面向图神经网络的隐私安全综述

陈晋音[1,2], 马敏樱[2], 马浩男[2], 郑海斌[1,2]

[1]浙江工业大学网络空间安全研究院 杭州 中国 310023
[2]浙江工业大学信息工程学院 杭州 中国 310023

摘要 图神经网络 (Graph Neural Network, GNN) 对图所包含的边和节点数据进行高效信息提取与特征表示, 因此对处理图结构数据具有先天优势。目前, 图神经网络已经在许多领域(如社交网络、自然语言处理、计算机视觉甚至生命科学等领域)得到了非常广泛的应用, 极大地促进了人工智能的繁荣与发展。然而, 已有研究表明, 攻击者可以发起对训练数据或目标模型的隐私窃取攻击, 从而造成隐私泄露风险甚至财产损失。因此探究面向 GNN 的隐私安全获得广泛关注, 陆续研究提出了一系列方法挖掘 GNN 的安全漏洞, 并提供隐私保护能力。然而, 对 GNN 隐私问题的研究相对零散, 对应的威胁场景、窃取方法与隐私保护技术、应用场景均相对独立, 尚未见系统性的综述工作。因此, 本文首次围绕 GNN 的隐私安全问题展开分析, 首先定义了图神经网络隐私攻防理论, 其次按照模型输入、攻防原理、下游任务、影响因素、数据集、评价指标等思路对隐私攻击方法和隐私保护方法进行分析归纳, 整理了针对不同任务进行的通用基准数据集与主要评价指标, 同时, 讨论了 GNN 隐私安全问题的潜在应用场景, 分析了 GNN 隐私安全与图像或自然语言处理等深度模型的隐私安全的区别与关系, 最后探讨了 GNN 的隐私安全研究当前面临的挑战, 以及未来潜在研究方向, 以进一步推动 GNN 隐私安全研究的发展和应用。

关键词 图神经网络; 推断攻击; 隐私保护; 重构攻击; 隐私安全
中图法分类号 TP18 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2025.05.08

A Survey of Privacy Security for Graph Neural Networks

CHEN Jinyin[1,2], MA Minying[2], MA Haonan[2], ZHENG Haibin[1,2]

[1]Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract Graph neural network (GNN) performs efficient information extraction and feature representation on the edge and node data contained in the graph, so it has inherent advantages in processing graph structure data. At present, graph neural networks have been widely used in many fields (such as social networks, natural language processing, computer vision and even life sciences), greatly promoting the prosperity and development of artificial intelligence. However, existing research has shown that attackers can launch privacy theft attacks on training data or target models, resulting in privacy leak risks and even property losses. Therefore, exploring the privacy security of graph neural network has attracted widespread attention, and a series of methods have been proposed to mine the security vulnerabilities of graph neural network and provide privacy protection capabilities. However, research on graph neural network privacy issues is relatively scattered, and the corresponding threat scenarios, stealing methods, privacy protection technologies, and application scenarios are relatively independent, and there is no systematic review work yet. Therefore, this article analyzes the privacy security issues of graph neural networks for the first time. Firstly, it defines the privacy attack and defense theory of graph neural network. Secondly, it analyzes and summarizes the privacy attack methods and privacy protection methods according to the ideas of model input, attack and defense mechanisms, downstream tasks, influencing factors, datasets, evaluations, organizes general benchmark datasets and main evaluations for different tasks. At the same time, the potential application scenarios of graph neural network privacy security issues were discussed, and the differences and relationships between graph neural network privacy security and privacy security of deep models such as image or natural language processing were analyzed. Finally, the current challenges faced by GNN privacy security research and the future were discussed. Potential research directions to further promote the development and application of graph neural network privacy security research.

Key words graph neural network; inference attack; privacy protection; reconstruction attack; privacy security

通讯作者: 郑海斌, 博士, 讲师, Email: haibinzheng320@gmail.com。
本课题得到浙江省自然科学基金(No. LDQ23F020001), 国家自然科学基金(No. 62072406, No. 62406286), 浙江省重点研发计划(No. 2022C01018), 国家重点研发计划(No. 2018AAA0100801)资助。
收稿日期: 2023-09-25; 修改日期: 2023-11-24; 定稿日期: 2025-03-04

1 引言

图结构通常可以对实体之间的复杂关系进行建模, 例如在医疗保健分析中, 蛋白质-蛋白质相互作用可以被建模为化学网络; 在社交场景下, 社交网络可以被建模为图, 其中节点是用户, 节点之间的边可以是用户之间的社交关系, 包含了用户的敏感信息。图神经网络(Graph Neural Network, GNN)[1-5]被作为分析图数据的重要技术之一, 能够对边和节点数据进行信息提取与特征表征。

GNN 目前已经在许多领域得到了广泛应用, 包括社交分析[6-7]、生物信息[8]、金融[9-10]等领域。GNN 的成功很大程度上取决于消息传递机制, 其中节点表征包含了节点特征、邻居信息和局部图结构, 这有助于各种图挖掘任务的进行如节点分类[11-13]、链路预测[14-16]、图分类[17-18]、社区检测[19-20]。越来越多的研究人员提出不同的 GNN 架构[3,11,21-26]如图卷积网络(Graph Convolutional Networks, GCN)[11]、图注意力网络(Graph Attention Networks, GAT)[3]等来提升任务的性能。

提升 GNN 的性能固然重要, 但对其安全的担忧也不容小视, GNN 模型很容易遭受到窃取隐私数据或者影响模型行为的攻击。例如, 在社交网络场景下, 攻击者可以根据节点嵌入推断出用户在社交网络中的属性信息以及与之相关联的朋友信息[27-28], 甚至他们还可以通过向网络中注入恶意节点, 轻松欺骗 GNN 模型使其错误预测[29]。在其他场景下也出现了这种情况, 如图 1 所示是现实场景中的某银行用户隐私泄露的例子, 该银行通过构建一个 GNN 模型来评估用户信誉度, 攻击者通过发动属性推断攻击推断出用户属性信息, 从而导致用户的隐私泄露, 给用户带来极大的困扰和麻烦。更糟糕的是, 一旦 GNN 在现实世界中广泛应用, 安全风险也在陡然上升, 尤其是在金融、医疗保健等一些高风险的场景下。另外, 本文概括了不同攻击行为的示意图, 如图 2 所示。

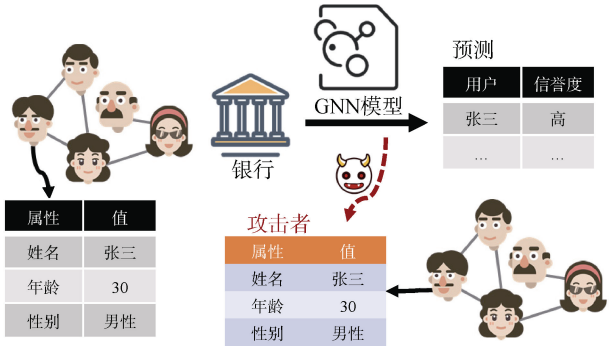


图 1 某银行用户隐私泄露的例子

Figure 1 An example of a bank user's privacy disclosure

有关机器学习模型隐私方面的相关研究[30-37]已经有了较全面的总结。但是他们所研究的隐私问题是针对文本、图像等数据, 并未讨论到图方面的隐私窃取与保护方法。Dai 等人[38]总结了可信 GNN 的相关工作, 包括隐私性、鲁棒性、公平性、可解释性等方面, 对于 GNN 是否可以信赖作了全面的调查。同样 Wu 等[39]也对可信图学习进行了一系列总结, 分别从可靠性、可解释性和隐私保护三个维度全面回顾了图学习领域。值得注意的是, 他们的研究主要关注可信 GNN, 专注于 GNN 隐私安全部分的篇幅较少, 且关于隐私安全的研究主要集中于分析现有的隐私攻击和防御方法原理并对其分类, 并未全面展开罗列不同工作的贡献尤其在隐私保护方法的阐述上。目前, 随着时间发展, 越来越多学者关注于这一方向。

因此, 为了深入探讨面向 GNN 的隐私安全问题, 本文收集了相关文献并进行了大量的研究。具体来说, 综述了大约 140 篇相关论文, 其中包括中英文学术数据库、知名会议和期刊上发表的论文, 以及发表

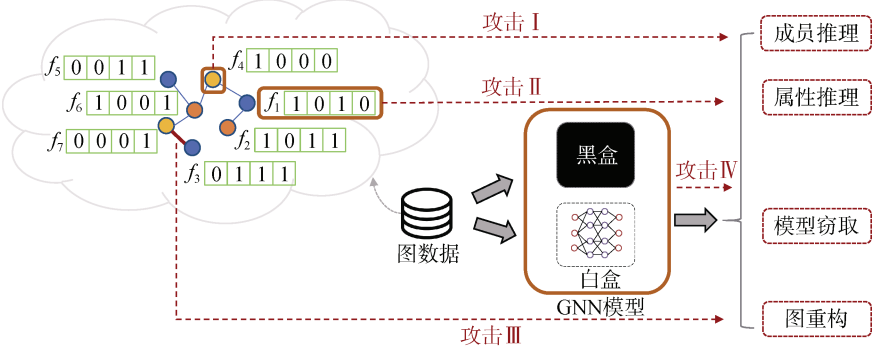


图 2 面向 GNN 的隐私攻击威胁场景的示意图

Figure 2 Schematic diagram of privacy attack threat scenario for graph neural network

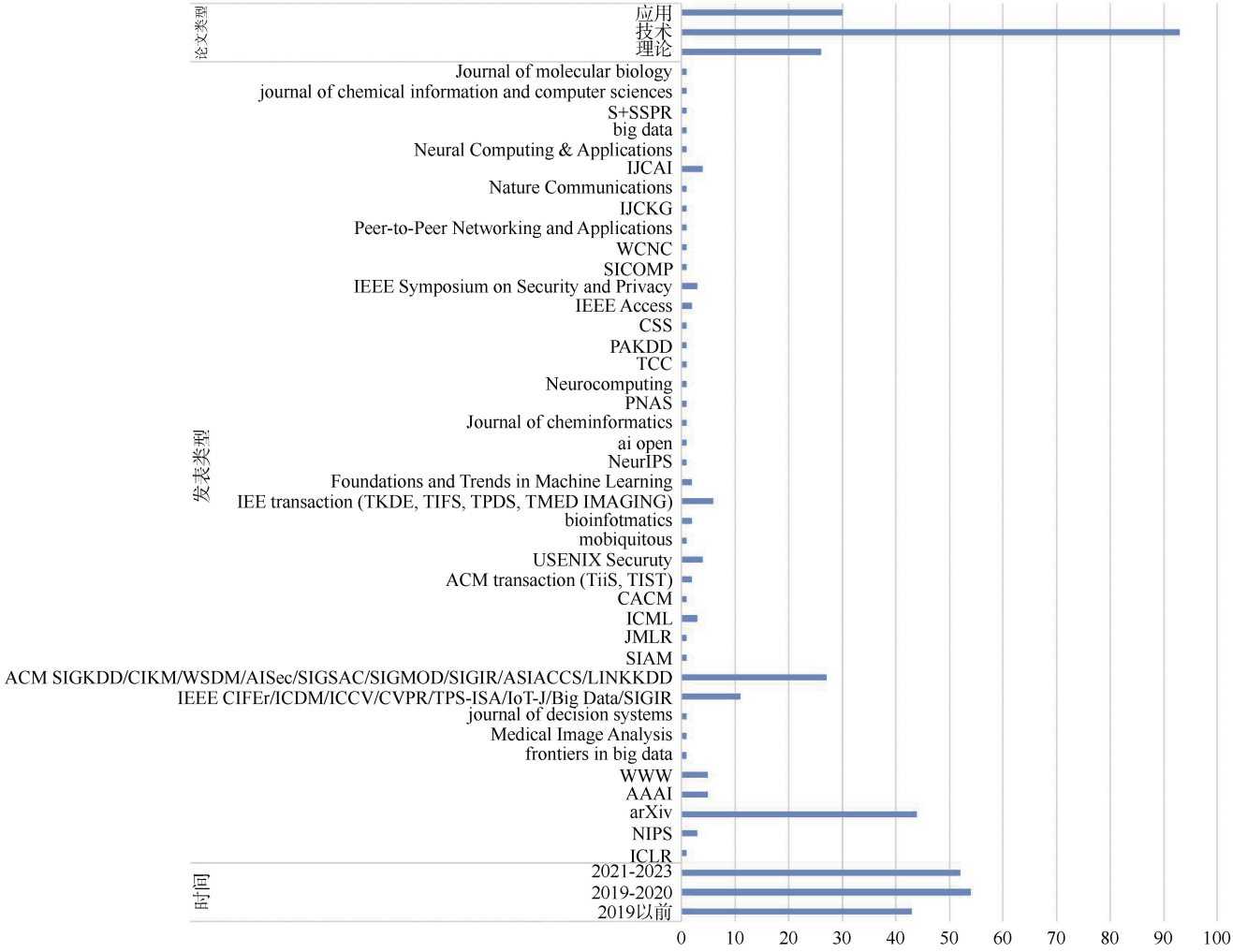


图 3　论文的详细统计情况

Figure 3　Detailed statistics of the papers collected

在专题讨论会上, 但被引次数较高的文章, 还包括在一些公共平台上近期发表的热点论文。图 3 展示了对收集到的论文进行的发表会议/期刊、论文类型、发表时间等方面的详细统计, 另外, 对于论文类型中的技术类型的论文(即第三第四章的论文)分为攻击防御两类, 图 4-6 为攻击防御方法和过去 5 年 GNN 隐私安全有关论文的详细统计情况。从图 6 中可以观察到 2021 年开始相关研究文献数量陡增, 相比去年增长了 91%, 之后一直保持在高数量, 这也说明了 GNN 的隐私安全问题正在引起越来越多人的重视。本文针对现有 GNN 的隐私安全类文献进行了具体综述, 分别详细介绍不同的隐私攻击、隐私保护方法与相应的评价指标和数据集, 还重点介绍了不同攻击和保护方法的现实应用场景, 最后总结 GNN 隐私安全未来可能的研究方向。

本文的结构章节安排如下: 第 2 节分析 GNN 隐私攻防理论, 第 3 节归纳面向 GNN 的隐私攻击方法, 第 4 节归纳面向 GNN 的隐私保护方法, 第 5 节总结 GNN 隐私安全的数据集及不同攻击保护方法的评估指标, 第 6 节总结了 GNN 隐私安全的实际应用场景, 第 7 节将 GNN 隐私安全与图像及文本等深度模型的隐私安全作了详细比较, 最后第 8 节总结并列举未来可能的研究方向, 论文章节内容与关系图如图 7 所示。

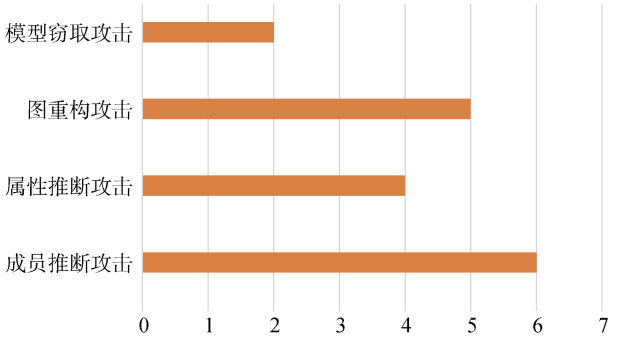


图 4　攻击方法的详细统计情况

Figure 4　Detailed statistics of the paper on attack methods

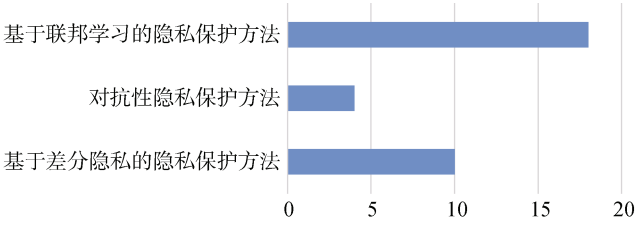


图 5 防御方法的详细统计情况

Figure 5 Detailed statistics of the paper on defense methods

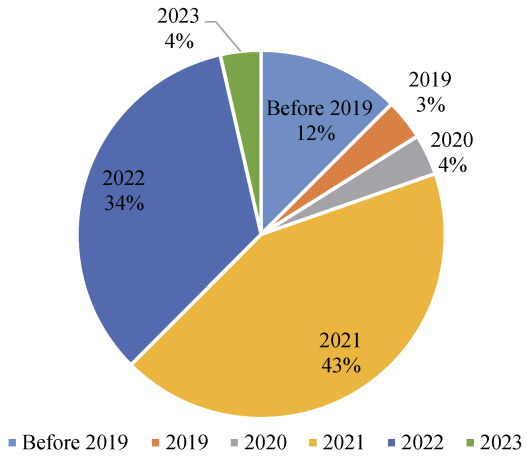


图 6 近五年深度学习隐私安全相关研究体量分析

Figure 6 Number of relevant studies in the past five years

2 图神经网络隐私攻防理论分析

本节首先介绍了一般图的定义及GNN领域常用的符号类型和其工作原理，然后分析了其主要应用，最后分别给出GNN隐私攻击和隐私保护按照不同分类方法的理论定义。

2.1 图神经网络

2.1.1 图基本概念及相关符号表示

通常图定义为 $G=\{V,E\}$，其中 $V=\{v_1,\cdots,v_N\}$ 表示 N 个节点的集合，$e_{i,j}=<v_i,v_j>\in E$ 表示节点 v_i 和节点 v_j 之间存在连边。进一步地，如果连边有特定的方向，那么就称该图为有向图，与之相反，无向图中的连边都是双向的；如果连边具有不同权重，那么就称该图为有权图，与之相反，无权图中的连边没有权重属性；图 G 可以是特征图，也可以是普通图，也就是说，除了以节点与连边集合构成的拓扑结构图，节点和连边也可以拥有自己的特征图，分别表示为 $X_{node}\in\mathbb{R}^{N\times D_{node}}$ 与 $X_{edge}\in\mathbb{R}^{N\times D_{edge}}$；如果图是由多种类型节点和连边构成的图，那么该图则为异构图，与之相反，同构图仅包含一个类型的节点与连边。$A\in\mathbb{R}^{N\times N}$ 是图 G 的邻接矩阵，在GNN中一般用于表示图中节点的关系，若节点 v_i 和节点 v_j 之间有边直接相连，那么 $A_{i,j}=1$，否则 $A_{i,j}=0$。为了更好地描述GNN及其隐私安全问题，表1更全面地列举了常用的具体符号及定义。

2.1.2 图神经网络基本理论

Scarselli 等人[40]首次提出GNN模型，用于处理图中表示的数据。由于原始的GNN在表示能力和训练效率方面存在局限，之后又有研究者提出不同的GNN变体[41-43]，以更好地学习表示和提高训练效率。Zhou 等人[44]总结了几种通用框架包括消息传递神经网络(Message Passing Neural Network, MPNN)[45]、非局部神经网络(Non-Local Neural Network, NLNN)[46]、图网络 (Graph Network, GN)[47]在内，旨在将不同模型集成到一个框架中。

一般来说，GNN 采用消息传递机制来学习节点表示，该节点表示包含节点特征信息和图拓扑信息。具体而言，GNN 将通过聚合来自其邻居节点的信息来更新节点表示。因此，对于更新后的第 k 层的节点表示的一般形式为:

$$h_v^{(k)}=COMBINE^{(k)}(h_v^{(k-1)},AGGREGATE^{(k-1)}(\{h_u^{(k-1)}:u\in N(v)\})) \tag{1}$$

其中 $h_v^{(k)}$ 为 GNN 第 k 层的节点表示，$N(v)$ 为节点 v 的邻居节点集合。

2.1.3 基于图神经网络的主要应用

GNN 的应用可以促进各种任务的应用与发展，如节点分类任务、链路预测任务、社区检测任务和图分类任务。

节点分类。许多现实世界的问题都可以被视为节点分类任务，例如社交媒体中的用户属性预测[48-49]、交易网络中的欺诈检测[50-51]、蛋白质-蛋白质相互作用网络中的蛋白质功能预测[52]等。在节点分类任务中，给定图 $G=\{V,E\}$，GNN 模型旨在推断节点标签，为半监督任务，即给定了部分标记节点 $V_L\subset V$ 及对应类标 $y_L=\{y_1,\cdots,y_{|V_L|}\}$ 和剩余未标记节点 $V_U=V-V_L$ 来训练一个节点分类模型。

