NORMAL-REFERENCED DISCRIMINATION FOR OPEN-SET GRAPH ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-set graph anomaly detection uses labeled normal nodes and a few labeled anomalies to detect both the labeled anomaly types and unseen types absent from the training labels. These labels help the detector learn normal and anomalous patterns, improving its detection accuracy. However, a detector fitted to the labeled types may mistake unseen anomalies for normal nodes. To address this problem, we propose Normal-Referenced Discrimination (NoRD), which trains a single anomaly score using both labeled and unlabeled nodes. The unlabeled nodes contain normal nodes and anomalies, including unseen types. Specifically, NoRD learns to distinguish these unlabeled nodes from labeled normal nodes alongside supervised anomaly detection, allowing unlabeled anomalies to contribute to training. Since both groups contain normal nodes, directly separating them would also encourage high anomaly scores for unlabeled normal nodes. Accordingly, NoRD first expresses each anomaly score relative to the labeled normal scores, providing a common reference for the comparison. It then applies a mixture link that accounts for normal nodes in both groups and reduces the pressure to increase already-low anomaly scores on unlabeled nodes. Experiments on seven public graphs demonstrate consistent improvements over competing methods in both overall and unseen-anomaly detection. Code is available at `https://anonymous.4open.science/r/NoRD/`.

1 INTRODUCTION

Graph anomaly detection (GAD) identifies anomalous nodes from their attributes and connections, with applications such as financial fraud and malicious-user detection (Ding et al., 2019; Tang et al., 2022; Wang et al., 2025). Supervised and semi-supervised GAD methods use labeled nodes to guide learning from node attributes and graph structure (Tang et al., 2022; 2023; Chen et al., 2024). These labels provide knowledge of normal and anomalous behavior, helping the detector learn relevant patterns and improve detection accuracy. However, the available anomaly labels may cover only part of the behavior that the detector must recognize. Open-set GAD addresses this setting: the training labels contain examples of *seen* anomaly classes, while evaluation also includes *unseen* classes with no labeled examples (Wang et al., 2025; Yu et al., 2026). The objective is therefore to assign higher scores to both seen and unseen anomalies than to normal nodes. Fitting the seen examples alone does not ensure that a graph neural network (GNN) learns this ordering for unseen classes.

To supplement limited anomaly labels, existing methods introduce additional training signals. Normal-supervised detectors learn from pseudo-anomalies or from diverse normal homophily patterns (Qiao et al., 2024; Ai et al., 2025). For open-set GAD, NSReg regularizes the shared representation by predicting structural relations among labeled normal nodes and between labeled normal and unlabeled nodes (Wang et al., 2025), and DEMO combines seen-anomaly synthesis with adaptive pseudo-labeling of unlabeled nodes (Yu et al., 2026). Beyond these training signals, the distributional difference between the unlabeled population and labeled normal nodes offers a further source of supervision for the anomaly score.

The unlabeled nodes contain both normal examples and anomalies, including types absent from the training labels. To learn from this mixture, we can use labeled normal nodes as a reference for normal behavior. When the normal nodes in both groups follow the same distribution, differences

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

between the two populations arise from the anomalous component. Learning to distinguish them can therefore provide a training signal that includes unseen anomaly types (Hido et al., 2008; Zhang & Tan, 2020). The challenge is to train the anomaly score with this population-level signal and the available anomaly labels. Applied directly to the raw score, population discrimination pushes unlabeled normal scores upward, whereas supervised detection favors low scores for normal nodes.

To address this conflict, we propose *Normal-Referenced Discrimination* (NoRD), which combines population discrimination with supervised detection to train a single graph anomaly score. First, *normal-reference normalization* expresses each score relative to the current labeled normal scores, allowing population discrimination to compare nodes against a reference for normal behavior. A *mixture link* then accounts for the normal component shared by the two populations and attenuates discrimination gradients on low-scoring unlabeled nodes. Experiments on seven public graphs demonstrate consistent gains in overall and unseen-anomaly detection.

Our contributions are as follows:

- We introduce NoRD for open-set GAD, combining supervised detection and normal-referenced discrimination to train a shared anomaly score from labeled nodes and the unlabeled population.
- We analyze how normal-reference normalization and the mixture link shape discrimination gradients: normalization couples score updates through the normal reference, while the link attenuates upward pressure on low-scoring unlabeled nodes. A density-ratio interpretation connects this objective to anomaly ranking under a shared normal distribution.
- Experiments on seven public graphs show that NoRD achieves the highest mean AUROC and AUPRC among the competing methods in both overall and unseen-anomaly detection. Ablations on Photo, Computers, and CS show that both components are needed to improve on supervised training alone.

2 Related Work

Graph anomaly detection. Unsupervised GAD methods use structural and edge-weight patterns (Akoglu et al., 2010), attributed neighborhoods (Perozzi & Akoglu, 2016), or joint attribute–structure models (Li et al., 2017; Peng et al., 2018); neural methods learn through reconstruction (Ding et al., 2019; Roy et al., 2024), contrastive learning (Jin et al., 2021), and local affinity (Qiao & Pang, 2023). Supervised GNNs exploit anomaly labels (Tang et al., 2022), with fraud detectors addressing camouflage (Dou et al., 2020), class imbalance (Liu et al., 2021), and class-dependent message passing (Zhuo et al., 2024). ConsisGAD extends supervision to unlabeled nodes through consistency training (Chen et al., 2024), and GADBench shows that tree ensembles with neighborhood-aggregated features are competitive (Tang et al., 2023). Normal-supervised methods generate pseudo-anomalies in GGAD (Qiao et al., 2024), model diverse homophily patterns in RHO (Ai et al., 2025), and align teacher scores under normality regularization in GraphNC (Zeng et al., 2026). Few-shot, cross-domain, and in-context variants transfer detectors across graphs (Ding et al., 2021; Wang et al., 2023; Liu et al., 2024). Open-set GAD lies between the label-free and the supervised settings: it uses the labeled anomaly classes without being confined to them.

Open-set anomaly detection. Open-set recognition combines known-category classification with unknown-category rejection (Scheirer et al., 2013; 2014; Bendale & Boult, 2016; Vaze et al., 2022). Open-set anomaly detection assigns seen and unseen anomaly types the same binary label; detectors for Euclidean data learn complementary abnormality patterns (Ding et al., 2022) or heterogeneous anomaly distributions (Zhu et al., 2024). Class-imbalanced node classification balances the labeled classes while assuming that all evaluation classes have training labels (Zhao et al., 2021; Park et al., 2022). Graph out-of-distribution detection instead identifies nodes that depart from the training distribution (Wu et al., 2023). Open-set GAD evaluates abnormality relative to a task-defined normal class, treating seen and unseen anomaly types alike (Wang et al., 2025). NSReg regularizes a shared graph representation by predicting connected and unconnected normal relations, together with unconnected relations between labeled normal and unlabeled nodes (Wang et al., 2025), and DEMO combines seen-anomaly mixup, energy-gradient weighting, and adaptive pseudo-labeling (Yu et al., 2026). NoRD uses membership in the labeled normal and unlabeled sets as an additional training

target. It combines discrimination between these sets with supervised anomaly detection through a shared anomaly score.

Learning from mixtures. With normal nodes as positives, our labeled normal nodes, labeled seen anomalies, and unlabeled nodes correspond to positives, biased negatives, and unlabeled data. Learning from these samples can follow two approaches. Two-sample classification relates class probabilities to density ratios (Elkan & Noto, 2008; Menon & Ong, 2016), enabling novelty and outlier detection by comparing normal data with an unlabeled mixture (Blanchard et al., 2010; Hido et al., 2008). This connection also underlies classification between mixtures with different class proportions (Metodiev et al., 2017). The exponential tilt mixture model provides an exponential class-density ratio for labeled and unlabeled samples that share class-conditional distributions (Zhang & Tan, 2020). In open-set GAD, however, labeled anomalies cover only seen subclasses and need not follow the distribution of the full anomalous component in the unlabeled population.

Alternatively, positive–unlabeled learning estimates classification risks from positive examples and an unlabeled mixture (du Plessis et al., 2015; Kiryo et al., 2017). PUbN additionally uses biased labeled negatives through risk weighting based on a separately estimated selection posterior (Hsieh et al., 2019); in open-set GAD, this bias arises from anomaly labels covering only a seen subclass. Deep SAD uses labeled normal, labeled anomalous, and unlabeled Euclidean samples in a hypersphere objective (Ruff et al., 2020). NoRD combines supervision on seen anomalies with two-sample discrimination through one anomaly score, using empirical reference normalization and a learned mixture link to shape the discrimination gradients.

3 METHOD

3.1 PROBLEM STATEMENT

We consider open-set GAD in a transductive setting. Let $G = (V, E, X)$ be an attributed graph with node set V, edge set E, and node attributes $X \in \mathbb{R}^{|V|\times d}$. The normal and anomalous node sets are V_N and V_A. We partition V_A into V_A^{seen}, whose classes are represented in the anomaly training labels, and V_A^{unseen}, whose classes have no labeled training examples. The labeled training nodes comprise a normal reference set $R \subseteq V_N$ and an anomaly set $S \subseteq V_A^{\text{seen}}$ from one seen class. The unlabeled set U contains the unlabeled nodes to be scored, including normal nodes and both seen and unseen anomalies, and is disjoint from R and S. Node attributes and graph structure are available throughout G; normal and anomaly training labels are available only on R and S.

