Hans Journal of Data Mining 数据挖掘, 2025, 15(4), 295-309
Published Online October 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.154026

Hans汉斯

LE-DCBFD：基于图神经网络的链路增强带Dice损失的均衡一致欺诈检测器

黄辉林*，范永希，郑迪宇

温州大学数理学院，浙江 温州

收稿日期：2025年9月16日；录用日期：2025年10月9日；发布日期：2025年10月20日

摘　要

图神经网络(GNNs)因其卓越的数据表征能力以及探索社交网络中复杂关系的能力，已被广泛应用于欺诈检测领域。本文提出了一种新颖的图数据增强算法——半可学习启发式链接预测算法，该算法利用丰富的标签信息来解决因数据缺失和人为操纵(如欺诈伪装)导致的网络信息丢失问题。基于此算法，本文提出了一种欺诈检测模型：基于图神经网络的链路增强的带Dice损失的均衡一致欺诈检测器(LE-DCBFD)。在两个公开的真实世界欺诈检测数据集(Amazon和Yelp)上对LE-DCBFD模型进行了评估。结果表明，本文的模型优于多个基线模型，在规模更大的Yelp数据集上，欺诈检测性能提升了超过10%。在消融实验中，它也优于未采用本文所提出的链接增强器的DCBFD模型，这证实了链接增强器对性能提升的重要性。即使在使用较小的训练数据集时，LE-DCBFD也展现出优越性，证明它比DCBFD更有效。

关键词

图神经网络，欺诈检测，数据表征，复杂关系，链路预测

LE-DCBFD: Link Enhanced Dice-Loss Consistent Balanced Fraud Detector Using Graph Neural Networks

Huilin Huang*, Yongxi Fan, Diyu Zheng

School of Mathematics and Physics, Wenzhou University, Wenzhou Zhejiang

Received: September 16, 2025; accepted: October 9, 2025; published: October 20, 2025

*通讯作者。

文章引用: 黄辉林, 范永希, 郑迪宇. LE-DCBFD: 基于图神经网络的链路增强带 Dice 损失的均衡一致欺诈检测器[J]. 数据挖掘, 2025, 15(4): 295-309. DOI: 10.12677/hjdm.2025.154026

Abstract

Graph Neural Networks (GNNs) have been widely adopted in fraud detection due to their exceptional data representation capabilities and their ability to explore complex relationships in social networks. This paper introduces a novel graph data augmentation algorithm, the semi-learnable heuristic link prediction algorithm, which leverages rich label information to address network information loss caused by insufficient data and artificial manipulation, such as fraud camouflage. Based on this algorithm, we propose a fraud detection model: Link Enhanced Dice-loss Consistent Balanced Fraud Detector (LE-DCBFD). We evaluated the LE-DCBFD model on two public real-world fraud detection datasets, Amazon and Yelp. The results show that this model outperforms multiple baseline models, with the fraud detection performance on the larger Yelp dataset improving by over 10%. In the ablation experiments, it also surpasses the DCBFD model without our proposed Link Enhancer (a link prediction algorithm), which confirms the importance of the Link Enhancer for performance improvement. Even when using a smaller training dataset, LE-DCBFD demonstrates superiority, proving that it is more effective than DCBFD.

Keywords

Graph Neural Network, Fraud Detection, Data Representation, Complex Relationship, Link Prediction

1. 引言

欺诈是蓄意误导他人制造虚假认知的行为，虚拟网络中犯罪分子常借此谋取非法利益。近年来，欺诈在各行业愈发普遍且手段多样，包括谣言传播[1] [2]、电信诈骗[3]、金融诈骗[4] [5]、保险诈骗[6] [7]等。虽欺诈发生频率低于合法活动，但据 ACFE 研究，专业欺诈预计使企业损失 150 万美元，平均占企业年收入的 5% [8]，其影响造成巨大经济和社会损失，故欺诈检测研究紧迫且重要。为此，研究人员利用图信息表示和提取能力丰富欺诈数据表示形式，其中图神经网络(GNN)算法也在欺诈检测领域广泛应用。

基于同质图聚合模型的方法被广泛应用于欺诈检测。三种经典的同质图聚合模型包括图卷积网络(GCN) [9]、将注意力机制集成到 GCN 框架中的图注意力网络(GAT) [10]，以及基于 GCN 原理通过邻居采样来聚合特征的 GraphSAGE [11]。许多基础模型已被调整以适应特定的欺诈检测任务。例如，FdGars 模型[12]在提取日志特征后使用两层 GCN 进行学习。SemiGNN [13]通过节点级和视图级聚合器来融合节点－邻居信息。此外，胡等人[14]提出了一种端到端的桥接图(BTG)来解决电信诈骗连接稀疏的问题，核心在于利用用户的协同行为来重建连接。

然而，上述算法仍难以应对欺诈者与检测者之间的动态博弈关系，因此一直面临着来自异质信息干扰和标签不平衡的挑战。刘等人[15]指出，欺诈者擅长伪装，包括特征操纵和关系伪造(窦等人[16])。杨等人[17]也指出，社交媒体谣言传播者善于使用表情符号和文本缩写来躲避检测者。刘等人[18]强调，这些伪装会导致邻居信息不一致，包括特征不一致和节点标签不一致。此外，与合法活动相比，欺诈行为相对较少，这加剧了标签不平衡的问题。这种不平衡使得基于图神经网络的识别器在训练过程中倾向于合法类别，以优化整体准确率，这与准确、全面识别欺诈实体的目标相悖。许多神经网络模型试图通过

平滑邻居嵌入表示来解决这些问题(高等人[19])，但这种方法也会破坏特征中的有用信息，从而影响最终决策(唐等人[20])。

研究人员已探索了多种方法，以应对欺诈检测中异质信息干扰和标签不平衡等挑战。部分方法包括去除冗余连接、采用相似性采样，以及运用异质图表示技术。如 MHGSL [21]和 Player2Vec [22]模型利用图卷积网络(GCN)，通过采用多条元路径来整合不同的信息源。相比之下，GraphConsis [15]和 CARE-GNN [16]等模型则专注于缓解伪装导致的邻域特征不一致问题。GraphConsis 运用了上下文嵌入、邻居采样和加权注意力机制等技术，而 CARE-GNN 则使用了相似性感知邻居选择器和关系感知邻居聚合器。这两种模型在解决特征不一致问题上均展现出显著效果。本研究提出的模型整合了上述两种模型的关键模块，并创新性地引入了新模块，以更好地处理特征不一致问题。因此，本文提出的模型旨在实现特征一致性，这也体现在其名称“特征一致欺诈检测器(FCFD)”上。

此外，CFTNet [23]采用三元组网络反事实方法增强数据，用于信用卡欺诈检测。PCGNN [13]和 SCN_GNN [24]通过不同的采样策略缓解了欺诈检测中普遍存在的类别不平衡问题。IDGL [25]引入了一种可学习的双通道图卷积滤波器，并设计了标签感知节点和边采样器，以解决类不平衡问题。采样技术是解决类别不平衡问题的有效手段之一。本文采用简单随机下采样方法来实现类别间的平衡，使所提出的欺诈检测模型能够有效处理类别不平衡问题。顾名思义，该模型旨在实现平衡的欺诈检测。结合本文提出的网络架构，即使训练样本仅占总样本量的 5%，也能获得具有竞争力的算法性能，有助于降低相关成本。

