基于对比学习的图神经网络后门攻击防御方法

陈晋音[1,2]，熊海洋[2]，马浩男[2]，郑雅羽[2]
（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）

摘 要：针对现有的后门攻击防御方法难以处理非规则的非结构化的离散的图数据的问题，为了缓解图神经网络后门攻击的威胁，提出了一种基于对比学习的图神经网络后门攻击防御方法（CLB-Defense）。具体来说，基于对比学习无监督训练的对比模型查找可疑后门样本，采取图重要性指标以及标签平滑策略去除训练数据集中的扰动，实现对图后门攻击的防御。最终，在 4 个真实数据集和 5 主流后门攻击方法上展开防御验证，结果显示 CLB-Defense 能够平均降低 75.66%的攻击成功率（与对比算法相比，改善了 54.01%）。

关键词：图神经网络；后门攻击；鲁棒性；防御；对比学习
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023074

CLB-Defense: based on contrastive learning defense for graph neural network against backdoor attack

CHEN Jinyin[1,2], XIONG Haiyang[2], MA Haonan[2], ZHENG Yayu[2]
1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: For the problem that the existing backdoor attack defense methods are difficult to deal with irregular and unstructured discrete graph data to alleviate the threat of backdoor attacks, a backdoor attack defense method for GNN based on contrastive learning was proposed, namely CLB-Defense. Specifically, a contrastive model was built by using contrastive learning in an unsupervised way, which searches suspicious backdoored samples. Then the suspicious backdoored samples were reshaped by using the graph importance indexes and the label smoothing strategy, and the defense against graph backdoor attack was realized. Finally, extensive experimental results show that CLB-Defense realizes the effect of defense performance on four public datasets and five popular graph backdoor attacks, e.g., CLB-Defense can reduce the attack success rate by an average of 75.66% (compared with the baselines, an improvement of 54.01%).

Keywords: graph neural network, backdoor attack, robustness, defense, contrastive learning

0 引言

图数据在现实生活中随处可见，如社交网络[1]、推荐系统[2]等。随着深度学习的兴起，图神经网络（GNN, graph neural network）[3]在图表示学习领域取得了显著成功，成为挖掘图数据潜在信息的有效工具之一。图神经网络的相关算法按照输入数据是否有标签，主要可以分为无监督学习和有监督学

收稿日期：2022−11−17；**修回日期：**2023−03−08
通信作者：郑雅羽，yayuzheng@zjut.edu.cn
基金项目：国家自然科学基金资助项目（No.62072406）；浙江省自然科学基金资助项目（No.LDQ23F020001）；信息系统安全国家科技重点实验室基金资助项目（No.61421110502）；浙江省重点研发计划基金资助项目（No.2022C01018）
Foundation Items: The National Natural Science Foundation of China (No.62072406), The Zhejiang Provincial Natural Science Foundation (No. LDQ23F020001), The Chinese National Key Laboratory of Science and Technology on Information System Security (No.61421110502), The Key Research and Development Program of Zhejiang Province (No.2022C01018)

习。无监督学习是一种不依赖于数据标签的学习方式。对比学习是一种无监督学习方法，在没有数据标签的情况下，通过让图神经网络模型挖掘图数据之间的相似或差异来学习图数据的一般特征，如图对比编码[4]、图对比学习[5]、对抗性图对比学习[6]等。有监督学习是能够利用数据标签进行模型训练的学习方式，往往能实现良好的表现性能。

然而，文献[7-11]的研究表明图神经网络容易遭受后门攻击的威胁。攻击者精心设计特定结构的子图作为触发器，并将触发器嵌入良性数据并修改标签，得到后门数据。目标模型利用后门数据进行训练，使目标模型留下后门，即后门模型。在推断阶段，攻击者通过调用触发器激活后门，使后门模型的输出为预先指定的标签，以实现恶意目的。图神经网络的后门攻击是一种发生在模型训练阶段的攻击方式。与推断阶段相比，训练阶段存在更多的步骤，即数据收集、数据预处理、模型构建、模型训练等。这意味着攻击者会有更多的攻击机会。随着数据海量式的增长，图神经网络有着更强大的表达能力，但会带来更大的训练成本，因此用户会选择第三方提供的预训练模型或已发布的数据集等方式来降低训练成本。这会进一步加大图神经网络遭受后门攻击的风险。如图 1 所示，面对电子商务平台的推荐系统，攻击者按照特定的规则浏览相关商品来构成触发器（用户浏览自行车和电脑等商品），然后进行目标商品的虚假购买（购买手机商品）。推荐系统收集到相关浏览信息和购买信息，并用于更新系统参数，这一过程会使推荐系统留下后门。攻击者仅需调用预先设置的触发器激活推荐系统中的后门，就能误导真实消费者进行错误的消费。因此，面对图神经网络后门攻击，提高图神经网络的鲁棒性显得至关重要。

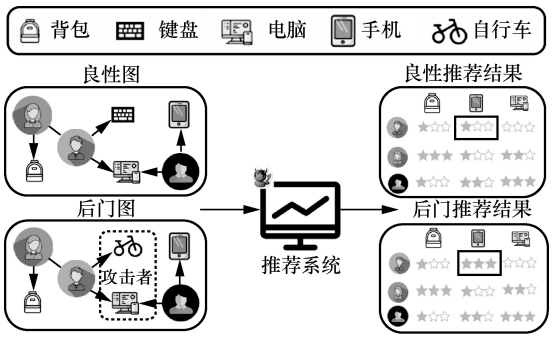


图 1　面对电子商务平台推荐系统的后门攻击示意

目前，已有的图神经网络后门攻击方法[7-11]按照触发器生成方式的不同可以划分为基于子图生成和基于梯度优化 2 种攻击方法。基于子图生成攻击方法[7-10]通过随机生成符合小世界网络、随机网络或优先连接网络分布的子图作为触发器，或者将特殊模体结构作为触发器。基于梯度优化的后门攻击方法[11]采用梯度优化生成器的方式生成有效的触发器发起后门攻击。值得注意的是，这些后门攻击方法不但向样本嵌入触发器，而且修改其标签为目标类，从而在模型训练阶段，使目标模型建立触发器与目标类之间的强相关性。

后门攻击防御方法按照防御阶段的差异可以分为两类，一类是是在训练阶段中过滤训练数据中的后门样本[12]，另一类是在推断阶段中检测样本中的触发器[13]和检测可疑模型[14]。它们是针对图像数据的后门攻击而设计的防御手段。图像是规则的、结构化的连续数据，而图是非规则的、非结构化的离散数据。这导致现有的后门攻击防御方法难以处理图神经网络的后门攻击。此外，图神经网络防御的相关研究聚焦于对抗攻击的防御。对抗攻击关注目标模型的推断阶段，而后门攻击针对目标模型的训练阶段。现有的图神经网络防御研究忽略了在训练阶段样本类标签被修改以及触发器被嵌入的问题，从而难以抵御图神经网络上的后门攻击。

基于上述问题，本文总结了目前针对图神经网络后门攻击的防御挑战：1) 标签混乱，后门攻击会篡改训练数据集中后门样本的标签，使其标签与攻击者选择的目标类一致，导致训练数据集中存在标签不匹配的情况；2) 触发器不确定性，后门样本中存在着由特定结构子图构成的触发器，难以直接确定触发器所在子图位置；3) 性能下降，后门训练数据集带有触发器扰动，这会降低基于该数据集训练的目标模型表现性能。

为了缓解图神经网络的后门攻击威胁，本文提出了一种基于对比学习的图神经网络后门攻击防御（CLB-Defense, contrastive learning defense for GNN against backdoored attack）方法。具体而言，针对标签混乱的挑战，采用对比学习的策略训练出不依赖数据标签的对比模型，并与目标模型结合，评估 2 个模型输入训练数据集和输出置信分数之间的差异，找出可疑的后门样本。其次，采取标签平滑策略得到新的标签并替代可疑后门样本原有的标签。再次，面对后门样本中触发器不确定性，利

用图重要性指标对图中的结构进行评估，并进行图重构操作，滤除存在于后门样本中的触发器，从而实现对图后门攻击的防御。关于性能下降的挑战，利用处理过可疑后门样本的训练数据，重新训练目标模型，进而减少模型受扰动的影响。最终，CLB-Defense 在 4 个真实数据集和 5 种主流后门攻击方法上验证了其防御的有效性，能够平均降低 75.66%的攻击成功率。

