

面向图神经网络的对抗攻击与防御综述

陈晋音^{1,2}, 张敦杰², 黄国瀚², 林翔², 鲍亮³

(1. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023;

2. 浙江工业大学信息工程学院, 浙江 杭州 310023;

3. 信息网络安全公安部重点实验室, 上海 200000)

摘要: 面向已有的图神经网络的攻击与防御方法, 较全面地综述了图神经网络对抗攻防技术与鲁棒性分析。首先, 综述了图神经网络在不同任务下的对抗攻击与基于不同策略的防御方法, 并全面介绍了鲁棒性分析技术; 随后, 介绍了常用的基准数据集与评价指标; 最后, 提出了未来可能的研究方向和发展趋势。

关键词: 图神经网络; 对抗攻击; 防御算法; 鲁棒性分析

中图分类号: TP29

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2021051

Adversarial attack and defense on graph neural networks: a survey

CHEN Jinyin^{1,2}, ZHANG Dunjie², HUANG Guohan², LIN Xiang², BAO Liang³

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China

2. The College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

3. Key Lab of Information Network Security, Ministry of Public Security, Shanghai 200000, China

Abstract: For the numerous existing adversarial attack and defense methods on GNN, the main adversarial attack and defense algorithms of GNN were reviewed comprehensively, as well as robustness analysis techniques. Besides, the commonly used benchmark datasets and evaluation metrics in the security research of GNN were introduced. In conclusion, some insights on the future research direction of adversarial attacks and the trend of development were put forward.

Keywords: graph neural networks, adversarial attack, defense algorithms, robustness analysis

收稿日期: 2020-07-02; 修回日期: 2020-10-09

通信作者: 鲍亮, baoliang@stars.org.cn

基金项目: 国家自然科学基金(62072406); 浙江省自然科学基金(LY19F020025); 信息网络安全公安部重点实验室开放课题项目(C20604)

Foundation Items: The National Natural Science Foundation of China (62072406), The Natural Science Foundation of Zhejiang Province (LY19F020025), Key Lab of Information Network Security, Ministry of Public Security (C20604)

论文引用格式: 陈晋音, 张敦杰, 黄国瀚, 等. 面向图神经网络的对抗攻击与防御综述[J]. 网络与信息安全学报, 2021, 7(3): 1-28.
CHEN J Y, ZHANG D J, HUANG G H, et al. Adversarial attack and defense on graph neural networks: a survey[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 1-28.

1 引言

随着人工智能技术的研究与发展,深度学习在自然语言处理^[1-3]、图像识别^[4]、信号处理^[5]、物体识别^[6]等领域中表现不俗,卷积神经网络就是其中的典型代表之一。图数据由于其强大的表达能力,在现实生活中有着广泛的应用^[7]。但是基于图结构不具有平移不变性,每一个节点周围结构可能大不相同,这使传统的深度学习模型难以发挥作用。针对具有不规则的空间结构的图数据,研究人员^[8-11]尝试将神经网络推广用以处理任意结构的图,图神经网络^[12-16]应运而生。

图神经网络可以巧妙地从业数据中提取特征,而这些被提取出来的特征可以完成许多图数据分析任务,如节点分类^[17-18]、链路预测^[19-20]、社区发现^[21-22]和图分类^[23]等。这些图数据分析任务又被广泛应用于社交网络^[24]、推荐系统^[25-26]、电子商务网络^[27]等实际场景。假如应用于这些实际场景的图神经网络算法存在明显的漏洞,可能会带来严重的后果。例如,在电子商务领域,网络水军的存在给真实用户带来了极大的困扰和误导;在社交网络中,一则虚假消息如果不能被有效地检测出,有可能导致谣言散播,造成不良影响。因此,图神经网络的安全性问题也是研究热点之一。

在对图神经网络的安全性研究过程中,研究者发现:图神经网络很容易被一些微小的扰动迷惑,这类扰动被定义为对抗性扰动。对抗攻击发生在模型的测试阶段,攻击者根据模型结构精心设计微小的扰动,并将其添加在原始数据中,从而使图神经网络模型误判,达到愚弄图神经网络模型的目的。而通过对图神经网络的攻击研究可以发现图神经网络存在的缺陷。根据这些漏洞提出相应的防御措施来增强模型的鲁棒性显得尤为重要。Zügner 等^[28]首次提出在图上进行对抗攻击,他们提出的 NETTACK 算法通过重要的数据特征(如度分布)选择候选的连边和特征,在评估函数的指导下选择修改得分最高的连边或特征来生成微小的对抗性扰动,并更新对抗性网络,以此使图神经网络失效。之后,面向图神经网络的对抗攻击方法的研究相继展开:Dai^[29]、

Wang^[30]、Zhou^[31]、Bojchevski^[32]、Sun^[33]、Chen^[34-37]等^[38-45]纷纷加入了对抗攻击的研究行列。此外,面向图神经网络的防御方法的研究陆续展开,如 Dai^[29]等尝试在训练中随机丢弃一些连边来达到一定的防御目的;Feng^[46]、Chen^[47]等利用对抗训练来使模型对对抗攻击具有一定的鲁棒性;而其他研究者分别从不同的角度提出了相应的防御方法^[48-56]。

Jin 等^[57]对现有的对抗攻击方法进行了一系列的描述和详细分类。同样地, Sun 等^[58]详细阐述了面向图卷积网络的对抗攻击方法。Chen 等^[59]根据攻击的不同应用场景和防御的不同原理分别对其进行了系统的分类,并强调了相关评价指标的重要性,对其进行了全面的调查和总结。值得注意的是,他们的研究主要集中于分析现有的对抗攻击和防御方法原理并对其分类。本文针对图挖掘的不同应用任务,分别详细介绍现有的对抗攻击、防御方法与相应的评价指标,并重点介绍模型鲁棒性分析技术,最后总结图挖掘攻防未来可能的研究方向。

本文旨在针对不同的攻击分类标准与防御策略分别定义对抗攻击与防御,并对已有的面向图神经网络的对抗攻击和防御方法、常用数据集及评价指标进行系统综述,并展望面向图神经网络攻防方法未来的可能研究方向和发展趋势。图神经网络鲁棒性分析方法作为面向图神经网络的防御方法中的一个分支,对模型或数据安全给出了更多的理论性分析与评估,本文同样对当前图神经网络的鲁棒性分析方法进行了详细的描述和分析。

2 图神经网络攻防理论分析

本节首先介绍了图神经网络常见的图数据类型和主要的下游图分析任务,随后给出了图神经网络攻击的统一建模,最后分别介绍了图神经网络攻击与防御的理论定义,并根据攻击与防御方法的不同层面进行分类与分析。

2.1 图神经网络

为了更好地描述图神经网络及其对抗攻防问题,表 1 列举了图神经网络常用的符号定义。通常定义 $G = \{V, E\}$ 为一个图,其中 $V = \{v_1, \dots, v_N\}$

表示 N 个节点的集合, $e_{i,j} = \langle v_i, v_j \rangle \in E$ 表示节点 v_i 和节点 v_j 间存在连边。一般情况下, 节点集合 V 和连边集合 E 所包含的信息通过邻接矩阵 $A \in \mathbb{R}^{N \times N}$ 来表示, 当节点 v_i 和节点 v_j 存在直接连边时, $A_{i,j} \neq 0$, 否则 $A_{i,j} = 0$ 。由 M 个图组成的图集合记作 $\mathcal{G} = \{G_1, \dots, G_M\}$, Y_n 为图 G_n 对应的类标。

2.1.1 图数据的类型

在实际应用中, 根据不同数据本身的特性, 图具有多种表现形式, 本节进一步介绍不同类型的图数据。

有向图与无向图: 有向图中节点对间的连边有特定的方向, 在邻接矩阵的表现形式中, $A_{i,j} \neq A_{j,i}$ 。与之相反, 无向图中的连边都是双向的, 即 $A_{i,j} = A_{j,i}$ 。有向图的典型代表是在线社交网络 Twitter 数据集^[60], 其中用户 A 到用户 B 的有向连边表示 A 跟随 B, 反之 B 并不一定跟随 A。Facebook 数据集^[61]作为典型的无向图, 其用户 A 与用户 B 之间的连边表示他们是否存在朋友关系, 因而这种关系是双向的。

有权图与无权图: 有权图中, 节点 v_i 到节点 v_j 的连边具有权重 $w_{i,j}$, 不同连边的权重值不同。无权图存在的连边没有权重属性, 所有连边被视为相同。

特征图: 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 $X_{\text{node}} \in \mathbb{R}^{N \times D_{\text{node}}}$ 与 $X_{\text{edge}} \in \mathbb{R}^{N \times D_{\text{edge}}}$ 。交通图^[62]具有连边特征图, 其包含道路的各种条件信息。在推荐系统^[63]中, 用户的各种信息构成了节点的特征图。

同构图与异构图: 在实际应用中, 节点与连边往往存在多种类型。由多个类型的节点/连边构成的图称为异构图, 如开放学术图 (OAG)^[64]存在论文、作者、领域、机构和地点 5 个类型的节点。类似于上述的 Facebook 数据集^[61], 若图仅包含一个类型的节点与连边, 则称之为同构图。

动态图与静态图: 动态图的节点、连边或节点/连边特征会随时间变化, 表示为 $G^{(t)}$ 。在静态图 G 中, 节点、连边或节点/连边特征是固定的。社交网络^[65]中用户的关系随时间改变, 是一种典型的动态图。静态图的典型例子是生物数据集^[66], 其分子结构不会随时间而改变。

图神经网络在实际应用中派生出了各种类型的图, 然而现有的攻防工作主要集中在无向、无权的静态同构图上, 本文以无向、无权, 并且具有节点属性 X 的静态同构图为例, 综述图神经网络的对抗攻击与防御工作。

表 1 常用符号及定义
Table 1 Definition of common symbols

符号	定义	符号	定义	符号	定义
$\mathcal{G}$	原始图集合	G	原始图	V	节点集合
E	连边集合	X	节点特征矩阵	A	邻接矩阵
M	图数量	N	节点数量	D	节点特征维度
$\mathcal{G}_L$	标记图集合	V_L	标记节点集合	E_L	标记连边集合
$e_{i,j}$	节点 v_i 与 v_j 间的连边	Y_n	n 的类标	Y_n	节点 n 的类标
$I_{i,j}$	连边 $e_{i,j}$ 的类标	C	社团集合	$\mathcal{T}$	目标实例集合
τ_i	目标实例	c_i	目标实例类标	$\mathcal{F}$	特征集合
$\mathcal{E}$	共同出现的特征矩阵	f	神经网络模型	$\tilde{f}$	防御模型
$\mathcal{L}$	损失函数	$\mathcal{L}_{\text{train}}$	训练损失函数	$\mathcal{L}_{\text{att}}$	目标攻击损失函数
$\hat{G}$	受扰动后的图	$\tilde{X}$	受扰动后的节点特征	$\hat{A}$	受扰动后的邻接矩阵
g	连边梯度矩阵	K	修改连边数	H	隐藏层中的特征映射数
Δ	扰动代价	Z	节点嵌入表示	$\ \cdot\ _0$	0 范数

2.1.2 不同的下游任务

图神经网络的发展促进了许多下游的图分析任务的应用, 本节根据下游分析任务关注的不同目标(节点、连边、图和社区结构), 主要介绍 4 类图挖掘任务: 节点分类任务、链路预测任务、图分类任务和社区检测任务。

节点分类: 节点分类任务在社交网络的用户识别中得到了广泛应用。给定一个具有部分标记节点 $V_L \subset V$ 和其他未标记节点 $V_U = V - V_L$ 的图 G , $y_L = \{y_1, \dots, y_{|V_L|}\}$ 为标记节点对应的节点类标。节点分类任务的目的是利用图 G 与标记节点 V_L 训练一个节点分类模型 $f^{\text{node}}(\cdot)$, 并利用该模型正确预测未标记节点 V_U 的类标。

链路预测: 链路预测任务在社交网络与交易网络中应用广泛, 如预测未来社交关系的变化或可能发生的交易。类似于节点分类任务, 针对一个图对象 G , 依据已知节点与连边关系, 预测未知的连边。以标记连边 $E_L \subset E$ 与对应的连边类标 $I = \{0, 1\}$ 替换 V_L 与 y_L , 通过训练链路预测模型 $f^{\text{link}}(\cdot)$ 预测未标记连边 $E_U = E - E_L$ 的类标, 其中 $I_{i,j} = 1$ 表示节点 v_i 与节点 v_j 间存在连边, 否则不存在。

图分类: 图分类任务常用于生物化学网络, 它将图看作一个整体, 聚合图内所有节点的信息并预测目标图的类别。图分类任务以图集合 $\mathcal{G} = \{G_1, \dots, G_M\}$ 作为输入, 利用具有部分标记图 $\mathcal{G}_L \subset \mathcal{G}$ 与对应的图类标 $Y_L = \{Y_1, \dots, Y_{|\mathcal{G}_L|}\}$ 训练一个图分类模型 $f^{\text{graph}}(\cdot)$ 来预测未标记图 $\mathcal{G}_U = \mathcal{G} - \mathcal{G}_L$ 。

社区检测: 社区检测任务通常根据节点间的连边关系将其划分到不同的社区中, 更好地揭示社区内的节点关系。给定一个图 G 与节点对应的社区 $C = \{C_1, \dots, C_p\}$, 社区检测任务利用图 G 训练社区检测模型 $f^{\text{comm}}(\cdot)$, 并利用该模型预测节点 V 的所属社区。

图分析任务学习的统一表示为

$$\theta = \arg \min_{\theta} \sum \mathcal{L}(f_{\theta}(G, \tau_i), c_i) \quad (1)$$

其中, $f_{\theta}(\cdot)$ 可以表示不同图分析模型, θ 为模型 $f_{\theta}(\cdot)$ 的参数。根据图神经网络下游任务的不同, 目标实例 τ_i 可以是节点 v_n 、连边 $e_{i,j}$ 和图 G_n , c_i 表示目标实例对应的类标(节点类标 y_n 、连边类标

$I_{i,j}$ 、图类标 Y_n 和节点所属社区 C_n)。

2.2 攻击统一建模

为了进一步描述图神经网络的对抗攻防问题, 本节针对图神经网络的各种下游图分析任务, 给出了统一的攻击定义。考虑到图数据包含的节点/连边特征信息, 将原始图表示为 $G = (A, X)$, 添加对抗扰动的图表示为 $\hat{G}$, 目标实例集合 $\mathcal{T} \subseteq S_U$, S_U 可以是 V_U 、 E_U 或 G_U , $f_{\theta}(\cdot)$ 为图神经网络模型, S_L 表示用于训练的实例集合。攻击者的目的在于尽可能使目标实例 $\tau_i \in \mathcal{T}$ 在 $f_{\theta}(\cdot)$ 上的损失最大化, 从而导致错误的预测结果。本文将在图神经网络中的攻击问题描述为

$$\begin{aligned} \max_{\tau_i \in \mathcal{T}} \sum \mathcal{L}_{\text{att}}(f_{\theta}(\hat{G}, \tau_i), c_i) \\ \text{s.t. } \theta^* = \arg \min_{\theta} \sum_{\tau_j \in S_L} \mathcal{L}_{\text{train}}(f_{\theta}(\tilde{G}, \tau_j), c_j) \end{aligned} \quad (2)$$

