Robust Heterophilic Graph Learning against Label Noise for Anomaly Detection

Junhang Wu[1,2], **Ruimin Hu**[1,3,*], **Dengshi Li**[4,1], **Zijun Huang**[1,2], **Lingfei Ren**[1,2], **Yilong Zang**[1,2]
[1] National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University
[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
[3] School of Cyber Science and Engineering, Wuhan University
[4] School of Artificial Intelligence, Jianghan University
{wjh920925, huangzijun, renlingfei, zangyl}@whu.edu.cn, hurm1964@gmail.com, reallds@jhun.edu.cn

Abstract

Given clean labels, Graph Neural Networks (GNNs) have shown promising abilities for graph anomaly detection. However, real-world graphs are inevitably noisy labeled, which drastically degrades the performance of GNNs. To alleviate it, some studies follow the local consistency (a.k.a homophily) assumption to conduct neighborhood-based label noise correction, and to dense raw graphs using raw features or representations learned by poisoned labels. But for the anomaly detection task, the graph is not always homophilic but more likely to be heterophilic, which would corrupt the above assumption due to complicating connection patterns and impairing the effects of message passing. To this end, we propose a novel label noise-resistant graph learning (NRGL) framework, which facilitates robust graph learning from the perspectives of structure augmentation and fine-grained label governance. Specifically, we first present an investigation to verify that increasing graph homophily could help resist label noise. Based on the observation, an unsupervised contrastive learning paradigm is then introduced so well that it cannot only adaptively extract the dual views from the raw graph as structure augmentation, but also enhance the robustness of node representations. Next, given robust node representations, the noisy labels are divided into three candidate sets based on the small-loss criterion for fine-grained noise governance. Furthermore, a node sampler is designed to take structure importance, class frequency, and confidence score into consideration, which helps select reliable and important nodes for training. Extensive experiments on real-world datasets demonstrate the effectiveness of our method.

1 Introduction

In real-world scenarios, a set of entities and their relationships can be naturally formed graph-like structures, which have been applied in various domains, like social science [Zhao *et al.*, 2023], financial transaction [Zheng *et al.*, 2023], and recommendation system [Wu *et al.*, 2022]. Recently, Graph Neural Networks (GNNs) have achieved promising performance in dealing with such graph structure data by introducing a message-passing mechanism to effectively aggregate information from its neighbors. This mechanism makes the supervision of labeled nodes propagated to the unlabeled nodes, which helps semi-supervised graph learning (e.g., node classification-based graph anomaly detection [Zhang *et al.*, 2021; Chai *et al.*, 2022; Tang *et al.*, 2022]).

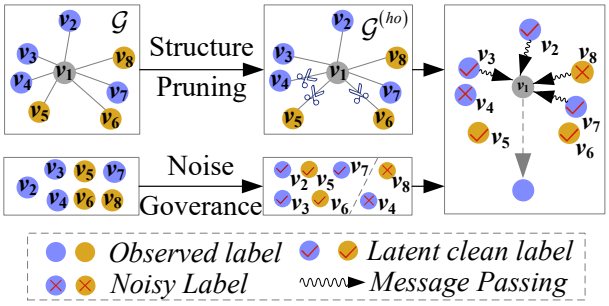


Figure 1: An flow of structure pruning-based graph augmentation and noise goverance for anomaly detection.

Although achieving promising progress, most existing GNN-based models assumed that the training label is clean. However, in real-world scenarios, the label annotation is labor-intensive, expensive, and full of subjective judgments (e.g., medical knowledge, fake news, and fraudulent comments), so the node labels always inevitably contain noise. However, it has already been reported that deep learning models would overfit the noisy labels and cause poor generalization performance [Arpit *et al.*, 2017; Han *et al.*, 2018]. Currently, robust GNNs against graph perturbations and attacks have been widely studied [Li *et al.*, 2022; Jin *et al.*, 2023], but label noise on graphs still remains under-explored.

To resist label noise, extensive approaches have been proposed, e.g., loss correction [Goldberger and Ben-Reuven, 2016; Patrini *et al.*, 2017], sample selection [Jiang *et al.*, 2018; Yu *et al.*, 2019; Huang *et al.*, 2019], and robust loss function [Ghosh *et al.*, 2017; Zhang and Sabuncu, 2018; Wang *et al.*, 2019]. Although they have achieved satisfac-

*Corresponding Author

tory results, they are dedicated to independent and identically distributed data (e.g., images), which may not be directly applicable to handling noisy labels on graphs. That is because the message-passing mechanism would make the noisy information propagate to the whole graph across the structure, which would negatively pollute other unlabeled neighbors [Li *et al.*, 2024].

To mitigate the effects of label noise on the graph, existing methods rely on local consistency (a.k.a homophily) assumption to make neighborhood-based label noise correction or dense graph for augmentation to facilitate sufficient message passing. Specifically, NRGNN [Dai *et al.*, 2021] can be treated as the pioneering work that bridges links between labeled and unlabeled nodes for graph augmentation. Based on it, RTGNN [Qian *et al.*, 2023] is further designed to make more noise correction and mine accurate pseudo-labels for enhanced supervision. Furthermore, CGNN [Yuan *et al.*, 2023] integrates contrastive learning with neighbor-based noisy label correction. Though they have achieved much progress, there still exist some drawbacks: (1) ***non-homophilic structure.*** Existing methods predominantly rely on the homophily assumption, where connected nodes tend to have the same labels. But for anomaly detection, the graph is not always homophilic but more likely to be heterophilic, which would complicate connection patterns and impair the effects of message passing. (2) ***non-robust node representation.*** Existing methods make graph augmentation and noise correction using either raw features or representations learned by noisy labels, which would not only neglect structure information but also suffer poor performance from polluted labels. (3) ***non-balanced distribution.*** Graph imbalance is another challenge for anomaly detection [Liu *et al.*, 2021], which would exacerbate the difficulty in the presence of label noise governance, but existing label noise-resistant methods have not taken it into full consideration.