综上所述，本文的主要贡献如下。

1) 针对图神经网络面对后门攻击鲁棒性不足的问题，提出了一种基于对比学习的防御方法（CLB-Defense），根据对比学习训练的对比模型找出可疑后门样本，并采用标签平滑和图重构的策略去除训练数据集中的扰动。

2) 从多个角度对 CLB-Defense 的防御有效性进行了深层次的分析，并可视化了重构后的后门样本，为所提方法的防御能力提供解释。

3) 在 4 个真实数据集和 5 种主流图神经网络后门攻击方法上进行防御验证，结果表明 CLB-Defense 能够平均降低 75.66%的攻击成功率（与对比算法相比，改善了 54.01%），验证了所提方法对多样性后门攻击有良好的防御性能。

1 相关工作

本节简要介绍图神经网络后门攻击和对比学习方面的相关工作。

1.1 面向图神经网络后门攻击

依据触发器的生成方式，已有的后门攻击方法可以分为两类，即基于子图生成方法和基于梯度优化方法。针对基于子图生成方法，Zhang 等[7]首先提出了随机生成满足小世界网络、随机网络或优先连接网络分布的子图作为触发器进而发动后门攻击。进一步地，Xu 等[8]采用模型可解释工具（如 GNNExplainer）来搜索触发器注入的位置，以实现有效的后门攻击。此外，Sheng 等[9]利用图的重要性指标从图结构的角度来评估图中节点的重要性，以此找到触发器节点的位置。Zheng 等[10]从模体的角度分析图数据的子图分布特点进而快速确定触发器，并结合模型层面和图结构层面确定触发器的注入位置。针对基于梯度优化方法，Xi 等[11]设计了一种根据图自身结构构建相适应的触发器的图后门攻击方法，采用双层优化的策略来更新触发器生成器的参数。

图神经网络上的后门攻击相关研究已经表明，发动后门攻击的方式尽管存在差异，但是都能高效地通过预先设置的触发器激活后门模型中的后门，从而误导目标模型的预测结果。这进一步凸显了面向图神经网络后门攻击防御方法的重要性。

1.2 面向图神经网络对比学习

图对比学习是一种针对图数据的自监督学习方法，能够在大量无标注图数据下训练出表现性能良好的图编码器。已有的图对比学习方法按照数据增强策略的不同可以分为两类：启发式的数据增强方法和可学习的数据增强方法。针对启发式的数据增强方法，Qiu 等[4]提出了一种随机采样子图的策略进行图对比编码，来学习潜在的、可迁移的图结构信息。基于不同视图的图增强方式，Hassani 等[15]将邻接矩阵转换为扩散矩阵，并将这 2 个矩阵视为同一图结构的 2 个全等视图。You 等[5]对连边、节点或节点特征进行操作（如随机删除连边等）实现图数据增强，并探究了不同数据增强方式所带来的性能影响。Zhu 等[16]在数据增强方面提出根据节点的中心性设定删除连边的概率，不重要的边被删除的概率更高。

针对可学习的数据增强方法，Suresh 等[6]使用的对抗图增强策略来避免在训练过程中捕获冗余信息，从而达到自动、自适应、动态地数据增强的目的。You 等[17]将增强数据中的离散先验扩展到图生成器的参数空间中的可学习的连续先验，来实现自适应数据增强的方式。现有的图对比学习方法已被证明能够在大量无标注图数据的情况下，训练得到良好性能的图嵌入模型。

2 基础知识

本节介绍图、图分类任务的图神经网络、图神经网络后门防御的相关定义。

定义 1 图。将图表示为 $G=\{V,E,\boldsymbol{X}\}$，其中，$V=\{v_1,\cdots,v_n\}$ 表示 n 个节点的集合，E 表示连边集合，特征矩阵 $\boldsymbol{X}$ 表示图中节点所包含的特征信息。邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 表示图所包含的网络拓扑信息，当节点 v_i 和节点 v_j 存在直接连边时，$A_{i,j}=1$；否则 $A_{i,j}=0$。为此，图可以简洁地表示为 $G=\{\boldsymbol{A},\boldsymbol{X}\}$。

定义 2 图分类任务的图神经网络。图分类数据集表示为 $\mathcal{G}=\{(G_1,y_1),\cdots,(G_N,y_N)\}$，包含 N 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。$Y=\{c_1,\cdots,c_L\}$ 表示数据集有 L 类标签空间。图神经网络模型 $F_\theta(\cdot)$ 是一个图分类器，其目标是通过已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 $F_\theta:\mathcal{G}\to Y$ 。

定义 3　图神经网络后门防御。给定一个图分类数据集 $\mathcal{G}$、良性模型 $F(\cdot)$、后门模型 $F_b(\cdot)$。攻击者通过混合函数 $M(\cdot)$ 将触发器 g 注入良性样本 G 中生成后门样本 G_b，使后门模型 $F_b(\cdot)$ 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 G_b，防御下的目标模型 $F_d(\cdot)$ 仍可以进行准确的分类，表示为

$$\begin{cases} F_b(M(G,g))=y_t \\ F_d(M(G,g))=F(G) \end{cases} \quad \text{s.t.}\, G\in\mathcal{G} \tag{1}$$

其中，$M(\cdot)$ 是负责将触发器 g 注入良性样本中的混合函数，y_t 是攻击者选定的目标类，防御的目标是使攻击者无法通过触发器误导目标模型，同时目标模型与良性模型有着相同的表现性能。

3　CLB-Defense 介绍

本节详细介绍提出的基于对比学习的后门攻击防御方法 CLB-Defense。现有的图神经网络后门攻击方法[11-15]是利用触发器与标签之间训练搭建的强相关系作为攻击媒介，即在部分训练数据中注入触发器以及网络标签修改为攻击目标类，进而将后门保留在模型中。根据后门攻击工作原理，CLB-Defense 针对训练数据集纠正样本错误标签，过滤样本中的触发器，得到干净的训练数据集，从而实现有效防御图神经网络后门攻击。CLB-Defense 的框架如图 2 所示，可分为 3 个阶段：对比学习构建对比模型、差值查找可疑样本、图重构及标签平滑。

具体来说，第一阶段为对比学习构建对比模型，采用对比学习的策略训练出不依赖数据标签的对比模型，可以避免受到后门攻击。第二阶段为差值查找可疑样本，计算目标模型和对比模型关于训练数据集的输出置信分数之间的差异，通过差值查找到可疑的后门样本。第三阶段为图重构及标签平滑，针对可疑的后门样本，利用网络重要性指标对样本的拓扑结构进行重构，过滤不合理的连边，并使用标签平滑的策略对后门样本的标签进行重置，通过对比模型和目标模型输出的置信分数平滑得到样本的重构标签。

3.1　对比学习构建对比模型

图神经网络后门攻击会篡改部分训练数据的标签，首先基于对比学习的策略构建一个不依赖训练数据标签的对比模型，如图 2 中的阶段 1 所示。具体来说，受文献[5]启发，本文对训练数据进行图

