UniFORM: Towards Unified Framework for Anomaly Detection on Graphs

Chuancheng Song[1,2], **Xixun Lin**[1,2], **Hanyang Shen**[1,2], **Yanmin Shang**[1,2], **Yanan Cao**[1,2*]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{songchuancheng, linxixun, shenhanyang, shangyanmin, caoyanan}@iie.ac.cn

Abstract

Graph anomaly detection has attracted significant attention due to its critical applications, such as identifying money laundering in financial systems and detecting fake reviews on social networks. However, two major challenges persist: (1) anomaly detection at the node, edge, and graph levels is often addressed in isolation, hindering the integration of complementary information to identify anomalies arising from collective behaviors; and (2) the inherent label sparsity in graph data, coupled with the difficulty of obtaining high-quality annotations, exacerbates bias in detection. To address these challenges, we propose UniFORM, a unified self-supervised anomaly detection framework comprising three modules: UIO, UMC and UPL. UIO unifies node-, edge-, and graph-level tasks from a subgraph perspective, leveraging an energy-based GNN for iterative multi-granular anomaly detection. UMC enhances meta-learning through contrastive learning and employs Langevin dynamics to generate phantom samples as substitutes for anomalous data, reducing reliance on labeled data. UPL design unified loss in intra-inter perspectives. Extensive experiments on real-world datasets demonstrate that UniFORM significantly outperforms state-of-the-art methods across multiple granularities.

Introduction

Anomaly detection is a data mining process that aims to identify these unusual patterns that deviate from the majority and represent a specific case of out-of-distribution detection (Ma et al. 2021; Yang et al. 2024; Wang et al. 2024). Recently, anomaly detection on graphs has garnered significant attention due to its critical applications in financial risk control (Shi et al. 2022; Motie and Raahemi 2023), social networks (Lin et al. 2024; Tang et al. 2024), etc.

In recent years, state-of-the-art advances have leveraged contrastive techniques increasingly. CoLA (Liu et al. 2021) employs node-subgraph pairs in contrastive learning to generate node representations, with anomaly scores derived from agreement between these pairs. SL-GAD (Zheng et al. 2021) constructs contextual views around a target node and integrates generative attribute regression with multi-view contrastive training. GOOD-D (Liu et al. 2023a), as an unsupervised framework, employs hierarchical contrastive learning to detect out-of-distribution graphs by capturing latent patterns. GRADATE (Duan et al. 2023a) and NLGAD (Duan et al. 2023b) also use the idea of contrastive learning, aiming to use multi-scale contrasts and multi-view augmentations to construct diverse positive and negative pairs.

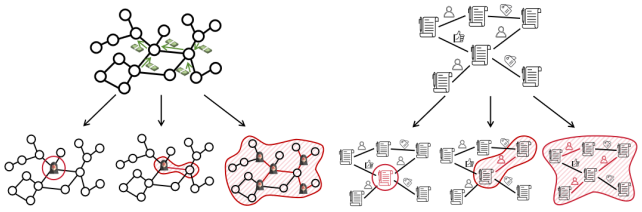

Figure 1: Toy examples for anomaly detection on graphs in node-, edge-, and graph-level. (Left) Money laundering in financial systems: fraudster, abnormal transaction, and criminal activity. (Right) Fake reviews on social networks: fake review, spammer, and collaborative group.

Synchronously, more researches consider the relationship between the locality and globality of anomalies in the network. AANE (Duan et al. 2020) consider anomalous nodes and edges with employing GCN to generate node embeddings and calculates link probability using the hyperbolic tangent of these embeddings. GLocalKD (Ma et al. 2022) utilizes knowledge distillation to learn comprehensive global and local normal patterns for detecting anomalous graphs. BOURNE (Liu et al. 2024b) represents a unified framework that employs self-supervised bootstrap learning to detect anomalies at the node and the edge using graph and hypergraph representations.

Despite recent advancements, graph anomaly detection faces two critical challenges: ***(1) Limited integration of multi-granular anomalies.*** Current methods often treat node-, edge-, and graph-level anomalies as separate tasks, overlooking their interconnectedness in real-world scenarios. For example, in financial networks, fraudsters (node-level), suspicious transactions (edge-level), and overarching criminal activities (graph-level) are interdependent. Insights from node-level anomalies can enhance edge-level detection (Liu et al. 2024b), and vice versa, while graph-level anomalies often emerge from collective behaviors across nodes and

*Corresponding author.

edges (Mukherjee, Liu, and Glance 2012). Failing to integrate these perspectives restricts comprehensive analysis of individual and community information. ***(2) Label sparsity in graph data.*** Graph data suffer from pronounced class imbalance, where anomalous samples are vastly outnumbered by normal ones. State-of-the-art methods, such as contrastive learning, depend on abundant negative samples, exacerbating the challenges of imbalance. Moreover, obtaining high-quality labels for graph data is difficult. Existing approaches that address adaptable tasks often rely on classical meta-learning paradigms, requiring manually constructed meta-tasks with labeled support and query sets, further limiting their scalability.

We propose **UniFORM** (**Uni**fied **F**ramew**OR**k for Ano**M**aly Detection on Graphs), comprising three modules: Unified Input and Output among multi-levels (UIO), Unified Meta-learning with Contrastive learning (UMC), and Unified Intra-Inter Perspective for Loss (UPL). For ***challenge (1)***, UIO leverages dual subgraph sampling methods in the purpose of making full use of multi-granular information, and employs an energy-based GNN to compute sample energy for anomaly characterization. For ***challenge (2)***, UMC design a self-supervised meta-learning framework inspired by contrastive learning to reduce label dependency and introduce Langevin dynamics-based sampling to mitigate reliance on negative samples. Finally, UPL construct loss functions to guide model learning from both intra-level and inter-level perspectives. Our main contributions are as follows:

- To the best of our knowledge, UniFORM is the first framework that integrates node-level, edge-level, and graph-level anomaly detection tasks.
- UniFORM synergizes meta-learning with contrastive learning to be a self-supervised framework and reduce its dependence on labeled data. Additionally, it innovatively introduces Langevin Dynamics for sampling enhancement, thereby improving contrastive learning and effectively addressing the imbalance challenge.
- Extensive experiments on real-world datasets demonstrate the superior performance of UniFORM compared to existing state-of-the-art methods.