其中, $\tilde{G}$ 既可以是原始图 G , 也可以是经过扰动的对抗图 $\hat{G}$ 。攻击者通过最小化目标模型的训练损失 $\mathcal{L}_{\text{train}}$ 得到神经网络参数 θ^* , 并最大化攻击目标节点 v_i 的损失 $\mathcal{L}_{\text{att}}$ 实现攻击。控制扰动大小的扰动代价可描述为

$$\|\hat{A} - A\|_0 + \|\hat{X} - X\|_0 \leq \Delta \quad (3)$$

其中, $\|\cdot\|_0$ 表示向量中的非零元素个数, 该式将攻击前后网络邻接矩阵或节点特征矩阵的修改总数限制在 Δ 内。

图 1 为对抗攻击的示意图(以修改连边为例), 图 2 为攻击前后对后续任务的影响示意图。

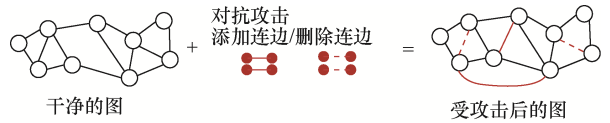


图 1 针对图卷积神经网络的对抗攻击
Figure 1 Adversarial attack on graph neural network

2.3 图神经网络攻击原理

2.3.1 不同知识背景的攻击

白盒攻击: 目标模型的参数、训练数据、类标和预测输出均可知, 攻击者根据这些已知信息进行攻击策略的选择, 如利用模型的梯度信息产生攻击。在实际情况中往往无法获得这些信息, 因此白盒攻击常用于评估目标模型在极端恶劣的

情况下的鲁棒性。

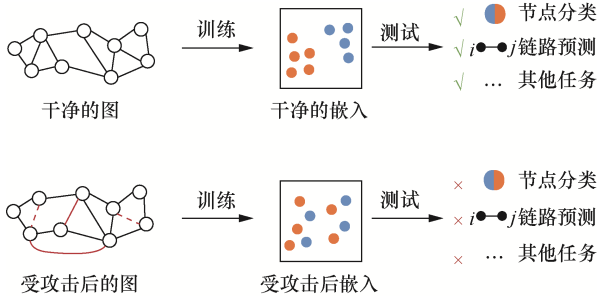


图2 对抗攻击对后续任务的影响
Figure 2 Impacts of adversarial attacks on downstream tasks

黑盒攻击：攻击者对目标模型的参数、训练数据、类标和预测输出均不可知，仅通过目标模型的输入和输出信息实施攻击，因此黑盒攻击更具有挑战性。

灰盒攻击：在模型结构不能完全已知情况下，使用带有类标的数据训练一个替代模型，近似目标模型结构，并对此模型实施白盒攻击生成，完成对真实目标模型的攻击，称之为灰盒攻击。

2.3.2 不同攻击策略

修改连边：图数据的拓扑结构在很大程度上决定了图的特性，在指定的攻击代价下增/删网络中的连边数量以达到对图数据拓扑结构的攻击，往往能对图神经网络产生明显的影响。连边攻击的代价可以表示为

$$\|\hat{A} - A\|_0 \leq \Delta \quad (4)$$

修改特征：节点或连边特征在图神经网络中同样起到关键的作用。在图数据中，节点/连边的属性可以表示为二值的 $X \in \{0,1\}^{N \times D}$ 或是连续的 $X \in \mathbb{R}^{N \times D}$ 。对于二值特征，攻击者通过反转部分特征值实现特征修改，而对于连续特征，通常对特征值做微小调整。特征攻击的代价可以表示为

$$\|\hat{X} - X\|_0 \leq \Delta \quad (5)$$

添加虚假节点及对应连边：攻击者为了保存网络原有节点与连边的信息，通过向网络中添加虚假节点和对应的部分连边实现攻击，添加虚假节点通常伴随着对应连边的增加，其攻击代价可以表示为

$$\|\hat{A} - A\|_0 + \|\hat{X} - X\|_0 \leq \Delta \quad (6)$$

2.3.3 不同的攻击能力

对抗攻击：在测试阶段添加对抗样本，而目标模型是通过原始网络 G 中的训练实例 $\tau_j \in S_L$ 与对应的实例类标 c_j 训练得到以 θ 作为参数的模型 f_θ 。在攻击的过程中，目标模型的参数 θ 保持不变，对测试阶段中的对抗样本做出错误判定，定义为

$$\begin{aligned} & \max_{\tau_i \in J} \sum \mathcal{L}_{\text{att}}(f_\theta(\hat{G}, \tau_i), c_i) \\ \text{s.t. } & \theta = \operatorname{argmin}_{\theta} \sum_{\tau_j \in S_L} \mathcal{L}_{\text{train}}(f_\theta(G, \tau_j), c_j) \end{aligned} \quad (7)$$

中毒攻击：针对训练阶段的一种攻击方式。攻击者在训练阶段改变训练数据集以影响目标模型的性能。以节点分类为例，利用经过扰动的网络 $\hat{G}$ 训练目标模型，得到以 θ^* 作为参数的模型 f_{θ^*} ，实现对模型的中毒攻击，表示为

$$\begin{aligned} & \max_{\tau_i \in I} \sum \mathcal{L}_{\text{att}}(f_{\theta^*}(\hat{G}, \tau_i), c_i) \\ \text{s.t. } & \theta^* = \operatorname{argmin}_{\theta} \sum_{\tau_j \in S_L} \mathcal{L}_{\text{train}}(f_\theta(\hat{G}, \tau_j), c_j) \end{aligned} \quad (8)$$

与式(2)不同的是，中毒攻击利用经过扰动的网络 $\hat{G}$ 训练模型参数 θ^* 。

2.3.4 不同的攻击目标

无目标攻击：以节点分类任务为例，无目标攻击并不关心目标样本的预测结果，只要预测是错误的，就认为攻击成功。

目标攻击：要求图神经网络将目标样本误分类成攻击者指定的类标。从难度上来说，有目标攻击的实现难于无目标攻击。

2.3.5 不同的攻击任务

根据上文所述的图攻击能力，中毒攻击需要改变目标模型的参数，这使中毒攻击相较于对抗攻击更加复杂。本文以中毒攻击为例，介绍攻击不同图神经网络任务的原理。

节点分类攻击：给定一个扰动网络 $\hat{G} = (\hat{A}, \hat{X})$ ，对于目标节点 v_i ，使训练得到目标节点分类模型 f_θ ，对扰动网络 $\hat{G}$ 中的目标节点 v_i 做出错误分类，可以表示为

$$\begin{aligned} & \max_{v_i \in V_U} \mathcal{L}_{\text{att}} \left(f_{\theta^*}(\hat{G}, v_i), y_i \right) \\ & \text{s.t. } \theta^* = \operatorname{argmin}_{\theta} \sum_{v_j \in V_L} \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}, v_j), y_j \right) \end{aligned} \quad (9)$$

图分类攻击：不同于节点分类，网络图分类攻击需要输入多个扰动网络集合 $\hat{G} = \{\hat{G}_1, \dots, \hat{G}_n\}$ 与相应网络类标 $Y_{\hat{G}_n}$ 来训练目标图分类模型 f_{θ^*} ，并对目标网络 $\hat{G}_i$ 做出错误分类，对式(9)变换得到

$$\begin{aligned} & \max_{\hat{G}_i \in \mathcal{G}_U} \mathcal{L}_{\text{att}} \left(f_{\theta^*}(\hat{G}, \hat{G}_i), Y_i \right) \\ & \text{s.t. } \theta^* = \operatorname{argmin}_{\theta} \sum_{\hat{G}_j \in \mathcal{G}_L} \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}, \hat{G}_j), Y_j \right) \end{aligned} \quad (10)$$

以多个网络作为输入，实现对某个网络的攻击。

链路预测攻击：给定扰动网络 $\hat{G} = (\hat{A}, \hat{X})$ 与标记连接 E_L ，给定需要预测的目标连边 E_U ，使目标链路预测模型 f_{θ^*} 对扰动网络 $\hat{G}$ 中的目标连接 E_U 做出错误预测，可表示为

$$\begin{aligned} & \max_{e_{i,j} \in E_U} \mathcal{L}_{\text{att}} \left(f_{\theta^*}(\hat{G}, e_{i,j}), I_{i,j} \right) \\ & \text{s.t. } \theta^* = \operatorname{argmin}_{\theta} \sum_{e_{u,v} \in E_L} \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}, e_{u,v}), I_{u,v} \right) \end{aligned} \quad (11)$$

其中， I 表示连边存在状态，值为 0 或 1，根据现有信息降低存在连边的可能性。

社区检测攻击：给定扰动网络 $\hat{G} = (\hat{A}, \hat{X})$ 与原网络的社团集合 $C = \{C_1, \dots, C_p\}$ ，目标社区检测模型 f_{θ^*} 将扰动网络 $\hat{G}$ 中的目标节点 v_i 划分到错误的社区中，用公式可以表示为

$$\begin{aligned} & \max_{v_i \in V_U} \mathcal{L}_{\text{att}} \left(f_{\theta^*}(\hat{G}, v_i), C_i \right) \\ & \text{s.t. } \theta^* = \operatorname{argmin}_{\theta} \sum_{v_j \in V_L} \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}, v_j), C_j \right) \end{aligned} \quad (12)$$

其中， C_i 为节点 v_i 所属的社团。

2.4 图神经网络防御原理

随着图神经网络攻击研究的逐步深入，面向图神经网络隐私的防御方法的研究也进展迅速，针对不同的对象提出了防御策略，可将防御问题描述如下。

$$\begin{aligned} & \min_{\tau_i \in T} \mathcal{L}_{\text{att}} \left(f_{\theta^*}(\hat{G}, \tau_i), c_i \right) \\ & \text{s.t. } \theta^* = \operatorname{argmin}_{\theta} \sum_{\tau_j \in S_L} \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}, \tau_j), c_j \right) \end{aligned} \quad (13)$$

其中， $\tilde{G}$ 既可以是原始网络 G ，也可以是经过扰动的网络 $\hat{G}$ 。防御的目标即最小化受攻击的模型的损失函数，使其与未受攻击的模型的损失尽可能相近。

本节根据防御策略的不同对当前的防御方法进行系统分类，并对一些面向图神经网络的鲁棒性方法进行原理分析。

2.4.1 不同的防御策略

1) 对抗训练：常用防御方法之一，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，该防御方法受对抗攻击方法的限制，即对未知的对抗攻击无法实现对抗训练完成防御。对抗训练分为两种类型。

① 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，目标可以描述为

$$\min_{\theta} \max_{\tau_i \in T} \sum \mathcal{L}_{\text{train}}(\tilde{f}_{\theta}(\tilde{G}, \tau_i), c_i) \quad (14)$$

其中， $\tilde{G}$ 是原始网络 G 与经过扰动的网络 $\hat{G}$ 的集合，对抗训练最小化原始网络 $\hat{G}$ 的目标函数，同时最大化扰动网络 $\hat{G}$ 的目标函数。

② 直接加入经过扰动的网络 $\hat{G}$ 进行对抗训练，额外学习对抗样本的属性。

$$\min_{\theta} \sum_{\tau_i \in T} \mathcal{L}_{\text{train}}(\tilde{f}_{\theta}(\hat{G}, \tau_i), c_i) \quad (15)$$

2) 对抗性扰动检测：一些工作更侧重于对抗性攻击的检测，此类方法通过研究对抗性样本和干净样本之间的差异实现对抗性扰动检测。对抗性扰动检测虽然不能起到直接防御的作用，但是可以作为一种对应图神经网络攻击的监控手段，提前发现对抗性扰动以更有效率地进行下一步的防御措施。

3) 启发式：此类方法在可接受的花费（指计算时间和空间）下给出图神经网络训练中待解决组合优化问题的每一个实例的一个可行解。

4) 图纯化：此类方法主要用于防御中毒攻击。图纯化方法将受到中毒攻击的图数据进行纯化，然后在纯化后的图上训练图神经网络模型。

5) 注意力机制：此类方法学习一种注意力机制，用来区分对抗性扰动和干净的样本，通过惩罚对抗性节点或连边上的权重来训练出鲁棒的图神经网络模型。

2.4.2 面向图神经网络的鲁棒性分析

图神经网络的鲁棒性分析同样是攻防研究的重点方向，可以理解为图神经网络的鲁棒性分析是图神经网络防御研究的一个方向。与其他防御方法不同，面向图神经网络的鲁棒性分析在防御对抗攻击的同时，对模型或数据的安全性进行评估。这种评估可以衡量不同网络结构和图深度模型对对抗扰动的敏感性，在网络结构重要性分析和图深度模型的可解释性研究中均具有关键性作用。

3 面向图神经网络的对抗攻击方法

上文对图神经网络攻击方法从不同几个方面进行了分类，并给出了对应的原理分析，为了加深对图神经网络攻击的理解，本节总结图神经网络攻击的相关研究，并详细介绍现有图神经网络攻击技术，对算法中一些和图神经网络攻击相关性较弱的部分只进行简单的探讨。不同攻击方法的侧重点不同，根据攻击所针对的任务类型，可以将这些方法分为：针对节点分类的攻击方法、针对图分类的攻击方法、针对链路预测的攻击方法和针对社区检测的攻击方法。需要指出的是，图神经网络攻击方法可以在不同的分类标准上做出多种分类结果，由于同种攻击目标任务的攻击方法在攻击结果上有更高的相似性，本节以不同的攻击目标任务作为分类标准，对现有的图神经网络攻击方法进行回顾。表2是对图神经网络攻击方法的归纳，为保证行文的流畅性，本节的论述主要是按时间顺序组织的^[67]。

3.1 节点分类攻击

3.1.1 NETTACK

针对节点分类任务，Zügner等^[28]首次提出在图上进行对抗性攻击，并提出了NETTACK算法，迭代生成对抗性网络，以此对GCN^[68]进行攻击。为了确保对抗扰动不明显，本节从3个角度分析扰动大小限制问题。

(1) 修改代价：为了确保攻击者不能完全修改图数据，算法首先通过一个代价限制允许的变化数量，限制节点属性和邻接矩阵的改变量的总和不能大于一个上限。

$$\sum_u \sum_i |X_{ui} - \hat{X}_{ui}| + \sum_{u < v} |A_{uv} - \hat{A}_{uv}| \leq \Delta \quad (16)$$

其中， X_{ui} / A_{uv} ($\hat{X}_{ui} / \hat{A}_{uv}$) 分别表示攻击前(后)节点 u 的第 i 个属性和节点 u, v 的连边关系。该式将图数据攻击前后的节点特征与图结构变化数量限制在预算 Δ 内。

(2) 图结构特征：NETTACK 引用幂律分布的统计双样本检验，使用似然比检验来估计 G 与 $\hat{G}$ 的度分布是否来自相同的分布，在图结构的度分布上确保扰动的隐蔽性。

(3) 节点属性特征：NETTACK 对来自 G 节点特征的共现图 $F = (\mathcal{F}, \mathcal{E})$ 使用随机游走得到从最初为节点 u 提供的特性开始到游走到下一个节点特征 i 的概率，并根据该概率大小来确定节点特征上扰动的大小。

在对抗攻击阶段，给定当前图的状态 $G^{(t)}$ 和符合图结构幂律分布的候选集 C_{struct} 和符合节点特征共现性的特征修改集 C_{feat} ，计算修改图结构或节点特征的得分函数，选择得分最高的操作将图更新为 $G^{(t+1)}$ ，重复进行直至超出修改代价。

3.1.2 RL-S2V

Dai等^[29]提出一种基于强化学习的攻击方法(RL-S2V)，学习可推广的攻击策略，同时只需要目标分类器的预测类标。具体而言，将对抗攻击算法的决策过程看作一个有限层的马尔可夫决策过程MDP，并使用Q-learning来学习这个决策过程。RL-S2V对时刻 t 的状态 $(\hat{G}_t, v_t)$ 添加或删除一条连边，计算目标分类器的预测结果，若目标分类器的预测结果跟类标不一致，返回值为1的正反馈，反之返回值为-1的负反馈，当修改连边数目达到预设值停止攻击。RL-S2V将原来直接选取一条连边进行修改的操作分解为先后选取这条边的两个节点，通过两个为节点评分的 Q 函数 Q^1 与 Q^2 ，先后选取两个使最后反馈值最大的节点对进行连边攻击。