受李等人[26]提出的样本增强策略以及可学习启发式链接预测算法 WLNM [27]的启示，我们旨在设计一个链接预测模块，以解决数据不足和欺诈者伪装的问题。因此，本文深入研究了现有的链接预测模型，发现链接预测的探索始于启发式算法。Islam 等人[28]总结了基于节点对相似度的启发式链接预测方法。这些方法包括经典的方法，如共同邻居(CN) [29]、资源分配(RA) [30]和杰卡德指数(JA) [31]。此外，张和陈指出启发式链接预测受预设条件的限制，并提出了两种可学习的启发式链接预测算法，即 WLNM [27]和一种新的 γ-衰减启发式理论(SEAL) [32]。WLNM 引入了一种基于颜色细化的快速哈希 Weisfeiler-Lehman (WL)图标记算法，该算法用于链接预测，从网络拓扑结构中学习启发式链接规则。SEAL 将广泛的启发式方法统一到一个框架中，并证明了局部子图保留了丰富的与链接相关的信息。

综上所述，现有的链接预测算法难以直接应用于欺诈检测任务，可学习的启发式链接预测算法仅关注网络拓扑信息，无法从标签一致角度有效提取链接。为此，本研究目标是开发半可学习的启发式链接预测算法，将传统启发式方法与深度学习技术结合，优化标签信息嵌入以获取标签一致的节点特征，运用启发式原则进行链接预测，期望识别有意义的链接改善特征异质性、缓解特征模糊问题。受 CARE-GNN [16]标签感知相似度度量方法启发，对优化后的节点进行链接挖掘，最终构建类别平衡、特征和标签一致的欺诈检测模型，类别平衡通过简单随机下采样实现，特征和标签一致性通过结合提出的链接增强器与现有邻居消除器和邻居采样器达成。

本文的主要贡献如下：

1) 提出了一种名为链接增强器(Link Enhancer)的可学习半启发式链接预测算法。该算法将启发式链接预测与反向传播神经网络相结合，并利用标签信息优化节点–链接嵌入表示。通过学习得到的高质量链接嵌入表示用于挖掘有效链接。

2) 将链接增强器应用于欺诈检测，并提出了一种带有链接预测功能的欺诈检测器——LE-DCBFD。链接增强器通过挖掘强相似链接来增强原始图网络，以解决因数据丢失或欺诈性伪装导致的特征不一致问题。

3) 提出了一种组合损失函数，即交叉熵损失和 Dice 损失的组合，旨在提高模型的抗干扰能力和识别准确率。

4) 在 Amazon 和 Yelp 等两个公开数据集上进行了实验，结果表明所提出的方法是有效的，并达到了当前的先进性能水平。

本文的内容结构安排如下：第 1 节阐述研究背景并对该领域的相关工作进行综述。随后，在第 2 节中，详细阐述与本研究相关的基本定义，包括问题的阐释和符号表示的说明。第 3 节深入全面地介绍模型组件，展示完整的网络框架图。第 4 节公布相关实验结果，包括模型性能的对比分析、敏感性评估以及消融实验。最后，第 5 节总结本文得出的关键结论，并指明未来研究工作的方向。

2. 问题定义

定义 1. 异质图。定义一个异质图 $\mathcal{G}=\left(\mathcal{V},\mathcal{X},\{\mathrm{E}_r\}\middle|_{r=1}^{R},Y\right)$。其中，$\mathcal{V}=\{v_1,v_2,\cdots,v_n\}$ 表示图中的节点集合，每个节点对应一个待分类的观测对象。$\mathcal{X}=\{x_1,x_2,\cdots,x_n\}$ 表示节点的特征属性集合。ε_r 表示关系 r 下的边集合，每条边用 $e_{u,v}^{r}$ 表示。如果节点 v 和 u 在关系 r 下存在链接，则 $e_{u,v}^{r}=1$。Y 表示 V 中每个节点的标签组成的集合。

定义 2. 链接预测。给定一个异质图 $\mathcal{G}$，其中 $\mathcal{V}$ 表示顶点集合，ε 表示边集合，链接预测的任务是估计每对未连接节点 $u,v\in\mathcal{V}$ 形成边的可能性 $(u,v)\notin\mathcal{E}$。通过应用一个阈值来识别潜在的新链接，这个过程会更新原始的连接边集合 $\varepsilon_r^{(l)}$，得到一个增强的连接边集合 $\varepsilon_r^{(l)}$。

定义 3. 欺诈检测。利用图网络信息，包括节点属性和邻居关系，进行一个半监督的二分类任务。目标是找到一个映射函数 $f\to y_v\in\{0,1\}$ 来完成节点分类预测。类别 0 代表良性用户，而类别 1 代表欺诈者。

为了更好地说明本文提出的算法，表 1 列出了关键元素的符号和定义，这些将在后续章节中使用。

Table 1. Symbol description
表 1. 符号说明

符号	描述
$\mathcal{G};\ \mathcal{V};\ \mathcal{X};\mathcal{E}$	图；节点集；节点特征集；边集
$R;L;B;E;N$	关系总数；网络层数；批次数；总训练轮数；采样邻居数量
$\mathcal{V}train;\mathcal{V}b$	训练节点集；当 $B=b$ 时的点集
$y_v;Y$	节点 v 的真实标签；节点标签集
$\mathcal{E}_r^{(l)}$	第 l 层关系 r 下的边集
$S_{link}^{(l)}(v,v')$	链接增强器中节点 v 和节点 v' 的相似度得分
$\mathcal{L}_{link}$	优化链接预测嵌入表示的损失函数
Neighbor$AGG_r^{(l)}$	第 l 层关系 r 下的邻居聚合器
$AttAGG^{(l)}$	第 l 层的注意力机制聚合器
$S_{fr}^{(l)}(u,v)$	邻居消除器中第 l 层关系 r 下节点 v 和 u 之间的相似度得分
$p_r^{(l)}(u,v)$	第 l 层关系 r 下节点 v 和 u 之间的采样概率
$h_{v,r}^{(l)}$	第 l 层关系 r 下节点 v 的嵌入

续表

$h_v^{(l)}$	第 l 层节点 v 的嵌入
$\alpha_r^{(l)}$	第 l 层关系 r 下的注意力得分
$z_v^{(l)}$	节点 v 的最终嵌入
$\mathcal{L}_{LE-DCBFD}$	模型 LE-DCBFD 的损失函数

3. 提出模型

受 CARE-GNN [15]中标签感知相似度度量的启发，本文提出采用链接预测来寻找潜在的高价值链接以增强原始图信息。具体而言，利用标签信息优化节点嵌入，随后使用优化后的嵌入表示来计算链接相似度，最后选择相似度高的链接来扩展图网络的邻接关系。通过运用链接增强器进行欺诈检测，提出了 LE-DCBFD 模型。

3.1. 神经网络框架

本文链接增强器处理过程如图 1 所示：将节点原始关系和特征输入多层感知机(MLP)生成隐藏链接嵌入与预测输出，计算链接嵌入向量相似度得分，整合得分排名前 10%的链接增强原始网络，利用预测与实际标签计算损失并反向传播更新参数，提升模型一致性。完整模型框架如图 2 所示，其中图 1 展示于指定橙色矩形框内，其基本流程为：先算节点与一阶邻居相似度得分，用链接增强器对原始图网络数据增强，将增强图输入邻居消除器去除高异质性关系，归一化相似度得分需采样概率，用邻居别名采样器获取中心节点最终邻接关系，通过邻居聚合器跨关系合并邻居信息生成节点各关系下嵌入表示，用注意力机制聚合不同关系邻居信息获取每层嵌入表示，经 L 层操作后将嵌入表示输入 MLP 分类器产生预测结果，算损失反向传播优化模型。为评估链接预测算法有效性，设计不使用链接增强器的框架 DCBFD，在图 2 中用蓝色矩形框表示。