To this end, we study robust graph learning in the presence of such challenges: (1) How to derive robust node representations? (2) How to mitigate the graph heterophily issue? (3) How to deal with label imbalance when making noise governance? To address such challenges, we first present an investigation to study how graph homophily or heterophily affects the robustness of models to combat label noise, as shown in 2. We can observe that graph heterophily exacerbates the effects of label noise on the model and increasing graph homophily (by structure pruning) can help the model combat label noise, and more details can be found in Section 3.3. Based on this, we propose a novel label noise-resistant graph learning framework named NRGL, which makes robust graph learning from the perspectives of structure augmentation and label governance, as shown in Fig. 1. Specifically, to address challenges one and two, an edge discrimination-based dual encoder is first introduced to divide the raw graph $\mathcal{G}$ into a homophilic view $\mathcal{G}^{(ho)}$ and a heterophilic one $\mathcal{G}^{(he)}$. Given $\mathcal{G}^{(ho)}$ and $\mathcal{G}^{(he)}$, considering the inherent noisy nature of labels, an unsupervised contrastive learning paradigm is adopted to derive robust node representations. By doing so, we can not only derive robust node representations but also a subgraph with high homophily by structure pruning. Then, given the robust node representations from two views, a mutual cross-entropy is introduced to divide the noisy label set into three subsets (i.e., clean, confident, and remaining ones) based on the small-loss criterion. Furthermore, a novel node sampler is designed to take node popularity, class frequency, and confidence score into consideration, which helps select reliable and important nodes for model training.

Contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to study label noise-resistant graph learning for anomaly detection in the presence of graph heterophily and imbalance. Furthermore, we find that increasing graph homophily can help resist label noise on graphs.
- We develop a novel NRGL model, which facilitates robust graph learning from the perspectives of structure augmentation with reliable node representations, and imbalance-oriented sampler for label governance.
- Experiments on two real-world datasets have verified the advantages of our proposed method.

2 Related Work

There have been many robust deep learning studies on non-graph data from the perspectives of loss correction [Goldberger and Ben-Reuven, 2016; Patrini *et al.*, 2017], sample selection [Jiang *et al.*, 2018; Yu *et al.*, 2019; Huang *et al.*, 2019], and robust loss function [Ghosh *et al.*, 2017; Zhang and Sabuncu, 2018; Wang *et al.*, 2019], but the GNN with robustness to resist label noise is still under-explored. Jin et al. [NT *et al.*, 2019] first proved that GNNs are vulnerable to label noise and further proposed a noise-tolerant method by introducing backward loss correction. Afterward, NRGNN was proposed to learn a robust GNN by linking the unlabeled node to the labeled one. In addition, the pseudo labels are also adopted to help alleviate the limited label issue. Recently, based on graph homophily assumption, RTGNN [Qian *et al.*, 2023] and CGNN [Yuan *et al.*, 2023] are developed by introducing graph augmentation methods and further conduct noise governance and correction to facilitate robust Graph learning.

Such methods have achieved much progress in robust GNN learning based on the graph homophily assumption, but they may oversimplify the complexity of the graph because the real-world networks are not always homophilic but more likely to be heterophilic, where the connected nodes tend to be different classes. Furthermore, most of these methods rely on raw features or learned representations for graph augmentation or label correction, but it would either neglect structure information or suffer poor performance from mislabeled labels. Hence, it prompts us to study a unified framework to deal with graph heterophily issue and derive robust node representations for noise governance.

3 Preliminaries

3.1 Definition

A graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ represents the node set, $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ refers to the edge set. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ represent the raw feature matrix of all nodes. Furthermore, the adjacency matrix of $\mathcal{G}$ is

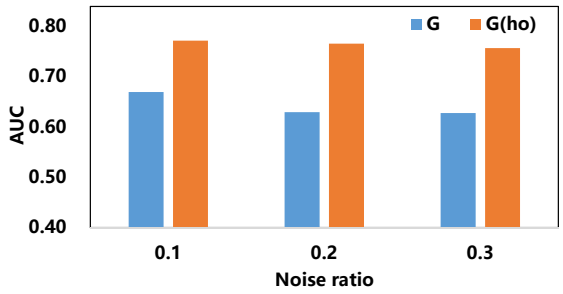


Figure 2: Performance (AUC) comparison of GCNs fed by the raw graph and processed homophilic graph on Elliptic dataset with noisy label ratio varying from 0.1 to 0.3.

denoted as $\mathbf{A} \in \mathbb{R}^{N\times N}$, where $\mathbf{A}_{ij} = 1$ if there exists a connection between the node v_i and v_j, and $\mathbf{A}_{ij} = 0$ otherwise.

3.2 Problem Statement

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as defined above, $\mathcal{V}_L = \{v_1, v_2, ..., v_l\}$ is the set of training nodes labeled as $\{y_1, y_2, ..., y_l\}$, where $y_i \in \{0, 1\}$, and $\mathcal{V}_U = \mathcal{V} - \mathcal{V}_L$ is the remaining set of unlabeled nodes. However, the given labels of the training set are corrupted by noise, i.e., y_i may be incorrect for some of the nodes in $\mathcal{V}_L$. To this end, the task is to train a robust GNN to identify whether the remaining unlabeled nodes are anomalous or normal.

3.3 How Does Graph Homophily Affect Node Classification in the Case of Label Noise?

Here we will explore the effect of graph homophily on combating label noise with empirical experiments. Specifically, we conduct experiments on the real-world dataset Elliptic. In the dataset, heterophilic connections between connected nodes are removed to add graph homophily, and the raw and processed graphs are named G and $G(ho)$, respectively. Then, we randomly sample 40% of the total nodes as the training set with known labels, and the validation set and test set are divided according to 1:2. Next, labels of the nodes in the training set are corrupted by randomly flipping the true labels to another class with a probability of p (p = 0.1, 0.2, 0.3). Finally, we use GCN [Kipf and Welling, 2017] for model training and testing, and the performance (AUC) is shown in Fig. 2. We can observe that performance on $G(ho)$ is significantly better than that on G, which means increasing graph homophily can help the model facilitate resisting label noise.

