REND: REJECT-ENCODER NOVELTY DIFFUSION FOR OUT-OF-DISTRIBUTION DETECTION ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for out-of-distribution (OOD) detection on graphs read their score from a representation that was trained for another purpose, namely classifier logits, classifier embeddings, or diffused class evidence. When the observed classes are fewer than the embedding dimensions, every score read from the logits is invariant along a residual subspace of the embedding. Scores that read the whole embedding escape this invariance, but classification training never requires any direction of the representation to separate novel nodes from familiar ones, so whatever novelty structure a borrowed representation retains is incidental. We propose REND (Reject-Encoder Novelty Diffusion), in which a dedicated reject encoder reads the raw features and the graph structure directly and is trained by novelty objectives alone. Since OOD labels are not available at training time, a self-supervised surrogate-outlier loss supplies the missing supervision, and stop-gradients keep the reject branch from reaching the classifier, which leaves the accuracy unchanged. A personalized-PageRank kernel then smooths the learned reject propensity, and its teleport floor guarantees that every node keeps a fixed share of its own novelty. REND shows excellent performance on twelve homophilous, heterophilous, and heterogeneous benchmarks.

1 INTRODUCTION

Motivation. Graph neural networks are the standard tool for semi-supervised node classification (Kipf & Welling, 2017; Hamilton et al., 2017), and they are now used in fraud screening, content moderation, and scientific discovery. In these applications the graph often contains out-of-distribution (OOD) nodes from classes that no training label covers, such as a new fraud pattern or an emerging research topic, and a classifier trained on the known classes still assigns each such node a known label, often with high confidence (Hendrycks & Gimpel, 2017). A deployed model therefore needs a per-node score that flags such nodes, so that it can abstain on them or defer them to a human reviewer.

Limitations of Existing Methods. Many methods now supply such a score for graphs. Some carry confidence, energy, and distance scores over from the independent and identically distributed (i.i.d.) setting (Hendrycks & Gimpel, 2017; Liu et al., 2020; Lee et al., 2018), and the evidential recipe of predicting a Dirichlet distribution over the classes comes from there as well (Sensoy et al., 2018; Charpentier et al., 2020). Others are graph-specific: they diffuse node energies along the edges (Wu et al., 2023), carry the Dirichlet parametrization onto the graph itself (Zhao et al., 2020; Stadler et al., 2021), or estimate uncertainty through stochastic diffusion and conformal calibration (Lin et al., 2024; Huang et al., 2023). All of these methods measure their novelty score in a representation trained to classify. The cross-entropy objective asks nothing of the representation beyond separating the observed classes, so the space from which every logit, entropy, or density score is read is free to discard the off-manifold directions along which an unfamiliar node departs (Lee et al., 2018), and no term of the objective rewards keeping them. When uncertainty is coupled to the label computation, as in the evidential objectives that replace the softmax with a Dirichlet head (Sensoy et al., 2018; Zhao et al., 2020), the detector and the classifier share one set of parameters, and on the many-class Facebook graph this coupling costs the evidential network sixteen points of accuracy against a plain softmax classifier. A score propagated along the edges is a weighted sum of local scores and inherits their invariances. Neighborhood averaging also spreads a concentrated anomaly back into its surroundings through the same ego–neighbor mixing that weakens the classification signal under heterophily (Zhu et al., 2020), and the posterior-network family diffuses class evidence with a

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.

restart (Stadler et al., 2021), so the retention guarantee of its kernel protects each node's confidence and not its novelty.

Proposed Approach. We propose **REND**, short for Reject-Encoder Novelty Diffusion, which learns the novelty score in a separate encoder instead of reading it from the classifier. This lightweight reject encoder reads the raw features and the graph directly, so it can keep the directions to which the classifier's scores cannot respond. Because OOD labels are unavailable, a self-supervised surrogate-outlier objective trains it by corrupting the features of randomly chosen nodes and requiring the reject score to rise on them. Stop-gradients and a branch-separated optimizer decouple the reject branch from the classifier, so the classifier is trained by the cross-entropy alone and the detector can be retuned or removed without changing a single prediction. A personalized-PageRank kernel then smooths the learned reject propensity under a teleport floor, which keeps a fixed share of every node's own novelty at any depth. The same floor caps the novelty that a familiar node can borrow from its novel neighbors, so one learned rate bounds both what a node keeps and what it imports. Since the floor holds for every row-stochastic transition, a fixed uniform walk suffices, and each diffusion step is one sparse message pass with cost linear in the number of edges. Both reject losses act on the smoothed score, so the encoder is trained through the diffusion for the very score that REND reports.

Contributions. The main contributions of this paper are summarized as follows.

- We show that every score read from the logits of a classifier is invariant on an embedding subspace whose dimension grows as the number of classes decreases, and that no step of classification training requires any direction to separate novel nodes from familiar ones (Proposition 1). We therefore give the detector its own reject encoder, train it by novelty objectives alone through a *self-supervised surrogate-outlier loss*, and decouple it from the classifier down to the optimizer, at no measurable cost in accuracy.
- We apply the standard personalized-PageRank kernel to the reject propensity instead of class evidence, so its teleport floor guarantees that every node keeps a fixed share of its own reject propensity at every horizon (Proposition 2). Because the floor holds for every row-stochastic transition and every horizon, the walk and the depth can stay fixed, and no parameters are spent on learning the transition.
- *REND* attains excellent results on twelve homophilous, heterophilous, and metapath-induced heterogeneous benchmarks, and component ablations trace its margin to the learned reject representation, not to the operator that propagates it.

2 Related Work

Uncertainty Quantification on Graphs. Graph uncertainty methods fall into four broad families, and wherever one of them produces a novelty score, that score is read from a representation supervised only by the observed class labels. Discriminative scores take the softmax response (Hendrycks & Gimpel, 2017) or the free energy of the logits (Liu et al., 2020) directly from the classifier, and the graph variant propagates that energy along the edges (Wu et al., 2023). Evidential networks place a Dirichlet distribution on the class probabilities (Sensoy et al., 2018; Zhao et al., 2020), and the graph posterior network derives density-based pseudo-counts (Charpentier et al., 2020) and diffuses them by personalized PageRank (Stadler et al., 2021). Post-hoc estimators score a frozen representation by Mahalanobis distance (Lee et al., 2018), while conformalized graph networks turn the classifier's outputs into prediction sets with a coverage guarantee and train a topology-aware correction model to keep those sets small (Huang et al., 2023). A fourth family learns uncertainty jointly through stochastic diffusion, using a forcing network trained alongside the drift (Lin et al., 2024). Even the open-world line, which lets training acknowledge unseen-class nodes (Wu et al., 2020; Song & Wang, 2022), ties its decision to the classifier trained beside it, either by thresholding that classifier's softmax or by attaching a single outlierness unit to its latent space and keeping the two in agreement. In each of these methods the uncertainty score is computed from a network optimized to classify, and none of them trains a separate representation for novelty.

Graph Diffusion, Outlier Exposure, and Heterophily. Diffusion has supported semi-supervised learning since harmonic label propagation fixed the labeled nodes and solved for the rest (Zhu et al., 2003), and a retention term prevents the diffusion from dissipating that evidence. Partially absorbing walks hold back a share of the mass at each node through a per-node absorption rate (Wu et al., 2012),

and in that view the teleport of personalized PageRank is the special case of one rate at every node, while predict-then-propagate architectures carry that teleport into a neural propagation (Gasteiger et al., 2019). All of these methods diffuse label evidence. REND applies the same off-the-shelf kernel to a reject propensity, so its retention floor bounds how much of a node's own novelty survives the diffusion. Outlier exposure trains a detector against surrogate anomalies (Hendrycks et al., 2019), which can be synthesized in feature space (Du et al., 2022) or, on graphs, by the adversarial latent generator of GOLD (Wang et al., 2025). The surrogate used here is simpler, since the feature-shuffle corruption of deep graph infomax (Veličković et al., 2019) is reused for rejection, and the propensity it supervises originates in classification with a reject option (Chow, 1970; Cortes et al., 2016). A separate line of sampling-based and ensemble-based estimators obtains epistemic uncertainty at the cost of repeated model queries (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Finally, the heterophily literature questions the neighborhood averaging that all of the above rely on (Zhu et al., 2020), although well-tuned standard architectures remain competitive with the operators proposed to replace them (Platonov et al., 2023). We therefore leave the transition fixed and let the teleport floor, which holds for every transition, carry the guarantee.

3 Proposed Method

We consider transductive node classification on an undirected attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes, adjacency $\boldsymbol{A}$, features $\boldsymbol{X} \in \mathbb{R}^{N \times F}$, and K classes. A subset $\mathcal{V}_L \subset \mathcal{V}$ carries labels $y_i \in \{1, \ldots, K\}$, and the remaining nodes are unlabeled. The setting is transductive in the strong sense, because the graph seen during training already contains the unlabeled mixture of in-distribution and OOD test nodes.