For node i, the model combines a graph representation learner ψ_{θ_g} and a scalar scoring network h_{θ_s}:

$$f_i = f_\theta(G, i) = h_{\theta_s}\big(\psi_{\theta_g}(G, i)\big), \qquad \theta = (\theta_g, \theta_s). \tag{1}$$

The learning objective is to rank unlabeled anomalies $U \cap V_A$, including unseen classes, above unlabeled normal nodes $U \cap V_N$.

3.2 OVERVIEW OF NORD

NoRD trains one anomaly score with two objectives (Figure 1). Supervised detection separates the labeled seen anomalies from labeled normal nodes. Normal-referenced discrimination separates the unlabeled set U from the reference R: each node in $R \cup U$ carries a membership label, $d_i = 0$ for $i \in R$ and $d_i = 1$ for $i \in U$, observed without any anomaly annotation. Applied to the raw score, this discrimination objective would ask every unlabeled node, most of which are normal, to lie above the reference. NoRD therefore expresses scores relative to the reference and passes them through a mixture link that retains a shared normal component before computing the discrimination loss. Both losses update the scoring network and the graph encoder.

Let $\ell(t, y) = \log(1 + e^t) - yt$ denote binary cross-entropy for logit t and label y, and let $B_R \subseteq R$ be the normal subset sampled for supervised detection. NoRD minimizes

$$\min_{\theta,\eta} \quad \mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{seen}} + \lambda\mathcal{L}_{\text{dis}}, \tag{2}$$

$$\mathcal{L}_{\text{seen}} = \frac{\sum_{i\in B_R} \ell(f_i, 0) + \sum_{i\in S} \ell(f_i, 1)}{|B_R| + |S|}, \tag{3}$$

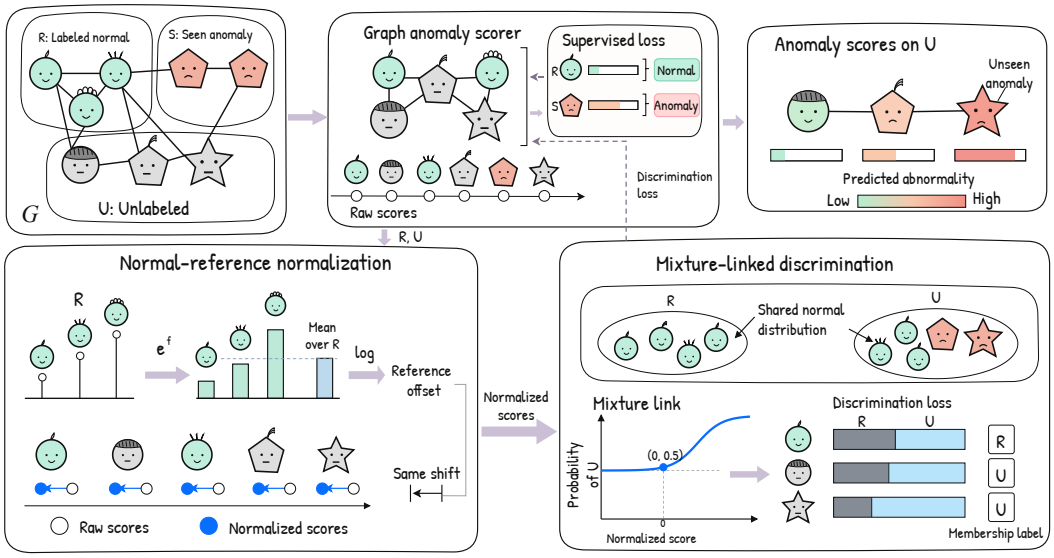

Figure 1: Overview of NoRD. The shared scorer supports supervised detection (top) and discrimination between labeled normal nodes R and unlabeled nodes U (bottom). Normalization subtracts the log-mean-exponential score over R; the mixture link maps normalized scores to membership probabilities, shown at initialization ($\eta = -2$).

where $\lambda \in [0, 1]$ weights the two objectives, $\mathcal{L}_{\text{dis}}$ is the discrimination loss of Equation (8), and η is the scalar parameter of its mixture link. At inference, the anomaly score is $s_i = \sigma(f_\theta(G, i))$ with σ the sigmoid function; higher values indicate greater predicted abnormality.

3.3 Discrimination and Anomaly Ranking

The following proposition applies the standard density-ratio relationship for two-sample classification (Menon & Ong, 2016) to a normal reference and an unlabeled mixture (Hido et al., 2008; Zhang & Tan, 2020; Metodiev et al., 2017). In a fixed representation space, let z denote a node representation and $\alpha \in (0, 1)$ the anomaly proportion in the unlabeled population, and write $p_R(z)$, $p_U(z)$, $p_N(z)$, and $p_A(z)$ for the reference, unlabeled, normal, and aggregate anomaly densities.

Proposition 1 (Membership logit under a shared normal distribution) *Suppose $p_R = p_N$ and $p_U = (1 - \alpha)p_N + \alpha p_A$. Consider the population binary cross-entropy for distinguishing the two populations with equal total weight assigned to each. Among unrestricted logit functions, its pointwise minimizer where $p_N(z) > 0$ is*

$$a^*(z) = \log \frac{p_U(z)}{p_R(z)} = \log \left[(1 - \alpha) + \alpha \frac{p_A(z)}{p_N(z)} \right], \tag{4}$$

which is strictly increasing in the aggregate anomaly-to-normal density ratio $p_A(z)/p_N(z)$.

Appendix A gives the proof and the connection to the normalized parameterization. If unlabeled anomalies comprise K classes with densities p_{A_k} and mixture weights $\omega_k \geq 0$, $\sum_{k=1}^{K} \omega_k = 1$, then

$$\frac{p_U(z)}{p_N(z)} = (1 - \alpha) + \alpha \sum_{k=1}^{K} \omega_k \frac{p_{A_k}(z)}{p_N(z)}, \qquad p_N(z) > 0. \tag{5}$$

Each class enters the optimal logit through its prevalence ω_k and its density relative to normal nodes, whether or not it has labels. The unlabeled population thus supplies a ranking target that includes classes absent from the anomaly labels, while the supervised loss directly fits the seen class.

This distinction also affects the target assigned to normal nodes. Under Proposition 1, a region with $p_N(z) > 0$ and $p_A(z) = 0$ has the raw-score discrimination optimum $f^*(z) = \log(1 - \alpha)$,

corresponding to an anomaly probability of $(1-\alpha)/(2-\alpha)$. Supervised detection instead favors probabilities approaching zero in this region. This difference motivates learning the discrimination signal relative to the normal reference while allowing supervision to determine the absolute score level.

3.4 NORMAL-REFERENCE NORMALIZATION

To parameterize the ratio in Equation (4), NoRD uses an exponential score normalized over the reference set. For the current scores, the reference offset b_R and the normalized score u_i are

$$b_R = \log\left(\frac{1}{|R|}\sum_{j\in R} e^{f_j}\right), \qquad u_i = f_i - b_R, \tag{6}$$

so that $\frac{1}{|R|}\sum_{j\in R} e^{u_j} = 1$. This matches the unit mean of a density ratio under its denominator distribution when p_A is absolutely continuous with respect to p_N; the constraint fixes the aggregate reference scale while individual normal scores remain free to differ. Subtracting a common offset preserves node ordering and cancels common score shifts, so discrimination acts on relative scores while $\mathcal{L}_{\text{seen}}$ sets their absolute level. The reference also participates in backpropagation: each discrimination term contributes to reference-node updates through b_R, with larger weights for higher-scoring reference nodes.

3.5 MIXTURE-LINKED DISCRIMINATION

To account for the normal component shared by R and U, we adopt the mixture form of exponential tilt mixture models (Zhang & Tan, 2020). With e^{u_i} parameterizing the anomaly-to-normal density ratio in Equation (4), the membership logit a_i and membership probability q_i under equal population weighting are

$$\rho = \sigma(\eta), \qquad a_i = g_\eta(u_i) = \log\big[(1-\rho) + \rho e^{u_i}\big], \qquad q_i = \sigma(a_i). \tag{7}$$

The constant term represents the shared normal nodes, and the exponential term grows with the normalized score. Several mixture representations yield the same membership distributions, so discrimination alone does not identify the anomaly proportion; we therefore learn ρ jointly as a link parameter rather than as an estimate of α. For any fixed $\rho \in (0,1)$, $g_\eta(0) = 0$ and $g_\eta(u) \to \log(1-\rho)$ as $u \to -\infty$: a node scored at the reference level has equal membership probabilities, and the membership logit approaches a finite lower bound at low normalized scores. The discrimination loss is

$$\mathcal{L}_{\text{dis}} = \frac{1}{2|R|}\sum_{i\in R}\ell(a_i, 0) + \frac{1}{2|U|}\sum_{i\in U}\ell(a_i, 1), \tag{8}$$

which assigns each population total weight one half, matching the population risk in Proposition 1 and making the loss independent of the size ratio $|U|/|R|$. The link adds one learned scalar, η, and is used only during training; Algorithm 1 gives the full procedure.