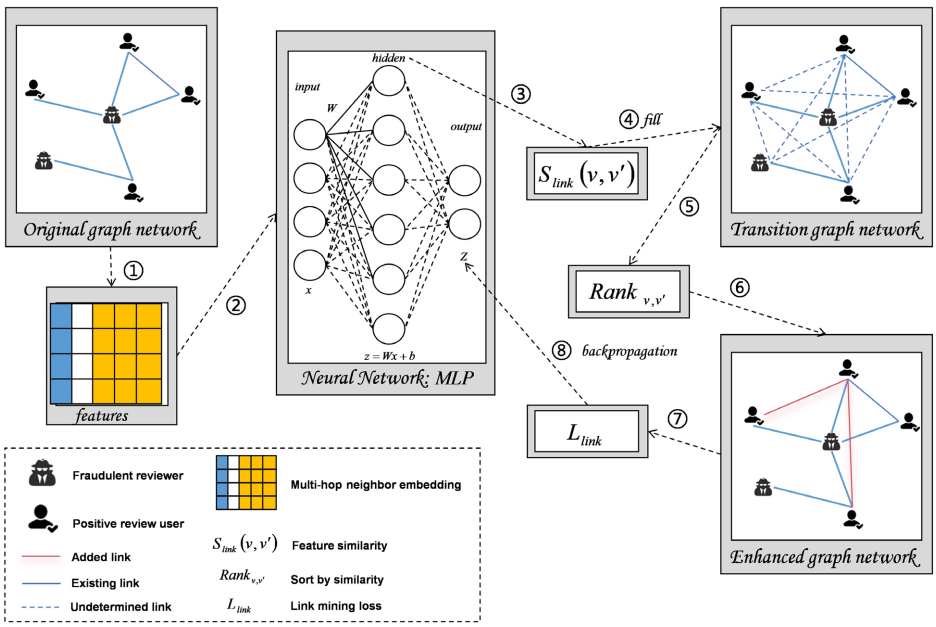

Figure 1. Schematic diagram of the link prediction module
图 1. 链路预测模块示意图

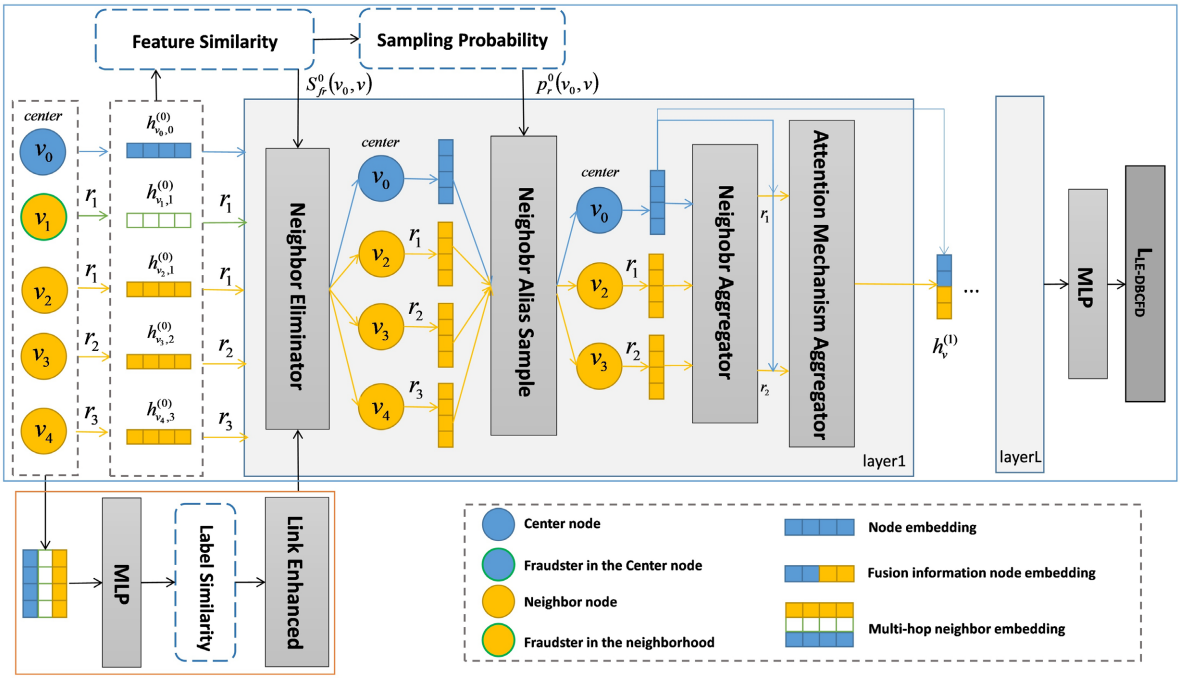

Figure 2. The model framework of this paper
图 2. 本文的模型框架

3.2. 模块分解

3.2.1. 链接增强器

引入链接增强器的目的是通过扩展缺失的链接来增强图网络数据，以应对数据遗漏或人工干预带来的挑战。具体方法如下：

首先，节点特征通过多层感知机(MLP)进行处理，并由激活函数激活，以获得链接嵌入表示。接下来，基于这些链接嵌入表示计算链接相似度，具体如公式(1)所示。

$$S_{link}^{(l)}(v,v') = 1 - \mathcal{D}^{(l)}(v,v') \tag{1}$$

$$\mathcal{D}^{(l)}(v,v') = \left\| ReLU\left(MLP\left(embed_{v,r}^{(l)}\right)\right) - ReLU\left(MLP\left(embed_{v',r}^{(l)}\right)\right) \right\|_1 \tag{2}$$

在公式(1)中，$S_{link}^{(l)}(v,v')$ 表示第 l 层节点 v 和节点 v' 之间的链接相似度。$\mathcal{D}^{(l)}(v,v')$ 表示第 1 层节点 v 和节点 v' 嵌入表示之间的距离，该距离使用 L_1 范数进行度量，如公式(2)所定义。在公式(2)中，*embed* 表示第 l 层关系 r 下节点 v 的链接嵌入表示，初始链接嵌入使用节点的属性特征表示。*MLP* 指的是多层感知机，激活函数定义为 $\mathrm{ReLU}(x) = \max(0,x)$。

随后，将计算得到的链接相似度按降序排序以确定链接排名。通过将此排名与预设的排名阈值进行比较，选择排名前 $\left\lceil r_{le} * \left|\mathcal{E}_r^{(l)}\right| \right\rceil$ 的链接来增强图网络内的连接。在本文中，链接预测的数量基于该关系内现有链接的数量，并根据特定比例进行计算，该比例被视为一个超参数。

$$rank_{v,v'} = Sorted\left(S_{link}^{(l)}(v,v')\right) \tag{3}$$

$$\tilde{\mathcal{E}}_r^{(l)}(v,v') = \begin{cases} 1, \text{if } rank_{v,v'} \le \left\lceil r_{le} * \left|\mathcal{E}_r^{(l)}\right| \right\rceil, \\ 0, \qquad \text{otherwise.} \end{cases} \tag{4}$$

其中，$[\cdot]$表示向下取整函数，$\left|\mathcal{E}_r^{(l)}\right|$表示第$l$层关系$r$下现有链接的数量，$\tilde{\mathcal{E}}_r^{(l)}$表示链接预测后得到的连接边集合，$r_{le}$表示链接预测的比例系数。具体的链接预测过程详见公式(4)。

