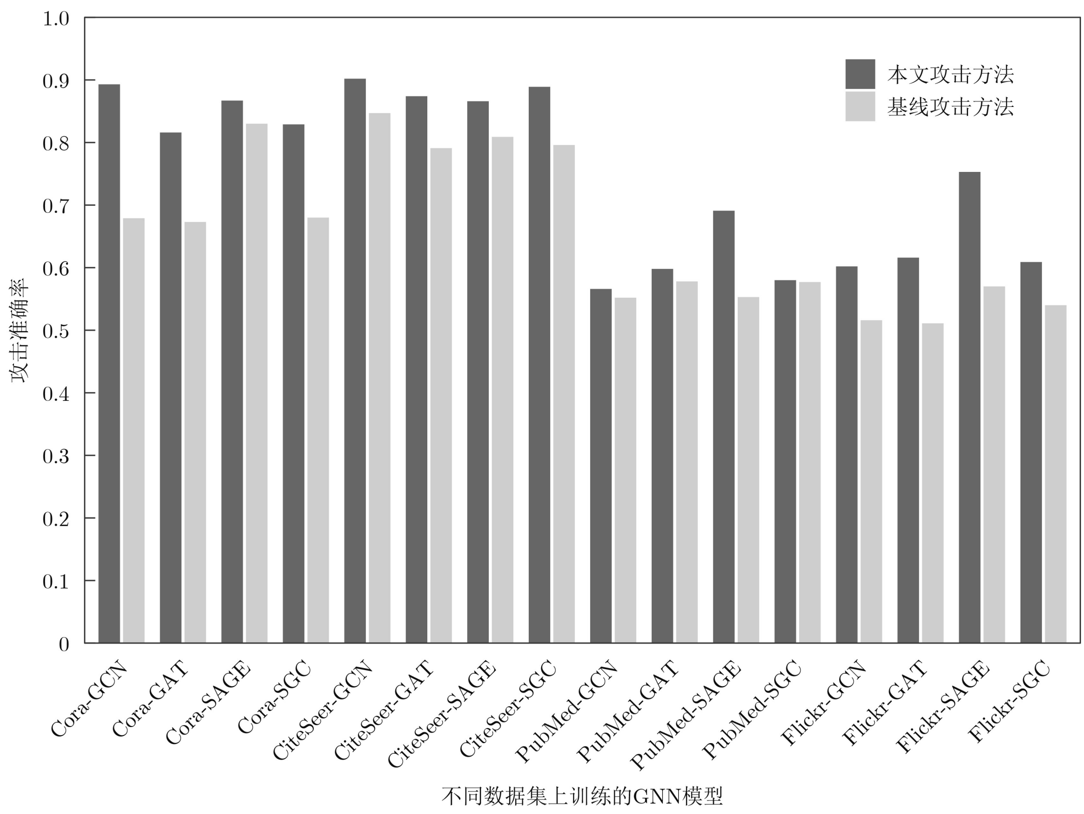

图 4 攻击模式2下MIAs的攻击准确率

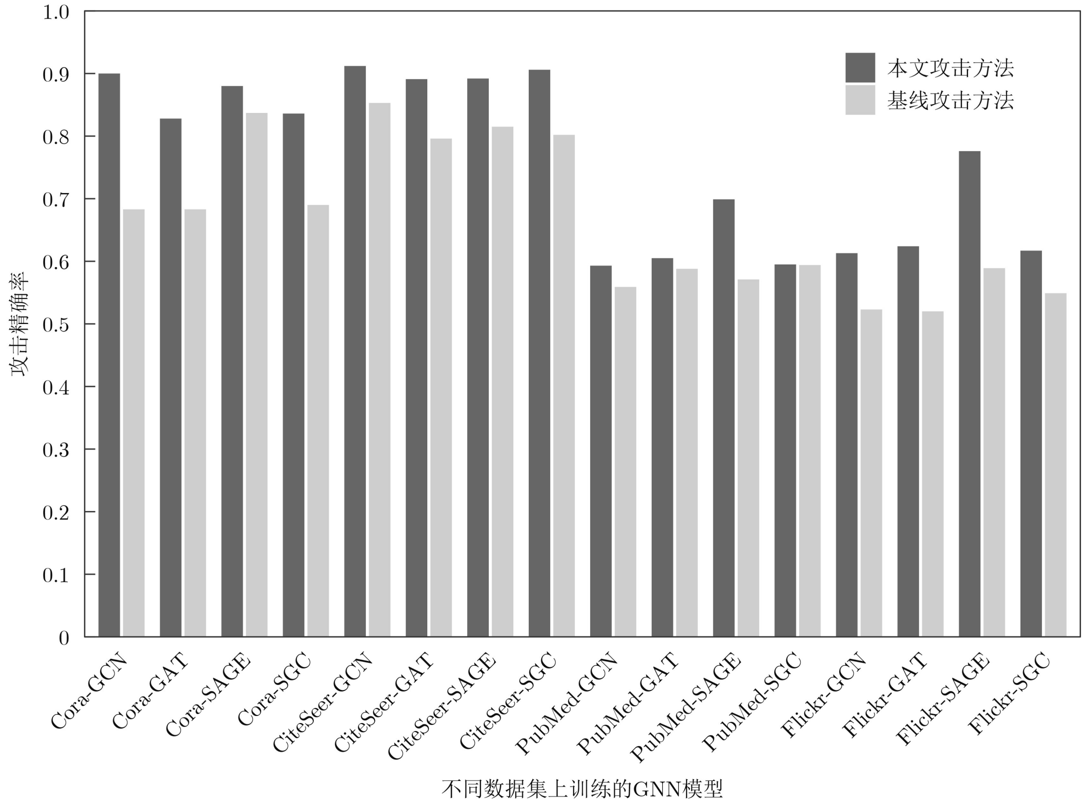

图 5 攻击模式2下MIAs的攻击精确率

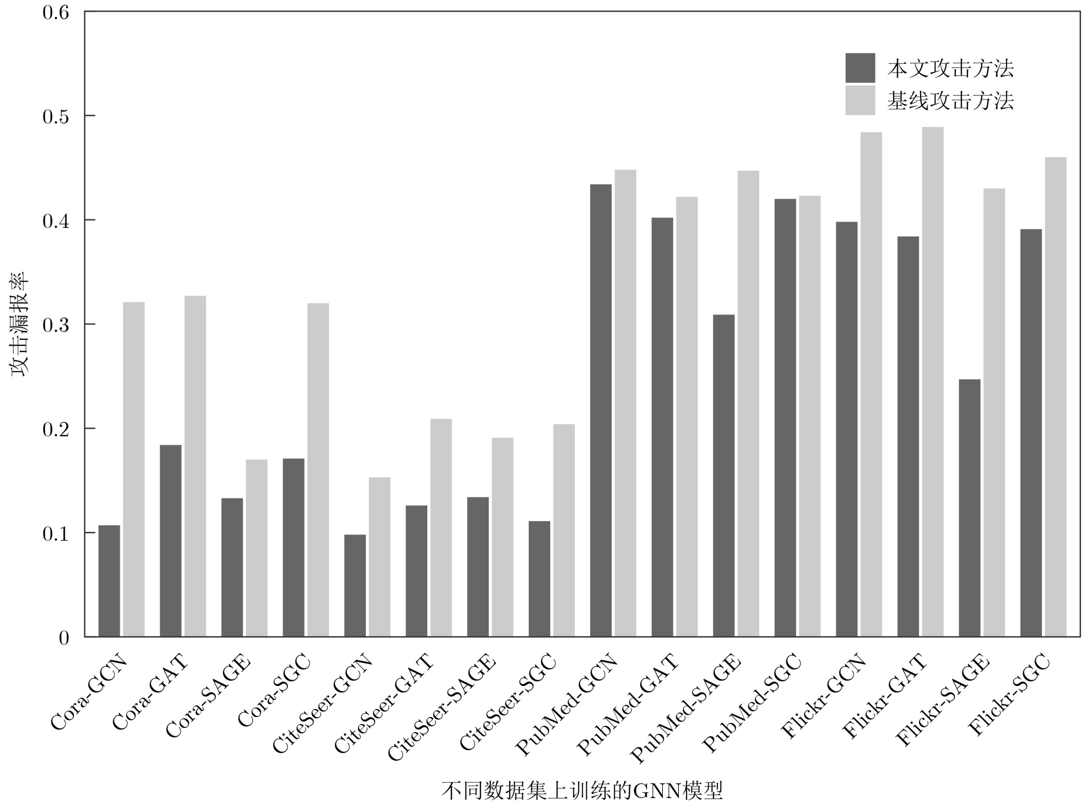

图 6 攻击模式2下MIAs的攻击漏报率

6.2.3 两种攻击模式下的实验结果对比分析

从攻击模式1和2下的实验结果中可以看出，本文所提攻击方法MIAs-MC在攻击模式1下进行成员推理攻击实验的攻击性能普遍优于在攻击模式2下进行成员推理攻击实验的攻击性能，其原因分析如下。