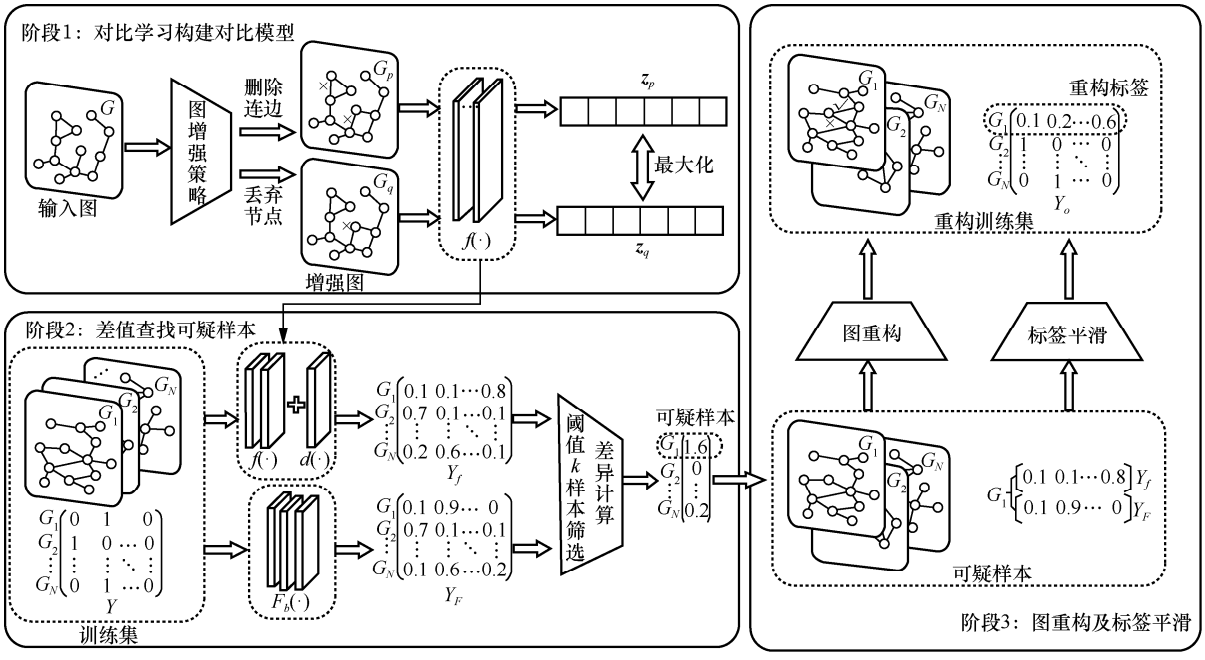


图 2　CLB-Defense 框架

增强操作构建正负样本，考虑到大多数触发器是以子图结构的形式被注入样本中，这里图增强操作选用随机删除连边和丢弃节点这 2 种方式来构建正负样本。这样不仅可以快速得到大量的增强数据，也能够破坏数据中存在的触发器，表示为

$$G_p, G_q = D_{\text{aug}}(G) \quad , \text{s.t.}\, G \in \mathcal{G}_{\text{train}} \tag{2}$$

其中，$D_{\text{aug}}(\cdot)$ 为图增强操作，$\mathcal{G}_{\text{train}}$ 为训练数据集，包含 N 个图样本，G 为 $\mathcal{G}_{\text{train}}$ 中的某样本。G_p 和 G_q 分别为由删除连边和丢弃节点所得到的增强图，共有 $2N$ 个样本。其中，来自同一样本的增强图认定为正样本，而来自不同样本的增强图认定为负样本。

然后，构建一个基于图神经网络的编码器 $f(\cdot)$，用来提取增强图的特征，可以得到增强图 G_p 和 G_q 对应的嵌入特征 $\boldsymbol{z}_p$ 和 $\boldsymbol{z}_q$。对比学习的损失函数定义为最大化正样本之间的一致性，使正样本在特征空间中相互靠近，而负样本在特征空间中相互远离。这里使用归一化温度表示交叉熵损失，计算式为

$$l_{p,q} = -\log \frac{\exp\left(\dfrac{\text{sim}(\boldsymbol{z}_p, \boldsymbol{z}_q)}{\tau}\right)}{\displaystyle\sum_{m=1}^{2N} \mathbb{I}[m \neq q] \exp\left(\dfrac{\text{sim}(\boldsymbol{z}_p, \boldsymbol{z}_m)}{\tau}\right)} \tag{3}$$

其中，$\text{sim}(\boldsymbol{z}_p, \boldsymbol{z}_q) = \dfrac{\boldsymbol{z}_p^{\mathrm{T}} \boldsymbol{z}_q}{\|\boldsymbol{z}_p\| \|\boldsymbol{z}_q\|}$ 为 $\boldsymbol{z}_p$ 和 $\boldsymbol{z}_q$ 这 2 个特征的余弦相似度；τ 为温度参数；$\mathbb{I}[m \neq q]$ 为指示函数，当 $m \neq q$ 时，指示函数值为 1，否则为 0。最后优化的损失值为将一小批次的正样本对损失的累加。

3.2 差值查找可疑样本

基于对比学习构建的对比模型编码器 $f(\cdot)$ 是在无标签的情况下构建正负样本对，使正样本对的嵌入特征相互靠近，负样本对的嵌入特征相互远离，而且编码器的训练不依赖于数据标签。为了构建端到端的预测模型，在编码器后加入两层的多层感知器作为解码器 $d(\cdot)$，用于实现下游任务。为了使基于对比模型搭建的分类器有着良好表现性能，防御方提供了对比样本率 ς 的训练数据集带有正确标签，用于解码器参数训练。为此，对比模型 $f(\cdot)$ 与依赖训练数据集（存在后门样本）标签所得到的目标模型 $F_b(\cdot)$ 对训练数据集中的后门样本预测置信分数会有着不同的表现。直观上，带有后门的目标模型 $F_b(\cdot)$ 会将带有触发器的后门样本预测为攻击者的目标类，而对比模型 $f(\cdot)$ 与解码器 $d(\cdot)$ 构成的分类器则不会受到触发器的干扰，计算这 2 种端到端模型在面对后门样本时所表现的性能差异，以此来找到可疑的后门样本。首先分别得到输入训练数据集到相应分类器的输出置信分数，表示为

$$\begin{cases} Y_f = d(f(\mathcal{G}_{\text{train}})) \\ Y_F = F_b(\mathcal{G}_{\text{train}}) \end{cases} \tag{4}$$

其中，$\mathcal{G}_{\text{train}}$ 表示训练数据集，包含 N 个图样本和 L 类标签；$Y_f \in \mathbb{R}^{N \times L}$ 表示基于对比模型的分类器输出的训练数据集对应的置信分数；$Y_F \in \mathbb{R}^{N \times L}$ 表示带有后门的目标模型输出的训练数据集对应的置信分数。采用曼哈顿距离[18]来计算 Y_F 和 Y_f 之间的差异，表示为

$$\text{Diff}_i = \sum_{j=1}^{L} \left| Y_f(i,j) - Y_F(i,j) \right|, \text{s.t.}\, i \in \{1, 2, \cdots, N\} \tag{5}$$

其中，Diff_i 表示第 i 个图样本输入图分类器中得到输出置信分数之间的差异，L 为数据集的类别数。根据差异阈值 k 筛选出可疑的后门样本，表示为

$$\text{Sus\_idx} = \text{Choose}(\text{Diff}, k) \tag{6}$$

其中，Sus_idx 是通过差异阈值筛选出来的训练数据集中可疑后门样本的序号。

3.3 图重构及标签平滑

经过差值查找出可疑后门样本后，紧接着需要过滤可疑后门样本中的触发器并对其原有标签进行重塑，进而实现对后门攻击的防御。具体分为 2 个步骤，即图重构和标签平滑。

图重构。首先，对可疑样本中的图结构进行处理，达到过滤可疑样本中触发器的目的。具体来说，对于可疑的后门样本，利用边介数中心性指标计算图中的所有连边重要性程度。边介数中心性[19]是复杂网络研究中的一种经典结构相似度算法。它是通过计算图中所有节点对之间的最短路径，并计算经过每条连边的次数得到的。连边 e 经过的最短路径越多，其边介数中心性值越大，表明连边 e 越重要。边介数中心性定义为

$$\text{BEC}_e = \sum_{i \neq j} \frac{\sigma_{ij}(e)}{\sigma_{ij}} \tag{7}$$

其中，σ_{ij} 表示节点 v_i 到节点 v_j 的最短路径数，$\sigma_{ij}(e)$ 表示从 v_i 到 v_j 通过连边 e 的最短路径数。对于可疑的后门样本，利用边介数中心性指标计算图中

的所有连边重要性程度，进行升序排序，筛选出不重要的连边并删除，表示为

$$\mathcal{G}_{\text{del}} = R_{\text{del}}(\mathcal{G}_{\text{train}}, \text{Sus\_idx}, \text{del\_rate}) \tag{8}$$

其中，$\mathcal{G}_{\text{train}}$ 表示训练数据集；Sus_idx 表示训练数据集中可疑样本的序号；del_rate 表示删除连边的数量占图中连边数的比例，即连边丢弃率；$\mathcal{G}_{\text{del}}$ 表示经过删除连边处理的训练数据集。