最后，我们通过交叉熵损失函数纳入节点标签信息，该函数用于训练更高质量的链接嵌入。损失函数记为$\mathcal{L}_{link}$，其公式如公式(5)所示。在3.2.6节中，我们引入一个加权因子将其纳入最终的优化目标函数。

$$\mathcal{L}_{link} = \sum_{v\in\mathcal{V}} -\log\left(y_v \cdot \sigma\left(MLP\left(embed_v\right)\right)\right) \tag{5}$$

其中，embed_v表示节点v的最终链接嵌入表示。在本文中，激活函数σ均为如公式(6)所定义的LeakyReLU函数。

$$\text{LeakyReLU}\left(x\right) = \begin{cases} x, & x > 0, \\ ax, & x \le 0. \end{cases} \tag{6}$$

3.2.2. 邻居剔除器

该模块采用Graphconsis [14]提出的方法来计算第l层节点之间的相似度，如公式(7)所示。通过设置一个相似度阈值，我们为中心节点过滤掉一些邻居。

$$S_{fr}^{(l)}\left(u,v\right) = \exp\left(-\left\|\mathbf{h}_u^{(l-1)} - \mathbf{h}_v^{(l-1)}\right\|_2^2\right) \tag{7}$$

其中，$S_{fr}^{(l)}\left(u,v\right)$的定义见表1。

对于相似度低于阈值的两个节点，其邻接关系将被消除；否则，该关系将被保留。对第1层关系r下的所有节点完成邻居剔除后，得到一个新的邻接边集合，如公式(8)所示。

$$\tilde{\mathcal{E}}_r^{(l)}\left(u,v\right) = \begin{cases} 1, \text{if } S_{fr}^{(l)}\left(u,v\right) \le \Theta, \\ 0, \quad \text{otherwise.} \end{cases} \tag{8}$$

其中，$\tilde{\mathcal{E}}_r^{(l)}$表示第$l$层关系$r$经过邻居排除过程过滤后的新邻接边集合。若链接被保留其值为1，否则为0。Θ表示相似度阈值。

3.2.3. 邻居别名采样器

为中心节点进行邻居采样时，通过对经过修剪后的邻居节点集的相似度得分进行归一化处理，以得到采样概率。计算如公式(9)所示。

$$p_r^{(l)}\left(u,v\right) = \frac{S_{fr}^{(l)}\left(u,v\right)}{\sum_{u,v\in\tilde{\mathcal{E}}_r^{(l)}} S_{fr}^{(l)}\left(u,v\right)} \tag{9}$$

其中，$p_r^{(l)}\left(u,v\right)$表示在第$l$层关系$r$下，中心节点$v$的采样邻居节点$u$的概率。

在本文中，采用了Schwartz博客文章[33]中公布的别名采样方法，以提高采样的效率和准确性。引入别名采样来优化图神经网络的采样策略，旨在有效获取高质量的邻居样本，以进一步应对欺诈伪装问题。与传统的均匀分布或随机采样方法相比，别名采样基于节点之间的相似度构建别名表，使采样概率更符合节点相似度的分布。因此，使用别名邻居采样能够保留重要的邻居节点信息。此外，与复杂度为$O\left(n\right)$的均匀采样策略以及平均复杂度为$O\left(\log n\right)$、最坏情况下复杂度高达$O\left(n\right)$的基于二叉搜索树(BST)的采样策略[34]相比，别名采样算法的复杂度仅为$O\left(1\right)$，这显著提升了模型的性能和训练效率。

3.2.4. 邻居聚合器

获取邻居节点的采样概率后，将其用以表示中心节点与其邻居之间的紧密程度，由此通过邻居聚合器为中心节点实现更一致的特征表示。在这个过程中，采用逐元素求和的方式进行邻居聚合。第 l 层关系 r 下的邻居聚合器定义如公式(10)所示。

$$\begin{aligned}\boldsymbol{h}_{v,r}^{(l)} &= \sigma\left(\boldsymbol{h}_{v,r}^{(l-1)} \oplus WeightAGG_r^{(l)}\left(\left\{\boldsymbol{h}_{v',r}^{(l-1)} : (v,v') \in \tilde{\mathrm{E}}_r^{(l)}\right\}\right)\right) \\ &= \sigma\left(\boldsymbol{h}_{v,r}^{(l-1)} + \sum_{(v,v')\in \mathrm{E}_r^{(l)}} w_r^{(l)}(v,v') \cdot \boldsymbol{h}_{v',r}^{(l-1)}\right),\end{aligned} \tag{10}$$

其中，$\oplus$ 表示加法聚合；$w_r^{(l)}(v,v')$ 是第 l 层关系 r 下节点 v 和节点 v' 之间的信息聚合权重。

3.2.5. 注意力机制聚合器

经过邻居聚合器获取中心节点在同一关系下邻居节点的嵌入表示。目标转向解决不同关系之间的不一致性问题。为此我们引入注意力机制来计算各种关系的注意力得分，并将得分作为邻居聚合的权重。此外，在聚合过程中采用逐元素求和的方式。因此，第 l 层中心节点的嵌入表示如公式(11)所示。

$$\begin{aligned}\mathbf{h}_v^{(l)} &= \sigma\left(\mathbf{h}_v^{(l-1)} \oplus AttAGG^{(l)}\left(\mathbf{h}_{v,r}^{(l-1)}\right)\right) \\ &= \sigma\left(\mathbf{h}_v^{(l-1)} + \sum_{r=1}^{R} \alpha_r^{(l)} \mathbf{h}_{v,r}^{(l-1)}\right)\end{aligned} \tag{11}$$

其中，$\oplus$ 表示基于加法的聚合。$\alpha_r^{(l)}$ 表示第 l 层关系 r 的注意力得分，其定义如公式(12)所示。

$$\alpha_r^{(l)} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}^{(l)} \,\middle\|\, \mathbf{W}\vec{h}_r^{(l)}\right]\right)\right)}{\sum_{r\in\{1,\cdots,R\}} \exp\left(\sigma\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}^{(l)} \,\middle\|\, \mathbf{W}\vec{h}_r^{(l)}\right]\right)\right)} \tag{12}$$

在公式(12)中，$\vec{\boldsymbol{a}}$ 和 $\boldsymbol{W}$ 是网络的可学习参数，“||”表示矩阵的水平拼接。

3.2.6. 分类器与优化目标

本文采用多层感知机(MLP)对事先得到的嵌入表示向量进行训练，进而预测分类结果。为增强检测器对难以区分类别的关注，本文提出如下组合损失函数：

首先将 Dice 损失[35]与交叉熵损失函数相结合，分别引入超参数 λ_1 和超参数 λ_2 来调控 Dice 损失和 $\mathcal{L}_{link}$ 对最终训练的影响。由此得到 LE-DCBFD 算法的损失函数，记为 $\mathcal{L}_{LE-DCBFD}$，如公式(13)所示。

$$\mathcal{L}_{LE-DCBFD} = \sum_{v\in\mathcal{V}} -y_v \log\left(\sigma\left(MLP\left(z_v\right)\right)\right) + \lambda_1 \cdot Loss_{Dice} + \lambda_2 \cdot \mathcal{L}_{link} \tag{13}$$