RL-S2V适用于黑盒攻击，而在其他能获得更多信息的场景下，本文提出基于遗传算法的GeneticAlg与基于梯度的白盒攻击GradArgmax。图3与图4展示以上两种攻击的攻击步骤。GeneticAlg将不同的对抗样本组成种群，通过适应性选择、交叉和变异完成进化，其中，适应性分数由目标模型的损失函数计算得到。GradArgmax根据邻接矩阵 A 对损失函数 $\mathcal{L}$ 的导数绝对值大小选择目标连边，再根据其符号增/删该连边。

表 2 对抗攻击算法分类
Table 2 The classification of adversarial attack algorithm

任务分类	攻击方法	攻击背景	扰动类型	原理	攻击目标
节点分类	NETTACK	灰盒	修改连边/修改特征	基于数据特征修改连边/特征	GCN/CLN/DeepWalk
	RL-S2V	黑盒	修改连边	基于强化学习修改连边	GNN
	FGA	白盒	修改连边	根据模型梯度信息修改连边	GCN/Deepwalk/ Node2vec/LINE
	Greedy-GAN	白盒	添加节点	利用 GAN ^[67] 和贪心算法生成虚假节点	GCN
	$\mathcal{A}_{DW}$	黑盒	修改连边	利用矩阵摄动理论进行连边翻转	GCN/Node2vec/ DeepWalk
	Metattack	灰盒	修改连边	通过元梯度来构造攻击	GNN/CLN/DeepWalk
	Manipulating	白盒	修改连边	基于图的优化问题修改连边	DeepWalk/GCN/ LINE/LBP/JW/LinBP
	DAGAER	白盒	修改特征	基于编码器-解码器框架修改特征	GCN
	IG-FGSM/JSMA	白盒	修改连边/修改特征	根据模型梯度信息修改连边与特征	GCN
	梯度投影攻击	白盒	修改连边	根据模型梯度信息修改连边	GCN
	GF-Attack	黑盒	修改连边	基于近似的图滤波器和属性构造攻击	Network Embedding
	ReWatt	黑盒	重布线	基于强化学习, 重新布线	GCN
	NIPA	灰盒	添加节点	基于强化学习向网络中添加虚假节点	GCN
	GUA	白盒	修改连边	利用梯度迭代寻找攻击向量翻转连边	GCN/DeepWalk/ Node2vec/GAT
	POISONPROBE	白盒	修改特征	基于梯度, 攻击中毒效率分数最高的非直接邻居的节点特征	GCN
	MGA	白盒	重布线	在基于梯度的迭代过程中加入动量项	GCN/DeepWalk/ Node2vec
链路预测	Triads Attack	白盒	修改连边	基于图结构的邻居评分与启发式算法增加或删除连边, 降低连边相似性	similarity metrics
	IGA	白盒	修改连边	根据模型梯度信息修改连边	GAE
	Opt-attack	白盒	修改连边	基于一阶 KKT 条件和投影梯度下降构造攻击	Network Embedding
	Approx-Local	白盒	修改连边	删除有限子集的连边, 以最小化一组目标链路的总加权相似度得分	Local/Global similarity metrics
	TGA	白盒	重布线	根据模型梯度信息重布线	Deep dynamic network embedding model
图分类	RL-S2V	黑盒	修改连边	基于强化学习修改连边	GNN
	多层次图池化神经网络攻击	白盒	修改特征/修改连边	基于梯度攻击图池化层保留的节点	SAG/ HGP-SL
	PA 后门攻击	黑盒	修改连边	生成随机子图进行后门攻击	GIN
	GTA	黑盒	修改特征/修改连边	为每个图生成自适应的子图进行后门攻击	GCN/GraphSAGE/GAT
社区检测	FGA	白盒	修改连边	根据模型梯度信息修改连边	GCN/Deepwalk/ Node2vec/LINE
	Q-Attack	白盒	重布线	基于遗传算法, 重新布线	Community Detector
	EPA	白盒	修改连边	基于遗传算法修改连边	Community Metrics
	EDA	白盒	修改连边	基于遗传算法 (GA) 的欧几里得距离攻击策略 (EDA) 修改连边	HOPE/LPA/EM/ DeepWalk
	MGA	白盒	重布线	在基于梯度的迭代过程中加入动量项	Community Detector
	CD-ATTACK	黑盒	修改连边	利用图自编码作为生成器生成不明显的扰动, 代理算法对扰动加以限制	Surrogate community detection model
	DICE	白盒	重布线	利用启发式算法, 删除部分连边, 随后通过添加连边恢复其影响力	Community Detector

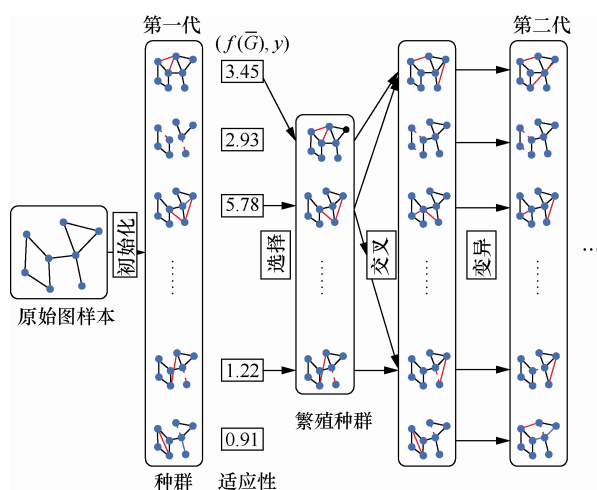


图3 GeneticAlg 示意

Figure 3 Illustration of GeneticAlg

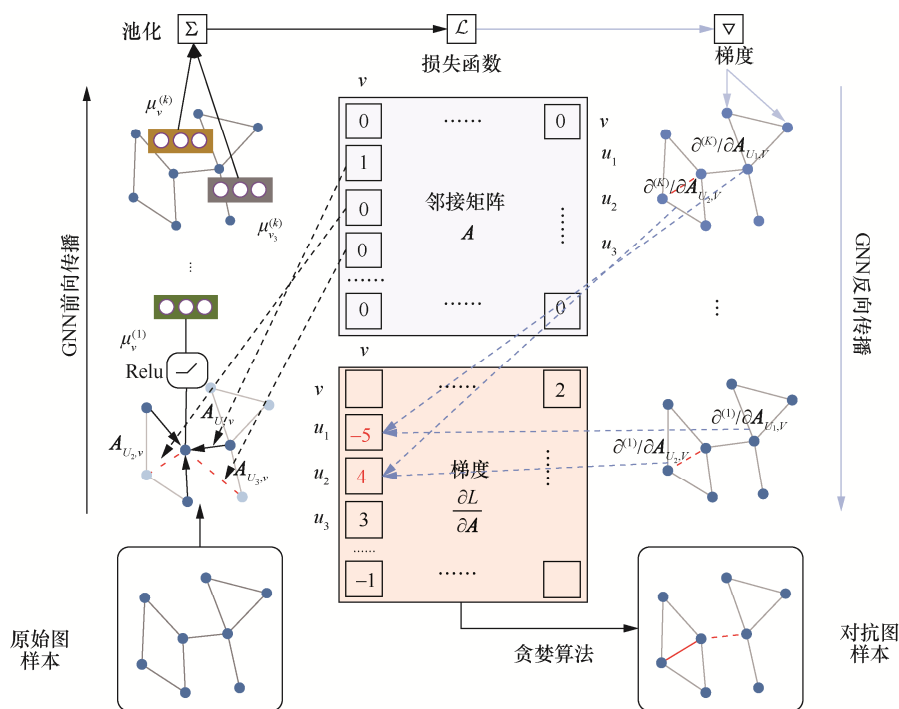


图4 GradArgmax 示意

Figure 4 Illustration of GradArgmax

3.1.3 FGA

Chen 等^[34]提出的快速梯度攻击（FGA）在训练好的 GCN 模型上，根据邻接矩阵元素计算目标损失函数 $\mathcal{L}_t$ 的偏导数 $\partial \mathcal{L}_t / \partial A_{ij}$ 作为连接梯度网络（LGN） $\hat{g}$ 。每次迭代过程中，FGA 利用对抗网络的邻接矩阵 $\hat{A}^{h-1}$ 生成第 $h-1$ 个 LGN，即 $\hat{g}^{h-1}$ 。接着基于 $\hat{g}^{h-1}$ 选择一对节点 (u, v) 的最大连接梯度 $\hat{g}_{u,v}^{h-1}$ ，根据其正/负梯度对节点 (u, v) 添加/删除连接，当连

接修改数量达到一定阈值时结束攻击。FGA 对多种常见的网络嵌入方法进行了普通节点、中心节点、桥节点的节点分类攻击，其中包括对目标节点的直接攻击、间接攻击与无限制攻击。大量的实验验证了算法的攻击有效性以及可迁移性。

3.1.4 Greedy-GAN

Wang 等^[30]指出，在很多情景下，攻击现有的网络结构或特征矩阵是不现实的。他们通过恶

意添加虚假节点，并将虚假节点连接到已有节点实现对图数据的攻击。针对假节点邻接矩阵和特征矩阵设计中的离散输入空间问题，本文提出了一种贪婪算法来生成恶意节点的连边及其对应特征，以最小化目标节点的分类精度。为了确保扰动的隐蔽性，Greedy-GAN 通过添加一个判别器，利用判别器与攻击者的迭代训练确保虚假节点与真实节点的相似性。

3.1.5 $\mathcal{A}_{DW}$

Bojchevski 等^[32]首次对广泛使用的基于随机游动的方法进行了对抗性脆弱性分析，推导出了有效的对抗性扰动。基于 RW (rand walk) 的网络嵌入方法的中间过程是个随机采样的过程，作者使用了奇异值分解 (SVD) 得到基于 RW 的优化嵌入矩阵 $\mathbf{Z}^*$ ，利用 $\mathbf{Z}^*$ 得到 Deepwalk 的损失函数 $\mathcal{L}_{DW_i}(\mathbf{A}, \mathbf{Z}^*)$ ，以此避免随机游走所产生的随机性。

由于 SVD 和优化问题的离散型，最大化 $\mathcal{L}_{DW_i}$ 仍然具有挑战性。对此提出两种方法：① 将对特征值的扰动看作取损失函数 $\mathcal{L}$ 的梯度，近似提出 $\mathcal{L}_{DW_2}$ ；② 通过分析一些中间矩阵的频谱变化，推导出共现矩阵频谱变化的边界，得到损失函数 $\mathcal{L}$ 的估计 $\mathcal{L}_{DW_3}$ 。这两种方法在相对较小的扰动和限制下，成功毒害图数据，同时此类攻击可以在不同的模型之间进行转移。

3.1.6 Metattack

Zügner 等^[40]研究了针对节点分类的图神经网络，扰乱离散图结构的训练中毒问题。文章提出了首个破坏全局节点分类性能的中毒攻击 Metattack，并将其表示为一个二次优化问题。

$$\begin{aligned} \min_{\hat{G} \in \Phi(G)} \mathcal{L}_{\text{atk}}(f_{\theta^*}(\hat{G})) \\ \text{s.t. } \theta^* = \arg \min_{\theta} \mathcal{L}_{\text{train}}(f_{\theta}(\hat{G})) \end{aligned} \quad (17)$$

其中， $\mathcal{L}_{\text{atk}}$ 表示攻击者要优化的损失函数， $\mathcal{L}_{\text{train}}$ 表示分类模型的训练损失函数， $\hat{G}$ 表示符合限制条件 $\Phi(\cdot)$ 的中毒样本。由于测试数据类标不可用导致 $\mathcal{L}_{\text{atk}}$ 无法直接优化，提出两种方法降低模型在未标记节点上的泛化性能：① 最大化训练节点的损失；② 利用由标记节点训练的模型预测未标记节点的类标，最大化未标记节点的损失。

Metattack 的核心思想是使用元梯度来解决上述问题，本质上是将图结构作为超参数，构造元梯度来优化。

$$\nabla_G^{\text{meta}} := \nabla_G \mathcal{L}_{\text{atk}}(f_{\theta^*}(G)) \quad \text{s.t. } \theta^* = \text{opt}_{\theta}(\mathcal{L}_{\text{train}}(f_{\theta}(G))) \quad (18)$$

其中， $\text{opt}(\cdot)$ 是一个可微优化过程（如梯度下降及其随机变量），攻击者利用元梯度构造评分函数，得到图数据的微小扰动对攻击损失 $\mathcal{L}_{\text{atk}}$ 的影响，最终在 NETTACK 中的攻击约束下，根据评分函数选择对连边与节点特征的修改。

3.1.7 控制图结构的图分类攻击

Wang 等^[69]研究了基于协作分类的节点分类攻击，并证明其在基于图神经网络的节点分类模型上具有良好的攻击效果。他们将攻击描述为一个基于图的优化问题，为解决图数据的离散型给该优化问题带来的挑战，采用两种解决方式：① 将连边情况松弛为一个取值区间为 $[0,1]$ 的连续变量，在解决优化问题后将其转为二进制；② 为了减小攻击预算，将优化空间缩小到目标节点与其余节点之间的连边以及目标节点之间的连边。

3.1.8 DAGAER

Avishek 等^[70]将对抗性攻击视为一个生成模型问题，他们提出一种统一的编码器-解码器框架 (DAGAER) 来生成对抗性样本的完整分布，能够有效地为单个给定的输入生成一组不同的攻击。

DAGAER 由 4 个部分组成：概率编码器网络、概率解码器网络、组合函数和定义在 $\mathcal{X}$ 上的相似度函数 Δ 。生成网络的基本流程如图 5 所示，给定输入 $x \in \mathcal{X}$ ，首先使用图卷积网络 (GCN) 作为概率编码器，采样得到一个潜在“扰动编码” $z \in \mathbb{R}^d$ ，之后鉴于这个采样编码，利用多层感知器 (MLP) 作为解码器生成一个扰动 $\delta = p_{\theta}(z)$ ，然后使用组合函数结合扰动与原始输入，生成对抗样本 $x' = c(\delta, x)$ 。

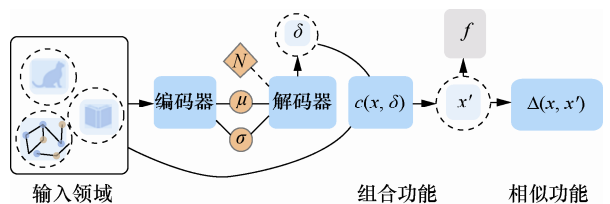


图 5 DAGAER 的主要组成部分与上对抗性例子
Figure 5 The main components of DAGAER and the forward generation of an adversarial example

3.1.9 IG-FGSM/JSMA

Wu 等^[71]证明通过引入综合梯度可以很容易地解决图数据攻击中的离散性问题，该梯度可以准确地反映扰动某些特征或连边的效果，同时可充分利用并行计算提高效率。他们提出两种梯度攻击方法 IG-FGSM 与 IG-JSMA，前者基于目标模型的损失函数获得梯度，而后者根据 GCN 模型输出的预测结果获得梯度。不同于 FGA 的梯度攻击，该方法考虑到对节点属性的攻击，根据梯度计算连边与节点属性的重要性，选择更重要的修改项（连边或属性），进一步实现攻击操作。

3.1.10 梯度投影攻击

Xu 等^[72]提出了一种新的基于梯度的攻击方法，使处理离散图数据变得更加容易。他们提出一阶攻击生成框架的两种攻击场景：① 攻击一个预定义的 GNN；② 攻击一个可再训练的 GNN。针对这两种场景，利用一阶优化，提出两种新的拓扑攻击：投影梯度下降（PGD）拓扑攻击和最小最大（Min-max）拓扑攻击。