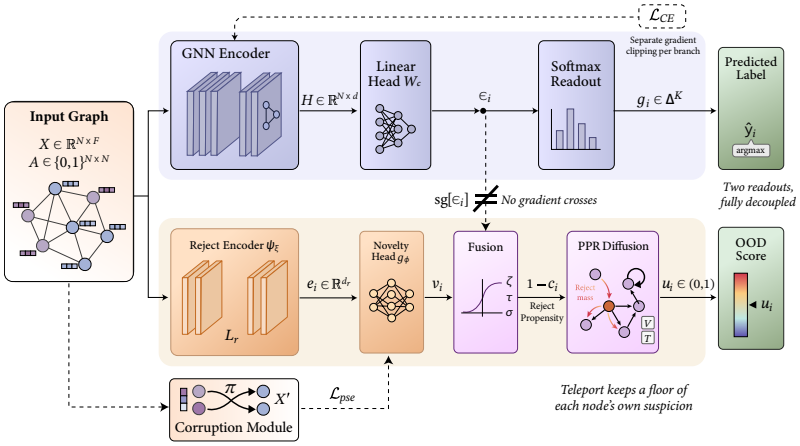

Figure 1: General overview of the proposed REND architecture. The reject branch (bottom) fuses the learned novelty ν_i with the detached energy $\mathrm{sg}[\epsilon_i]$ into the reject propensity $1 - c_i$, which the PPR kernel $\boldsymbol{V}$ spreads into the OOD score u_i, with no gradient path to the classifier (top).

Nodes later scored as OOD therefore contribute features and edges, though never labels, to every forward pass. The learner thus sees the OOD population but not its annotation, and the inductive regime, in which that population arrives only after training, is studied in Appendix .9. Besides predicting the labels of the unlabeled nodes, we require a scalar uncertainty score $u_i \in (0, 1)$ that is large when node i is OOD, so that unreliable predictions can be identified or abstained upon. Structure-aware detectors read a local signal at each node and then spread it along the edges, so that the decision at a node reflects its neighborhood, and existing detectors differ mainly in the operator that performs this spreading. The local signal, however, limits what any such operator can achieve, since a smoothed score is a weighted combination of local scores, and combinations of scores that are invariant to novelty remain invariant to it. REND therefore learns the local signal with a reject channel that has its own encoder and training signal and no gradient path to the classifier, and it spreads the resulting propensity with an off-the-shelf personalized-PageRank kernel (Fig. 1).

3.1 Reject Channel

Let f_θ be any message-passing encoder producing node embeddings $\boldsymbol{H} = f_\theta(\boldsymbol{X}, \boldsymbol{A}) \in \mathbb{R}^{N \times d}$, from which a linear head yields class logits $\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{W}_c \in \mathbb{R}^{N \times K}$ and a predictive distribution $\boldsymbol{g}_i = \mathrm{softmax}(\boldsymbol{z}_i)$. The most direct per-node novelty cue comes from the classifier itself. Its free energy $\epsilon_i = -\log \sum_k \exp(z_{ik})$ is a strong novelty statistic for discriminative models (Liu et al.,

2020; Wu et al., 2023), since $-\epsilon_i$ is a smooth maximum of the logits and therefore retains the magnitude the softmax normalizes away, leaving ϵ_i large when no class logit is large. REND keeps the free energy as one of two cues. Like every score read from the logits, however, it inherits the following invariance of the classification head.

Proposition 1 (Invariance of logit scores)**.** *Let $\boldsymbol{z}_i = \boldsymbol{W}_c^\top \boldsymbol{h}_i$ with $\boldsymbol{W}_c \in \mathbb{R}^{d\times K}$, and let $s(\boldsymbol{z}_i)$ be any score read from the logits, such as the free energy, the maximum softmax probability, or the entropy. Then (i) s is invariant to every embedding shift $\boldsymbol{\delta} \in \ker(\boldsymbol{W}_c^\top)$, which is a subspace of dimension at least $d - K$ and hence non-trivial whenever $K < d$, so that two embeddings differing by an arbitrarily large displacement within it receive identical scores; and (ii) at every training step the cross-entropy gradient at the embedding, $\nabla_{\boldsymbol{h}_i}\mathrm{CE} = \boldsymbol{W}_c(\boldsymbol{g}_i - \boldsymbol{y}_i)$ with $\boldsymbol{y}_i$ the one-hot label, lies in the column space of the current $\boldsymbol{W}_c$, so that the loss of that step is invariant to the component of $\boldsymbol{h}_i$ in the concurrent null space and no encoder update rewards or penalizes it.*

By part (i), no logit score can react to a displacement inside the null space, however large, and that space is widest when only a handful of classes are seen during training. Part (i) limits what a logit score can express and makes no claim about where novel nodes lie, and since it holds for every function of the logits, replacing the energy with the maximum softmax probability or the entropy cannot recover what the logits discard. Appendix .1 gives the proofs of all formal results together with a discussion of their scope.

A density score computed on $\boldsymbol{h}_i$ itself, such as the Mahalanobis distance (Lee et al., 2018), reads the whole embedding and is therefore not subject to part (i). By part (ii), however, no encoder update rewards or penalizes the null-space component that such a score reads beyond the logits, and no term of the objective asks any direction to separate novel nodes from familiar ones, so the structure the score relies on is only a side effect of classification training. We find that side effect too weak to rely on, as REND outperforms Mahalanobis on every benchmark of Table 1 and placing the same novelty head on the detached classifier embedding costs about twelve points of mean detection (Section 4.2). We therefore give the reject branch a separate encoder, a lightweight stack of L_r mean-aggregator layers ψ_ξ that reads the raw features and the graph directly, $\boldsymbol{e} = \psi_\xi(\boldsymbol{X}, \boldsymbol{A})$, is trained only by the reject objectives described below, and feeds a small network g_ϕ that outputs a novelty score $\nu_i = g_\phi(\boldsymbol{e}_i)$. Since no classification gradient reaches ψ_ξ, the reject objectives alone determine its representation, which can therefore carry the directions to which, by Proposition 1, no logit score responds. We standardize each cue against the labeled nodes and add them, which gives a per-node in-distribution affinity

$$a_i = \underbrace{\zeta\big(-\epsilon_i\big)}_{\text{energy}} + \underbrace{\zeta\big(-\nu_i\big)}_{\text{novelty}}, \qquad \zeta(v_i) = \frac{v_i - \mathrm{mean}_{j\in\mathcal{V}_L} v_j}{\mathrm{std}_{j\in\mathcal{V}_L} v_j}, \tag{1}$$

where high a_i marks an in-distribution node. A learnable steepness τ and the graph-adaptive midpoint $\mathrm{med}(a) + b$, the median of the affinity over all nodes shifted by a learnable offset b, map the affinity to an acceptance propensity, where σ is the logistic sigmoid,

$$c_i = \sigma\big(\tau\,(a_i - \mathrm{med}(a) - b)\big) \in (0, 1), \tag{2}$$

so that a confident node has $c_i \to 1$ whereas a novel one has $c_i \to 0$. The median adapts the boundary to the scale of the affinity on each graph, and the offset, initialized at zero and trained by the reject losses, lets that boundary move away from an even split of the nodes. The complement $1 - c_i$ is the node's *reject propensity*, that is, its purely local tendency to be rejected, and it is the only quantity the rest of the model propagates. The class distribution never travels along the edges, so the diffusion cannot change a prediction.

3.2 Personalized PageRank Smoothing

The propensity $1 - c_i$ is purely local, yet a node whose own evidence looks unremarkable may still lie inside a novel region, where its neighbors carry the evidence that it lacks. Spreading $1 - c$ along the edges recovers such nodes, but an operator that only averages also dilutes each node's own score with those of its neighbors at every step, so each additional step extends its reach and lowers the share of the node's own score that it retains. This loss is mild at the shallow depths that deployed detectors use and complete in the depth limit (Proposition 2(i)), whereas the personalized-PageRank kernel retains a fixed share at every depth (Proposition 2(ii)).

Proposition 2 (Score retention under diffusion)**.** *Fix a graph without isolated nodes. (i) If scores are propagated by averaging over a zero-diagonal row-stochastic transition $\boldsymbol{P}$ supported on the edges, $\boldsymbol{s}^{(\ell+1)} = \alpha\, \boldsymbol{s}^{(\ell)} + (1-\alpha)\, \boldsymbol{P}\boldsymbol{s}^{(\ell)}$ with $\alpha \in (0,1)$, then a single step strictly decreases the score of any node that strictly exceeds its neighbors, and as $\ell \to \infty$ the scores on each connected component converge to a component-wise constant, so that every ranking read within a component collapses to chance. (ii) If instead the score is smoothed by the personalized-PageRank kernel with teleport probability $1-m$ over* any *row-stochastic transition $\tilde{\boldsymbol{P}}$,*

$$\boldsymbol{u} = \boldsymbol{V}(\mathbf{1} - \boldsymbol{c}), \qquad \boldsymbol{V} = (1-m)(\boldsymbol{I} - m\tilde{\boldsymbol{P}})^{-1}, \tag{3}$$

then $\boldsymbol{V}$ is row-stochastic with $V_{ii} \geq 1-m$, and, writing $\boldsymbol{u}^{(T)} = \sum_{t=0}^{T} m^t \tilde{\boldsymbol{P}}^t (1-m)(\mathbf{1}-\boldsymbol{c})$ for the T-step truncation of (3)*, the diffused mass satisfies, at every horizon $T \geq 0$,*

$$(1-m)(1-c_i) \;\leq\; u_i^{(T)} \;\leq\; u_i \;\leq\; \max_j (1-c_j);$$

and, moreover, whenever some other node is at least as novel, that is, $\max_{j\neq i}(1-c_j) \geq 1-c_i$,

$$u_i \;\leq\; (1-m)(1-c_i) + m\, \max_{j\neq i}(1-c_j),$$

so that the diffused novelty never falls below a fixed fraction of the novelty of the node itself, at any horizon, and the mass that a node borrows from its surroundings never exceeds m times the strongest reject propensity among them.