链路预测。链路预测任务一般广泛应用于社交媒体中的朋友推荐[53]、交易网络中可能发生的交易预测[54]等。在链路预测任务中，给定图 $G=\{V,E\}$，GNN 模型旨在预测节点之间是否有连边存在，与节点分类任务类似也为半监督任务，即给定了部分标记连边 $E_L\subset E$ 及对应类标 $I=\{0,1\}$ 和剩余未标记连边 $E_U=E-E_L$ 来训练一个链路预测模型。

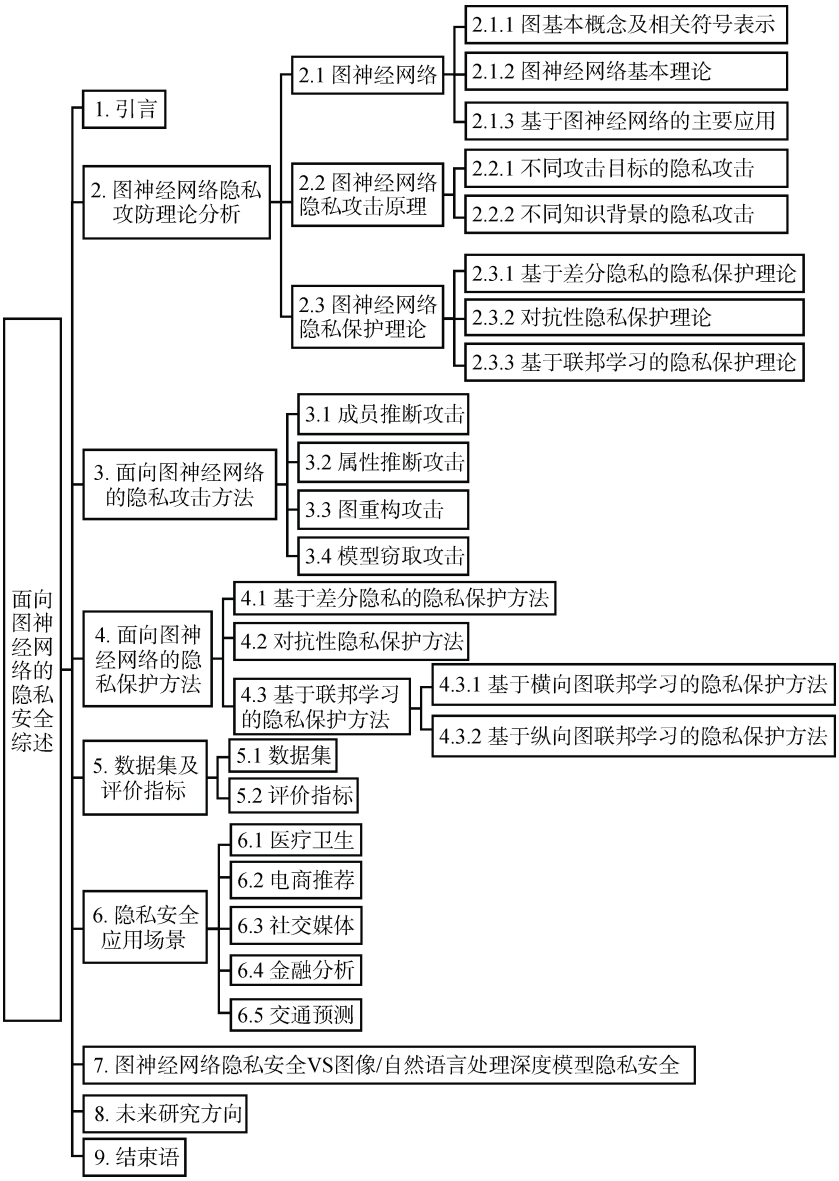


图 7　论文章节内容与关系图

Figure 7　Thesis Chapter Content and Relationship Diagram

表 1　常用的具体符号及定义

Table 1　Common specific symbols and definitions

符号	定义	符号	定义	符号	定义
G	原始图	v_i	节点	f_T	目标模型
E	连边集合	$e_{i,j}$	两节点间连边	f_s	影子模型
V	节点集合	$N(v)$	邻居节点集合	$l(\cdot)$	损失函数
$\mathcal{G}$	原始图集合	h_v	节点表示	D	数据集
A	邻接矩阵	y	节点类标集合	$\mathcal{C}$	属性集合
X	特征矩阵	I	连边类标集合	ε	隐私预算
$\mathcal{M}$	随机函数	k	客户端数目	p_k	客户端权重
C	社区集合	u	邻居节点	δ	松弛项

图分类。许多现实世界的问题如药物的性质预测[55]等都可以被视为图分类任务。在图分类任务中，给定图集合 $\mathcal{G}=\{G_1,\cdots,G_M\}$，GNN 模型旨在预测图的标签，与节点分类任务类似也为半监督任务，即

给定了部分标记图及对应类标和未标记图来训练一个图分类模型。

社区检测。社区检测任务同样可以用于各种领域, 如社交网络分析[56]和大脑功能区识别[57]等。社区是网络的子图, 它们之间的连接比网络中的其他节点更紧密。形式上, 网络中的社区集合可以用 $\mathrm{C}=\{C_1,\cdots,C_K\}$ 表示, 这些社区可能是脱节的, 也可能是重叠的。在社区检测任务中, 给定图 $G=\{V,E\}$, GNN 模型旨在推断或识别每个节点的社区, 通常是无监督任务[58-60]。

2.2 图神经网络隐私攻击原理

2.2.1 不同攻击目标的隐私攻击威胁模型

面向GNN的隐私攻击是为了获取用户本不希望被分享的信息。这些目标信息往往是训练 GNN 所用的训练数据的信息, 如成员关系, 节点的敏感属性以及节点之间的连边关系。此外也有一些攻击者的攻击目标是获取 GNN 的模型参数。因此, 基于攻击者的目的, 可以将隐私攻击分为四类: 成员推断攻击, 属性推断攻击, 图重构攻击, 模型窃取攻击。

成员推断攻击。成员推断攻击的目的是推断目标样本是否参与目标 GNN 模型的训练。成员推断攻击能够成功的主要原因是训练数据在目标模型上产生了过拟合, 导致训练样本和测试样本的输出向量产生了不同的分布。因此攻击者可以通过模型的输出向量来判断一个目标样本是否在目标模型中。为了实现成员推断攻击, 攻击者往往会训练一个影子模型来获得成员推断攻击所需要的一部分标签, 并通过获得标签来训练一个攻击模型。攻击者通过使用训练数据集的影子数据集 D_{shadow}^{train} 来训练一个影子模型 f_s 以模仿目标模型的功能, 其过程可以被表示为:

$$\min_{\theta}\sum_{G_i\in D_{shadow}^{train}} l(f_T(G_i),f_S(G_i)) \tag{2}$$

其中 G_i 属于影子训练数据集 D_{shadow}^{train}, $f_T(G_i)$ 和 $f_S(G_i)$ 代表目标模型和影子模型的输出向量。$l(\cdot)$ 表示损失函数如交叉熵损失函数。由于 D_{shadow}^{train} 被用于训练影子模型 f_s, 因此 D_{shadow}^{train} 的输出结果可以作为正样本来帮助攻击模型判断 $f_S(G_i)$ 是否过拟合, 另外使用非训练数据集的影子数据集 D_{shadow}^{out} 在影子模型 f_s 上的输出向量作为攻击模型的负样本。最终, 攻击者就可以将得到的正负样本用于攻击模型的训练, 使其能够推断出待测样本是否在目标样本的训练集中。

属性推断攻击。属性推断攻击的目的是推断训练数据集的属性。由于 GNN 的信息传递机制, 目标模型学习到的节点嵌入捕获了节点的属性。因此攻击者可以利用影子数据集的属性以及影子数据集在目标模型的输出来训练一个攻击模型。该模型的目标可以被表示为:

$$\min_{\theta}\sum_{G_i\in D_s} l(f_A(f_T(G_i)),p_i) \tag{3}$$

其中 G_i 属于影子数据集 D_s, $f_T(\bullet)$ 和 $f_A(\bullet)$ 代表目标模型和影子模型, p_i 为目标图 G_i 的属性, $l(\cdot)$ 表示损失函数如 MSE 或者交叉熵损失函数, 使攻击模型的输出结果与影子数据集的属性尽可能相似。

图重构攻击。由于 GNN 普遍存在的消息传递机制, 使得图嵌入捕获了关于图嵌入丰富的结构信息, 例如相邻的节点的嵌入具有较强的相关性。因此攻击者能够使用影子数据集的节点嵌入和影子数据集的连边关系来训练一个攻击模型。攻击者首先将具有相同连边分布的影子数据集输入到目标模型, 获得目标模型的预测结果或者嵌入。其次, 攻击者通过使用目标模型的嵌入作为攻击模型的输入, 影子数据集的邻接矩阵作为攻击模型的类标对攻击模型进行训练, 其过程可以被表示为

$$\min_{\theta}\sum_{G_i\in D_{shadow}^{out}} l(f_A(f_T(G_i)),A_i) \tag{4}$$

其中 G_i 属于影子数据集 D_s, $f_T(\bullet)$ 和 $f_A(\bullet)$ 代表目标模型和影子模型, A_i 为目标图 G_i 的邻接矩阵, $l(\cdot)$ 表示损失函数, 如 MSE 或者交叉熵损失函数, 使攻击模型的输出结果与影子数据集的属性尽可能相似。

模型窃取攻击. 模型窃取攻击旨在学习与目标模型行为相似的代理模型。训练代理模型的过程往往包括在成员关系推理攻击中。攻击者首先查询目标模型以获得对影子数据集的预测。然后它利用阴影数据集和相应的预测来训练模型提取攻击的代理模型, 其过程可以被表示为

$$\min_{\theta}\sum_{G_i\in D_{shadow}^{train}} l(f_T(G_i),f_S(G_i)) \tag{5}$$

其中 G_i 属于影子训练数据集 D_{shadow}^{train}, $f_T(G_i)$ 和 $f_S(G_i)$ 代表目标模型和影子模型的输出向量。$l(\cdot)$ 表示损失函数如交叉熵损失函数, 使影子模型的输出结果与目标模型的输出结果相似。

2.2.2 不同知识背景的隐私攻击的威胁模型

为了进行隐私窃取攻击, 攻击者通常会拥有目标 GNN 模型或者数据集的辅助性知识。本小节根据目标GNN的模型参数是否可用分为两类即白盒攻击和黑盒攻击。

白盒攻击。在白盒攻击中, 攻击者可以获得模型的参数以及训练之间的梯度信息。除此之外, 攻击

者可能需要一些其他的背景知识, 如在图重构攻击中, 攻击者可能需要目标样本的节点特征或者一个影子数据集(与目标样本有着相同分布的数据集)。

黑盒攻击。和白盒攻击相比, 黑盒攻击中的目标 GNN 模型参数是不可知的。这种情况下, 攻击者通常被允许查询目标 GNN 模型, 以获取目标样本在输入 GNN 后的预测向量或者嵌入。黑盒攻击的一个具体例子就是攻击 API 服务, 该服务在接受用户的查询时发送 GNN 的输出结果。

2.3 图神经网络隐私保护原理

2.3.1 基于差分隐私的隐私保护理论

差分隐私(Differential Privacy, DP)是由 Dwork 等人[61]提出的一种隐私保护机制, 旨在提高数据查询的准确性, 同时最小化查询统计数据库时识别其记录的机会, 其实现原理主要是通过匿名、扰乱、混淆等方式为数据添加噪声。比较常用的方法是加高斯噪声[62]和拉普拉斯噪声[61]。差分隐私可以提供训练数据的隐私保证, 因此被广泛应用于机器学习和数据挖掘等领域。

形式上, 给定 $\varepsilon>0$, $\delta\geqslant 0$, 如果一个输入的随机函数 $\mathcal{M}$ 满足 $(\varepsilon,\delta)-differential\ privacy$, 简写为 $(\varepsilon,\delta)-DP$, 那么该随机函数 $\mathcal{M}$ 可使其在单个条目中不同的任意一组相邻数据集 D 和 D' 上得到的任意输出集合 E 的概率满足:

$$P[\mathcal{M}(D)\in E]\leqslant \exp(\varepsilon)P[\mathcal{M}(D')\in E]+\delta \tag{6}$$

其中, ε 为隐私预算, 以权衡效用和隐私, δ 为松弛项, 当 δ=0 时, 则称随机函数 $\mathcal{M}$ 满足 $\varepsilon-DP$, 即无松弛项的差分隐私。

2.3.2 对抗性隐私保护理论

对抗性隐私保护是常用的隐私保护方法之一, 简单来说就是将对抗样本打上正确类标对模型进行训练, 使模型具备对应攻击方法的防御能力。

形式上, 假定 $\mathbf{H}=f_E(\mathcal{G};\theta)$ 为 GNN 模型学习的节点表示集合, 那么攻击者 f_A 的主要目标就是去推断 $\mathbf{H}$, 而 GNN 模型 f_E 就是使 f_A 不能够推断出 $\mathbf{H}$。这实际上是最小−最大博弈过程, 即在两种相反的目标函数(最小化和最大化)的指导下, 在连续的最小−最大博弈中逐步优化模型, 该目标函数可以描述为:

$$\min_{\theta_E}\max_{\theta_A}\mathcal{L}_{utility}(f_E(\mathcal{G};;\theta_E))-\beta\mathcal{L}_{Adversarial}(f_A(\mathbf{H};;\theta_A)) \tag{7}$$

其中, θ_E 和 θ_A 分别为 f_E 和攻击者模型 f_A 的参数, $\mathcal{L}_{utility}$ 是主任务的损失函数, $\mathcal{L}_{Adversarial}$ 是对抗损失, β 则是平衡以上两个损失项的超参数。

2.3.3 基于联邦学习的隐私保护理论

GNN 虽然是一种处理图数据的有效工具, 但它像大多数深度学习方法一样, 都需要集中存储用户数据来进行训练。然而, 近年来出于对数据隐私的保护, 许多国家都出台了数据保护法规限制数据的采集和直接传输。另外, 由于担心信息泄露, 用户可能不愿意将他们的数据上传到平台服务器[63-64]。受限于种种原因, 拥有数据的参与方之间无法以直接交换原始数据的方式集中训练模型, 从而造成了“数据孤岛”现象, 极大地制约了 GNN 模型的性能。为了解决以上问题, 许多研究引入联邦学习, 通过服务器聚合参与方之间共享的中间数据来联合训练模型。

由于联邦学习在GNN模型训练过程中的原始数据始终离不开数据拥有者, 因此降低了联邦学习参与方在训练阶段隐私泄露的风险, 满足数据隐私法规的要求。形式上旨在优化以下目标函数:

$$\min_{w}\sum_{k=1}^{n}p_k\mathcal{L}_k\left(\mathcal{D}_k,w\right) \tag{8}$$

其中, k 是指客户端总数目, $\mathcal{D}_k$ 是第 k 个客户端本地的数据集, $\mathcal{L}_k$ 是第 k 个客户端本地的目标函数, w 为模型参数, 第 k 个客户端的权重为 p_k, 且 p_k 满足 $p_k\geqslant 0$ 及 $\sum_k p_k=1$。

3 面向图神经网络的隐私攻击方法

上文对 GNN 的隐私攻击方法从不同角度进行了分类, 并对对应的攻击原理进行了介绍, 为了加深对 GNN 隐私攻击的理解, 本节总结 GNN 隐私攻击的相关研究, 并详细介绍现有的隐私攻击技术。根据隐私攻击的不同的攻击目标可以将这些隐私攻击方法分为: 成员推断攻击方法、属性推断攻击方法、图重构攻击方法、模型窃取攻击方法。需要指出的是即使是相同的隐私攻击方法在针对不同的 GNN 的下游任务时也存在区别, 因此对于某一隐私攻击方法, 先介绍针对节点分类任务方法, 再介绍针对图分类任务的攻击方法。本节以不同的隐私攻击目标作为分类标准, 对现有的 GNN 隐私攻击方法进行归纳。表 2 对现有 GNN 的隐私攻击进行归纳, 本节的归纳顺序是根据图下游任务以及时间顺序组织的, 其中, 表中的“背景知识”指的是攻击者推断过程中所需要的辅助知识如影子数据集, 目标样本的节点特征等, 不同类型的攻击方法所需的背景知识也不相同; “主性能和窃取攻击平衡关系”这一列的“/”表示原论文中未涉及这一部分的实验或者理论; “影响因

表 2 攻击方法分类与优缺点总结

Table 2 Classification of Attack Methods and Summary of Advantages and Disadvantages

分类	名称	输入	背景知识	原理	下游任务	影响因素	数据集	评价指标	平衡主性能和窃取效果	防御机制
成员推断(MI)	成员推断攻击量化[33]	节点类标嵌入	辅助图	通过置信度和影子模型对图网络的成员推断进行量化	节点分类	目标模型层数	Pubmed, Citeseer, Cora	Accuracy	/	×
	MIA[64]	节点类标嵌入	辅助图	通过影子模型对多种 GNN 的成员推断攻击进行鲁棒实验	节点分类	数据集类别数、阴影模型隐藏层神经元数量	Cora, CiteSeer, PubMed, Flickr, Reddit	AUC ROC scores, Precision, and Recall	/	×
	Node-Level MIA[65]	节点类标嵌入	辅助图	依靠连边关系提高成员推断攻击效果	节点分类	/	Cora, Cora-ML, Pub-Med, Citeseer, Pol-blogs	ASR	/	×
	面向对抗鲁棒模型成员推断攻击[66]	节点类标嵌入	辅助图	对抗鲁棒 GNN 与成员推断攻击的影响	节点分类	节点类型(如节点的度、密度、特征相似度)	Cora, Citeseer, Cora-full, Lastfm	Accuracy	否(攻击性能好, 原分类任务不准确)	×
	Label only MIA[67]	节点类标	辅助图	利用节点结构属性实现仅类标对 GNN 进行成员推断	节点分类	过拟合水平、采样策略、攻击模型选择	Cora_ML, Citeseer, DBLP, PubMed	Average Accuracy, Precision, Recall, AUC, F1	/	×
	AMIA[31]	图嵌入	辅助图	对图分类任务进行成员推断	图分类	过拟合水平	PROTEIN full, DD, ENZYMES, OGBG-PPA, CIFAR10, MNIST, NCI	precision, recall and F1 score	/	×
属性推断(PI)	属性推断攻击量化[33]	节点类标嵌入	节点特征	对图网络的属性推断进行量化	节点分类	攻击者辅助知识了解	LastFM,Facebook	F1	/	×
	Inference Attack[28]	图嵌入	图特征	对图级任务进行属性推断	图分类	攻击者辅助数据集了解及分布所属、属性类别数量	DD, ENZYMES, AIDS, NCI1, and OVCAR-8H	Accuracy	/	×
	Group PIA[70]	节点类标嵌入	群特征	多个影子模型进行推断 GNN 的群属性	节点分类	攻击分类器、聚合方法、阴影模型隐藏层神经元数量、属性组大小比例、训练测试数据之间的节点重叠与否	Pubmed, Pokec, Face-book	Accuracy	/	√
	Black-box AIA[98]	图嵌入	节点特征	使用特征传播算法来推断敏感属性	节点分类	训练数据大小	Credit Cora Pubmed Facebook LastFM Texas	hamming distance, MSE	/	×
图重构(GR)	图重构量化[33]	节点类标嵌入	辅助图	量化图网络的图重构	节点分类	攻击者辅助知识了解	Pubmed, Citeseer, Cora	precision, ROC-AUC Score	/	×