The link is strictly increasing, $g'_\eta(u) = \sigma(\eta + u) > 0$, and preserves the ordering of fixed scores. It modifies the training gradients as follows:

$$\frac{\partial \ell(a_i, d_i)}{\partial u_i} = (q_i - d_i)\,\sigma(\eta + u_i). \tag{9}$$

For an unlabeled node ($d_i = 1$), this derivative favors increasing u_i, but its magnitude tends to zero at both score extremes. In particular, $\sigma(\eta+u_i)$ attenuates the upward pressure on low-scoring nodes, which receive the largest gradient magnitudes under an identity link. At the reference level $u_i = 0$, the magnitude is $\rho/2$, so the learned link weight controls the discrimination pressure at that level.

With reference normalization, the score gradients of $\mathcal{L}_{\text{dis}}$ sum to zero, consistent with invariance to a common score shift; Appendix A gives the full derivatives and the mixture parameter interpretation. In the joint objective, supervision lowers labeled normal scores and raises labeled anomaly scores, while discrimination learns from the unlabeled population through relative scores, with attenuated upward pressure on low-scoring nodes.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate NoRD on seven public graphs. Photo, Computers, and CS (Shchur et al., 2018) use semantic classes to define anomaly subclasses; ogbn-arxiv uses paper topics and ogbn-proteins uses species (Hu et al., 2020). Classes containing less than 5% of nodes are anomalous on Photo, Computers, CS, and ogbn-proteins, and ogbn-arxiv uses classes containing 3%–5% (Wang et al., 2025); the remaining classes are normal. For binary-labeled Yelp (Rayana & Akoglu, 2015) and T-Finance (Tang et al., 2022), we retain the normal/anomaly labels and partition anomalies into two subclasses using k-means on Deep Graph Infomax (DGI) (Veličković et al., 2019) representations. We also evaluate on a large-scale internal payment graph constructed from three months of payment data, referred to as Internal Payment. Appendix B.1 provides dataset statistics and preprocessing details; Appendix C.4 reports the internal-graph results.

Baselines. We compare with eleven GAD methods and one graph-adapted PUbN baseline. TAM (Qiao & Pang, 2023) is unsupervised; GGAD (Qiao et al., 2024), RHO (Ai et al., 2025), and GraphNC (Zeng et al., 2026) use normal labels. BWGNN (Tang et al., 2022), ConsisGAD (Chen et al., 2024), SpaceGNN (Dong et al., 2025), PMP (Zhuo et al., 2024), and XGBGraph (Tang et al., 2023) use normal and seen-anomaly labels. NSReg (Wang et al., 2025) and DEMO (Yu et al., 2026) specifically address open-set GAD with both label types. GraphSAGE+PUbN (Hsieh et al., 2019) uses two-stage biased-negative risk learning with a GraphSAGE encoder and known priors matched to the normal population and biased-negative sampling mechanism. Appendix B.3 describes the baselines.

Implementation Details. NoRD uses two GraphSAGE (Hamilton et al., 2017) layers with 64 hidden units each. We train for 200 epochs on Photo, Computers, and CS and 300 epochs on the remaining datasets, fixing $\lambda = 0.5$ and using $k = 50$ labeled seen anomalies by default. Appendix B.2 provides hyperparameter selection and additional implementation details.

Evaluation protocol. Each anomaly subclass serves as seen in turn, with all other anomaly subclasses treated as unseen (Wang et al., 2025; Yu et al., 2026). Each training set contains 50 seen anomalies and 5% of normal nodes; validation uses another 30 anomalies from the same subclass and 1% of normal nodes. We perform five independent random runs on every dataset and average over the seen-class settings, reporting mean $\pm$ standard deviation (Appendix B.2). All methods use the same partitions and metric definitions, and every baseline is run under this protocol rather than quoted from its paper. We report the area under the receiver operating characteristic curve (AUROC) and the precision–recall curve (AUPRC) under *All* and *Unseen*: All evaluates seen and unseen anomalies jointly against unlabeled normal nodes, while Unseen excludes seen anomalies and retains the same normal nodes.

4.2 EMPIRICAL RESULTS

GAD Performance. NoRD achieves the highest mean AUROC and AUPRC on Photo, Computers, CS, and Yelp under both All and Unseen (Table 1). NSReg and DEMO are the two strongest competitors when performance is averaged across these graphs. Compared with the best and second-best competing methods for each metric, NoRD improves average AUROC by 4.3% and 4.5%, and AUPRC by 5.5% and 7.2%, respectively, under All. Its advantage remains after seen anomalies are excluded from evaluation, with relative gains of 8.2%–8.8% in average AUPRC over these two methods. These results show that NoRD's overall advantage extends to anomaly types absent from the training labels.

Large-scale Evaluation. Table 2 reports results on three large graphs, T-Finance, ogbn-arxiv, and ogbn-proteins; the complete comparison appears in Appendix C.1. NoRD maintains its advantage across all three graphs, achieving the highest mean AUROC and AUPRC for both overall and unseen-anomaly detection. In particular, its overall AUROC and AUPRC, averaged across these graphs, exceed those of the strongest baseline by 2.4% and 5.2%, respectively. Together, these

Table 1: AUROC and AUPRC under All and Unseen, in percent (mean ± standard deviation). Bold and underlined entries denote the highest and second-highest means, respectively.

Method	AUROC				AUPRC			
	Photo	Computers	CS	Yelp	Photo	Computers	CS	Yelp
				All				
TAM	59.67±0.66	46.56±0.53	56.31±0.44	30.47±0.25	10.81±0.16	14.43±0.18	28.68±0.33	13.26±0.03
GGAD	55.34±4.03	50.38±4.97	65.73±2.11	32.63±2.77	9.73±1.20	14.89±2.07	39.70±4.41	10.74±0.69
RHO	53.73±2.07	37.32±1.58	73.69±4.30	54.97±2.06	8.85±0.44	11.80±0.42	39.81±3.96	18.53±1.01
GraphNC	55.73±4.55	50.49±5.04	65.67±2.13	67.66±0.66	9.82±1.43	14.93±2.08	39.55±4.41	23.77±0.69
BWGNN	75.63±1.02	60.78±5.79	85.30±3.99	63.00±2.92	43.14±8.18	33.42±3.82	74.73±2.40	28.25±3.42
ConsisGAD	70.49±2.88	60.90±8.82	70.29±5.51	56.28±5.38	41.45±6.92	31.87±4.17	59.89±2.17	30.66±2.28
SpaceGNN	78.31±2.01	72.15±2.18	85.59±6.70	74.98±4.60	49.58±3.64	47.69±1.42	77.58±2.79	37.33±3.01
PMP	75.68±2.49	70.46±4.27	83.76±3.38	71.63±4.16	47.34±3.90	44.27±6.73	73.45±3.58	29.97±3.86
XGBGraph	70.13±1.47	72.29±4.01	76.31±2.53	65.22±4.05	49.61±6.32	46.91±6.64	56.24±1.49	32.98±4.42
GraphSAGE+PUbN	83.02±3.32	69.42±2.67	93.36±1.26	66.75±3.52	54.07±1.77	40.08±2.46	88.08±2.40	30.26±1.78
NSReg	83.18±1.87	78.97±2.13	95.62±0.88	71.59±1.70	53.85±1.44	53.59±1.99	90.63±2.97	36.25±1.53
DEMO	86.04±1.29	80.34±1.30	92.98±1.12	70.94±2.01	55.21±1.81	55.26±0.79	88.45±3.47	31.64±1.77
NoRD	**88.05±2.18**	**82.29±2.50**	**97.23±0.54**	**76.77±1.65**	**58.41±3.82**	**58.19±1.18**	**92.73±1.97**	**37.89±1.95**
				Unseen				
TAM	59.45±4.56	46.57±2.47	56.38±1.96	46.18±1.02	6.66±0.18	12.33±0.93	26.49±1.28	7.69±0.40
GGAD	55.12±5.10	50.28±5.85	65.69±2.94	47.28±1.16	5.75±0.95	12.77±2.27	37.16±5.74	6.44±0.58
RHO	53.46±4.70	37.23±2.28	73.81±4.89	54.55±2.32	5.24±0.81	10.06±0.93	37.33±4.51	9.68±0.49
GraphNC	55.52±5.83	50.38±5.93	65.62±2.95	46.77±2.20	5.84±1.19	12.80±2.28	37.01±5.73	12.70±0.82
BWGNN	60.26±5.06	53.56±7.24	83.67±4.04	48.76±1.02	6.87±1.65	17.61±5.75	69.19±4.57	7.67±1.08
ConsisGAD	49.86±2.49	53.68±9.79	67.28±7.76	58.97±1.48	5.11±0.48	16.44±6.45	53.15±5.54	7.98±0.19
SpaceGNN	62.42±5.28	66.48±5.52	84.04±6.99	38.57±3.11	7.47±2.23	32.18±4.32	72.63±4.13	6.19±0.25
PMP	57.84±5.64	64.71±9.18	82.12±6.13	55.52±3.36	6.98±2.48	28.63±2.10	67.90±4.78	10.36±1.57
XGBGraph	47.60±2.70	67.03±2.35	73.52±5.63	43.17±3.74	8.36±2.61	32.08±6.60	45.60±3.65	6.74±1.65
GraphSAGE+PUbN	70.37±6.37	63.70±3.10	92.72±1.39	47.14±3.51	12.78±2.00	23.61±2.26	85.89±2.82	7.61±0.79
NSReg	82.30±2.28	76.77±1.24	95.30±0.92	69.04±2.58	23.52±1.76	39.57±4.37	86.17±4.29	20.75±2.48
DEMO	81.64±2.67	79.50±2.78	95.10±1.59	70.18±2.98	20.98±0.48	43.60±2.78	88.10±3.21	18.33±3.89
NoRD	**82.69±2.49**	**81.63±3.40**	**96.97±0.61**	**71.90±3.05**	**27.09±1.41**	**44.98±3.13**	**90.81±2.67**	**22.07±2.62**

Table 2: AUROC and AUPRC under All and Unseen, in percent (mean ± standard deviation). Bold and underlined entries denote the highest and second-highest means, respectively. Results for all baselines are provided in Appendix Table 5.