REND propagates the reject propensity by (3) alone, with $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{D}}^{-1}(\boldsymbol{A}+\boldsymbol{I})$, the uniform random walk over the self-loop-augmented adjacency, where $\tilde{\boldsymbol{D}}$ is its degree matrix. The kernel is evaluated by the truncated Neumann recursion $\boldsymbol{x}^{(t+1)} = (1-m)(\mathbf{1}-\boldsymbol{c}) + m\tilde{\boldsymbol{P}}\boldsymbol{x}^{(t)}$ started at $\boldsymbol{x}^{(0)} = (1-m)(\mathbf{1}-\boldsymbol{c})$, whose iterates are exactly the truncations $\boldsymbol{u}^{(t)}$ of Proposition 2, and which is the predict-then-propagate iteration of Gasteiger et al. (2019) applied to the reject propensity. Each step is a single sparse message pass that costs $\mathcal{O}(|\mathcal{E}|)$ and is fully differentiable, and a single global rate $m = \sigma(\eta)$ is learned. Read as a walk, u_i is the probability that a lazy walker leaving node i, halting with probability $1-m$ at each step, stops on a node that the reject channel judges novel.

Training and evaluation use a fixed truncation of $T = 5$ steps, which discards only a geometrically vanishing fraction of the mass, at most m^{T+1} per node (Lemma 1, Appendix .1). The diagonal bound $V_{ii} \geq 1-m$ is the classical teleport floor, and it holds for whatever quantity the kernel smooths. REND applies it to the reject propensity, so it bounds how much of a node's own novelty survives, whereas the graph posterior network applies it to class evidence (Stadler et al., 2021), where it bounds how much of a node's own *confidence* survives. Because a floor bounds retained evidence only from below, in that network it guarantees the certainty of a familiar node, while the lack of evidence at a novel node is governed by the import bound of the same kernel, which at the released teleport of 0.1 still admits nine tenths of the largest evidence among its neighbors. When scores are averaged without a restart, as in the energy-propagation detectors, the retention carries no depth-uniform floor at all, because the guaranteed self-weight α^ℓ vanishes as the horizon grows.

Under the teleport floor, by contrast, the receptive field can grow without lowering the retained share. Since the weight that a node assigns to all other nodes never exceeds m, an in-distribution node bordering a novel cluster can also import at most m times the strongest propensity its walk can reach, so the single learned rate bounds both the novelty that a node loses and the novelty that it borrows. Because the floor holds for *every* row-stochastic $\tilde{\boldsymbol{P}}$, a learned transition would add parameters without strengthening the guarantee, and REND keeps the uniform walk. Appendix .1 shows where the dilution of part (i) is most severe and turns the two bounds of part (ii) into a sufficient margin condition for the ranking that detection metrics evaluate.

3.3 Self-Supervised Outlier Training

Since no OOD *label* is available at training time, the reject propensity has no direct supervision, and an unsupervised smoothing prior alone leaves the reject mass uninformative. We therefore generate surrogate outliers. This follows outlier exposure (Hendrycks et al., 2019; Du et al., 2022) but needs neither an auxiliary corpus nor a density model, because the outliers are drawn from the graph's own nodes. At each training step a fraction π of the nodes is sampled and their features are replaced by those of other nodes drawn at random from $\mathcal{V}$. This is the feature shuffle of deep

graph infomax (Veličković et al., 2019) restricted to a sampled subset, and it produces a corrupted input $\boldsymbol{X}'$ whose perturbed nodes carry features that do not match their neighborhoods while staying on the scale of real features. Under the transductive protocol both the corrupted positions and their feature donors are drawn from the full node set. A genuine held-out-class feature vector can therefore be copied into an in-distribution neighborhood and act, without any label, as a surrogate outlier. In Appendix .9, we isolate the contribution of this access by restricting both targets and donors to the labeled in-distribution nodes. We pass $\boldsymbol{X}'$ through the model and require the reject mass to rise on the corrupted set $\mathcal{C}$ and to stay low on the labeled in-distribution nodes outside it. A labeled node drawn into $\mathcal{C}$ serves at that step as a surrogate outlier rather than as an anchor. The resulting loss is

$$\mathcal{L}_{\text{pse}} = -\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \log u_i' - \frac{1}{|\mathcal{V}_L \setminus \mathcal{C}|}\sum_{i\in\mathcal{V}_L\setminus\mathcal{C}} \log(1-u_i'), \tag{4}$$

where u_i' is the reject mass under $\boldsymbol{X}'$. It is produced by a full forward pass in which the free energy that feeds the propensity is recomputed on the corrupted input and enters, as always, through the stop-gradient. Like the anchor term of (5) below, this objective acts on the diffused mass and not on the local propensity, so the reject encoder ψ_ξ, the novelty head g_ϕ, and the rate m are trained for the score that is finally reported. The reject branch touches the classifier only through detached quantities: the free energy in (1) enters through a stop-gradient, and every learned reject component, namely the encoder ψ_ξ, the head g_ϕ, the steepness τ, the offset b, and the rate m, is optimized by the reject losses alone. Neither $\mathcal{L}_{\text{pse}}$ nor the diffusion can therefore perturb the classifier, trained by cross-entropy alone, so REND keeps the accuracy of a plain graph network and adds a structured OOD detector to it.

3.4 Training Objective and Uncertainty Score

Training minimizes the classifier likelihood together with the two reject terms,

$$\mathcal{L} = \underbrace{\text{CE}(\boldsymbol{Z}_{\mathcal{V}_L}, y)}_{\text{accuracy}} + \lambda_{\text{id}}\,\tfrac{1}{|\mathcal{V}_L|}\sum_{i\in\mathcal{V}_L} u_i + \lambda_{\text{pse}}\,\mathcal{L}_{\text{pse}}, \tag{5}$$

where the middle term anchors the reject mass of the labeled nodes to zero, so u_i measures novelty with respect to the distribution rather than classification difficulty alone. Neither reject term reaches a classifier parameter. The only classifier quantity inside u_i and u_i' is the free energy, which enters the affinity through the stop-gradient of Algorithm 1, so the gradient of (5) with respect to θ and $\boldsymbol{W}_c$ is that of the cross-entropy alone. At inference the OOD score is the diffused reject mass u_i of (3), while the predicted label is read from the local belief, $\hat{y}_i = \arg\max_k g_{ik}$. No choice made in the reject branch affects the local belief, so the detector can be retuned or removed without changing a single prediction. A single global gradient clip would still couple the two branches, since a large reject gradient would shrink the classifier's step. The branches are therefore clipped *separately*, so the size of the classifier's step never depends on the reject gradient. Algorithm 1 in Appendix .2 collects the complete training step.

4 Experiments

Datasets and Setup. We evaluate on twelve benchmarks spanning three families: seven homophilous citation, co-purchase, and attributed social graphs; the heterophilous Chameleon and Flickr; and the metapath-homogenized ACM, IMDB, and DBLP-H. Distribution shift follows the leave-out-classes protocol of Zhao et al. (2020), in which selected classes leave the training and validation sets while their nodes stay in the graph as the out-of-distribution pool. The last $\lfloor K/2 \rfloor$ classes by index are withheld on every graph, so no dataset receives a tailored choice, and Appendix .3 reports how this differs from the released splits. Every method therefore operates transductively on the same unlabeled mixture, and we quantify the value of that access in Appendix .9. The classifier is a two-layer GraphSAGE network (Hamilton et al., 2017) with 64 hidden units, the reject encoder adds $L_r = 3$ mean-aggregator layers, and the diffusion runs for $T = 5$ steps, with one configuration shared by all twelve datasets (Appendices .3 and .4).