针对第一种场景的攻击，类似于文献[69]等，PGD 同样先将连边情况松弛为一个取值区间在 $[0,1]$ 的连续变量，在解决优化问题后将其转为二进制。而对第二种可再训练的模型参数的场景，攻击生成问题变成最小最大化的问题，外部最小化原始数据的损失由 PGD 解决，内部最大化生成对抗数据的损失由梯度上升来解决。

3.1.11 GF-Attack

Chang 等^[38]指出由于实际系统中模型预测的不可达性，使白盒攻击在实际的图数据学习系统中难以实现，提出了一种更通用、更灵活的基于黑盒驱动的攻击各种图嵌入模型的框架。GF-Attack 将特征矩阵看作具有通道 l 的图信号，并将图嵌入模型看作根据图滤波器和特征变换生成新的图信号。

$$\tilde{X} = \mathcal{H}(X), X' = \sigma(\tilde{X}\theta) \quad (19)$$

其中， $\mathcal{H}$ 是图信号滤波器，通常可以构建为一个关于图平移滤波器 S 的多项式函数 $\mathcal{H} = h(S)$ ，图平移滤波器 S 反映了图的一些局部特性。GF-Attack 是由图滤波器和特征矩阵构造的，在优化过程中，利用低秩近似与奇异值分解对图信号

滤波器 $h(S)$ 进行处理，降低计算复杂度。

GF-Attack 通过构造对应的图滤波器来描述图嵌入模型，以黑箱攻击方式仅对图过滤器进行攻击，改变节点的特征值，进而影响网络嵌入方法的输出结果。

3.1.12 ReWatt

Ma 等^[39]指出，图神经网络的攻击通常是通过添加或删除一些边来实现的，即使修改的边的数量很少，也可能会引起注意，因此提出了一个图形重布线的操作（ReWatt）实现对图神经网络的攻击，它对图的影响相对不明显。ReWatt 使用强化学习来学习基于重布线操作的攻击策略。这种策略的一个明显优点是它不会改变图的节点数、边数和总度。然而，单一的添加或删除操作可能会改变这些属性。

3.1.13 NIPA

类似于 Greedy-GAN，Sun 等^[33]提出的基于强化学习的方法 NIPA 也将目光放在对现有图中注入虚假节点上。NIPA 基于层次 Q 学习，其关键思想在于训练深度强化学习代理，迭代地执行破坏图的操作，包括添加对抗性边和修改注入节点的类标。代理需要首先从注入节点集 V_A 中选择一个节点，并从中毒节点集 V' 中选择另一个节点，在二者间添加敌对的连边，修改注入节点的节点类标，实现对分类器 f 的攻击。

3.1.14 GUA

Zang 等^[73]定义了锚节点，这类节点通过翻转任何目标受害者的连边，可以破坏一个训练好的图神经网络。他们提出 GUA 的算法来识别锚节点并进行攻击。GUA 引入了新的攻击向量概念，节点数目为 n 的图攻击向量 $\mathbf{p}$ 为一个长度 n 的二进制向量，锚节点为 1，否则为 0。把向量 $\mathbf{p}$ 推广到一个 $n \times n$ 矩阵 $\mathbf{P}$ ， $\mathbf{P}$ 的第 i 行和第 i 列与 $\mathbf{p}$ 相同，其他地方为 0。以攻击矩阵 $\mathbf{P}$ 的 (i, j) 元素表示节点对 (i, j) 的连接是否翻转，当值为 1 时翻转。GUA 在攻击过程中，将攻击向量 $\mathbf{P}$ 的二进制元素松弛到 $[0,1]$ 的连续值，值的大小表示变化的强度，迭代访问每个训练节点 i ，通过 DeepFool 计算目标节点 i 到另一个类决策边界的最小扰动，并遵循最小扰动，利用梯度方法寻找的攻击向量 $\mathbf{p}$ ，实现攻击。

3.1.15 POISONPROBE

Takahashi 等^[74]指出,在图卷积分类器中,一个节点的分类结果可能会受到它邻居节点的影响。他们提出 POISONPROBE,如图 6 所示,该方法攻击目标节点两跳甚至多跳邻居节点的节点特征,实现对目标节点的有效攻击。

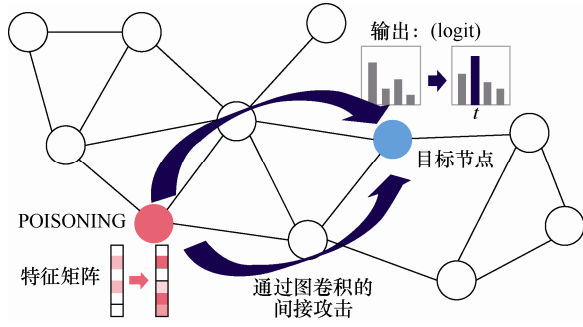


图 6 通过毒害单一节点的非直接对抗攻击
Figure 6 Indirect adversarial attack by poisoning a single node

POISONPROBE 首先给出了中毒节点的选择方法:定义中毒效率分数,类似于将中毒信息从候选者传递到目标路径的带宽,如果分数很小,中毒信息将通过路径缩小,限制扰动的大小。POISONPROBE 算法由外环和内环两部分组成。内环发现较小的扰动,在几个参数固定的情况下可以达到预期的目标攻击。然后,外环利用二叉搜索迭代地发现参数。如果内部循环成功找到对抗的扰动,它将降低内环中初始常数 λ 。

3.1.16 MGA

Chen 等^[75]发现,基于梯度信息的对抗攻击方法能通过重布线降低基于深度学习模型的图嵌入算法的有效性,但也容易陷入局部最优。他们针对 GCN 模型提出了一种动量梯度攻击 (MGA) 方法,该方法在 FGA 中的连边梯度网络 (LGN) 上做了进一步扩展,提出了带有衰减系数的连边动量网络 (LMN),通过在梯度方向积累速度矢量更新第 k 个 LMN。该方法能够以较少的重布线连边实现更强的攻击。与直接利用梯度信息对原始网络进行更新相比,在迭代过程中加入动量项可以稳定更新方向,使模型跳出较差的局部最优,增强了方法的可移植性。

3.2 链路预测攻击

3.2.1 Triads Attack

Wanick 等^[76]指出链路预测攻击问题本质上

是一个优化问题,链路预测问题涉及一个“探索者”,它根据相似度指数 (ROC 曲线下面积 AUC、平均精度 AP) 对非连边进行排序,高排名的连边可能在未来形成连边。他们提出了两种启发式算法:① 删除连边的封闭三元组删除 CTR;② 添加连边的 OTC。CTR、OTC 的目的为重新连接网络,最小化那些原本相似度排名高的连边。如图 7 所示的 CTR 方法,网络中包含 3 个三元组 (x, v, w) 、 (y, x, v) 和 (z, v, w) ,通过删除三元组中的 (v, w) , (x, w) 、 (y, w) 和 (z, w) 的相似性得分将会减少。图 8 展示了 OTC 的攻击方法, (v, w) 的添加创建了两个开放的三元组:一个包含节点 (x, v, w) , 另一个包含 (y, x, v) , 因此, (x, w) 和 (y, v) 的相似度得分增加,而 (w, u) 的相似度得分减少。

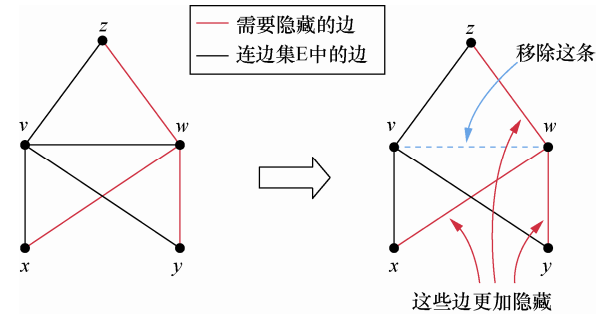


图 7 CTR 的主要思想说明
Figure 7 Illustration of the main idea of the CTR

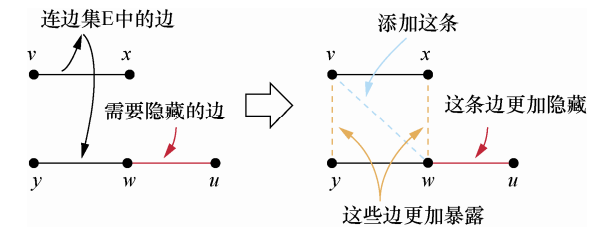


图 8 OTC 的主要思想说明
Figure 8 Illustration of the main idea of the OTC

3.2.2 IGA

针对链路预测任务,Chen 等^[35]提出了一种基于已训练的图形自动编码器 (GAE) 中梯度信息的迭代梯度攻击 (IGA)。IGA 针对现实网络的复杂性和网络攻击的局限性,提出了无限攻击和单节点攻击两种对抗攻击策略,对 GAE 的损失函数计算得到损失函数关于连边的梯度矩阵,根据梯度矩阵对应位置元素的正负值与绝对值选择需要修改的连边来实现攻击。

3.2.3 Opt-Attack

Sun 等^[41]介绍了一种数据中毒攻击的节点嵌入方法, 针对两种流行的无监督节点嵌入方法 (DeepWalk 和 LINE) 提出一种基于投影梯度下降攻击无监督节点嵌入算法的高效算法 Opt-attack。他们指出, DeepWalk 和 LINE 隐式地进行矩阵分解^[77], 可以根据一阶卡罗需-库恩-塔克条件 (KKT, Karush Kuhn Tucker) 条件推导出反向传播梯度。Opt-attack 有两个步骤。① 投影梯度下降 (PGD) 步骤: 将图视为一个加权图, 因而邻接矩阵是连续的, 对加权邻接矩阵进行梯度下降。② 投影步骤: 将加权邻接矩阵投影到 $\{0,1\}^{p \times p}$ 上, 对接近 1 对应的邻接矩阵元素置 1 以实现增边, 反之将接近 0 对应元素置 0 实现删边。

3.2.4 Approx-Local

Zhou 等^[31]系统地研究攻击者操纵链路预测的能力, 通过删除有限数量观察到的连边子集, 以最小化一组目标链路的总加权相似度得分; 重点研究了相似性度量的两个重要子类: 仅利用目标链路的局部信息的局部度量和使用全局网络信息的全局度量。

对于局部度量, 他们提出了一种最小化局部度量上界的算法, 对应于在基数约束下最大化子模函数, 确定了两种特殊情况: ① 攻击单个连边, 确保对所有局部度量的最佳攻击; ② 攻击一组节点而第二种情况确保对共同邻居度 (CND, common neighbor degree) 指标的最佳攻击。对于全局度量, 证明即使在攻击单个连边时, 最小化 Katz 相似度和最大化 ACT 全局相似性度量都是 NP 困难的。针对这两个问题, 他们分别提出了有效的贪婪算法 (Greedy-Katz) 和启发式算法 ((Local-ACT))。

3.2.5 TGA

Chen 等^[78]指出, 先前的工作没有考虑到真实世界系统的动态特性。他们提出第一个动态网络链路预测 (DNLP) 的对抗攻击方法——时间感知梯度攻击 (TGA)。TGA 利用跨快照的深度动态网络 (DDNE) 嵌入生成的梯度信息, 计算目标损失函数 L_i 对第 i 个网络快照 $S(i,:)$ 的导数 $g_i(i, j)$, 在一个迭代中, 首先根据每个快照中节点对 (i, j) 的导数绝对值 $|g_{ij}|$ 对连边进行排序, 随

后在每个快照中选择排序最靠前的连边, 根据 g_{ij} 的正/负值对连边进行增加/删除操作, 使 DDNE 无法对目标链路进行准确预测。

3.3 图分类攻击

图分类常见于对生物结构的分类任务, 通过学习图整体结构的低维嵌入对图整体进行分类。针对图分类的攻击, 旨在使图分类器分类错误, 从而降低图分类模型的准确率和召回率。与节点分类攻击相比, 针对图分类模型的攻击研究相对较少。

3.3.1 RL-S2V

在 3.1.2 节中, Dai 等^[29]除了对节点分类任务进行了攻击。在图分类任务中, RL-S2V 利用图分类模型得到马尔可夫决策过程 (MDP) 中的奖励值, GeneticAlg 和 GradArgmax 通过图分类模型的损失函数计算适应性分数与梯度。

3.3.2 多层次图池化神经网络的对抗性攻击

Tang 等^[79]探讨了多层次图池化 (HGP) 神经网络的脆弱性, 提出了对基于多层次 GCN 的图分类模型的对抗攻击。多层次 GCN 图分类模型中的多层次池化算子倾向于保留对图分类更重要的节点, 作者利用一个卷积算子和池化算子构建代理模型, 并将其池化算子保留的节点设置为攻击目标, 基于梯度方法添加/删除连边或修改节点特征以产生对抗样本, 从而欺骗多层 GCN 中的池化算子, 使其保留错误节点。在扰动的隐蔽性层面, 作者通过编辑距离^[80]与二阶 DELTACON0 图距离^[81]限制在图结构上的扰动, 并以 L1 范数对节点特征扰动加以限制。

3.3.3 PA 后门攻击

Zhang 等^[82]针对基于 GNN 的图分类提出了一种基于子图的黑盒后门攻击, 生成随机子图作为后门触发器, 具体使用 4 个参数设定后门攻击: ① 设置子图节点数目作为触发器大小; ② 子图中连边总数与节点对数的比值为触发器密度; ③ 采用优先依附模型生成具有与真实图相似幂律分布的后门触发器; ④ 以测试图被毒化的分数作为中毒强度。一旦攻击者向测试图注入一个预定义的后门触发器, GNN 分类器就会预测攻击者为测试图选择的目标类标。

3.3.4 GTA

Xi 等^[83]对基于 GNN 的图分类提出了另一种

黑盒后门攻击, 相比于 Zhang 等^[82]的工作, 他们将触发器定义为特定的子图, 包括拓扑结构和节点/连边特征, 并动态地为不同的图自适应设计后门触发器。具体来说, GTA 在不知道下游分类器的情况下, 利用 GNN 模型 f_θ 的损失作为中间表示, 交替优化模型参数 θ 与后门触发子图 g_i , 随后采用基于回溯的算法 VF2^[84], 在给定图 G 中寻找与 g_i 相似的子图 g 。根据节点特征与图拓扑结构, 基于注意力机制提取 g 的特征, 并以此特征调整 g_i , 最后使用 g_i 替换 g 实现高效的后门攻击。

3.4 社区检测攻击

3.4.1 Q-Attack

Chen 等^[36]为了更有效地攻击社区检测任务, 提出了两种启发式攻击策略: ① 通过结合特定的社区检测算法和社区划分结果提出了社区检测攻击算法 (CDA), 根据社区划分结果, 断开目标节点与社区内邻居节点的连边, 添加与社区间非邻居节点的连边; ② 基于度的攻击 (DBA), 选择网络中度值较高的节点实现 CDA 攻击。在此基础上, 提出基于遗传算法的 Q-Attack 方法, 利用模块性度量值 Q 设计适应性函数 $f = e^{-Q}$, 低模块性表明个体具有更大的适应度。选择重新布线攻击策略来维持目标节点的度, 在删除目标节点的连边的同时, 为其增加指向目标节点的连边, 保持网络攻击的隐蔽性。

3.4.2 EPA

Chen 等^[37]首次将社区检测攻击问题化为攻击整个网络、攻击特定社区和攻击特定节点 3 个尺度, 将这些攻击视为一个优化问题, 提出了一种基于遗传算法的新的进化扰动攻击 (EPA)。他们将改变的节点连边集定义为染色体, 重连的连