Dataset	Metric	GGAD	SpaceGNN	XGBGraph	GraphSAGE+PUbN	NSReg	DEMO	NoRD
					All			
T-Finance	AUROC	71.94±8.97	92.94±0.94	88.48±1.36	90.75±0.61	90.03±1.62	88.48±0.57	**93.26±0.67**
	AUPRC	15.67±2.80	76.94±1.11	73.40±2.83	72.86±3.30	73.70±1.74	74.41±1.66	**78.05±2.43**
ogbn-arxiv	AUROC	50.62±1.16	60.41±4.84	60.92±3.41	60.30±1.24	62.43±2.22	60.45±2.77	**63.25±2.50**
	AUPRC	16.05±0.49	27.69±4.31	27.80±4.50	28.79±0.69	29.05±4.03	28.50±4.32	**29.57±4.69**
ogbn-proteins	AUROC	34.39±0.68	74.73±3.60	80.17±3.51	74.42±1.71	81.95±7.31	76.88±6.87	**83.59±6.62**
	AUPRC	5.81±0.03	57.98±6.94	61.08±7.17	60.32±1.15	64.11±4.46	62.53±2.42	**67.96±6.06**
					Unseen			
T-Finance	AUROC	73.07±3.67	85.28±4.49	90.34±8.33	92.74±0.40	93.83±1.36	91.40±1.03	**94.77±1.02**
	AUPRC	13.35±8.98	61.53±1.78	60.39±1.07	72.88±3.24	77.29±2.09	80.55±2.31	**81.84±2.96**
ogbn-arxiv	AUROC	50.65±2.86	51.95±1.99	52.36±1.01	51.33±1.69	54.56±1.45	54.98±0.82	**56.21±0.46**
	AUPRC	12.63±0.91	13.86±2.25	13.94±2.55	14.61±0.68	14.49±0.17	14.83±0.31	**14.97±0.76**
ogbn-proteins	AUROC	29.95±1.11	45.59±2.32	49.08±1.23	63.20±1.72	74.50±0.48	72.16±1.91	**74.91±0.45**
	AUPRC	3.29±0.69	11.61±0.49	10.03±0.57	41.18±0.37	41.44±0.65	43.06±0.53	**43.94±0.56**

results demonstrate that NoRD maintains effective open-set detection on large graphs, improving overall detection while recognizing anomaly classes absent from the training labels.

Data Efficiency. To further evaluate how detection performance changes with anomaly-label availability, we vary the number of labeled training anomalies $k \in \{5, 10, 20, 50, 100\}$ while fixing the normal and validation budgets. Figure 2 presents unseen-anomaly detection on Photo, Computers, and CS; complete results appear in Appendix C.2. NoRD already leads unseen AUPRC on Photo and CS with only five labeled anomalies and maintains this advantage across all tested budgets. On Computers, it overtakes DEMO from twenty labels onward. Although additional labels improve detection over much of the range, increasing the budget from fifty to one hundred lowers unseen AUROC on Computers and CS. Increasing supervision within the same seen class therefore does not guarantee better detection of unseen classes, while NoRD remains effective with few labeled training anomalies.

Ablation Study. To assess the contribution of each part of NoRD, we compare three ablations on Photo, Computers, and CS: (1) *w/o discrimination objective* removes population discrimination and uses supervised training alone; (2) *w/o normalization* applies the mixture link directly to the raw score; and (3) *w/o mixture link* uses the normalized score directly as the membership logit. Table 3

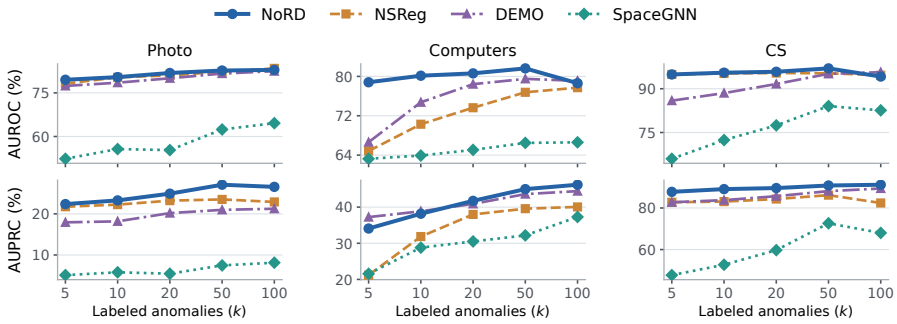

Figure 2: Unseen AUROC and AUPRC versus the number of labeled training anomalies on Photo, Computers, and CS. Points show means; budget positions are equally spaced.

shows that removing discrimination reduces all four measures on every graph, with larger losses under Unseen than under All. On Photo, the relative reductions in AUROC are 18.2% on unseen anomalies and 8.6% overall. This greater loss on unseen anomalies highlights the contribution of population discrimination beyond fitting the labeled class.

Table 3: Ablation results under All and Unseen, in percent (mean ± standard deviation). The entry with the highest mean is boldfaced.

Scope	Ablation / method	Photo		Computers		CS	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
All	NoRD	**88.05±2.18**	**58.41±3.82**	**82.29±2.50**	**58.19±1.18**	**97.23±0.54**	**92.73±1.97**
	w/o discrimination objective	80.48±0.78	50.44±0.13	70.87±4.10	41.74±1.48	94.96±0.61	86.09±1.39
	w/o normalization	78.62±1.58	48.03±1.12	68.52±3.59	39.97±1.07	83.90±0.82	80.27±1.96
	w/o mixture link	75.31±1.79	44.17±4.44	66.19±1.93	40.05±1.31	81.60±0.87	77.50±1.86
Unseen	NoRD	**82.69±2.49**	**27.09±1.41**	**81.63±3.40**	**44.98±3.13**	**96.97±0.61**	**90.81±2.67**
	w/o discrimination objective	67.62±1.48	10.76±0.26	67.13±3.93	21.68±3.38	94.25±0.71	83.48±1.91
	w/o normalization	60.58±3.03	4.75±0.47	60.13±4.36	13.54±2.58	85.18±0.96	76.26±2.75
	w/o mixture link	62.81±3.41	7.17±1.72	61.49±2.33	16.84±1.58	83.65±1.01	71.99±2.71

Retaining discrimination while removing either normalization or the mixture link causes an even larger loss: both variants fall below supervised training alone on every reported metric. This pattern is consistent with the gradient analysis: normalization makes discrimination invariant to common score shifts, and the mixture link weakens upward pressure on low-scoring unlabeled nodes. The improvement therefore depends on how the two objectives are combined, with both components needed to turn population discrimination into a useful training signal on these graphs.

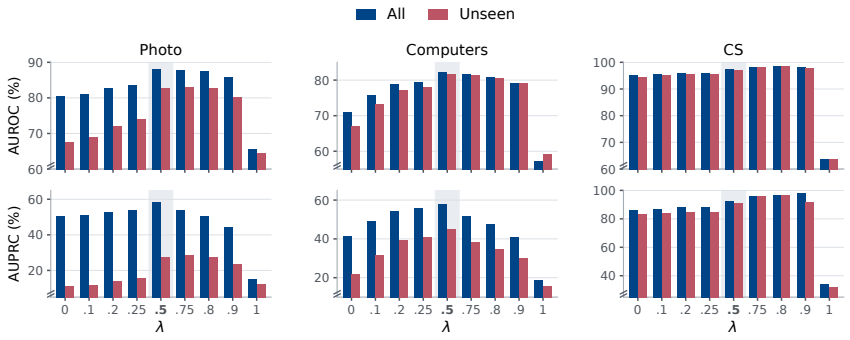

Figure 3: All and Unseen performance across nine discrimination weights on Photo, Computers, and CS. Bars show means (%).

Sensitivity Analysis. To examine how the two training objectives interact, we vary the discrimination weight λ from zero to one, with the endpoints using supervision or discrimination alone. Figure 3 presents results on Photo, Computers, and CS, and Appendix C.3 provides the remaining

four graphs. At the default weight of 0.5, joint training improves all four measures over supervision alone on every graph and outperforms discrimination alone on six graphs. The exception is ogbn-arxiv, where discrimination alone gives the strongest unseen results, although its overall AUPRC is lower than at $\lambda = 0.9$. The benefit of combining the objectives is particularly clear on ogbn-proteins: unseen AUROC rises from 5.5% with supervision alone to 59.9% with discrimination alone and 74.9% with both. The gain over both endpoints shows that seen-class supervision can help identify unseen anomalies when combined with population discrimination, even when its standalone ranking is poor.