Baselines and Metrics. Every baseline shares REND's GraphSAGE encoder and single fixed configuration, so the comparison isolates how uncertainty is read out. We compare the free-energy

Method	CiteseerF			PubmedF			DBLP			Photo		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
GCN	63.7 ± 11.9	76.7 ± 9.3	90.7 ± 5.2	89.7 ± 3.0	61.9 ± 5.5	94.9 ± 1.1	82.6 ± 5.9	65.3 ± 7.8	89.5 ± 3.0	72.0 ± 7.0	73.5 ± 8.7	95.2 ± 2.1
Softmax	51.8 ± 7.8	81.3 ± 4.0	92.9 ± 1.6	87.2 ± 3.6	61.4 ± 5.3	94.7 ± 0.9	85.2 ± 4.4	64.4 ± 7.2	89.0 ± 2.8	64.5 ± 6.5	82.2 ± 9.5	97.1 ± 2.2
GNNSafe	47.0 ± 8.8	83.8 ± 4.0	93.7 ± 1.9	84.5 ± 4.9	62.7 ± 5.4	95.1 ± 0.9	81.3 ± 6.7	68.3 ± 9.9	90.5 ± 4.0	55.5 ± 8.4	84.2 ± 10.8	96.7 ± 2.4
Maha.	48.4 ± 5.6	82.7 ± 1.8	93.4 ± 0.7	88.2 ± 3.6	63.9 ± 4.1	95.2 ± 0.6	82.0 ± 3.9	66.9 ± 1.9	89.8 ± 0.7	43.5 ± 11.4	85.2 ± 6.1	97.2 ± 1.4
EDL	61.2 ± 12.4	77.6 ± 6.3	92.0 ± 2.5	87.5 ± 3.1	61.9 ± 4.4	95.0 ± 0.7	90.1 ± 3.9	60.2 ± 7.1	88.3 ± 2.3	75.2 ± 6.7	74.1 ± 8.6	95.1 ± 2.9
GPN	65.3 ± 17.8	71.1 ± 14.0	87.6 ± 6.9	90.1 ± 5.2	61.0 ± 6.2	95.0 ± 1.4	84.3 ± 10.3	59.5 ± 8.4	87.1 ± 3.1	60.3 ± 19.8	75.4 ± 12.5	95.2 ± 2.6
NODESAFE	51.9 ± 12.7	79.7 ± 8.6	91.6 ± 4.4	85.3 ± 6.2	62.1 ± 7.0	94.7 ± 1.4	**79.9** ± 8.7	69.3 ± 5.6	90.9 ± 2.2	42.1 ± 12.3	90.0 ± 2.5	98.0 ± 0.5
OODGAT	57.1 ± 16.8	68.2 ± 14.9	84.0 ± 7.9	85.7 ± 6.1	51.8 ± 5.3	91.7 ± 1.8	87.6 ± 2.8	48.6 ± 8.2	81.5 ± 3.8	86.6 ± 5.6	49.0 ± 6.9	87.1 ± 2.7
GOLD	48.7 ± 9.9	81.8 ± 3.7	92.8 ± 1.6	77.3 ± 9.8	76.3 ± 7.9	97.3 ± 1.3	84.4 ± 10.6	65.3 ± 9.5	89.8 ± 3.1	46.6 ± 18.5	89.0 ± 5.5	98.3 ± 0.9
REND	**40.7** ± 5.4	**86.8** ± 1.4	**95.2** ± 0.6	**71.2** ± 6.1	**78.5** ± 4.1	**97.7** ± 0.5	81.8 ± 6.5	**70.5** ± 4.8	**91.9** ± 1.5	**6.3** ± 3.5	**97.9** ± 0.4	**99.6** ± 0.1

Method	Wiki			BlogCat			Facebook			Chamel.		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
GCN	85.9 ± 2.6	63.5 ± 2.1	67.6 ± 1.5	60.2 ± 7.1	77.1 ± 4.2	93.1 ± 1.3	93.1 ± 1.5	50.8 ± 0.9	85.1 ± 0.4	93.8 ± 1.6	52.4 ± 3.7	77.9 ± 3.0
Softmax	**52.6** ± 4.7	76.1 ± 4.3	74.9 ± 5.4	68.0 ± 10.5	76.4 ± 9.0	92.8 ± 4.0	81.9 ± 5.7	87.5 ± 1.6	97.7 ± 0.5	68.7 ± 3.7	67.9 ± 4.3	84.0 ± 2.6
GNNSafe	58.1 ± 5.4	71.1 ± 4.1	71.9 ± 3.0	57.8 ± 10.2	80.3 ± 11.2	94.1 ± 5.0	84.5 ± 13.4	88.9 ± 3.2	97.8 ± 1.4	81.4 ± 6.3	66.5 ± 4.5	85.4 ± 3.0
Maha.	61.0 ± 7.8	69.3 ± 4.9	68.4 ± 4.1	53.6 ± 7.6	80.9 ± 3.8	94.0 ± 1.7	62.4 ± 3.7	86.4 ± 2.0	96.1 ± 1.0	94.0 ± 3.1	40.7 ± 8.0	69.5 ± 3.4
EDL	53.7 ± 8.5	79.4 ± 3.8	80.5 ± 3.8	71.4 ± 14.8	75.5 ± 17.3	92.8 ± 7.1	97.6 ± 2.8	38.4 ± 15.3	80.6 ± 5.9	68.8 ± 7.8	65.4 ± 6.0	81.5 ± 2.9
GPN	89.6 ± 3.7	61.9 ± 5.4	67.8 ± 4.6	92.2 ± 2.8	58.8 ± 3.6	88.9 ± 1.2	88.4 ± 4.9	67.6 ± 5.9	90.8 ± 1.7	95.8 ± 1.2	55.6 ± 5.0	78.8 ± 4.0
NODESAFE	58.5 ± 4.3	71.7 ± 1.6	72.5 ± 1.6	54.2 ± 10.8	85.0 ± 5.7	96.2 ± 1.9	71.9 ± 16.3	87.0 ± 4.2	95.8 ± 2.5	89.1 ± 7.2	59.8 ± 5.2	81.5 ± 3.6
OODGAT	70.0 ± 7.2	60.8 ± 4.7	61.3 ± 3.3	86.1 ± 8.4	48.3 ± 7.5	80.8 ± 4.5	94.2 ± 3.1	52.5 ± 13.5	84.7 ± 4.8	69.8 ± 8.3	62.8 ± 6.1	79.7 ± 3.8
GOLD	67.4 ± 7.1	75.4 ± 3.6	78.3 ± 3.6	67.1 ± 11.6	69.9 ± 11.3	90.8 ± 3.5	**45.9** ± 20.4	**90.8** ± 4.1	**98.1** ± 1.1	91.2 ± 2.5	62.6 ± 3.8	83.7 ± 2.3
REND	53.7 ± 10.4	**86.0** ± 3.5	**89.2** ± 3.3	**34.8** ± 8.2	**92.2** ± 2.0	**98.3** ± 0.5	72.4 ± 5.1	87.1 ± 1.8	97.8 ± 0.4	**55.1** ± 3.8	**81.0** ± 2.0	**91.3** ± 1.5

Method	Flickr			ACM			IMDB			DBLP-H		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
GCN	91.5 ± 3.7	55.4 ± 4.1	87.8 ± 1.6	58.4 ± 5.9	79.2 ± 2.8	68.2 ± 3.5	94.8 ± 0.3	59.4 ± 0.4	48.0 ± 0.6	87.3 ± 2.0	60.9 ± 1.7	65.6 ± 1.4
Softmax	74.1 ± 3.9	68.0 ± 5.7	90.8 ± 2.4	71.4 ± 10.9	68.8 ± 7.1	56.9 ± 6.9	93.4 ± 0.4	58.2 ± 0.9	48.2 ± 1.0	81.9 ± 1.2	67.5 ± 1.0	70.1 ± 0.7
GNNSafe	77.4 ± 4.2	68.0 ± 5.5	91.8 ± 1.8	70.4 ± 12.0	70.5 ± 7.8	58.0 ± 7.7	94.7 ± 0.1	59.7 ± 1.4	50.1 ± 1.4	82.0 ± 1.2	69.0 ± 1.1	72.3 ± 0.7
Maha.	73.9 ± 3.8	64.1 ± 4.7	89.2 ± 2.3	86.1 ± 2.7	65.9 ± 2.8	55.3 ± 2.9	**90.0** ± 0.6	56.5 ± 1.6	45.6 ± 1.3	95.1 ± 0.6	51.6 ± 2.1	58.0 ± 1.6
EDL	69.3 ± 6.3	69.8 ± 9.8	91.3 ± 3.6	63.5 ± 8.0	74.1 ± 10.6	64.0 ± 10.0	92.2 ± 0.7	58.6 ± 0.6	47.8 ± 0.7	83.4 ± 1.0	67.7 ± 1.3	72.3 ± 1.1
GPN	94.3 ± 1.1	49.1 ± 1.0	86.4 ± 0.4	71.5 ± 10.8	67.3 ± 10.0	55.8 ± 7.7	93.3 ± 1.0	54.4 ± 1.8	46.0 ± 1.8	93.4 ± 3.8	55.6 ± 8.3	62.3 ± 7.0
NODESAFE	75.4 ± 3.4	71.5 ± 2.3	92.9 ± 0.6	66.4 ± 17.6	59.5 ± 19.0	50.1 ± 15.8	92.3 ± 2.3	57.5 ± 4.3	47.8 ± 3.8	89.5 ± 4.2	64.6 ± 4.3	69.5 ± 2.6
OODGAT	91.5 ± 1.9	49.5 ± 1.8	85.2 ± 0.8	51.4 ± 2.6	62.2 ± 1.3	45.3 ± 1.1	95.0 ± 1.0	52.8 ± 1.3	41.2 ± 0.8	75.4 ± 5.2	63.5 ± 1.2	62.4 ± 2.0
GOLD	88.8 ± 5.8	53.6 ± 9.2	87.2 ± 3.0	54.1 ± 12.0	80.2 ± 9.4	71.3 ± 11.5	92.4 ± 1.9	57.6 ± 2.9	47.6 ± 1.8	87.5 ± 2.7	74.5 ± 4.2	78.2 ± 4.1
REND	**68.5** ± 2.2	**76.3** ± 2.1	**94.3** ± 0.6	**20.4** ± 4.1	**95.1** ± 0.8	**93.1** ± 1.2	91.6 ± 1.1	**62.6** ± 1.6	**52.9** ± 1.7	**65.7** ± 6.5	**85.6** ± 4.1	**88.8** ± 3.5