续表

分类	名称	输入	背景知识	原理	下游任务	影响因素	数据集	评价指标	平衡主性能和窃取效果	防御机制
图重构(GR)	LSA[27]	节点类标嵌入	辅助图、节点特征、影子数据	从八个场景根据节点嵌入相似性设计攻击方法	节点分类	不同架构的阴影目标模型	Citeseer, Cora, Pubmed, AIDS, COX2, DHFR, ENZYMES, and PROTEINS_full	AUC	/	×
	GraphMI[72]	节点类标嵌入	节点特征	通过模型反演实现图重构	节点分类	节点标签比例和边	Cora, Citeseer, Polblogs, USA, Brazil, AIDS, ENZYMES	AUC, AP	/	×
	LINKTELLER[73]	节点类标嵌入	节点特征	通过节点之间的影响力实现图重构	节点分类	节点度	Pubmed, Citeseer, Cora, PPI, Flickr, Twitch-ES, Twitch-RU, Twitch-DE, Twitch-FR, Twitch-ENGB, Twitch-PTBR	precision, recall, F1, AUC	否(效用高的模型更容易受到攻击)	×
模型窃取(ME)	面向图分类的图重构攻击[28]	图嵌入	辅助图	对图分类任务通过编码解码器实现图重构	图分类	迭代次数	AIDS, ENZYMES, and NCI1	Degree Distribution, Local Clustering Coefficient, Betweenness Centrality, Closeness Centrality	/	×
	Model Stealing Attacks[74]	节点类标嵌入	辅助图	提出通用的图网络代理模型实现模型窃取	节点分类	查询预算	DBLP, Pubmed, Citeseer, Coauthor, ACM, Amazon	Accuracy, fidelity	/	×
	Model Extraction Attacks[135]	节点类标嵌入	辅助图、节点特征、影子数据	针对7种攻击场景设计7种攻击类型	节点分类	攻击节点数量、合成节点的使用与否	Pubmed, Citeseer, Cora	Accuracy, fidelity	/	×

素”指的是原论文中对影响实验结果的因素进行了实验或者说明的部分。

3.1 成员推断攻击

成员推断攻击是机器学习中常见的隐私攻击, 其目标是为了推断目标样本如节点是否参与模型训练。为了对 GNN 上的成员推断攻击进行研究, Duddu 等人[33]首次对 GNN 上的成员推断攻击进行了量化, 并且将成员推断攻击的场景分成黑盒场景以及白盒场景, 如图 8 所示。在黑盒设置中, 假设对手在给定节点时只能访问目标模型的输出概率。因此, 在这种情况考虑利用训练(即成员和非成员数据)的预测置信度之间的统计差异。并基于此分别考虑了两种攻击方法(即影子攻击和信任攻击), 实验还验证了在黑盒环境下, 置信度攻击比阴影攻击的性能要好得多。在白盒设置中, 攻击者可以访问目标模型的中间输出, 并提出了一种将中间嵌入映射到单个隶属度值的无监督方法。但是该工作缺乏清晰的攻击方法及实验细节包括模型设置及选择等。

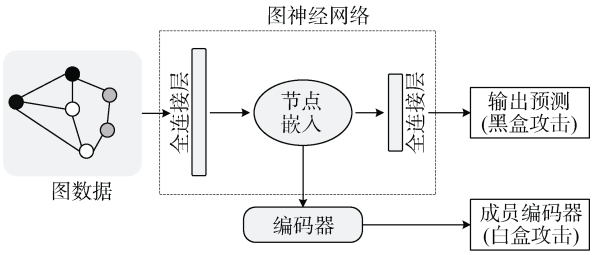


图 8 黑盒与白盒场景下的成员推断攻击原理

Figure 8 Membership Inference Attack Principles in Black-box and White-box Scenarios

Olatunji 等人[64]借助阴影模型来实施成员推理攻击, 其方法框图如图 9 所示, 该工作中的阴影模型与 Duddu 等人[33]提出的在黑盒设置中的阴影模型具有相似的思想, 通过监督的方式使用由训练好的影子模型生成的训练数据来构建攻击模型, 此外, 他们还尝试另一种方法来获得阴影模型。他们使用原

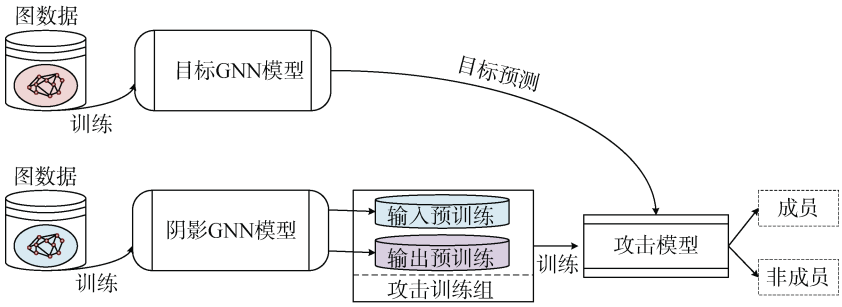


图 9　GNN 成员推断攻击框图

Figure 9　Block Diagram of GNN Member Inference Attack

始节点的真实标签来训练阴影模型，而不是查询目标模型的置信度。他们发现两种方式的攻击成功率并没有显著差异，因此能够说明阴影模型可以不需要与目标模型具有完全相同的架构。

Liu 等人[66]则分析了对抗攻击与成员推断攻击之间的关系，即在对抗训练得到的鲁棒模型是否能提高成员推断攻击的攻击效果，其方法框图如图 10 所示。他们经过大量的实验，发现通过对抗性训练获得的鲁棒模型可以显著提高成员推理攻击的攻击成功率，且造成这一现象的主要原因是模型的损失函数。

考虑到子图的邻接关系会给成员推断攻击带来不同的攻击效果，He 等人[65]系统地对攻击者拥有的背景知识进行系统的分类，同时他们发现不同的子图邻接关系也会对成员推断攻击的效果产生影响，如图 11 所示。实验表明，子图密度越高的目标节点越容易实现成员推理。这是由于密集子图驱动节点更多地参与 GNN 训练的聚合过程，这放大了节点在目标 GNN 模型中的影响力。

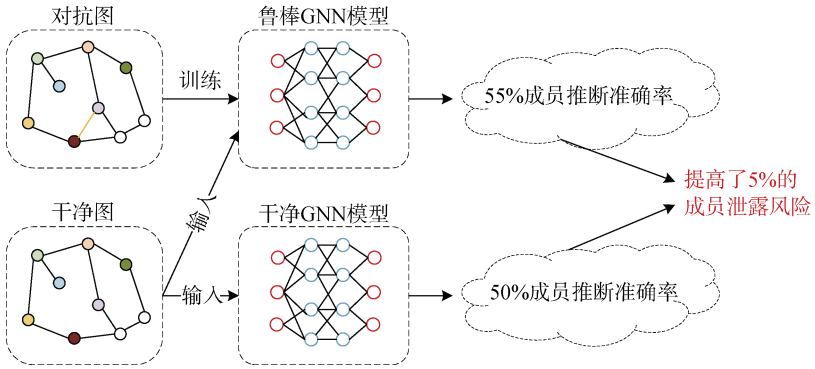


图 10　对抗鲁棒 GNN 成员推断框图

Figure 10　Block diagram of member inference attack against robust GNN

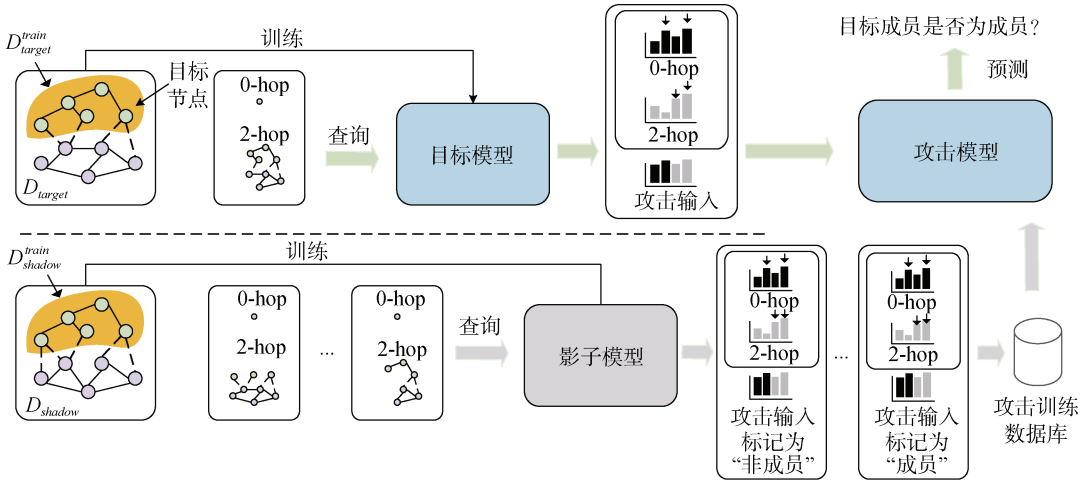


图 11　基于邻接子图的 GNN 成员推断框图

Figure 11　Block Diagram of Membership Inference Attack of GNN Based on Adjacency Subgraph

此前的方法往往都是利用GNN的类标嵌入作为成员推断攻击的输入, 但实际场景中攻击者往往不能获得模型内部的参数。受到子图邻接关系对攻击性能会产生影响的启发, Conti 等人[67]提出仅使用目标模型输出的类标实现成员推断攻击, 其将目标节点的邻居节点的预测结果作为目标节点的辅助特征, 并以此作为攻击模型的输入来提高攻击模型的性能。除此之外, 节点的度也作为节点的辅助特征来一起训练攻击模型。其方法框图如图 12 所示。实验结果表明, 使用这些特征能够有效的提高攻击模型的性能, 使成员推断攻击在真实场景下也存在应用的可能。

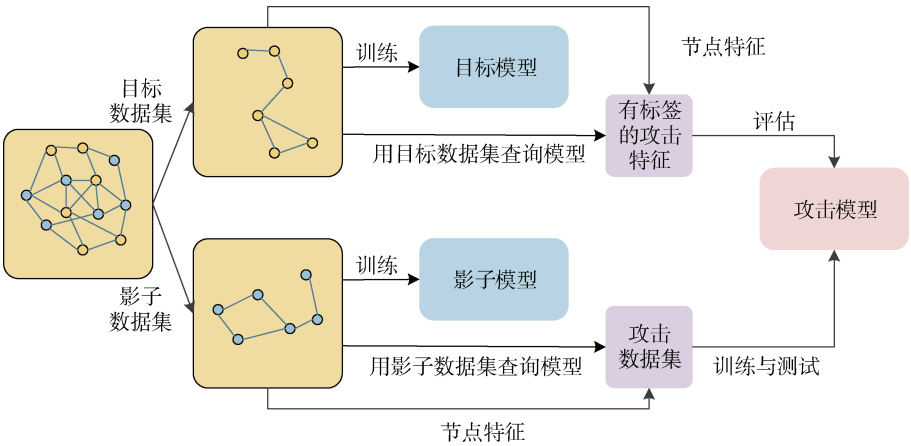


图 12　仅类标场景下 GNN 成员推断框图

Figure 12　The Block Diagram of Member Inference Attack in the Class-marking Only Scenario

针对成员推断攻击在图分类任务上的研究, Wu 等人[31]首次在图分类任务中实现了成员推断攻击, 并根据攻击者拥有的知识提出了两种不同的攻击方法, 分别是基于影子模型和基于置信度的攻击方法, 并通过大量实验表明GNN模型比非图结构的模型更容易受到成员推断攻击。此外, 与前面几篇关注节点分类任务的工作不同, 图分类任务中的成员推断攻击更多地与 GNN 的过拟合水平相关, 而不是与训练图的统计特性相关。

总而言之, 现有的针对 GNN 的成员推断攻击在节点分类任务和图分类任务中都被归类于两种方法, 即基于影子模型的方法和基于置信度的方法。基于影子模型的方法往往需要使用一部分辅助数据集来训练一个影子模型, 并通过训练一个攻击模型来判断目标样本是否用于模型训练, 其中辅助数据集往往需要包括训练数据集, 这使得攻击者需要较强的背景知识。基于置信度分数的方法则并不需要辅助数据集来训练一个影子模型, 仅仅只需要通过模型的类标置信度分数来判断目标数据集是否用于目标模型的训练, 这种方法虽然不需要强大的背景知识, 但是其判断的效果较差。因此在背景知识匮乏的情况下[68-69], 现有的方法难以实现准确的成员推断。此外, 现有的方法集中于研究成员推断攻击对不同 GNN 模型以及不同的背景知识的攻击性能。因此提出更多样的成员推断攻击来提高攻击性能是很有必要的。

3.2　属性推断攻击

属性推断攻击的研究目前仍然处于初级阶段, 仅有少量的方法被提出。为了对 GNN 上的属性推断攻击进行研究, Duddu 等人[33]首次对 GNN 的属性推断攻击进行了量化, 并对节点分类任务中属性推断攻击的可行性做了阐释。由于节点嵌入会捕获节点的特征和连边关系, 从而导致节点嵌入与节点的属性具有强相关性, 因此利用节点嵌入推断节点的属性是可行的。他们通过使用三种攻击模型, 即神经网络、随机森林、支持向量机对目标节点的属性进行推断, 验证了属性推断攻击在节点分类任务上的可行性。Olatunji 等人[148]进一步探索了黑盒属性推断攻击, 通过改变对抗性知识和假设来发起攻击来观察是否会造成更大的隐私风险。

图分类任务与节点分类任务的对象不同, 节点分类任务的对象是节点, 其往往具有明显的属性, 而图分类任务的对象是图, 一个图的属性是隐性的。Zhang 等人[28]为了验证成员推断攻击在图分类任务上的效果, 首次在图分类任务上实现了对图的属性的推断, 如图的密度, 节点数量等, 其目标如图 13 所示。他们通过将具有相同分布的影子图数据输入到目标图嵌入模型中, 将模型输出的图嵌入作为攻击模型的输入样本, 并且将影子数据集自身的属性作为标签, 对攻击模型进行训练, 使攻击模型能够

根据图嵌入实现对图属性的推断。

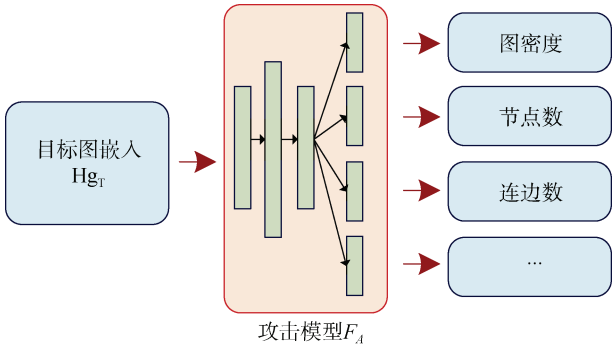


图 13 图分类任务下 GNN 属性推断目标

Figure 13 Attribute Inference Attack of GNN Target under Graph Classification Task

Wang 等人[70]在图属性的基础上对群属性实现了隐私推断，该攻击推断图中节点以及链路的分布如图中男性节点与女性节点之间的比例。他们首先需要同时训练多个影子模型，其次将每个影子模型的输出进行聚合作为攻击模型的样本，并且将影子数据集的特征作为攻击模型的类标，用于攻击模型的训练，其方法框图如图 14 所示。

综上所述，现有的属性推断攻击根据下游任务的不同可以分为两类，即针对节点分类任务的属性推断攻击和针对图分类任务的属性推断攻击。他们的重点都是将获取的属性作为攻击模型训练的类标，将目标模型的类标嵌入作为攻击模型的输入来训练一个目标模型。这两类方法的主要区别是属性是否能直接被观察到，节点分类任务的属性如性别、年龄等[71]往往能够直接被观察到，而图分类任务的属性如图的密度往往不太容易直接被观察到，因此针对图分类任务的属性推断攻击更加困难，对背景知识的要求也更高。

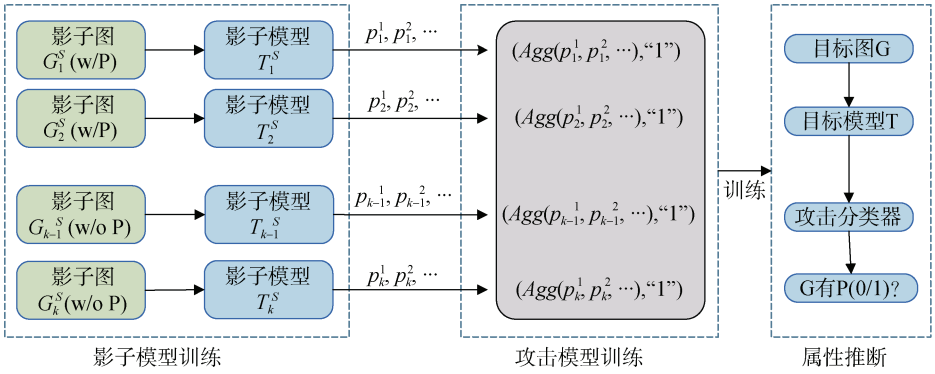


图 14 GNN 群属性推断框图

Figure 14 Block Diagram of Group Attribute Inference Attack of GNN

3.3 图重构攻击

为了更好的对图重构攻击进行解释，Duddu 等人[33]首次对 GNN 的图重构攻击进行了量化，并对图重构攻击实现的原因进行了说明，如图 15 所示。他们指出图嵌入捕获了关于图丰富的语义和结构信息，例如保持与相邻节点的接近度。因此攻击者在知道辅助子图的情况下能够依靠辅助子图训练一个以图卷积网络为架构的图自编码器，并通过解码器对图嵌入进行解码实现对图网络的重构。

为了进一步研究不同背景知识下图重构攻击带来的风险，He 等人[27]沿着三个维度系统地描述对手的背景知识，包括节点属性、目标数据集的部分子图和影子数据集，因此攻击者可以拥有 8 种不同类型的攻击场景。同时，根据具有相似节点嵌入的节点容易存在连边关系这一点，为每一种攻击场景分别设计了一种攻击方法。例如当目标数据集的部分图可用时，可以使用监督学习来训练一个二分类器作为攻击模型，该分类器从两个节点的属性中总结特征并以此实现对连边的预测。

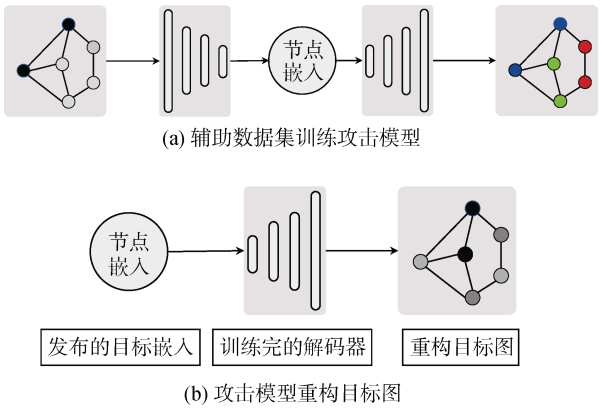


图 15 GNN 图重构攻击原理框图

Figure 15 GNN graph reconstruction attack principle block diagram

受到图像上模型反演的启发，Zhang 等人[72]将模

型反演方法应用到图重构攻击中。他们将一个全为零的邻接矩阵输入到目标模型中, 通过梯度的反向传播, 优化邻接矩阵使输出结果与正确输出结果接近, 最终实现对目标邻接矩阵的重构, 如图 16 所示。此外, 他们还研究了边缘影响力与模型反演之间的关系并得出结论: 边缘影响力越大的连边更容易被重构。

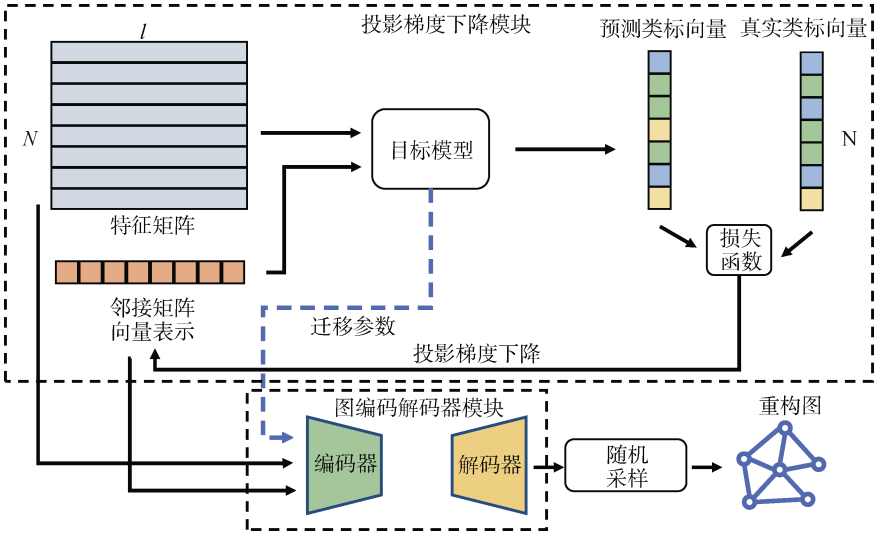


图 16 基于模型反演 GNN 图重构攻击框图

Figure 16 Block Diagram of GNN Graph Reconstruction Attack Based on Model Inversion