Related Work

Anomaly Detection on Graphs

Node-level anomaly detection has dominated the existing research. For example, H^2-FDetector (Shi et al. 2022) believes that homophilic and heterophilic connections propagate completely different information, which should be considered in the anomaly detection model. ACT (Wang et al. 2023) applies contrastive learning to align the representations of labeled nodes to facilitate the learning of complex normal node relations without relying on predefined anomaly distributions. *Edge-level anomaly detection* focuses on identifying anomalous connections that signal unexpected or irregular relationships between entities. (Ouyang, Zhang, and Wang 2020) models edge distributions to detect existing edges with the lowest likelihood of occurrence as anomalies. RGSE (Liu et al. 2023b) leverages robust graph structure embedding to generate both global and local stable views, facilitating the detection of anomalous links within contaminated graphs. *Graph-level anomaly detection* seeks to identify entire graphs that significantly deviate from the majority within a graph set. OCGTL (Qiu et al. 2022) leverages transformation learning principles to improve anomaly detection accuracy. SIGNET (Liu et al. 2024c) applies the information bottleneck theory to develop a self-interpretable method.

Despite these innovations in detecting anomalies at the node-, edge-, or graph-level, addressing these multi-granular tasks concurrently remains an underexplored area.

Graph Contrastive Learning

Contrastive learning identifies latent data representations by maximizing the similarity between related samples and minimizing it between unrelated ones. Early works like DGI (Velickovic et al. 2019), MVGRL (Hassani and Khasahmadi 2020) and BiGI (Cao et al. 2021) maximize mutual information between local and global graph features or between different graph views to learn node representations. Recent work emphasizes graph augmentation, as in GRACE (Zhu et al. 2020), which improves node-level agreement between corrupted views, and GraphCL (You et al. 2020), which uses contrastive loss to align representations of augmented graph views. Advanced approaches integrate contrastive learning with other techniques, such as GACN (Wu et al. 2023a), which combines it with graph generative adversarial networks for automatic view augmentation, and COLA (Liu et al. 2024a), which pairs it with meta-learning for few-shot node classification by identifying semantically similar nodes.

Overall, graph contrastive learning is crucial for addressing the challenge of insufficient valid samples, particularly in graph data.

Preliminary

Problem Definition

Let $\mathcal{G} = \{G_1, G_2, \cdots, G_N\}$ denote N graphs. For each graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{X})$, $\mathcal{V}$ is the node set and node number $|\mathcal{V}| = n$; $\mathcal{E}$ is the edge set and edge number $|\mathcal{E}| = m$; $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_{n+m+1}, \mathbf{x}_{n+m+2}\}$ contains feature embeddings of all nodes, edges and the graph itself, i.e. $\mathcal{X} \in \mathbb{R}^{(n+m+1)\times d}$. For each entry $\mathbf{x}_i \in \mathbb{R}^d$ in $\mathcal{X}$, we aim to learn a score function $s = \psi(\cdot)$ to generate an anomaly detection score. The larger the s value, the more likely $\mathbf{x}_i$ is anomalous. Associated with $\psi(\cdot)$, binary decision function $\Delta(\cdot)$ can be formulated as:

$$\Delta(\mathbf{x}_i; \psi) = \begin{cases} 0, & \mathbf{x}_i \in \mathcal{X}^n, \\ 1, & \mathbf{x}_i \in \mathcal{X}^a, \end{cases} \tag{1}$$

where $\mathcal{X}^n$ represent normal samples and $\mathcal{X}^a$ represent anomaly samples and $\mathcal{X} = \mathcal{X}^n + \mathcal{X}^a$.

Graph Meta-learning Paradigm

The traditional meta-training samples classes by anchoring different labels from the training set $\mathcal{X}_{train}$ to generate

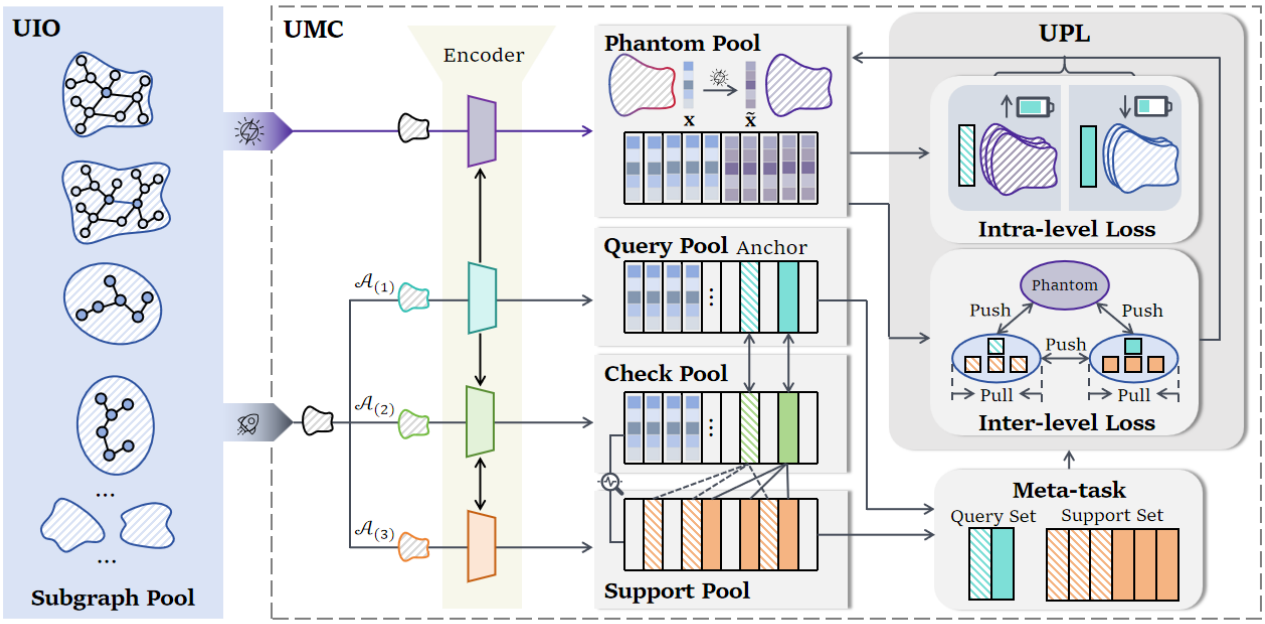

Figure 2: The overall framework of UniFORM.

meta-tasks (Lin et al. 2022). For a N-way k-shot meta-task $\mathcal{T}$, a subset C_{meta} with N classes is randomly sampled from C_{train} (classes of $\mathcal{X}_{train}$) to generate the *support set* $\mathcal{S}$ and *query set* $\mathcal{Q}$ synchronously:

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, y_i) | y_i \in C_{meta}, i = 1, \cdots, N \times k\} \\ \mathcal{Q} &= \{(\mathbf{x}_i, y_i) | y_i \in C_{meta}, i = 1, \cdots, N \times q\} \\ \mathcal{S} &\cap \mathcal{Q} = \emptyset, k \ll q \end{aligned} \tag{2}$$

Nevertheless, we shift from using labels as anchors for meta-tasks and propose a self-supervised meta-learning framework.