Table 1: Out-of-distribution detection under the leave-out-classes protocol (%, mean ± standard deviation over ten runs). Best per metric in bold, second best underlined.

classifier (Liu et al., 2020), reported as Softmax with a GCN (Kipf & Welling, 2017) variant, its energy-propagation form GNNSafe (Wu et al., 2023) and its bounded-and-uniform extension NODESAFE (Yang et al., 2024); post-hoc Mahalanobis on the final embedding (Lee et al., 2018); the evidential EDL (Sensoy et al., 2018) and graph posterior network (Stadler et al., 2021); and the exposure and open-world systems GOLD (Wang et al., 2025) and OODGAT (Song & Wang, 2022). No release covers all twelve graphs, so each is ported with its method-defining hyperparameters at their released values (Appendix .4). We omit the stochastic diffusion of Lin et al. (2024), which reads uncertainty from class supervision alone, and the conformalized network of Huang et al. (2023), whose coverage guarantee requires an exchangeability that out-of-distribution nodes violate. Metrics are OOD AUROC, with FPR95 and AUPR in Table 1, accuracy on the observed classes, and misclassification-detection AUROC (Hendrycks & Gimpel, 2017) from each model's own confidence, all reported as means and standard deviations over ten runs (Appendix .5).

4.1 Main Results

Detection Performance. REND attains the highest AUROC and AUPR among the compared methods on eleven of the twelve benchmarks (Table 1). The mean AUROC rises to 83.3% from the 73.1% of the strongest compared baseline, and a Wilcoxon signed-rank test confirms the improvement at $p < 10^{-3}$ against every competitor. REND also raises the mean AUPR from 84.5% to 90.8% and lowers the mean FPR95 from 71.0% to 55.2%, so that at a 95% detection rate it wrongly rejects about sixteen fewer in-distribution nodes in every hundred. The gains are largest where the classifier's own logits are least informative and the structural evidence that REND learns matters most, reaching fifteen AUROC points on the metapath graph ACM, thirteen on the heterophilous Chameleon, and eleven on DBLP-H, while on Photo REND cuts the FPR95 of the strongest baseline from 42.1% to 6.3%.

Consistency with the Retention Bound. The dilution of part (i) of Proposition 2 dominates where neighborhoods mix labels, and GNNSafe falls below the local energy it smooths on two such graphs, Wiki and Chameleon. REND leads GNNSafe by about fifteen AUROC points on both, and the

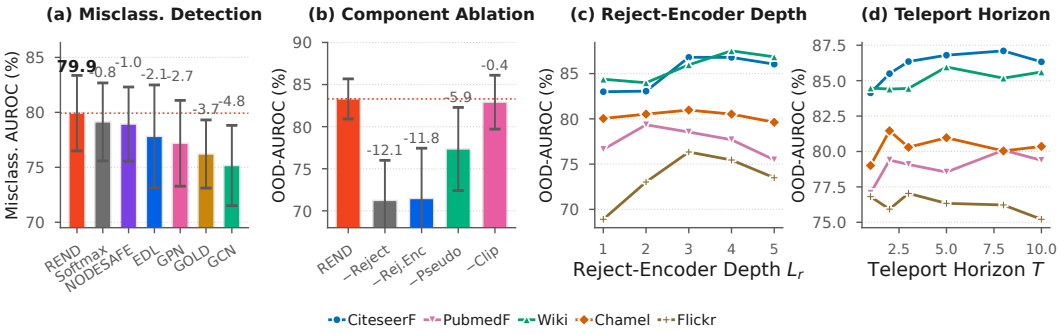

Figure 2: Misclassification-detection AUROC (a) and component contributions (b), averaged over the twelve graphs; reject-encoder depth (c) and teleport horizon (d) on the five-dataset core suite.

ablations of Section 4.2 trace part of this lead to the teleport floor, which keeps at least a $(1 - m)$ share of every node's own reject propensity however strongly its neighbors disagree. On most other graphs the held-out classes form clusters whose members reinforce one another, so a shallow average usually helps. Our feature-corruption study (Fig. 4c) covers isolated anomalies, where every neighbor of a novel node scores lower and this dilution is most severe.

Comparison with Baselines. The kernel of GPN spreads class evidence, so at a novel node on a mixed boundary the evidence imported from confident neighbors masks the absence of the node's own. REND applies the same kernel to the reject mass instead, of which a novel node keeps at least a $(1 - m)$ share whatever its neighbors carry. Mahalanobis falls to chance or below on Chameleon and DBLP-H, where the by-product structure it must model is weakest. The synthetic exposure of GOLD falls to near chance on the dense attributed Flickr graph. The evidential losses become unstable on the many-class Facebook graph, where EDL falls 49 points below the plain softmax it replaces.

Classification Accuracy. The reject branch is detached from the classifier down to the optimizer, so by construction no reject gradient can affect the training of the primary task, and the remaining gap is run-to-run noise. The mean accuracy is 81.9% against the 81.7% of the plain network, a gap of one tenth of a point before rounding, a small fraction of the spread across runs and indistinguishable from zero under a paired test ($p = 0.70$). On misclassification detection REND attains the highest suite mean among the compared methods, with a consistent per-dataset profile (Fig. 2a; Tables 6 and 7 report every dataset). The baselines closest to REND on this task are not those closest on novelty detection, since NODESAFE scores one point below REND in misclassification detection but twelve points below it in novelty detection, whereas REND leads the suite means of both tasks at once, at no cost to the classifier.

4.2 Ablation Study

Reject Channel. Removing the reject channel removes the diffusion from the readout, and the uncertainty reverts to the classifier's local score, a monotone map of its free energy, so the variant matches the local energy baseline (71.2 against 71.6). The channel therefore raises mean detection by twelve points (Fig. 2b) and improves it on all twelve graphs. The largest gains, nearly thirty points on ACM and nineteen on DBLP-H, arise on metapath graphs, where the reject encoder reads the composite edges directly and the diffusion carries its evidence along them.

Reject Encoder. Reading the novelty score from the detached classifier embedding, with the input score, the diffusion, and both losses unchanged, isolates the representation and separates the two parts of Proposition 1. Such a head is not subject to the invariance of part (i), yet it scores 11.8 mean points below REND and below even the purely logit-driven variant discussed next. Reading the null-space directions therefore does not help when, as part (ii) states, nothing in classification training organizes them for detection. The drop is largest on the metapath graphs, whose leave-out split trains on only two classes, and the embedding head reaches only 54.8 on DBLP-H, where REND attains 85.6. Even a single-layer reject encoder, shallower than the two-layer classifier and of the same width, scores five points above the embedding head on the core suite of Fig. 2c, so the gap reflects what each representation is trained for and not its capacity.

Surrogate-Outlier Loss. Without the surrogate loss, the anchor term of (5) still drives the reject mass of the labeled nodes toward zero, but nothing specifies where it should rise, so only the inductive bias of the encoder determines the propensity. Mean detection then falls by six points, by eighteen on the heterophilous Chameleon, and by eleven on BlogCatalog. Restricting every feature donor to a labeled in-distribution node, so that no genuine OOD vector can enter a neighborhood, still leaves REND at a mean AUROC of 79.7, 6.6 points above the strongest baseline of Table 1, and the citation and co-purchase graphs lose less than one point (Appendix .9). The margin of REND over the compared methods therefore does not rest on copying genuine held-out features into the graph.

Novelty Input. Feeding the diffusion with the free energy alone, with everything downstream identical, spreads a statistic that sees the embedding only through the K-dimensional image of the classification head. The diffused score is then a weighted sum of functions of the logits and shares their invariance under part (i) of Proposition 1, and the resulting drop is largest where a node's logits are least informative (Fig. 4a). The learned novelty score reads the reject encoder instead, whose directions are not confined to that image, so the same diffusion receives per-node evidence that the free energy cannot provide.