Wu 等人[73]考虑到一种特定场景, 即一方持有图的连边信息, 另一方持有图的特征信息。持有图特征信息的一方作为攻击者, 并试图对连边信息进行窃取。他们分析节点对预测结果的影响来实现对潜在连边的判断。具体来讲, 他们在节点特征中加入一定噪声, 并计算噪声对其他节点嵌入的影响力, 之后对影响力进行排序, 将相互之间影响力较大的节点对判断为存在连边, 反之则不存在连边。

此前对图重构攻击的研究都集中在节点分类任务中, 但图分类任务也面临着图重构攻击的风险。因此, Zhang 等人[28]提出了在图分类任务上的图重构攻击, 其攻击方法与 Duddu 等人[33]在节点分类任务上的图重构攻击方法结构相似, 都使用了辅助子图来训练图自编码器, 并通过解码器实现对图嵌入的解码, 从而实现对原始图网络的重构。

目前对 GNN 的图重构攻击主要集中于节点分类任务以及图分类任务, 其核心思想就是两个节点的特征或者嵌入越相似, 那么这两个节点就越有可能相连。而在实施方法上主要可以分为两个大类, 即基于图自编码器的方法和基于相似度的方法。基于图自编码器的方法往往需要得到部分子图信息来对自编码器进行训练, 对背景知识的要求更高, 而基于相似度的方法在仅拥有目标模型节点嵌入的情况下也能够实现对目标连边的判断, 在现实场景中更容易实现。

3.4 模型窃取攻击

目前对 GNN 的模型窃取攻击仍处于初级阶段[74-75], 目前仅有两种攻击方法被提出, 这两种攻击方法分别根据不同的背景知识发动对目标模型的窃取攻击。

Shen 等人[74]首次提出了针对 GNN 的模型窃取攻击, 以填补模型窃取攻击在图数据领域上的空缺。他们将攻击者的背景知识分为两个维度, 询问图和目标模型响应, 并根据攻击者拥有的背景知识类型提出了 6 种不同的攻击场景, 同时提出一种通用的攻击方法使得其攻击在这 6 种攻击场景下都能够实现模型窃取。具体来说, 该方法首先通过 IDGL 算法根据获得到的节点特征生成图的邻接矩阵, 其次, 将得到的邻接矩阵输入到目标模型中来获得目标模型的响应, 最终使用生成的邻接矩阵和目标模型的响应来训练代理模型, 使得代理模型具有和目标模型相似的功能。

为了使模型窃取攻击在真实场景下更具有可行性, Wu 等人[135]考虑了三种维度下的背景知识, 即节点属性、由节点组成的辅助图和影子图包括其图结构和属性信息。他们根据攻击者拥有的不同背景知识将攻击场景分为 7 类, 并通过自适应的攻击策略为 7 种攻击场景设计了不同的模型提取攻击。以攻击者可以获得邻接矩阵和部分节点的特征为例, 如图 17 所示, 攻击者能够使用邻接矩阵以及节点特征对目标模型进行询问以获得目标模型的嵌入。同时,

由于攻击者只知道攻击节点的属性, 攻击者通过合成二阶邻居节点的属性来表示未知节点的属性, 从而补全子图上所有节点的属性。在知道邻接矩阵, 节点特征和目标模型的嵌入后, 攻击者就能够通过有监督的方式来训练一个代理模型以获得目标模型的功能。

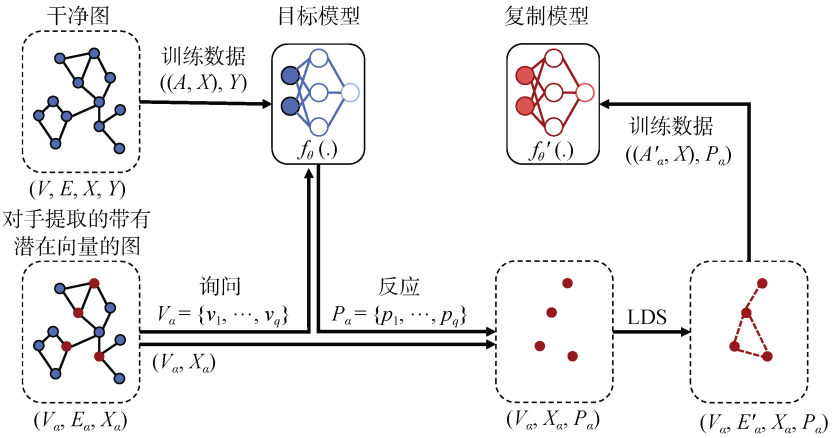


图 17 基于 LDS 的模型窃取攻击原理框图

Figure 17 Block Diagram of Model Stealing Attack based on LDS

目前已有的两种模型窃取方法分别根据不同背景知识对 GNN 发动了模型窃取攻击。他们的核心思路都是通过部分已知背景知识和目标模型的输出对一个代理模型进行有监督的训练, 从而使得代理模型获得目标模型的能力。他们方法主要的不同点在于如何根据已有的部分背景知识对辅助数据集实现补全, 其中, Shen 等人[74]集中于对网络结构的补全, 而 Wu 等人[135]则集中于对节点特征的补全。

4 面向图神经网络的隐私保护方法

第 2.3 节按照不同的隐私保护策略作为分类标准对现有的面向 GNN 隐私保护方法进行分类即分为基于差分隐私的隐私保护方法、基于联邦学习的隐私保护方法和对抗性隐私保护方法三大类, 并对对应的隐私保护原理进行了介绍, 为了总结不同隐私保护方法, 本节按照上述分类标准对现有针对 GNN 隐私保护方法进行进一步展开。具体隐私保护方法如表 3 所示。

4.1 基于差分隐私的隐私保护方法

为了保护图结构数据的隐私, 许多研究者首先集中于如何在网络嵌入中保护个人隐私[63,68]。Xu 等人[63]提出了如何在网络嵌入中保护差分隐私的第一项工作即开发了一种差分专用网络嵌入方法(DPNE), 对学习网络嵌入的目标函数施加扰动。但是该方法是基于矩阵分解的方法, 因此其存在一个潜在限制——不能扩展到大型网络。Zhang 等人[68]提出了一种也是基于矩阵分解的扰动梯度下降方法(PPGD), 它保证了通过矩阵分解学习的图嵌入的隐私性。然而, 此方法假设矩阵分解的目标函数已知, 并未考虑目标函数未知的情况。在 GNN 模型广泛应用之后, 本节具体介绍几种用于 GNN 的有效差分隐私方法。

Sajadmanesh 和 GaticaPerez[69]提出了一种基于局部差分隐私(LDP)[75]的新的隐私保护 GNN 学习框架 LPGNN。该方法可以在节点特征和标签都是私有的或者两者之一私有时使用, 并且可以独立地与任何 GNN 架构组合。他们首先提出了一种 LDP 机制, 允许服务器私有地收集节点特征, 并使用噪声特征估计 GNN 的第一层图卷积。然后, 为了进一步降低估计误差还引入了 KProp, 这是一个简单的图卷积层, 它聚合了来自高阶邻居的特征。最后, 为了学习带有扰动标签的模型, 还提出了一种称为 Drop 的学习算法, 该算法利用 KProp 进行标签去噪。图 18 展示了 LPGNN 的总体框架, 图中的 MB 指 LDP 机制, RR 指随即响应, 用户分别对其私有特性和标签运行图上步骤, 并将输出发送到服务器, 然后开始训练。绿色实心箭头和红色虚线箭头分别表示训练和验证路径。然而, 该方法虽然已经涉及了对于节点特征和标签的隐私保护, 但是图拓扑方面并没有提及受到隐私保护。

Lin 等人[76]对其进行了扩展提出了 Solitude, 可在 LDP 设置中保护链路信息, 然而, 他们的链路 LDP 机制并不是原则性的。Zhu 等人[77]在链路 LDP 概念与其相同的情况下改进并提出了一种在 GNN 中实现链路 LDP 的新框架 BLINK, 它允许服务器在较低的隐私预算下估计较少的链路, 从而减少误报链路估计, 相比前两个工作[69,76]来说, 它达到的分类准

表 3　隐私保护方法分类与优缺点总结

Table 3　Classification of Defense Methods and Summary of Advantages and Disadvantages

保护策略	二级分类	保护方法	任务	原理	目标模型	数据集	评价指标	保护目标
基于差分隐私的隐私保护	/	LPGNN[69]	节点分类	利用节点特征的多跳聚合作为去噪机制	GNN	Cora, Pubmed, Facebook, LastFM	Accuracy	节点特征和标签
		Solitude[76]	节点分类	LPGNN 的扩展并在校准分散图中引入噪声以保护链路信息	GNN	Cora, Citeseer, LastFM	Accuracy	节点特征和拓扑结构
		BLINK[77]	链路预测	在 GNN 中实现链路 LDP 以减少误报链路估计	GNN	Cora, Citeseer, LastFM, Facebook	Accuracy	拓扑结构
		PrivGNN[79]	节点分类	使用不相交的数据集训练学生教师模型并与两种噪声机制相结合保证隐私	GNN	Amazon, ArXiv, Reddit	Accuracy	节点标签和模型优化过程
		GERAI[80]	评级预测	研究输入阶段用户特征扰动和优化阶段的损失扰动	GCN	ML-100K	F1, Hit@K, NDCG@K	节点特征和模型优化过程参数
		HeteDP[81]	节点分类/链路预测	设计隐私保护特征编码器和基于差分隐私的梯度扰动异构链路重构器以容忍数据多样性并抵御攻击	GNN	ACM, DBLP, Amazon, IMDB	ROC-AUC score	节点特征和拓扑结构
		图分割算法利用 SGD-DP[82]	节点分类	将图划分为子图同时避免对原始数据进行额外查询	GNN	Cora, Citeseer, PubMed, Reddit, Pokec	F1	节点特征和拓扑结构
		DP-SGD 对图分类的扩展[85]	图分类	用指纹数据集进行图分类并用差分隐私随机梯度下降(DP-SGD)	GNN	Synthetic, Fingerprints, Molbace, ECG	ROC AUC Score, Accuracy, Sensitivity, Specificity, F1	拓扑结构
		GAP[87]	节点分类/链路预测	利用一阶邻居之外的多跳聚合, 并在不增加额外成本的情况下保证推理隐私	GNN	Facebook, Reddit, Amazon	Accuracy, AUC	拓扑结构
		ProGAP[88]	节点分类	使用渐进训练方案来限制隐私成本	GNN	Facebook, Reddit, Amazon, Facebook-100, Wenet	Accuracy	拓扑结构
对抗性隐私保护方法	/	NetFense[89]	节点分类	通过改变邻接矩阵欺骗攻击者同时保持下游任务的效用	GNN	Cora, Citeseer, PIT	Accuracy	节点标签
		GAL[91]	节点分类/链路预测	通过任务解码器和一个模拟最坏情况的攻击者之间在嵌入和属性上的极小极大博弈来防御节点和邻域推理攻击	GNN	Citeseer, Pubmed, QM9, ML-1M, FB15K-237, WN18RR	AUC, Macro-F1, RMSE	节点标签

续表

保护策略	二级分类	保护方法	任务	原理	目标模型	数据集	评价指标	保护目标
基于差分隐私的隐私保护	/	基于互信息的图表示学习框架[92]	节点分类/链路预测	以互信息的角度提出目标函数	GNN	Cora, Citeseer, Pubmed	AUC, Accuracy	节点标签与拓扑结构
		APGE[90]	属性分类	用隐私解纠缠和隐私清除机制从学习的节点表示中删除用户的私有信息	GCN	Yale, Rochester	Accuracy, Macro-F1	节点特征
		FL-AGCNS[96]	节点分类	采用联邦进化优化策略有效搜索高质量 GCN 架构	GCN	Cora, CiteSeer, PubMed, CoraFull and Physics	FLACC	节点特征
		GCFL[97]	图分类	将各种强大的 GNN 模型集成到聚类 FL	GNN	MUTAG, BZR, COX2, DHFR, PTC_MR, AIDS, NCI1, ENZYMES, DD, PROTEINS, COLLAB, IMDB-BINARY, IMDBMULTI	Accuracy	拓扑结构和节点特征
	基于横向图联邦学习的隐私保护方法	FedGraph[99]	节点分类	使用跨客户端图卷积操作在共享之前压缩嵌入隐藏私人信息	GCN	Cora, Citeseer, PubMed, Reddit	Accuracy	拓扑结构和节点特征
		Fedsage/Fedsage+[100]	节点分类	可生成缺失的邻居节点并模拟出子图间的关联性, 以及改善子图间的分布差异	GNN	Cora, Citeseer, PubMed, MSAcademic	Accuracy	拓扑结构和节点特征
对抗性隐私保护方法		FedGL[102]	节点分类	用服务器端边缘生成器通过 FL 客户端上传节点嵌入生成全局伪图	GCN	Cora, Citeseer, ACM, Wiki	Accuracy	拓扑结构
		ASFGNN[103]	节点分类	在客户端应用了模型插值技术	GNN	Cora, Pubmed, Citeseer	Accuracy	拓扑结构和节点特征
		PPGNN[109]	节点分类	将私有数据和非私有数据分开计算同时采用安全多方计算技术	GNN	Cora, Pubmed, Citeseer	Accuracy	拓扑结构, 节点特征和标签
	基于纵向图联邦学习的隐私保护方法	FedVGCN[110]	节点分类	将计算图数据分成两部分并采用加性同态加密	GCN	Cora, Pubmed, Citeseer	Accuracy	拓扑结构
		FedSGC[111]	节点分类	在无服务器端的情况下在模型参数更新之前对敏感信息进行加性同态加密	SGC	Cora, Citeseer	Accuracy	拓扑结构
		SGNN[112]	节点分类	计算基于动态时间扭曲(Dynamic Time Warping, DTW)算法的相似矩阵来传递相同的图拓扑	GNN	Aminer, Brazil, Europe	Accuracy	拓扑结构
		GraphFL[114]	节点分类	遵循模型不可知的元学习(MAML)的训练方案在服务器上学习全局模型	GCN	Cora, Citeseer, Coauthor CS, Amazon2M	Accuracy	拓扑结构

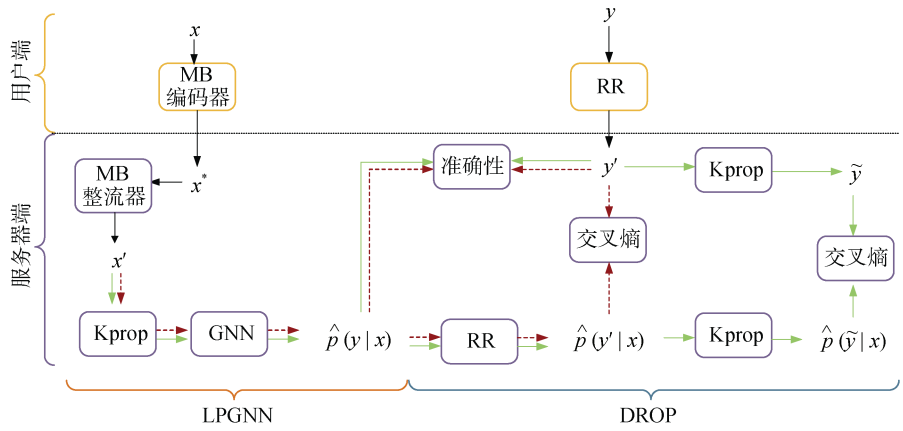


图 18　LPGNN 的总体框架

Figure 18　The Overall Framework of LPGNN

确率较高, 例如在 Cora 数据集上使用 GraphSAGE 模型的分类准确率高达 85%, 但是对于隐私效用界限等理论见解还较少, 并且它并没有扩展到节点信息的保护。

由于训练好的模型会存在隐私泄露的风险, Papernot 等人[78]提出了一种普遍使用的方法 Private Aggregation of Teacher Ensembles (PATE), 为训练数据提供强有力的隐私保证。该方法以黑盒的方式结合了使用不相交的数据集训练的多个模型, 例如来自不同用户子集的记录。因为它们直接依赖于敏感数据, 所以这些模型没有公布, 而是被用作“学生”模型的“教师”。学生学习预测通过在所有教师中进行嘈杂投票选择的输出, 并且不能直接访问单个教师或基本数据或参数。学生的隐私属性既可以直观地理解(因为没有一个老师, 也没有一个数据集决定学生的训练), 也可以从形式上理解, 即差分隐私。即使对手无论询问学生, 还是检查其内部工作, 这些属性仍然有效。不过该方法仅针对图像数据, 在图结构数据上并不适用, 因此 Olatunji 等人[79]基于类似 PATE 的假设即存在一个有标签的隐私图和一个无标签的公共图, 提出了一个称作 PrivGNN 的隐私保护框架, 用于发布满足 DP 的 GNN 模型, 如图 19 所示, PrivGNN 需要把在隐私图上训练的“教师”模型知识转移到仅在公共图上训练的“学生”模型中, 最终将其与噪声机制结合。由于隐私数据的子采样以及公共数据的嘈杂标签, 隐私得到了直观的保证。通过利用师生训练范式, PrivGNN 对 GNN 模型的攻击具有鲁棒性, 包括成员推理攻击和模型窃取攻击, 即它保证了节点标签与模型优化训练过程的隐私性。然而, 该方法对公共图数据的依赖限制了它们方法的适用性。

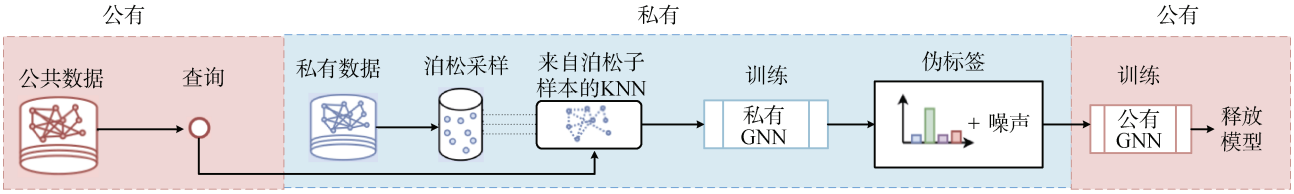


图 19　PrivGNN 的整体流程

Figure 19　The Overall Process of PrivGNN

接下来针对不同领域提出具体差分隐私保护方案, 在推荐领域, Zhang 等人[80]研究了输入阶段的用户特征扰动和优化阶段的损失扰动, 具体来说就是提出了一种基于 GCN 的推荐系统框架 GERAI, 将差分隐私中的信息扰动机制与图卷积网络的推荐能力相结合。如图 20 所示, GERAI 首先屏蔽用户的特征包括敏感信息, 然后将差分隐私纳入 GCN, 这有效地桥接了用户偏好和生成安全推荐的特征, 使得恶意攻击者无法从用户的交互历史和推荐中推断出他们的私有属性, 他们也用实验证明了对于属性推断攻击的防御有效性。

在社交领域, 随着更多辅助信息的加入, 攻击者的推理能力可能会增强, 现有的方法很难适应异构图的多样性。为了处理社交网络的异构性, Wei 等人[81]提出了一种新的无监督的基于差分隐私机制的异构 GNN 隐私保护方法——HeteDP, 该方法在图特