Methodology

In this section, we exhibit UniFORM in detail and introduce components UIO, UMC and UPL in the following subsections.

Unified Input and Output among Multi-levels

UIO leverages dual subgraph sampling strategy to make full use of multi-granular information, and employs an energy-based GNN for anomaly characterization.

Dual Subgraph Sampling To standardize input for anomaly detection tasks of varying granularities and maximize information capture, we designed a dual subgraph sampling strategy that collects (1) node- or edge-centered ego-graphs (see Fig. 3a) and (2) subgraph fragments from random walks (see Fig. 3b).

Ego-graphs capture fine-grained features of central nodes or edges but are insufficient for representing holistic properties. To Enrich the information representation, we employ the RWR algorithm (Duan et al. 2006) (the transition probability is proportional to the node's information content in practice) to generate subgraph fragments. Formally, nodes and edges derive sample sets $S_{\mathcal{V}}$ and $S_{\mathcal{E}}$ via ego-graph function $f_e(\cdot)$, while graphs obtain sample sets S_G through random walk function $f_r(\cdot)$ as follows:

$$\begin{aligned} &S_{\mathcal{V}} = f_e(\mathcal{V}),\ S_{\mathcal{E}} = f_e(\mathcal{E}),\ S_G = f_s(G), \\ &\mathcal{P}_s = \{\mathbf{x}_i | s_i \in S_{\mathcal{V}} \cup S_{\mathcal{E}} \cup S_G\}, \end{aligned} \tag{3}$$

where all these ego-graphs and subgraph fragments are then merged into the `Subgraph Pool` $\mathcal{P}_s = \{\mathbf{x}_1, \cdots, \mathbf{x}_{|\mathcal{P}_s|}\}$ to become unified input samples in the subsequent pipeline. After input unification, the graph is restructured from "node-node" to "sample-sample" connections, increasing complexity as nodes may appear in multiple samples. It is reasonable that high-information nodes naturally receive more attention during sampling due to their richer content.

Energy-based GNN Research on anomaly scores is extensive, with approaches based on energy gaining attention recently (LeCun et al. 2006; Wu et al. 2023b). The energy-based GNN $g_e(\cdot)$ encodes each input sample s_i to produce its feature $\mathbf{x}_i = g_e(s_i)$. Then the energy function $E_\theta(\mathbf{x}_i) : \mathbb{R}^d \to \mathbb{R}$ maps this feature vector to a scalar value termed *energy*, derived from a *Boltzmann distribution* (details in *Theorem 4.1*). We extend this to the graph domain by designing an $g_e(\cdot)$ to compute the sample energy as an anomaly score. The reconstructed GNN architecture consists of K layers, where $\mathbf{x}_i$ is transformed via a linear layer:

$$\mathbf{h}_i^{(0)} = \mathbf{W}_0 \mathbf{x}_i + \mathbf{b}_0, \tag{4}$$

where $\mathbf{h}_i^{(0)}, \mathbf{b}_0 \in \mathbb{R}^f$ and $\mathbf{W}_0 \in \mathbb{R}^{f \times d}$. f is the hidden dimension of the sample representations. The representation of sample s_i in layer k of the GNN, denoted as $\mathbf{h}_i^{(k)}$, is then derived by aggregating information from its neighborhood

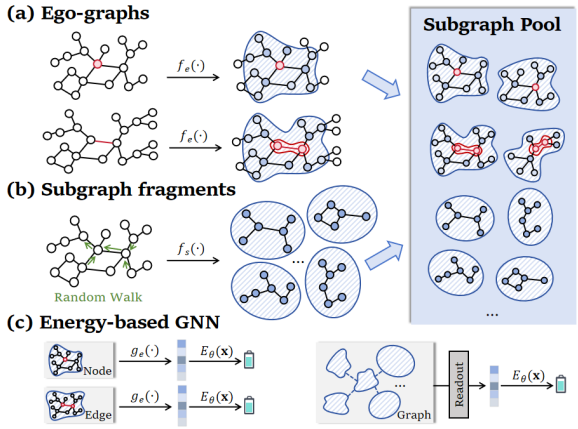

Figure 3: UIO: Unified Input and Output among Multi-levels: (a) Dual Subgraph Sampling: ego-graph; (b) Dual Subgraph Sampling: random walk; (c) Energy-based GNN

(samples containing neighbor nodes), formalized as:

$$\mathbf{h}_i^{(k)} = \gamma^{(k)}(\mathbf{h}_i^{(k-1)} + \sum_{j \in \mathcal{N}(i)} \frac{\phi_{ij}}{\sqrt{(d_i+1)(d_j+1)}} \mathbf{h}_j^{(k-1)}), \tag{5}$$

where $\gamma^{(k)} \in \mathbb{R}^{f \times f}$, d_i and d_j denote the degrees of sample s_i and s_j, respectively, standardizing the information flow, and $\phi_{ij} \in [0, 1]$ encapsulates the connectivity between these samples, thereby regulating the influence of neighboring samples:

$$\phi_{ij} = \frac{\cos(\mathbf{U}\mathbf{x}_i, \mathbf{U}\mathbf{x}_j) + 1}{2}, \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{f \times d}$ is a learnable parameter, and this design can mitigate the over-smoothing issue inherent in GNNs (Steck, Ekanadham, and Kallus 2024). Subsequently, the node feature $\mathbf{x}_{node} = \{\mathbf{x}_i, s_i \in S_\mathcal{V}\}$ and edge feature $\mathbf{x}_{edge} = \{\mathbf{x}_i, s_i \in S_\mathcal{E}\}$ are represented by their ego-graph samples, and the graph feature $\mathbf{x}_{graph}$ in layer k can be read out as:

$$\mathbf{x}_{graph}^{(k)} = Readout^{(k)}(\mathcal{P}_s) = \sigma(\frac{1}{|\mathcal{P}_s|} \sum_{i=1}^{|\mathcal{P}_s|} \mathbf{W}^{(k)} \mathbf{h}_i^{(k)}), \tag{7}$$

where σ is a sigmoid function. Drawing inspiration from PageRank-based strategies (Yang et al. 2021), the energy function $E_\theta(\mathbf{x})$ computes the sample energy by aggregating the representations across all GNN layers as follows:

$$E_\theta(\mathbf{x}_i) = -\log \sum_{j=1}^{f} \exp\{\sum_{k=0}^{K} \zeta_k \mathbf{x}_i^{(k)}\}_{[j]}, \tag{8}$$

where $\zeta_k \in \mathbb{R}$ is a a learnable parameter, and $\cdot_{[j]}$ is the j-th component of the vector. A larger $E_\theta(\mathbf{x})_i$ means this sample is more likely from $\mathcal{X}^a$.

Unified Meta-Learning with Contrastive Learning

UMC designs a self-supervised meta-learning framework inspired by contrastive learning to reduce label dependency and introduce Langevin dynamics-based sampling to mitigate reliance on negative samples.

Langevin Dynamics Langevin Dynamics (Chan 2024) is innovatively exploited to generate phantom samples to alleviate imbalance nature in this paper. Phantom samples in `Phantom Pool` $\mathcal{P}_P = \{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{|\mathcal{P}_P|}\}$ act as mediators between abnormal and normal samples, thus improving their discriminability. Moreover, Langevin Dynamics is appropriately effective for energy-based evaluation of anomaly detection on graphs theoretically:

Theorem 4.1 Boltzmann Distribution. The Boltzmann Distribution establishes a mapping from energy $E_\theta(\cdot)$ to probability distribution $p_\theta(\cdot)$ and is related to Langevin Dynamics inextricably: the diffusion of stochastic process $x(t)$ obeying the Brownian Motion converges to a stationary distribution, which is called the Boltzmann distribution $p_\theta(\mathbf{x})$:

$$p_\theta(\mathbf{x}) \propto exp(-E_\theta(\mathbf{x}))/Z_\theta \tag{9}$$

where $Z_\theta = \int exp(-E_\theta(\mathbf{x}))$ is the partition function.

Specifically, the phantom $\tilde{s}$ is sampled from model distribution p_θ approximately by stochastic gradients perturbing iteratively. Formally, the phantom attribute $\tilde{\mathbf{x}}$ is updated from $(t-1)$-step to t-step as follows (Qin et al. 2022):

$$\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \epsilon \psi_\theta(\tilde{\mathbf{x}}_{t-1}) + \sqrt{2\epsilon}\mathbf{z}, \ \mathbf{z} \sim \mathcal{N}(0, I), \tag{10}$$

where ϵ is the step size and $\mathbf{z}$ is the Gaussian noise added at each iteration to facilitate gradual oscillation convergence and improve the effect. $\psi_\theta(\cdot)$ is the score function (Swersky et al. 2011; Hyvärinen and Dayan 2005) of the observed data distribution $\mathbf{x}$, defined as

$$\psi_\theta(\mathbf{x}) = \nabla_\mathbf{x} \log p_\theta(\mathbf{x}) = -\nabla_\mathbf{x} E_\theta(\mathbf{x}) \tag{11}$$

where ∇ is the gradient operator, $p_\theta(\mathbf{x})$ is the data distribution density. Apparently, samples at multi-grained levels iteratively generate phantom samples $\tilde{s}$, introducing perturbation in the energy gradient. The gradient descent in Eq. 10 decreases the energy of normal samples and increases the energy of phantom samples. In practice, phantom samples are generated for all samples and used in contrastive learning based on the selection of anchors.

Label-free Meta-Task Construction Given graph G, $\mathcal{A}(G)$ denotes the distribution of graph data augmentations, typically involving Node Dropping, Edge Perturbation, Attribute Masking, etc (You et al. 2020). Following (Liu et al. 2024a), we apply three distinct augmentations $\mathcal{A}_{(1)}, \mathcal{A}_{(2)}, \mathcal{A}_{(3)} \sim \mathcal{A}$ (a single or a combination of multiple augmentation methods), generating corresponding augmented views $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \mathbf{x}_{(3)}$ to construct:

- `Query Pool`: $\mathcal{P}_Q = \{\mathbf{x} = g_e(\mathcal{A}_{(1)}(s))\}$.
- `Check Pool`: $\mathcal{P}_C = \{\mathbf{x} = g_e^{ema}(\mathcal{A}_{(2)}(s))\}$.
- `Support Pool`: $\mathcal{P}_S = \{\mathbf{x} = g_e^{ema}(\mathcal{A}_{(3)}(s))\}$.

where the embeddings in $\mathcal{P}_Q$ are generated by a trainable graph encoder $g_e(\cdot)$. The embeddings in $\mathcal{P}_C$ and $\mathcal{P}_S$ are generated by a momentum encoder $g_e^{ema}(\cdot)$, of which the weights are exponential moving average (EMA) (Nakano, Takahashi, and Takahashi 2017) from $g_e(\cdot)$. They are used to construct *query set* and *support set* for label-free meta-tasks, and their specific roles will be described in detail later.