其中，y_v 表示节点 v 的真实标签，z_v 表示最终输出的节点 v 的嵌入表示。

$$Loss_{Dice} = \left(1 - \frac{2\sum_{v\in\mathcal{V}} y_v \cdot \hat{y}_v}{\sum_{v\in\mathcal{V}} y_v + \sum_{v\in\mathcal{V}} \hat{y}_v + smooth}\right) \tag{14}$$

其中，$\hat{y}_v$ 表示节点 v 的预测标签，smooth 的引入是为防止除零错误发生，通常设置为正整数，本文设置为 10。

特别地，当 λ_2 被设置为 0 且相邻边集合保持为原始集合 $\mathcal{E}_r^{(l)}$ 时，LE-DCBFD 模型不包含链接增强器，此时该模型简称为 DCBFD。

4. 实验

4.1. 实验准备

4.1.1. 数据描述

本研究在经典的欺诈检测数据集 Yelp [36]和 Amazon [37]上对 LE-DCBFD 模型进行了实证测试。两个数据集具有不同的多关系连接以及极度类别不平衡的欺诈数据，已有大量的前期研究可供比较，并且由于它们在商业领域的重要性，因此受到学术界和工业界的广泛关注。Yelp 数据集包含 Yelp 上关于酒店和餐厅评论的垃圾评论和合法评论，其中约 14.5%为垃圾评论。每条评论生成一个 32 维的特征向量。Amazon 数据集包含乐器类别的产品评论，其中只有约 9.5%为垃圾评论。该数据集中的每条评论生成一个 25 维的特征向量。

Table 2. Descriptive statistical data of Yelp and Amazon
表 2. Yelp 和 Amazon 的描述性统计数据

数据集	节点(欺诈%)	关系	边	$AvgS^f$	$AvgS^l$
Yelp	45,954 (14.5%)	ALL	3,846,979	0.77	0.07
		R-U-R	49,315	0.83	0.09
		R-T-R	573,616	0.79	0.05
		R-S-R	3,402,743	0.77	0.07
Amazon	11,944 (9.5%)	ALL	4,398,392	0.65	0.05
		U-P-U	175,608	0.61	0.19
		U-S-U	3,566,479	0.64	0.04
		U-V-U	1,036,737	0.71	0.03

如表 2 所示，在 Yelp 数据集中，欺诈性评论者在多个维度上存在关联，包括用户、产品、评论文本和时间方面。在图表示中，评论被表示为节点，并建立了三种不同类型的关系：1) R-U-R：该关系连接同一用户发布的评论；2) R-S-R：该关系连接对同一产品给出相同星级评分(1~5 星范围内)的评论；3) R-T-R：该关系连接在同一个月内发布的对同一产品的两条评论。

对于 Amazon 数据集，用户在图中被视为节点，并建立了与 Yelp 数据集类似的三种关系，包括：1) U-P-U：该关系连接至少对一款相同产品进行过评论的用户；2) U-S-V：该关系连接在一周内至少有一次相同星级评分的用户；3) U-V-U：该关系连接在评论文本相似度方面排名前 5%的所有用户，注意相似度使用 TF-IDF 进行度量。此外，表格引用了 Graph Cosis [5]提出的平均特征相似度和平均标签相似度公式，如公式(15)和公式(16)所示。

$$AvgS_r^{(f)} = \sum_{(u,v)\in E_r} \frac{\exp\left(-\left\|x_u - x_v\right\|_2^2\right)\cdot d}{\left|E_r\right|} \tag{15}$$

$$AvgS_r^{(l)} = \sum_{(u,v)\in E_r} \frac{\left(1-\mathbb{I}\left(u \sim v\right)\right)}{\left|E_r\right|} \tag{16}$$

在公式(15)中，d 代表特征的维度。在公式(16)中，$\mathbb{I}(\cdot)$ 表示指示函数，若节点 v 和 u 的标签不同，则该函数值为 1，否则为 0。

对比表 2 的相似度结果表明，两个数据集不仅在特征上存在差异，在标签上也存在差异。进一步说明存在欺诈者隐藏在良性用户之中并伪装其身份现象。此外，为解决标签不平衡问题，本研究在进行实验前对数据进行了欠采样处理。

4.1.2. 实验设置

通过在 Python 3.9.13 上实现 LR [38]、SVC [39]和 DT [40]模型。对于包括 GCN [9]、GAT [10]、RGCN [41]、GraphSAGE [11]、GeniePath [42]、Player2Vec [22]、SemiGNN [13]和 GraphConsis [14]在内的基线模型，主要参考了 CARE-GNN [15]论文中的实验结果。PCGNN [17]模型以及本文提出的 LE-DCBFD 模型均在 PyTorch 中实现，并同样在 Python 3.9.13 上运行。实验所用的计算机配置如下：16 GB 内存、第 12 代英特尔(R)酷睿(TM)i5-12500 3.00 GHz 处理器、64 位 Windows 操作系统。

4.1.3. 性能指标

本文提出了一个用于欺诈检测的图节点分类问题。因此，我们选择了两个经典的分类性能指标：ROC-AUC [43](简称 AUC)和召回率[44]。AUC 指标通过计算受试者工作特征(ROC)曲线下的面积来评估二分类模型的性能，该面积表示正样本排名高于负样本的概率。AUC 的值介于 0 到 1 之间，AUC 值越高表示分类准确率越好。召回率指标衡量模型正确预测正样本的准确性。召回率值越高，表示在识别正实例方面的性能越好，因此它是评估分类模型的重要指标之一。召回率指标由公式(17)定义。

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

其中，TP 和 FN 均源自混淆矩阵。TP 表示真正例的数量，即被正确预测为正类的正样本；FN 表示假负例的数量，即被错误预测为负类的正样本。

4.1.4. 实验参数设置

Table 3. Model parameter settings
表 3. 模型参数设置

	rle	hop	λ_1	λ_2	Θ	num_neigh	embed_dim	out_dim
Amazon	[1, 1, 1]	3	3	0.5	0.7	2	32	128
Yelp	[0.005, 0.005, 0.005]	3	0.5	0.1	0.6	5	32	64

表 3 给出了两个数据集上四个模型的参数设置，相关结果将在 4.2 节中进行讨论。重要参数包括权衡因子 λ_1 和 λ_2、链接预测比率 r_{le}、邻居跳数 hop、链接插入维度 out_dim。为作比较，使用了 PCGNN [22]模型，该模型可在 GitHub (https://github.com/PonderLY/PC-GNN)上公开获取。

4.2. 实验结果比较

在评估过程中，将基线模型与本文提出的模型进行了比较。对于不同规模的训练集，分别计算了模型的性能指标，结果分别显示在表 4 和表 5 中。另外，LE-DCBFD 模型的数值代表了在固定数据集划分下进行 10 次随机初始化运行的平均结果和标准误差。在表中，粗体字体表示最佳性能，下划线表示次佳性能。根据表 4 和表 5 中的数据，DCBFD 表现出次佳性能。同时，将 LE-DCBFD 与 PCGNN 进行了比较，并将性能提升的百分比填写在括号内。负类(欺诈者)下采样后的实验结果表明，机器学习算法(如决策树、逻辑回归和支持向量机)显著优于图信息融合模型，如 GCN、GAT、RGCN 和 GraphSAGE 等，甚至超过了一些专门为欺诈检测任务提出的模型，如 GeniePath、Player2Vec 和 GraphConsis 等。其中，支

持向量机(SVC)的分类效果甚至可与 PCGNN 模型相媲美。

Table 4. Test set performance (%) for fraud detection using Amazon training datasets of different proportions
表 4. 针对不同比例(百分比)的 Amazon 训练数据集进行欺诈检测的测试集性能(%)