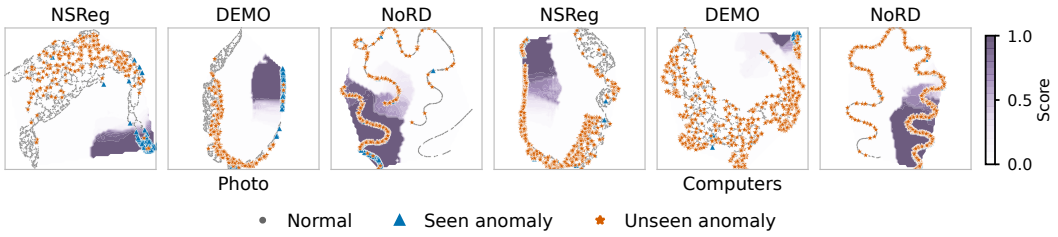

Figure 4: Node representations and reconstructed score contours on Photo and Computers. Gray dots, blue triangles, and orange stars denote normal, seen-anomaly, and unseen-anomaly nodes. Darker backgrounds indicate higher scores on reconstructed representations; each model uses its own projection.

Visualization Results. Figure 4 overlays reconstructed score contours on node representations from Photo and Computers. The high-score regions of NSReg and DEMO are concentrated near seen anomalies, whereas NoRD's also extend to groups of unseen anomalies, with some overlap with normal nodes remaining. Together with the detection results, this qualitative pattern supports NoRD's ability to recognize anomalies beyond the types represented in the training labels.

5 Conclusion

This paper presents NoRD for transductive open-set graph anomaly detection. It jointly learns an anomaly score through supervised detection and discrimination between a labeled normal reference and the unlabeled set. Normal-reference normalization and a mixture link adapt this population objective to anomaly scoring by setting the reference level and controlling discrimination gradients. Experiments on seven public graphs demonstrate higher mean AUROC and AUPRC than twelve competing methods under both overall and unseen-anomaly evaluation. Ablations on Photo, Computers, and CS further show that both components are needed to improve upon supervised training alone. Together, these findings demonstrate the value of directly learning from the unlabeled population to detect anomaly types that training labels do not cover.

AI use statement

In this work, we used generative AI tools to assist with text editing, grammar polishing, and code debugging. The manuscript and the code were first written by the authors and then refined with this assistance.

Reproducibility statement

The training procedure is provided in Algorithm 1. Dataset descriptions, evaluation protocols, hyperparameter settings, and implementation details are provided in Appendix B. The code is available at `https://anonymous.4open.science/r/NoRD/`.

Ethics statement

Photo, Computers, CS, Yelp, T-Finance, ogbn-arxiv, and ogbn-proteins are publicly available datasets provided by third parties for academic research. Internal Payment is an internal dataset.

This study did not involve the collection of new data from human participants for any of these datasets.

REFERENCES

Guoguo Ai, Hezhe Qiao, Hui Yan, and Guansong Pang. Semi-supervised graph anomaly detection via robust homophily learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 133988–134013, 2025.

Leman Akoglu, Mary McGlohon, and Christos Faloutsos. oddball: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining*, pp. 410–421, 2010.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Nan Chen, Zemin Liu, Bryan Hooi, Bingsheng He, Rizal Fathony, Jun Hu, and Jia Chen. Consistency training with learnable data augmentation for graph anomaly detection with limited supervision. In *International Conference on Learning Representations*, 2024.

Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7388–7398, 2022.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602, 2019.

Kaize Ding, Qinghai Zhou, Hanghang Tong, and Huan Liu. Few-shot network anomaly detection via cross-network meta-learning. In *Proceedings of the Web Conference 2021*, pp. 2448–2456, 2021.

Xiangyu Dong, Xingyi Zhang, Lei Chen, Mingxuan Yuan, and Sibo Wang. SpaceGNN: Multi-space graph neural network for node anomaly detection with extremely limited labels. In *International Conference on Learning Representations*, 2025.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 315–324, 2020.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1386–1394, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 223–232, 2008.

Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2820–2829, 2019.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. ANEMONE: Graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 3122–3126, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jundong Li, Harsh Dani, Xia Hu, and Huan Liu. Radar: Residual analysis for anomaly detection in attributed networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2152–2158, 2017.

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 the Web Conference 2021*, pp. 3168–3177, 2021.

Yixin Liu, Shiyuan Li, Yu Zheng, Qingfeng Chen, Chengqi Zhang, and Shirui Pan. ARC: A generalist graph anomaly detector with in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50772–50804, 2024.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 304–313, 2016.

Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: learning from mixed samples in high energy physics. *Journal of High Energy Physics*, 2017(10):174, 2017.

Joonhyung Park, Jaeyun Song, and Eunho Yang. GraphENS: Neighbor-aware ego network synthesis for class-imbalanced node classification. In *International Conference on Learning Representations*, 2022.

Zhen Peng, Minnan Luo, Jundong Li, Huan Liu, and Qinghua Zheng. ANOMALOUS: A joint modeling approach for anomaly detection on attributed networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3513–3519, 2018.

Bryan Perozzi and Leman Akoglu. Scalable anomaly ranking of attributed neighborhoods. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 207–215, 2016.

Hezhe Qiao and Guansong Pang. Truncated affinity maximization: One-class homophily modeling for graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49490–49512, 2023.

Hezhe Qiao, Qingsong Wen, Xiaoli Li, Ee-Peng Lim, and Guansong Pang. Generative semi-supervised graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4660–4688, 2024.

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

Amit Roy, Juan Shu, Jia Li, Carl Yang, Olivier Elshocht, Jeroen Smeets, and Pan Li. GAD-NR: Graph anomaly detection via neighborhood reconstruction. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 576–585, 2024.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013.

Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 21076–21089, 2022.

Jianheng Tang, Fengrui Hua, Ziqi Gao, Peilin Zhao, and Jia Li. GADBench: Revisiting and benchmarking supervised graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29628–29653, 2023.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Qizhou Wang, Guansong Pang, Mahsa Salehi, Wray Buntine, and Christopher Leckie. Cross-domain graph anomaly detection via anomaly-aware contrastive alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4676–4684, 2023.

Qizhou Wang, Guansong Pang, Mahsa Salehi, Xiaokun Xia, and Christopher Leckie. Open-set graph anomaly detection via normal structure regularisation. In *International Conference on Learning Representations*, 2025.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *International Conference on Learning Representations*, 2023.

Caiyang Yu, Wei Ju, Haixin Wang, Yifan Wang, and Ziyue Qiao. Dynamic multi-sample mixup with gradient exploration for open-set graph anomaly detection. In *International Conference on Learning Representations*, 2026.

Guolei Zeng, Hezhe Qiao, Guoguo Ai, Jinsong Guo, and Guansong Pang. Normality calibration in semi-supervised graph anomaly detection. In *International Conference on Machine Learning*, 2026.

Xinwei Zhang and Zhiqiang Tan. Semi-supervised logistic learning based on exponential tilt mixture models. *Stat*, 9(1):e312, 2020.

Tianxiang Zhao, Xiang Zhang, and Suhang Wang. GraphSMOTE: Imbalanced node classification on graphs with graph neural networks. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 833–841, 2021.

Jiawen Zhu, Choubo Ding, Yu Tian, and Guansong Pang. Anomaly heterogeneity learning for open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17616–17626, 2024.

Wei Zhuo, Zemin Liu, Bryan Hooi, Bingsheng He, Guang Tan, Rizal Fathony, and Jia Chen. Partitioning message passing for graph fraud detection. In *International Conference on Learning Representations*, 2024.

A Proof and interpretation of Proposition 1

Proof. For an unrestricted membership logit $a(z)$, the population counterpart of Equation (8) is

$$\mathcal{J}(a) = \tfrac{1}{2}\mathbb{E}_{p_R}\ell(a(z),0) + \tfrac{1}{2}\mathbb{E}_{p_U}\ell(a(z),1). \tag{10}$$

Where $p_N(z) > 0$, the assumptions imply that both population densities are positive. The pointwise risk is strictly convex in the logit, and its first-order condition is

$$\tfrac{1}{2}p_R(z)\sigma(a(z)) + \tfrac{1}{2}p_U(z)(\sigma(a(z)) - 1) = 0. \tag{11}$$

Solving gives $\sigma(a^*(z)) = p_U(z)/(p_R(z) + p_U(z))$ and $a^*(z) = \log[p_U(z)/p_R(z)]$. Substituting $p_R = p_N$ and $p_U = (1-\alpha)p_N + \alpha p_A$ yields Equation (4). For $r = p_A/p_N \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}r}\log[(1-\alpha) + \alpha r] = \frac{\alpha}{1-\alpha+\alpha r} > 0, \tag{12}$$

which establishes the stated ordering. □

Reference distribution shift. Allowing $p_R \neq p_N$, define $\delta(z) = \log[p_R(z)/p_N(z)]$ where both densities are positive. In the same fixed representation space, the mixture model gives the unrestricted population-optimal logit

$$a^*(z) = \log\left[(1-\alpha) + \alpha\frac{p_A(z)}{p_N(z)}\right] - \delta(z).$$

Reference shift therefore enters the ranking through $-\delta(z)$. If $|\delta(z)| \leq \varepsilon$ on this support, pairs whose logit gap under the shared normal distribution exceeds 2ε retain their ordering: the change in their gap is bounded by $|\delta(z_1) - \delta(z_2)| \leq 2\varepsilon$. Our benchmarks sample normal labels uniformly; biased normal labels were not evaluated.