4 Method

In this section, we will introduce our model NRGL in detail, and the illustration of it is shown in Fig. 3. We can observe that NRGL is composed of three modules. First, an edge discrimination-based dual view encoder is introduced by designing an edge predictor to discriminate homophilic or heterophilic edges for graph division. Then the dual channel encoders with low- and high-pass filters are adopted to derive robust node representations from corresponding homophilic and heterophilic graph views; Second, given dual-frequency representations of nodes, the labeled nodes are divided into three subsets for noise governance based on small-loss criterion; Third, an imbalance-oriented sampler is designed to help select reliable and important nodes for training.

4.1 Edge Discrimination-Based Dual View Encoder

The core of our proposed method is how to exploit the supervision of clean labels out of the noisy ones. "Co-training" [Han *et al.*, 2018] paradigm has achieved much progress in noisy labeled image processing, which maintains two different networks and alternately searches for useful knowledge from each view to their peer networks for parameters update. Intuitively, different networks have different decision boundaries, and different abilities to filter out the label noise.

Based on this exchange strategy, it can reduce the effect of the error flows. Inspired by this, we divide the original graph into a pair of homophilic and heterophilic views, and construct two different GNNs. Specifically, a direct tool is to design an edges discriminator to estimate the homophily probability w_{ij} between connected node v_i and v_j. For w_{ij}, we need to consider the features of both node itself and its neighbors. Take the raw features $\mathbf{x}$ as the input, the homophily probability of each edge is estimated as follows:

$$\begin{aligned} &\mathbf{h}_i = \mathrm{MLP}_1(\mathbf{x}_i), \mathbf{h}_j = \mathrm{MLP}_1(\mathbf{x}_j), \\ &w_{ij} = \mathrm{Sigmoid}\left(\frac{\mathbf{W}_a^{\mathrm{T}}[\mathbf{h}_i \| \mathbf{h}_j] + \mathbf{W}_a^{\mathrm{T}}[\mathbf{h}_j \| \mathbf{h}_i]}{2\sqrt{d}}\right), \end{aligned} \tag{1}$$

where $\mathrm{MLP}_1(\cdot)$ denotes the multilayer perceptron, $\cdot \| \cdot$ represents the concatenation operation, $\sqrt{d}$ (d is the dimension of $\mathbf{h}_i$ or $\mathbf{h}_j$) acts a scaling factor, $\mathbf{W}_a \in \mathbb{R}^d$ denotes the shared weight matrice, and sigmoid function can naturally limit the value of w_{ij} in the range of 0 to 1. With the estimated homophily probability indicator, the original graph $\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathbf{X})$ is divided into two graph views, i.e., the homophilic one $\mathcal{G}^{(ho)} = \left(\mathcal{V}, \mathbf{A}^{(ho)}, \mathbf{X}\right)$ and the heterophilic one $\mathcal{G}^{(he)} = \left(\mathcal{V}, \mathbf{A}^{(he)}, \mathbf{X}\right)$. To make the edge discriminator trainable, a soft weight is assigned to each connection as follows:

$$\mathbf{A}_{ij}^{(ho)} = w_{ij}, \mathbf{A}_{ij}^{(he)} = 1 - w_{ij}, e_{ij} \in \mathcal{E}. \tag{2}$$

Given homophilic and heterophilic graph views, two different encoders are introduced to perform low- and high-pass graph signal filters, which helps retain commonalities between similar pair nodes and filters out the irrelevant information from dissimilar neighborhoods.

On the homophilic view, similar nodes are connected together with a larger homophily probability. A low-pass graph filter can be deployed to smooth the node representations along the homophilic structure, which facilitates graph learning by retaining the commonalities between connected similar nodes. Therefore, a simple low-pass filter is introduced for aggregation as follows:

$$\mathbf{H}_0^{(ho)} = \mathrm{MLP}_1^{(ho)}(\mathbf{X}), \ \mathbf{H}_l^{(ho)} = \left(\mathbf{I} + \tilde{\mathbf{A}}^{(ho)}\right)\mathbf{H}_{l-1}^{(ho)}, \tag{3}$$

where $\tilde{\mathbf{A}}^{(ho)}$ refers to the symmetric normalized homophilic adjacency matrix of $\mathbf{A}^{(ho)}$, $l \in \{1, ..., L\}$ represents the index of layer, and $\mathbf{H}_L^{(ho)}$ (a.k.a. $\mathbf{H}^{(ho)}$) is the final representation from the homophilic view.

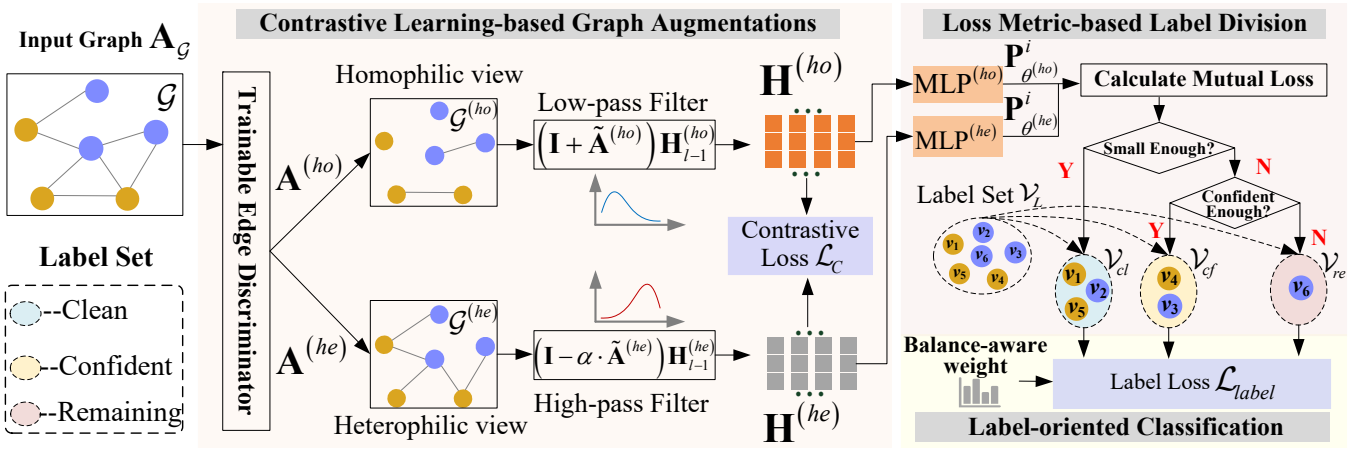


Figure 3: An illustration of proposed framework.