Source of the Improvement. Driven by the free energy, REND's diffusion is personalized-PageRank smoothing of the same score GNNSafe averages lazily. The two match on average, and the teleport floor gains about eight points on Wiki and three on Chameleon, where averaging falls below the local score, as part (i) of Proposition 2 predicts. Restoring the learned input then adds more than ten mean points, so nearly all of REND's advantage comes from the learned reject representation and not from the diffusion operator.

Energy Cue. The two cues enter (1) standardized against the labeled nodes, so neither dominates by scale. They are complementary, since the energy reads classifier confidence while the reject encoder reads directions that no logit score can express (Proposition 1). The energy cue adds three points on Wiki, whose dense word features keep the classifier's confidence informative, and the novelty cue adds nearly eighteen on DBLP-H, whose composite metapath edges leave the logits confidently incorrect. The sum of both cues attains the best mean detection (Fig. 4a), and because the energy enters through a stop-gradient, the classifier supplies this second cue without receiving any gradient from the reject objectives.

Robustness and Computational Cost. Appendices .6, .7, and .8 report further studies (Figs. 2c,d and 4). The reject-encoder depth peaks at three layers. The teleport horizon saturates once the truncated tail m^{T+1} is negligible, and longer horizons do not degrade detection because $V_{ii} \geq 1 - m$ holds at every horizon. A learned attention transition adds variance without improving detection, since the floor already holds for every row-stochastic transition. Detection changes by at most 1.1 points across the surrogate fraction and loss weights and depends much less on the classifier backbone than accuracy does. REND stays ahead of the softmax, evidential, and posterior-network baselines under every redraw of the held-out classes, and its margin holds down to five labels per class. The novelty that an in-distribution node can import from its neighbors is capped by Proposition 2, and in practice REND's false-positive rate stays below that of energy propagation in every exposure bucket. Each diffusion step is one sparse message pass, and most training runs take a few seconds on a single GPU (Appendix .4).

5 Conclusion

We have presented REND, a graph out-of-distribution detector that learns its novelty score with a dedicated reject encoder instead of reading it from a classifier. Because every score read from the logits is invariant along a subspace of the embedding, and classification training never asks any direction to separate novel nodes from familiar ones, REND trains its encoder by novelty objectives alone, with self-supervised surrogate outliers standing in for the missing OOD labels, and decouples it from the classifier down to the optimizer, so that the classifier is still trained by the cross-entropy alone. A fixed personalized-PageRank kernel then smooths the learned reject propensity, and its teleport floor keeps a fixed share of each node's own novelty at every horizon while capping the novelty that any node can borrow from its neighbors, so the receptive field can grow without washing out a concentrated anomaly. The same decoupling lets the reject branch re-anchor itself when novel nodes join the graph after training, without a single new label or any change to the classifier, and extending this label-free re-anchoring to entirely new graphs is a promising direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with implementation of the method and evaluation pipeline, refinement of hypotheses and their framing, design of the experimental protocol, and polishing the presentation and formulation of mathematical claims. All mathematical claims and AI-assisted work were reviewed and verified by the authors. The authors take responsibility for the final content of the submission.

REPRODUCIBILITY

All datasets are public and are loaded through standard libraries. The appendices specify the data processing, the model configurations, the training procedures, the baseline settings, the software and hardware, and complete proofs of the theoretical claims. Everything needed to reproduce the reported results is in the paper and in the anonymized source code submitted as supplementary material.

ETHICS STATEMENT

This work collects no new data and recruits no human subjects. The benchmarks are public graphs with anonymized features. The supplementary code reproduces the reported evaluation and gives no capability for identifying individuals in a live system.

REFERENCES

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Aleksandar Bojchevski and Stephan Günnemann. Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations (ICLR)*, 2018.

Kaidi Cao, Maria Brbić, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, 2016.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. MAGNN: Metapath aggregated graph neural network for heterogeneous graph embedding. In *The Web Conference (WWW)*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations (ICLR)*, 2019.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

Kexin Huang, Ying Jin, Emmanuel Candès, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiao Huang, Jundong Li, and Xia Hu. Label informed attributed network embedding. In *ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 731–739, 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Xixun Lin, Wenxiao Zhang, Fengzhao Shi, Chuan Zhou, Lixin Zou, Xiangyu Zhao, Dawei Yin, Shirui Pan, and Yanan Cao. Graph neural stochastic diffusion for estimating uncertainty in node classification. In *International Conference on Machine Learning (ICML)*, 2024.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking, and refining heterogeneous graph neural networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

Julian McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations (ICLR)*, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop, NeurIPS*, 2018.

Yu Song and Donglin Wang. Learning on graphs with out-of-distribution nodes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. GOLD: Graph out-of-distribution detection via implicit adversarial latent generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *The World Wide Web Conference (WWW)*, 2019.

Man Wu, Shirui Pan, and Xingquan Zhu. OpenWGL: Open-world graph learning. In *IEEE International Conference on Data Mining*, 2020.

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 (ICLR)*, 2023.

Xiao-Ming Wu, Zhenguo Li, Anthony Man-Cho So, John Wright, and Shih-Fu Chang. Learning with partially absorbing random walks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y. Chang. Network representation learning with rich text information. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Renchi Yang, Jieming Shi, Xiaokui Xiao, Yin Yang, Juncheng Liu, and Sourav S. Bhowmick. Scaling attributed network embedding to massive graphs. *Proceedings of the VLDB Endowment*, 14(1): 37–49, 2020.

Shenzhi Yang, Bin Liang, An Liu, Lin Gui, Xingkai Yao, and Xiaofang Zhang. Bounded and uniform energy-based out-of-distribution detection for graphs. In *International Conference on Machine Learning (ICML)*, 2024.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, 2003.

APPENDIX

.1 DEFERRED PROOFS

Notation. Throughout, $\tilde{\boldsymbol{P}}$ is the row-stochastic transition supported on the edges of the graph, $m \in [0, 1)$ is the diffusion rate, $\boldsymbol{c} \in (0, 1)^N$ collects the acceptance propensities of (2), $\boldsymbol{r} = (1-m)(\mathbf{1}-\boldsymbol{c})$ is the teleport source of (3), and

$$\boldsymbol{u}^{(T)} = \sum_{t=0}^{T} m^t \tilde{\boldsymbol{P}}^t \boldsymbol{r} \tag{6}$$

is the T-step truncation of the diffused mass $\boldsymbol{u}$. The vector $\mathbf{1}$ has every entry equal to one, and inequalities between vectors or matrices hold entrywise. The proofs rely on one standard fact about finite Markov chains, the fundamental convergence theorem, by which the powers of an irreducible, aperiodic, row-stochastic matrix converge to the rank-one projector onto its stationary distribution (Levin & Peres, 2017, Theorem 4.9).

PROOF OF PROPOSITION 1

Proposition 1 (Restated). *Let $\boldsymbol{z}_i = \boldsymbol{W}_c^\top \boldsymbol{h}_i$ with $\boldsymbol{W}_c \in \mathbb{R}^{d\times K}$, and let $s(\boldsymbol{z}_i)$ be any score read from the logits, such as the free energy, the maximum softmax probability, or the entropy. Then (i) s is invariant to every embedding shift $\boldsymbol{\delta} \in \ker(\boldsymbol{W}_c^\top)$, which is a subspace of dimension at least $d - K$ and hence non-trivial whenever $K < d$, so that two embeddings differing by an arbitrarily large displacement within it receive identical scores; and (ii) at every training step the cross-entropy gradient at the embedding, $\nabla_{\boldsymbol{h}_i}\mathrm{CE} = \boldsymbol{W}_c(\boldsymbol{g}_i - \boldsymbol{y}_i)$ with $\boldsymbol{y}_i$ the one-hot label, lies in the column space of the current $\boldsymbol{W}_c$, so that the loss of that step is invariant to the component of $\boldsymbol{h}_i$ in the concurrent null space and no encoder update rewards or penalizes it.*

Proof. Part (i). Take any $\boldsymbol{\delta} \in \ker(\boldsymbol{W}_c^\top)$. Then $\boldsymbol{W}_c^\top(\boldsymbol{h}_i + \boldsymbol{\delta}) = \boldsymbol{W}_c^\top \boldsymbol{h}_i = \boldsymbol{z}_i$, so the logit vector, and with it any function of the logits alone, is unchanged. The free energy $-\log\sum_k \exp(z_{ik})$, the maximum softmax probability $\max_k g_{ik}$, and the predictive entropy $-\sum_k g_{ik}\log g_{ik}$ are all compositions of fixed maps with $\boldsymbol{z}_i$.

By the rank–nullity theorem,

$$\dim\ker(\boldsymbol{W}_c^\top) = d - \mathrm{rank}(\boldsymbol{W}_c^\top) = d - \mathrm{rank}(\boldsymbol{W}_c) \geq d - K, \tag{7}$$

since the rank of a $d \times K$ matrix is at most K. Whenever $K < d$ this subspace is non-trivial, and two embeddings differing by an arbitrarily large displacement within it receive identical scores from every logit-based detector. Appending a bias, $\boldsymbol{z}_i = \boldsymbol{W}_c^\top \boldsymbol{h}_i + \boldsymbol{b}$, does not affect this argument.