Normal-reference parameterization. Under Proposition 1, assume additionally that p_A is absolutely continuous with respect to p_N. Then $\mathbb{E}_{p_N}[p_A/p_N] = 1$, whose empirical counterpart is Equation (6). Setting $\rho = \alpha$ and $u = \log(p_A/p_N)$ recovers the population-optimal membership logit, with zero ratios represented by $u \to -\infty$. NoRD combines this parameterization with seen-anomaly supervision.

Interpretation of the mixture parameter. The same membership ratio can admit multiple mixture representations. Let $t(z) = p_U(z)/p_N(z)$ and choose any $\widetilde{\alpha} \in (0,1)$ satisfying $t(z) \geq 1 - \widetilde{\alpha}$ for p_N-almost every z. Under the absolute-continuity assumption above, the function

$$\widetilde{r}(z) = \frac{t(z) - (1-\widetilde{\alpha})}{\widetilde{\alpha}}$$

is nonnegative, has unit expectation under p_N, and satisfies $t(z) = (1-\widetilde{\alpha}) + \widetilde{\alpha}\widetilde{r}(z)$. Thus, discrimination alone does not uniquely identify the anomaly proportion, and ρ serves as a learned link parameter.

Joint discrimination gradients. For $i \in R \cup U$, let $w_i = 1/(2|R|)$ on R and $w_i = 1/(2|U|)$ on U. Denote the weighted direct derivative by c_i and the reference-offset derivative $\partial b_R/\partial f_j$ by β_j:

$$c_i = w_i(q_i - d_i)\sigma(\eta + u_i), \qquad \beta_j = \frac{e^{f_j}}{\sum_{k\in R} e^{f_k}} \quad (j \in R). \tag{13}$$

Since $u_i = f_i - b_R$, the chain rule gives, with η held fixed,

$$\frac{\partial \mathcal{L}_{\mathrm{dis}}}{\partial f_j} = \begin{cases} c_j - \beta_j \displaystyle\sum_{i\in R\cup U} c_i, & j \in R, \\ c_j, & j \in U. \end{cases} \tag{14}$$

The link modulates each direct derivative, while b_R couples every discrimination term to the reference scores. Since $\sum_{j\in R}\beta_j = 1$, the gradients sum to zero, consistent with invariance to a common score shift.

B EXPERIMENTAL DETAILS

B.1 DATASETS

Table 4 summarizes the seven public datasets and Internal Payment. The public graphs cover co-purchase, co-authorship, review, transaction, citation, and biological association networks.

- Photo and Computers (Shchur et al., 2018) are Amazon co-purchase graphs. Nodes represent products, edges connect products purchased together, attributes encode product reviews, and labels identify product categories.
- CS (Shchur et al., 2018) is a co-authorship graph with researchers as nodes, co-authorships as edges, and research fields as class labels.
- Yelp (Rayana & Akoglu, 2015) contains spam and legitimate reviews. We use the Review–User–Review (RUR) edge view (Wang et al., 2025).
- T-Finance (Tang et al., 2022) is a transaction graph with accounts as nodes, transactions as edges, and normal/anomalous account labels.
- ogbn-arxiv (Hu et al., 2020) is a citation graph of computer science papers, whose topics define the semantic classes.
- ogbn-proteins (Hu et al., 2020) contains biological associations between proteins. Species define the classes for open-set evaluation.
- Internal Payment contains payment data collected over three months, with over two million nodes, over fifteen million edges, and five relation types. Data-security and partner-confidentiality requirements restrict further disclosure.

Table 4: Summary of dataset statistics. Graph sizes follow the public dataset releases; Yelp uses the RUR view reported by Wang et al. (2025). Anomaly statistics follow the open-set class selection described in the experimental settings. — denotes statistics withheld for data confidentiality.

Dataset	# Dim.	# Nodes	# Edges	# Ano. classes	# Anomalies	Anomalies (%)
Photo	745	7,650	238,162	2	700	9.15
Computers	767	13,752	491,722	5	2,064	15.01
CS	6,805	18,333	163,788	8	4,159	22.69
Yelp	32	45,954	98,630	2	6,677	14.53
T-Finance	10	39,357	21,222,543	2	1,803	4.58
ogbn-arxiv	128	169,343	1,166,243	4	27,830	16.43
ogbn-proteins	8	132,534	39,561,252	2	10,693	8.07
Internal Payment	18	$>$ 2M	$>$ 15M	—	—	0.03

Anomaly subclass construction. The class-frequency rules in Section 4 determine the anomaly subclasses and hence the seen-class choices on Photo, Computers, CS, ogbn-arxiv, and ogbn-proteins.

For Yelp and T-Finance, we follow the anomaly subclass construction of Wang et al. (2025). DGI (Veličković et al., 2019) learns node representations with a two-layer GraphSAGE encoder, using 128 hidden units per layer, Adam with a learning rate of 0.001, a batch size of 512, and 1,000 training epochs. We apply k-means to the anomalous-node representations to obtain two anomaly subclasses. Labeled training anomalies are sampled with probabilities proportional to their distances from the corresponding cluster centroids. The original binary normal/anomaly labels are retained for evaluation.

B.2 EVALUATION PROTOCOL AND IMPLEMENTATION DETAILS

Each random run traverses every anomaly subclass as the seen class. For each choice, the detector is trained using that subclass and the labeled normal nodes, and the remaining anomaly subclasses are unseen. We repeat the class-rotation experiment five times on each dataset. Photo, Yelp, T-Finance, and ogbn-proteins each have two seen-class settings, giving ten sub-experiments per dataset; Computers has five settings and twenty-five sub-experiments; CS has eight settings and forty sub-experiments; ogbn-arxiv has four settings and twenty sub-experiments. Performance is averaged over the seen-class settings, and tables report means and standard deviations from the repeated experiments.

The training and validation budgets are those specified in Section 4, with normal nodes sampled uniformly and anomaly labels confined to the selected seen subclass. Training, validation, and unlabeled sets are disjoint. Under All, evaluation includes unlabeled normal nodes and both seen and unseen anomalies. Under Unseen, seen anomalies are removed and the same unlabeled normal nodes are retained. Labels of nodes in U are used for evaluation and, together with the label-sampling mechanism, to determine the two aggregate priors supplied to GraphSAGE+PUbN; the baseline receives these proportions without node-level labels from U. TAM, NSReg, DEMO, ConsisGAD, GraphNC, and GraphSAGE+PUbN use the unlabeled nodes during training, as NoRD does. Every baseline is run under these partitions with its public implementation; the reported numbers are therefore not those of the original papers, whose partitions differ.

For the label-budget experiment, we use $k \in \{5, 10, 20, 50, 100\}$ labeled training anomalies. The normal training fraction remains 5%, and validation retains 30 anomalies from the same seen subclass and 1% of normal nodes. Each budget follows the same class-rotation procedure. The budget k counts training anomalies and excludes the fixed validation labels.

Hyperparameter selection. For NoRD and the neural baselines, including both PUbN stages, we use Adam (Kingma & Ba, 2015) and grid-search learning rates $\{0.0001, 0.0005, 0.001, 0.005\}$ and dropout rates $\{0, 0.1, 0.2, 0.5\}$. Each configuration is trained for 200 epochs on Photo, Computers, and CS and 300 epochs on Yelp, T-Finance, ogbn-arxiv, ogbn-proteins, and Internal Payment. We select the configuration with the highest arithmetic mean of validation AUROC and AUPRC, evaluated using its final-epoch model.

GraphSAGE+PUbN. GraphSAGE+PUbN treats R as positives, S as biased negatives, and U as unlabeled data. It receives the split-specific normal prior π and biased-negative prior $\rho_{\mathrm{bN}} = \Pr(y = -1, b = 1)$, where $b = 1$ denotes selection into the biased-negative component; we determine ρ_{bN} from the negative-label sampling mechanism. An auxiliary GraphSAGE is trained with nnPU using prior $\pi + \rho_{\mathrm{bN}}$; its posterior is then fixed to train an independently initialized detector with the PUbN risk. The detector matches NoRD's input processing, GraphSAGE architecture, and scoring network. We additionally search $\tau \in \{0.5, 1, 2\}$, which sets the fraction $\tau(1-\pi-\rho_{\mathrm{bN}})$ of U used to determine the posterior quantile cutoff, retaining candidates with $1 \le \lfloor \tau(1-\pi-\rho_{\mathrm{bN}})|U| \rfloor < |U|$.

B.3 Baselines

The twelve baselines differ in their available supervision and training objectives; each is run with its public implementation under the partitions and evaluation rules of Appendix B.2.