(1) 在攻击模式1下，即使GNN模型的训练过程仅在子图上进行，但攻击者在模型的测试阶段依然可以访问完整的图结构。这意味着攻击者在进行成员推理攻击时，可以利用完整图数据的信息，包括训练子图外的节点和边。通过利用这些额外的信息，攻击者能够更准确地推断出目标GNN模型的成员节点，从而对GNN模型训练数据的隐私安全造成更大的威胁。

(2) 相比于攻击模式1下攻击者在GNN模型测试阶段可以访问完整图数据的信息，攻击模式2下攻击者在测试阶段被限制为仅能访问与模型训练时相同的子图。这限制了攻击者观察模型更多行为和模式的能力，因此在推断成员时可供攻击者利用的信息较少，导致在攻击模式2下，本文攻击方法和基线攻击方法进行的成员推理攻击性能均劣于在攻击模式1下进行的成员推理攻击性能。

6.2.4 攻击效果差异的原因分析

上述实验结果表明，本文所提攻击方法MIAs-MC在不同的GNN架构和数据集上均取得了显著的攻击效果，性能指标优于现有基线攻击方法。但是，在不同的数据集和模型架构下，MIAs-MC的攻击效果存在一定差异。该方法的攻击效果存在差异的原因为：

(1) 图数据集结构特性的不同。攻击效果存在差异与图数据集的结构特性密切相关。节点数和边数较少的图数据集(如Cora数据集和CiteSeer数据集)节点密度较低、拓扑结构相对简单，在此类图数据集上构建的影子GNN模型与目标GNN模型的预测行为相似度较高。通过利用与目标GNN模型预测行为相似度较高的影子GNN模型构建攻击模型，使其能够从影子GNN模型的输出中准确提取成员信息，从而有效地推断哪些节点属于GNN模型的训练数据集。相反，对于节点密度较高、拓扑结构相对复杂的图数据集(如PubMed数据集和Flickr数据集)，在此类图数据集上构建的影子GNN模型对目标GNN模型的模拟效果可能不佳。此时，攻击模型难以从影子GNN模型的输出中提取有效成员信息，导致攻击模型在推断成员状态时误差较大，攻击准确率较低。

(2) GNN架构的不同。攻击效果存在差异与不同GNN架构在信息聚合和特征提取过程中的机制差异密切相关。本文研究的4种GNN模型中，GCN和SGC采用较简单的图卷积操作，通过邻居节点加权求和实现节点特征的聚合。GCN和SGC采用的信息聚合机制使影子GNN模型能够直接捕捉目标GNN模型输出中的关键信息，其概率输出与训练数据的成员标签匹配度较高。因此，攻击模型可有效地从影子GNN模型中学习成员标签与数据特征之间的映射关系，提高攻击的准确性。相对而言，GraphSAGE和GAT采用更复杂的信息聚合机制，前者通过样本化邻居节点进行特征聚合，后者利用自注意力机制动态调整邻居节点的权重。这些复杂机制使影子GNN模型的概率输出包含更多噪声特征，降低了影子GNN模型的概率输出与训练数据成员标签之间的匹配度，导致攻击模型较难从影子GNN模型中提取影响成员推理攻击的关键信息，降低攻击的准确性。

7　结束语

针对GNN模型预测的欠自信状态导致成员推理攻击难度大和攻击漏报率高等问题，本文提出一种基于GNN模型校准的成员推理攻击方法。通过设计一种基于因果推断的GNN模型校准方法构建因果关联图，利用因果关联图构建影子GNN模型和攻击模型，从而对目标GNN模型实施成员推理攻击。实验结果表明，本文提出的攻击方法对目标GNN模型的攻击效果优于现有攻击方法，并能够有效降低成员推理攻击的漏报率。

本文提出的攻击方法中基于因果推断的模型校准方法依赖于因果关联图的生成过程，该过程在大规模数据集上时间开销较大，导致整体攻击过程耗时较长。在未来的研究中，将探索更高效、更准确的GNN模型校准方法，进一步提升成员推理攻击的攻击效果并降低攻击过程的时间开销。

参 考 文 献

[1] SHOKRI R, STRONATI M, SONG Congzheng, *et al*. Membership inference attacks against machine learning models[C]. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, USA, 2017: 3–18. doi: 10.1109/SP.2017.41.

