基于模体增强对比学习的图神经网络后门防御方法

陈晋音 [1] 穆文博 [2] 郑海斌 [1, 3, 4]

摘 要 图神经网络在图数据挖掘任务中表现出卓越性能, 因此广泛应用于社交网络、商品推荐等领域. 在图分类任务中, 模型决策高度依赖全局拓扑结构, 使得图神经网络易受后门攻击. 已有研究表明, 在训练数据中注入中毒信息使得训练获得的模型容易被触发样本欺骗, 严重威胁模型安全. 然而, 现有防御方法仍然存在一些挑战, 即对不同后门攻击的防御泛化性弱、无法有效均衡主任务性能与防御成功率等问题. 为此, 首次提出一种基于模体增强对比学习的图神经网络后门攻击防御方法 (Motif-Defense), 可高效防御多种未知类型的后门攻击, 且主任务性能仅略有下降. 首先, 设计模体角度增强图的对比学习模型, 选取可疑后门样本. 其次, 使用 Jaccard 相似度和标签平滑策略将可疑后门样本净化为干净样本, 实现对图后门攻击的防御. 最终, 在四个真实数据集上展开防御实验, Motif-Defense 平均降低 84.70% 的攻击成功率, 且分类准确率平均仅下降 2.53%.

关键词 图神经网络; 后门攻击; 防御; 对比学习; 模体

引用格式 陈晋音, 穆文博, 郑海斌. 基于模体增强对比学习的图神经网络后门防御方法. 自动化学报, 2026, **52**(4): 765−779

DOI 10.16383/j.aas.c240767 **CSTR** 32138.14.j.aas.c240767

Motif-augmented Contrastive Learning-based Defense Against Backdoor Attack on Graph Neural Networks

CHEN Jin-Yin[1] MU Wen-Bo[2] ZHENG Hai-Bin[1, 3, 4]

Abstract Graph neural networks (GNNs) have demonstrated strong performance in graph data mining tasks and are widely applied in social networks and recommendation systems. In graph classification, model decisions rely heavily on global topological structures, making GNNs vulnerable to backdoor attacks. Existing studies show that injecting poisoned samples into the training set can implant hidden triggers, causing the trained model to be misled by trigger patterns at inference time and thus posing serious security threats to model security. However, existing defense methods still suffer from limited generalization to different backdoor attacks and difficulties in balancing defense success rate with main task performance. To address these challenges, we first propose a motif-augmented contrastive learning-based defense method against backdoor attacks on graph neural networks, termed Motif-Defense, which can effectively defend against multiple unknown backdoor attacks while incurring only a slight performance degradation on the main task. Specifically, a motif-oriented enhanced contrastive learning framework is designed to identify suspicious backdoor samples, which are then purified into clean samples using Jaccard similarity and label smoothing strategies, thereby achieving defense against graph backdoor attacks. Extensive experiments on four real-world datasets against six backdoor attack methods and five defense baselines show that Motif-Defense reduces the average attack success rate by 84.70%, while the classification accuracy decreases by only 5.32%.

Keywords graph neural networks; backdoor attacks; defense; contrastive learning; motif

Citation Chen Jin-Yin, Mu Wen-Bo, Zheng Hai-Bin. Motif-augmented contrastive learning-based defense against backdoor attack on graph neural networks. *Acta Automatica Sinica*, 2026, **52**(4): 765−779

收稿日期 2024-12-02 录用日期 2026-01-04

Manuscript received December 2, 2024; accepted January 4, 2026

国家自然科学基金 (62406286, 62072406), 工业和信息化部电子第五研究所重点实验室开放课题 (HK00202503455), 北京生命科技研究院有限公司开放基金 (2024200CD0210), 四川大学数据安全防护与智能治理教育部重点实验室开放课题 (SCUSAKFKT202402Z), 浙江省自然科学基金 (LDQ23F020001, LD22F020002) 资助

Supported by National Natural Science Foundation of China (62406286, 62072406), Key Laboratory of the Fifth Research Institute of Electronics, Ministry of Industry and Information Technology (HK00202503455), Beijing Life Science Academy (BLSA) (2024200CD0210), Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University (SCUSAKFKT202402Z), and Zhejiang Provincial Natural Science Foundation (LDQ23F020001, LD22F020002)

本文责任编委 俞扬

Recommended by Associate Editor YU Yang

1. 浙江工业大学计算机科学与技术学院 (软件学院、人工智能学院) 杭州 310023 2. 浙江工业大学信息工程学院 杭州 310023 3. 北京生命科技研究院有限公司 北京 102206 4. 四川大学数据安全防护与智能治理教育部重点实验室 成都 610065

1. College of Computer Science and Technology (College of Software, College of Artificial Intelligence), Zhejiang University of Technology, Hangzhou 310023 2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023 3. Beijing Life Science Academy, Beijing 102206 4. Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University, Chengdu 610065

图结构数据广泛存在于现实生活中, 其所包含的大量节点特征与连边关系能为图分类、节点分类、链路预测、社区发现等任务提供丰富信息, 例如社交媒体网络[1]、金融网络[2]、交通网络[3]. 为了高效实施图数据挖掘, 基于深度学习的图神经网络 (graph neural network, GNNs)[4] 应运而生. 得益于神经消息传递框架, GNNs 能够高效表征图结构的关键特征, 从而得到广泛应用.

本文聚焦于图分类任务下的图神经网络后门防御. 图分类任务旨在通过学习图的拓扑结构与节点属性信息, 实现对图数据的类别划分, 广泛应用于医疗诊断[5]、药物发现[6] 和网络安全[7] 等领域. 由于图分类模型需聚合全局结构信息以形成图级表示, 其决策过程对特定子图模式高度敏感. 因此, 针对图分类任务设计高效、鲁棒的后门防御机制, 不仅具有理论研究价值, 更对保障关键应用的安全可信运行具有重要意义. 图神经网络在图表示学习领域取得优异表现, 然而最近的研究表明图神经网络容易受到后门攻击. 即通过在图的特定子图或者节点上植入触发器, 导致模型在推理时错误地将包含触发器的整个图分类为目标类别. 目前, 已有的图神经网络后门攻击方法[8−12] 多数是以子图生成触发器的方式进行攻击. 攻击者可以通过设计特定的子图 (如某种结构的节点和边) 作为触发器. 这种子图在训练时被标记为特定类别或特征, 使目标模型建立触发器与目标类之间的强相关性, 当模型遇到相同结构的子图时, 产生错误的推理结果.

为保护 GNN 免受后门攻击的影响, 一个直观的想法是删除后门触发器或减少负面影响. 在这个方向上已经有一些研究成果[13−19]. 例如, Dai 等[13] 修剪连接低余弦相似度的边来破坏触发结构和附加边. Jiang 等[17] 在连接节点具有较高相似度的假设下, 使用 Jaccard 相似度对图数据进行剪枝. Liu 等[15] 对由触发模型激活的神经元进行微调, 稀释污染数据的影响. 这些方法主要依赖于模型参数或可解释性工具来删除图数据中的节点或连边, 从而去除后门触发器. 然而, 这些方法依然面临两个问题: 1) 针对特定类型的后门攻击进行设计, 导致对其他攻击或未知攻击的适应性差; 2) 在提高模型对后门攻击的防御能力时, 主任务性能下降严重, 无法有效均衡主任务性能与防御成功率.

针对这些问题, 本文提出一种面向图神经网络后门攻击的对比学习防御方法, 称为 Motif-Defense. 本方法灵感来源于文献 [12] 的发现: 对于图后门攻击, 使用数据集中不存在或不频繁的模体 (motif) 作为触发器, 其攻击效果通常优于使用常见模体. 基于这一发现, Motif-Defense 设计基于模体增强的对比学习模型, 旨在预先学习数据集中存在最少模体结构下的正常嵌入, 从而破坏后门攻击后目标类与触发器之间的强相关性. 首先, 针对问题 1), 设计基于模体 (motif) 角度增强图的对比学习模型, 模体增强方式不依赖于特定攻击方法的具体细节, 而是基于图结构和节点特征进行分析, 因此它能够有效地应对多种不同类型的后门攻击. 然后, 针对问题 2), Motif-Defense 通过比较目标模型和对比学习模型在相同训练数据集下的输出置信度, 准确地检测出可疑后门样本, 并使用标签平滑策略得到可疑后门样本的新标签, 从而减少对非后门样本的影响, 达到保护目标模型主任务性能的效果. 标签平滑策略能够有效减轻后门攻击中篡改训练数据标签带来的负面影响, 减少模型对错误标签的依赖, 从而提高模型的鲁棒性.

本文的主要贡献总结如下:

1) 设计基于模体角度增强图的对比学习模型用于防御. 它的增强方式不同于常规的删除节点、连边的增强方式, 而是在对数据进行模体分析后, 选择部分节点按照特定结构进行连接.

2) 根据对比学习模型输出的置信度快速地定位后门图数据, 并通过删除低排序节点和标签平滑策略消除后门威胁. 不仅大大提高 GNNs 对各种后门攻击的鲁棒性, 而且模型主任务性能保持较好.

3) 在四个真实数据集、六种后门攻击方法和五种后门防御方法上对 Motif-Defense 的防御性能进行评估, 结果表明该方法能够平均降低 84.70% 的攻击成功率, 且分类准确率平均仅下降 2.53%.

1 相关工作

本节简要介绍图神经网络的后门攻击、后门防御、对比学习和 motifs 应用方面的相关工作.