Part (ii). The per-node loss is written as

$$\mathrm{CE}_i = -\log g_{iy_i} = -z_{iy_i} + \log\sum_k \exp(z_{ik}). \tag{8}$$

Its partial derivatives are the classical softmax–cross-entropy gradient,

$$\frac{\partial\,\mathrm{CE}_i}{\partial z_{ik}} = g_{ik} - \delta_{ky_i}, \qquad \nabla_{\boldsymbol{z}_i}\mathrm{CE}_i = \boldsymbol{g}_i - \boldsymbol{y}_i. \tag{9}$$

Since $\boldsymbol{z}_i$ depends on the embeddings only through $\boldsymbol{h}_i$, the chain rule yields

$$\nabla_{\boldsymbol{h}_i}\mathrm{CE} = \frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{h}_i}^\top (\boldsymbol{g}_i - \boldsymbol{y}_i) = \boldsymbol{W}_c(\boldsymbol{g}_i - \boldsymbol{y}_i), \tag{10}$$

which is a vector in the column space of $\boldsymbol{W}_c$, whose dimension is at most K, and unlabeled nodes contribute no gradient at all. The encoder parameters θ in turn receive

$$\nabla_\theta \mathrm{CE} = \frac{1}{|\mathcal{V}_L|}\sum_{i\in\mathcal{V}_L}\Big(\frac{\partial \boldsymbol{h}_i}{\partial\theta}\Big)^\top \boldsymbol{W}_c(\boldsymbol{g}_i - \boldsymbol{y}_i). \tag{11}$$

At any single step, therefore, every data-dependent training signal reaches the representation through K-dimensional projections onto the current $\mathrm{range}(\boldsymbol{W}_c)$, and the loss of that step is unchanged by

any perturbation of $\boldsymbol{h}_i$ inside the concurrent null space, which proves part (ii). Explicit regularizers such as weight decay act isotropically and independently of the data, so they supply no such signal either. One gradient does read the null-space component, namely the gradient of the head itself,

$$\nabla_{\boldsymbol{W}_c} \mathrm{CE}_i = \boldsymbol{h}_i \big(\boldsymbol{g}_i - \boldsymbol{y}_i\big)^\top. \tag{12}$$

It changes when $\boldsymbol{h}_i$ is displaced inside $\ker(\boldsymbol{W}_c^\top)$, even though the loss that it descends does not, so the displacement steers where the column space of the next step points without ever being rewarded or penalized itself. This is the head-rotation channel discussed after the proof, and it is the reason that part (ii) is stated for the loss and the encoder update rather than for the whole parameter vector. □

Scope of Proposition 1

Single-Step Scope. Part (ii) describes single steps rather than the whole trajectory, so it does not claim that the final null space is untrained. The head is trainable, and $\mathrm{range}(\boldsymbol{W}_c)$ rotates during training while the head gradient $\nabla_{\boldsymbol{W}_c}\mathrm{CE}$ reads the null-space component of every embedding. A direction in $\ker(\boldsymbol{W}_c^\top)$ at convergence may therefore have belonged to the column space earlier, since a sequence of K-dimensional subspaces can together span $\mathbb{R}^d$. The shared encoder adds a second channel, because an update $\Delta\theta$ moves every embedding by $\Delta\boldsymbol{h}_i \approx (\partial \boldsymbol{h}_i/\partial\theta)\,\Delta\theta$ and nothing confines that change to a particular subspace. Training therefore does move the final $d-K$ directions, but the loss of no step evaluates their concurrent component, and no term of the objective asks any direction to separate novel nodes from familiar ones.

Implications. By part (i), no score computed from the final logits can react to a displacement inside $\ker(\boldsymbol{W}_c^\top)$, so an OOD node displaced there receives the same energy, confidence, and entropy as an in-distribution node that differs from it only in those directions. By part (ii), nothing aligns the structure of those directions with the distinction a detector needs. Both parts are informative whenever $K < d$, which holds on every benchmark of the suite except the 96-way Facebook graph. The proposition predicts no particular detection margin, and the ablations of Section 4.2 measure what a separately trained representation contributes.

Density-Based Scores. Scores that read the whole embedding, such as the Mahalanobis distance, fall outside part (i). The proposition does not make the null-space directions inoperative for density estimation. It shows only that no encoder update rewards or penalizes them, so whether they support detection must be tested empirically. REND outperforms Mahalanobis on all twelve benchmarks, Mahalanobis falls below chance on the heterophilous Chameleon graph and to near chance on the metapath graph DBLP-H (Table 1), and the identical novelty head mounted on the detached classifier embedding loses 11.8 mean points (Section 4.2). Sub-chance Mahalanobis scores also appear outside this setup, at 45.7 and 40.4 on the Twitch-ES and Twitch-FR cross-graph splits of Wu et al. (2023), which suggests that an inverted ordering is a recurring failure mode of density scores read from the embedding of a classifier.

Well-Posedness and Truncation Error

Lemma 1 (Well-posedness and truncation error)**.** *For $0 \le m < 1$ the spectral radius satisfies $\rho(m\tilde{\boldsymbol{P}}) \le m$, so the inverse in* (3) *exists, and the truncated iterates increase monotonically to their limit with $0 \le u_i - u_i^{(T)} \le m^{T+1}$ for every node i.*

Proof. Since $\|m\tilde{\boldsymbol{P}}\|_\infty = m < 1$, the spectral radius obeys $\rho(m\tilde{\boldsymbol{P}}) \le m$, so $\boldsymbol{I} - m\tilde{\boldsymbol{P}}$ is invertible with the Neumann series

$$(\boldsymbol{I} - m\tilde{\boldsymbol{P}})^{-1} = \sum_{t\ge 0} m^t \tilde{\boldsymbol{P}}^t. \tag{13}$$

Unrolling the recursion

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{r} + m\tilde{\boldsymbol{P}}\boldsymbol{x}^{(t)}, \qquad \boldsymbol{x}^{(0)} = \boldsymbol{r}, \tag{14}$$

gives the partial sums $\boldsymbol{x}^{(T)} = \sum_{t=0}^{T} m^t \tilde{\boldsymbol{P}}^t \boldsymbol{r} = \boldsymbol{u}^{(T)}$ of (6), so the iterates of the recursion are exactly the truncations, and they increase entrywise to $\boldsymbol{u} = \boldsymbol{V}(\boldsymbol{1} - \boldsymbol{c})$ because every increment is non-negative. For the tail, $\boldsymbol{r} \le (1-m)\boldsymbol{1}$ and each $\tilde{\boldsymbol{P}}^t$ is row-stochastic, so

$$0 \le u_i - u_i^{(T)} = \sum_{t>T} m^t \big(\tilde{\boldsymbol{P}}^t \boldsymbol{r}\big)_i \le (1-m) \sum_{t>T} m^t = m^{T+1}. \tag{15}$$

The bound is absolute rather than relative, and at $T = 5$ the discarded tail is at most m^6 per node, which is below 0.016 for $m = \frac{1}{2}$. □

PROOF OF PROPOSITION 2

Proposition 2 (Restated). *Fix a graph without isolated nodes. (i) If scores are propagated by averaging over a zero-diagonal row-stochastic transition $\boldsymbol{P}$ supported on the edges,*

$$\boldsymbol{s}^{(\ell+1)} = \alpha\, \boldsymbol{s}^{(\ell)} + (1-\alpha)\, \boldsymbol{P}\boldsymbol{s}^{(\ell)}, \qquad \alpha \in (0,1), \tag{16}$$

then a single step strictly decreases the score of any node that strictly exceeds its neighbors, and as $\ell \to \infty$ the scores on each connected component converge to a component-wise constant, so that every ranking read within a component collapses to chance. (ii) If instead the reject propensity is smoothed by the personalized-PageRank kernel (3) *over any row-stochastic transition $\tilde{\boldsymbol{P}}$, then $u_i = \sum_j V_{ij}(1-c_j)$ with $\boldsymbol{V} = (1-m)(\boldsymbol{I} - m\tilde{\boldsymbol{P}})^{-1}$ row-stochastic and $V_{ii} \geq 1-m$, and the diffused mass satisfies at every horizon $T \geq 0$*

$$(1-m)(1-c_i) \;\leq\; u_i^{(T)} \;\leq\; u_i \;\leq\; \max_j (1-c_j); \tag{17}$$

moreover, whenever $\max_{j\neq i}(1-c_j) \geq 1-c_i$,

$$u_i \;\leq\; (1-m)(1-c_i) + m \max_{j\neq i} (1-c_j). \tag{18}$$

Proof. Support convention. In part (i), "supported on the edges" means that $\boldsymbol{P}_{ij} > 0$ exactly when $(i,j) \in \mathcal{E}$ and that $\boldsymbol{P}_{ii} = 0$, which matches the propagation operators that this part models. Every row must place its unit mass on a non-empty neighborhood, so the convention itself forbids isolated nodes. Part (ii) invokes only row-stochasticity, so it covers in particular the self-loop-augmented uniform walk $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{D}}^{-1}(\boldsymbol{A} + \boldsymbol{I})$ of (3).