From the perspective of the heterophilic view, a low-pass filter, which smooths the features, will result in the loss of discriminative attributes between nodes, and the high-pass filter can preserve the high-frequency signals by sharpening the difference between dissimilar nodes. For signal processing on images, the Laplacian kernel filter $\mathbf{L} = \mathbf{I} - \alpha \cdot \tilde{\mathbf{A}}^{(he)}$ is commonly applied to image sharpening tasks. Along this line to graph signal processing, the normalized Laplacian matrix can also be considered as a high-pass filter to extract the high-frequency component $\mathbf{x}^h$ of the given graph signal, as follows:

$$\mathbf{H}_0^{(he)} = \mathrm{MLP}_1^{(he)}(\mathbf{X}),\ \mathbf{H}_l^{(he)} = \left(\mathbf{I} - \alpha \cdot \tilde{\mathbf{A}}^{(he)}\right)\mathbf{H}_{l-1}^{(he)}, \tag{4}$$

where α controls the strength of high-pass filter, $\tilde{\mathbf{A}}^{(he)}$ is the symmetric normalized homophilic adjacency matrix of $\mathbf{A}^{(he)}$, and $\mathbf{H}_L^{(he)}$ (a.k.a. $\mathbf{H}^{(he)}$) represents the final representation from the heterophilic view.

Furthermore, to learn a robust edge discriminator, we use a robust contrasting mechanism to facilitate the model generate consistent node representations from two different graph views, and the contrastive learning loss of $h_i^{(ho)}$ and $h_i^{(he)}$ from the perspective of homophilic view is denoted as follows:

$$\mathcal{L}^{(ho)}\left(\mathbf{h}_i^{(ho)}, \mathbf{h}_i^{(he)}\right) = -\log \frac{\exp\left(\mathrm{sim}\left(\mathbf{h}_i^{(ho)}, \mathbf{h}_i^{(he)}\right)/\tau\right)}{\sum_{i\neq j}\exp\left(\mathrm{sim}\left(\mathbf{h}_i^{(he)}, \mathbf{h}_j^{(he)}\right)/\tau\right)}, \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ represents the cosine similarity between pair nodes, and τ denotes a temperature coefficient of 0.5. By combining heterophilic view, the total contrasting learning loss is denoted as:

$$\mathcal{L}_C = \frac{1}{2|\mathcal{V}|}\sum_{v_i\in\mathcal{V}}\left[\mathcal{L}^{(ho)}(\cdot) + \mathcal{L}^{(he)}(\cdot)\right]. \tag{6}$$

4.2 Loss Metric-Based Label Division

As [Arpit *et al.*, 2017] reported that DNNs tend to memorize the easy instances with clean labels, then gradually adapt to or overfit the hard ones with noisy labels, which means there exist different loss distributions between the clean and noisy items. To this end, the small-loss criterion mechanism [Han *et al.*, 2018] was proposed for training set division, which reduces the error flows by alternatively viewing small-loss instances during the exchange procedure. In our paper, as homophilic and heterophilic graph views have different learning abilities and decision boundaries, they are naturally treated as peer networks to exclude the label noise. First, the mutual cross-entropy loss is defined for the node i as follows:

$$\begin{aligned}\mathbf{p}_{\theta^{(ho)}}^i &= \mathrm{MLP}_2^{(ho)}\left(\mathbf{H}_L^{(ho)}\right), \mathbf{p}_{\theta^{(he)}}^i = \mathrm{MLP}_2^{(he)}\left(\mathbf{H}_L^{(he)}\right),\\ \mathcal{L}_{mul}^i &= -y^i\left[\log\left(\mathbf{p}_{\theta^{(ho)}}^i\right) + \log\left(\mathbf{p}_{\theta^{(he)}}^i\right)\right]\\ &= -y^i\log\left(\mathbf{p}_{\theta^{(ho)}}^i \cdot \mathbf{p}_{\theta^{(he)}}^i\right).\end{aligned} \tag{7}$$

$\mathcal{L}_{mul}^i$ measures the confidence of the prediction where the lower value of it, the higher probability of correct prediction. Intuitively, as the clean labels are easier to learn, the instance with small-loss are more likely to be correctly labeled. Given the $\mathcal{L}_{mul}^i$ of each node, a crucial issue is how to build a reliable enough classifier for indeed clean instance selection. It needs to be carefully considered twofold. First, the "memorization" effect of the deep network makes the model learn clean and easy patterns in the initial epochs, which means more instances should be included for sufficient training to learn a reliable pattern at the beginning of training. Second, with the epoch going large, the model would overfit on the noisy labels, which means we should gradually exclude the node with the large loss value [Han *et al.*, 2018].