1.1 图神经网络后门攻击

已有的图神经网络后门攻击方法多数以基于子图生成触发器的方式进行攻击. 首先, Zhang 等[8] 探讨了在 GNNs 中嵌入后门的可行性, 通过将特定子图作为触发器嵌入到随机选择的训练样本中, 并将这些样本的标签更改为目标类, 实现对 GNNs 模型的攻击. Xu 等[9] 利用 GNN 可解释方法分析 GNNs 的预测, 并基于分析结果选择对模型预测影响较小的节点或节点特征作为触发注入位置. Xi 等[10] 通过数据中毒将后门嵌入到训练数据中, 设计了动态训练的后门触发器生成器并采用双层优化训练范式, 同时训练模型和触发器生成器. Sheng 等[11] 设计了

基于数据特征的子图触发器, 并通过融合结构特征选择攻击节点. Zheng 等[12] 基于触发子图和正常子图的分布模式差异性, 选择偏离正常分布的子图作为触发器, 利用节点重要性确定最优触发器注入位置. Dai 等[13] 设计了一种高效的攻击方法, 通过精心选择中毒节点和生成难以被检测的触发器, 以有限的攻击成本实现对目标节点的有效攻击.

1.2　图神经网络后门防御

目前针对 GNNs 后门攻击的防御方法主要是对测试样本进行过滤和检测. Dai 等[13] 在后门攻击者创建的边连接不同节点的事实前提下, 通过修剪连接低余弦相似度的边来破坏触发结构和附加边. Yang 等[16] 提出一种基于数据过滤的防御方法, 利用了后门攻击中子图触发器通常与正常图结构不同的特征分布事实. 该方法利用分布差异, 识别并过滤掉异常子图, 从而修正中毒图数据的分布. Jiang 等[17] 使用可解释性工具评估恶意样本和良性样本的差异, 并以此删除带有触发器的子图. 陈晋音等[18] 采用对比学习模型以及图重要性指标去除训练数据集中的扰动. Xiao 等[19] 通过删除度量空间中表现异常的低值边来识别并消除被干扰的边. 同时, 为保留图结构的宝贵信息并防止删除过程中出现孤立节点, 采用最小连通性原则作为终止条件.

1.3　图神经网络对比学习

图对比学习 (graph contrastive learning, GCL) 近年来在无监督和半监督学习领域取得了显著进展[20−25]. GCL 引入对比损失, 促进模型在学习过程中获得更具辨识度的图表示 (即节点或图级别的嵌入), 从而能够有效地捕捉图的结构信息和节点特征.

现有的 GCL 按照数据增强方式可分为基于结构视图和基于属性视图两类. 基于结构视图的方法主要关注学习图的结构信息, 例如节点和边的连接关系、子图结构等. You 等[20] 使用四种不同的图增强方法来生成不同视图, 达到增强不同视图下图表示一致性的目的. Qiu 等[21] 从图中采样子图作为增强的图实例, 模型可以学习到不同图之间的结构相似性. Hassani 等[22] 利用扩散过程将节点与其连接的邻居节点连接起来, 促使模型学习到图的全局结构信息. 基于属性视图的方法主要关注学习节点的属性信息, 例如节点的标签、特征等. Yu 等[23] 使用随机噪声进行数据增强, 通过随机掩码节点的特征改变节点的属性信息. Cai 等[24] 对节点的特征进行线性变换, 帮助模型学习节点不同属性特征之间的关系. Xu 等[25] 利用生成模型生成新的节点特征, 从而学习节点的潜在特征空间.

1.4　模体在图神经网络中的应用

近年来, 随着图神经网络在图结构数据分析中的广泛应用, 模体作为图结构中重复出现的、统计上显著的子图模式, 逐渐引起了研究者的关注. Sankar 等[26] 提出 Motif-CNN, 利用注意力模型融合多个模体模式提取的特征, 捕捉高阶的结构和特征信息, 从而提升图分类的性能. Yang 等[27] 提出一个分层网络嵌入方法, 结合模体过滤和卷积神经网络, 捕捉网络中的精确小结构, 提升图分类的性能. Zhao 等[28] 利用模体捕捉同类型节点之间的高阶关系, 并设计了基于模体的推荐模型, 提升推荐系统的性能. Dareddy 等[29] 利用模体捕捉异构图中的高阶、重复和统计上显著的连接模式, 更好地学习节点表示或嵌入. Shao 等[30] 通过顶点和各种模体的学习嵌入, 在统一的表示中连接连通性和结构相似性, 提升节点嵌入的质量. Zhao 等[31] 统一模体的高阶结构和原有的低阶结构, 通过模体感知图表示学习, 提升 GNNs 对对抗样本的鲁棒性. Wang 等[32] 提出一种新的嵌入算法, 该算法结合模体来捕获网络中的高阶结构以进行链路预测. 此外, Zheng 等[12] 从模体的角度重新审视图后门攻击, 并提出基于模体的后门攻击框架 Motif-Backdoor, 它通过分析数据集中模体的分布, 选择合适的模体作为触发器, 并结合图结构和 GNNs 模型反馈优化触发器注入位置, 最终实现高效的后门攻击.

2　预备知识

本节介绍图分类任务上的图神经网络、图神经网络后门防御、模体的相关定义.

定义 1 (图分类任务上的图神经网络). 给定一个图, 其中, V 和 E 分别表示节点集和边集. 也可以使用 $G=(A,\ X)$ 来表示一个图, 其中, $A\in\{0,\ 1\}^{N\times N}$ 表示邻接矩阵, X 用来表示节点特征矩阵, N 表示节点数量. 图分类数据集表示为 $\mathcal{G}=\{(G_1,\ y_1),\ \cdots,\ (G_M,\ y_M)\}$, 包含 M 个图, 其中 G_i 表示第 i 个图, y_i 表示第 i 个图对应的标签. $Y_l=\{c_1,\ \cdots,\ c_L\}$ 表示数据集有 L 类的标签空间. 图神经网络模型 $f_\theta(\cdot)$ 是一个图分类器, 其目标是通过现有标签的数据训练图神经网络模型 $f_\theta(\cdot)$ 来预测数据集中无标签的图, 即构建映射函数为 $f_\theta:\mathcal{G}\to Y_l$.

定义 2 (图神经网络后门防御). 给定一个图分类的数据集 $\mathcal{G}$、良性模型 $f_\theta(\cdot)$ 和后门模型 $f_{\hat{\theta}}(\cdot)$. 攻击者通过混合函数 $M(\cdot)$ 将触发器 t 注入到良性样

本 G 中生成后门样本 $\hat{G}$, 使得后门模型 $f_{\hat{\theta}}(\cdot)$ 预测标签为预设的目标类 y_t. 因此, 防御图神经网络后门攻击的目标是攻击者生成后门样本 $\hat{G}$, 但防御下的目标模型 $f_{\hat{\theta}}(\cdot)$ 仍可以进行准确的分类, 由式 (1) 表示如下:

$$\begin{cases} f_{\hat{\theta}}(M(G,\ t)) = y_t, \\ f_{\hat{\theta}}(M(G,\ t)) = f_{\theta}(G), \end{cases} G \in \mathcal{G} \tag{1}$$

其中, $M(\cdot)$ 是负责将触发器 t 注入到良性样本中的混合函数, y_t 是攻击者选定的目标类. 防御的目标是使得攻击者无法通过中毒样本误导目标模型, 同时确保目标模型能够正常分类良性样本.

定义 3 (模体). 给定一个图 $G = \{V,\ E\}$, 模体是在图 G 中重复出现的子图 $G' = \{V',\ E'\}$, 其中 $V' \subset V$, $E' \subset E$, 且 $|V'| \leq |V|$. 模体反映了特定网络类型的底层过程, 它比随机子图更频繁地存在于一种类型的网络中. 模体的结构复杂性通常与其拓扑规模正相关. 遵循网络科学中的普遍认知[14, 33], 模体的复杂度由其节点数量 $|V'|$ 和连边数量 $|E'|$ 共同决定, 节点越多、边越多的模体被视为结构更复杂的模体. 本文依照文献 [12] 的研究, 重点关注如图 1 所示的三节点和四节点模体, 为便于分析与数据增强操作的选择, 依据模体的结构复杂度将其从低到高进行排序, 即按照节点数优先、边数次之的原则进行升序排列, 并依次标记为 M_{31}, M_{32}, $\cdots$, M_{46}.

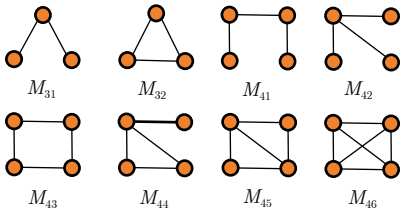


图 1 三节点、四节点模体示意图

Fig. 1 Illustration of three-node and four-node motifs

3 方法

现有的图神经网络后门攻击主要依赖于攻击者精心设计的触发器与目标类标签之间的强关联性. 攻击者将特定结构的子图作为触发器嵌入到良性数据中, 并将标签修改为目标类. 这样, 目标模型在训练过程中就会建立起触发器与目标类之间的关联, 从而在推断阶段被攻击者利用触发器激活后门, 进而误导模型的预测结果. 为了有效应对这种攻击, 本文提出一种面向图神经网络后门攻击的对比学习防御方法, 该方法采用了基于模体角度的数据增强方式, 称为 Motif-Defense. Motif-Defense 的总体框架如图 2 所示, 主要由以下三部分组成:

1) 对比学习模型构建. 分析图数据中模体的分布情况, 根据分布情况构建基于模体角度增强的对比学习模型, 该模型能够学习到图数据的潜在特征.

2) 可疑后门样本识别. 利用对比学习模型和目标模型对训练数据进行预测, 并计算两者输出置信分数之间的差异. 根据差异阈值, 筛选出可疑的后门样本. 这些样本可能包含被篡改的标签和嵌入的触发器.

3) 可疑后门样本处理. 对可疑后门样本进行处理, 以消除触发器并纠正标签. 具体来说, 通过计算可疑后门样本中节点对的 Jaccard 相似度, 删除相似度较低的节点对之间的连边. 同时, 使用标签平滑策略对可疑后门样本的原始标签进行修改, 以减少模型对错误标签的依赖. 通过以上步骤, Motif-Defense 能够有效地消除后门样本中的触发器, 提高模型的鲁棒性.

3.1 对比学习模型构建

首先使用 Orca 算法[34] 对目标数据集进行模体分析, 主要是确定定义 3 中描述的 M_{31} 到 M_{46} 模体的分布情况. 然后基于模体分布情况选择合适的模体结构进行数据增强, 构建对比学习模型. 模体分布的分析结果可分为两类情形: 情形一为所有候选模体结构均存在于图数据中; 情形二为部分候选模体结构在图数据中缺失. 本文建立如下模体选择准则:

1) 针对情形一, 优先选择出现频率最低的模体结构.