Metric	AUC				Recall			
Train %	5%	10%	20%	40%	5%	10%	20%	40%
DT	85.52	90.95	89.40	88.55	85.52	**90.95**	89.40	88.55
LR	86.53	85.79	93.00	93.56	86.87	87.97	89.74	88.78
SVC	93.15	93.49	94.02	94.81	86.16	87.62	87.94	88.98
GCN	74.44	75.25	75.13	74.34	65.54	67.81	66.15	67.45
GAT	73.89	74.55	72.10	75.16	63.22	65.84	67.13	65.51
RGCN	75.12	74.13	75.58	74.68	64.23	67.22	65.08	67.68
GraphSAGE	70.71	73.97	73.97	75.27	69.09	69.36	70.30	70.16
GeniePath	71.56	72.23	71.89	72.65	65.56	66.63	65.08	65.41
Player2Vec	76.86	75.73	74.55	56.94	50.00	50.00	50.00	50.00
SemiGNN	70.25	76.21	73.98	70.35	63.29	63.32	61.28	62.89
GraphConsis	85.46	85.29	85.50	85.50	85.49	85.38	85.59	85.53
CARE-GNN	89.54	89.44	89.45	89.73	88.34	88.29	88.27	88.48
PCGNN	93.01	94.62	95.17	**95.86**	87.00	88.21	87.72	**90.30**
DCBFD	94.80	94.84	94.86	95.11	89.39	89.84	88.97	89.27
LE-DCBFD	**95.57 ± 0.000 (2.75)**	**95.27 ± 0.002 (0.69)**	**95.38 ± 0.001 (0.22)**	95.60 ± 0.001 (−0.27)	**89.71 ± 0.002 (3.11)**	89.86 ± 0.002 (1.88)	**89.32 ± 0.002 (1.82)**	89.69 ± 0.002 (−0.68)

本文提出的欺诈检测模型 LE-DCBFD 相较于最优的 PCGNN 模型展现出显著的性能提升。具体而言，在规模更大的 Yelp 数据集中，LE-DCBFD 模型在 AUC 指标上平均提升了约 9%，在召回率指标上平均提升了约 12%。在 Amazon 数据集中，其在 AUC 和召回率指标上均实现了约 1%的平均提升。上述结果清晰地表明，LE-DCBFD 超越了现有的最优模型。总之，可以认为 LE-DCBFD 模型显著提升了欺诈检测性能，即使在处理大规模数据集和稀疏负类数据时，也能给出稳定的结果。

Table 5. Test set performance (%) for fraud detection using Yelp training datasets of different proportions
表 5. 针对不同比例(百分比)的 Yelp 训练数据集进行欺诈检测的测试集性能(%)

Metric	AUC				Recall			
Train %	5%	10%	20%	40%	5%	10%	20%	40%
DT	63.21	65.43	67.04	68.51	63.21	65.43	67.04	68.51
LR	73.91	74.60	74.19	75.24	68.14	68.45	68.38	69.13
SVC	76.48	78.35	80.00	81.41	69.42	71.46	72.84	73.56
GCN	54.98	50.94	53.15	52.47	53.12	51.10	53.87	50.81
GAT	56.23	55.45	57.69	56.24	54.68	52.34	53.20	54.52
RGCN	50.21	55.12	55.05	53.38	50.38	51.75	50.92	50.43

续表

GraphSAGE	53.82	54.20	56.12	54.00	54.25	52.23	52.69	52.86
GeniePath	56.33	56.29	57.32	55.91	52.33	54.35	54.84	50.94
Player2Vec	51.03	50.15	51.56	53.65	50.00	50.00	50.00	50.00
SemiGNN	53.73	51.68	51.55	51.58	52.28	52.57	52.16	50.59
GraphConsis	61.58	62.07	62.31	62.07	62.60	62.08	62.35	62.08
CARE-GNN	71.26	73.31	74.45	75.70	67.53	67.77	68.60	71.92
PCGNN	75.17	77.13	78.43	81.69	69.46	63.04	71.03	74.20
DCBFD	82.17	82.17	82.11	82.25	74.59	74.54	74.55	74.86
LE-DCBFD	**85.22 ± 0.001 (13.37)**	**85.10 ± 0.001 (10.33)**	**85.13 ± 0.001 (8.54)**	**85.24 ± 0.001 (11.85)**	**77.69 ± 0.005 (23.03)**	**77.56 ± 0.004 (8.87)**	**77.33 ± 0.006 (8.54)**	**77.42 ± 0.002 (4.35)**

4.3. 消融实验

图 3 为 LE-DCBFD 模型在 Amazon 和 Yelp 上的链路增强器消融实验，展示了在两个不同数据集上，包含链接增强器的 LE-DCBFD 模型与不包含链接增强器的 DCBFD 模型的性能对比分析。观察结果显示，标记为“•”的性能曲线始终高于标记为“+”的曲线，特别是在训练过程的最后 100 个轮次中，这一趋势更为明显。这清楚地表明，链接增强器的加入显著提升了模型识别欺诈行为的能力，从而验证了该模块设计的有效性和实用价值。图 4 展示了损失函数的消融实验结果。可以看出在两个不同数据集上综合评估 AUC 和召回率指标时，本文提出的组合损失函数 ce_dice 能使模型达到最优性能。其次是单独使用交叉熵损失，而单独使用 Dice 损失的性能极差。

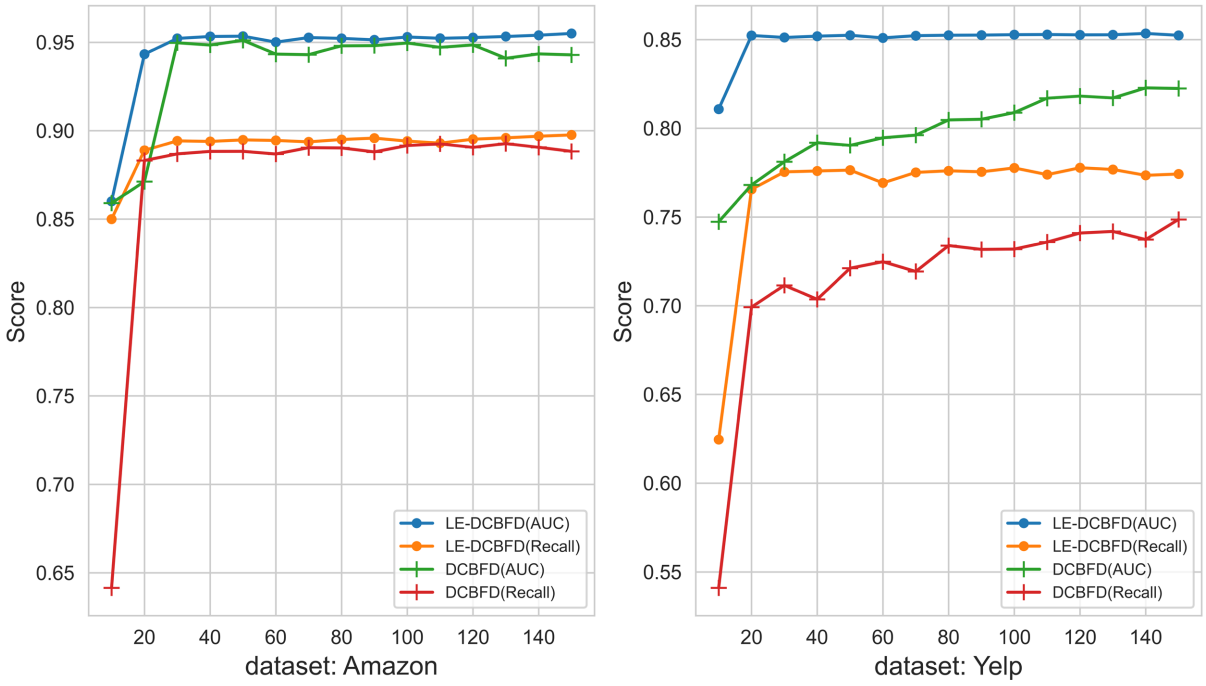

Figure 3. The link enhancer ablation experiment of the LE-DCBFD model on Amazon and Yelp datasets
图 3. LE-DCBFD 模型在 Amazon 和 Yelp 上的链路增强器消融实验