Clean Label Set Extraction

First, we adopt a linear function to derive a dynamic percentage gating threshold with the increasing epochs as follows:

$$\mathcal{L}_{t_{epoch}}^{thre} = \mathrm{Percentile}\left(\mathcal{L}_{mul}^i, 1 - 0.5\times\frac{t_{epoch}}{T_{max}}\right), \tag{8}$$

where T_{max} represents the total number of epochs, t_{epoch} denotes the t-th epoch, $0.5 \leq \left(1 - 0.5\times\frac{t_{epoch}}{T_{max}}\right) < 1$ is a ratio gating threshold, $\mathrm{Percentile}(\mathcal{L}, p)$ is the value which

$(100 \times p)\,\%$ of the loss value in $\mathcal{L}$ fall, and $\mathcal{L}_{mul}^{thre}$ is the threshold value of mutual loss. Furthermore, the average of mutual loss is treated as another gating threshold for clean labels selection:

$$\mathcal{L}_{t_{epoch}}^{avg} = \frac{1}{|\mathcal{V}_L|} \sum_{v_i \in \mathcal{V}_L} \mathcal{L}_{mul}^{i}. \tag{9}$$

Given $\mathcal{L}_{t_{epoch}}^{thre}$ and $\mathcal{L}_{t_{epoch}}^{avg}$ as the upper bounds of mutual loss, the clean and noisy label sets can be divided as follows:

$$\mathcal{V}_{cl} = \left\{ v_i \,\middle|\, \mathcal{L}_{mul}^{i} < \max\left(\mathcal{L}_{t_{epoch}}^{thre}, \mathcal{L}_{t_{epoch}}^{avg} \right) \right\}, \tag{10}$$

where $\mathcal{V}_{cl}$ can be treated as the selected clean label set in which the mutual loss of each element is below the threshold. In Eq. 10, we can observe that $\mathcal{L}_{t_{epoch}}^{thre}$ allows more instances for sufficient training at the initial stage, and the number and quality of clean labels are gradually affected by the interaction of both $\mathcal{L}_{t_{epoch}}^{thre}$ and $\mathcal{L}_{t_{epoch}}^{avg}$ with the epoch decreasing. Overall, it makes the trade-off between early sufficient training and later noise-resistant training.

Confident Label Set Extraction

Given $\mathcal{V}_{cl}$ as the selected clean label set, the remaining label set is denoted as $\mathcal{V}_{ns} = \mathcal{V}_L - \mathcal{V}_{cl}$, which is vulnerable to be noisy. Inspired by the progress of RTGNN [Qian *et al.*, 2023] in dealing with noise governance, we observe that the model gradually has the ability to predict correct labels, which can be used to further divide a subset $\mathcal{V}_{cf} \in \mathcal{V}_{ns}$ where the prediction of two predictors (i.e., $\mathbf{P}_{\theta^{(ho)}}^{i}$ and $\mathbf{P}_{\theta^{(he)}}^{i}$) is confident but different from their labels, denoted as follows:

$$Z^i = \arg\max_{c=0,1} \mathbf{P}_{\theta^{(ho)}}^{i,c} = \arg\max_{c=0,1} \mathbf{P}_{\theta^{(he)}}^{i,c} \neq Y^i, \tag{11}$$

where Y^i denotes the observed label of the node v_i, and Z^i is the predicted label but different from the labeled one. Next, we want both $\mathbf{P}_{\theta^{(ho)}}^{i,Z^i}$ and $\mathbf{P}_{\theta^{(he)}}^{i,Z^i}$ be of the greater value with higher confidence. A natural idea is the value of $\sqrt{\mathbf{P}_{\theta^{(ho)}}^{i,Z^i} \cdot \mathbf{P}_{\theta^{(he)}}^{i,Z^i}}$ is greater than a threshold value Th. Furthermore, the threshold Th should be dynamic as we should tighten the condition at the early training and gradually loosen it with increasing epochs. Based on this, the confident label set can be divided by formulating Th with the variable epoch (t_{epoch}) as follows:

$$\mathcal{V}_{cf} = \left\{ v_i \,\middle|\, \sqrt{\mathbf{P}_{\theta^{(ho)}}^{i,Z^i} \cdot \mathbf{P}_{\theta^{(ho)}}^{i,Z^i}} > Th \right\}, Th = 1 - \frac{t_{epoch}}{C \cdot T_{max}}, \tag{12}$$

where $C = 2$ is the number of samle classes (i.e., normal and abnormal), $0.5 \leq Th < 1$ falls back from 1 to 0.5 with t_{epoch} increasing from 1 to T_{max}, which means gradually loosen the restrictions. Furthermore, a predicted confidence score $\mu(i)$ is defined as follows:

$$\mu(i) = \sqrt{\mathbf{P}_{\theta^{(ho)}}^{i,Z^i} \cdot \mathbf{P}_{\theta^{(ho)}}^{i,Z^i}} \tag{13}$$

where the higher $0.5 < \mu(i) < 1$, the higher the probability that v_i is incorrectly labeled. Given above $\mathcal{V}_{cl}$ and $\mathcal{V}_{cf}$, the remaining training set is denoted as

$$\mathcal{V}_{re} = \mathcal{V}_L - \mathcal{V}_{cl} - \mathcal{V}_{cf}. \tag{14}$$

4.3 Label-Oriented Classification

So far, we have divided the training set into three subsets, i.e., clean set $\mathcal{V}_{cl}$, confident set $\mathcal{V}_{cf}$ and the remaining set $\mathcal{V}_{re}$. By combining all of them, the loss of such labeled nodes is formulated as follows:

$$\mathcal{L}_{label} = \frac{1}{|\mathcal{V}_L|} \sum_{v_i \in \mathcal{V}_L} \xi(i)\, \hat{y} \log\left(\mathbf{P}_{\theta^{(ho)}}^{i} \cdot \mathbf{P}_{\theta^{(he)}}^{i} \right), \tag{15}$$

where

$$\xi(i) = \begin{cases} 1, \hat{y} = y^i & if\ v_i \in \mathcal{V}_{cl}; \\ \mu(i), \hat{y} = z^i & if\ v_i \in \mathcal{V}_{cf}; \\ 0.5, \hat{y} = y^i & if\ v_i \in \mathcal{V}_{re}. \end{cases} \tag{16}$$