		Research Network			Social Network					Commercial Network	
		Cora	Pubmed	COLLAB	BlogCatalog	Flickr	Enron	IMDB	Reddit	Yelp	IBM-AML
Node-level	ANOMALOUS	0.577	0.7316	0.5851	0.7237	0.7434	0.7083	-	0.5589	0.5277	0.5363
	DOMINANT	0.8155	0.8081	OOM	0.7468	0.7442	0.7610	-	0.5518	OOM	OOM
	CoLA	0.8779	0.9512	0.7996	0.7854	0.7513	0.8273	-	0.5462	0.7034	0.7207
	SL-GAD	0.9130	**0.9672**	0.7912	0.8184	0.7966	0.8355	-	0.4963	0.7463	0.7521
	BOURNE	0.9116	0.9561	0.8285	0.8145	0.7821	0.7804	-	0.6227	0.7834	0.7109
	UniFORM (Ours)	**0.9246**	0.9428	**0.8392**	**0.8635**	**0.8035**	**0.8524**	-	**0.6332**	**0.8103**	**0.7782**
Edge-level	AAGAE	0.7057	0.8162	0.7550	0.7868	0.6691	0.8541	-	-	0.7010	0.6333
	AAGCN	0.7125	0.8025	0.7768	0.7937	0.6453	0.8243	-	-	0.6909	0.6149
	SEAL	0.9136	0.7471	0.8291	0.8159	0.7464	0.8916	-	-	0.7462	0.7028
	RGSE	0.9394	0.7564	0.8326	0.9014	0.7745	**0.9679**	-	-	0.7541	0.7153
	BOURNE	0.8585	0.9765	0.8021	0.7433	0.8038	0.8415	-	-	0.7879	0.7662
	UniFORM (Ours)	**0.9487**	**0.9771**	**0.8615**	**0.9023**	**0.8278**	0.8756	-	-	**0.8125**	**0.7890**
Graph-Level	InfoGraph-iF	-	-	0.4627	-	-	-	0.5653	0.6850	0.6467	0.4298
	GraphCL-iF	-	-	0.4761	-	-	-	0.5654	0.7180	0.6153	0.4497
	OCGTL	-	-	0.6070	-	-	-	0.6019	0.7593	0.6348	0.6339
	GLocalKD	-	-	0.5294	-	-	-	0.5209	0.7785	0.7415	0.7433
	GOOD-D	-	-	0.7208	-	-	-	0.6588	0.8867	0.7113	0.7306
	UniFORM (Ours)	-	-	**0.7472**	-	-	-	**0.6832**	**0.8993**	**0.7842**	**0.7561**

Table 1: Node/Edge/Graph-level anomaly detection results with AUC values (%). "OOM" denotes the detector is out of memory on the dataset, and "-" denotes the detector is not applicable to the dataset. The best performance is **boldfaced**.

Based on these `Pool` tools, the construction of label-free meta-tasks is as follows: first, select N samples, treating them as N classes to form the *query set* $\mathcal{Q} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, where s_i is the query sample of the i-th way. For a N-way k-shot meta-task, the *support set* $\mathcal{S}$ should include k samples with similar semantics to the query sample in each of the N classes. In order to identify similar samples, we first obtain embeddings for the query samples from $\mathcal{P}_C = \{\mathbf{x}_1^C, \cdots, \mathbf{x}_N^C\}$. For each $i \in [1, \cdots, N]$, we measure the similarity between $\mathbf{x}_i^C$ and all embeddings in $\mathcal{P}_S = \{\mathbf{x}_1^S, \cdots, \mathbf{x}_{|\mathcal{P}_S|}^S\}$. The k most similar embeddings from $\mathcal{P}_S$ are then selected as *support set*:

$$\mathcal{S} = \{\mathbf{x}_1^S, \cdots, \mathbf{x}_{N\times k}^S\} \rightarrow \{S_{\mathbf{x}_i^1}, \cdots, S_{\mathbf{x}_i^k}\}_{i=1}^N, \quad (12)$$

where $S_{\mathbf{x}_i^j}$ is the j-th support sample for the i-th query. Finally, embeddings in $\mathcal{P}_Q$ that correspond to $\mathcal{P}_C$ are selected as the *query set*:

$$\mathcal{Q} = \{\mathbf{x}_1^Q, \cdots, \mathbf{x}_N^Q\} \rightarrow \{Q_{\mathbf{x}_1}, \cdots, Q_{\mathbf{x}_N}\}, \quad (13)$$

where the arrow right sides only ensure consistency between the *support set* and *query set* formally.

In summary, the N-way k-shot task $\mathcal{T}$ can be represented as $\mathcal{T} = \{Q_{\mathbf{x}_i}, \{S_{\mathbf{x}_i^j}\}_{j=1}^k\}_{i=1}^N$ and Fig. 2 illustrates a 2-way 3-shot meta-task example.

Unified Intra-Inter Perspective for Loss

In UPL, for each way i, the query embedding $Q_{\mathbf{x}_i}$ serves as the anchor, with support embeddings $\{S_{\mathbf{x}_i^1}, \cdots, S_{\mathbf{x}_i^k}\}$ as positive samples and $\{S_{\mathbf{x}_r^1}, \cdots, S_{\mathbf{x}_r^k}\}_{r\neq i}$ from other ways as negative samples. The inter-level loss for each N-way k-shot meta-task is formally expressed as

$$\mathcal{L}_{inter} = -\sum_{i=1}^{N} \frac{1}{k} \sum_{j=1}^{k} \log \frac{exp(Q_{\mathbf{x}_i} \cdot S_{\mathbf{x}_i^j})}{U_S + U_P}, \quad (14)$$

where $Q_{\mathbf{x}_i}$ is the query sample in i-th way, and $S_{\mathbf{x}_i^j}$ is the j-th support sample. $U_S = \sum_{\mathbf{x}\in\mathcal{S}} exp(Q_{\mathbf{x}_i} \cdot \mathbf{x})$ and $U_P = \sum_{\tilde{\mathbf{x}}\in\mathcal{P}_P} exp(Q_{\mathbf{x}_i} \cdot \tilde{\mathbf{x}})$. Finally, the inter-loss for each meta-train episode is the average loss across multiple meta-tasks. Then for the intra-level, the model aims to decrease the energy of normal samples and increase the energy of phantom samples:

$$\mathcal{L}_{intra} = \frac{1}{|\mathcal{X}_n|} \sum_{\mathbf{x}_i\in\mathcal{X}_n} E_\theta(\mathbf{x}_i) - \frac{1}{|\mathcal{P}_P|} \sum_{\mathbf{x}_j\in\mathcal{P}_P} E_\theta(\mathbf{x}_j), \quad (15)$$

where $\mathcal{X}_n$ abbreviates $\mathcal{X}_{norm}$. Additionally, we apply regularization to the energy magnitudes for stable training, using the squared energy terms as in (Liu et al. 2021):

$$\mathcal{L}_{reg} = \frac{1}{|\mathcal{X}_n|} \sum_{\mathbf{x}_i\in\mathcal{X}_n} E_\theta(\mathbf{x}_i)^2 + \frac{1}{|\mathcal{P}_P|} \sum_{\mathbf{x}_j\in\mathcal{P}_P} E_\theta(\mathbf{x}_j)^2 \quad (16)$$

In summary, the overall loss function of UniFORM is given by $\mathcal{L} = \lambda_1\mathcal{L}_{inter} + \lambda_2\mathcal{L}_{intra} + \lambda_3\mathcal{L}_{reg}$, where λ_1, λ_2 and λ_3 are balance parameters.