2) 针对情形二, 优先选择缺失的模体结构. 当存在多个缺失模体时, 为降低防御成本, 进一步优先选择复杂度最低者 (依据定义 3).

在模体结构选择完毕后, 采用两种图增强方式构建正负样本, 图增强方式一为随机丢弃节点或者随机删除连边, 图增强方式二为模体角度的增强操作. 具体流程如下:

假设选择 M_{43} 模体结构, 首先, 按照度中心性 (degree centrality, DC)[35] 将图数据中所有节点降序排列. 接着, 选取其中前四个节点并按照 M_{43} 结构进行连接. 此时需要注意, 如果原图中已经存在上述连接, 则舍弃排列中的第一个节点, 继续选取前四个节点按照结构进行连接. 重复此过程, 直到选择的连接方式在原图中不存在为止. 度中心性是图论中常用的一种指标, 用于评估节点在网络中的重要性. DC 的定义表示如下:

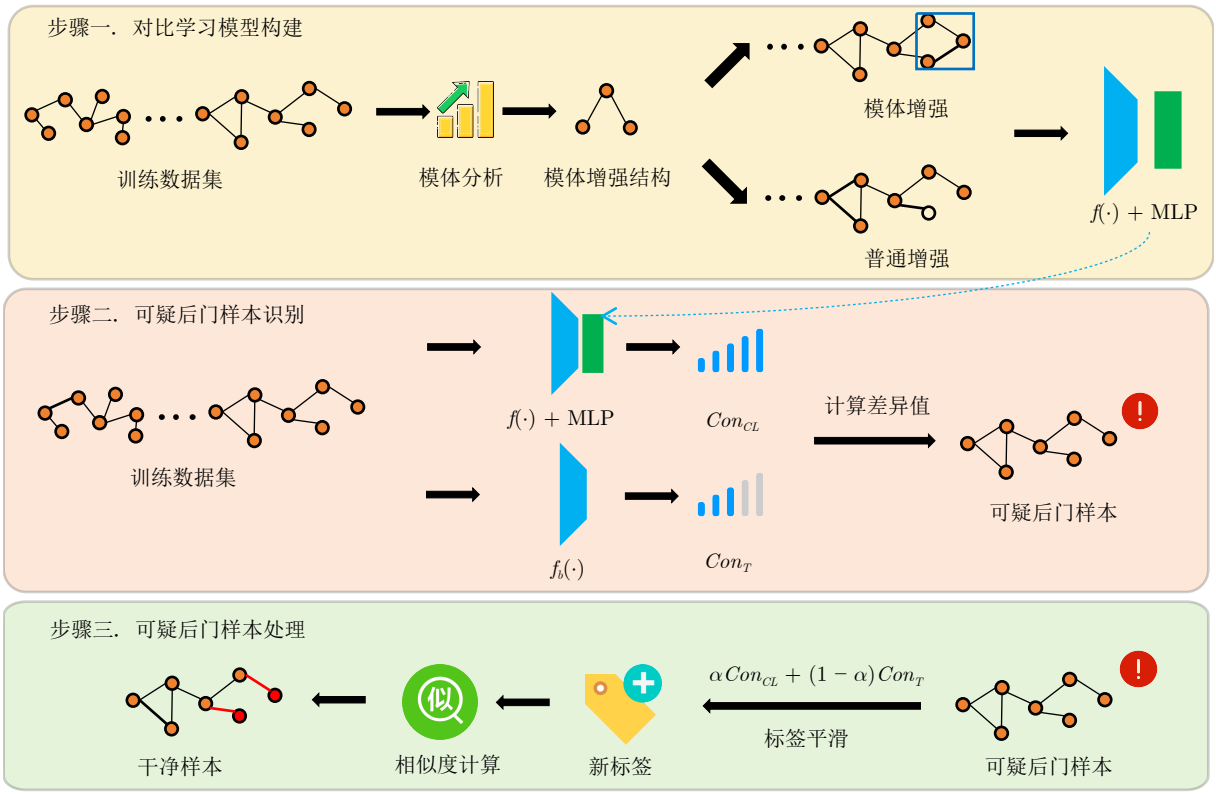


图 2　Motif-Defense 系统框架

Fig. 2　The framework of Motif-Defense system

$$\mathrm{DC}_i = \frac{d_i}{N-1} \tag{2}$$

其中, d_i 是节点 i 的度, N 是图中的节点数.

至此, 正负样本构建完毕, 将构建过程简要描述如下:

$$\begin{cases} \tilde{G}_1 = \mathrm{Aug}_1(G), \\ \tilde{G}_2 = \mathrm{Aug}_2(G), \end{cases} \quad G \in \mathcal{G}_{\mathrm{train}} \tag{3}$$

其中, $\mathrm{Aug}_1(\cdot)$, $\mathrm{Aug}_2(\cdot)$ 为图增强操作; $\mathcal{G}_{\mathrm{train}}$ 为训练数据集, 共有 M 个样本; $\tilde{G}_1$ 和 $\tilde{G}_2$ 表示属于 $\mathcal{G}_{\mathrm{train}}$ 中的 G 经过增强操作后的增强图, 其中, 来自同一样本的增强图认定为正样本, 而不同样本增强图之间认定为负样本.

在获得正负样本后, 使用图神经网络 $f(\cdot)$ 和多层感知器 (multi-layer perceptron, MLP) 学习它们的嵌入, 如下所示:

$$z_i = \mathrm{MLP}(f(\tilde{G}_i)) \tag{4}$$

接着, 采用噪声对比估计损失 (noise contrastive estimation, NCE)[36] 作为对比模型的损失函数. 本节定义了一个负样本库 $\mathcal{Q}$, 其中对于每一个正样本对 $\{z_i, z_j\}$, 包含 K 个负样本嵌入 $\{z_k\}_{k=1}^{K}$, 使用点积相似度 $\mathrm{sim}(\cdot, \cdot)$ 计算正负样本对之间的相似度. 基于此, NCE 损失定义如下:

$$\mathcal{L}_{NCE} = -\log \frac{\exp\left(\frac{\mathrm{sim}(z_i, z_j)}{\tau}\right)}{\exp\left(\frac{\mathrm{sim}(z_i, z_j)}{\tau}\right) + \sum_{k=1}^{K} \exp\left(\frac{\mathrm{sim}(z_i, z_k)}{\tau}\right)} \tag{5}$$

其中, τ 是温度参数, 用于调节正负样本间相似度分布的锐化程度. 负样本库规模 K 的设置参考了 MoCo[37], 采用一个动态队列来存储历史批次中生成的图嵌入, 形成大规模且多样化的负样本集合, 从而突破单批次负样本数量限制, 增强模型判别能力.

在训练过程中, 每经过一个训练步骤, 就根据评分函数 $S(\cdot)$ 对负样本库 $\mathcal{Q}$ 中的负样本进行升序排序, 分数越高代表该负样本越难被学习, 应该在训练后期被采用. 设总训练步数为 V, 则在每个训练步骤前从排序后的负样本库 $\mathcal{Q}$ 中选取前 K/V 个负样本计算对比损失. 评分函数 $S(\cdot)$ 使用余弦相似度, 计算式为:

$$S(z_k) = \frac{|z_k \cdot z_i|}{|z_k||z_i|} \tag{6}$$

其中, 评分函数 $S(\cdot)$ 与总训练步数 V 共同实现负样本采样机制. 该机制模拟课程学习过程, 使模型从

易于区分的负样本开始学习, 逐步过渡到更困难的样本, 有助于提升训练稳定性和泛化能力.

3.2 可疑后门样本识别

第 3.1 节中基于模体增强的对比学习模型学习到的嵌入 z_i 没有用到标签, 该模型主要通过最大化正样本对之间的相似性, 同时最小化负样本对之间的相似性, 从而学习到正确的图表示. 这样, 对比学习模型能够识别出与正常图结构差异较大的后门样本, 从而有效地识别可疑的后门样本. 对比学习模型和目标模型 $f_b(\cdot)$ 在面对后门样本时的预测结果会存在显著差异. 主要原因是目标模型由于受到后门攻击的影响, 会将带有触发器的后门样本错误地预测为目标类别. 而对比学习模型由于其独立于数据标签的训练过程, 对触发器不敏感, 因此能够更准确地预测后门样本的真实类别.

利用这一差异识别可疑的后门样本, 分别使用对比学习模型和目标模型对训练数据集进行预测, 并计算两者输出置信分数之间的差异. 置信分数差异较大的样本被识别为可疑后门样本. 注意, 此处需要对对比学习的输出进行归一化, 使其与目标模型的输出置信分数处于相同的概率空间. 两者的输出置信分数可表示为:

$$\begin{cases} Con_{CL} = \mathrm{Nor}(\mathrm{MLP}(f(\tilde{\mathcal{G}}_{\mathrm{train}}))) \\ Con_T = f_b(\mathcal{G}_{\mathrm{train}}) \end{cases} \tag{7}$$

其中, $\mathcal{G}_{\mathrm{train}}$ 表示训练数据集, 共有 M 个图样本, L 个样本标签; $Con_{CL} \in \mathbf{R}^{M\times L}$ 和 $Con_T \in \mathbf{R}^{M\times L}$ 分别表示对比模型和目标模型的输出置信分数; $\mathrm{Nor}(\cdot)$ 为归一化操作. 采用曼哈顿距离计算 Con_{CL} 和 Con_T 之间的差异, 计算式为:

$$\Delta_i = \sum_{j=1}^{L} |Con_{CL}(i,\ j) - Con_T(i,\ j)|,$$
$$i \in \{1,\ 2,\ \cdots,\ M\} \tag{8}$$

其中, Δ_i 表示训练集中第 i 个图按照式 (7) 计算的输出置信分数之间的差异值, 将差异值大于阈值 o_1 的样本视为可疑后门样本, 即当 $\Delta_i > o_1$ 时, 此图被识别为可疑后门样本.