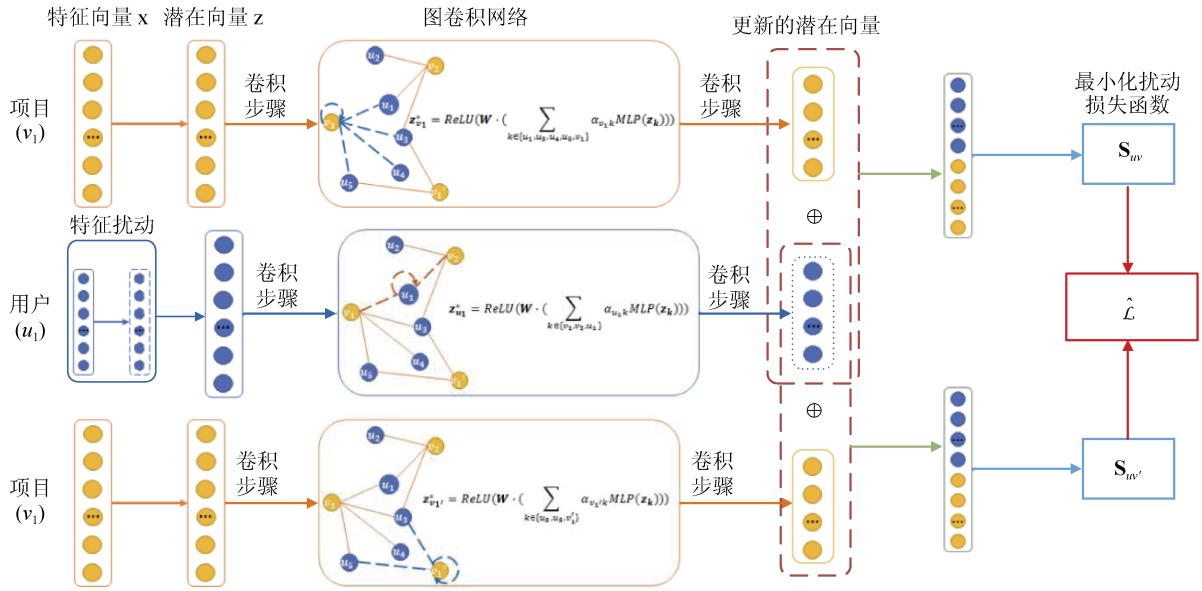


图 20 GERAI 的整体框架

Figure 20 Overall framework of GERAI

征和拓扑结构上提供了双重保证，如图 21 所示，并通过使用双层优化来设计自适应隐私预算分配，以平衡 HeteDP 的隐私和效用。他们首先定义了一种新的攻击方案来揭示异构图中的隐私泄露。具体来说，他们设计了一个两阶段的流水线框架，其中包括隐私保护特征编码器和基于差分隐私的梯度扰动异构链路重构器，以容忍数据多样性并抵御攻击。此外，为了更好地控制噪声并提高模型性能，他们还利用双层优化模式为上述两个模块分配合适的隐私预算。

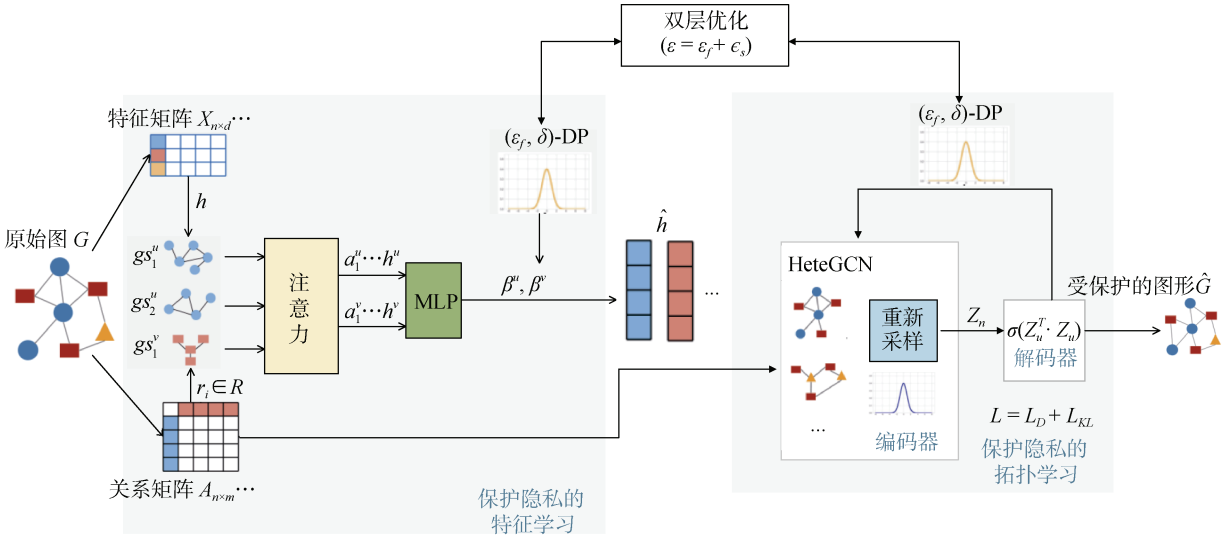


图 21 HeteDP 的整体框架

Figure 21 Overall Framework of HeteDP

在自然语言处理领域, Igamberdiev 和 Habernal[82] 提出对于以图形式出现的文档(如引文或社交网络), 隐私信息如以个人资料或关系为边缘的文档很容易被泄露, 因为经过训练的模型可能会泄露原始输入。因此他们探索了 GCN 的差分隐私训练, 展示了隐私效用权衡的本质。方法包括在 DP 设置中将一种易于实现的图分割算法应用于 GCN, 将图划分为子图, 同时避免对原始数据进行额外查询。除此之外, 他们还使 SGD-DP[83]适用于 GCN, 并且提出了一个差分隐私的 Adam[84]版本, 称为 Adam-DP。在未来, 可以进一步探索图拆分选项, 这些选项可以利用图结构, 而不是均匀的随机拆分。

Mueller 等人[85]正式将 DP-SGD 的应用扩展到图分类任务。如图 22 所示是在指纹数据集上进行图分类的差分隐私训练的方法概述。在步骤(1)中, 指纹图像被转换成图, 然后在步骤(2)中将图传递给 GNN 模型, 该模型用差分隐私随机梯度下降(DP-SGD)进行训练。对各个梯度进行修剪, 然后进行平均, 并添加高斯噪声。值得注意的是, 该工作研究了一些数据规模较小, 比较少见的图数据集以解决不同领域包括化学分子、指纹和心电图上的左束支传导阻滞检测的分类问题。上述几篇[80-82,85]拓宽了不同领域差分隐私与 GNN 的结合应用。

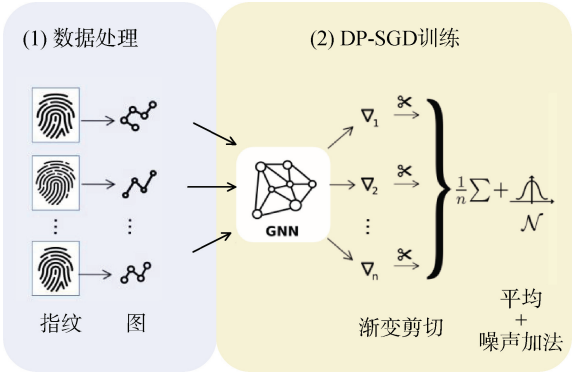


图 22 在指纹数据集上进行图分类的差分隐私训练方法

Figure 22 Differential Privacy Training Method for Graph Classification on the Fingerprint Dataset

Daigavane 等人[86]提出了一种节点级方法, 也是通过扩展标准 DP-SGD 算法和通过将结果二次采样到有界度图数据的隐私放大。然而, 他们的方法无法提供推理隐私, 并且仅限于 1 层 GNN, 因此无法利用高阶聚合。Sajadmanesh 等人[87]随后提出 GAP, 是第一种基于应用程序需求提供边缘级或节点级隐私保证的方法。

如图 23 所示为 GAP 的整体框架, GAP 所设计的 GNN 架构具体包括三个模块, 首先是编码器模块, 编码器仅使用节点特征(X)和标签(Y)进行训练, 不依赖于图结构, 然后是聚合模块, 编码的特征被提供给聚合模块以计算多跳聚合节点嵌入, 最后是分类模块, 在私有聚合上进行训练用于标签预测, 而无需进一步查询图的边缘。该方法不依赖于公共数据, 可以利用一阶邻居之外的多跳聚合, 并在不增加额外成本的情况下保证推理隐私, 并且可以很好的权衡准确性和隐私性。但是, 由于图的内在结构连通性, 在 GNN 中实现准确性和隐私之间的理想平衡仍然是具有挑战性的问题。

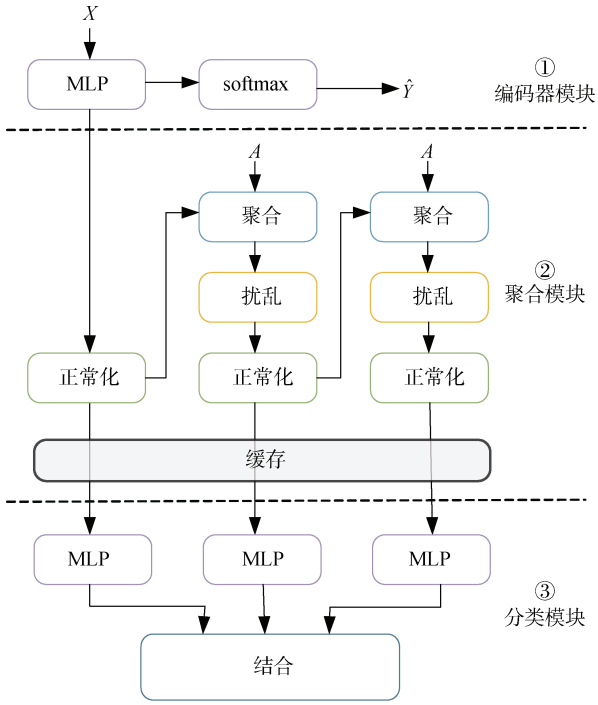


图 23 GAP 的整体框架

Figure 23 Overall Framework of GAP

4.2 对抗性隐私保护方法

Li 等人[90]对属性和链路都提出对抗性隐私保护方法, 给出了一种图对抗性训练框架(Adversarial Privacy Graph Embedding, 简称 APGE), 该框架集成了隐私解纠缠和隐私清除机制, 从学习的节点表示中删除用户的私有信息。

如图 24 所示为 APGE 的总体框图, 该方法保留了图的结构信息和效用属性, 同时隐藏了用户的私有属性以防御推理攻击, 其中链路预测损失和节点属性预测损失被组合在一起来发挥效果, 也就是说效用损失可以表示为:

$$\mathcal{L}_{utility}=\sum_{i=1}^{N}\sum_{j=1}^{N} l_{edge}(\mathrm{A}_{ij},\hat{\mathrm{A}})+\alpha\sum_{v\in\mathcal{V}}\sum_{c\in C} l_{attr}(x_v^c, f_c(\mathrm{h}_v)) \tag{9}$$

其中, x_v^c 指节点 v 的第 c 个属性, $f_c(\mathrm{h}_v)$ 指基于 v 节点的嵌入表示对第 c 个属性预测结果, l_{edge} 和 l_{attr} 则分别是对链路预测和属性推断的损失函数, C 指被推断的属性集合。对抗损失则可以表示成:

$$\mathcal{L}_{Adversarial}=\sum_{v\in\mathcal{V}_S}-[s_v\log(\hat{s}_v)+(1-s_v)\log(1-\hat{s}_v)] \tag{10}$$

其中, $\hat{s}_v$ 是指 $f_A(\mathrm{h}_v)$, 而 $\mathcal{V}_S$ 指带有敏感属性的节点集。

Hsieh 等人[89]提出了一种针对基于 GNN 的隐私攻击的对抗性防御模型 NetFense, 该模型能够降低攻击者的预测性能(即隐私保护), 并保持数据和模型的效用(即维持下游任务的性能)。与前述工作[90]不同

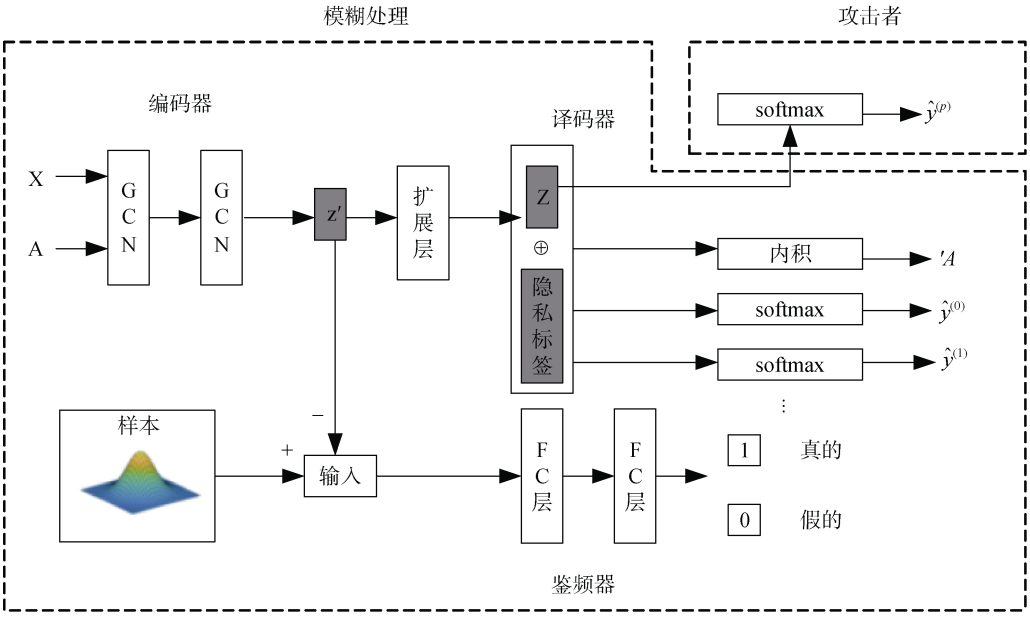


图 24 APGE 的总体框图

Figure 24 Overall Block Diagram of APGE

的是, 他们提出了一种基于图扰动的方法, 旨在对原始图进行扰动(即改变邻接矩阵)以欺骗攻击者, 同时保持下游任务(即节点分类)的效用。NetFense 在给定预训练的分类模型和攻击模型的情况下, 修改输入图的结构, 得到一个能够保护隐私的摄动图。另一个区别是 Netfense 假定隐私标签是二进制的。因此, 它的目标是将攻击者的预测精度降低到 0.5, 而不仅仅是最大化攻击者的损失函数。

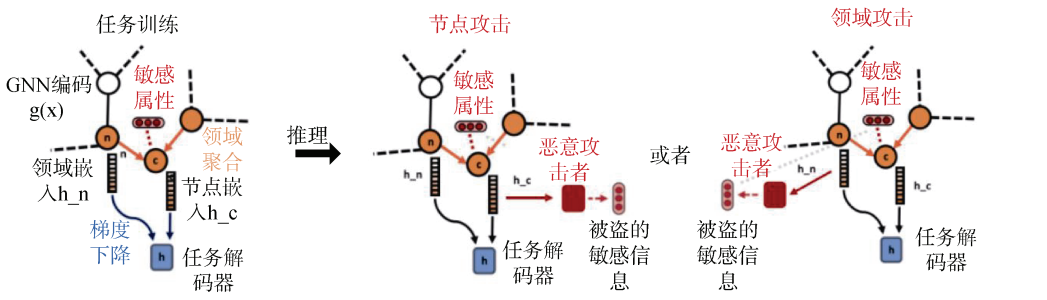
Liao 等人[91]研究了在使用图结构数据进行学习时, 利用信息混淆来保护敏感属性的问题。他们提出了一个对抗性训练框架(Graph Adversarial Networks, GAL), 用于局部过滤掉预先确定的敏感属性, 如图 25 所示, (a)是图学习算法存在的信息泄露问题, (b)则是对此提出的解决方法。图 25(a)为攻击者提供了一种从 GNN 获取敏感信息的方法: 在通过邻域聚合(1)进行训练后, 通过在图中生成的节点(2)及其邻居节点(3)的嵌入上运行神经网络, 他们能够恢复敏感信息。另一方面, 图 25(b)中 GAL 通过任务解码器(蓝色)和一个模拟最坏情况的攻击者(黄色)之间在嵌入(下降)和属性(上升)上的极小极大博弈来防御节点和邻域推理攻击。恶意对手在推理时很难从使用该框架训练的 GNN 嵌入中提取敏感属性。

Wang 等人[92]从互信息的角度提出了一种保护隐私的图表示学习框架, 研究了两个问题, 每个问题都涉及一个主要的学习任务和一个隐私保护任务。目标是学习节点表示, 以便它们可用于实现主要学习任务的高性能, 同时获得接近随机猜测的隐私保护任务的性能。具体来说, 首先通过互信息进行目标定义。其次, 为互信息项导出易处理的变分边界, 其中每个边界都可以通过神经网络进行参数化。第三, 训练这些参数化神经网络来近似真实的互信息并学习隐私保护节点表示。以上过程涉及三个模块: 节点嵌入函数(用于节点表示学习)、链接预测器(使用节点表示进行链接预测)和节点分类器(使用节点表示进行节点分类), 如图 26 所示。

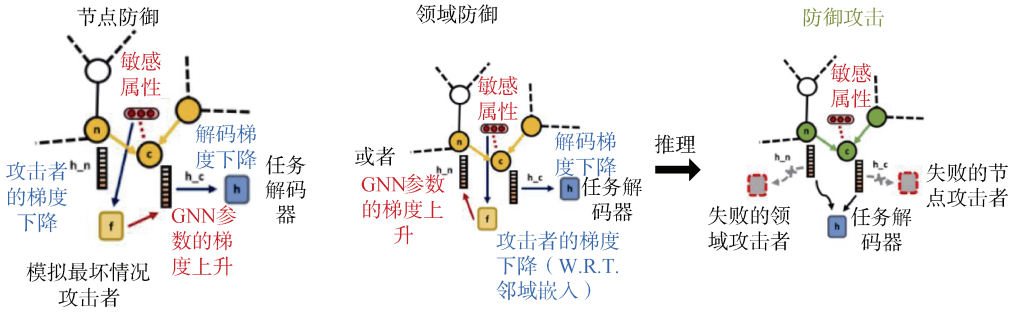
总结来说, 对抗性隐私保护方法的目标不仅仅需要防止目标信息(如节点标签、特征、图拓扑结构等)的泄露, 还需要保证分类任务的准确性, 从而达到两者之间的平衡。由于所保护的目标不同, 所设计的方法框架也会有所区别。

4.3 基于联邦学习的隐私保护方法

本节具体介绍不同的图联邦学习框架, 在此之前, 已有学者提出了图联邦学习或称联邦图机器学习(Federated Graph Learning or Federated Graph Machine Learning, FGL 或 FGML)的定义和挑战, 并给出了不同的分类方案。例如, Zhang 等人[93]根据图数据在客户端之间的分布方式将 FGL 分为图间 FGL, 图内 FGL 和图结构 FGL, 其中图内 FGL 被进一步划分为水平 FGL 和垂直 FGL, 虽然该论文分类法介绍得很详细, 但是所提到的代表性参考文献很少, 几乎只是集中于介绍分类标准; 而 Fu 等人[94]则将 FGL 分为两类, 分别是具有结构化数据的 FGL 和结构化的 FGL; 还有 Liu 等人[95]提出了二维分类法, 第一个维度侧重于 FL 和 GNN 的集成, 第二个维度



(a)客户端–服务器威胁模型中的信息泄露



(b)使用 GAL 进行信息混淆

图 25　图学习算法存在的信息泄露问题及解决方案

Figure 25　The Information Leakage Problem and Solution in the Graph Learning Algorithm

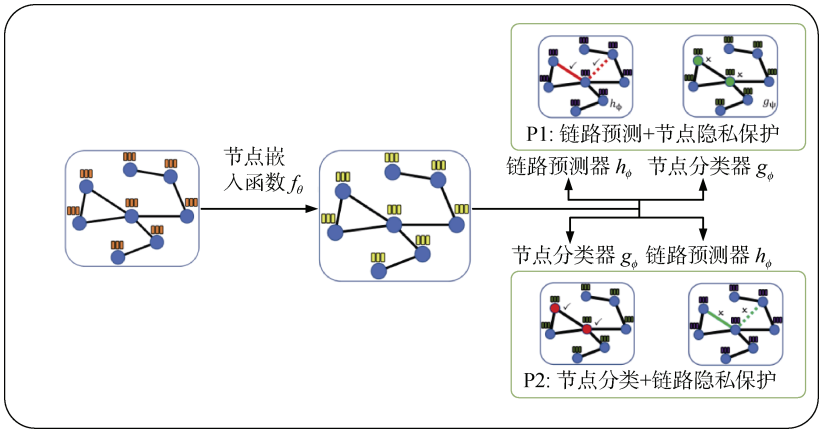


图 26　基于互信息的图表示学习框架

Figure 26　Graph Representation Learning Framework based on Mutual Information

则侧重于处理不同级别的图数据异构性的FL聚合解决方案。为了分类现有的关于图联邦学习的参考文献，我们将其像联邦学习的分类法类似大致分为基于横向联邦学习的隐私保护方法和基于纵向图联邦学习的隐私保护方法。然而，联邦虽然避免了数据共享，但还是有隐私泄露的风险，因此为了进一步保护用户隐私，已有学者提出在基本的图联邦学习框架上加入密码学等手段。