同时，为了避免错误删除的连边，采用复杂网络研究中的共同邻居数指标[20]来增加连边。共同邻居数是一种衡量结构相似度的指标，2 个节点的公共邻居越多，则说明两者在网络中的关系越近，定义为

$$\text{CN} = \left|\Gamma(i) \cap \Gamma(j)\right| \tag{9}$$

其中，$\Gamma(i)$ 和 $\Gamma(j)$ 分别表示节点 v_i 和节点 v_j 周围的邻居节点集合。对于可疑的后门样本，利用共同邻居数指标计算图中的所有连边重要性程度，进行降序排序筛选出重要的连边并进行添加，表示为

$$\mathcal{G}_{\text{add}} = R_{\text{del}}(\mathcal{G}_{\text{del}}, \text{Sus\_idx}, \text{add\_rate}) \tag{10}$$

其中，$\mathcal{G}_{\text{del}}$ 表示经过删除处理的训练数据集；Sus_idx 表示训练数据集中可疑样本的序号；add_rate 表示增加连边的数量占图中连边数的比例，即连边增强率；$\mathcal{G}_{\text{add}}$ 表示经过添加连边处理的训练数据集，即对于可疑样本完成了图重构的数据集。

标签平滑。对于可疑后门样本的标签，利用基于对比模型的图分类器与后门目标模型输出的置信分数相匹配，得到平滑后的置信分数作为可疑后门样本新的标签，表示为

$$Y_o(i) = \begin{cases} Y_f(i) + \alpha Y_F(i), & i \in \text{Sus\_idx} \\ Y(i), & \text{其他} \end{cases} \tag{11}$$

其中，α 为标签平滑率，表示新的标签中对比模型输出的置信分数在 $Y_o(i)$ 中所占比例；i 为训练数据集中对应样本的序号；Sus_idx 为训练数据集中可疑样本的序号；Y_f 为对比模型构建的图分类器输出的置信分数；Y_F 为后门目标模型输出的置信分数；Y 为训练数据集原有的标签。

为了便于描述，将经过对可疑后门样本的图重构和标签平滑处理的训练数据集 $\mathcal{G}_{\text{train}}$ 称为防御后的训练数据集 $\mathcal{G}_{\text{def}}$。

4　实验与分析

为了验证所提出的基于对比学习的图神经网络后门攻击防御方法 CLB-Defense 的性能，本文在 4 个真实数据集以及 5 种主流图神经网络后门攻击方法上进行实验，分别从防御有效性实验和 CLB-Defense 有效性分析的实验 2 个部分进行详细阐述。

4.1　数据集及评价指标

本文在 4 个广泛使用的真实数据集上评估 CLB-Defense 的防御性能，分别是生物网络中的 PROTEINS[9]数据集、小分子网络中的 AIDS[11]和 NCI1[10]数据集，以及社交网络中的 DBLP_v1[10]数据集。表 1 总结了 4 个图数据集的基本统计信息。其中，图标签分布 663[0], 450[1]表示属性为 0 的图像有 663 张，属性为 1 的图像有 450 张。

为了客观准确地衡量 CLB-Defense 的防御性能，本文采用攻击成功率[7]（ASR, attack success rate）、平均防御置信分数[11]（ADC, average defense confidence）和分类准确率[3]（ACC, accuracy）3 个评价指标。其中 ASR 表示攻击样本被目标模型预测为目标类的数量与攻击样本数量的比例，表示为

$$\text{ASR} = \frac{N_{\text{suc}}}{N_{\text{att}}} \tag{12}$$

其中，N_{suc} 表示被成功攻击样本的数量，N_{att} 表示攻击样本的数量。ASR 指标能直观地反映攻击性能的变化程度，进而刻画防御方法的有效性。

ADC 指标表示被成功防御样本的平均输出置信分数，表示为

$$\text{ADC} = \frac{\sum_{n=1}^{N_{\text{suc-def}}} \text{Con}_n}{N_{\text{suc-def}}} \tag{13}$$

其中，$N_{\text{suc-def}}$ 表示成功防御的攻击样本数量，Con_n

表 1　　**4 个图数据集的基本统计信息**

数据集	图数量/个	平均节点数/个	连边数/条	图标签分布	目标类
PROTEINS	1 113	39.06	72.82	663[0],450[1]	1
AIDS	2 000	15.69	16.20	400[0],1 600[1]	0
NCI1	4 110	29.87	32.30	2 053[0],2 057[1]	0
DBLP_v1	19 456	10.48	19.65	9 530[0],9 926[1]	0

表示样本对应标签类的置信分数。ADC 的值越高，表明防御性能越好。

ACC 指标是样本中正确预测数量与总的样本数量的比例，表示为

$$\mathrm{ACC}=\frac{N_{\mathrm{cor}}}{N_{\mathrm{total}}} \tag{14}$$

其中，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数量。ACC 可以反映模型的表现性能，ACC 的值越高，表明模型性能越好。

4.2 后门攻击方法及目标模型

本文选用了 5 种主流后门攻击方法，即 ER-B（erdős-rényi backdoor）[7]、MIA （most important nodes selecting attack）[8]、MaxDCC[9]、GTA （graph trojaning attack）[11]、Motif-Backdoor[10]。

因为攻击方法[7-11]都以 GIN（graph isomorphism network）[3]模型作为攻击的目标模型，GIN 模型通过引入多层感知器利用可学习的参数来确保注入性，同时采用累加池化的方法来聚合图级信息，在图分类任务上有着出色的表现性能，所以本文选择 GIN 模型作为目标模型。

4.3 对比防御方法

为了验证 CLB-Defense 的有效性，而目前没有针对图神经网络后门攻击防御的相关研究。为此，本文迁移了 3 种有效防御图神经网络对抗攻击的方法作为对比防御方法，分别为 Jaccard-Based[21]、Label-Smooth[22]和 Adv-Training[23]。其中，Jaccrard-Based 计算图中连边的 Jaccard 相似度，删除可疑连边；Label-Smooth 将图对应的真实标签改为软标签；Adv-Training 通过 ER-B 攻击方法构建训练数据集。

4.4 实验设置

受先前工作[3]启发，本文采用十折交叉验证的方式去评估目标模型的表现性能。对于目标模型训练，采用 Adam 优化器去优化模型参数，其余参数遵循原工作[3]设计。参考后门攻击方法设置，将触发器大小设置为 4 个节点，中毒比例为 10%的训练数据集。对于防御方法，Jaccard-Based 删除的连边数占样本连边总数的 10%。Label-Smooth 平滑比例 p 设置为 0.7。Adv-Training 构建训练数据集 10%的样本作为对抗样本。CLB-Defense 中的差异阈值 k 设置为 0.5，标签平滑率 α 设置为 0.7，对比样本率 ς 设置为 0.1，连边丢弃率 del_rate 设置为 0.05，连边增强率 add_rate 设置为 0.02。所有实验均在一台搭载了 2×Intel(R) Xeon(R) Gold 5218R CPU @ 2.10 GHz、384 GB 系统内存和 8×NVIDIA A100 Tensor Core 40G GPU 的服务器上进行。

4.5 防御实验

为了验证 CLB-Defense 的防御有效性和可用性，首先需要评估以下 2 个主要的问题：1) 面对多样化的图神经网络后门攻击方法，CLB-Defense 是否能起到令人满意的防御效果；2) 面对良性样本，CLB- Defense 是否会大幅度降低图神经网络模型的表现性能。

1) 针对后门攻击的防御效果

选择 4 个广泛真实的数据集和 5 种图神经网络后门攻击方法来评估 CLB-Defense 和 3 种对比防御方法的防御性能，结果如表 2～表 5 所示。其中，在 PROTEINS、AIDS、NCI1、DBLP_v1 数据集上，目标模型在无攻击情况下实现的 ACC 分别为 76.23%、98.92%、77.01%、80.83%；5 种图神经网络后门攻击方法在 4 个数据集上平均能达到 81.86%的 ASR 和 80.29%的分类准确率 ACC。