- **TAM** (Qiao & Pang, 2023) maximizes local node affinity on iteratively truncated graphs and scores nodes according to their affinity to neighbors. It learns without node labels.
- **GGAD** (Qiao et al., 2024) generates pseudo-anomalous nodes using asymmetric local affinity and egocentric closeness, providing negative samples for a classifier trained with labeled normal nodes.
- **RHO** (Ai et al., 2025) uses adaptive spectral filters to model varying normal-node homophily and aligns channel-wise and cross-channel representations to learn normality from limited normal labels.
- **GraphNC** (Zeng et al., 2026) calibrates a pretrained normal-supervised detector through anomaly-score distribution alignment and perturbation-based normality regularization, using labeled and unlabeled nodes.
- **BWGNN** (Tang et al., 2022) uses beta-wavelet band-pass filters with spectral and spatial localization to capture frequency components associated with graph anomalies.
- **ConsisGAD** (Chen et al., 2024) combines supervised detection with consistency training on unlabeled nodes and learns graph augmentations to exploit limited supervision.
- **SpaceGNN** (Dong et al., 2025) learns node projections into multiple spaces, performs distance-aware propagation, and combines information across spaces for anomaly detection with limited labels.

- **PMP** (Zhuo et al., 2024) distinguishes neighbors of different classes through node-specific aggregation functions, allowing the detector to use both homophilic and heterophilic neighborhoods.
- **XGBGraph** (Tang et al., 2023) combines neighborhood-aggregated features with an XGBoost classifier, incorporating graph information into tree-based supervised detection.
- **NSReg** (Wang et al., 2025) regularizes supervised representation learning by predicting connected normal, unconnected normal, and unconnected normal-to-unlabeled relations.
- **DEMO** (Yu et al., 2026) combines multi-sample anomaly mixup, energy-gradient-based weighting, and adaptive pseudo-labeling to learn from seen anomalies and unlabeled nodes in open-set GAD.
- **GraphSAGE+PUbN** (Hsieh et al., 2019) trains a GraphSAGE detector through positive–unlabeled risk learning with biased labeled negatives.

B.4 Training algorithm and implementation details

We optimize Equation (2) with Adam and initialize the mixture-link parameter at $\eta = -2$. Each update includes all labeled seen anomalies and a uniformly sampled normal subset with $|B_R| = \min(512, |R|)$. The reference offset uses all of R, and the discrimination loss uses all of R and U. After training, we score the unlabeled nodes with the final model in evaluation mode, as summarized in Algorithm 1.

The ablation variants use the hyperparameter configuration selected for the full model.

Algorithm 1 Training and scoring with NoRD

Input: Graph G, labeled normal set R, labeled seen-anomaly set S, unlabeled set U, epochs T, discrimination weight λ.
Output: Anomaly scores for unlabeled nodes.

Initialize θ and $\eta = -2$.
Assign $d_i = 0$ for $i \in R$ and $d_i = 1$ for $i \in U$.
for each of the T training epochs **do**
  Sample $B_R \subseteq R$ uniformly with $|B_R| = \min(512, |R|)$.
  Compute $f_i = f_\theta(G, i)$ for $i \in R \cup S \cup U$.
  Compute b_R from the full R and $u_i = f_i - b_R$ using Equation (6).
  Compute membership logits $a_i = g_\eta(u_i)$ for $i \in R \cup U$ using Equation (7).
  Compute $\mathcal{L}_{\text{seen}}$ on $B_R \cup S$ using Equation (3).
  Compute $\mathcal{L}_{\text{dis}}$ on the full R and U using Equation (8).
  Update (θ, η) with Adam on $(1-\lambda)\mathcal{L}_{\text{seen}} + \lambda\mathcal{L}_{\text{dis}}$, differentiating through b_R.
end for
Recompute $f_\theta(G, i)$ for $i \in U$ with the final model in evaluation mode.
return $s_i = \sigma(f_\theta(G, i))$ for $i \in U$.

Visualization procedure. For Figure 4, we project the representations entering each model's scoring head to 50 dimensions with PCA and then to two dimensions with UMAP. We construct a uniform 64×64 grid and retain points within the convex hull of the projected nodes. These points are mapped back through the UMAP and PCA inverse transforms and evaluated by the original scoring head to obtain the reconstructed score contours.

C Additional Experimental Results

C.1 Complete public benchmark results

Table 5 includes all eleven GAD baselines and GraphSAGE+PUbN on T-Finance, ogbn-arxiv, and ogbn-proteins, extending the selected comparisons in Table 2. Several unseen-anomaly margins on ogbn-arxiv and ogbn-proteins are small relative to the reported standard deviations, and NoRD is nearly tied with DEMO in unseen AUPRC on ogbn-arxiv.

Table 5: AUROC and AUPRC under All and Unseen, in percent (mean ± standard deviation). Bold and underlined entries denote the highest and second-highest means, respectively.

Method	AUROC			AUPRC		
	T-Finance	ogbn-arxiv	ogbn-proteins	T-Finance	ogbn-arxiv	ogbn-proteins
			All			
TAM	65.66±1.09	56.82±0.04	43.71±1.04	12.10±0.88	18.88±0.01	16.60±1.48
GGAD	71.94±8.97	50.62±1.16	34.39±0.68	15.67±2.80	16.05±0.49	5.81±0.03
RHO	60.66±3.74	45.01±0.61	56.61±2.55	25.49±0.57	14.47±0.13	13.84±0.89
GraphNC	62.44±4.22	50.35±0.55	34.80±0.88	29.89±9.06	16.94±0.21	10.86±0.04
BWGNN	72.90±2.76	60.80±3.01	68.64±8.88	36.93±4.69	27.97±4.53	54.98±4.52
ConsisGAD	57.23±0.68	60.65±1.43	60.56±4.21	22.15±2.10	27.86±4.10	45.96±1.14
SpaceGNN	92.94±0.94	60.41±4.84	74.73±3.60	76.94±1.11	27.69±4.31	57.98±6.94
PMP	90.06±4.41	60.68±2.99	66.97±6.67	75.79±7.98	25.25±4.07	46.55±3.70
XGBGraph	88.48±1.36	60.92±3.41	80.17±3.51	73.40±2.83	27.80±4.50	61.08±7.17
GraphSAGE+PUbN	90.75±0.61	60.30±1.24	74.42±1.71	72.86±3.30	28.79±0.69	60.32±1.15
NSReg	90.03±1.62	62.43±2.22	81.95±7.31	73.70±1.74	29.05±4.03	64.11±4.46
DEMO	88.48±0.57	60.45±2.77	76.88±6.87	74.41±1.66	28.50±4.32	62.53±2.42
NoRD	**93.26±0.67**	**63.25±2.50**	**83.59±6.62**	**78.05±2.43**	**29.57±4.69**	**67.96±6.06**
			Unseen			
TAM	68.53±2.10	50.82±1.65	41.47±2.30	11.34±8.27	13.42±2.35	2.61±0.64
GGAD	73.07±3.67	50.65±2.86	29.95±1.11	13.35±8.98	12.63±0.91	3.29±0.69
RHO	62.92±3.30	45.06±3.29	56.80±2.28	23.08±1.40	11.35±0.76	5.52±0.20
GraphNC	64.89±2.86	50.33±1.74	30.36±1.13	28.76±2.53	13.34±0.91	3.32±0.71
BWGNN	86.12±3.30	51.89±1.55	37.37±1.19	69.12±1.00	13.65±2.31	3.57±0.37
ConsisGAD	68.11±1.27	49.25±2.35	42.33±2.83	30.19±1.86	13.07±1.99	6.23±0.99
SpaceGNN	85.28±4.49	51.95±1.99	45.59±2.32	61.53±1.78	13.86±2.25	11.61±0.49
PMP	89.21±3.86	54.65±2.78	32.53±1.29	71.11±1.33	14.41±2.15	5.20±0.27
XGBGraph	90.34±8.33	52.36±1.01	49.08±1.23	60.39±1.07	13.94±2.55	10.03±0.57
GraphSAGE+PUbN	92.74±0.40	51.33±1.69	63.20±1.72	72.88±3.24	14.61±0.68	41.18±0.37
NSReg	93.83±1.36	54.56±1.45	74.50±0.48	77.29±2.09	14.49±0.17	41.44±0.65
DEMO	91.40±1.03	54.98±0.82	72.16±1.91	80.55±2.31	14.83±0.31	43.06±0.53
NoRD	**94.77±1.02**	**56.21±0.46**	**74.91±0.45**	**81.84±2.96**	**14.97±0.76**	**43.94±0.56**

C.2 Number of labeled training anomalies

Figure 5 complements the Unseen analysis with All performance under the same fixed validation budget. NoRD leads All AUROC at every budget on Photo and Computers, and All AUPRC on Photo; DEMO has a slightly higher Computers All AUPRC at ten labels. With only five training anomalies, NoRD exceeds the three baselines trained with one hundred in All AUROC on Computers and CS and in All AUPRC on Photo and CS.