边 ID 作为染色体上的基因, 使遗传算法更适合寻找合适的攻击目标并有效减少种群的存储空间。考虑到重新连接的连边数量也是一个变量, EPA 采用了一种不相等交叉的策略。如图 9 所示, 初始染色体 O_i 和 O_j 从各自的增边差集 A_i^E / A_j^E 中根据交叉数量 r_i / r_j 随机选择增/删 (图中的方块/三角) 操作, 交叉得到新的染色体 ($\bar{O}_i$ 和 $\bar{O}_j$)。在突变过程中引入一种基于网络结构特性的搜索机制 (包括成对节点间的节点介数和最短路径长度) 来加快遗传算法的收敛速度, 更快地获得近似最优解。

3.4.3 EDA

Yu 等^[85]提出了一种新的无监督网络嵌入攻击策略 EDA, 以嵌入向量间的欧几里得距离为参考, 采用遗传算法 (GA) 寻找 EDA 的最优解。EDA 算法由 3 部分组成: 网络编码、适应度选择、交叉和变异操作。① 选择翻转后的连边作为基因, 基因的组合作为个体, 表示对抗性扰动的不同解, 不同的个体组成一个种群; ② 通过攻击获取嵌入空间中向量欧几里得距离的相对变化, 并通过所提出的自适应度函数计算个体的适应度; ③ 选择适应度较高的个体作为亲本, 通过交叉和变异操作产生新的个体, 最终实现对图数据的攻击。

3.4.4 CD-Attack

Li 等^[86]专注于对社区检测问题的黑盒攻击, 旨在隐藏目标个体, 提出一种新的图生成神经网络 CD-ATTACK, 通过攻击代理模型生成满足离散约束条件的对抗样本。CD-ATTACK 由两个主要模块的交替训练完成优化: ① 一个敌对的攻击者 $g(\cdot)$, 以图自编码器 VGAE 为基础, 在一定的扰动限制下生成对抗样本; ② 一个 GCN 与全连

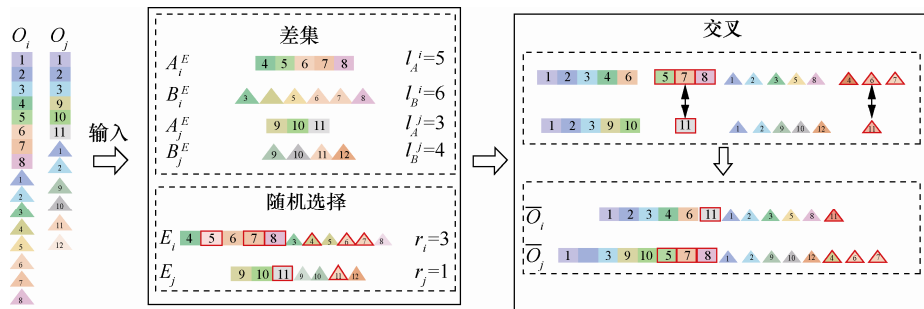


图 9 不相等交叉示意

Figure 9 Illustration of non-equal crossover operation

接网络组成的代理社区检测算法 $f(\cdot)$ ，以无监督的方式划分个体的社区并测量图的不相似度。这两个模块是高度相关的，敌对攻击者 $g(\cdot)$ 需要依赖 $f(\cdot)$ 反馈来检查是否成功隐藏个体，社区检测算法 $f(\cdot)$ 依赖于对抗样本来增强其鲁棒性。

3.4.5 DICE

Wanick 等^[87]研究如何在个体不影响其网络影响力的情况下，规避网络中心性分析以保护隐私，提出一种启发式算法 DICE。文章首先定义度中心性、接近度中心性、中介中心性 3 个中心性度量，并以独立级联模型、线性阈值模型建立的影响模型，通过减少 $v^\dagger$ 的中心性而不影响其对网络的影响（根据上述的影响模型）来隐藏 $v^\dagger$ 的重要性。给定一个网络和一个源节点 $v^\dagger$ ，DICE 在不超过最大修改连边数 K 的情况下，通过两个阶段重新连接网络实现攻击：① 将攻击代价 K 的一部分 b 用于最小化上述中心性指标，在此期间， $v^\dagger$ 的影响可能会减小，表示为中心性最小化问题；② 花费剩余的代价 $K-b$ 来尽可能地恢复 $v^\dagger$ 的影响，同时避免添加在中心性最小化阶段中删除的任何连边。

4 面向图神经网络的防御方法

上文以不同攻击目标任务为分类标准，详细介绍了现有图神经网络攻击技术，本节以不同的防御策略作为分类标准，将现有的面向图神经网络防御方法分为：对抗性训练、对抗性扰动检测、启发式防御、图纯化防御和注意力机制。本节按照上述分类标准对现有图神经网络防御方法进一步展开，并详细介绍部分面向图神经网络的鲁棒性分析方法。一些经典的防御方法如表 3 所示。

4.1 对抗性训练

4.1.1 BVAT

Deng 等^[88]指出，在面对输入周围的局部扰动时，GCN 没有考虑模型输出分布的平滑性，如图 10(a)所示，他们定义了节点 u 的接受域 RF_u （红色节点）， u 上的虚拟对抗扰动将改变接受域 RF_u 中所有节点的嵌入特征；提出基于样本的 S-BVAT 和基于优化的 O-BVAT 两种算法，利用批量梯度下降法训练 BVAT。为了限制对抗性扰动的传播以防止不同节点之间的相互干扰，如

图 10(b)所示，S-BVAT 从两个彼此不重叠接受域中采样节点 u, v ，对节点子集 $V_s \subset V$ ， $\forall u, v \in V_s$ 产生不会互相作用的虚拟对抗性扰动。而在 O-BVAT 中，如图 10(c)所示，利用优化过程为节点集 V 中的所有节点生成虚拟的对抗性扰动。这两种算法分别产生不同范围的虚拟对抗性扰动进行对抗训练，提高 GCN 分类器的平滑性。

4.1.2 GCN-GATV

Feng 等^[46]提出图对抗训练（GAT）的方法以及 GAT 的扩展 GATV。GAT 的每次迭代可以看作在进行一个极大极小的博弈：① 从干净样本中生成图的对抗性扰动，最大限度地破坏连接节点之间的平滑性；② 增加一个正则化器，通过激励对抗样本和对应干净样本预测结果之间的平滑性来最小化图神经网络在图对抗样本上的目标函数，提升模型对于通过图传播扰动的鲁棒性。

与 GAT 相比，GATV 可以看作 3 个玩家共同玩两个极大极小游戏，在训练目标函数中额外加入一个虚拟对抗正则化器。当图对抗性扰动是输入特征变化的方向时，可以最大限度地攻击图对抗性正则化器，通过在当前的模型参数下最大化图对抗正则化器来生成虚拟对抗样本，平滑每个干净样本周围的预测分布，进一步增强模型的鲁棒性。

4.1.3 平滑防御

Chen 等^[47]针对全局和目标的对抗攻击提出了不同的对抗训练策略 Global-AT 和 Target-AT，不断地用所有选定的目标对节点更新全局/目标对抗网络 G_{adv} ，利用对抗训练来提高 GNN 的鲁棒性。

文献[47]还提出两种特殊的平滑防御策略：

- ① 平滑蒸馏（SD）以温度为 T 的 softmax 层来训练初始的 GCN，使用初始训练的 GCN 模型的输出置信度作为软类标对训练节点进行编码，利用该编码对蒸馏 GCN 进行训练，训练目标函数由软损失函数和原始损失函数组成；
- ② 平滑交叉熵损失函数（SCEL）鼓励 GCN 对真实的类标返回一个高置信度分布，而对每个节点的错误类标返回一个平滑的置信度分布以实现梯度隐藏，使攻击者难以通过模型的梯度来进行攻击。

表 3 防御方法分类
Table 3 The classification of defense methods

防御策略	防御方法	任务	原理	方式	目标模型
对抗训练	随机防御	节点分类	在训练期间随机丢弃一些连边	修改图结构	GNN
	BVAT	节点分类	通过对远离彼此的节点子集或所有节点产生虚拟的对抗性扰动来提高 GCN 分类器的平滑性	修改模型	GCN
	GCN-GATV	节点分类	通过最大化图虚拟对抗正则化器生成对抗样本	修改图结构	GCN
	Global/Target-AT/平滑防御	节点分类	针对全局/目标的对抗攻击进行对抗训练, 用平滑蒸馏和平滑交叉熵损失函数实现梯度平滑	修改模型	GNN
	S/DVAT	节点分类	将基于标记和未标记数据的虚拟对抗训练应用在 GCN 的损失函数上, 进行虚拟对抗训练	修改模型	GCN
	GraphDefense	节点分类	利用原始分类器为未类标数据做出预测类标, 使用所有节点进行对抗训练	修改模型	GCN
对抗性 扰动检测	KL 散度检测	节点分类	通过计算节点及其邻域的 softmax 概率之间的 KL 散度的平均值来衡量差异	修改图结构	GCN, GAT
	GraphSAC	异常检测	随机绘制节点子集, 并依赖于图形感知标准, 过滤掉被异常节点污染的集	修改图结构	异常模型
启发式防御	IDOpt/Rank	链路预测	将防御者和攻击者之间的交互建模为一个非零和的贝叶斯 Stackelberg 博弈	修改模型	基于相似度的链路预测模型
图纯化防御	Low-Rank Defense	节点分类	对图进行低秩逼近 (用顶部奇异分量进行重构)	修改图结构	GCN
注意力 机制	PA-GNN	节点分类	限制通过扰动连边的消息传递, 使聚合函数更加专注于真实邻居节点	修改模型	GNN
	RGCN	节点分类	使用高斯分布作为图卷积层中节点的隐藏表示, 并根据它们的方差为邻居节点分配注意力权重	修改模型	GCN
鲁棒性 分析	VPN	节点分类	使用可变功率算子代替 Laplacian 算子	—	GCN
	CRIAGE	链路预测	识别要在知识图中添加或删除的事实, 模型经过重新训练后改变对目标事实的预测	—	知识图嵌入
	AGCN	节点分类	通过抖动 (增加或删除) 连边并选择概率来增强 GCN 的鲁棒性	—	GCN
	具有径向基函数核的 SVM	节点分类	将节点的属性附加到表示节点的欧几里得向量上而获得增强节点特征向量	—	GCN
	可训练邻接矩阵	节点分类	通过在训练过程中学习到选定的连边的权重来充分修改图结构	修改图结构	GCN
	Pro-GNN	节点分类	通过保持图的低秩性、稀疏性和特征光滑性, 迭代地重构干净图并优化 GNN 参数	修改图结构/ 修改模型	GNN

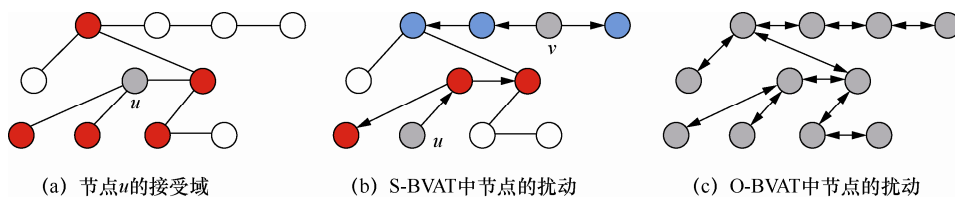


图 10 BVAT 扰动示意
Figure 10 Illustration of perturbations for BVAT

4.1.4 S/DVAT

Sun 等^[48]指出, GCN 中的参数更新仅来自标记节点, 缺乏对未标记数据的利用。作者将基于标记和未标记数据的虚拟对抗训练 (VAT) 应用在 GCN 的损失函数上, 对稀疏特征与稠密特征进行了虚拟对抗扰动, 提出 GCNSVAT 与 GCNDVAT 算法来提高 GCN 的局部平滑性。

他们对 GCN 的局部频谱卷积的一阶逼近进行了重点分析, 并在 GCN 的基本损失函数上添加虚拟对抗损失, 形成了 GCNVAT 算法。在此基础上, 根据节点特征的稀疏特性, GCNSVAT 对每个节点的特定稀疏元素进行虚拟对抗扰动, 而 GCNDVAT 将稀疏特征矩阵转换为稠密特征矩阵来扰动每一个特征元素。实验表明, 经过虚拟对抗训练的 GCN 模型的性能得到了进一步的提升。

4.1.5 Graph Defense

Wang 等^[89]指出, 在对抗训练中, 只有少数节点得到对抗样本的训练, 当测试节点与训练集中的节点较远时, 对抗训练通常是无效的。为此, 他们提出了一种防御对抗扰动的图防御 (Graph Defense) 方法, 该方法可以保持半监督学习设置, 而不需要很大的类标率。Graph Defense 利用原始分类器预测未标记数据的类标 (这个类标可能是正确的, 也可能是错误的)。将标记数据和未标记数据的损失以权重 α 进行组合, 使训练集在对抗性训练中包含所有节点。文章还证明了特征的对抗性训练等价于具有小扰动的边的对抗性训练。实验结果表明, Graph Defense 有效地提高了图卷积网络的鲁棒性。

4.2 对抗性扰动检测

此类方法通过研究对抗性样本和干净样本之间的差异来对对抗性扰动进行检测。

4.2.1 KL 散度差异检测

Zhang 等^[49]通过研究发现, 对图的拓扑扰动比特征扰动更有研究意义, 因为图数据深度学习模型的预测能力取决于图数据中固有的数据关系。鉴于 NETTACK 的原理, 他们提出了一种基于图拓扑结构的攻击检测方法, 通过计算节点及其邻居节点间的 softmax 概率的 KL 散度差异平均值来衡量差异。应用 Neyman-Pearson 引理设置检测阈值, 将原假设分布建模为正态分布, 通过最大似

然法将其拟合到不受扰动的训练数据上, 将尾部概率与目标误报率匹配来确定适当的检测阈值 τ , 并使用逆累积分布函数 (CDF) 进行阈值计算。

4.2.2 GraphSAC

Ioannidis 等^[90]提出的 GraphSAC 随机绘制节点子集, 依赖于图感知标准, 在使用半监督学习 (SSL) 模块估计每个节点的标称类分布之前过滤掉被异常节点污染的集合。由于每次绘制的复杂性与连边的数量呈线性增长关系, GraphSAC 通过限定所需的绘制数量, 实现高效的 SSL。

图 11 展示了 GraphSAC 的运作过程, 第一行表示可用的图和类标。第二行表示采样的类标。第三行的 SSL 模块包含输出预测类标。第四行的绿色虚线表示分类不正确的节点。第五行的过滤器判断损失函数是否包含异常。第六行的红色框表示含有大量错误分类节点的预测被丢弃。最后通过结合来自干净子集的预测结果来检测异常。

4.3 启发式防御

Zhou 等^[91]为了提高基于相似性的链路预测的鲁棒性, 提出一种新的方法来准确地度量被查询连边的存在可能。他们定义分析者 (防御者) 与攻击者两个概念, 将分析者和攻击者之间的交互建模为一个非零和的贝叶斯 Stackelberg 博弈, 分析者首先提交一组可靠的查询, 攻击者在观察分析者的决策后选择要删除的连边集。文章提出了两种防御方法。① 独立损伤优化 IDOpt: 将防御者的问题表示为可利用标准技术线性化的非线性整数规划, 利用其解得到在独立损伤近似下的防御者对称度量最优; ② 独立损害排序 IDRank: 根据累积损害对每个连边的重要性进行排序, 避免了优化问题, 显著提高了可伸缩性。

4.4 图纯化防御

Entezari 等^[50]探索了 NETTACK 扰动的性质, 以寻找有效的防御措施。他们指出 NETTACK 在频谱中存在特定行为: 仅影响图的高阶 (低值) 奇异成分, 由此提出对图进行低秩近似 (仅使用顶部奇异分量进行重构) 以极大地减少 NETTACK 的影响, 提高 GCN 在防御对抗性攻击的能力。该方法根据 SVD 将原始样本分解得到邻