从整体防御性能角度分析，CLB-Defense 使 5 种后门攻击方法在 4 个数据集上实现的平均 ASR 为 19.92%，同时平均置信分数 ADC 达到 0.9293，而 Jaccard-Based、Label-Smooth 和 Adv-Training 的平均 ASR 分别为 58.66%、52.58%和 48.89%，平均置信分数 ADC 分别为 0.833 8、0.876 1 和 0.895 6。CLB-Defense 都能实现最优的防御性能，带来这个现象的原因主要有 2 个。其一，CLB-Defense 利用对比学习构建对比模型，通过差值查找出了训练数据集中标签存在错误的样本，并对其标签进行平滑操作，使模型对于这部分数据进行软标签的学习。其二，CLB-Defense 基于 2 个网络重要性指标对可疑的后门样本进行了图重构，删除部分网络不重要的连边，增加网络中重要的连边。CLB-Defense 清洗可疑的后门样本中的触发器以及标签，使目标模型基于这一批数据集训练时无法被留下后门，从而实现防御的目的。此外，CLB-Defense 使 5 种攻击方法在 PROTEINS、AIDS、NCI1 和 DBLP_v1 数据集上的 ASR 分别降低了 88.50%、91.21%、49.22%和 75.61%。值得注意的是，在 AIDS 数据集上，CLB-Defense 起到最佳防御效果，原因主要是图神经网络模型 GIN 在 AIDS 数据集的 ACC 为 98.92%，这就意味着数据集中不同类之间的分类边界划分是准确的，因此利用对比模型识别数据集中的标签也较准确，从而高效、精确地筛选出数据集中的后门样本，达到防御目的。

从图神经网络模型表现性能的角度分析，在 4 个广泛真实的数据集和 5 种图神经网络后门攻击的方

表 2　PROTEINS 数据集上不同防御方法的防御性能

后门攻击方法	ASR					ADC					ACC				
	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense
ER-B	64.57%	29.41%±2.19%	26.39%±3.74%	23.86%±2.58%	9.41%±2.08%	—	83.06%±2.07%	84.35%±1.74%	84.78%±1.54%	86.29%±0.54%	71.16%	72.51%±2.25%	73.16%±2.71%	73.85%±1.87%	74.24%±1.03%
MIA	64.89%	26.72%%±2.76%	17.48%±2.04%	17.14%±2.37%	4.54%±1.08%	—	81.07%±1.19%	79.58%±1.12%	81.54%±2.04%	82.28%±0.57%	70.82%	70.73%±2.58%	73.29%±1.70%	73.54%±1.67%	73.95%±1.48%
MaxDCC	84.75%	58.82%±3.29%	55.46%±3.16%	48.76%±3.28%	15.79%±3.05%	—	90.95%±2.35%	91.16%±2.87%	91.27%±2.23%	96.48%±0.58%	72.06%	72.85%±1.60%	72.92%±1.16%	73.04%±2.26%	74.39%±0.36%
GTA	86.72%	24.98%±2.35%	26.05%±3.64%	18.64%±1.15%	3.36%±1.06%	—	92.06%±0.31%	90.21%±1.05%	92.53%±1.41%	95.59%±0.16%	71.56%	71.53%±1.46%	72.13%±1.28%	74.35%±2.05%	74.76%±1.21%
Motif-Backdoor	89.46%	75.62%±2.37%	63.87%±1.52%	61.27%±2.06%	11.76%±1.68%	—	88.21%±1.92%	91.07%±1.46%	92.81%±3.17%	98.51%±0.36%	71.43%	72.08%±1.27%	73.26%±1.32%	74.68%±1.02%	75.03%±1.34%

表 3　AIDS 数据集上不同防御方法的防御性能

后门攻击方法	ASR					ADC					ACC				
	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense
ER-B	93.62%	73.33%±0.53%	56.82%±2.34%	48.32%±2.53%	4.31%±0.91%	—	72.89%±1.18%	91.48%±0.53%	83.67%±1.98%	91.85%±0.12%	96.28%	96.58%±0.47%	96.92%±0.63%	97.48%±0.72%	98.46%±0.18%
MIA	95.48%	81.87%±1.52%	72.31%±2.38%	70.81%±1.32%	1.06%±0.15%	—	88.05%±0.95%	79.58%±1.12%	92.26%±1.13%	98.27%±0.02%	96.85%	96.15%±0.61%	96.98%±0.46%	97.65%±0.34%	98.39%±0.23%
MaxDCC	96.57%	86.25%±1.45%	79.75%±0.96%	71.81%±1.36%	8.63%±1.25%	—	81.09%±2.46%	82.65%±0.83%	87.79%±0.72%	96.14%±0.16%	98.12%	98.34%±0.54%	98.52%±0.47%	98.64%±0.79%	98.72%±0.62%
GTA	98.52%	89.06%±1.57%	88.82%±1.39%	85.75%±1.42%	7.94%±2.11%	—	90.56%±0.43%	90.97%±0.56%	92.89%±0.42%	99.76%±0.05%	97.39%	97.58%±0.45%	97.92%±0.73%	98.41%±0.48%	98.74%±0.27%
Motif-Backdoor	99.86%	90.75%±1.62%	88.72%±1.82%	82.93%±1.06%	20.63%±2.12%	—	85.72%±2.18%	87.68%±1.86%	95.87%±0.68%	99.89%±0.03%	97.64%	97.83%±0.64%	97.92%±0.58%	98.25%±0.62%	98.68%±0.34%

表 4　NCI1 数据集上不同防御方法的防御性能

后门攻击方法	ASR					ADC					ACC				
	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense
ER-B	78.32%	72.25%±2.64%	62.97%±2.65%	55.07%±2.08%	45.39%±2.51%	—	85.85%±2.18%	98.50%±0.57%	98.45%±0.37%	99.25%±0.08%	73.85%	73.04%±0.61%	74.57%±1.02%	75.38%±1.53%	75.74%±1.08%
MIA	96.98%	87.63%±2.32%	72.85%±1.29%	71.59%±2.13%	57.53%±2.18%	—	78.52%±2.51%	93.73%±2.16%	96.24%±1.35%	98.07%±0.17%	73.36%	74.09%±0.55%	75.47%±1.12%	75.78%±0.48%	76.06%±0.12%
MaxDCC	98.95%	78.54%±1.86%	62.07%±2.38%	60.25%±1.98%	42.47%±2.53%	—	84.67%±2.48%	99.03%±0.16%	99.39%±0.34%	99.92%±0.07%	74.38%	75.03%±0.82%	75.64%±1.17%	75.78%±1.54%	76.53%±0.63%
GTA	100%	74.25%±1.26%	72.31%±3.01%	69.40%±3.12%	46.48%±1.87%%	—	91.34%±0.24%	92.46%±0.73%	96.24%±1.35%	99.54%±0.13%	74.05%	74.67%±0.59%	75.18%±0.43%	76.30%±0.48%	76.87%±0.76%
Motif-Backdoor	100%	92.64%±2.37%	82.86%±2.38%	81.64%±2.72%	48.94%±2.38%	—	75.63%±2.16%	96.29%±0.80%	98.15%±0.79%	99.46%±0.21%	73.25%	73.82%±0.96%	75.35%±1.33%	75.40%±0.72%	75.78%±0.57%

表 5 **DBLP_v1 数据集上不同防御方法的防御性能**

后门攻击方法	ASR					ADC					ACC				
	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense	无防御	Jaccard-Based	Label-Smooth	Adv-Training	CLB-Defense
ER-B	41.28%	21.57%±2.58%	20.28%±1.73%	17.03%±2.24%	15.28%±1.35%	—	75.01%±2.43%	78.39%±2.52%	78.54%±3.26%	80.11%±0.15%	78.52%	78.92%±0.12%	79.05%±0.17%	79.83%±0.28%	80.01%±0.19%
MIA	43.65%	15.17%±0.47%	13.05%±2.72%	11.04%±1.73%	10.84%±1.04%	—	66.98%±0.17%	68.95%±0.92%	70.53%±0.54%	72.20%±0.58%	78.46%	78.86%±0.23%	78.25%±0.71%	79.90%±0.46%	79.98%±0.38%
MaxDCC	62.86%	22.54%±2.46%	26.79%±3.24%	23.25%±2.23%	9.45%±1.02%	—	84.11%±1.38%	82.17%±0.68%	84.98%±1.94%	86.03%±0.18%	79.23%	78.95%±0.75%	78.39%±0.86%	79.36%±0.77%	79.87%±0.54%
GTA	68.42%	18.95%±1.16%	15.74%±1.19%	13.59%±1.05%	10.28%±0.87%	—	85.94%±0.15%	86.23%±0.62%	86.42%±0.38%	87.19%±0.48%	78.49%	78.06%±0.28%	78.26%±0.67%	79.65%±0.52%	79.93%±0.51%
Motif-Backdoor	71.84%	52.78%±1.07%	46.92%±1.25%	45.76%±2.87%	24.39%±0.56%	—	85.91%±0.25%	86.73%±1.26%	86.94%±2.04%	91.69%±0.21%	78.85%	79.20%±0.12%	80.18%±0.08%	80.23%±0.16%	80.48%±0.17%