where $\xi(i)$ acts as a confidence score to assign each subset with a different weight. Furthermore, to alleviate the influence of the imbalance problem, a sampler selector $P(v_i)$ is introduced to calculate $\mathcal{L}_{label}$ as follows:

$$P(v_i) \propto \frac{\sqrt{d_i} \cdot \xi(i)}{Z(C(v_i))}, \tag{17}$$

where d_i represents the degree node v_i, $Z(\mathcal{C}(v_i))$ denotes the label frequency of class $\mathcal{C}(v_i)$, φ_i is the confidence score of each node as defined in Eq.16. In Eq.17, we can observe that d_i means the popularity,$Z(C(v_i))$ means the rarity and φ_i means the credibility. Note that, we use $\sqrt{(\cdot)}$ as the scaling operator on d_i to smooth the uneven degree distribution of nodes. Summarily, the nodes with high popularity, rarity, and credibility are more likely to be selected. Finally, the total loss can be calculated as follows:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{label}, \tag{18}$$

where $\mathcal{L}_C$ and $\mathcal{L}_{label}$ represent the contrastive learning loss and label loss. Note that $\mathcal{L}_C$ and $\mathcal{L}_{label}$ are trained in a mutually boosting manner, and we adopt an alternating training strategy to iteratively optimize them. Finally, we use the $\mathcal{G}^{(ho)}$-based GNN for inference in the paper.

5 Experiment

5.1 Experimental Setup

Two widely used datasets are utilized to evaluate NRGL, and their statistics are shown in Table 2.

- Eliiptic [Weber *et al.*, 2019]: It is a Bitcoin transaction network where transactions and flows are the nodes and edges. The task is to predict illegal nodes (transactions).
- Yelp [Rayana and Akoglu, 2015]: It collects the reviews of hotels or restaurants on the Yelp platform, and the reviews are seen as nodes to be connected if they are posted by the same user. The task is to detect fake nodes (reviews).

Following [Chai *et al.*, 2022], we adopt the same dataset division. Furthermore, following [Dai *et al.*, 2021], we use **Uniform Noise** to corrupt the training set, where the label of each node is uniformly flipped with a probability of p.

Method	Elliptic						Yelp					
	10%		20%		30%		10%		20%		30%	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
GCN	0.6698	0.4779	0.6297	0.4736	0.6279	0.4758	0.5341	0.4787	0.5324	0.4622	0.5233	0.4605
GraphSAGE	0.6834	0.4886	0.6577	0.4849	0.6295	0.4795	0.6207	0.4818	0.6097	0.4809	0.6088	0.4702
NRGNN	0.7141	0.5737	0.6641	0.5341	0.6914	0.5191	0.6566	0.5496	0.6347	0.5309	0.6166	0.5029
RTGNN	0.7373	0.5975	0.7149	0.5635	0.6805	0.5285	0.6744	0.5612	0.6457	0.5313	0.6322	0.5182
FRAUDRE	0.8183	0.7072	0.7603	0.6824	0.7397	0.5674	0.7032	0.6176	0.6861	0.5908	0.6868	0.5884
AMNet	0.8451	0.7432	0.8353	0.7232	0.8042	0.6984	0.7086	0.6385	0.6851	0.6231	0.6692	0.5906
NRGL(Ours)	**0.9409**	**0.8567**	**0.9303**	**0.8427**	**0.9146**	**0.8037**	**0.7636**	**0.6532**	**0.7554**	**0.6443**	**0.7468**	**0.6335**

Table 1: Performance of anomaly detection on Elliptic and Yelp datasets under various noise rates p of 10%, 20% and 30%.

Dataset	#nodes	#edges	#features	Anomaly(%)
Elliptic	46,564	73,248	93	9.76
Yelp	45,954	7,693,958	32	14.53

Table 2: Dataset statistics information.

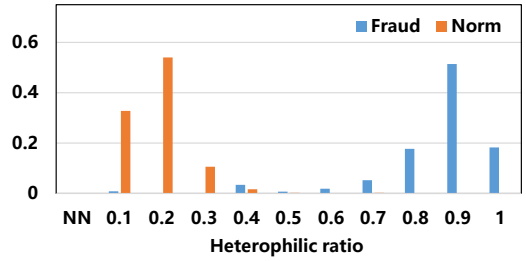


Figure 4: Heterophily evidence in Yelp dataset. "NN" denotes the isolated nodes.

5.2 Baselines

We compare NRGL with the three groups of baseline methods: (1) general GNN models, including GCN [Kipf and Welling, 2017] and GraphSage [Hamilton *et al.*, 2017]; (2)advanced GNN-based anomaly detection methods, including FRAUDRE [Zhang *et al.*, 2021] and AMNet [Chai *et al.*, 2022]; (3) robust GNN models which are specifically designed to resist label noise, including NRGNN [Dai *et al.*, 2021] and RTGNN [Qian *et al.*, 2023].

Evaluation Metrics

Two widely-used evaluation metrics are used for performance evaluation: the Area Under Curve (AUC) and Macro-F1.

Implementation Details

All baseline methods are initialized with the same parameters suggested by their official codes and have been carefully fine-tuned. We deploy the batch size of 512 for both Elliptic and Yelp, the initial learning rate of 0.01, and the high-pass filter strength controller coefficient α of 0.1. The source code of our model is available[1].

5.3 Performance Comparison

To demonstrate the effectiveness of NRGL, we compare it with the above baseline methods on two datasets by varying the noise rate p from 0.1 to 0.3, and the results are shown in Table 1. We have the following observations.