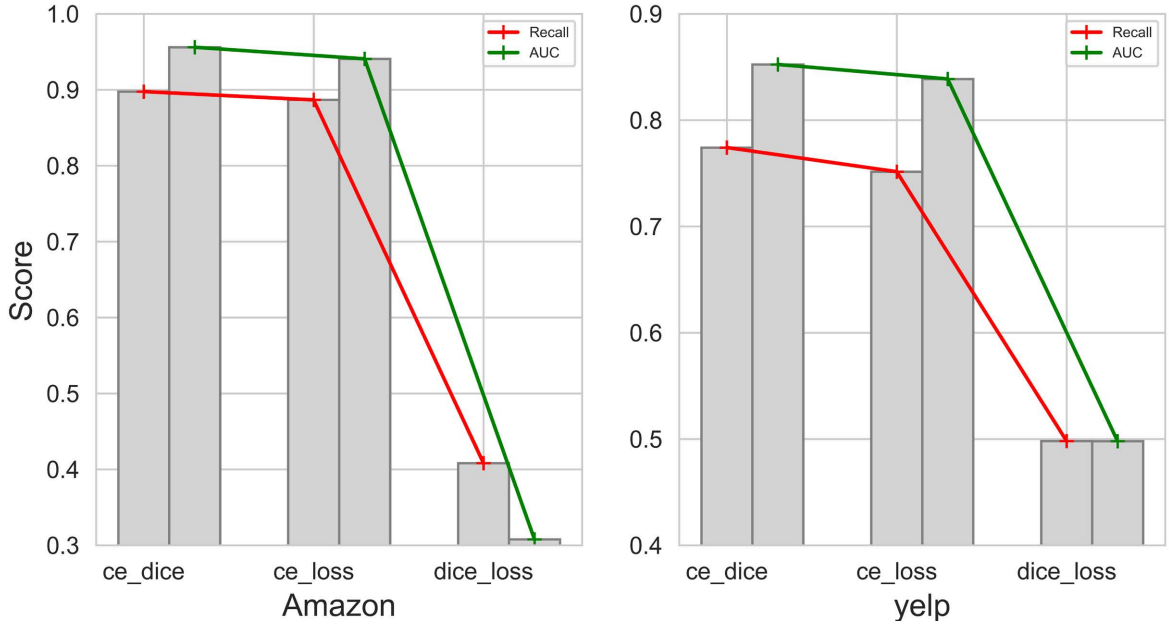

Figure 4. Loss function ablation experiments of the LE-DCBFD model on the Amazon and Yelp datasets
图 4. LE-DCBFD 模型在 Amazon 和 Yelp 数据集上的损失函数消融实验

5. 结论与未来工作

本文聚焦链接预测对异构图欺诈检测算法有效性的影响，提出创新链接预测算法——链接增强器，它结合启发式链接预测与反向传播神经网络，利用标签信息优化节点链接嵌入表示并挖掘潜在连接。在此基础上开发了链接增强的 Dice 损失一致平衡欺诈检测器(LE-DCBFD)，并设计组合损失函数提升模型性能。在 Amazon 和 Yelp 数据集实验显示，LE-DCBFD 有效且优于基线模型，在两个数据集上的欺诈检测性能分别平均提高 1%、10%以上，处理大型数据集和稀疏负类样本时鲁棒性好，链接增强器还降低了人工标注成本。未来研究将聚焦聚合算子选择和模型可解释性，探索引入强化学习发现最优链接预测比率、替换注意力机制，还会开展采样方法和网络对抗策略设计研究以提升检测器能力。

数据可用性

在此分享本文所使用公共数据的链接。Amazon 数据集和 Yelp 数据集(https://github.com/YingtongDou/CARE-GNN/tree/master/data)。

基金项目

本研究得到了浙江省“十四五”教学改革项目(编号：jg20220522)、浙江省大学生科技创新活动计划(新苗计划) (编号：2024R429C065)、浙江省自然科学基金(编号：LY17A010013)和国家自然科学基金(编号：11201344)的资助。

参考文献

[1] Han, Q., Wen, H. and Miao, F. (2018) Rumor Spreading in Interdependent Social Networks. *Peer-to-Peer Networking and Applications*, **11**, 955-965. https://doi.org/10.1007/s12083-017-0616-y

[2] Sreenivasulu, V. and Wajeed, M.A. (2021) Image Based Classification of Rumor Information from the Social Network Platform. *Traitement du Signal*, **38**, 1413-1421. https://doi.org/10.18280/ts.380516

[3] Wu, J., Hu, R., Li, D., Ren, L., Huang, Z. and Zang, Y. (2024) Beyond the Individual: An Improved Telecom Fraud

Detection Approach Based on Latent Synergy Graph Learning. *Neural Networks*, **169**, 20-31. https://doi.org/10.1016/j.neunet.2023.10.019

[4] Abdul Salam, M., Fouad, K.M., Elbably, D.L. and Elsayed, S.M. (2024) Federated Learning Model for Credit Card Fraud Detection with Data Balancing Techniques. *Neural Computing and Applications*, **36**, 6231-6256. https://doi.org/10.1007/s00521-023-09410-2

[5] Mao, X., Sun, H., Zhu, X. and Li, J. (2022) Financial Fraud Detection Using the Related-Party Transaction Knowledge Graph. *Procedia Computer Science*, **199**, 733-740. https://doi.org/10.1016/j.procs.2022.01.091

[6] Bayerstadler, A., van Dijk, L. and Winter, F. (2016) Bayesian Multinomial Latent Variable Modeling for Fraud and Abuse Detection in Health Insurance. *Insurance*: *Mathematics and Economics*, **71**, 244-252. https://doi.org/10.1016/j.insmatheco.2016.09.013

[7] Yan, C., Li, M., Liu, W. and Qi, M. (2020) Improved Adaptive Genetic Algorithm for the Vehicle Insurance Fraud Identification Model Based on a BP Neural Network. *Theoretical Computer Science*, **817**, 12-23. https://doi.org/10.1016/j.tcs.2019.06.025

[8] Rybalchenko, L., Ryzhkov, E. and Ciobanu, G. (2022) Global Consequences of the Loss of Business in Countries around the World Caused by Fraud. *Philosophy*, *Economics and Law Review*, **2**, 118-126.

[9] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. *Proceedings of the International Conference on Learning Representations* (*ICLR*). arXiv:1609.02907.

[10] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y. (2017) Graph Attention Networks. *Proceedings of the International Conference on Learning Representations* (*ICLR*). arXiv:1710.10903.

[11] Hamilton, W.L., Ying, R. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. *Proceedings of the* 31*st Conference on Neural Information Processing Systems* (*NeurIPS*), Long Beach, 4-9 December 2017, 1025-1035.