法上，CLB-Defense 实现的平均 ACC 为 82.33%，而 Jaccard-Based、Label-Smooth 和 Adv-Training 实现的平均 ACC 分别为 80.54%、81.17%和 81.88%。与无防御方法下的平均 ACC（80.29%）相比，CLB-Defense 实现了分类准确率的最优提升，但与良性模型实现的平均 ACC（83.25%）还有差距。造成该现象的原因如下：首先，CLB-Defense 通过构建对比模型和差值查找可疑样本的策略，筛选出数据集中可疑的后门样本；其次，采用图重要性指标对样本进行重构，滤除样本中可能存在的触发器，同时对样本中的真实标签用标签平滑策略得到的新标签进行替换。CLB-Defense 重构了数据集中可能存在的后门样本，使目标模型的 ACC 得到了改善。而与良性模型的 ACC 仍有差距的原因是经过 CLB-Defense 防御之后的目标模型还存在一些后门攻击的样本。

2) 防御方法在良性样本上的表现性能

在评估 CLB-Defense 防御有效性的过程中，除了验证防御方法面对图神经网络后门攻击所表现出的防御性能外，还需要考虑防御方法面对不带有后门样本的训练数据时，对图神经网络模型正常性能的影响。因此，为了探究所提防御方法在良性样本上的表现性能，将带有防御方法的良性模型在 4 个真实的良性数据集进行图分类的实验，记录模型的 ACC，实验结果如图 3 所示。

针对 4 个数据集，良性模型在无防御方法下平均 ACC 为 83.25%，在 Jaccard-Based、Label-Smooth、Adv-Training、CLB-Defense 防御方法下平均 ACC 分别为 80.69%、83.13%、82.67%、82.54%。造成该现象的原因主要是 CLB-Defense 并非对数据集中所有的样本都进行修改，而是利用对比模型与目标模型进行差异查询，找出可疑的后门样本，再进行图重构、标签平滑等操作。因此，这能够最大限度上减小 CLB-Defense 防御方法对良性样本的影响。

4.6 CLB-Defense 有效性分析

通过实验验证了 CLB-Defense 防御有效性后，需要进一步分析的是防御有效性深层次的原因。为此，本文进行了 4 个方面的实验，即纠错能力、消融实验、后门样本重构的可视化以及时间复杂度。

1) 纠错能力

为了研究 CLB-Defense 能够实现有效防御的深层次原因，在面对最优攻击性能的 Motif-Backdoor 后门攻击时，分析了 CLB-Defense 防御方法对后门数据集的修改情况。具体来说，定义了 4 个指标来直观描述，即正后率、后训率、改正率、改误率。正后率表示 CLB-Defense 成功纠正后门样本对应标签的样本数量与后门数据集本存在标签错误的后门样本数量之间的比例。后训率表示 CLB-Defense 修改后训练数据集仍然存在的标签错误后门样本数量与训练数据集样本数量之间的比例。改正率表示 CLB-Defense 修改的样本中为正确标签所占的比例。改误率表示 CLB-Defense 修改的样本中为错误标签所占的比例。实验结果如图 4 所示。

面对 4 个真实数据集，CLB-Defense 防御下实现的平均正后率为 80.59%，这表明 CLB-Defense 防御方式能够将后门数据集中大部分标签错乱的