First, as the general GNNs, GCN and GraphSAGE don't perform well, which implies that they have poor resistance to label noise. Furthermore, GraphSAGE performs better than GCN as [Zhu *et al.*, 2020] found that the aggregation of higher-order neighborhoods can help alleviate the effects of graph heterophily. Second, NRGNN and RTGNN, as the advanced label noise-resistant GNNs, perform better than GCN and GraphSAGE but still not well enough. That is because graph heterophily would diminish the effectiveness of homophilic assumption-based graph augmentation, which makes them suffer noise not only from labels but also structures. Third, GNN-based anomaly detection methods (FRAUDRE and AMNet) achieve better performance than robust GNNs by dealing with graph inconsistency and imbalance issues, which means how to alleviate graph heterophily and imbalance problems is the key to performance improvement. Finally, NRGL significantly outperforms all baseline methods on both two datasets under various ratios of label noise, which can be boiled down to the following points. First, NRGL divides the input graph into homophilic and heterophilic views via the unsupervised ranking loss, which provides robust node representations. By doing so, the homophilic graph obtained by structure pruning can better hinder the propagation of misinformation. Next, the labeled nodes are divided into three subsets for fine noise governance, and an imbalance-oriented label sampler is designed to take label confidence, frequency, and structure importance into consideration, which helps select a reliable supervised signal.

5.4 Evidence of Graph Heterophily

Here, we calculate the ratio of heterophilic edges to all adjacent edges for both normal and anomalous nodes on the Elliptic dataset, and then count the rate of the number of nodes with the corresponding heterophily ratio to all nodes of normality and anomaly in Fig. 4. We can observe that over 80% normal nodes have smaller than 20% heterophilic ratio and 90% anomalous nodes have more than 80% heterophilic ratio, which means different social structure patterns between them. Consequently, heterophilic edges of anomalous nodes are widespread in graph anomaly detection, which corrupts the effectiveness of traditional GNNs based on graph homophily assumption. Furthermore, given the incorrect super-

[1]https://github.com/Shzuwu/NRGL

Ablation	10%		20%		30%	
	AUC	F1	AUC	F1	AUC	F1
w/o St. Aug	0.9062	0.8151	0.8806	0.7849	0.8438	0.7461
w/o La. Div	0.9219	0.8347	0.8943	0.8189	0.8603	0.7657
NRGL	**0.9409**	**0.8567**	**0.9303**	**0.8427**	**0.9146**	**0.8037**

Table 3: Results of ablation study on the Elliptic dataset.

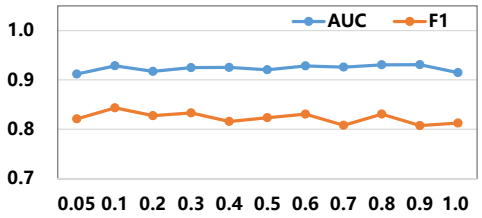


Figure 5: Performance with varing α on Elliptic dataset with the noise ratio p of 0.2.

vised information from label noise, it would further confuse the message passing between connected nodes, which exacerbates the difficulty of anomaly detection.

5.5 Ablation Study

Here, we compare NRGL with two variants on the Elliptic dataset to validate the effectiveness of each component, and similar results can also be observed on the Yelp dataset. Specifically, we remove structure augmentation (w/o St. Aug) and label division (w/o La. Div), respectively.

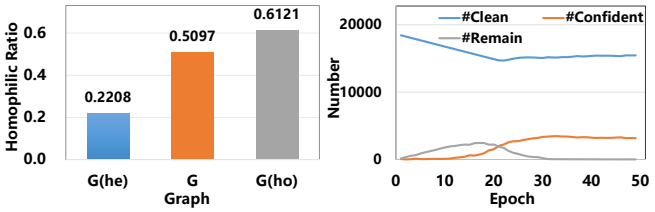
From the results in Table 3, we can observe that NRGL notably outperforms (w/o St. Aug) and (w/o La. Div) on both AUC and Macro-F1, which demonstrates the effectiveness of such two modules. Furthermore, (w/o La. Div) does less well than (w/o St. Aug), which means structure pruning-based graph augmentation and robust node representations can make more contributions to robust graph learning.

5.6 Hyper-Parameters Sensitivity

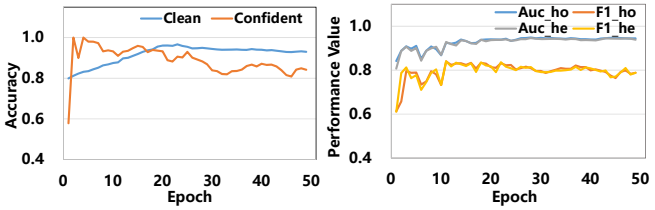
Here, we will study how α affects the performance of NRGL. More specifically, α controls the strength of the high-pass filter, and we range it from 0.05 to 1. The results are reported on the Elliptic dataset with noise rate p of 0.2 in Fig. 5. We can observe that NRGL achieves the best performance when $\alpha = 0.1$, that may be because a high-pass filtering with a larger strength would lead to the loss of original graph signals. Based on this, we finally set α of 0.1.

5.7 Case Study

In Fig. 6 (a), we will show the effectiveness of dual view division on the Elliptic dataset with a noise ratio of 0.2. Specifically, we denote the input graph as $\mathbf{G}$, the homophily-view graph as $\mathbf{G}\,(\mathrm{ho})$ (i.e, remove edges with $\mathbf{A}_{ij}^{(ho)} < 0.5$) and heterophily-view graph as $\mathbf{G}\,(\mathrm{he})$ (i.e., remove edges with $\mathbf{A}_{ij}^{(he)} < 0.5$). Then, we calculate the label consistency coefficient (i.e., homophilic ratio) between each anomaly node and its neighbors. In $\mathbf{G}\,(\mathrm{ho})$ and $\mathbf{G}\,(\mathrm{he})$, we expect this ratio is greater and smaller than that in the original graph, respectively. As expected, the average homophily ratio 0.6121



(a) Changes of Homophily ratio (b) Number changes of subsets



(c) Accuracy changes of subsets (d) AUC and F1 of changes

Figure 6: Case study on the Elliptic dataset with noise ratio p of 0.2.