Experiments

We would answer the following research questions:

- **RQ1**: Can UniFORM handle node, edge, and graph-level tasks, outperforming SOTA methods at each level?
- **RQ2**: How does the integration of intra-level and inter-level perspectives enhance detection?
- **RQ3**: How do key parameters influence the experimental outcomes?

Experiment Setup

Datasets We conduct experiments using datasets from three distinct domains: Research Networks (Cora, Pubmed,

COLLAB), Social Networks (BlogCatalog, Flickr, Enron, IMDB, Reddit), and Commercial Networks (Yelp, IBM-AML). These datasets include two types of benchmark graphs used in previous anomaly detection studies: ground-truth anomaly graphs, which are rare, and injected anomaly graphs, which are more commonly analyzed. To rigorously evaluate the effectiveness of UniFORM, we employ ten real-world graph datasets, comprising both ground-truth (Enron, Yelp, IBM-AML) and injected anomaly graphs (Cora, Pubmed, COLLAB, BlogCatalog, Flickr, Enron, IMDB, Reddit). Given the scarcity of ground-truth anomalies, we also used artificially injected anomalies, following the perturbation scheme outlined in (Liu et al. 2021). Specifically, a dataset with a single graph can only perform node-level and edge-level anomaly detection tasks, while a dataset with multiple graphs can support all tasks. Notably, the IBM-AML dataset is temporal, where each graph is constructed by slicing the data at specific time intervals.

Baselines

- **Node-level**: The field of node anomaly detection is extensive. For our comparative analysis, we select the most representative and state-of-the-art models: ANOMALOUS, DOMINANT, CoLA, SL-GAD, and BOURNE.
- **Edge-level**: Research on edge anomaly detection is relatively limited. We include two variants of AANE (AAGAE, AAGCN), along with SEAL, RGSE, and BOURNE as our baselines. Notably, BOURNE is applicable to both node and edge anomaly detection.
- **Graph-level**: For graph-level anomaly detection, our selected baselines include InfoGraph-iF, GraphCL-iF, OCGTL, GLocalKD, and GOOD-D.

Experimental Setting and Implementation For efficiency and performance, we fixed the sampled community size c (central component plus $c-1$ hop neighbors for ego-graph, and random walk steps for subgraph fragments) to 4. For isolated nodes or those in smaller communities, nodes are repeatedly sampled until an overlapping community of the desired size is formed. In Langevin Dynamics, we select $\epsilon = 0.3$ and $T = 25$, justified below. The energy-based GNN uses 2 layers ($K = 2$) to extract information from small communities, with an embedding dimension $f = 64$. Batch size is set to 300 for each dataset. All models are optimized using the Adam optimizer. Training epochs are 200 for Cora, Pubmed, BlogCatalog, and Flickr; 400 for Enron, IMDB, and Reddit; and 600 for COLLAB, Yelp, and IBM-AML. Learning rates are 0.001 for Cora, Pubmed, BlogCatalog, and Flickr, and 0.003 for the others. All models were run on Python 3.9.19, NVIDIA Tesla V100 GPU, 629GB RAM, and 2.20GHz Intel Xeon E5-2650 CPU.

Evaluation Metric The ROC curve demonstrates the model's stability by balancing the True Positive Rate (TPR) and False Positive Rate (FPR). Consequently, the AUC (Area Under the Curve) $\in [0, 1]$ serves as the evaluation metric and a higher AUC indicates better model performance.

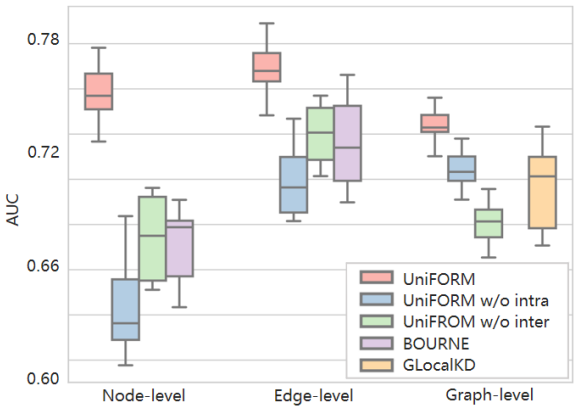

Figure 4: Ablation analysis for UniFORM model on IBM-AML dataset.

Performance Comparison (RQ1)

To address RQ1, we compare UniFORM's performance with state-of-the-art methods across all three levels. The results, shown in Table 1, reveal the following:

- For node-level evaluation, UniFORM achieves the best results on most datasets, except on the Pubmed with SL-GAD. This discrepancy is likely due to the high number of anomalies injected into Pubmed, while our energy-based model aims to detect anomalies from different distributions. Despite this, UniFORM leads on nearly all datasets, with a maximum improvement of 4.51% on BlogCatalog.
- For edge-level evaluation, UniFORM also outperforms most methods. As a unification pioneer, BOURNE integrated node- and edge-level anomaly detection, but UniFORM surpasses it, with an improvement of up to 2.89%.
- For graph-level anomaly detection, UniFORM outperforms all the methods, with a maximum improvement of 4.27%. Although most datasets are specific to a single level, we adapt the more comprehensive datasets (COLLAB, Yelp, and IBM-AML) to support tasks at all levels, where UniFORM demonstrated overall superiority.

Ablation Study (RQ2)

In the loss design, the comprehensive integration of intra-level and inter-level perspectives across multi-grained levels captures crucial associations between individuals and communities, both of which are indispensable. To demonstrate the importance of this inter-intra design, we conducted ablation studies on the IBM-AML dataset, comparing UniFORM with two variants:

- UniFORM w/o intra loss: Excludes intra-level loss and replaces Langevin dynamics with random feature perturbation for contrastive learning.
- UniFORM w/o inter loss: Excludes inter-level loss, replacing meta-learning with independent training of encoders for each level.