应当说明的是, 这种基于置信度差异的识别方法, 其本质是通过统计阈值来区分正常样本与后门样本. 因此, 该方法不可避免地存在将部分正常样本误判为后门样本的风险, 这也是大多数基于统计或学习的检测方法共同面临的挑战[16−17]. 在 Motif-Defense 中, 一旦正常样本被错误识别, 系统将对其执行标签平滑等净化操作, 这种对正常样本的不当处理, 直接干扰了其原始、正确的监督信号, 可能对模型的主任务性能产生负面影响. 为缓解此问题, Motif-Defense 基于模体的对比学习增强检测的准确性, 并采用局部化、精细化的样本净化策略 (详见第 3.3 节), 旨在最大限度地减少此类误判, 并在发生误判时将负面影响降至最低.

3.3 可疑后门样本处理

在可疑后门样本识别步骤成功筛选出可能包含后门触发器的可疑后门样本后, 需要对可疑后门样本进行处理, 以消除触发器并纠正标签, 从而提高模型的鲁棒性. 针对可疑后门样本, 首先利用节点重要性指标评估图中每个节点的重要性. 常用的节点重要性指标包括度中心性[35]、介数中心性[38]、特征向量中心性[39]、Jaccard 相似度[40] 等. 这些指标可以帮助识别图中的关键节点, 从而判断触发器可能存在的位置. 本文采用 Jaccard 相似度来区分中毒边和正常边, 该相似度是通过计算两个节点共同邻居数与总邻居数之比得出. 对于一对存在连边的节点 v_i 和 v_j, 它们的 Jaccard 相似度定义如下:

$$J(v_i,\ v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|} \tag{9}$$

其中, $N(v_i)$ 和 $N(v_j)$ 分别表示节点 v_i 和 v_j 的邻居集合. 使用式 (9) 对可疑样本中所有存在连边的节点对计算相似度后, 将相似度小于阈值 o_2 的节点对之间的连边识别为中毒边并删除. 在删除操作后, 需要更新图数据的邻接矩阵和节点特征矩阵.

为了纠正可疑后门样本的错误标签, Motif-Defense 采用标签平滑策略. 具体而言, 利用对比模型和目标模型输出的置信分数, 通过加权平均的方式得到平滑后的置信分数, 表示为:

$$Con_{new}(i) = \begin{cases} \alpha Con_{CL}(i) + (1-\alpha) Con_T(i), \\ \qquad\qquad \Delta_i > o_1 \\ Con_T(i), \qquad \text{其他} \end{cases} \tag{10}$$

其中, α 为标签平滑率, i 为使用式 (8) 识别为可疑后门样本的序号. 将平滑后的置信分数向量中最大值对应的类别作为可疑后门样本的新标签. 该操作不仅可以减轻攻击中篡改训练数据标签带来的负面影响, 还可以防止模型过拟合训练数据中的噪声, 从而提高模型的泛化能力.

3.4 Motif-Defense 时间复杂度分析

Motif-Defense 的时间成本主要来自四个部分, 包括模体分析的时间成本 (T_{motif})、对比学习模型训练的时间成本 (T_{cl})、可疑样本识别的时间成本 (T_{sus}) 以及可疑样本处理的时间成本 (T_{mod}). 因此, Motif-Defense 的时间复杂度可表示为:

$$\mathrm{O}(T_{\mathrm{motif}})+\mathrm{O}(T_{\mathrm{cl}})+\mathrm{O}(T_{\mathrm{sus}})+\mathrm{O}(T_{\mathrm{mod}})\sim\mathrm{O}(N) \quad (11)$$

其中, $\mathrm{O}(T_{\mathrm{motif}})$ 在实际操作中会限制模体大小, 选择较小的模体进行分析, 且对图数据进行采样, 只分析部分节点或边, 可以显著降低计算时间; $\mathrm{O}(T_{\mathrm{cl}})$ 取决于对比模型训练的时间, 训练完成后防御过程中不需要再训练; $\mathrm{O}(T_{\mathrm{sus}})$ 取决于数据大小和图的规模; $\mathrm{O}(T_{\mathrm{mod}})$ 取决于删除率.

4　实验

在本节中, 为了评估 Motif-Defense 的防御能力, 在四个真实数据集上, 对六种后门攻击方法和五种后门防御方法展开防御实验. 具体而言, 本文实验旨在回答以下六个关键问题:

1) RQ1: 在面临多种未知后门攻击时, Motif-Defense 的防御性能是否最优?

2) RQ2: 模体角度数据增强方式和标签平滑策略对 Motif-Defense 防御效果的影响?

3) RQ3: Motif-Defense 如何影响模型实现防御效果?

4) RQ4: Motif-Defense 的时间复杂度如何?

5) RQ5: Motif-Defense 在面临自适应攻击时是否仍然有效?

6) RQ6: Motif-Defense 的性能是否受其超参数选择的影响?

4.1　实验设置

4.1.1　数据集及目标模型

本节在四个广泛使用的真实数据集上评估 Motif-Defense 的防御性能, 分别为生物信息学的 PROTEINS[11]、来自小分子的 AIDS[10]、来自小分子的 NCI1[12] 和来自社交网络的 DBLP_v1[12]. 基本统计数据总结在表 1 中 (图标签分布列的 [0] 和 [1] 分别表示标签为 0 和标签为 1 的图样本数量, 如 DBLP_v1 包含 9530 个 0 类图和 9926 个 1 类图). GIN (graph isomorphism network)[41] 模型通过设计具有不变性和区分性的聚合函数来更新节点特征, 从而有效捕捉图的结构信息. GIN 模型在节点分类和图分类任务中表现出色, 因其结构简单且参数效率高, 能够在不牺牲性能的情况下处理较大规模的图数据. 所以本文选择 GIN 模型作为攻击的目标模型.

4.1.2　后门攻击及对比防御方法

本节选取了六种攻击性能最优的后门攻击方法和五种前沿的后门防御方法. 其中防御方法为 Prune[13]、BloGGaD[16]、ES[17]、CLB-Defense[18] 和 MD-GNN[19], 攻击方法为 Erdős–Rényi backdoor (ER-B)[8]、most important nodes selecting attack (MIA)[9]、MaxDCC[11]、graph trojaning attack (GTA)[10]、Motif-Backdoor[12] 以及 unnoticeable graph backdoor attack (UGBA)[13]. 关于上述六种攻击方法已在第 1.1 节中进行详细阐述, 故不再赘述, 下面主要对上述五种防御方法进行简要介绍:

Prune[13]**.** 通过修剪连接低余弦相似度节点的边来识别和移除后门触发器, 从而降低攻击的成功率.

BloGGaD[16]**.** 通过高斯混合模型识别具有异常分布的触发节点, 并利用聚类算法识别和清除触发特征.

ES[17]**.** 利用解释性指标来区分良性样本和恶意样本. 通过计算验证数据集上的解释性得分并将其用作检测阈值来识别恶意样本. 然后, 使用可解释性方法找到并删除触发器.

CLB-Defense[18]**.** 采用对比学习模型以及图重要性指标去除训练数据集中的扰动, 从而实现对图后门攻击的防御.

MD-GNN[19]**.** 研究了扰动边和正常边在度量空间上的特征差异, 并根据基于 Jaccard 相似度的度量值准确地识别中毒边, 同时尽可能地保存图拓扑信息.

4.1.3　评价指标

为了准确地评估 Motif-Defense 的防御效果, 本文使用攻击成功率 (attack success rate, ASR) 与主任务准确率 (accuracy, ACC) 作为评估指标, 其定义如下:

$$\mathrm{ASR}=\frac{N_{\mathrm{suc}}}{N_{\mathrm{att}}} \quad (12)$$

表 1　数据集基本统计数据

Table 1　The basic statistics of datasets

数据集	图样本数	节点数	链路数	图标签分布	目标类	网络类型
PROTEINS	1113	39.06	72.82	663 [0], 450 [1]	1	生物信息
AIDS	2000	15.69	16.20	400 [0], 1600 [1]	0	小分子
NCI1	4110	29.87	32.30	2053 [0], 2057 [1]	0	小分子
DBLP_v1	19456	10.48	19.65	9530 [0], 9926 [1]	0	社交网络

式中, N_{suc} 是被成功攻击样本的数量, N_{att} 表示被攻击样本的总数. ASR 直接反映了攻击者是否成功地在目标模型中植入了后门, 高 ASR 表明攻击策略有效, 攻击者可以成功控制模型行为.

$$\text{ACC} = \frac{N_{\text{cor}}}{N_{\text{total}}} \tag{13}$$

式中, N_{cor} 表示模型正确预测的样本数量, N_{total} 表示模型预测的样本总数. ACC 用于评估模型在正常任务上的表现, 在实施防御措施后, 如果 ACC 保持较高水平, 表明防御策略未显著影响模型的整体性能.

4.1.4 实验设置

本文实验采用十折交叉验证法, 即将数据集随机划分为 10 个相同大小的子集, 每次选取其中 1 个子集作为测试集, 另外 9 个子集作为训练集. 攻击方法设置中, 中毒率为训练数据的 10%, 考虑到后门攻击的隐蔽性, 触发大小设置为 4 个节点以内. ER-B 采用随机图模型生成一个子图作为触发器, 其触发密度设为 0.8. GTA 采用三层全连接神经网络作为触发器生成器. 防御方法设置中, Prune 的连边删减率设置为 10%, BloGGaD 中节点分布类别和聚类类别数都设置为 2, 其余方法采用原文中默认的参数设置. 使用学习率为 0.01 的 Adam 优化器对 GIN 模型进行训练. Motif-Defense 中差异值阈值设置为 0.5, 相似度阈值在满足删除 10% 连边要求下进行动态调整, 标签平滑率设置为 0.5. 实验环境的具体配置如下: CPU 型号为 Intel Xeon E3-1231 v3@3.40 GHz (4 核 8 线程), GPU 型号为 TITAN Xp 12 GiB × 2, 运行内存为 32 GB × 4, 操作系统为 Ubuntu 16.04, 编程语言为 Python 3.7, 使用的深度学习框架为 PyTorch-1.4.0, 主要的深度学习包为 Torchvision-0.5.0.