($\mathbf{G}\,(\mathrm{ho})$) and 0.2208 ($\mathbf{G}\,(\mathrm{he})$) are larger and smaller than 0.5097 ($\mathbf{G}$), which implies the model does help extract homophilic views from the input graph.

Next, we will visualize the changes in label division and performance with epoch increases on the Elliptic dataset with a noise ratio of 0.2 and a training set ratio of 0.4. As introduced above, we divide the labeled nodes into three subsets, i.e., clean set $\mathcal{V}_{cl}$, confident set $\mathcal{V}_{cf}$ and the remaining set $\mathcal{V}_{re}$. In Fig. 6(b), we report the changes in node numbers in each subset. In Fig. 6(c), we show the accuracy of the clean set and confident set, and the corresponding performance of AUC and Macro-F1 derived by both homophilic and heterophilic views, as shown in Fig. 6(d). We can observe that the training span can be divided into three periods. First, with the epoch getting large, the number of clean and confident nodes decreases and increases respectively, and the accuracy of clean and confident sets rises especially the confidence set, and the performance improves rapidly as the model gets quickly learned; Second, all items are of slight fluctuations, and performance increase slowly as the model gradually fits the data; Finally, the model gradually stabilizes where the number of three subsets remains unchanged, but the risk of overfitting makes the accuracy fluctuation of confident subset cause slight performances degradation in F1-score.

6 Conclusion

In the paper, we study an under-researched yet crucial issue of robust graph learning against label noise for anomaly detection in the presence of graph heterophily and imbalance problems. Based on empirical experiments, we find that increasing graph homophily can help resist label noise. To this end, we develop a novel NRGL model, which facilitates robust graph learning from the perspectives of structure augmentation with reliable node representations and an imbalance-oriented sampler for fine-grained label governance. Experimental results on two real-world datasets show the effectiveness of NRGL with varying different ratios of label noise.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (No.U22A2035, U1803262, U1736206) and National Social Science Fund of China (No.19ZDA113).

References

[Arpit *et al.*, 2017] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of International conference on machine learning(ICLR)*, pages 233–242. PMLR, 2017.

[Chai *et al.*, 2022] Ziwei Chai, Siqi You, Yang Yang, Shiliang Pu, Jiarong Xu, Haoyang Cai, and Weihao Jiang. Can abnormality be detected by graph neural networks. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 23–29, 2022.

[Dai *et al.*, 2021] Enyan Dai, Charu Aggarwal, and Suhang Wang. Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 227–236, 2021.

[Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1919–1925, 2017.

[Goldberger and Ben-Reuven, 2016] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *Proceedings of International conference on learning representations*, pages 1–9, 2016.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of Advances in neural information processing systems*, pages 1024–1034, 2017.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of Advances in neural information processing systems*, pages 8536–8546, 2018.

[Huang *et al.*, 2019] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334, 2019.

[Jiang *et al.*, 2018] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of International conference on machine learning*, pages 2304–2313. PMLR, 2018.

[Jin *et al.*, 2023] Di Jin, Bingdao Feng, Siqi Guo, Xiaobao Wang, Jianguo Wei, and Zhen Wang. Local-global defense against unsupervised adversarial attacks on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8105–8113, 2023.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of 5th International Conference on Learning Representations(ICLR)*, pages 1–14, 2017.

[Li *et al.*, 2022] Kuan Li, Yang Liu, Xiang Ao, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Reliable representations make a stronger defender: Unsupervised structure refinement for robust gnn. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 925–935, 2022.

[Li *et al.*, 2024] Xianxian Li, Qiyu Li, Haodong Qian, Jinyan Wang, et al. Contrastive learning of graphs under label noise. *Neural Networks*, 172:1–12, 2024.

[Liu *et al.*, 2021] Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: a gnn-based imbalanced learning approach for fraud detection. In *Proceedings of International World Wide Web Conferences*, pages 3168–3177, 2021.

[NT *et al.*, 2019] Hoang NT, Choong Jun Jin, and Tsuyoshi Murata. Learning graph neural networks with noisy labels. *arXiv preprint arXiv:1905.01591*, pages 1–5, 2019.

[Patrini *et al.*, 2017] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[Qian *et al.*, 2023] Siyi Qian, Haochao Ying, Renjun Hu, Jingbo Zhou, Jintai Chen, Danny Z Chen, and Jian Wu. Robust training of graph neural networks via noise governance. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 607–615, 2023.

[Rayana and Akoglu, 2015] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International CSonference on Knowledge Discovery and Data Mining*, pages 985–994, 2015.

[Tang *et al.*, 2022] Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of International Conference on Machine Learning*, pages 21076–21089. PMLR, 2022.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019.

[Weber *et al.*, 2019] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in Finance Workshop*, pages 1–7, 2019.

[Wu *et al.*, 2022] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.

[Yuan *et al.*, 2023] Jingyang Yuan, Xiao Luo, Yifang Qin, Yusheng Zhao, Wei Ju, and Ming Zhang. Learning on graphs under label noise. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of Advances in neural information processing systems*, pages 8792–8802, 2018.

[Zhang *et al.*, 2021] Ge Zhang, Jia Wu, Jian Yang, Amin Beheshti, Shan Xue, Chuan Zhou, and Quan Z Sheng. Fraudre: Fraud detection dual-resistant to graph inconsistency and imbalance. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 867–876. IEEE, 2021.

[Zhao *et al.*, 2023] Ruoyan Zhao, Zhou Shao, Wenhu Zhang, Jiachen Zhang, and Chunming Wu. A multi-channel multi-tower gnn model for job transfer prediction based on academic social network. *Applied Soft Computing*, 142:1–10, 2023.

[Zheng *et al.*, 2023] Wen Zheng, Bingbing Xu, Emiao Lu, Yang Li, Qi Cao, Xuan Zong, and Huawei Shen. Midlg: Mutual information based dual level gnn for transaction fraud complaint verification. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5685–5694, 2023.

[Zhu *et al.*, 2020] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Proceedings of Advances in neural information processing systems*, pages 7793–7804, 2020.