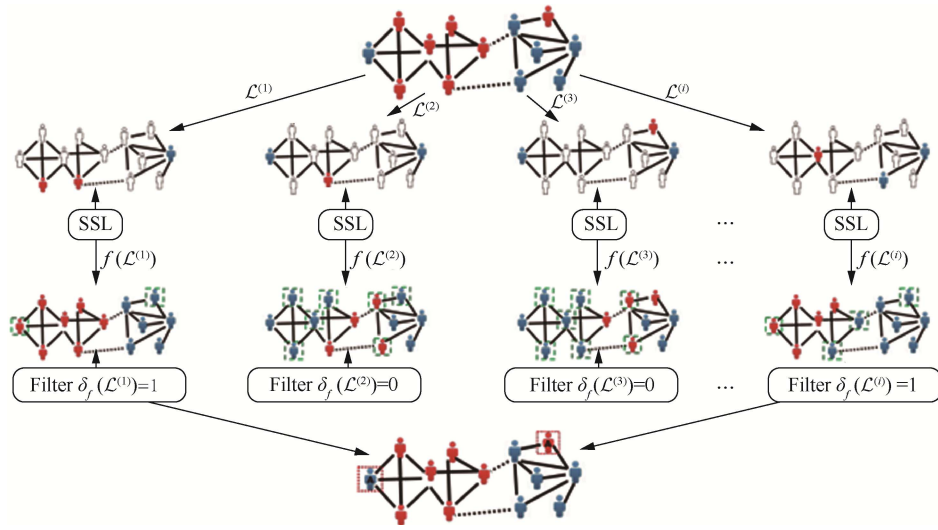


图 11 GraphSAC 的运作过程
Figure 11 Operation process of GraphSAC

接矩阵和特征矩阵的 r 阶近似 A_r 与 X_r , 利用近似矩阵对 GCN 进行再训练以提高其鲁棒性。实验验证不同秩 r 的选择对模型鲁棒性的提升效果, 只要保持邻接矩阵的前 10 个奇异值就足以防御 NETTACK 并提高模型鲁棒性。

4.5 注意力机制

4.5.1 PA-GNN

Tang 等^[51]指出, 图神经网络容易受到攻击的原因在于图数据受到扰动时, 聚合函数将“假”邻居节点识别为真实邻居节点, 并将其信息传播更新其他节点。他们提出一种具有惩罚性聚合机制的新型图神经网络 (PA-GNN), 如图 12 所示, PA-GNN 首先利用对抗攻击方法将扰动连边注入干净的图 $\{G_1, \dots, G_M\}$ 中, 所产生的对抗样本作为监督知识, 通过为扰动连边分配较低的注意力系数来训练一个具有惩罚能力的元模型, 并利用该元模型初始化 PA-GNN, 以保持 PA-GNN 对目标中毒图的惩罚能力使聚合函数更加专注于真实邻居节点, 从而使图神经网络正常工作。

4.5.2 RGCN

Zhu 等^[52]提出一种鲁棒的图神经网络 RGCN, 不同于以往的 GCN, RGCN 采用高斯分布替代特征向量作为每个卷积层中节点的隐藏表示。当图受到攻击时, RGCN 可以自动地吸收高斯分布的方差变化带来的不利影响。为了弥补对抗性攻击在 GCN 中的传播, RGCN 使用一种基

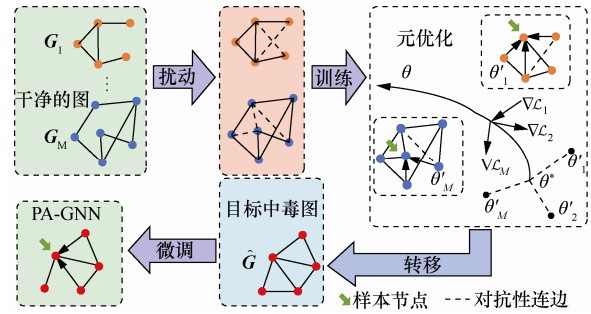


图 12 PA-GNN 的整体框架
Figure 12 Overall framework of PA-GNN

于方差的注意机制, 在执行卷积时根据节点邻域的方差分配不同的权值。在以前的 GCN 中, 当一个节点受到攻击时, 这种对抗效应会通过卷积传播到它的邻居, 从而影响图数据的很大一部分。在 RGCN 中, 由于被攻击的节点对其他节点的影响往往方差较大, 权值较小, 这种传播的影响会大大降低。

4.6 面向图神经网络的鲁棒性分析

图神经网络的鲁棒性分析属于图神经网络防御研究的一个方向, 差别在于图神经网络的鲁棒性分析对模型或数据安全给出了更多的理论性分析与评估。本节对几种有效的图神经网络鲁棒性分析方法进行介绍。

4.6.1 VPN

Jin 等^[53]强调了现有 GCN 的基本构建块 Laplacian 算子的缺陷: ① 空间域中信息融合范

围有限；② 离群特征值会模糊稀疏状态下的 Laplacian 或邻接矩阵的频谱，削弱有用的频谱信号。为了提高图频谱特征应对噪声的鲁棒性，他们提出了可变幂算子，它可以被看作提取隐藏在局部不规则状态下频谱的核函数，将 GCN 中的图卷积算子替换为可变幂算子，得到可变幂网络 (VPN)，同时学习不同距离处的特征变换函数和全局影响参数，提升图神经网络的鲁棒性。

如图 13 所示，在原始图的基础上生成一系列图， $A^{[k]}$ 代表距离为 k 的邻接矩阵，当且仅当节点 i 与 j 的最短距离为 k 时 $[A^{[k]}]_{ij}=1$ ，通过测量影响强度的参数将这些图加权组合，得到一个可变的幂图。该方法大大增加了图运算符的搜索空间，作为一类新的图数据融合算法，具有更广泛的空间融合范围和更强的谱鲁棒性。实验证明，提出的方法可以提高在干净样本和对抗样本中的准确性。

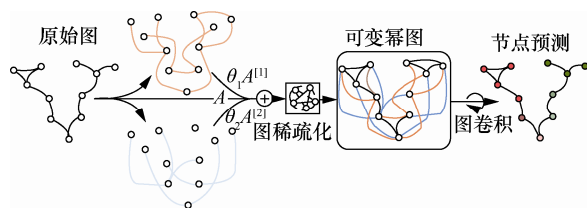


图 13 距离为 k 的可变幂算子
Figure 13 A variable power of distance k

4.6.2 CRIAGE

Pezeshkpour 等^[92]指出，大部分现有的嵌入方法主要侧重于提高准确性，而忽略了鲁棒性分析和可解释性等方面。他们针对知识图，提出链路预测模型的对抗性修正方法 CRIAGE，在模型重训练后，确定要添加到知识图中或从中删除的事实，从而更改目标事实的预测。使用泰勒展开近似嵌入的变化来评估单个修改对图的影响，确定最具影响力的预测事实，并评估模型对添加虚假

事实的敏感性。为了避免对所有可能的事实进行组合搜索，训练了一个网络来解码它们对应的图组件嵌入，使用基于梯度的优化来识别对抗性修改。

以主语、谓语、宾语三部分组成三元组 $\langle s, r, o \rangle$ ，表示知识图中的事实，限定仅能修改事实中的谓语和宾语，得到 $\langle s', r', o \rangle$ 。如图 14(a)所示，蓝色虚线表示原始子图的预测结果，图 14(b)中删除目标的一个邻居连接，导致其中一个事实的消失和预测的变化，图 14(c)展示为目标增加了一个事实三元组后的改变。他们总结每个关系中最有影响力的事实，结合 CRIAGE 研究知识图表示的可解释性，并指出对事实进行影响力评估可以作为检测知识图错误的手段。

4.6.3 AGCN

Ioannidis 等^[54]提出自适应图卷积网络 (AGCN)，给定被扰动的无权图，通过概率选取抖动（增加或删除）连边来绘制多个辅助图，增强 GCN 的鲁棒性。

如图 15 所示，辅助图与节点特征 X 一同输入 AGCN 中，在第 l 个隐藏层，AGCN 首先考虑一个邻域聚合模块 (NAM)，将一个图邻域内的节点特征线性组合，通过多次操作不断扩大邻域，对每个辅助图 i 的 k 跳邻域矩阵 A_i^k 进行参数化映射得到提取的特征向量 $h_{n,i}^{(l)}$ ，捕获每个边抖动图的扩散特征。为了保证模型鲁棒性，提出图自适应模块 (GAM) 将较大的权重分配给未扰动的邻域特征，得到图的自适应特征 $g_{n,i}^{(l)}$ ，再利用特征聚合模块 (FAM) 将 $g_{n,i}^{(l)}$ 与一个可训练向量混合得到该 AGCN 层的输出 $\tilde{Z}^{(l)}$ ，最后连接所有的 AGCN 层，利用数据的剩余输入来促进特征在图中的扩散，更灵活地将输入 X 分散到 L 跳图邻域中，并使用鲁棒的基于图的正则化器以防止过度

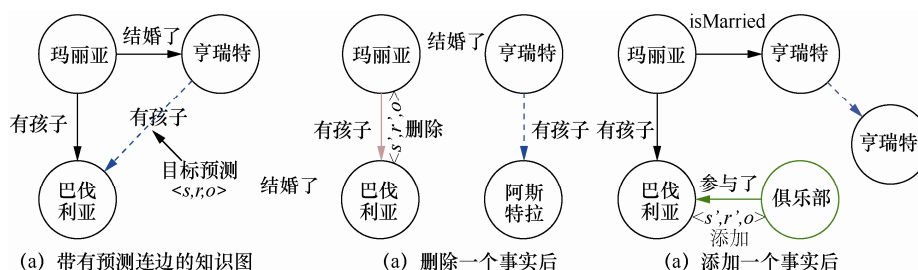


图 14 CRIAGE 的示意
Figure 14 Illustration of CRIAGE

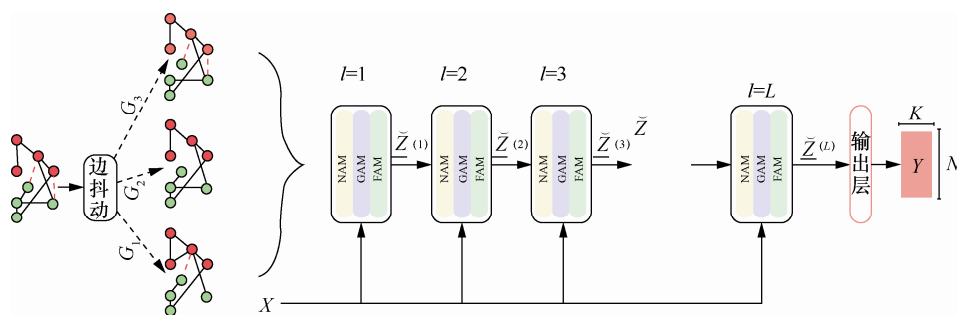


图 15 AGCN 的框架

Figure 15 The framework of AGCN

拟合, 进一步解释底层图的拓扑结构。

4.6.4 具有径向基核函数的 SVM

针对使用图卷积网络进行节点分类任务时易受到针对性的中毒攻击的问题, Miller 等^[55]进行了两个方面的研究。① 使用具有径向基函数内核的支持向量机作为替代分类器, 根据图的结构生成一个 $N \times d_e$ 的新特征矩阵 X_s , 连接到原有节点属性上, 将新的特征矩阵 $X_f = [X_s, X]$ 传递给分类器。② 选择训练数据的方法有两种, 一种是在训练集中简单地选择一定比例的最高度节点, 其余节点按比例划分为测试和验证机; 另一种使用贪婪方法, 从一个空的训练集和阈值 $k = 0$ 开始, 迭代地添加邻居 (在训练集中最多连接 k 个节点) 数量最大的节点。当没有这样的邻居时, 增加 k , 直到得到训练所需的数据集, 并将其余数据随机分为测试和验证集, 以确保每个节点在训练集中至少具有一定数量的邻居。文章表明, 改变训练集可以使网络攻击更难。为了保持一个给定的攻击成功的概率, 对手必须使用更多的扰动, 这通常是随机训练基线的 24 倍。即使在攻击者相对容易成功的情况下, 分类模型和选择训练数据集的方法可以使分类性能更缓慢地下降, 对被攻击节点的错误预测也更弱。

4.6.5 可训练邻接矩阵

Wu 等^[56]指出, 如 GCN 这类图模型的弱点在于这些模型本质上是根据图结构来聚集特征, 在对目标节点进行预测时, 此类模型严重依赖最近的邻居信息。他们提出使邻接矩阵可训练的防御方法, 在训练过程中不断学习连边的权值, 以改变攻击者制作的对抗样本, 从而使其在普通 GCN 上制造的攻击无效; 根据 Jaccard 相似度评分度量

特征的相似性, 并以此观察修改连边/节点特征、增/删连边等攻击的特征, 对上述防御方法的有效性进行了大量的理论分析。他们还提出, 对图的邻接矩阵进行简单的预处理就可以识别攻击后的连边, 通过移除连接低相似度节点的连边, 能够在不降低 GCN 模型准确性的情况下防御有针对性地对攻击。

4.6.6 Pro-GNN

Jin 等^[93]指出, 真实世界的图形共享一些内在属性。例如, 许多现实世界的图是低秩和稀疏的, 两个相邻节点的特征往往是相似的。实际上, 许多对抗攻击很可能会违反这些图的性质。作者应用最先进的图中毒攻击 metattack 来扰动图数据并可视化 metattack 前后的图属性。

如图 16 所示, 在图 16(a)中, metattack 放大了邻接矩阵的奇异值, 图 16(b)说明 metattack 快速增加了邻接矩阵的秩。此外, 当分别从扰动图中移除对抗边和正常边时, 可以观察到移除对抗边比移除正常边降低秩的速度更快, 如图 16(c)所示。图 16(d)中描绘了受攻击图的连通节点特征差异的密度分布, 发现 metattack 倾向于连接特征差异较大的节点。

作者进一步提出了一个通用框架 Pro-GNN, 该框架可以从受这些特性指导的扰动图联合学习结构图和鲁棒图神经网络模型。Pro-GNN 通过保持图的低秩性、稀疏性和特征光滑性, 以中毒图为基础, 迭代地重构干净图, 并通过交替模式求解优化问题, 同时更新重构图上的 GNN 参数, 减少对抗性结构的负面影响。

4.6.7 鲁棒性分析研究总结

鲁棒性分析研究同样提高了图神经网络模型

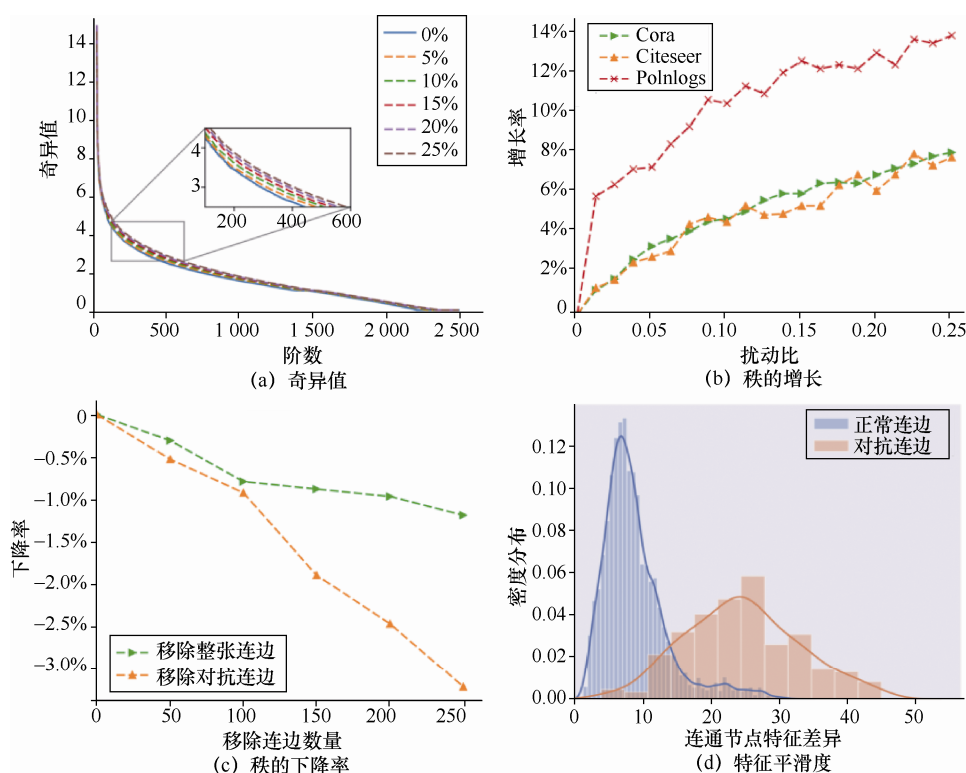


图 16 对抗性攻击导致邻接矩阵属性变化的一个示例
Figure 16 An illustrative example on the property changes of the adjacency matrix by adversarial attacks