4.3.1　基于横向图联邦学习的隐私保护方法

横向图联邦学习指的是一个联邦学习设置，参与者共享相同的特征和标签空间，但是样本空间不同。

由于神经结构搜索(NAS)技术并不适用图联邦场景，且需要从头开始训练许多候选 GCN 模型，这对于横向联邦学习来说效率很低。为了解决这些挑战，Wang 等人[96]提出了 FL-AGCNS，一种适用于联邦场景的高效 GCN NAS 算法。如图 27 所示是 FL-AGCNS 的整体框架，其中的联邦进化优化策略是用于在联邦框架下有效地搜索高质量的GCN架构，也就是说该策略使分布式代理能够协同设计强大的 GCN 模型，同时将个人信息保留在本地设备上，而 GCN SuperNet 可用于降低 GCN NAS 的搜索成本，做到了有效性和高效率的结合。

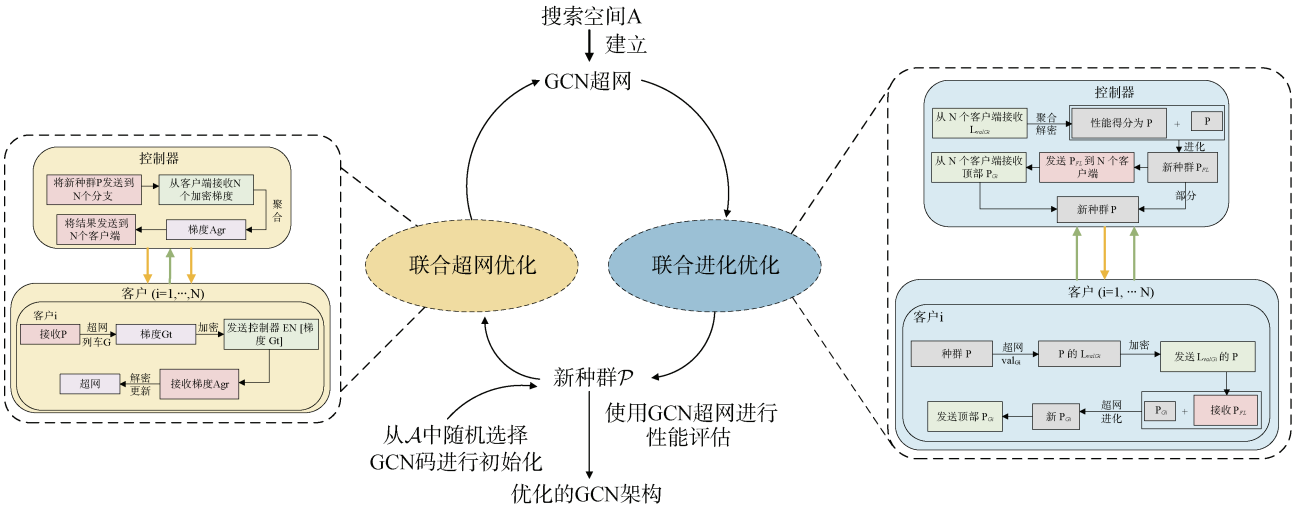


图 27　FL-AGCNS 的整体框架

Figure 27　Overall Framework of FL-AGCNS

虽然图数据集之间存在共同的模式, 但仍然可以观察到一定的异质性。Xie 等人[97]考虑此, 提出了一个用于图分类任务的图聚类联邦学习(Graph Clustered Federated Learning, GCFL)框架, 每个客户端拥有多图, 通过将各种强大的 GNN 模型集成到聚类 FL 中, 其中服务器可以根据 GNN 的梯度动态聚类客户端, 而无需额外的先验知识, 同时协作训练多个 GNN 作为同构客户端集群的必要条件。

随着图规模的扩大, 将图拆分成多个本地子图进行数据收集和存储不可避免, 那么若是每个客户端拥有全局图的一个子图, 一些节点可能有属于其他客户端的邻居。由于隐私问题, 一个节点只能聚合客户端内邻居的特征, 而不能访问位于其他客户端的特征, 导致节点表示不足。在这种情况下, Chen 等人[99]提出了 FedGraph 作为一种新的联邦图系统来实现保护隐私的分布式 GCN 学习方法, 为了解决 GCN 训练过程涉及在客户端之间嵌入共享的问题, 该方法使用了一种新颖的跨客户端图卷积操作, 在共享之前压缩嵌入, 从而可以很好地隐藏私人信息, 并且还提出了一种智能图采样算法来减少训练花费。Zhang 等人[100]同样也是在每个客户端拥有全局图的一个子图的情况下提出 Fedsage 和 Fedsage+这两个技术, 其中 FedSage 算法是将 GlobalSage 算法应用在 Subgraph FL 中实现的, 而 Fedsage+是在 FedSage 算法基础上的改进, 因为子图间缺少的关联可能使得 FedSage 算法无法很好地学习到全局特征, 受到子图间分布差异, 子图间的关联性无法通过 FedSage 捕捉等因素影响, 而 FedSage+通过对 Missing Neighbor 进行生成, 模拟出子图间的关联性, 以及改善子图间的分布差异, 弥补了 FedSage 算法的缺陷。FedSage+在客户端提出了一个节点特征生成器。为了训练生成器, 客户端在局部图中随机伸出一些现有的边。该生成器配有高斯噪声发生器, 训练其预测缺失邻域节点的数量并重建保留邻域节点特征。Chen 等人[102]提出了 FedGL, 它用服务器端边缘生成器通过 FL 客户端上传节点嵌入, 生成全局伪图, 分发给客户端修改其局部图, 用于 GNN 模型训练。Zheng 等人[103]提出了一种自动分离联邦 GNN 学习范式 ASFGNN, 该方法在客户端应用了模型插值技术。客户机的最终模型是全局模型和局部模型的组合。局部模型在更新过程中的百分比由混合权值控制。另外, 面对将整个图划分为几个子图的设置, 知识图任务场景中也有采用通过将一部分节点或边随机分配给一个客户端进行重叠分区方法来处理, 如 FedE[104], FKGE[105]等。

在推荐系统领域中, 用户数据的集中存储也可能会引起隐私问题和风险, 放在图领域上来说, 就是每个客户端中存储的用户物品图中的隐私泄露问题, Wu 等人提出基于 GNN 的隐私保护推荐的联邦框架 FedGNN[106]和 FedPerGNN[107], 以更安全的方式保护局部图隐私而不影响评分预测任务性能, 具体来说就是引入了一个第三方服务器, 该服务器只处理客户端的图形扩展。原始中央服务器首先生成并向客户端发送用于本地节点 IDs 和嵌入加密的公钥。然后客户端将密文上传到第三方服务器。第三方服务器通过检查交互节点 id 的密文来定位交互节点, 并将加密节点嵌入分发给客户端, 修改客户端的本地图, 用于后续的本地 GNN 训练。该方法的基本框图如图 28 所示, 首先从本地用户项交互数据推断出用户项图, 再基于推断出的图在每个用户客户端中

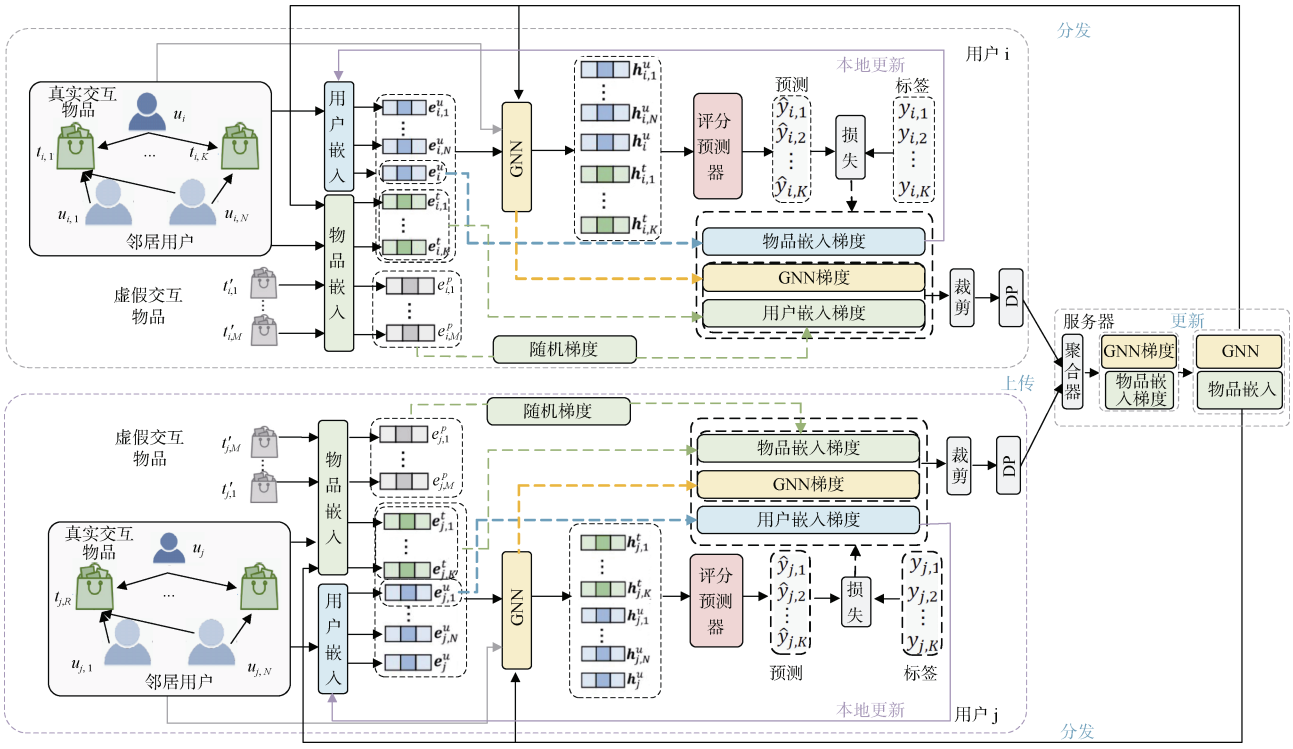


图 28　FedGNN 的基本框图

Figure 28　The Basic Block Diagram of FedGNN

局部训练 GNN 模型。每个客户端将 GNN 的局部梯度上传到服务器进行聚合，然后发送给用户客户端更新本地 GNN 模型。由于局部梯度可能包含隐私信息，我们将局部差分隐私技术应用于局部梯度以保护用户隐私。此外，为了保护用户与之交互的项目，我们建议将随机抽样的项目合并为伪交互项目，以实现匿名性。同样，FeSoG[108]也应用了局部差分隐私技术，使用动态本地差分隐私(LDP)将加密梯度上传到服务器进行 FL 聚合。

4.3.2　基于纵向图联邦学习的隐私保护方法

垂直图联邦学习是一种样本空间相同，但不同特征和标签空间的联邦学习设置。

Zhou 等人[109]提出一种用于垂直图联邦学习隐私保护的 GNN 节点分类学习范式 PPGNN。如图 29 所示是 PPGNN 的流程图，PPGNN 分为三个计算图(Computational graph, CG)，即私有特征和边缘相关计算(CG1，红色表示)、非私有数据相关计算(CG2，绿色表示)和数据持有者上私有标签相关计算(CG3，蓝色表示)，包含了三个关键步骤，即生成初始节点嵌入(步骤 1)、生成局部节点嵌入(步骤 2)和生成全局节点嵌入(步骤 3)。该方法通过将图分成几个部分，用于不同的客户端，同时采用安全多方计算，保证了数据的保密性和效率。但是假如攻击者是服务器，那么该学习范式可能就会有隐私泄露的风险。

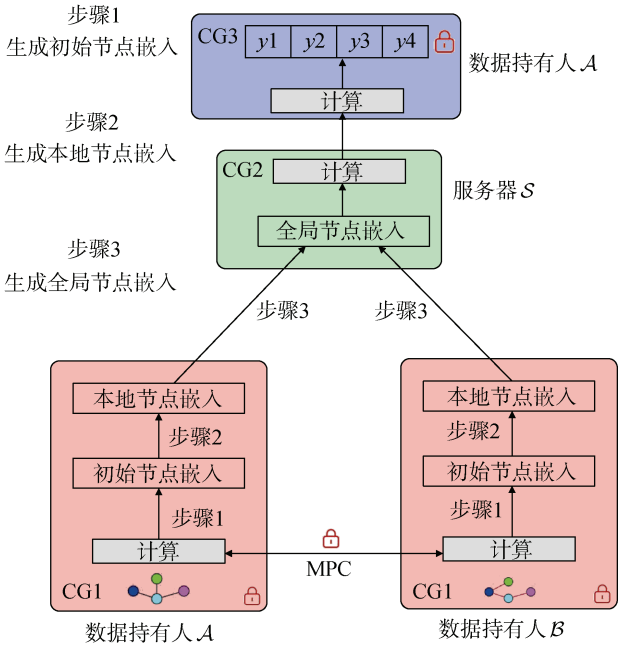


图 29　PPGNN 的流程图

Figure 29　Flowchart of PPGNN

FedVGCN[110]也是一种用于数据垂直分割设置下隐私保护节点分类任务的联邦 GCN 学习范式，可以推广到现有的 GCN 模型中。它假设每个客户端拥有一张图，同时为了体现隐私保护，该方法采用了加性同态加密(AHE)。整个过程将计算图数据分成两部分。

对于训练过程的每次迭代, 双方在同态加密下相互传递中间结果。但是对于其他更复杂的 GNN 模型并未涉及, 而且其并未考虑不同的攻击者假设造成的隐私泄露后果。FedSGC[111]在客户端在将敏感信息发送给另一方进行 GNN 模型参数更新之前, 也使用加性同态加密(AHE)对敏感信息进行加密, 它假设只有两个客户端没有中央服务器。图拓扑和节点特征由两个客户端拥有。拥有节点标签的客户端是创建加密密钥对的活动方。SGNN[112]不同于前面的密码学方法, 也没有共享图的原始邻接矩阵, 而是计算了一个基于动态时间扭曲(Dynamic Time Warping, DTW)算法的相似矩阵来传递相同的图拓扑, 但隐藏了原始结构。为了保护节点特征的隐私性, 采用单热编码将原始特征映射到矩阵中。然后将来自不同客户端的信息上传到服务器, 用于训练节点分类任务的全局 GNN 模型。FMLST[113]假设存在一个由具有相同节点的所有客户机共享的全局模式图。他们使用多层感知器(MLP)将局部时空(ST)模式和全局时空(ST)模式融合在一起, 并将连接的模式作为输入。客户端利用全局模式通过评估全局和局部模式图之间的差异来个性化他们的本地模式图。Wang 等人[114]针对节点分类任务提出了一种新的图联邦学习框架 GraphFL。该框架将一种模型不可知的元学习(MAML)纳入图联邦学习处理图数据, 具体来说首先通过遵循 MAML 的训练方案在服务器上学习全局模型, 再利用现有的联邦学习方法进一步更新全局模型, 使其能够在测试节点上实现良好的泛化。除此之外, 为了解决现有的联邦学习方法主要关注训练数据和测试数据共享同一标签域的问题, 在联邦学习框架中重新制定 MAML, 也就是说重新定义了一个不同于现有联邦学习方法的新目标函数。这样就可以在服务器上为所有客户端学习一个共享的全局模型, 这样全局模型可以快速适应标签域与训练节点不同的测试节点。

为了减少客户端数据分布的统计异质性, FLIT[115]通过根据预测置信度对样本重新加权来解决非独立同分布数据问题。为了使局部训练在客户端之间更加一致并避免局部数据的过拟合, 它将更多的权重放在样本上, 并且不同于前面的节点分类任务, 它解决的是图分类任务, 并且假设每个客户端拥有多个图。

总结来说, 跨客户端操作往往会带来巨大的通信开销, 因此, 在考虑基于联邦学习的隐私保护方法时, 我们通常需要保证方法性能和效率两者的平衡, 不仅如此, 联邦化只能保证本地数据的隐私性, 并不能保证传输过程的隐私性, 因此它往往会和其他隐私保护方法(如差分隐私, 同态加密等)结合以更好地保证数据隐私。

另外, 除上述分类标准之外, 若按照联邦客户端架构分类, 还可以将基于联邦学习的隐私保护方法分为去中心化的架构[116-117]和客户端-服务端架构[118-119]。一般的横向和纵向联邦图学习框架基本设置都是客户端-服务端架构, 再将客户端-服务端架构细分成服务器端进行 GNN 训练和客户端进行 GNN 训练。此外, He 等人[120]鉴于目前没有合适的平台用于 GNN 的联邦学习, 还提出了用于联邦 GNN 的开源 FL 基准系统 FedGraphNN, 包括了来自不同领域的广泛的数据集、流行的 GNN 模型和 FL 算法。

5 数据集及评价指标

已有的面向 GNN 的隐私攻击以及防御, 采用相对固定的数据集与评价指标, 方便不同方法之间的性能比较, 因此本节总结了常用的数据集和评价指标。

5.1 数据集

GNN 常用的数据集可以分为引文数据集, 交易记录数据集, 社交网络数据集, 生物化学数据集。这些数据集的详细内容如表 4 所示, 我们将其具体代码开源在 https://github.com/13estdeda/graphdata.git。其中, 由于生物化学网络数据集由多个大小不同的图构成, 因此本文展示了每个数据集中的平均节点和连边数。Cora, Citeseer, Pubmed 这三个引文数据集是节点分类中最常用的数据集, 因此这三个数据集也被广泛应用于多种隐私攻击场景。共同作者网络以及共同购买者网络也被应用于节点分类, 目前只在这些数据集上进行模型窃取攻击。由于存在多种大小的 Facebook 数据集, 表中省略了它的统计数据。交通网络只记录了节点之间的连边关系, 并没有记录节点的特征和类标。生物化学数据集通常包含多个图, 因此被广泛应用于图分类任务下的隐私攻击。

5.2 评价指标

现有的评价指标主要有准确度、*F*1 评分、AUC 等, 以从不同的角度反映分类准确率。

准确率(*Accuracy*). 准确率是图数据领域较为常见的评价指标, 其在图下游任务中会被应用于评价节点分类任务以及图分类任务。准确度在隐私攻击中常被用于评价成员推断攻击以及属性推断攻击, 其计算过程可以被表示为:

$$Accuracy = \frac{TP + TN}{P + N} \tag{11}$$

其中, TP 表示实际为正样本被划分为正样本的样本数, TN 表示实际为负样本被划分为负样本的样本数, P 表示为正样本数, N 表示为负样本数。

表 4 数据集内容

Table 4 The details of datasets

类别	任务	数据集	图数	节点数	连边数	特征数	类标数
引文网络	MI/GR/ME	Cora[11]	1	2708	5429	1433	7
	MI	Cora_ML[136]	1	2995	10674	2879	7
	MI	Cora-full[136]	1	19793	65311	8710	70
	MI /GR/ME	Citeseer[11]	1	3327	4732	3703	6
	MI/PI/GR/ME	Pubmed[11]	1	19717	88648	500	3
共同作者网络	MI/ME	DBLP[137]	1	17716	105734	1639	4
	ME	Coauthor Physics[138]	1	34493	495924	8415	5
	ME	ACM[81]	1	3025	26256	1870	3
共同购买者网络	ME	Amazon Photos[139]	1	7650	143663	745	8
社交网络	PI	Facebook[33]	1	—	—	—	—
	MI/PI	LastFM[140]	1	7624	27806	7842	18
	MI/GR	Flickr[141]	1	89250	449878	500	7
	MI	Reddit[64]	1	232965	57307946	602	41
	GR	Twitch-ES[142]	1	4648	59382	3170	2
	GR	Twitch-RU[142]	1	4385	37304	3170	2
	GR	Twitch-DE[142]	1	9498	153138	3170	2
	GR	Twitch-FR[142]	1	6549	112666	3170	2
	GR	Twitch-ENGB[142]	1	7126	35324	3170	2
	GR	Twitch-PTBR[142]	1	1912	31299	3170	2
	MI/GR	Polblogs[143]	1	1222	19090	1490	2
交通网络	GR	USA[144]	1	1190	13599	—	—
	GR	Brazil[144]	1	131	1038	—	—
	PI	Pokec[70]	1	45036	170964	5	2
	GR	PPI[43]	1	14755	225270	50	121
生物化学网络	GR/PI	AIDS[145]	2000	15.69	16.20	42	2
	MI/PI/GR	NCI1[146]	4110	29.87	32.30	37	2
	GR	COX2[147]	467	41.22	43.44	3	8
	GR	DHFR[147]	756	42.42	44.54	3	9
	GR/MI/PI	ENZYMES[101]	600	32.63	62.14	21	6
	MI/GR	PROTEINS[32]	1113	39.06	72.81	4	2
	PI	OVCAR-8H[146]	4052	46.67	48.70	65	2
	MI/PI	DD[146]	1178	284.33	715.66	89	2
	PI	PC3[28]	2751	25.36	28.49	37	2
	PI	MOLT-4H[28]	3977	46.70	48.74	65	2