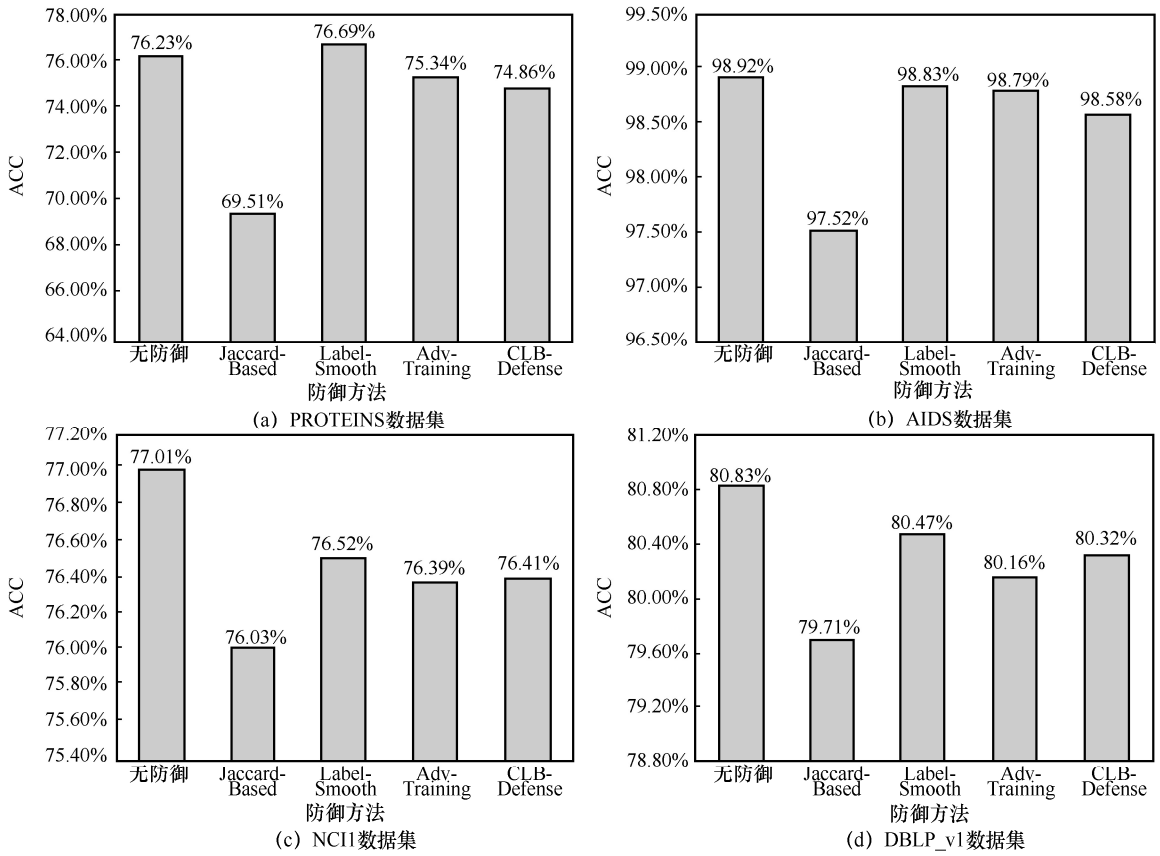


图 3　防御方法对良性样本的影响

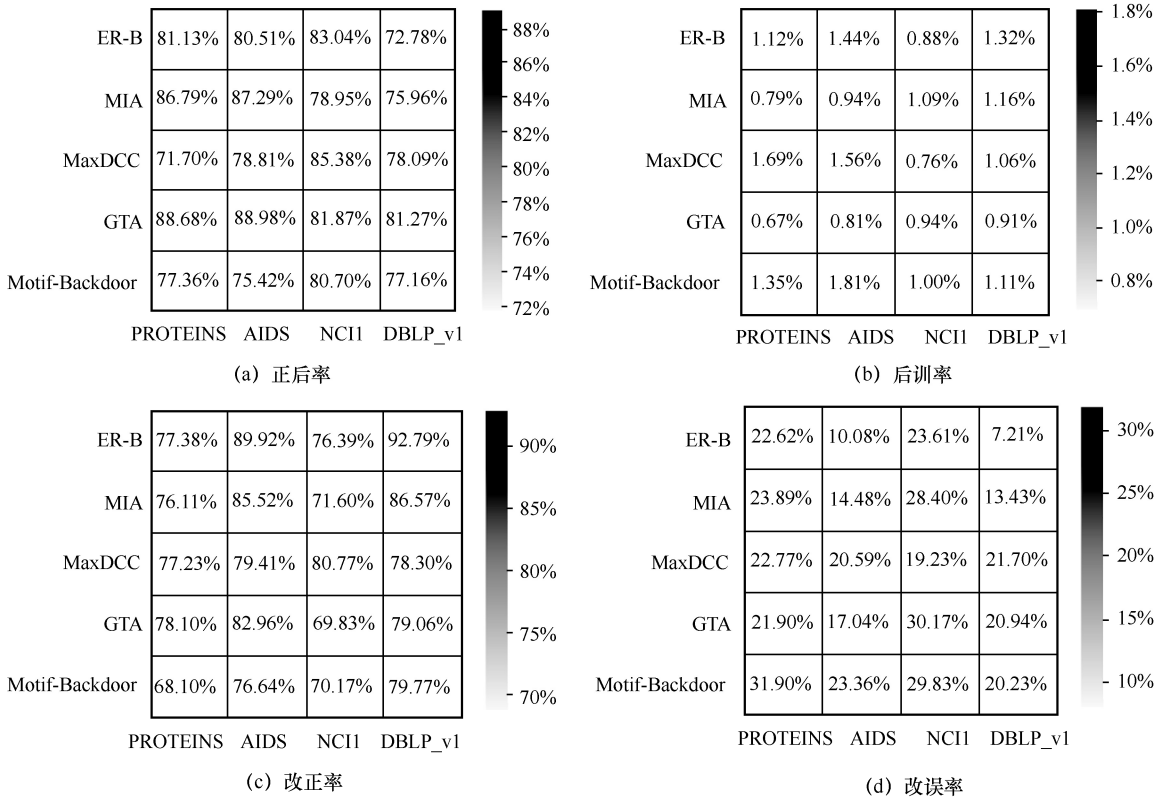


图 4　CLB-Defense 纠错能力

后门样本查找到并修改为正确标签，以此来保证防御的有效性。其中，针对 AIDS 数据集，CLB-Defense 防御方法实现了 4 个数据集中最高的平均正后率，即 82.20%，同时这也对应防御实验中 CLB-Defense 在该数据集上实现的攻击成功率下降最显著，即 91.21%。此外，在 4 个数据集上，CLB-Defense 防御下实现的后训率为 1.12%，意味防御下的训练数据集，标签错乱的后门样本仅占训练样本的 1.12%，使后门攻击方法难以在目标模型上留下后门。这也进一步验证了 CLB-Defense 中标签平滑策略对防御后门攻击方法的有效性。值得注意的是 CLB-Defense 在 4 个数据集上平均实现的改正率和改误率分别为 78.83% 和 21.17%。这表示 CLB-Defense 并不会大量修改良性样本的标签，从而保障了目标模型的分类准确率，进一步验证 CLB-Defense 中采用对比模型差值查找可疑的后门样本进行处理的有效性。

2) 消融实验

CLB-Defense 在实现防御后门攻击的过程中，有着 2 个重要的模块，即图重构和标签平滑。为了进一步探究各个模块对防御性能的影响，面对 Motif-Backdoor 后门攻击方法，本文进行了 CLB-Defense 的消融实验。具体来说，CLB-Aug 是将 CLB-Defense 方法中的图重构模块保留、标签平滑模块删除的方法，CLB-Label 是将 CLB-Defense 方法中的标签平滑模块保留、图重构模块删除的方法，实验结果如图 5 所示。

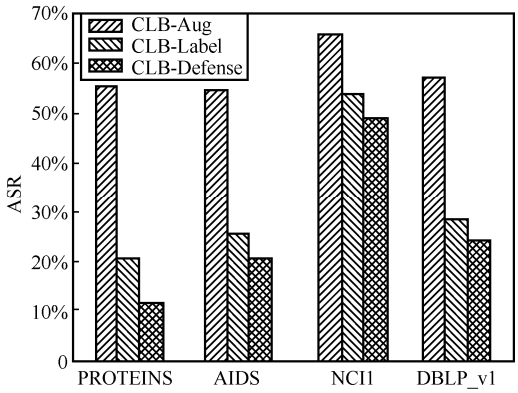


图 5 CLB-Defense 消融实验

在 CLB-Defense 防御方法下，Motif-Backdoor 攻击在 4 个数据集上实现的平均 ASR 为 26.43%，而在 CLB-Label 和 CLB-Aug 防御方法下，平均 ASR 分别达到 32.20%和 58.26%。CLB-Label 方法专注于可疑样本中的标签部分，CLB-Aug 方法专注于可疑样本图结构部分。实验结果表明 CLB-Label 方法的防御性能优于 CLB-Aug 方法，这意味着标签平滑模块对 CLB-Defense 的防御贡献程度更大。同时揭露了现有后门攻击方法中修改图标签环节对攻击有效性起着重要的作用。因此，在防御图神经网络后门攻击时，应该重点关注训练数据集标签是否被篡改的问题，提高防御成功率。

3) 后门样本重构图可视化

CLB-Defense 在查找到可疑的后门样本后，会利用图重要性指标对样本的结构进行重构，目的是过滤样本中的触发器。因此，在 PROTEINS 数据集和 NCI1 数据集上，面对 Motif-Backdoor 的后门攻击，本文采用 Gephi 工具可视化了 CLB-Defense 成功防御的后门样本，即重构图。如图 6 所示，良性图表示未添加任何扰动的图。后门图表示带有触发器的图。重构图表示经过 CLB-Defense 防御方法处理后的图。在后门图中，圈出了图中触发器的位置。对于重构图，虚线代表被删除的连边，加粗的实线代表增加的连边。在图 6(a)中的重构图中，删除的连边分别为节点 2 和节点 3、节点 5 和节点 6、节点 1 和节点 4 构建的连边，对应的边介数指标值分别为 0.001 8、0.003 3、0.003 6。增加的连边为节点 7 和节点 8 构建的连边，对应的共同邻居数指标值为 4。在图 6(b)中的重构图中，删除的连边为节点 2 和节点 4 构建的连边，对应的边介数指标值为 0.016 9。增加的连边为节点 1 和节点 5 构建的连边，对应的共同邻居数指标值为 2。

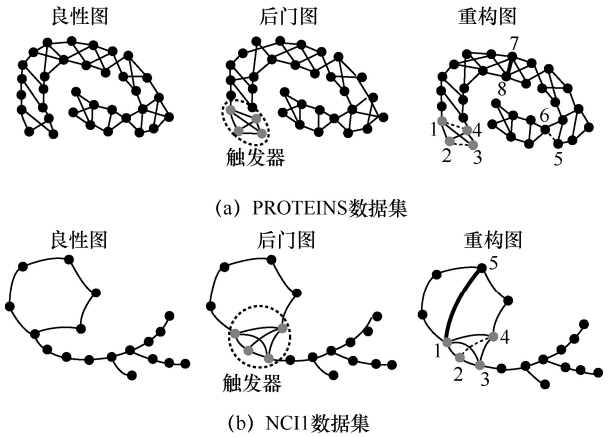


图 6 可视化 CLB-Defense 防御下重构图

从图 6 可以看出，CLB-Defense 能够破坏后门图中的触发器结构，删除后门图中扰动，使目标模型难以被后门攻击方法留下后门，从而起到防御的作用。这进一步验证了 CLB-Defense 中图重构模块的有效性，利用图重要性指标对样本的结构进行重

构，删除图中非正常相连的连边，补充图中可能存在的连边，有效地滤除后门样本中存在的触发器，进而防御图后门攻击方法。

4) 时间复杂度

本节对 Jaccard-Based、Label-Smooth、Adv-Training 和 CLB-Defense 进行时间复杂度的实验分析。在 4 个数据集上，记录了 CLB-Defense 防御方法面对 Motif-Backdoor 攻击方法时，实施防御所需要的时间，结果如表 3 所示。CLB-Defense 面对 4 个数据集实现防御平均运行时间为 127.07 s，而 Jaccard-Based、Label-Smooth、Adv-Training 的平均运行时间分别为 9.72 s、2.82 s、6.45 s。与其他防御方法相比，CLB-Defense 在实现更有效的防御性能同时，也需要消耗更长的时间。