[2] SALEM A, ZHANG Yang, HUMBERT M, *et al*. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models[C]. The Network and Distributed System Security Symposium (NDSS), San Diego, USA, 2019: 24–27.

[3] LONG Yunhui, WANG Lei, BU Diyue, *et al*. A pragmatic approach to membership inferences on machine learning models[C]. 2020 IEEE European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, 2020: 521–534. doi: 10.1109/EuroSP48549.2020.00040.

[4] KO M, JIN Ming, WANG Chenguang, *et al*. Practical membership inference attacks against large-scale multi-modal models: A pilot study[C]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 4848–4858. doi: 10.1109/ICCV51070.2023.00449.

[5] CHOQUETTE-CHOO C A, TRAMER F, CARLINI N,

et al. Label-only membership inference attacks[C]. The 38th International Conference on Machine Learning, 2021: 1964–1974.

[6] LIU Han, WU Yuhao, YU Zhiyuan, *et al*. Please tell me more: Privacy impact of explainability through the lens of membership inference attack[C]. 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, USA, 2024: 120–120. doi: 10.1109/SP54263.2024.00120.

[7] 吴博, 梁循, 张树森, 等. 图神经网络前沿进展与应用[J]. 计算机学报, 2022, 45(1): 35–68. doi: 10.11897/SP.J.1016.2022.00035.

WU Bo, LIANG Xun, ZHANG Shusen, *et al*. Advances and applications in graph neural network[J]. *Chinese Journal of Computers*, 2022, 45(1): 35–68. doi: 10.11897/SP.J.1016.2022.00035.

[8] OLATUNJI I E, NEJDL W, and KHOSLA M. Membership inference attack on graph neural networks[C]. 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, USA, 2021: 11–20. doi: 10.1109/TPSISA52974.2021.00002.

[9] HE Xinlei, WEN Rui, WU Yixin, *et al*. Node-level membership inference attacks against graph neural networks[EB/OL]. https://arxiv.org/abs/2102.05429, 2021.

[10] WU Bang, YANG Xiangwen, PAN Shirui, *et al*. Adapting membership inference attacks to GNN for graph classification: Approaches and implications[C]. 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021: 1421–1426. doi: 10.1109/ICDM51629.2021.00182.

[11] WANG Xiuling and WANG W H. Link membership inference attacks against unsupervised graph representation learning[C]. The 39th Annual Computer Security Applications Conference, Austin, USA, 2023: 477–491. doi: 10.1145/3627106.3627115.

[12] WANG Xiao, LIU Hongrui, SHI Chuan, *et al*. Be confident! Towards trustworthy graph neural networks via confidence calibration[C]. The 35th Conference on Neural Information Processing Systems, 2021: 1820.

[13] HSU H H H, SHEN Y, TOMANI C, *et al*. What makes graph neural networks miscalibrated?[C]. The 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1001.

[14] LIU Tong, LIU Yushan, HILDEBRANDT M, *et al*. On calibration of graph neural networks for node classification[C]. 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022: 1–8. doi: 10.1109/IJCNN55064.2022.9892866.

[15] YANG Zhilin, COHEN W W, and SALAKHUTDINOV R. Revisiting semi-supervised learning with graph embeddings[C]. The 33rd International Conference on International Conference on Machine Learning, New York, USA, 2016: 40–48.

[16] ZENG Hanqing, ZHOU Hongkuan, SRIVASTAVA A, *et al*. GraphSAINT: Graph sampling based inductive learning method[C]. The 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020: 1–19.

[17] KIPF T N and WELLING M. Semi-supervised classification with graph convolutional networks[C]. The 5th International Conference on Learning Representations, Toulon, France, 2017: 1–14.

[18] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, *et al*. Graph attention networks[C]. The 6th International Conference on Learning Representations, Vancouver, Canada, 2018: 1–12.

[19] HAMILTON W, YING Z, and LESKOVEC J. Inductive representation learning on large graphs[C]. Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 1025–1035.

[20] WU F, SOUZA A, ZHANG Tianyi, *et al*. Simplifying graph convolutional networks[C]. The 36th International Conference on Machine Learning, Long Beach, USA, 2019: 6861–6871.

谢丽霞：女，硕士，教授，研究方向为网络信息安全.

史镜琛：男，硕士生，研究方向为人工智能安全.

杨宏宇：男，博士，教授，博士生导师，研究方向为网络与系统安全、软件安全、网络安全态势感知.

胡　泽：男，博士，讲师，研究方向为人工智能、自然语言处理、网络信息安全.