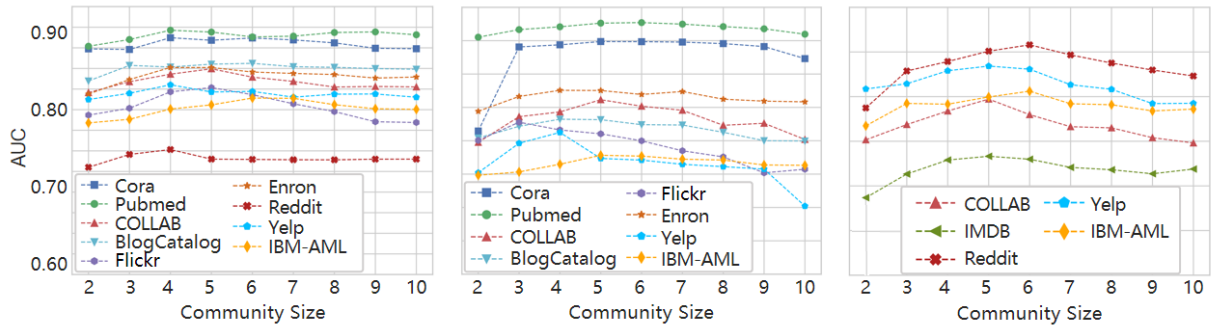

Figure 5: Parameter study for c: the impact of different community sizes in three-times detection w.r.t. AUC values on IBM-AML for (left) node-level, (middle) edge-level, (right) graph-level.

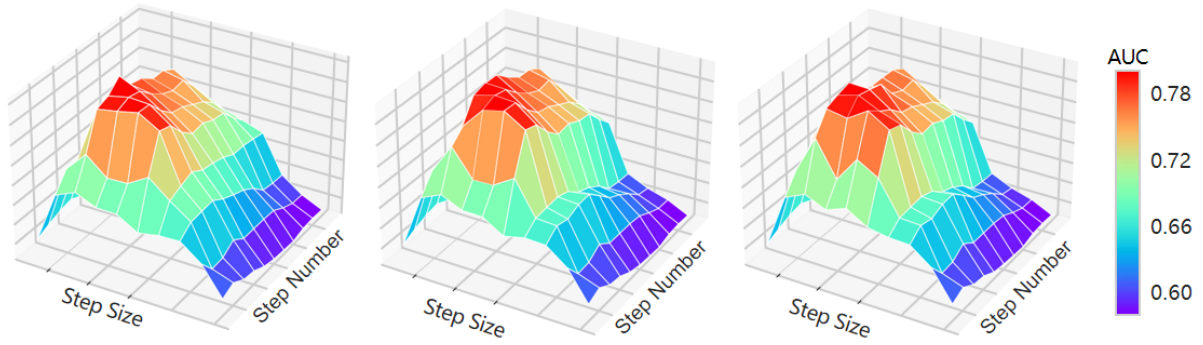

Figure 6: Parameter study for ϵ and T: the impact of different step size and step number w.r.t. AUC values on IBM-AML for (left) node-level, (middle) edge-level, (right) graph-level.

Given that BOURNE is the first method to jointly probe node- and edge-level anomalies, and GLocalKD the first to integrate global and local information for graph-level tasks, we compare UniFORM against them across all three levels, as shown in Fig. 4. Clearly, the two components complement each other, and removing either significantly impacts UniFORM's performance. Notably, the blue and green boxes are inverted for the different levels of the task:: w/o intra performs worse than w/o inter on node- and edge-level tasks, while the opposite is true for graph-level tasks. This difference arises because intra-level analysis focuses on individual sample information, while inter-level analysis captures potential interactions between samples, demonstrating the rationale behind our design.

Parameter Study (RQ3)

We further analyze the impact of key parameters, such as community size c, step size ϵ, and step number T.

Since UniFORM samples both ego-graphs and subgraph fragments, the community size c significantly impacts the results. As shown in Fig. 5, a very small c leads to low AUC due to insufficient structural information, as the community includes only the target node and one-hop neighbors. AUC improves with larger communities but decreases when $c > 6$ due to noise from distant nodes. Optimal performance is achieved when $c \in [3, 6]$. Fine-grained tasks perform best with $c \in [3, 4]$, while holistic tasks perform best with $c \in [4, 6]$. For efficiency and robustness, we set c to 4 for all datasets.

The iteration of `Phantom Pool` is a critical factor and step size ϵ and step number T are the core elements. Therefore, we study the effect of them, as depicted in Fig. 6. Warm colors indicate high values and cool colors have low values. Apparently, extremes in either direction result in suboptimal outcomes. The optimal step size $\epsilon \in [0.2, 0.4]$, and the best step number $T \in [20, 30]$. To balance model stability and performance, we select $\epsilon = 0.3$ and $T = 25$.

Conclusion

In this paper, we present UniFORM, the first unified framework integrating node-level, edge-level, and graph-level anomaly detection tasks on graph. The label-free meta-learning framework with contrastive learning alleviates label scarcity and inaccessibility. Experimental results demonstrate its superiority. We believe this research offers valuable new insights into anomaly detection on graphs.

Acknowledgments

This work is supported by the Key Program of the General Technology Joint Fund of the National Natural Science Foundation of China (No. U2336202) and the National Natural Science Foundation of China (No. 62402491). We thank the reviewers for their insights and useful advice during the different reviewing processes.

References

Cao, J.; Lin, X.; Guo, S.; Liu, L.; Liu, T.; and Wang, B. 2021. Bipartite graph embedding via mutual information maximization. In *Proceedings of the 14th ACM international conference on web search and data mining*, 635–643.

Chan, S. H. 2024. Tutorial on Diffusion Models for Imaging and Vision. arXiv:2403.18103.

Duan, D.; Tong, L.; Li, Y.; Lu, J.; Shi, L.; and Zhang, C. 2006. Fast random walk with restart and its applications. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-06)*, 613–622.

Duan, D.; Tong, L.; Li, Y.; Lu, J.; Shi, L.; and Zhang, C. 2020. Aane: Anomaly aware network embedding for anomalous link detection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-20)*, 1002–1007.

Duan, J.; Wang, S.; Zhang, P.; Zhu, E.; Hu, J.; Jin, H.; Y., L.; and Dong, Z. 2023a. Graph Anomaly Detection via Multi-Scale Contrastive Learning Networks with Augmented View. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*, 7459–7467.