表 3　　CLB-Defense 防御方法运行时间

防御方法	时间/s			
	PROTEINS	AIDS	NCI1	DBLP_v1
Jaccard-Based	4.88	2.36	6.08	25.55
Label-Smooth	1.34	0.93	2.14	6.86
Adv-Training	2.62	1.45	6.76	14.98
CLB-Defense	130.48	116.7	125.63	135.45

为了进一步研究 CLB-Defense 防御方法的时间复杂度，分析了 CLB-Defense 各个模块所耗费的时间，结果如图 8 所示。从图 8 可知，构建对比模型训练模块是 CLB-Defense 耗费时间占比最大的模块，在 PROTEINS、AIDS、NCI1、DBLP_v1 的占比分别为 91.69%、99.51%、94.79%、97.38%。而 CLB-Defense 在完成对比模型训练之后，在防御过程中不需要再进行对比模型训练，这意味着该模块的时间不会随着检测样本数量的增多而增多，是可控的常量。

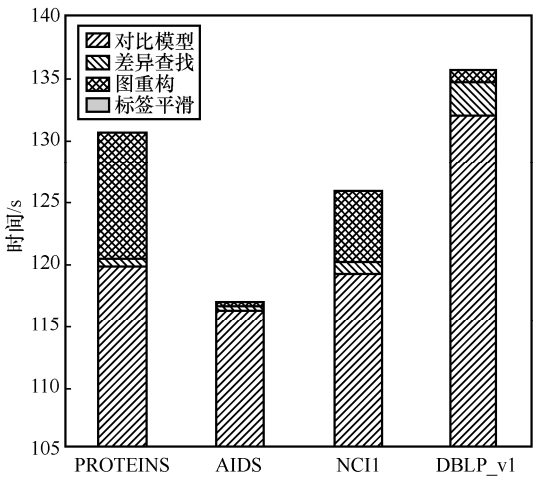


图 8　CLB-Defense 各模块运行时间

5　结束语

为了缓解图神经网络后门攻击的威胁，本文提出了一种基于对比学习的图神经网络后门攻击防御方法 CLB-Defense，利用对比学习构建对比模型，基于输出置信分数的差值查找可疑后门样本，然后采用图重要性指标重构训练数据中可疑后门样本的结构，采用标签平滑策略对可疑后门样本的标签进行重塑。同时，在 4 个真实的数据集上展开丰富的防御实验，验证了 CLB-Defense 面对多样性后门攻击方法情况下防御的有效性，且不影响良性模型的正常表现性能。此外，与其他的防御方法相比，CLB-Defense 方法能够实现更有效的防御性能，但也存在更高复杂度的问题。为此，后续的研究工作将聚焦于实现更高效的后门攻击防御方法，进一步降低图神经网络后门攻击带来的影响，提升图神经网络的鲁棒性。

参考文献：

[1] 王璿, 张瑜, 周军锋, 等. 基于社交网络的影响力最大化算法[J]. 通信学报, 2022, 43(8): 151-163.
WANG X, ZHANG Y, ZHOU J F, et al. Influence maximization algorithm based on social network[J]. Journal on Communications, 2022, 43(8): 151-163.

[2] 任永功, 张云鹏, 张志鹏. 基于粗糙集规则提取的协同过滤推荐算法[J]. 通信学报, 2020, 41(1): 76-83.
REN Y G, ZHANG Y P, ZHANG Z P. Collaborative filtering recommendation algorithm based on rough set rule extraction[J]. Journal on Communications, 2020, 41(1): 76-83.

[3] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[J]. arXiv Preprint, arXiv: 1810.00826, 2018.

[4] QIU J Z, CHEN Q B, DONG Y X, et al. GCC: graph contrastive coding for graph neural network pretraining[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1150-1160.

[5] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 5812-5823.

[6] SURESH S, LI P, HAO C, et al. Adversarial graph augmentation to improve graph contrastive learning[J]. arXiv Preprint, arXiv: 2106.05819, 2021.

[7] ZHANG Z X, JIA J Y, WANG B H, et al. Backdoor attacks to graph neural networks[C]//Proceedings of the 26th ACM Symposium on Access Control Models and Technologies. New York: ACM Press, 2021: 15-26.

[8] XU J, XUE M, PICEK S. Explainability-based backdoor attacks against graph neural networks[C]//Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2021: 31-36.

[9] SHENG Y, CHEN R, CAI G Y, et al. Backdoor attack of graph neural

networks based on subgraph trigger[C]//International Conference on Collaborative Computing: Networking, Applications and Worksharing. Berlin: Springer, 2021: 276-296.

[10] ZHENG H, XIONG H, CHEN J, et al. Motif-backdoor: rethinking the backdoor attack on graph neural networks via motifs[J]. arXiv Preprint, arXiv: 2210.13710, 2022.

[11] XI Z H, PANG R, JI S L, et al. Graph backdoor[C]//Proceedings of the 30th USENIX Security Symposium. Berkeley: USENIX Association, 2021: 1523-1540.

[12] ZENG Y, PARK W, MAO Z M, et al. Rethinking the backdoor attacks' triggers: a frequency perspective[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 16453-16461.

[13] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv Preprint, arXiv: 1811.03728, 2018.

[14] WANG B L, YAO Y S, SHAN S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 707-723.

[15] HASSANI K, KHASAHMADI A H. Contrastive multi-view representation learning on graphs[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press, 2020: 4116-4126.

[16] ZHU Y Q, XU Y C, YU F, et al. Graph contrastive learning with adaptive augmentation[C]//Proceedings of the Web Conference 2021. New York: ACM Press, 2021: 2069-2080.

[17] YOU Y N, CHEN T L, WANG Z Y, et al. Bringing your own view: graph contrastive learning without prefabricated data augmentations[C]//Proceedings of the International Conference on Web Search & Data Mining International Conference on Web Search & Data Mining. New York: ACM Press, 2022: 1300-1309.

[18] 窦家维, 葛雪, 王颖囡. 保护隐私的曼哈顿距离计算及其推广应用[J]. 计算机学报, 2020, 43(2): 352-365.
DOU J W, GE X, WANG Y N. Secure Manhattan distance computation and its application[J]. Chinese Journal of Computers, 2020, 43(2): 352-365.

[19] 黄海平, 王凯, 汤雄, 等. 基于边介数模型的差分隐私保护方案[J]. 通信学报, 2019, 40(5): 88-97.
HUANG H P, WANG K, TANG X, et al. Differential privacy protection scheme based on edge betweenness model[J]. Journal on Communications, 2019, 40(5): 88-97.

[20] NEWMAN M J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.

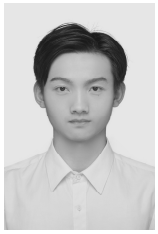
[21] WU H, WANG C, TYSHETSKIY Y, et al. Adversarial examples on graph data: deep insights into attack and defense[J]. arXiv Preprint, arXiv: 1903.01610, 2019.

[22] LUKASIK M, BHOJANAPALLI S, MENON A K, et al. Does label smoothing mitigate label noise? [C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press, 2020: 6448-6458.

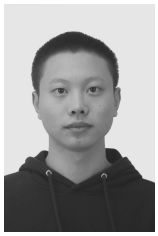
[23] TRAMÈR F, BONEH D. Adversarial training and robustness for multiple perturbations[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2019: 5866-5876.

[作者简介]

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



熊海洋（1998– ），男，江西南昌人，浙江工业大学硕士生，主要研究方向为深度学习、人工智能安全和图数据挖掘。



马浩男（2000– ），男，浙江杭州人，浙江工业大学硕士生，主要研究方向为深度学习、人工智能安全和图数据挖掘。

郑雅羽（1978– ），男，浙江温州人，博士，浙江工业大学副教授、硕士生导师，主要研究方向为嵌入式软硬件应用开发、视频图像处理算法、服务器网络技术等。