成　翔：男，博士，讲师，研究方向为网络与系统安全、网络安全态势感知、APT攻击检测.

责任编辑：余　蓉

Membership Inference Attacks Based on Graph Neural Network Model Calibration

XIE Lixia[①] SHI Jingchen[①] YANG Hongyu[①②] HU Ze[②] CHENG Xiang[③]

[①](*School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*)

[②](*School of Safety Science and Engineering, Civil Aviation University of China, Tianjin* 300300, *China*)

[③](*School of Information Engineering, Yangzhou University, Yangzhou* 225127, *China*)

Abstract:

Objective Membership Inference Attacks (MIAs) against machine learning models represent a significant threat to the privacy of training data. The primary goal of MIAs is to determine whether specific data samples are part of a target model's training set. MIAs reveal potential privacy vulnerabilities in artificial intelligence models, making them a critical area of research in AI security. Investigating MIAs not only helps security researchers assess model vulnerabilities to such attacks but also provides a theoretical foundation for establishing guidelines for the use of sensitive data and developing strategies to improve model security. In recent years, Graph Neural Network (GNN) models have become a key focus in MIAs research. However, GNN models often exhibit under-confidence in their predictions, marked by cautious probability distributions in model outputs. This issue prevents existing MIAs methods from fully utilizing posterior probability information, resulting in reduced attack accuracy and higher false negative rates. These challenges significantly limit the effectiveness and applicability of current attack methods. Therefore, addressing the under-confidence problem in GNN predictions and developing enhanced MIA approaches to improve attack performance has become both necessary and urgent.

Methods Given that GNN models are often characterized by under-confidence in their predictions, which hampers the implementation of MIAs and resulting in high false negative rates, an MIAs method based on GNN Model Calibration (MIAs-MC) is proposed (Fig. 1). First, a GNN model calibration method based on causal inference is designed and applied. This approach involves extracting causal graphs using an attention mechanism, decoupling causal and non-causal graphs, applying a backdoor adjustment strategy, and generating causal association graphs, which are then used to train the GNN model (Fig. 2). Next, a shadow GNN model is constructed using shadow causal association graphs that share the same data distribution as the target causal association graph, enabling the shadow models to mimic the performance of the target GNN model. Finally, posterior probabilities from the shadow GNN model are used to create an attack dataset, which is employed to train an attack model. This attack model is then used to infer whether a target node is part of the training data of the target GNN model, based on the posterior probabilities generated by the target GNN model.

Results and Discussions To assess the feasibility and effectiveness of the proposed attack method, two attack modes are implemented in the experiment, and MIAs are conducted under both modes. The experimental results demonstrate that the proposed method consistently outperforms the baseline attack method across various metrics. In Attack Mode 1, the proposed method is evaluated on the Cora, CiteSeer, PubMed, and Flickr datasets, with comparative results presented against the baseline method (Table 2 and Table 3). Compared to the baseline attack method, the proposed method achieves improvements in attack accuracy and attack precision for GCN, GAT, GraphSAGE, and SGC models, ranging from 3.4% to 35.0% and 1.2% to 34.6%, respectively. Furthermore, the results indicate that after GNN model calibration, the shadow model more effectively mimics the prediction behavior of the target model, contributing to an increased success rate of MIAs on the target model (Table 4 and Table 5). Notably, the GAT model exhibits high robustness against MIAs, both for the proposed and baseline methods. In Attack Mode 2, the attack performance of the proposed method is compared with the baseline method across the same datasets (Cora, CiteSeer, PubMed, and Flickr) (Fig. 4, Fig. 5, and Fig. 6). The proposed method improves attack accuracy by 0.4% to 32.1%, attack precision by 0.3% to 31.8%, and reduces the average attack false negative rate by 31.2%, compared to the baseline methods. Overall, the results from both attack modes indicate that calibrating the GNN model and training the attack model with the calibrated GNN posterior probabilities significantly enhances the performance of MIAs. However, the attack performance varies across different datasets and model architectures. Analysis of the experimental results reveals that the effectiveness of the proposed method is influenced by the structural characteristics of the graph datasets and the specific configurations of the GNN architectures.

Conclusions The proposed MIAs method, based on GNN model calibration, constructs a causal association graph using a calibration technique rooted in causal inference. This causal association graph is subsequently used to build shadow GNN models and attack models, facilitating MIAs on target GNN models. The results verify that GNN model calibration enhances the effectiveness of MIAs.

Key words: Graph Neural Network (GNN); Membership Inference Attacks (MIAs); Model calibration; Causal inference; Privacy risk