***F*1-评分。**$F1$-评分是分类问题的一个衡量指标, 是精密度和召回率的调和平均数, 其计算过程可以被表示为:

$$F1-score=\frac{2PR}{P+R} \tag{12}$$

其中精密度表示为 $P=TP/(TP+FP)$,召回率表示为 $R=TP/(TP+FN)$, FP 表示实际为负样本且被分类器划分为正样本的样本数, FN 表示实际为正样本且被分类器划分为负样本的样本数

AUC。AUC 表示 ROC 曲线下与坐标轴围成的面积, ROC 曲线的横坐标表示为 $TPR=TP/(TP+FN)$, 其纵坐标表示为 $FPR=FP/(FP+TN)$。ROC 曲线通常情况下都高于 $y=x$, 因此 AUC 的值处于 0.5 到 1 之间。

AP。AP 表示 PR 曲线下与坐标轴的面积, 其横坐标为召回率 R, 纵坐标为 P。PR 曲线就是综合考虑了模型要尽可能的召回所有的正样本, 但又不能预判错误太多。

Fidelity。在模型窃取任务中, 常用的分类指标并不完全适用, 因此通常会使用 Fidelity 指标来评价

模型之间的相似度。Fidelity 指标测量了目标模型与代理模型的预测同意度, 即预测为相同结果的数量在样本总数中的占比。

结合攻防方法, 隐私攻击方法对于隐私窃取的分类准确率(如 Accuracy)可以降低 17%以内的范围[74], 攻击准确率(如 AUC, AP 等)在不同数据集上最高能达到 95%以上[27], 现有文献对于不同背景知识的攻击分类已经较为详细, 但是他们往往只考虑一种下游任务, 因此对于任务迁移方面在未来还有较大的提升空间; 在隐私保护方法方面, 若分类准确率在隐私保护之后能够保持原分类任务准确率那么就说明该隐私保护方法能够平衡隐私性和任务性能, 如果是在引入联邦的情况保证较少的通信代价, 做到效率与性能的平衡, 那么也能说明该隐私保护方法的有效性, 现有的文献已经可以在保证性能不损失的情况下提高通信效率, 但是他们往往只考虑一个优化目标或一种场景, 对这一点来说还具有较大的提升空间。另外, 对于不同隐私保护方法的结合未来还有很大的提升空间, 以进一步平衡隐私性与任务性能。

6 图神经网络隐私安全应用场景

6.1 医疗卫生

图结构数据在医疗卫生领域普遍存在, 例如蛋白质分子网络, 抗病毒药物网络以及病人网络。目前已有大量的 GNN 在这些私人的医疗卫生网络上进行了训练, 并且被用于各种现实应用。例如, GNN 被用于挖掘慢性病患者的信息, 并通过用户的连接关系来预测一个用户是否有类似的慢性疾病或类似的症状[121]; GNN 也被用于建立病人电子健康记录的模型, 并且根据病人的历史记录对病人进行诊断预测[122]。但是这些应用往往会收集用户大量的数据, 这其中也会包含用户的隐私数据。因此, 为了确保病人敏感数据的隐私, 在医疗保健领域的应用需要对隐私进行保护。

6.2 电商推荐

GNN 也被应用于电商推荐领域, 这主要是由于大部分推荐系统数据集都具有图结构。例如, 电商平台应用程序中的交互数据可以由用户和项目节点之间的二分图表示, 观察到的交互由链接表示, 甚至用户行为序列中的项目转换也可以构造为图。此外, 相比于传统的推荐系统方法往往只考虑到一阶邻居来提升用户嵌入学习, GNN 能够学习到更多阶邻居嵌入, 且学习高阶邻居以被证明有利于推荐系统[123-125]。依靠 GNN 的这些优势, 有大量研究者将其应用于推荐系统, 例如, GNN 被用于学习用户历史会话记录, 并将用户的下一项推荐表述为图分类问题[126], 但是, 得到一个高性能的电商推荐系统需要大量的用户信息, 这些信息可能从基于 GNN 的电商推荐系统泄露。因此, 为了保护用户的隐私, 已经提出了各种隐私保护电商推荐系统。例如, 通过联邦学习方法对模型进行分布式训练, 并通过伪标签技术和项目抽样来对用户隐私进行保护等等。

6.3 社交媒体

图结构数据普遍存在于社交媒体领域[30,140-143], 如通过用户之间的点赞、评论、分享等互动建立的社交网络。利用这些数据, GNN 可以分析社交网络中的图结构和节点特征, 从而导致潜在的隐私泄露风险和攻击行为; GNN 也可以通过设计隐私保护的节点分类算法或者图生成算法, 来保护个人隐私信息, 对用户数据进行匿名化处理, 以保护个人隐私; 利用 GNN 还可以帮助发现用户数据共享与隐私保护之间的最佳平衡点, 从而更好地保护用户隐私的同时推动数据共享和社交媒体的良好发展。

6.4 金融分析

最近, 不少研究者将 GNN 应用于金融应用, 如违约风险预测[129-130]和钓鱼节点检测[131]。在贷款违约风险中, 通过 GNN 对担保网络或者用户关系网络进行学习, 能够得到更准确的预测结果。在钓鱼节点检测任务中, GNN 中还对交易网络进行了学习。与其他领域的 GNN 类似, 交易网络中也包含了大量的用户隐私数据, 如用户的性别和年龄。因此为了保护用户的隐私数据, 在金融分析领域需要对隐私进行保护。

6.5 交通预测

交通预测问题涉及大数据量、高维、多动态, 包括交通事故等紧急情况, 并且常常具有基于图结构的数据形式[144], 例如道路图以道路交叉点为节点, 道路连接点为边。因此 GNN 非常适合于交通预测问题, 利用它可以捕捉空间依赖性。然而, 交通网络中包含着许多隐私数据包含用户个人隐私信息、道路信息等, 若是攻击者利用该漏洞入侵并篡改信息, 很有可能造成交通秩序的紊乱, 因此交通数据的安全与公民个人隐私安全联系紧密, 甚至关乎着公共利益和国家安全。

7 图神经网络隐私安全 VS 图像/自然语言处理深度模型隐私安全

本节我们将 GNN 与自然语言处理及图像领域中的深度模型进行比较并从不同角度进行分析。

数据类型: GNN 模型主要用于处理图结构数据, 可以对图中的节点和边进行建模和分析。而图像/自然语言处理的模型广泛应用于处理图像和文本数据。

数据规模和维度: GNN 处理的图结构数据通常具有不确定的规模和维度, 节点和边的数量可以根据具体应用而变化。相比之下, 图像和自然语言处理深度模型处理的数据通常具有确定的规模和维度, 如固定大小的图像或固定长度的文本。

隐私泄露的形式: GNN 的隐私泄露主要涉及图中节点和边的身份和关联信息, 可能导致实体和关系的泄露。而图像/自然语言处理深度模型的隐私泄露更关注原始数据的泄露, 可能导致图像或文本中的敏感信息暴露。

威胁模型: GNN 面临的威胁模型可能涉及对图结构和节点身份的推断、重构攻击等。而图像/自然语言处理深度模型可能面临的威胁模型更多涉及数据泄露、对抗性攻击、模型欺骗等。

隐私保护技术的适用性: 由于数据类型和处理方式的不同, GNN 和图像/自然语言处理深度模型在隐私保护技术的适用性上也有所区别。例如, 对于 GNN, 差分隐私技术可以直接应用于节点和边的属性或连接关系, 而对于图像/自然语言处理深度模型, 差分隐私技术可能需要在模型训练过程中进行调整或结合其他技术。

数据共享和协作: GNN 通常用于分析和挖掘图数据, 涉及多方数据的共享和协作。在隐私安全方面, 涉及多方数据的共享和协作可能增加隐私泄露的风险。而图像/自然语言处理深度模型可能更多地集中在单一数据源上, 数据共享和协作的问题相对较少。

8 未来研究方向

本节我们针对面向 GNN 的隐私攻击和保护, 探讨其在未来的研究发展方向, 从不同角度分析之后可发展的研究内容。

GNN 预训练的隐私攻击和保护: 模型预训练中进行隐私的攻击和保护已成为一种常用的方案, 以使得缺乏标签得下游任务受益。近年来, GNN 的监督任务[132]和自监督任务[127-128]的预训练都取得了很大的成功。预训练后的 GNN 参数将用于下游任务, 可能导致私有信息泄露。然而, 现有的隐私攻击大多集中在黑盒设置上, 并没有对模型发布导致的信息泄露进行调查。因此, 需要对预训练的 GNN 模式进行隐私攻击以及相应的防御方法的探索。

隐私攻击和保护性能与主任务性能间的权衡: 虽然已经提出了应用差分隐私、联邦学习或对抗学习的方法来保护训练数据的隐私, 但很少讨论隐私保护性能与预测精度之间的关系。例如, 在差分隐私 GNN[69,79-80]中, 通常不评估防御各种隐私攻击的实际性能。而在对抗性隐私保护中[90-92], 如何控制预测性能与隐私保护之间的平衡仍然没有得到很好的讨论。此外, 往往缺乏评估针对隐私攻击的防御性能与模型性能指标之间权衡的标准方法。

分布式学习环境下的 GNN 的隐私安全: GNN 结合联邦学习在保护数据隐私方面显示出了良好的结果[96-97,99-100,102-103,109-112,114], 特别是在医疗保健和金融应用中。大多数现有工作的重点是在不损害原始数据隐私的情况下构建全局 GNN 模型的体系结构学习和知识共享。已有研究针对联邦学习攻击和防御的隐私性和鲁棒性进行了全面调查包括威胁模型, 隐私攻击和防御, 以及中毒攻击和防御。然而, 很少有研究关注图联邦下的隐私攻击及隐私保护。

GNN 公平性与隐私性的平衡: 模型在学习过程中可能会因为弱势群体占比小而产生歧视, GNN 的公平性就是确保模型的决策和预测对不同个体和群体都是公正和无偏的, 可是在解决公平性问题的过程中需要考虑某些隐私属性如性别、年龄等, 从而造成隐私数据的泄露, 最近, 已有研究开始探索 GNN 的隐私和公平领域[71,148]。从不平衡的角度来看, 了解哪些节点更容易受到隐私攻击, 以及它们如何根据敏感特征和/或类标签与多数/少数群体保持一致, 将是比较有趣的关注点。平衡公平性和隐私性是一个复杂的问题, 并且在不同的应用场景和背景下可能存在权衡和取舍。

针对复杂图结构数据的隐私安全: 虽然大多数研究 GNN 中隐私攻击和保护的努力都集中在简单图[69,76]上, 但在许多现实世界的应用中, 复杂系统可以更好地用复杂图来表示, 在这些应用中, 已有研究证明 GNN 已经做出了专门的努力例如超图[133]、多维图[134]、动态/时间图[113]、知识图[104-105]等, 那么对于复杂图结构数据中就会有可能存在不同级别的隐私安全问题。

生成式 AI 对图结构数据隐私安全的影响: 最近在图像/NLP 领域出现的生成式 AI 引起了许多隐私问题, 因为生成的图像可能包含侵犯公司版权的敏感信息, 导致机密信息的泄露。由于这些生成技术可以很容易地适应于图结构数据[42], 同样的隐私问题也可能出现。具体举例来说, 在分子生成中, 生成的分子可能包含机密的子结构; 在社交网络领域, 如果生成过程涉及用户嵌入, 则生成的内容可能会反映该用户邻居的个人资料信息。

9 结束语

随着 GNN 在处理图数据上的广泛应用, GNN 隐私安全问题也逐渐成为研究热点。本文总结了面向 GNN 的隐私安全问题, 包括隐私攻击方法和隐私保护方法, 总结了 GNN 的隐私安全所用到的常用数据集以及评价指标。同时讨论了当前 GNN 隐私保护的应用场景, 并对未来的研究方向进行讨论, 旨在为推动 GNN 隐私安全的进一步发展和应用提供帮助。

致 谢 本课题得到浙江省自然科学基金(No. LDQ23F020001), 国家自然科学基金(Nos. 62072406, 62406286), 浙江省重点研发计划(No. 2022C01018), 国家重点研发计划(No. 2018AAA0100801)资助。

参考文献

[1] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[C]. *2nd International Conference on Learning Representations,* 2014.

[2] Defferrard M, Bresson X, Vandergheynst P, et al. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering[C]. *The 30th International Conference on Neural Information Processing Systems*, 2016: 3844-3852.

[3] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[EB/OL]. 2017: 1710.10903. https://arxiv.org/abs/1710.10903v3.

[4] Cao S S, Lu W, Xu Q K, et al. Deep Neural Networks for Learning Graph Representations[C]. *The Thirtieth AAAI Conference on Artificial Intelligence*, 2016: 1145-1152.

[5] Kipf T N, Welling M. Variational Graph Auto-Encoders[EB/OL]. 2016: 1611.07308. https://arxiv.org/abs/1611.07308v1.

[6] Fan W Q, Ma Y, Li Q, et al. Graph Neural Networks for Social Recommendation[C]. *The World Wide Web Conference*, 2019: 417-426.

[7] Tan Q Y, Liu N H, Hu X. Deep Representation Learning for Social Network Analysis[J]. *Frontiers in Big Data*, 2019, 2: 2.

[8] Li X X, Zhou Y, Dvornek N, et al. BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis[J]. *Medical Image Analysis*, 2021, 74: 102233.

[9] Harl M, Weinzierl S, Stierle M, et al. Explainable Predictive Business Process Monitoring Using Gated Graph Neural Networks[J]. *Journal of Decision Systems*, 2020, 29(sup1): 312-327.

[10] Lv L, Cheng J B, Peng N B, et al. Auto-Encoder Based Graph Convolutional Networks for Online Financial Anti-Fraud[C]. *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2019: 1-6.

[11] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[EB/OL]. 2016: 1609.02907. https://arxiv.org/ abs/1609.02907v4.

[12] Tang J, Qu M, Mei Q Z. PTE: Predictive Text Embedding through Large-Scale Heterogeneous Text Networks[C]. *The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 1165-1174.

[13] Wang S H, Tang J L, Aggarwal C, et al. Linked Document Embedding for Classification[C]. *The 25th ACM International on Conference on Information and Knowledge Management*, 2016: 115-124.

[14] Chen J Y, Shi Z Q, Wu Y Y, et al. Link Prediction Adversarial Attack[EB/OL]. 2018: 1810.01110. https://arxiv.org/abs/1810.01110v2.

[15] Perozzi B, Al-Rfou R, Skiena S, et al. DeepWalk[C]. *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014: 701-710.

[16] Wang S, Tang J, Aggarwal C, et al. Signed network embedding in social media[C]. *Proceedings of the 2017 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2017: 327-335.

[17] Jin M, Chang H, Zhu W W, et al. Power Up! Robust Graph Convolutional Network *via* Graph Powering[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(9): 8004-8012.

[18] Lee J B, Rossi R, Kong X N, et al. Graph Classification Using Structural Attention[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 1666-1674.

[19] Tian F, Gao B, Cui Q, et al. Learning Deep Representations for Graph Clustering[C]. *The Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014: 1293-1299.

[20] Allab K, Labiod L, Nadif M. A Semi-NMF-PCA Unified Framework for Data Clustering[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(1): 2-16.

[21] Baehrens D, Schroeter T, Harmeling S, et al. How to Explain Individual Classification Decisions[J]. *Journal of Machine Learning Research*, 2010, 11: 1803-1831.

[22] Bojchevski A, Günnemann S. Adversarial attacks on node embeddings via graph poisoning[C]. *International Conference on Machine Learning*, 2019: 695-704.

[23] Chang H, Rong Y, Xu T Y, et al. A Restricted Black-Box Adversarial Framework towards Attacking Graph Embedding Models[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(4): 3389-3396.

[24] Funke T, Khosla M, Rathee M, et al. Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 35(8): 8687-8698.

[25] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144.

[26] Harper F M, Konstan J A. The MovieLens Datasets: History and Context[J]. *ACM Transactions on Interactive Intelligent Systems*, 2015, 5(4): 1-19.

[27] He X, Jia J, Backes M, et al. Stealing Links from Graph Neural Networks[C]. *USENIX Security Symposium.* 2021: 2669-2686.

[28] Zhang Z K, Chen M, Backes M, et al. Inference Attacks Against Graph Neural Networks[EB/OL]. 2021: 2110.02631. https://arxiv.org/abs/2110.02631v1.

[29] Sun Y W, Wang S H, Tang X F, et al. Adversarial Attacks on Graph Neural Networks *via* Node Injections: A Hierarchical Reinforcement Learning Approach[C]. *Proceedings of The Web Conference*

2020, 2020: 673-683.

[30] Rigaki M, Garcia S. A Survey of Privacy Attacks in Machine Learning[EB/OL]. 2020: 2007.07646. https://arxiv.org/abs/2007.07646v3.

[31] Wu B, Yang X W, Pan S R, et al. Adapting Membership Inference Attacks to GNN for Graph Classification: Approaches and Implications[C]. *2021 IEEE International Conference on Data Mining*, 2021: 1421-1426.

[32] Borgwardt K M, Ong C S, Schönauer S, et al. Protein Function Prediction *via* Graph Kernels[J]. *Bioinformatics*, 2005, 21(Suppl 1): i47-i56.

[33] Duddu V, Boutet A, Shejwalkar V, et al. Quantifying Privacy Leakage in Graph Embedding[C]. *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems*: *Computing*, *Networking and Services*, 2021: 76-85.

[34] Al-Rubaie M, Chang J M. Privacy-Preserving Machine Learning: Threats and Solutions[J]. *IEEE Security & Privacy*, 2019, 17(2): 49-58.

[35] Ji Z L, Lipton Z C, Elkan C, et al. Differential Privacy and Machine Learning: A Survey and Review[EB/OL]. 2014: 1412.7584. https://arxiv.org/abs/1412.7584v1.

[36] Kairouz P, McMahan H B, Avent B, et al. Advances and Open Problems in Federated Learning[J]. *Foundations and Trends® in Machine Learning*, 2021, 14(1/2): 1-210.

[37] Yang M M, Lyu L J, Zhao J, et al. Local Differential Privacy and Its Applications: A Comprehensive Survey[EB/OL]. 2020: 2008.03686. https://arxiv.org/abs/2008.03686v1.

[38] Dai E Y, Zhao T X, Zhu H S, et al. A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability[EB/OL]. 2022: 2204.08570. https://arxiv.org/abs/2204.08570v2.

[39] Wu B Z, Li J T, Yu J C, et al. A Survey of Trustworthy Graph Learning: Reliability, Explainability, and Privacy Protection [EB/OL]. 2022: 2205.10014. https://arxiv.org/abs/2205.10014v2.

[40] Scarselli F, Gori M, Tsoi A C, et al. The Graph Neural Network Model[J]. *IEEE Transactions on Neural Networks*, 2009, 20(1): 61-80.

[41] Masci J, Boscaini D, Bronstein M M, et al. Geodesic Convolutional Neural Networks on Riemannian Manifolds[C]. *2015 IEEE International Conference on Computer Vision Workshop*, 2015: 832-840.

[42] Liu C Y, Fan W Q, Liu Y Q, et al. Generative Diffusion Models on Graphs: Methods and Applications[EB/OL]. 2023: 2302.02591. https://arxiv.org/abs/2302.02591v3.

[43] Hamilton W L, Ying R, Leskovec J, et al. Inductive Representation Learning on Large Graphs[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 1025-1035.

[44] Zhou J, Cui G Q, Hu S D, et al. Graph Neural Networks: A Review of Methods and Applications[J]. *AI Open*, 2020, 1: 57-81.