4.2 实验结果与分析

4.2.1 RQ1: Motif-Defense 的防御性能分析

为了验证 Motif-Defense 的防御性能, 本文选择了四个真实数据集和六种后门攻击方法, 评估 Motif-Defense 与五种对比防御算法的防御性能, 具体实验结果见表 2, 其中加粗的数据表示效果最优. 以下是对表 2 结果的详细分析.

首先, PROTEINS、AIDS、NCI1、DBLP_v1 四个数据集在 GIN 模型上的原始 ACC 分别为 76.23%、98.92%、80.41%、80.83%. MIA、GTA、ER-B、MaxDCC、Motif-Backdoor、UGBA 六种攻击方法在四个数据集上, 在无防御的情况下平均 ASR 和 ACC 分别为 86.00% 和 79.42%.

从面对后门攻击时的防御效果方面分析, Motif-Defense 的表现在大多数情况下都优于其他五种对比防御方法, 尤其在 PROTEINS 和 AIDS 数据集上, 在所有攻击场景下始终优于所有对比防御方法. 具体来说, Motif-Defense 在四种数据集上 ASR 平均值为 13.16%, 平均降低 84.70% 的攻击成功率. 而 Prune、BloGGaD、MD-GNN、CLB-Defense、ES 的 ASR 平均值分别为 43.97%、20.18%、27.70%、19.77%、22.33%. 对此, 本文分析 Motif-Defense 优异的防御效果得益于以下两个原因: 1) 模体增强的对比学习模型能够更有效地学习图的表示, 通过选择数据集中不存在的模体结构进行数据增强, 对比学习模型能够学习到更丰富的图结构信息, 并构建更具区分度的节点嵌入. 这使得 Motif-Defense 在预测时能够更准确地识别出后门样本, 即使触发器隐藏在复杂的图结构中, 也能被识别出来. 2) 标签平滑策略能够减轻攻击中篡改训练数据标签带来的负面影响, Motif-Defense 利用对比学习模型和目标模型输出的置信分数, 通过加权平均的方式得到平滑后的置信分数, 并以此得到可疑后门样本的新标签. 这种标签平滑策略能够减少模型对错误标签的依赖, 从而提高模型的鲁棒性. 此外, Motif-Defense 在 AIDS 数据集上表现最好, MIA、GTA、ER-B、MaxDCC、Motif-Backdoor、UGBA 的 ASR 分别降低了 99.26%、95.85%、90.28%、92.35%、89.25%、90.14% (相比无防御). 本文分析其原因为 AIDS 数据集节点特征丰富且图结构相对简单, 这使得 Motif-Defense 能够更好地发挥其使用的对比学习机制, 除此之外, 图结构的简单性减少了图数据中噪声和冗余信息对 Motif-Defense 的影响, 使其能够更专注于学习图结构和节点特征之间的关系.

从防御后目标模型的整体性能方面分析, Motif-Defense 的表现在大多数情况下同样优于其他五种对比防御方法, 尤其在 AIDS 和 DBLP_v1 数据集上, 在所有攻击场景下始终优于所有对比防御方法. 具体来说, Motif-Defense 在四种数据集上 ACC 平均值为 82.05%, 分类准确率平均仅下降 2.53%. 而 Prune、BloGGaD、MD-GNN、CLB-Defense、ES 的 ACC 平均值分别为 79.51%、79.87%、79.04%、81.60%、80.45%. 这表明 Motif-Defense 在保持模型分类准确率方面具有显著优势, 有效地平衡了防御性能和主任务性能. 本文认为原因在于 Motif-Defense 采用了基于模体角度的数据增强方式, 使得对比学习模型能够学习到更加丰富的图结构信息, 从而在识别可疑样本时更加准确, 减少对正常样本的影响. 同时, Motif-Defense 利用 Jaccard 相似度来识别中毒边, 并删除相似度较低的节点对之

表 2　不同攻击场景下 Motif-Defense 的防御性能 (%)

Table 2　The defense performance of Motif-Defense in different attack scenarios (%)

数据集	评价指标	防御方法	后门攻击方法					
			MIA	GTA	ER-B	MaxDCC	Motif-Backdoor	UGBA
PROTEINS (76.23)	ASR	无防御	68.07	95.80	72.23	93.28	94.12	98.94
		Prune	20.56	18.63	28.57	10.76	79.85	80.69
		BloGGaD	20.67	9.75	13.92	11.26	24.87	23.63
		MD-GNN	22.35	16.47	18.15	30.00	25.43	16.64
		CLB-Defense	4.86	3.58	9.39	15.86	11.76	8.96
		ES	13.61	12.35	9.71	11.76	16.64	15.37
		Motif-Defense	**4.03**	**2.26**	**8.93**	**9.07**	**10.92**	**4.58**
	ACC	无防御	66.82	65.02	64.13	66.37	74.89	73.95
		Prune	73.45	72.38	72.56	73.00	**76.23**	72.72
		BloGGaD	73.90	72.56	72.56	70.76	75.93	74.52
		MD-GNN	66.69	69.23	65.47	69.06	71.33	70.97
		CLB-Defense	73.35	72.44	72.24	**74.39**	75.03	71.45
		ES	71.57	71.30	68.43	72.55	76.23	71.14
		Motif-Defense	**73.69**	**73.93**	**73.49**	69.24	72.11	**74.69**
AIDS (98.92)	ASR	无防御	92.19	99.34	80.94	91.88	98.75	96.72
		Prune	38.99	47.33	35.63	22.43	39.63	93.47
		BloGGaD	14.75	20.75	14.63	9.19	19.87	22.41
		MD-GNN	16.87	25.19	16.00	23.62	27.75	19.25
		CLB-Defense	23.99	17.56	25.38	8.13	20.75	15.84
		ES	1.21	9.63	4.33	11.33	19.63	18.25
		Motif-Defense	**0.68**	**4.12**	**7.87**	**7.03**	**10.62**	**9.54**
	ACC	无防御	94.00	98.25	94.00	95.25	99.00	98.65
		Prune	96.50	96.33	96.25	82.33	97.45	94.93
		BloGGaD	96.30	96.25	95.25	97.10	97.20	95.92
		MD-GNN	95.45	94.47	94.75	95.85	98.11	96.38
		CLB-Defense	97.45	97.07	96.57	97.86	98.16	97.05
		ES	95.45	95.25	96.10	97.00	98.15	96.72
		Motif-Defense	**97.55**	**97.15**	**97.55**	**97.95**	**98.25**	**97.65**
NCI1 (80.41)	ASR	无防御	96.98	100.00	98.33	100.00	100.00	100.00
		Prune	50.50	45.10	38.20	46.00	53.30	98.32
		BloGGaD	33.26	39.44	27.36	30.25	**20.64**	28.34
		MD-GNN	37.23	41.04	34.92	38.54	62.66	32.63
		CLB-Defense	55.36	46.85	45.39	43.58	48.96	23.09
		ES	39.57	40.31	39.46	32.07	41.00	18.36
		Motif-Defense	**32.74**	**30.38**	**25.58**	**23.58**	34.36	**13.64**
	ACC	无防御	73.36	76.52	70.80	76.89	80.41	81.45
		Prune	76.23	76.74	73.48	**77.47**	74.96	78.33
		BloGGaD	74.31	73.55	74.92	74.33	73.84	**80.84**
		MD-GNN	75.50	72.90	77.10	75.00	73.80	77.97
		CLB-Defense	76.06	76.87	75.74	76.53	75.78	79.36
		ES	75.06	77.20	73.58	77.13	75.11	80.33
		Motif-Defense	**77.41**	**78.99**	**77.88**	77.34	**76.47**	79.67

表 2 不同攻击场景下 Motif-Defense 的防御性能 (%) (续表)

Table 2 The defense performance of Motif-Defense in different attack scenarios (%) (continued table)

数据集	评价指标	防御方法	后门攻击方法					
			MIA	GTA	ER-B	MaxDCC	Motif-Backdoor	UGBA
DBLP_v1 (80.83)	ASR	无防御	48.75	62.17	62.29	69.86	70.84	72.56
		Prune	19.20	11.50	25.00	23.00	60.50	68.00
		BloGGaD	**8.60**	10.90	19.20	17.50	24.00	19.20
		MD-GNN	18.40	20.10	32.77	24.00	29.50	35.40
		CLB-Defense	10.33	10.28	15.28	13.66	24.39	18.52
		ES	18.80	10.36	18.89	28.25	31.00	26.79
		Motif-Defense	14.21	**6.60**	**13.67**	**8.75**	**22.33**	**10.35**
	ACC	无防御	73.46	67.78	79.52	76.23	78.85	80.36
		Prune	76.00	75.20	74.05	78.00	73.70	70.03
		BloGGaD	75.90	72.80	76.50	74.10	72.50	75.03
		MD-GNN	77.20	73.60	75.80	79.30	74.70	76.32
		CLB-Defense	79.52	79.33	79.62	79.78	78.54	78.14
		ES	77.92	77.72	77.01	77.84	75.11	76.94
		Motif-Defense	**79.68**	**79.39**	**79.87**	**79.96**	**80.24**	**79.11**

注: 数据集列括号内数值为 GIN 模型原始准确率.

间的连边, 有效地消除了触发器, 避免了误删正常边, 从而在去除触发器的同时, 最大程度地保留了图的结构信息, 进一步提升了模型的整体性能. 此外, 其他对比防御方法缺乏针对标签的优化策略, 容易导致模型性能下降.