Duan, J.; Zhang, P.; Wang, S.; Hu, J.; Jin, H.; Zhang, J.; Zhou, H.; and Liu, X. 2023b. Normality Learning-based Graph Anomaly Detection via Multi-Scale Contrastive Learning. In *Proceedings of the ACM International Conference on Multimedia (MM-23)*, 7502–7511.

Hassani, K.; and Khasahmadi, A. 2020. Contrastive multi-view representation learning on graphs. In *Proceedings of the International Conference on Machine Learning (ICML-20)*, 4116–4126.

Hyvärinen, A.; and Dayan, P. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 695–709.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1–59.

Lin, X.; Li, Z.; Zhang, P.; Liu, L.; Zhou, C.; Wang, B.; and Tian, Z. 2022. Structure-aware prototypical neural process for few-shot graph classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4): 4607–4621.

Lin, X.; Zhang, W.; Shi, F.; Zhou, C.; Zou, L.; Zhao, X.; Yin, D.; Pan, S.; and Cao, Y. 2024. Graph Neural Stochastic Diffusion for Estimating Uncertainty in Node Classification. In *Forty-first International Conference on Machine Learning*.

Liu, H.; Feng, J.; Kong, L.; Tao, D.; Chen, Y.; and Zhang, M. 2024a. Graph Contrastive Learning Meets Graph Meta Learning: A Unified Method for Few-shot Node Tasks. In *Proceedings of Proceedings of the ACM on Web Conference (WWW-2024)*, 365–376.

Liu, J.; He, M.; Shang, X.; Shi, J.; Cui, B.; and Yin, H. 2024b. Bourne: Bootstrapped self-supervised learning framework for unified graph anomaly detection. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE-24)*, 2820–2833.

Liu, Y.; Ding, K.; Liu, H.; and Pan, S. 2023a. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM-2023)*, 339–347.

Liu, Y.; Ding, K.; Lu, Q.; Li, F.; Zhang, L. Y.; and Pan, S. 2024c. Towards self-interpretable graph-level anomaly detection. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS-24)*, volume 36.

Liu, Y.; Li, Z.; Pan, S.; Gong, C.; Zhou, C.; and Karypis, G. 2021. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6): 2378–2 392.

Liu, Z.; Zuo, W.; Zhang, D.; and Feng, X. 2023b. Rgse: Robust graph structure embedding for anomalous link detection. *IEEE Transactions on Big Data*, 9(5): 1420–1429.

Ma, R.; Pang, G.; Chen, L.; and Hengel, A. 2022. Deep graph-level anomaly detection by glocal knowledge distillation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM-2022)*, 704–714.

Ma, X.; Wu, J.; Xue, S.; Yang, J.; Zhou, C.; Sheng, Q. Z.; Xiong, H.; and Akoglu, L. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12012–12038.

Motie, S.; and Raahemi, B. 2023. Financial fraud detection using graph neural networks: A systematic review. *Expert Systems With Applications*, 240: 122156.

Mukherjee, A.; Liu, B.; and Glance, N. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of Proceedings of the ACM on Web Conference (WWW-2012)*, 191–200.

Nakano, M.; Takahashi, A.; and Takahashi, S. 2017. Generalized exponential moving average (EMA) model with particle filtering and anomaly detection. *Expert Systems with Applications*.

Ouyang, L.; Zhang, Y.; and Wang, Y. 2020. Unified graph embedding-based anomalous edge detection. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN-20)*, 1–8.

Qin, L.; Welleck, S.; Khashabi, D.; and Choi, Y. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS-22)*, volume 35, 9538–9551.

Qiu, C.; Kloft, M.; Mandt, S.; and Rudolph, M. 2022. Raising the bar in graph-level anomaly detection. In *In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*.

Shi, F.; Cao, Y.; Shang, Y.; Zhou, Y.; Zhou, C.; and Wu, J. 2022. H^2-fdetector: A gnn-based fraud detector with homophilic and heterophilic connections. In *Proceedings of the ACM on Web Conference (WWW-22)*, 1486–1494.

Steck, H.; Ekanadham, C.; and Kallus, N. 2024. Is cosine-similarity of embeddings really about similarity? In *Proceedings of the ACM on Web Conference (WWW-2024)*, 887–890.

Swersky, K.; Ranzato, M. A.; Buchman, D.; Freitas, N. D.; and Marlin, B. M. 2011. On autoencoders and score matching for energy based models. In *Proceedings of the International Conference on Machine Learning (ICML-11)*, 1201–1208.

Tang, J.; Hua, F.; Gao, Z.; Zhao, P.; and Li, J. 2024. Gadbench: Revisiting and benchmarking supervised graph anomaly detection. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS-24)*, volume 36.

Velickovic, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2019. Deep graph infomax. In *Proceedings of the International Conference on Learning Representations (ICLR-19)*, 4.

Wang, Q.; Pang, G.; Salehi, M.; Buntine, W.; and Leckie, C. 2023. Cross-domain graph anomaly detection via anomaly-aware contrastive alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*, volume 37, 4676–4684.

Wang, Y.; Liu, Y.; Shen, X.; Li, C.; Ding, K.; Miao, R.; Wang, Y.; and Pan, X., S. amd Wang. 2024. Unifying Unsupervised Graph-Level Anomaly Detection and Out-of-Distribution Detection: A Benchmark. arXiv:2406.15523.

Wu, C.; Wang, C.; Xu, J.; Liu, Z.; Zheng, K.; Wang, X.; Song, Y.; and Gai, K. 2023a. Graph Contrastive Learning with Augmentations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2023)*, 2721–2730.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023b. Energy-based Out-of-Distribution Detection for Graph Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR-23)*.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 1–28.

Yang, M.; Wang, H.; Wei, Z.; Wang, S.; and Wen, J. R. 2021. Efficient Algorithms for Personalized PageRank Computation: A Survey. *IEEE Transactions on Knowledge and Data Engineering*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph Contrastive Learning with Augmentations. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS-20)*, volume 33, 5812–5823.

Zheng, Y.; Jin, M.; Liu, Y.; Chi, L.; Phan, K. T.; and Chen, Y. P. P. 2021. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12220–12233.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2020. Deep graph contrastive representation learning. arXiv:2006.04131.