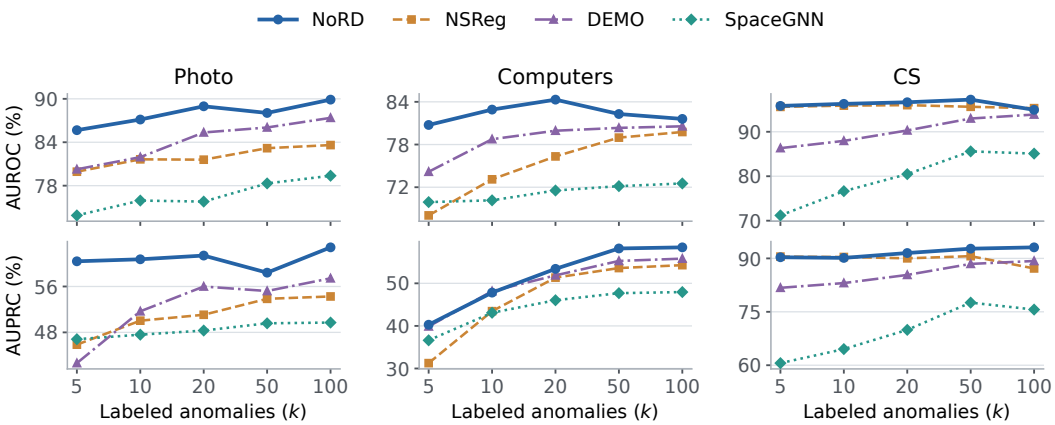

Figure 5: All AUROC and AUPRC versus the number of labeled training anomalies on Photo, Computers, and CS. Points show means; budget positions are equally spaced.

Figures 6 and 7 report the remaining four graphs. T-Finance illustrates the distinction between overall and unseen detection: NoRD trails some baselines in All AUROC at most budgets but leads Unseen AUPRC throughout. Its Unseen AUROC lead ends at one hundred labels, where NSReg is slightly higher, while its All AUPRC lead has one exception at twenty labels. Yelp retains the Unseen lead across all budgets. On ogbn-proteins, the Unseen AUPRC lead persists through fifty labels, indicating that additional seen-class labels do not uniformly strengthen the advantage on unseen anomalies.

Computers benefits more from additional seen-class examples: its All AUPRC has the largest decrease when the training budget falls from fifty to five anomalies, and its Unseen AUPRC overtakes DEMO only from twenty labels onward.

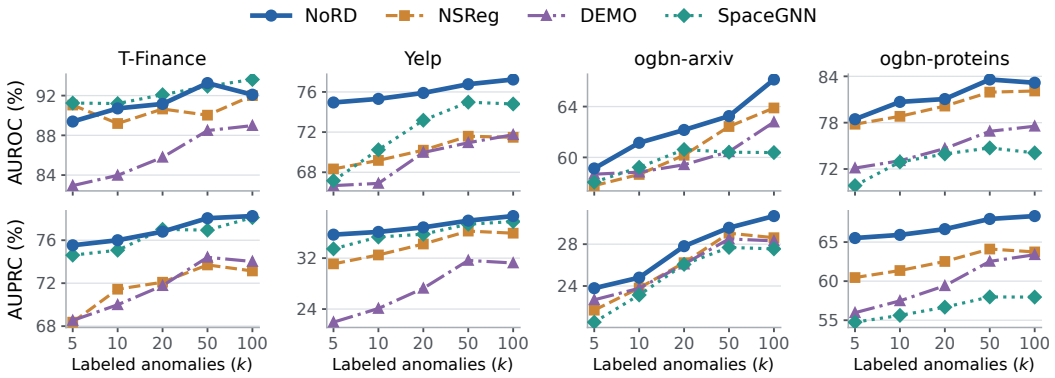

Figure 6: All AUROC and AUPRC versus the number of labeled training anomalies on T-Finance, Yelp, ogbn-arxiv, and ogbn-proteins. Points show means; budget positions are equally spaced.

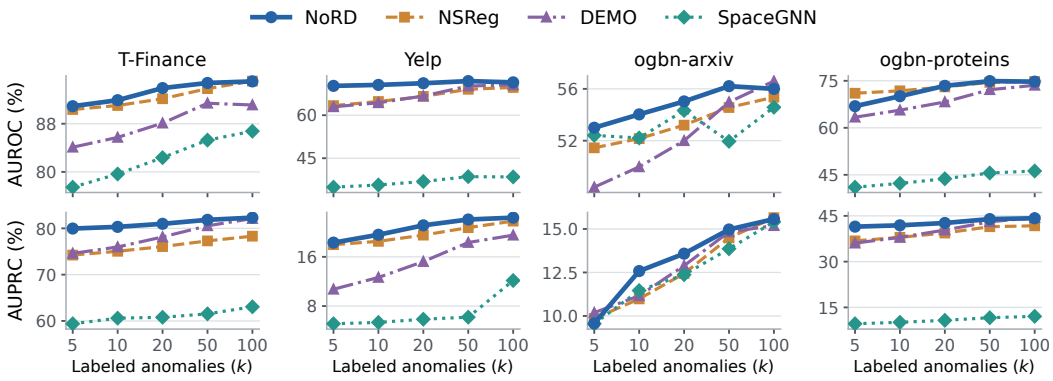

Figure 7: Unseen AUROC and AUPRC versus the number of labeled training anomalies on T-Finance, Yelp, ogbn-arxiv, and ogbn-proteins. Points show means; budget positions are equally spaced.

C.3 Discrimination weight

Figure 8 evaluates the remaining four graphs at the same nine weights as Figure 3. T-Finance remains stable over $\lambda = 0.1$–0.9, while ogbn-proteins reaches a plateau over 0.5–0.9; removing supervision at $\lambda = 1$ reduces all four metrics below their default values on both graphs. On Yelp, increasing λ from 0.5 to 0.9 improves Unseen AUPRC while leaving All AUPRC nearly unchanged, but removing supervision lowers both. On ogbn-arxiv, All AUROC and both Unseen metrics improve through $\lambda = 1$, whereas All AUPRC peaks at $\lambda = 0.9$. The Yelp and arxiv results therefore show that the weight favoring unseen detection need not favor overall detection.

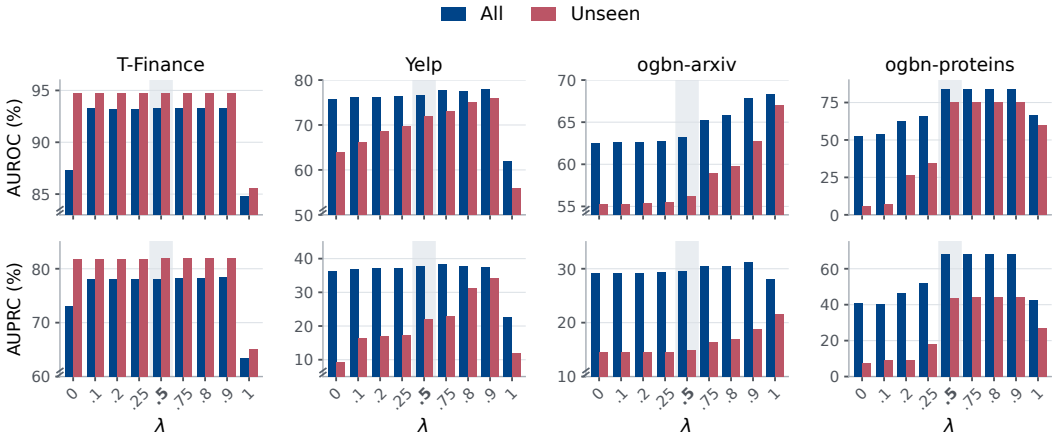

Figure 8: All and Unseen performance across nine discrimination weights on T-Finance, Yelp, ogbn-arxiv, and ogbn-proteins. Bars show means (%).

C.4 Experimental results on Internal Payment

Table 6 compares NoRD with eleven GAD baselines on Internal Payment under the protocol in Section 4.

Table 6: All and Unseen results on Internal Payment, in percent (mean ± standard deviation). Bold and underlined entries denote the highest and second-highest means, respectively.

Method	All		Unseen	
	AUROC	AUPRC	AUROC	AUPRC
TAM	98.81±0.18	32.18±1.08	97.63±0.31	17.36±1.17
GGAD	97.35±0.31	15.27±1.46	95.73±0.48	7.83±1.21
RHO	97.78±0.27	18.63±1.31	96.34±0.42	9.94±1.09
GraphNC	98.10±0.25	21.84±1.24	96.64±0.39	11.26±1.13
BWGNN	98.61±0.20	28.71±1.02	97.46±0.33	16.08±0.96
ConsisGAD	98.33±0.23	24.56±1.16	97.05±0.36	13.42±1.04
SpaceGNN	<u>99.35±0.10</u>	<u>45.82±0.55</u>	98.42±0.19	25.76±0.61
PMP	97.91±0.28	19.92±1.38	96.54±0.43	10.81±1.16
XGBGraph	99.28±0.11	43.67±0.62	98.43±0.18	25.91±0.54
NSReg	99.32±0.09	44.91±0.57	<u>98.46±0.17</u>	<u>26.32±0.49</u>
DEMO	98.71±0.19	30.41±0.94	97.69±0.29	17.92±0.88
NoRD	**99.38±0.08**	**47.12±0.39**	**98.49±0.16**	**26.85±0.42**

NoRD obtains the highest mean AUROC and AUPRC under both evaluations. SpaceGNN is the strongest All baseline, and NSReg leads the Unseen baselines. NoRD's AUROC gain is 0.03 percentage points in each case; its AUPRC gains are 1.30 points under All and 0.53 under Unseen. With anomalies constituting 0.03% of nodes, the leading methods have similar AUROC values, while their AUPRC values differ more. The comparison covers a larger payment graph with lower anomaly prevalence than the public benchmarks.