4.2.2 RQ2: 消融实验

为了进一步探究 Motif-Defense 中不同模块对防御效果的影响, 本节进行了消融实验. 具体而言, 将 Motif-Defense 中的两个关键模块, 即基于模体增强的对比学习模型和标签平滑策略, 分别进行消融, 以验证其对防御成功率的影响. 本节将完整的 Motif-Defense 记为 MD, 采用随机删除连边、丢弃节点的增强方式和不采用标签平滑策略的 Motif-Defense 分别记为 MD^{NO_M} 和 MD^{NO_S}. 在 Motif-Backdoor 攻击方法和四个数据集 (PROTEINS、AIDS、NCI1、DBLP_v1) 上进行了实验, 记录攻击成功率. 实验结果如图 3 所示. 在无防御情况下, Motif-Backdoor 在四个数据集上的平均 ASR 为 92.06%, 在经过 MD、MD^{NO_M} 以及 MD^{NO_S} 的防御后, 平均 ASR 分别降为 12.91%、20.94%、35.81%. 因此, 得到两个结论: 一是两种简化版的防御方法虽然都能够在一定程度上降低攻击成功率, 但效果明显不如 Motif-Defense; 二是两者相比来说, MD^{NO_M} 展现出较好的防御性能. 本文将主要原因总结为两点: 一是单一防御策略局限性较高, 只有模体增强和标签平滑策略共同使用, 才能有效地破坏触发器并减轻错误标签的影响, 实现最佳的防御效果. 二

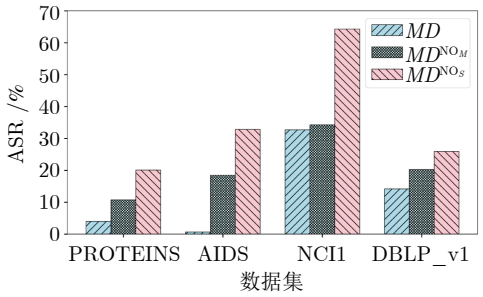


图 3 Motif-Defense 消融实验结果

Fig. 3 Ablation experiment results for Motif-Defense

是标签平滑策略通过优化可疑样本的标签, 使其更接近真实类别, 从而减少模型对错误标签的依赖. 这不仅能够有效防止模型过拟合训练数据中的噪声, 还能降低攻击者通过修改标签对模型进行误导的可能性. 而且 MD^{NO_M} 虽然没有使用模体角度的数据增强方式, 但保留了对比学习模型, 所以防御效果优于 MD^{NO_S}.

4.2.3 RQ3: 决策边界及神经元激活值可视化分析

为了更直观地理解 Motif-Defense 的防御效果, 本节进行了可视化分析. 首先, 对决策边界的变化进行分析, 通过对比干净样本、中毒样本和防御后样本的分类概率分布, 展示了 Motif-Defense 如何将中毒样本从目标类重新分类为正常类, 恢复了模型的决策边界. 其次, 可视化了模型的最后一层神经元输出, 观察了干净样本、中毒样本和防御后样本在神经元激活上的差异, 进一步验证了 Motif-

Defense 的有效性.

1) 为了展现 Motif-Defense 带来的模型决策边界变化, 随机选择 100 个干净样本, 使用 GTA 攻击对选择的样本进行中毒得到中毒样本, 再通过 Motif-Defense 对中毒样本进行防御得到净化样本. 将三类样本分别用于输入目标、中毒、防御后的模型训练, 每个模型训练 100 epochs, 记录样本分类为正常样本的平均概率分布. 选择 GIN 模型和 AIDS 数据集, 结果如图 4 所示. 结果显示, 经过攻击后, 中毒样本会被分类为目标类, 但经过 Motif-Defense 防御后, 中毒样本会被逐步分为正常类, 该结果充分说明了 Motif-Defense 在决策边界上的有效性. 攻击后的模型将中毒样本错误地归类为目标类, 这是因为攻击者在样本中注入了触发器, 导致模型将触发器与目标类标签建立了关联. 而经过 Motif-Defense 防御后, 模型能够正确地将中毒样本归类为正常类, 这是因为防御方法有效地移除了触发器, 并纠正了错误标签, 从而恢复了模型的正常决策边界.

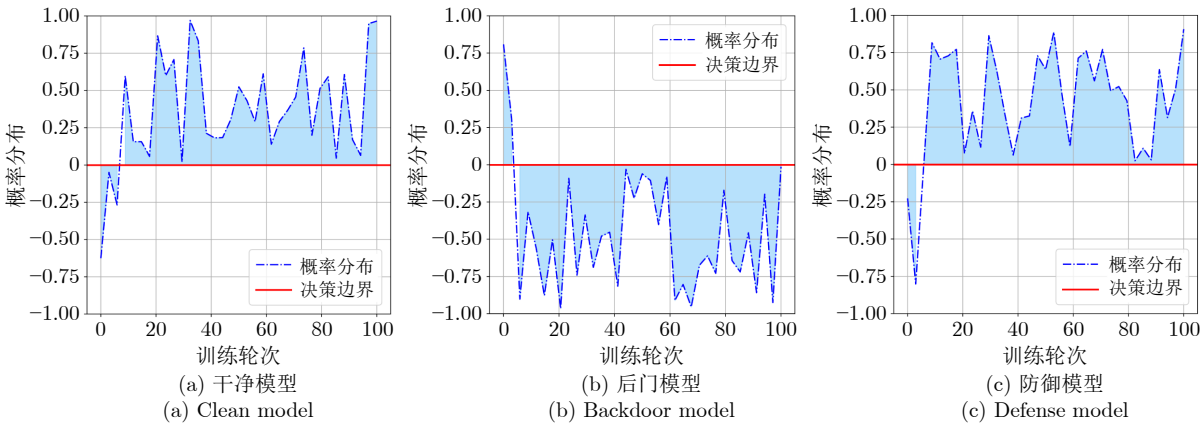
2) 本文对模型最后一层神经元输出的分布进行了可视化, 具体来说, 分别对攻击前、攻击后、经过防御后节点样本的神经元激活情况进行了可视化, 分别对应图 5 中干净、后门、防御测试集的实验结果. 攻击方法选用 GTA, 数据集使用 PROTEINS、AIDS、NCI1. 图中颜色越深代表神经元对图数据的分类结果越重要. 以 PROTEINS 为例进行分析, 从干净测试集的结果可以看出, 3 号和 8 号神经元是较为重要的神经元, 但经过 GTA 攻击后, 3 号和 8 号神经元上的颜色发生了明显变化, 即中毒模型中神经元对后门样本的激活发生了显著的改变, 从而导致攻击成功. 后续, 经过 Motif-Defense 防御后, 中毒样本被修复, 模型对样本的激活分布又重新集中于 3 号和 8 号神经元, 此结果说明经过防御后, 模型的性能得到了恢复. 具体而言, 经过 Motif-Defense 防御后, 模型重要神经元的激活情况显著改善, 有效地减少了对后门样本的过度激活, 帮助模型重新聚焦于重要特征.

4.2.4　RQ4: 时间复杂度分析

为了验证 Motif-Defense 的时间复杂度分析, 本节对 Motif-Defense 各步骤的运行时间和占比进行统计和分析. 其中对于 $O(T_{\text{motif}})$, 只查找 M_{31} 到 M_{46} 的模体分布情况, 且只采样 10% 的图数据进行分析. 相对于 $O(T_{\text{motif}})$ 和 $O(T_{\text{cl}})$, $O(T_{\text{sus}})$ 和 $O(T_{\text{mod}})$ 的时间消耗小得多, 所以本节将两者时间消耗计算在一起. 在 Motif-Backdoor 攻击下进行实验, 结果如图 6 所示. 结果显示, 可疑样本识别和修改时间占比最少, 这与第 3.4 节的分析一致. 模体分析和构建对比学习模型在 Motif-Defense 防御过程中时间占比最多, 在四个数据集上平均占比分别为 44.51%、52.59%. 主要原因在于, 模体分析需要对图数据进行深度挖掘, 识别图中重复出现、统计上显著的子图模式, 即便已经采用了限制模体大小和图采样操作, 但并不能完全消除子图枚举的复杂度. 此外, 构建对比学习模型需要对样本进行增强操作, 会对图结构进行修改, 并重新计算节点特征和邻接矩阵, 计算量较大.

4.2.5　RQ5: 自适应攻击

为评估 Motif-Defense 在攻击者知晓防御机制情况下的鲁棒性, 本节设计了一种基于模体结构的自适应攻击 (MIA-Adaptive). 该攻击假设攻击者已知 Motif-Defense 依赖模体结构进行数据增强, 因此主动选择模型最可能利用的模体结构作为触发器, 并结合 GNNs 解释性方法优化注入位置, 以规