的鲁棒性，这些工作更加注重模型鲁棒性的理论分析。例如，通过发现已有 GCN 模型的理论缺陷^[53,56]、挖掘图对模型的作用影响^[54,55,92]和探索真实图数据的性质特征^[89]来更深层次地研究图神经网络的鲁棒性。然而，现有的图神经网络鲁棒性分析研究还相对缺乏，如何更进一步通过理论分析来解释对抗攻击对图神经网络的影响，并从根本上增强图神经网络的鲁棒性，这是未来值得探索的一个方向。

5 数据集及评价指标

已有的面向图挖掘算法的对抗攻防方法研究，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

图神经网络研究中常见的数据集有引文网络数据集、社交网络数据集、生物-化学数据集等类别的数据集。这些数据集详细内容及结构如表 4

所示，其中，对于生物-化学数据集，本文展示了每个数据集中的平均节点数和连边数。前 4 个引文数据集与 PolBlogs 是最常用的节点分类任务数据集，Reddit、Facebook 与 Google+ 数据集也被用于节点分类攻击的工作^[69,89]。此外，Cora 和 Citeseer 数据集也被文献^[94]应用到链路预测攻击问题中。由于存在多种大小的 Facebook 数据集，表 4 中省略了它的统计数据^[95-103]。Email、Dolphin、Karate 和 Football 数据集常出现在社区检测任务^[26,37]中。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。

5.2 评价指标

图分析、对抗攻击或防御方法的效果需要通过不同的评价指标来衡量。本节首先介绍了图分析任务中的常用指标，随后从对抗攻击效果、攻击隐蔽性和防御性能 3 个角度介绍了攻防工作中提出的一些新的度量标准。

5.2.1 图分析任务中的常用指标

图神经网络的下游任务大多以分类问题为

表 4 数据集内容
Table 4 The details of datasets

类别	任务	数据集	来源	图数量	节点数	连边数	特征数	类标数
引文网络数据集	NC/LP	Cora	文献[95]	1	2 708	5 429	1,433	7
	NC	Cora-ML	文献[96]	1	2 995	8 416	2,879	3
	NC/LP	Citeseer	文献[95]	1	3 327	4 732	3,703	6
	NC	Pubmed	文献[95]	1	19 717	44 338	500	3
	NC/CD	DBLP	文献[64]	1	4 107 340	36 624 464	—	—
社交网络数据集	NC/LP	PolBlogs	文献[65]	1	1 490	19 025	—	2
	NC	Reddit	文献[97]	1	232 965	11 606 919	602	41
	NC/LP	Facebook	文献[61]	1	—	—	—	—
	NC/CD	Google+	文献[61]	1	107614	13 673 453	—	—
	CD	Email	文献[61]	1	1 005	25 571	—	—
	CD	Dolphin	文献[98]	1	62	159	—	—
	CD	Karate	文献[99]	1	34	78	—	—
	CD	Football	文献[100]	1	115	613	—	—
生物-化学数据集	GC	ENZYMES	文献[101]	600	32.63	62.14	18	6
	GC	NCI-1	文献[102]	4 110	29.87	32.30	37	2
	GC	MUTAG	文献[66]	188	17.93	19.79	7	2
	GC	D&D	文献[103]	1 178	284.31	715.65	82	2
	GC	PROTEIN	文献[101]	1 113	39.06	72.81	4	2
	GC	PTC	文献[103]	344	25.5	—	19	2

注: NC (node classification); LP (link prediction); CD (community detection); GC (graph classification)

主, 如节点分类、图分类。若把连边的存在情况视为连边类标, 链路预测任务也可以归为分类问题。分类问题通常是一个二类或多类分类问题。现有的基于准确度的度量方法如准确率、F1 评分、AUC 等是从不同的角度反映分类精度的。本节介绍常用的几种不同的准确度指标。

准确度 (Accuracy): 该领域研究中较为常见的技术指标, 它表示正确的预测占所有实例的百分比。

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad (20)$$

其中, TP 表示实际为正样本且被分类器划分为正样本的样本数, TN 表示实际为负样本且被分类器划分为负样本的样本数, P 为正样本数, N 为负样本数。

F1-score: 分类问题的一个衡量指标, 为精确度 precision 和召回率 recall 的调和平均数。

$$F1\text{-Score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (21)$$

其中, 精确度表示为 $\text{precision} = TP/(TP+FP)$, 召回率为 $\text{recall} = TP/(TP+FN)$, 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: ROC 曲线下坐标轴围成的面积, ROC 曲线的横轴表示为 $\text{TPR} = TP/(TP+FN)$, 纵轴为 $\text{FPR} = FP/(FP+TN)$ 。ROC 曲线一般处于 $y = x$ 条直线的上方, 因此 AUC 的取值范围在 0.5 和 1 之间。AUC 越接近 1.0, 检测方法真实性越高。

5.2.2 对抗攻击评价指标

误分类率为对抗攻击评价的最直接体现, 反映了攻击对目标模型性能的下降程度。为了从不同角度衡量图神经网络的攻击性能, 已有的工作在图数据的攻击中提出或使用了以下评价指标。

误分类率 (MCR, mis-classification rate): 描述攻击后目标图分类模型误分类实例数 N_{mis} 与总实例数 N_{all} 的比值。

$$\text{MCR} = \frac{N_{\text{mis}}}{N_{\text{all}}} \quad (22)$$

攻击成功率 (ASR, attack success rate) [24]: 描述在攻击前模型预测正确的情况下, 攻击后模型误判的节点数 N_s 与攻击前预测成功的总节点数 N 的比值。

$$\text{ASR} = \frac{N_s}{N} \quad (23)$$

平均修改连边数 (AML, average modified links) [24]: 描述攻击模型成功时, 网络的平均修改的连边数, 即攻击成功所需要付出的代价。

$$\text{AML} = \frac{1}{N_s} \sum_{i=0}^{N_s} L_i \quad (24)$$

其中, N_s 为攻击成功节点数, L_i 表示当前攻击节点成功需要的连边数。

分类裕度 (CM, classification margin) [28]: 体现了模型错误分类目标节点并且到正确类别边界具有的最大距离。

$$\text{CM} = \max_{y_{v_i} \neq y_{v_i}} Z'_{v_i} - Z_{v_i} \quad (25)$$

其中, v_i 为目标节点, Z 为模型的输出, Z'_{v_i} 为目标节点受攻击后的模型输出, y_{v_i} 为节点 v_i 的类别。因此, CM 仅用于节点分类任务。

平均最坏情况裕度 (AWM, averaged worst-case margin): 最坏情况裕度 (WM, worst case margin) 是分类裕度的最小值 [104], WM 的平均值是在训练过程中对一小批节点进行动态计算得到的。

$$\text{AWM} = \frac{1}{B_s} \sum_{i=1}^{i=B_s} \text{CM}_{\text{worst}_i} \quad (26)$$

其中, B_s 表示批量大小。

鲁棒性能 (RM, robustness merit) [71]: 计算 GNN 攻击前/后的精度差值, 对模型的鲁棒性进行评价。

$$\text{RM} = \text{Acc}_{V^{\text{attack}}}^{\text{post-attacked}} - \text{Acc}_{V^{\text{attack}}}^{\text{pre-attacked}} \quad (27)$$

其中, V^{attack} 表示理论上受攻击影响的节点集合。

攻击恶化 (AD, attack deterioration) [53]: 评估对预测模型的攻击效果, 由于 GNN 的空间范围限制, 任何添加/删除的边只能影响原始网络中空间范围内的节点。

$$\text{AD} = 1 - \frac{\text{Acc}_{V^{\text{attack}}}^{\text{post-attacked}}}{\text{Acc}_{V^{\text{attack}}}^{\text{pre-attacked}}} \quad (28)$$

5.2.3 隐蔽性评价指标

图结构具有离散性, 因此无法使用图像领域中的评价指标来衡量无穷小的细微变化。为了确保在图数据上的攻击是不容易觉察的, 已有工作提出了几种扰动隐蔽性评价指标来衡量对抗扰动的大小。

相似度分数 (similarity score): 用来推断连接存在可能性, 通常应用于链路级任务, 用来量化节点 u 和 v 的共同邻居、度等拓扑属性, 较高的相似度评分表明这对节点连接的可能性更大。相似度评分通常通过余弦相似矩阵 [94] 来衡量。

攻击预算 Δ (attack budget Δ): 通过对数似得分测量扰动的预算 Δ , 累积节点特征和邻接矩阵的变化, 将其限制在 Δ 范围内来限制扰动。

检验统计量 Λ (test statistic Λ): 利用节点度分布测量图之间的相似度, 图 G 中多组节点的度 U_G 服从幂律分布, 利用近似方法 [28] 得到主要参数分布 α_G , 计算对数似然得分 $l = (U_G, \alpha_G)$, 图的最终检验统计量表示为

$$\Lambda(G_{\text{pre}}, G_{\text{post}}) = -2 \cdot l(U_{\text{comb}}, \alpha_{G_{\text{comb}}}) + 2 \cdot l(U_{\text{post}}, \alpha_{G_{\text{post}}}) + 2 \cdot l(U_{\text{pre}}, \alpha_{G_{\text{pre}}}) \quad (29)$$

其中, $U_{\text{comb}} = U_{G_{\text{pre}}} \cap U_{G_{\text{post}}}$, $\alpha_{G_{\text{comb}}} = \alpha_{G_{\text{pre}}} \cap \alpha_{G_{\text{post}}}$ 。

5.2.4 防御方法评价指标

为了检验图神经网络防御方法的效果, 研究人员提出了以下几种防御性能评价指标。

平均防御率 (ADR, average defense rate) [47]: 表示为有/无防御情况下 ASR 的差值与无防御下 ASR 的比值。ADR 值越大, 防御效果越好。

$$\text{ADR} = \frac{\text{ASR}_{\text{defense}} - \text{ASR}_{\text{without-defense}}}{\text{ASR}_{\text{without-defense}}} \quad (30)$$

平均置信度差 (ACD, average confidence different) [47]: 表示节点在攻击前/后的平均置信度差。

$$ACD = \frac{1}{|N_s|} \sum_{v_i \in N_s} CD_i(\hat{A}_{v_i}) - CD_i(A) \quad (31)$$

其中,置信度差 CD 与式(24)的分类裕度等同, N_s 是测试集中攻击前被正确分类的节点集合。 ACD 的值越小,防御效果越好。

损害预防率 (DPR, damage prevention rate)^[91]: 用于评估防御方法可以容忍的攻击数量。

$$DPR = \frac{\mathcal{L}_A - \mathcal{L}_D}{\mathcal{L}_A - \mathcal{L}_0} \quad (32)$$

其中, $\mathcal{L}_0$ 为无攻击时的损失函数, $\mathcal{L}_A$ 为有明确攻击策略 A 下的损失函数, $\mathcal{L}_D$ 为有明确防御策略 D 下的损失函数。DPR 的值越大,该防御策略效果越好。

6 未来的研究方向

上文全面回顾了近年来关于图神经网络对抗攻击与防御的方法,详细描述了现有的攻防技术细节。本节就图神经网络未来的研究方向做出更多一般性的讨论。

6.1 攻击方法

攻击隐蔽性: 本文总结的针对图神经网络的对抗攻击方法主要利用图数据和模型的信息来修改图的结构或节点属性,通过添加扰动使图神经网络模型失效。由于在修改图结构时可能会将两个完全无关节点相连或产生孤立节点,且在网络结构稀疏的情况下修改连边容易被察觉,此类攻击方法存在隐蔽性不强的问题。如何以更隐蔽的方式和更小的代价对图神经模型进行有效攻击是未来可以研究的方向之一。

攻击效率: 现有的图神经网络攻击方法常用的数据集为 Cora 与 Citeseer,两者都是小规模图,当面对大规模图时,现有的方法往往很难高效率地实现攻击。在大规模图中,如何筛选出与攻击目标高度相关的图信息,仅利用少量信息实现对整个图的有效攻击,同样是未来值得思考的。

攻击可行性: 从表 2 可以看出,大部分现有的攻击算法是基于攻击者对目标模型与数据有完备知识条件下建立的,属于白盒或灰盒攻击,而真正的黑盒攻击相对较为缺乏。除此之外,在很多场景下,如社交网络的攻击,修改已有连边与

节点属性是难以实现的。这些问题可以归结为攻击在实际场景下的可行性。在未来的研究中,可以将重心放在如何利用图的某个部分训练一个替代模型,学习对图数据攻击的一般模式,以避免攻击的知识背景限制,并实现更符合真实应用场景的攻击策略上。

6.2 防御方法

防御任务多样化: 针对图神经网络的防御方法通过对抗性训练、对抗性扰动检测、启发式防御、图纯化防御或引入注意力机制等多种手段进行,从表 3 可以看出,大多数方法针对节点分类任务进行防御。研究对其他图神经网络下游任务的防御,或研究现有防御方法在其他任务中的可迁移性是研究人员未来值得考虑的方向。

防御可行性: 防御成本可以作为图神经网络防御方法可行性分析的一个指标,现有的防御方法大多关注提高模型的防御能力,而忽略了对防御成本的研究。在现实场景应用中,大规模的图数据将为高成本的防御方法带来严峻的挑战。通过更多的鲁棒性理论分析,研究在低防御成本下实现高效的防御,提高防御方法的可行性,这也是一个有价值的研究方向。

7 结束语

图神经网络是处理图数据的主要方法之一。随着图神经网络在具有不规则空间结构的图数据(如社交网络、知识图谱等)处理中得到广泛应用,图神经网络的安全问题逐渐成为研究热点。本文总结了面向图神经网络的对抗攻击、防御方法和相应的评价指标,综述了图神经模型的鲁棒性分析方法。分别总结了当前图神经网络攻防方法的局限性,并对未来的研究方向进行讨论。最后,通过引用的参考文献为本课题的研究指明更广阔的前景。

参考文献:

- [1] DENG L, LIU Y. A joint introduction to natural language processing and to deep learning[C]//Deep Learning in Natural Language Processing. Springer, Singapore, 2018: 1-22.
- [2] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning

- Research, 2011, 12(8): 2493-2537.
- [3] XIONG W, DROPO J, HUANG X, et al. Achieving human parity in conversational speech recognition[J]. arXiv Preprint Arxiv: 1610.05256, 2016.
- [4] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.
- [5] MOUSAVI A, BARANIUK R G. Learning to invert: signal recovery via deep convolutional networks[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017: 2272-2276.
- [6] KRISHNAMURTHY B, SARKAR M. Deep-learning network architecture for object detection: U.S. Patent 10,152,655[P]. 2018-12-11.
- [7] GOYAL P, FERRARA E. Graph embedding techniques, applications, and performance: a survey[J]. Knowledge-Based Systems, 2018, 151: 78-94.
- [8] FRASCONI P, GORI M, SPERDUTI A. A general framework for adaptive processing of data structures[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 768-786.
- [9] SPERDUTI A, STARITA A. Supervised neural networks for the classification of structures[J]. IEEE Transactions on Neural Networks, 1997, 8(3): 714-735.
- [10] GORI M, MONFARDINI G, SCARSELLI F. A new model for learning in graph domains[C]//2005 IEEE International Joint Conference on Neural Networks. 2005: 729-734.
- [11] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.
- [12] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[C]//2nd International Conference on Learning Representations. 2014.
- [13] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems 29. 2016: 3837-3845.
- [14] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
- [15] CAO S, LU W, XU Q. Deep neural networks for learning graph representations[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016: 1145-1152.
- [16] KIPF T N, WELLMING M. Variational graph auto-encoders[J]. arXiv preprint arXiv:1611.07308, 2016.
- [17] TANG J, QU M, MEI Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1165-1174.
- [18] WANG S, TANG J, AGGARWAL C, et al. Linked document embedding for classification[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016: 115-124.
- [19] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.
- [20] 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. 2017: 327-335.
- [21] TIAN F, GAO B, CUI Q, et al. Learning deep representations for graph clustering[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014: 1293-1299.
- [22] ALLAB K, LABIOD L, NADIF M. A semi-NMF-PCA unified framework for data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 2-16.
- [23] LEE J B, ROSSI R, KONG X. Graph classification using structural attention[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery. 2018: 1666-1674.
- [24] PEI Y, CHAKRABORTY N, SYCARA K. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015: 2083-2089.
- [25] CHEN L, LIU Y, ZHENG Z, et al. Heterogeneous neural attentive factorization machine for rating prediction[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 833-842.
- [26] XIE F, CHEN L, YE Y, ET AL. Factorization machine based service recommendation on heterogeneous information networks[C]//2018 IEEE International Conference on Web Services, ser. ICWS '18. 2018: 115-122.
- [27] WANG J, HUANG P, ZHAO H, et al. Billion-scale commodity embedding for e-commerce recommendation in Alibaba[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '18. 2018: 839-848.
- [28] ZÜGNER D, AKBARNEJAD A, AND GÜNNEMANN S. Adversarial attacks on neural networks for graph data[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '18. 2018: 2847C2856.
- [29] DAI H, LI H, TIAN T, et al. Adversarial attack on graph structured data[C]//Proceedings of the 35th International Conference on Machine Learning, ser. ICML '18. 2018: 1123-1132.
- [30] WANG X, EATON J, HSIEH C J, et al. Attack graph convolutional networks by adding fake nodes[J]. arXiv preprint arXiv:1810.10751, 2018.
- [31] ZHOU K, MICHALAK T P, WANIEK M, et al. Attacking similarity-based link prediction in social networks[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems, ser. AAMAS '19. 2019: 305-313.
- [32] BOJCHEVSKI A AND GÜNNEMANN S. Adversarial attacks on node embeddings via graph poisoning[C]//Proceedings of the 36th International Conference on Machine Learning, ICML 2019. 2019:

- 695-704.
- [33] SUN Y, WANG S, TANG X, et al. Node injection attacks on graphs via reinforcement learning[J]. arXiv preprint arXiv:1909.06543, 2019.
- [34] CHEN J, WU Y, XU X, et al. Fast gradient attack on network embedding[J]. arXiv preprint arXiv:1809.02797, 2018.
- [35] CHEN J, LIN X, SHI Z, et al. Link prediction adversarial attack via iterative gradient attack[J]. IEEE Transactions on Computational Social Systems, 2020, 7(4): 1081-1094.
- [36] CHEN J, CHEN L, CHEN Y, et al. Ga-based Q-attack on community detection[J]. IEEE Transactions on Computational Social Systems, 2019, 6(3): 491-503.
- [37] CHEN J, CHEN Y, CHEN L, et al. Multiscale evolutionary perturbation attack on community detection[J]. IEEE Transactions on Computational Social Systems, 2021, 8(1): 62-75.
- [38] CHANG H, RONG Y, XU T, et al. A restricted black-box adversarial framework towards attacking graph embedding models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 3389-3396.
- [39] MA Y, WANG S, WU L, et al. Attacking graph convolutional networks via rewiring[J]. arXiv preprint arXiv:1906.03750, 2019.
- [40] ZÜGNER D AND GÜNNEMANN S. Adversarial attacks on graph neural networks via meta learning[C]//7th International Conference on Learning Representations, ser. ICLR '19. 2019.
- [41] SUN M, TANG J, LI H, et al. Data poisoning attack against unsupervised node embedding methods[J]. arXiv preprint arXiv:1810.12881, 2018.
- [42] WANIEK M, ZHOU K, VOROBAYCHIK Y, et al. Attack tolerance of link prediction algorithms: How to hide your relations in a social network[J]. arXiv preprint arXiv:1809.00152, 2018.
- [43] FARD A M, WANG K, AND YU P. Limiting link disclosure in social network analysis through subgraph-wise perturbation[C]//15th International Conference on Extending Database Technology, ser. EDBT '12. 2012: 109-119.
- [44] FARD A AND WANG K. Neighborhood randomization for link privacy in social network analysis[J]. World Wide Web, 2015, 18(1): 9-32.
- [45] LI J, ZHANG H, HAN Z, et al. Adversarial attack on community detection by hiding individuals[C]//WWW '20: The Web Conference 2020. 2020: 917-927.
- [46] FENG F, HE X, TANG J, et al. Graph adversarial training: Dynamically regularizing based on graph structure[J]. IEEE Transactions on Knowledge and Data Engineering. 2019.
- [47] CHEN J, LIN X, XIONG H, et al. Smoothing adversarial training for gnn[J]. IEEE Transactions on Computational Social Systems, 2020.
- [48] SUN K, LIN Z, GUO H, Et al. Virtual adversarial training on graph convolutional networks in node classification[C]//Pattern Recognition and Computer Vision - Second Chinese Conference, PRCV 2019. 2019: 431-443.
- [49] ZHANG Y, KHAN S, AND COATES M. Comparing and detecting adversarial attacks for graph deep learning[C]//Proc Representation Learning on Graphs and Manifolds Workshop, Int Conf Learning Representations. 2019.
- [50] ENTEZARI N, AL-SAYOURI S A, DARVISHZADEH A, et al. All you need is low (rank): Defending against adversarial attacks on graphs[C]//WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining. 2020: 169-177.
- [51] TANG X, LI Y, SUN Y, et al. Transferring robustness for graph neural network against poisoning attacks[C]//WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining. 2020: 600-608.
- [52] ZHU D, ZHANG Z, CUI P, et al. Robust graph convolutional networks against adversarial attacks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 1399-1407.
- [53] JIN M, CHANG H, ZHU W, ET AL. Power up! robust graph convolutional network against evasion attacks based on graph powering[J]. arXiv preprint arXiv:1905.10029, 2019.
- [54] IOANNIDIS V N AND GIANNAKIS G B. Edge dithering for robust adaptive graph convolutional networks[J]. arXiv preprint arXiv:1910.09590, 2019.
- [55] MILLER B A, ÇAMURCU M, GOMEZ A J, et al. Improving robustness to attacks against vertex classification[C]//15th International Workshop on Mining and Learning with Graphs. 2019.
- [56] WU H, WANG C, TYSHETSKIY Y, et al. Adversarial examples for graph data: Deep insights into attack and defense[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019: 4816-4823.
- [57] JIN W, LI Y, XU H, et al. Adversarial attacks and defenses on graphs: a review and empirical study[J]. arXiv preprint arXiv:2003.00653, 2020.
- [58] SUN L, DOU Y, YANG C, et al. Adversarial attack and defense on graph data: A survey[J]. arXiv preprint arXiv:1812.10528, 2018.
- [59] CHEN L, LI J, PENG J, et al. A survey of adversarial learning on graph[J]. arXiv preprint arXiv: 2003.05730, 2020.
- [60] DOMENICO M, LIMA A, MOUGEL P, et al. The anatomy of a scientific rumor[J]. Scientific Reports, 2013, 3(10): 65.
- [61] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph evolution: Densification and shrinking diameters[J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2006,1.
- [62] LI Y, YU R, SHAHAB CI, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting[C]//6th International Conference on Learning Representations. 2018.
- [63] GONG Y, ZHU Y, DUAN L, et al. Exact-k recommendation via maximal clique optimization[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019. 2019: 617-626.
- [64] TANG J, ZHANG J, YAO L, et al. Arnetminer: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 990-998.
- [65] ADAMIC L A AND GLANCE N. The political blogosphere and the 2004 us election: divided they blog[C]//Proceedings of the 3rd International Workshop on Link Discovery. 2005: 36-43.

- [66] DEBNATH A K, COMPADRE R L, DEBNATH G, et al. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity[J]. *Journal of Medicinal Chemistry*, 1991, 34(2): 786-797.
- [67] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. 2014: 2672-2680.
- [68] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations, ser. ICLR '17. 2017.
- [69] WANG B AND GONG N. Attacking graphbased classification via manipulating the graph structure[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 2023-2040.
- [70] AVISHEK J, CIANFLONE A, HAMILTON W. Generalizable adversarial attacks using generative models[J].arXiv preprint arXiv:1905.10864, 2019.
- [71] WU H, WANG C, TYSHETSKIY Y, et al. Adversarial examples for graph data: deep insights into attack and defense[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019. 2019: 4816-4823.
- [72] XU K, CHEN H, LIU S, et al. Topology attack and defense for graph neural networks: An optimization perspective[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019. 2019: 3961-3967.
- [73] ZANG X, XIE Y, CHEN J, et al. Graph universal adversarial attacks: A few bad actors ruin graph learning models[J]. arXiv preprint arXiv:2002.04784, 2020.
- [74] TAKAHASHI T. Indirect adversarial attacks via poisoning neighbors for graph convolutional networks[C]//2019 IEEE International Conference on Big Data (Big Data). 2019: 1395-1400.
- [75] CHEN J, CHEN Y, ZHENG H, et al. Mga: Momentum gradient attack on network[J]. arXiv preprint arXiv:2002.11320, 2020.
- [76] WANIEK M, ZHOU K, VOROBAYCHIK Y, et al. Attack tolerance of link prediction algorithms: How to hide your relations in a social network[J].arXiv preprint arXiv:1809.00152, 2018.
- [77] QIU J, DONG Y, MA H, et al. Network embedding as matrix factorization[C]//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018. 2018: 459-467.
- [78] CHEN J, ZHANG J, CHEN Z, et al. Time-aware gradient attack on dynamic network link prediction[J]. arXiv preprint arXiv:1911.10561, 2019.
- [79] TANG H, MA G, CHEN Y, et al. Adversarial attack on hierarchical graph pooling neural networks[J]. arXiv preprint arXiv:2005.11560, 2020.
- [80] BÁLEK M, GOODALL A. Large networks and graph limits[J]. *Comput Sci Rev*, 2013, 10: 35-46.
- [81] FALOUTSOS C, KOUTRA D, VOGELSTEIN J. Deltacon: A principled massive-graph similarity function[C]//Proceedings of the 13th SIAM International Conference on Data Mining. 2013: 162-170.
- [82] ZHANG Z, JIA J, WANG B, et al. Backdoor attacks to graph neural networks[J]. arXiv preprint arXiv:2006.11165, 2020.
- [83] XI Z, PANG R, JI S, et al. Graph backdoor[C]//USENIX Security. 2021.
- [84] CORDELLA L, FOGGIA P, SANSONE C, et al. A (sub)graph isomorphism algorithm for matching large graphs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(10): 1367-1372.
- [85] YU S, ZHENG J, CHEN J, et al. Unsupervised euclidean distance attack on network embedding[C]//5th IEEE International Conference on Data Science in Cyberspace. 2020: 71-77.
- [86] LI J, ZHANG H, HAN Z, et al. Adversarial attack on community detection by hiding individuals[C]//WWW '20: The Web Conference 2020. 2020: 917-927.
- [87] WANIEK M, MICHALAK T, WOOLDRIDGE M, et al. Hiding individuals and communities in a social network[J]. *Nature Human Behaviour*, 2018, 2(2): 139-147.
- [88] DENG Z, DONG Y, ZHU J. Batch virtual adversarial training for graph convolutional networks[J]. arXiv preprint arXiv:1902.09192, 2019.
- [89] WANG X, LIU X, HSIEH C. Graphdefense: Towards robust graph convolutional networks[J]. arXiv preprint arXiv:1911.04429, 2019.
- [90] IOANNIDIS V, BERBERIDIS D, AND GIANNAKIS B. Graphsac: Detecting anomalies in large-scale graphs[J]. arXiv preprint arXiv:1910.09589, 2019.
- [91] ZHOU K, MICHALAK T, VOROBAYCHIK Y, et al. Adversarial robustness of similarity-based link prediction[J]. arXiv preprint arXiv:1909.01432, 2019.
- [92] PEZESHKPOUR P, IRVINE C, TIAN Y, et al. Investigating robustness and interpretability of link prediction via adversarial modifications[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 3336-3347.
- [93] JIN W, MA Y, LIU X, et al. Graph structure learning for robust graph neural networks[C]//KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Even. 2020: 66-74.
- [94] SUN M, TANG J, LI H, et al. Data poisoning attack against unsupervised node embedding methods[J]. arXiv preprint arXiv:1810.12881, 2018.
- [95] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. *AI Magazine*, 2008, 29(3): 93.
- [96] MCCALLUM A, NIGAM K, RENNIE J, et al. Automating the construction of internet portals with machine learning[J]. *Information Retrieval*, 2000, 3(2): 127-163.
- [97] HAMILTON W, YING Z, AND LESKOVEC J. Inductive representation learning on large graphs[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017: 1024-1034.
- [98] LUSSEAU D, SCHNEIDER K, BOISSEAU O, et al. The bottle-nose dolphin community of doubtful sound features a large propor-

tion of long-lasting associations[J]. Behavioral Ecology and Sociobiology, 2003, 54(4): 396-405.

- [99] ZACHARY W. An information flow model for conflict and fission in small groups[J]. Journal of anthropological research, 1977, 33(4): 452-473.

- [100] GIRVAN M AND NEWMAN M. Community structure in social and biological networks[C]//Proceedings of the National Academy of Sciences. 2002:7821-7826.

- [101] BORGWARDT K, ONG C, SCHOENAUER S, et al. Protein function prediction via graph kernels[J]. Bioinformatics, 2005, 21(1): 47-56.

- [102] WALE N, KARYPIS G. Comparison of descriptor spaces for chemical compound retrieval and classification[C]//Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006). 2006: 678-689.

- [103] DOBSON P, DOIG A. Distinguishing enzyme structures from non-enzymes without alignments[J]. Journal of molecular biology, 2003, 330(4): 771-783.

- [104] ZUGNER D AND GUNNEMANN S. Certifiable robustness and robust training for graph convolutional networks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 246-256.

[作者简介]



陈晋音(1982-), 女, 浙江象山人, 博士, 浙江工业大学副教授, 主要研究方向为人工智能安全、图数据挖掘和进化计算。



张敦杰(1997-), 男, 浙江苍南人, 浙江工业大学硕士生, 主要研究方向为数据挖掘与应用、智能计算。



黄国瀚(1997-), 男, 浙江台州人, 浙江工业大学硕士生, 主要研究方向为图数据挖掘和人工智能安全。



林翔(1996-), 男, 浙江宁波人, 浙江工业大学硕士生, 主要研究方向为图数据挖掘与应用。



鲍亮(1983-), 男, 安徽寿县人, 公安部第三研究所助理研究员, 主要研究方向为信息网络安全、网络防护技术、网络攻击数据分析。