[12] Wang, J., Wen, R., Wu, C., Huang, Y. and Xiong, J. (2019) FdGars: Fraudster Detection via Graph Convolutional Networks in Online App Review System. *Companion Proceedings of the* 2019 *World Wide Web Conference*, San Francisco, 13-17 May 2019, 310-316. https://doi.org/10.1145/3308560.3316586

[13] Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., *et al*. (2019) A Semi-Supervised Graph Attentive Network for Financial Fraud Detection. 2019 *IEEE International Conference on Data Mining* (*ICDM*), Beijing, 8-11 November 2019, 598-607. https://doi.org/10.1109/icdm.2019.00070

[14] Hu, X., Chen, H., Liu, S., Jiang, H., Chu, G. and Li, R. (2022) BTG: A Bridge to Graph Machine Learning in Telecommunications Fraud Detection. *Future Generation Computer Systems*, **137**, 274-287. https://doi.org/10.1016/j.future.2022.07.020

[15] Liu, Z., Dou, Y., Yu, P.S., Deng, Y. and Peng, H. (2020) Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event China, 25-30 July 2020, 1569-1572. https://doi.org/10.1145/3397271.3401253

[16] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H. and Yu, P.S. (2020) Enhancing Graph Neural Network-Based Fraud Detectors against Camouflaged Fraudsters. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, Virtual Event Ireland, 19-23 October 2020, 315-324. https://doi.org/10.1145/3340531.3411903

[17] Yang, X., Lyu, Y., Tian, T., Liu, Y., Liu, Y. and Zhang, X. (2020) Rumor Detection on Social Media with Graph Structured Adversarial Learning. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, 7-15 January 2021, 1417-1423. https://doi.org/10.24963/ijcai.2020/197

[18] Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., *et al*. (2021) Pick and Choose: A GNN-Based Imbalanced Learning Approach for Fraud Detection. *Proceedings of the Web Conference* 2021, Ljubljana, 19-23 April 2021, 3168-3177. https://doi.org/10.1145/3442381.3449989

[19] Gao, Y., Wang, X., He, X., Liu, Z., Feng, H. and Zhang, Y. (2023) Addressing Heterophily in Graph Anomaly Detection: A Perspective of Graph Spectrum. *Proceedings of the ACM Web Conference* 2023, Austin, 30 April 2023-4 May 2023, 1528-1538. https://doi.org/10.1145/3543507.3583268

[20] Tang, J., Hua, F.R., Gao, Z.Q., Zhao, P.L. and Li, J. (2023) GADBench: Revisiting and Benchmarking Supervised Graph Anomaly Detection. arXiv: 2306.12251.

[21] Hong, B., Lu, P., Xu, H., Lu, J., Lin, K. and Yang, F. (2024) Health Insurance Fraud Detection Based on Multi-Channel Heterogeneous Graph Structure Learning. *Heliyon*, **10**, e30045. https://doi.org/10.1016/j.heliyon.2024.e30045

[22] Zhang, Y., Fan, Y., Ye, Y., Zhao, L. and Shi, C. (2019) Key Player Identification in Underground Forums over Attributed Heterogeneous Information Network Embedding Framework. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, November 3-7, 2019, 549-558. https://doi.org/10.1145/3357384.3357876

[23] Kong, M., Li, R., Wang, J., Li, X., Jin, S., Xie, W., *et al*. (2024) CFTNet: A Robust Credit Card Fraud Detection Model Enhanced by Counterfactual Data Augmentation. *Neural Computing and Applications*, **36**, 8607-8623. https://doi.org/10.1007/s00521-024-09546-9

[24] Chen, J., Chen, Q., Jiang, F., Guo, X., Sha, K. and Wang, Y. (2024) SCN_GNN: A GNN-Based Fraud Detection Algorithm Combining Strong Node and Graph Topology Information. *Expert Systems with Applications*, **237**, Article 121643. https://doi.org/10.1016/j.eswa.2023.121643

[25] Wu, J., Hu, R., Li, D., Ren, L., Hu, W. and Zang, Y. (2024) A GNN-Based Fraud Detector with Dual Resistance to Graph Disassortativity and Imbalance. *Information Sciences*, **669**, Article 120580. https://doi.org/10.1016/j.ins.2024.120580

[26] Li, A., Qin, Z., Liu, R., Yang, Y. and Li, D. (2019) Spam Review Detection with Graph Convolutional Networks. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, November 3-7, 2019, 2703-2711. https://doi.org/10.1145/3357384.3357820

[27] Zhang, M. and Chen, Y. (2017) Weisfeiler-Lehman Neural Machine for Link Prediction. *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, 13-17 August 2017, 1835-1844. https://doi.org/10.1145/3097983.3097996

[28] Islam, M.K., Aridhi, S. and Smail-Tabbone, M. (2020) Appraisal Study of Similarity-Based and Embedding-Based Link Prediction Methods on Graphs. 10*th International Conference on Data Mining & Knowledge Management Process*, London, 25-26 July 2021, 81-92. https://doi.org/10.5121/csit.2021.111106

[29] Lorrain, F. and White, H.C. (1971) Structural Equivalence of Individuals in Social Networks. *The Journal of Mathematical Sociology*, **1**, 49-80. https://doi.org/10.1080/0022250x.1971.9989788

[30] Zhou, T., Lü, L. and Zhang, Y. (2009) Predicting Missing Links via Local Information. *The European Physical Journal B*, **71**, 623-630. https://doi.org/10.1140/epjb/e2009-00335-8

[31] Jaccard, P. (1901) Etude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 547-579.

[32] Zhang, M. and Chen, Y. (2018) Link Prediction Based on Graph Neural Networks. *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 5165-5175.

[33] Schwarz, K. (2011) Darts, Dice, and Coins: Sampling from a Discrete Distribution. https://www.keithschwarz.com/darts-dice-coins/

[34] Martínez, C. and Roura, S. (1998) Randomized Binary Search Trees. *Journal of the ACM*, **45**, 288-323. https://doi.org/10.1145/274787.274812

[35] Milletari, F., Navab, N. and Ahmadi, S. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 *Fourth International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3dv.2016.79

[36] Rayana, S. and Akoglu, L. (2015) Collective Opinion Spam Detection. *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, 10-13 August 2015, 985-994. https://doi.org/10.1145/2783258.2783370

[37] McAuley, J.J. and Leskovec, J. (2013) From amateurs to Connoisseurs. *Proceedings of the* 22*nd International Conference on World Wide Web*, Rio de Janeiro, 13-17 May 2013, 897-908. https://doi.org/10.1145/2488388.2488466

[38] Jindal, N. and Liu, B. (2008) Opinion Spam and Analysis. *Proceedings of the International Conference on Web Search and Web Data Mining*, Palo Alto, 11-12 February 2008, 219-230. https://doi.org/10.1145/1341531.1341560

[39] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

[40] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1023/a:1022643204877

[41] Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I. and Welling, M. (2018) Modeling Relational Data with Graph Convolutional Networks. In: Gangemi, A., *et al*., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 593-607. https://doi.org/10.1007/978-3-319-93417-4_38

[42] Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., *et al*. (2019) GeniePath: Graph Neural Networks with Adaptive Receptive Paths. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 4424-4431. https://doi.org/10.1609/aaai.v33i01.33014424

[43] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[44] Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997) Automatic Text Structuring and Summarization. *Information Processing & Management*, **33**, 193-207. https://doi.org/10.1016/s0306-4573(96)00062-3