(a) 干净模型　(b) 后门模型　(c) 防御模型
(a) Clean model　(b) Backdoor model　(c) Defense model

图 4　决策边界可视化

Fig. 4　Visualization of decision boundary change

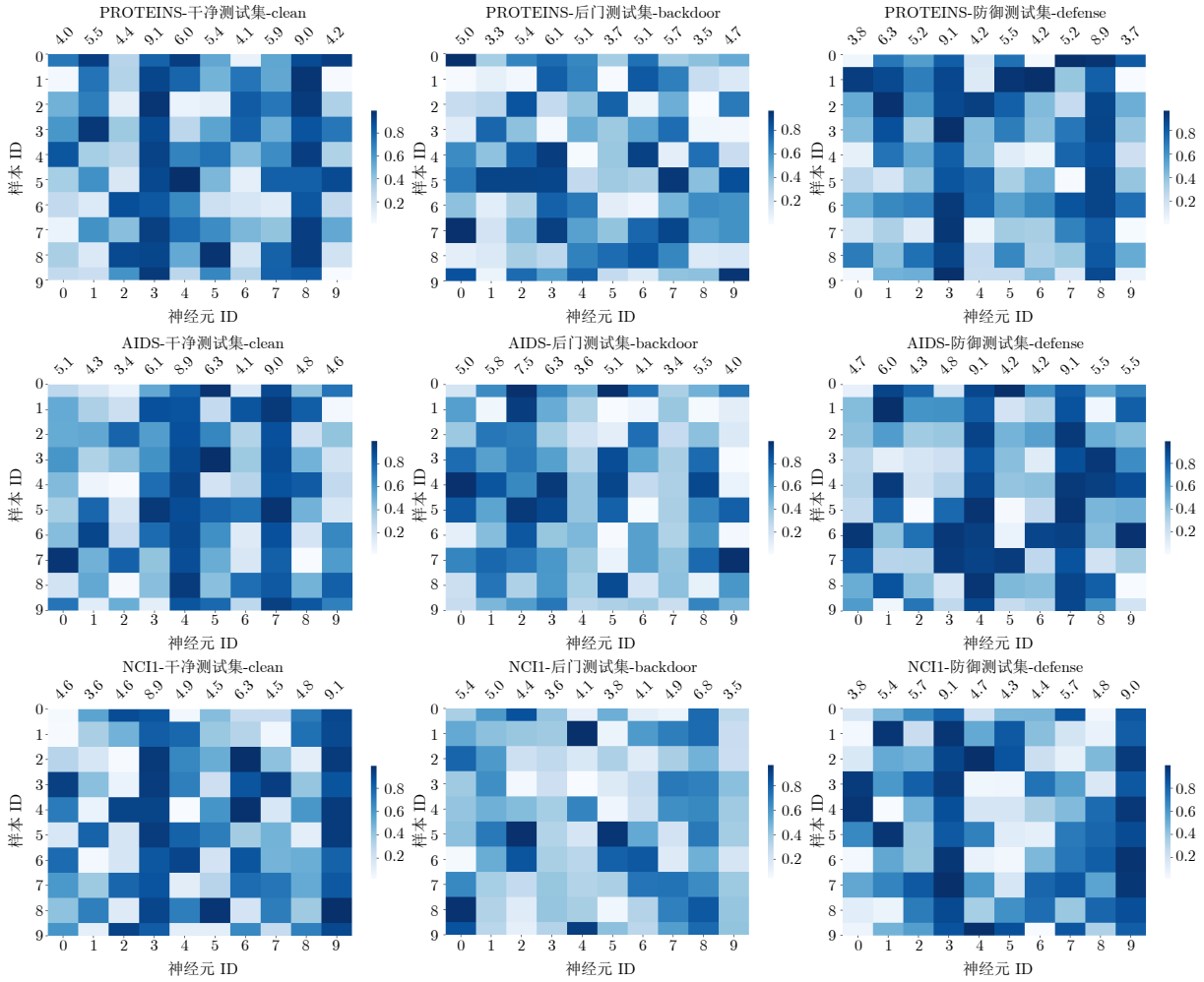


图 5 Motif-Defense 对后门样本神经元激活的影响

Fig. 5 The impact of Motif-Defense on the activation of neurons in backdoor samples

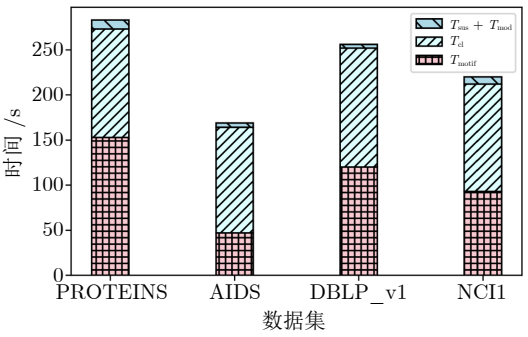


图 6 Motif-Defense 时间复杂度分析

Fig. 6 Time complexity analysis of Motif-Defense

避检测.

首先, 依照第 3.1 节中的方法对数据集进行模体分析, 选择 Motif-Defense 最有可能利用的模体结构作为触发器结构. 然后, 采用 MIA 使用的 GNNs 解释器分析目标模型的预测结果, 确定触发器注入位置. 最后, 根据触发器注入位置和模体结构生成触发器, 将其嵌入到良性样本中并修改标签. 采用 GIN 作为目标模型在四个数据集上进行实验, 其中 MIA-Adaptive 表示使用该方法进行攻击, MIA-Adaptive-Motif 表示使用 Motif-Defense 进行防御, 结果如图 7 所示. 结果显示, 在经过 Motif-Defense 防御后, 在 PROTEINS、AIDS、NCI1、DBLP_v1 数据集上攻击成功率分别降低了 91.10%、97.06%、61.14%、74.54%. 从防御结果来说, Motif-Defense 有效地抵御了本节提出的自适应攻击. 其原因是 Motif-Defense 的防御机制不仅仅依赖于模体结构的识别和增强. 在对抗本节提出的自适应攻击时, 发现模型在接受了基于模体结构的增强训练后, 对于图结构的变化更加敏感, 从而能够更好地识别和抵御经过精心设计的恶意模体结构. 此外, 在 Motif-Defense 中还采用随机删减连边和丢弃节点的数据增强方式, 该方式同样能够破坏中毒数据中存在的触发器, 从而实现有效防御.

上述结果表明, 即使攻击者构造出基于模体结

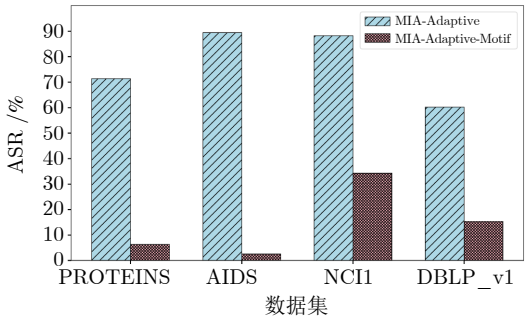


图 7　Motif-Defense 在面对自适应攻击下的表现

Fig. 7　The performance of Motif-Defense against adaptive attacks

构的隐蔽触发器, Motif-Defense 仍能有效识别并净化后门样本, 在自适应攻击场景下保持稳定的防御性能. 结合第 4.2.1 节对 Motif-Backdoor 攻击的防御结果 (相较于无防御方法, ASR 平均降低 77.94%), Motif-Defense 不仅适用于常规子图触发器攻击, 也能有效抵御以模体为基础的未知攻击. 这表明 Motif-Defense 并非局限于已知攻击方法, 也能够适用于基于模体的未知或新型攻击场景.

4.2.6　RQ6: 参数敏感性分析

为深入理解关键超参数 o_1 对 Motif-Defense 性能的影响, 并评估其鲁棒性, 本节在 AIDS 数据集上使用 GTA 攻击 (中毒率 10%) 训练后门模型 GIN. 将阈值 o_1 在 {0.3, 0.4, 0.5, 0.6, 0.7} 范围内调整, 并观察其对攻击成功率 (ASR)、模型准确率 (ACC)、检测率 (detection rate, DR) 和误检率 (false detection rate, FDR) 的影响. 每个实验重复 5 次, 结果取平均值. 其中, DR 和 FDR 分别表示被成功识别的中毒样本占所有中毒样本的比例以及被错误识别的正常样本占所有正常样本的比例. 结果如表 3 所示, 随着阈值 o_1 的增大, ASR 呈现先降后升的 U 型趋势, 在 $o_1 = 0.5$ 时达到最低 (4.12%); ACC 则持续上升至 97.86%; DR 和 FDR 均随 o_1 增大而下降. 其原因在于, 当 o_1 过低时, 宽松的阈值导致大量正常样本被误判 (FDR 高), 不必要的净化操作扰动了正常样本, 损害 ACC 且未能最优消除后门; 当 o_1 过高时, 严格的阈值使大量真实后门样本漏检 (DR 低), 导致 ASR 回升, 但因处理样本少, ACC 接近最高. 在 $o_1 = 0.5$ 时, 系统在检测真实后门 (DR = 95.84%) 与保护正常样本 (FDR = 1.53%) 之间达到最佳平衡, 实现了 ASR 最低与 ACC 较高的理想状态, 验证了方法的有效性.

需要指出的是, 可疑样本的识别依赖于其置信度差异 Δ_i (即后门模型对目标类与其他类预测置信度的差值), 而 Δ_i 的计算结果可能受模型训练过程中的随机性 (如参数初始化、数据采样顺序等) 影响而产生波动. 这意味着, 一个正常样本的 Δ_i 值在不同训练实例中可能略有不同. 如果该值因波动而偶然超过阈值 o_1, 就会导致误判, 从而影响 FDR. 然而, 如表 3 所示, FDR 随 o_1 的变化趋势是平滑且可预测的. 将 o_1 设置在经过验证的最优区间内, 系统性能对 Δ_i 的适度波动具有良好的鲁棒性, 整体误判风险是可控的.

表 3　不同阈值 o_1 下的详细性能结果 (AIDS + GTA) (%)

Table 3　The performance results under different threshold o_1 (AIDS + GTA) (%)

o_1	ASR	ACC	DR	FDR
0.3	8.53	95.81	99.00	6.81
0.4	5.27	96.59	97.51	3.02
0.5	4.12	97.15	95.84	1.53
0.6	6.81	97.45	92.07	0.82
0.7	10.34	97.86	86.23	0.45

5　结束语

针对现有图神经网络后门防御方法存在的不足, 例如对不同后门攻击的防御泛化性弱、无法有效均衡主任务性能与防御成功率等问题, 本文提出一种基于模体增强对比学习的图神经网络后门攻击防御方法, 称为 Motif-Defense. 该方法首先通过模体分析构建对比学习模型. 然后利用对比学习模型和目标模型输出的置信分数之差, 筛选出可疑后门样本. 最后利用 Jaccard 相似度以及标签平滑策略对可疑后门样本进行修改. 实验结果表明, Motif-Defense 在多种攻击场景下能够显著降低攻击成功率, 同时对模型分类准确率的影响较小.

然而, Motif-Defense 仍存在一些实际部署中的挑战. 首先, 在模体分析阶段, 尽管通过限制模体规模和对训练数据进行采样来控制计算复杂度, 该过程仍涉及子图枚举, 在面对超大规模图时可能面临性能瓶颈. 此外, 模体分析与对比学习模型的构建为离线预处理步骤, 虽不影响在线推理效率, 但在资源受限场景下仍需进一步优化. 为此, 未来工作将探索基于分布式计算的近似模体计数方法, 以提升算法的可扩展性.