[45] Gilmer J, Schoenholz S S, Riley P F, et al. Neural Message Passing for Quantum Chemistry[C]. *The 34th International Conference on Machine Learning - Volume 70*, 2017: 1263-1272.

[46] Wang X L, Girshick R, Gupta A, et al. Non-Local Neural Networks[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7794-7803.

[47] Battaglia P W, Hamrick J B, Bapst V, et al. Relational Inductive Biases, Deep Learning, and Graph Networks[EB/OL]. 2018: 1806.01261. https://arxiv.org/abs/1806.01261v3.

[48] Dai E Y, Wang S H, Dai E Y, et al. Say no to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information[C]. *The 14th ACM International Conference on Web Search and Data Mining*, 2021: 680-688.

[49] Zhao T X, Zhang X, Wang S H. GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks[C]. *The 14th ACM International Conference on Web Search and Data Mining*, 2021: 833-841.

[50] Wang J N, Zhang S, Xiao Y H, et al. A Review on Graph Neural Network Methods in Financial Applications[EB/OL]. 2021: 2111.15367. https://arxiv.org/abs/2111.15367v2.

[51] Xu B B, Shen H W, Sun B J, et al. Towards Consumer Loan Fraud Detection: Graph Neural Networks with Role-Constrained Conditional Random Field[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(5): 4537-4545.

[52] Yang F, Fan K J, Song D D, et al. Graph-Based Prediction of Protein-Protein Interactions with Attributed Signed Graph Embedding[J]. *BMC Bioinformatics*, 2020, 21(1): 323.

[53] Sankar A, Liu Y, Yu J, et al. Graph Neural Networks for Friend Ranking in Large-Scale Social Platforms[C]. *The Web Conference 2021*, 2021: 2535-2546.

[54] Arora S. A Survey on Graph Neural Networks for Knowledge Graph Completion[EB/OL]. 2020: 2007.12374. https://arxiv.org/abs/2007.12374v1.

[55] Jiang D J, Wu Z X, Hsieh C Y, et al. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery?A Comparison Study of Descriptor-Based and Graph-Based Models[J]. *Journal of Cheminformatics*, 2021, 13(1): 12.

[56] Girvan M, Newman M J. Community Structure in Social and Biological Networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99(12): 7821-7826.

[57] Garcia J O, Ashourvan A, Muldoon S F, et al. Applications of Community Detection Techniques to Brain Graphs: Algorithmic Considerations and Implications for Neural Function[J]. *Proceedings of the IEEE Institute of Electrical and Electronics Engineers*, 2018, 106(5): 846-867.

[58] Shchur O, Günnemann S. Overlapping Community Detection with Graph Neural Networks[EB/OL]. 2019: 1909.12201. https://arxiv.org/abs/1909.12201v1.

[59] Wang X F, Li J H, Yang L, et al. Unsupervised Learning for Community Detection in Attributed Networks Based on Graph Convolutional Network[J]. *Neurocomputing*, 2021, 456: 147-155.

[60] Zhang X T, Liu H, Li Q M, et al. Attributed Graph Clustering *via* Adaptive Graph Convolution[EB/OL]. 2019: 1906.01210. https://arxiv.org/abs/1906.01210v1.

[61] Dwork C, McSherry F, Nissim K, et al. Calibrating Noise to Sensitivity in Private Data Analysis[C]. Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 265-284.

[62] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[J]. *Foundations and Trends® in Theoretical Computer*

Science, 2014, 9(3/4): 211-407.

[63] Xu D P, Yuan S H, Wu X T, et al. DPNE: Differentially Private Network Embedding[C]. *Advances in Knowledge Discovery and Data Mining*, 2018: 235-246.

[64] Olatunji I E, Nejdl W, 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*, 2021: 11-20.

[65] He X L, Wen R, Wu Y X, et al. Node-Level Membership Inference Attacks Against Graph Neural Networks[EB/OL]. 2021: 2102.05429. https://arxiv.org/abs/2102.05429v1.

[66] Liu Z Y, Zhang X Y, Chen C Y, et al. Membership Inference Attacks Against Robust Graph Neural Network[C]. *Cyberspace Safety and Security*, 2022: 259-273.

[67] Conti M, Li J X, Picek S, et al. Label-Only Membership Inference Attack Against Node-Level Graph Neural Networks[C]. *The 15th ACM Workshop on Artificial Intelligence and Security*, 2022: 1-12.

[68] Zhang S, Ni W W. Graph Embedding Matrix Sharing with Differential Privacy[J]. *IEEE Access*, 2019, 7: 89390-89399.

[69] Sajadmanesh S, Gatica-Perez D, Sajadmanesh S, et al. Locally Private Graph Neural Networks[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 2130-2145.

[70] Wang X L, Wang W H, Wang X L, et al. Group Property Inference Attacks Against Graph Neural Networks[C]. *The 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022: 2871-2884.

[71] Wang X M, Gu T L, Bao X G, et al. Fair and Privacy-Preserving Graph Neural Network[C]. *Database Systems for Advanced Applications*, 2023: 731-735.

[72] Zhang Z X, Liu Q, Huang Z Y, et al. GraphMI: Extracting Private Graph Data from Graph Neural Networks[EB/OL]. 2021: 2106.02820. https://arxiv.org/abs/2106.02820v1.

[73] Wu F, Long Y H, Zhang C, et al. LINKTELLER: Recovering Private Edges from Graph Neural Networks *via* Influence Analysis[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 2005-2024.

[74] Shen Y, He X L, Han Y F, et al. Model Stealing Attacks Against Inductive Graph Neural Networks[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 1175-1192.

[75] Kasiviswanathan S P, Lee H K, Nissim K, et al. What Can we Learn Privately?[J]. *SIAM Journal on Computing*, 2011, 40(3): 793-826.

[76] Lin W Y, Li B C, Wang C. Towards Private Learning on Decentralized Graphs with Local Differential Privacy[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2936-2946.

[77] Zhu X C. Link Local Differential Privacy in GNNS *via* Bayesian Estimation[C]. *Companion of the 2023 International Conference on Management of Data*, 2023: 265-267.

[78] Papernot N, Abadi M, Erlingsson Ú, et al. Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data[EB/OL]. 2016: 1610.05755. https://arxiv.org/abs/1610.05755v4.

[79] Olatunji I E, Funke T, Khosla M. Releasing Graph Neural Networks with Differential Privacy Guarantees[EB/OL]. 2021: 2109.08907. https://arxiv.org/abs/2109.08907v2.

[80] Zhang S J, Yin H Z, Chen T, et al. Graph Embedding for Recommendation Against Attribute Inference Attacks[C]. *The Web Conference 2021*, 2021: 3002-3014.

[81] Wei Y C, Fu X C, Sun Q Y, et al. Heterogeneous Graph Neural Network for Privacy-Preserving Recommendation[EB/OL]. 2022: 2210.00538. https://arxiv.org/abs/2210.00538v2.

[82] Igamberdiev T, Habernal I. Privacy-Preserving Graph Convolutional Networks for Text Classification[EB/OL]. 2021: 2102.09604. https://arxiv.org/abs/2102.09604v3.

[83] Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 308-318.

[84] Kingma D P, Ba J, Hammad M M. Adam: A Method for Stochastic Optimization[EB/OL]. 2014: 1412.6980. https://arxiv.org/abs/1412.6980v9.

[85] Mueller T T, Paetzold J C, Prabhakar C, et al. Differentially Private Graph Classification with GNNS[EB/OL]. 2022: 2202.02575. https://arxiv.org/abs/2202.02575v2.

[86] Daigavane A, Madan G G, Sinha A, et al. Node-Level Differentially Private Graph Neural Networks[EB/OL]. 2021: 2111.15521. https://arxiv.org/abs/2111.15521v3.

[87] Sajadmanesh S, Shamsabadi A S, Bellet A, et al. GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation[EB/OL]. 2022: 2203.00949. https://arxiv.org/abs/2203.00949v3.

[88] Sajadmanesh S, Gatica-Perez D. ProGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees[EB/OL]. 2023: 2304.08928. https://arxiv.org/abs/2304.08928v2.

[89] Hsieh I C, Li C T. NetFense: Adversarial Defenses Against Privacy Attacks on Neural Networks for Graph Data[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(1): 796-809.

[90] Li K Y, Luo G C, Ye Y, et al. Adversarial Privacy-Preserving Graph Embedding Against Inference Attack[J]. *IEEE Internet of Things Journal*, 2021, 8(8): 6904-6915.

[91] Liao P, Zhao H, Xu K, et al. Information obfuscation of graph neural networks[C]. *International conference on machine learning. PMLR*, 2021: 6600-6610.

[92] Wang B H, Guo J Y, Li A, et al. Privacy-Preserving Representation Learning on Graphs: A Mutual Information Perspective[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 1667-1676.

[93] Zhang H D, Shen T, Wu F, et al. Federated Graph Learning — a Position Paper[EB/OL]. 2021: 2105.11099. https://arxiv.org/abs/2105.11099v1.

[94] Fu X B, Zhang B C, Dong Y S, et al. Federated Graph Machine Learning[J]. *ACM SIGKDD Explorations Newsletter*, 2022, 24(2): 32-47.

[95] Liu R, Xing P W, Deng Z C, et al. Federated Graph Neural Networks: Overview, Techniques and Challenges[EB/OL]. 2022: 2202.07256. https://arxiv.org/abs/2202.07256v2.

[96] Wang C N, Chen B Z, Li G, et al. FL-AGCNS: Federated Learning Framework for Automatic Graph Convolutional Network Search [EB/OL]. 2021: 2104.04141. https://arxiv.org/abs/2104.04141v1.

[97] Xie H, Ma J, Xiong L, et al. Federated Graph Classification over Non-IID Graphs[C]. *The 35th International Conference on Neural*

Information Processing Systems, 2021: 18839-18852.

[98] Olatunji I E, Hizber A, Sihlovec O, et al. Does Black-Box Attribute Inference Attacks on Graph Neural Networks Constitute Privacy Risk?[EB/OL]. 2023: 2306.00578. https://arxiv.org/abs/2306.00578v1.

[99] Chen F H, Li P, Miyazaki T, et al. FedGraph: Federated Graph Learning with Intelligent Sampling[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(8): 1775-1786.

[100] Zhang K, Yang C, Li X, et al. Subgraph federated learning with missing neighbor generation[J]. *Advances in Neural Information Processing Systems,* 2021, 34: 6671-6682.

[101] Dobson P D, Doig A J. Distinguishing Enzyme Structures from Non-Enzymes without Alignments[J]. *Journal of Molecular Biology*, 2003, 330(4): 771-783.

[102] Chen C, Hu W B, Xu Z Y, et al. FedGL: Federated Graph Learning Framework with Global Self-Supervision[EB/OL]. 2021: 2105.03170. https://arxiv.org/abs/2105.03170v1.

[103] Zheng L F, Zhou J, Chen C C, et al. ASFGNN: Automated Separated-Federated Graph Neural Network[J]. *Peer-to-Peer Networking and Applications*, 2021, 14(3): 1692-1704.

[104] Chen M Y, Zhang W, Yuan Z G, et al. FedE: Embedding Knowledge Graphs in Federated Setting[C]. *The 10th International Joint Conference on Knowledge Graphs*, 2022: 80-88.

[105] Peng H, Li H R, Song Y Q, et al. Differentially Private Federated Knowledge Graphs Embedding[C]. *The 30th ACM International Conference on Information & Knowledge Management*, 2021: 1416-1425.

[106] Wu C H, Wu F Z, Cao Y, et al. FedGNN: Federated Graph Neural Network for Privacy-Preserving Recommendation[EB/OL]. 2021: 2102.04925. https://arxiv.org/abs/2102.04925v2.

[107] Wu C H, Wu F Z, Lyu L J, et al. A Federated Graph Neural Network Framework for Privacy-Preserving Personalization[J]. *Nature Communications*, 2022, 13(1): 3091.

[108] Liu Z W, Yang L W, Fan Z W, et al. Federated Social Recommendation with Graph Neural Network[J]. *ACM Transactions on Intelligent Systems and Technology*, 2022, 13(4): 1-24.

[109] Zhou J, Chen C, Zheng L, et al. Privacy-preserving graph neural network for node classification[EB/OL]. arXiv preprint arXiv: 2005.11903, 2020.

[110] Ni X, Xu X L, Lyu L J, et al. A Vertical Federated Learning Framework for Graph Convolutional Network[EB/OL]. 2021: 2106.11593. https://arxiv.org/abs/2106.11593v1.

[111] Cheung T H, Dai W, Li S. Fedsgc: Federated simple graph convolution for node classification[C]. *IJCAI Workshops*, 2021.

[112] Mei G X, Guo Z Y, Liu S J, et al. SGNN: A Graph Neural Network Based Federated Learning Approach by Hiding Structure[C]. *2019 IEEE International Conference on Big Data*, 2019: 2560-2568.

[113] Li W Z, Wang S. Federated Meta-Learning for Spatial-Temporal Prediction[J]. *Neural Computing and Applications*, 2022, 34(13): 10355-10374.

[114] Wang B H, Li A, Pang M, et al. GraphFL: A Federated Learning Framework for Semi-Supervised Node Classification on Graphs[C]. *2022 IEEE International Conference on Data Mining*, 2022: 498-507.

[115] Zhu W, Luo J B, White A. Federated Learning of Molecular Properties with Graph Neural Networks in a Heterogeneous Setting [EB/OL]. 2021: 2109.07258. https://arxiv.org/abs/2109.07258v3.

[116] Pei Y, Mao R, Liu Y, et al. Decentralized federated graph neural networks[C]. *International Workshop on Federated and Transfer Learning for Data Sparsity and Confidentiality in Conjunction with IJCAI*, 2021.

[117] He C Y, Ceyani E, Balasubramanian K, et al. SpreadGNN: Serverless Multi-Task Federated Learning for Graph Neural Networks[EB/OL]. 2021: 2106.02743. https://arxiv.org/abs/2106.02743v1.

[118] Xing P W, Lu S T, Wu L F, et al. BiG-Fed: Bilevel Optimization Enhanced Graph-Aided Federated Learning[J]. *IEEE Transactions on Big Data*, 2024, 10(6): 903-914.

[119] Meng C Z, Rambhatla S, Liu Y, et al. Cross-Node Federated Graph Neural Network for Spatio-Temporal Data Modeling[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 1202-1211.

[120] He C Y, Balasubramanian K, Ceyani E, et al. FedGraphNN: A Federated Learning System and Benchmark for Graph Neural Networks[EB/OL]. 2021: 2104.07145. https://arxiv.org/abs/2104.07145v2.

[121] Baek J W, Chung K. Multi-Context Mining-Based Graph Neural Network for Predicting Emerging Health Risks[J]. *IEEE Access*, 2023, 11: 15153-15163.

[122] Li Y, Qian B Y, Zhang X L, et al. Graph Neural Network-Based Diagnosis Prediction[J]. *Big Data*, 2020, 8(5): 379-390.

[123] He X N, Deng K, Wang X, et al. LightGCN[C]. *The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 639-648.

[124] Wang X, He X N, Wang M, et al. Neural Graph Collaborative Filtering[C]. *The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019: 165-174.

[125] Ying R, He R N, Chen K F, et al. Graph Convolutional Neural Networks for Web-Scale Recommender Systems[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 974-983.

[126] Qiu R H, Li J J, Huang Z, et al. Rethinking the Item Order in Session-Based Recommendation with Graph Neural Networks[C]. *The 28th ACM International Conference on Information and Knowledge Management*, 2019: 579-588.

[127] Hu Z N, Dong Y X, Wang K S, et al. Gpt-Gnn[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 1857-1867.

[128] Qiu J Z, Chen Q B, Dong Y X, et al. GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 1150-1160.

[129] Cheng D, Tu Y, Ma Z W, et al. Risk Assessment for Networked-guarantee Loans Using High-order Graph Attention Representation[C]. *IJCAI,* 2019: 5822-5828.

[130] Liang T, Zeng G X, Zhong Q W, et al. Credit Risk and Limits Forecasting in E-Commerce Consumer Lending Service *via* Multi-View-Aware Mixture-of-Experts Nets[C]. *The 14th ACM*

International Conference on Web Search and Data Mining, 2021: 229-237.

[131] Chen J Y, Xiong H Y, Zhang D J, et al. TEGDetector: A Phishing Detector that Knows Evolving Transaction Behaviors[EB/OL]. 2021: 2111.15446. https://arxiv.org/abs/2111.15446v1.

[132] Hu W H, Liu B W, Gomes J, et al. Strategies for Pre-Training Graph Neural Networks[EB/OL]. 2019: 1905.12265. https://arxiv.org/abs/1905.12265v3.

[133] Feng Y F, You H X, Zhang Z Z, et al. Hypergraph Neural Networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 3558-3565.

[134] Ma Y, Wang S, Aggarwal C C, et al. Multi-dimensional graph convolutional networks[C]. *Proceedings of the 2019 siam international conference on data mining. Society for Industrial and Applied Mathematics*, 2019: 657-665.

[135] Wu B, Yang X W, Pan S R, et al. Model Extraction Attacks on Graph Neural Networks: Taxonomy and Realisation[C]. *The 2022 ACM on Asia Conference on Computer and Communications Security*, 2022: 337-350.

[136] Bojchevski A, Günnemann S. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking[EB/OL]. arXiv preprint arXiv:1707.03815, 2017.

[137] Pan S, Wu J, Zhu X, et al. Tri-party deep network representation[C]. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1895-1901.

[138] Shchur O, Mumme M, Bojchevski A, et al. Pitfalls of Graph Neural Network Evaluation[EB/OL]. 2018: 1811.05868. https://arxiv.org/abs/1811.05868v2.

[139] McAuley J, Targett C, Shi Q F, et al. Image-Based Recommendations on Styles and Substitutes[C]. *The 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015: 43-52.

[140] Rozemberczki B, Sarkar R, Rozemberczki B, et al. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models[C]. *The 29th ACM International Conference on Information & Knowledge Management*, 2020: 1325-1334.

[141] Zeng H Q, Zhou H K, Srivastava A, et al. GraphSAINT: Graph Sampling Based Inductive Learning Method[EB/OL]. 2019: 1907.04931. https://arxiv.org/abs/1907.04931v4.

[142] Rozemberczki B, Allen C, Sarkar R, et al. Multi-Scale Attributed Node Embedding[J]. *Journal of Complex Networks*, 2021, 9(1): 1-22.

[143] Adamic L A, Glance N. The Political Blogosphere and the 2004 U.S. Election: Divided they Blog[C]. *The 3rd International Workshop on Link Discovery*, 2005: 36-43.

[144] Ribeiro L F R, Saverese P H P, Figueiredo D R. *struc2vec*: Learning Node Representations from Structural Identity[C]. *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017: 385-394.

[145] Bai X, Hancock E, Ho T, et al. Structural, Syntactic, and Statistical Pattern Recognition[C]. *Lecture Notes in Computer Science*, 2008

[146] Morris C, Kriege N M, Bause F, et al. TUDataset: A Collection of Benchmark Datasets for Learning with Graphs[EB/OL]. 2020: 2007.08663. https://arxiv.org/abs/2007.08663v1.

[147] Sutherland J J, O'Brien L A, Weaver D F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships[J]. *Journal of Chemical Information and Computer Sciences*, 2003, 43(6): 1906-1915.

[148] Wang X M, Gu T L, Bao X G, et al. Individual Fairness for Local Private Graph Neural Network[J]. *Knowledge-Based Systems*, 2023, 268: 110490.

陈晋音 于 2009 年在浙江工业大学控制理论与控制工程专业获得博士学位。现任浙江工业大学信息工程学院教授, CCF 会员。研究领域为人工智能安全, 数据挖掘, 智能计算。Email: chenjinyin@zjut.edu.cn

马敏樱 于 2022 年在周口师范学院自动化专业获得学士学位。现在浙江工业大学控制科学与工程专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 图数据挖掘, 联邦学习等。Email: mmy021456@163. com

马浩男 于 2022 年在浙江工业大学电子信息专业获得学士学位。现在浙江工业大学控制科学与工程专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 图数据挖掘, 联邦学习等。Email: 13estdeda@gmail.com

郑海斌 于 2022 年在浙江工业大学控制科学与工程专业获得博士学位。现任浙江工业大学信息工程学院讲师, CCF 会员。研究领域为人工智能安全为深度学习, 人工智能安全, 图像识别。Email: haibinzheng320@gmail.com