另外, 方法中部分关键环节依赖阈值设置. 例如, 可疑后门样本的识别基于对比模型与目标模型输出置信度的差异, 其判定阈值目前依赖经验设定; 边剪枝操作中的相似度阈值虽采用删除 10% 连边的动态策略以适配不同图结构, 但仍需人为指定删除比例. 对预设阈值的依赖可能削弱方法在不同数据分布下的鲁棒性与可迁移性. 因此, 后续研究将

探索自适应机制, 例如利用验证集上的模型性能或置信度分布的统计特性 (如均值与标准差) 自动确定阈值, 从而减少对人工经验的依赖, 推动防御系统的自动化部署.

最后, 本文的研究集中于图分类任务, 所提出的可疑样本识别与处理策略依赖于图级表示. 对于节点分类、链接预测等任务, 其输入输出形式与图分类任务存在差异, 因此 Motif-Defense 需进行相应调整. 例如, 在节点分类中, 可基于局部子图构建对比视图, 并利用节点级置信度识别可疑节点; 在异构图中, 可基于元路径引导的子图采样设计增强策略. 这些扩展方向将作为未来的重要研究内容, 以推动 Motif-Defense 向更广泛的图学习任务迁移.

参考文献

1 Hamilton W L, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 1025−1035

2 Wang D X, Lin J B, Cui P, Jia Q H, Wang Z, Fang Y M, et al. A semi-supervised graph attentive network for financial fraud detection. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE, 2019. 598−607

3 Bai L, Yao L N, Kanhere S, Wang X Z, Liu W, Yang Z. Spatio-temporal graph convolutional and recurrent networks for city-wide passenger demand prediction. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: ACM, 2019. 2293−2296

4 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017. 1−14

5 Cui H J, Dai W, Zhu Y Q, Kan X, Gu A A C, Lukemire J, et al. BrainGB: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 2023, **42**(2): 493−506

6 Stokes J M, Yang K, Swanson K, Jin W G, Cubillos-Ruiz A, Donghia N M, et al. A deep learning approach to antibiotic discovery. *Cell*, 2020, **180**(4): 688−702

7 Yumlembam R, Issac B, Jacob S M, Yang L Z. IoT-based android malware detection using graph neural network with adversarial defense. *IEEE Internet of Things Journal*, 2023, **10**(10): 8432−8444

8 Zhang Z X, Jia J Y, Wang B H, Gong N Z. Backdoor attacks to graph neural networks. In: Proceedings of the 26th ACM Symposium on Access Control Models and Technologies. New York, USA: ACM, 2021. 15−26

9 Xu J, Xue M H, Picek S. Explainability-based backdoor attacks against graph neural networks. In: Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. Abu Dhabi, United Arab Emirates: ACM, 2021. 31−36

10 Xi Z H, Pang R, Ji S L, Wang T. Graph backdoor. In: Proceedings of the 30th USENIX Security Symposium. Berkeley, USA: USENIX Association, 2021. 1523−1540

11 Sheng Y, Chen R, Cai G Y, Kuang L. Backdoor attack of graph neural networks based on subgraph trigger. In: Proceedings of the 17th EAI International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2021). Virtual Event: Springer, 2021. 276−296

12 Zheng H B, Xiong H Y, Chen J Y, Ma H N, Huang G H. Motif-backdoor: Rethinking the backdoor attack on graph neural networks via motifs. *IEEE Transactions on Computational Social Systems*, 2024, **11**(2): 2479−2493

13 Dai E Y, Lin M H, Zhang X, Wang S H. Unnoticeable backdoor attacks on graph neural networks. In: Proceedings of the ACM Web Conference. Austin, USA: ACM, 2023. 2263−2273

14 Alon U. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 2007, **8**(6): 450−461

15 Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In: Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses (RAID). Heraklion, Crete, Greece: Springer, 2018. 273−294

16 Yang X, Li G L, Tao X Y, Zhang C F, Li J H. Black-box graph backdoor defense. In: Proceedings of the 23rd International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP). Tianjin, China: Springer, 2023. 163−180

17 Jiang B C, Li Z. Defending against backdoor attack on graph nerual network by explainability. arXiv preprint arXiv: 2209.02902, 2022.

18 Chen Jin-Yin, Xiong Hai-Yang, Ma Hao-Nan, Zheng Ya-Yu. CLB-Defense: Based on contrastive learning defense for graph neural network against backdoor attack. *Journal on Communications*, 2023, **44**(4): 154−166
(陈晋音, 熊海洋, 马浩男, 郑雅羽. 基于对比学习的图神经网络后门攻击防御方法. 通信学报, 2023, **44**(4): 154−166)

19 Xiao Y, Li J, Su W G. A lightweight metric defence strategy for graph neural networks against poisoning attacks. In: Proceedings of the 23rd International Conference on Information and Communications Security (ICICS). Chongqing, China: Springer, 2021. 55−72

20 You Y N, Chen T L, Sui Y D, Chen T, Wang Z Y, Shen Y. Graph contrastive learning with augmentations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 488

21 Qiu J Z, Chen Q B, Dong Y X, Zhang J, Yang H X, Ding M, et al. GCC: Graph contrastive coding for graph neural network pre-training. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2020. 1150−1160

22 Hassani K, Khasahmadi A H. Contrastive multi-view representation learning on graphs. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 4116−4126

23 Yu J L, Yin H Z, Xia X, Chen T, Cui L Z, Nguyen Q V H. Are graph augmentations necessary? Simple graph contrastive learning for recommendation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain: ACM, 2022. 1294−1303

24 Cai X H, Huang C, Xia L H, Ren X B. LightGCL: Simple yet effective graph contrastive learning for recommendation. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023. 1−15

25 Xu B W, Wang X L, Liu Z J, Kang L W. A GAN combined with graph contrastive learning for traffic forecasting. In: Proceedings of the 4th International Conference on Computing, Networks and Internet of Things. Xiamen, China: ACM, 2023. 866−873

26 Sankar A, Zhang X Y, Chang K C C. Motif-based convolutional neural network on graphs. arXiv preprint arXiv: 1711.05697, 2017.

27 Yang C, Liu M X, Zheng V W, Han J W. Node, motif and subgraph: Leveraging network functional blocks through structural convolution. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona, Spain: IEEE, 2018. 47−52

28 Zhao H, Zhou Y Q, Song Y Q, Lee D K. Motif enhanced recom-

mendation over heterogeneous information network. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: ACM, 2019. 2189–2192

29 Dareddy M R, Das M, Yang H. Motif2vec: Motif aware node representation learning for heterogeneous networks. In: Proceedings of the IEEE International Conference on Big Data (Big Data). Los Angeles, USA: IEEE, 2019. 1052–1059

30 Shao P, Yang Y, Xu S Y, Wang C P. Network embedding via motifs. *ACM Transactions on Knowledge Discovery From Data*, 2021, **16**(3): Article No. 44

31 Zhao M, Zhang Y L, Xia X W, Xu X. Motif-aware adversarial graph representation learning. *IEEE Access*, 2022, **10**: 8617–8626

32 Wang L, Ren J, Xu B, Li J X, Luo W, Xia F. MODEL: Motif-based deep feature learning for link prediction. *IEEE Transactions on Computational Social Systems*, 2020, **7**(2): 503–516

33 Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. *Science*, 2002, **298**(5594): 824–827

34 Hočevar T, Demšar J. A combinatorial approach to graphlet counting. *Bioinformatics*, 2014, **30**(4): 559–565

35 Freeman L C. Centrality in social networks conceptual clarification. *Social networks*, 1978, **1**(3): 215–239

36 Zhang Chong-Sheng, Chen Jie, Li Qi-Long, Deng Bin-Quan, Wang Jie, Chen Cheng-Gong. Deep contrastive learning: A survey. *Acta Automatica Sinica*, 2023, **49**(1): 15–39
(张重生, 陈杰, 李岐龙, 邓斌权, 王杰, 陈承功. 深度对比学习综述. 自动化学报, 2023, **49**(1): 15–39)

37 He K M, Fan H Q, Wu Y X, Xie S N, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9726–9735

38 Freeman L C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, **40**(1): 35–41

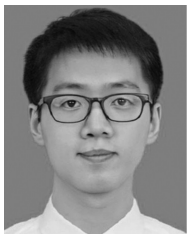
39 Bonacich P. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 1972, **2**(1): 113–120

40 Li Xiao-Qing, Tang Hao, Si Jia-Sheng, Miao Gang-Zhong. An improved semi-supervised FCM clustering method for mixed data sets. *Acta Automatica Sinica*, 2018, **44**(12): 2259–2268
(李晓庆, 唐昊, 司加胜, 苗刚中. 面向混合属性数据集的改进半监督 FCM 聚类方法. 自动化学报, 2018, **44**(12): 2259–2268)

41 Xu K, Hu W H, Leskovec J, Jegelka S. How powerful are graph neural networks? In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2019. 1–17

陈晋音 浙江工业大学计算机科学与技术学院教授. 主要研究方向为人工智能安全, 图数据挖掘和进化计算.
E-mail: chenjinyin@zjut.edu.cn
(**CHEN Jin-Yin** Professor at the College of Computer Science and Technology, Zhejiang University of Technology. Her research interests include artificial intelligence security, graph data mining and evolutionary computation.)

穆文博 浙江工业大学信息工程学院硕士研究生. 主要研究方向为人工智能安全, 深度学习和图数据挖掘.
E-mail: 211123030043@zjut.edu.cn
(**MU Wen-Bo** Master student at the College of Information Engineering, Zhejiang University of Technology. His research interests include artificial intelligence security, deep learning and graph data mining.)



郑海斌 浙江工业大学计算机科学与技术学院讲师. 主要研究方向为深度学习, 人工智能安全和图像识别. 本文通信作者.
E-mail: haibinzheng320@gmail.com
(**ZHENG Hai-Bin** Lecturer at the College of Computer Science and Technology, Zhejiang University of Technology. His research interests include deep learning, artificial intelligence security and image recognition. Corresponding author of this paper.)