Part (i), one-step dilution. Suppose that node i strictly dominates its neighborhood, that is, $s_i^{(\ell)} > s_j^{(\ell)}$ for all $j \in \mathcal{N}(i)$. The update at i is the convex combination

$$\begin{aligned} s_i^{(\ell+1)} &= \alpha\, s_i^{(\ell)} + (1-\alpha) \sum_{j\in\mathcal{N}(i)} \boldsymbol{P}_{ij}\, s_j^{(\ell)} \\ &\leq \alpha\, s_i^{(\ell)} + (1-\alpha) \max_{j\in\mathcal{N}(i)} s_j^{(\ell)} \;<\; s_i^{(\ell)}, \end{aligned} \tag{19}$$

where the last inequality is strict because $1-\alpha > 0$ and the maximum on the right is strictly below $s_i^{(\ell)}$, and the support convention guarantees $\mathcal{N}(i) \neq \emptyset$. The excess of the node over its neighborhood is thus eroded at every step at which it persists as a local maximum.

Part (i), collapse in the limit. The propagation unrolls to

$$\boldsymbol{s}^{(\ell)} = \boldsymbol{M}^{\ell}\boldsymbol{s}^{(0)}, \qquad \boldsymbol{M} = \alpha\boldsymbol{I} + (1-\alpha)\boldsymbol{P}, \tag{20}$$

where $\boldsymbol{M}$ is a row-stochastic matrix that moves no mass between connected components, so the limit may be taken on each component separately. Restricted to a component, the transition $\boldsymbol{P}$ is irreducible, and adding the positive diagonal preserves irreducibility while making the chain aperiodic, since $M_{ii} \geq \alpha > 0$ gives every state a self-loop. By the fundamental convergence theorem (Levin & Peres, 2017, Theorem 4.9), the restriction of $\boldsymbol{M}^{\ell}$ converges geometrically to $\mathbf{1}\boldsymbol{\pi}^{\top}$, where $\boldsymbol{\pi}$ is the unique stationary distribution of the restricted chain, and when $\boldsymbol{P}$ is moreover reversible, as the uniform walk over the edges is, the rate is governed by the second-largest eigenvalue modulus (Levin & Peres, 2017, Corollary 12.7). Moreover, the stationarity condition rearranges as

$$\boldsymbol{\pi}^{\top}\boldsymbol{M} = \boldsymbol{\pi}^{\top} \iff (1-\alpha)\big(\boldsymbol{\pi}^{\top}\boldsymbol{P} - \boldsymbol{\pi}^{\top}\big) = \mathbf{0}, \tag{21}$$

so $\boldsymbol{\pi}$ is also stationary for $\boldsymbol{P}$ itself. On every component we therefore have

$$\boldsymbol{s}^{(\ell)} \;\longrightarrow\; \big(\boldsymbol{\pi}^{\top}\boldsymbol{s}^{(0)}\big)\,\mathbf{1} \qquad (\ell \to \infty), \tag{22}$$

so that every node of the component receives the same limiting score and the initial configuration survives only through one scalar per component. Any detector that thresholds or ranks scores within

a component is then uninformative. With ties broken uniformly at random, its expected AUROC on any mixture drawn from a single component is exactly $\frac{1}{2}$, and on a disconnected benchmark such as DBLP-H all that survives the limit is the coarse ordering of whole components.

Part (ii), the teleport kernel. The teleport source of (3) is $\boldsymbol{r} = (1-m)(\mathbf{1}-\boldsymbol{c})$, so

$$\begin{aligned} \boldsymbol{u} &= (\boldsymbol{I} - m\tilde{\boldsymbol{P}})^{-1}(1-m)(\mathbf{1}-\boldsymbol{c}) = \boldsymbol{V}(\mathbf{1}-\boldsymbol{c}), \\ \boldsymbol{V} &= (1-m)\sum_{t\geq 0} m^t \tilde{\boldsymbol{P}}^t, \end{aligned} \tag{23}$$

where the series converges by Lemma 1. The matrix $\boldsymbol{V}$ has non-negative entries, and its rows sum to $(1-m)\sum_{t\geq 0} m^t = 1$ because each $\tilde{\boldsymbol{P}}^t$ is row-stochastic. Its diagonal dominates the $t=0$ term, so that $V_{ii} \geq (1-m)(\tilde{\boldsymbol{P}}^0)_{ii} = 1-m$. In probabilistic terms, consider a lazy walk that steps with probability m and halts in place with probability $1-m$. The entry

$$V_{ij} = (1-m)\sum_{t\geq 0} m^t \big(\tilde{\boldsymbol{P}}^t\big)_{ij} \tag{24}$$

is the probability that the walk launched at i halts at node j, so $\boldsymbol{V}$ is the personalized-PageRank occupancy with teleport $1-m$. Weighting each halting node by its local reject propensity $1-c_j$ and summing gives, by the law of total probability, the occupancy-weighted average $u_i = \sum_j V_{ij}(1-c_j)$. This is a convex combination of the local reject propensities, and it immediately yields the two outer bounds

$$(1-m)(1-c_i) \;\leq\; V_{ii}(1-c_i) \;\leq\; u_i \;\leq\; \max_j (1-c_j). \tag{25}$$

Since (2) places every c_j strictly inside $(0,1)$, the same pair of bounds keeps the diffused mass strictly inside the unit interval, which is what makes the logarithms of the surrogate objective (4) well defined at every step of training.

Part (ii), the borrowing cap. The same convex combination caps the mass that a node can borrow from its surroundings. Write $M_i = \max_{j\neq i}(1-c_j)$ for the strongest propensity around i. The borrowed portion of the reject mass then obeys, with no assumption on how i compares with its neighbors,

$$\sum_{j\neq i} V_{ij}(1-c_j) \;\leq\; (1-V_{ii})\, M_i \;\leq\; m\, M_i, \tag{26}$$

which is the borrowing cap of the statement. Unreachable nodes tighten it further without additional assumptions, since V_{ij} vanishes whenever no power of $\tilde{\boldsymbol{P}}$ connects i to j. The maximum may therefore be taken over only the nodes that the walk can reach, which is the form quoted in the main text. When moreover $M_i \geq 1-c_i$, splitting off the diagonal term sharpens this into the displayed bound,

$$\begin{aligned} u_i \;&=\; V_{ii}(1-c_i) + \sum_{j\neq i} V_{ij}(1-c_j) \\ &\leq\; V_{ii}(1-c_i) + (1-V_{ii})\, M_i \;=\; M_i - V_{ii}\big(M_i - (1-c_i)\big), \end{aligned} \tag{27}$$

and since $V_{ii} \geq 1-m$ and $M_i - (1-c_i) \geq 0$, the right-hand side is at most

$$M_i - (1-m)\big(M_i - (1-c_i)\big) = (1-m)(1-c_i) + m\, M_i. \tag{28}$$

Both bounds come from the self-weight alone: the inequality $V_{ii} \geq 1-m$ keeps the reject propensity of a novel node from being diluted below a $(1-m)$-share, and, because it leaves at most m for the off-diagonal weights, it also caps the novelty that an unremarkable node can import at m times the strongest propensity its walk can reach. Both guarantees therefore weaken together as $m \to 1$.

Part (ii), every horizon. Lemma 1 shows that $u_i^{(T)} = \sum_{t=0}^{T} m^t \big(\tilde{\boldsymbol{P}}^t \boldsymbol{r}\big)_i$ increases monotonically in T toward u_i, and its first term, at $t=0$, is $r_i = (1-m)(1-c_i)$. Every truncation therefore already satisfies the floor, which gives the full chain

$$(1-m)(1-c_i) \;\leq\; u_i^{(T)} \;\leq\; u_i \;\leq\; \max_j (1-c_j) \qquad \text{for every } T \geq 0. \tag{29}$$

The floor is strict, that is, $u_i > (1-m)(1-c_i)$, as soon as $m > 0$ and the walk can reach, in one or more steps, any node j with $c_j < 1$, which is the amplification by which a suspicious neighborhood raises the reject mass of a node above its own input score.

Contrast between the two parts. Under averaging, the self-weight $(\boldsymbol{M}^{\ell})_{ii}$ settles at the stationary mass π_i of its component, a quantity fixed by the graph and independent of α, typically of order $1/n$ on a component of n nodes, while the share that the operator guarantees, α^{ℓ}, decays to zero. A locally concentrated anomaly is therefore progressively averaged into its surroundings. The teleport self-weight V_{ii}, by contrast, is bounded below by $1 - m$ uniformly in the horizon, because the $t = 0$ term of the Neumann series alone contributes $1 - m$ to the diagonal. Part (ii) uses no property of $\tilde{\boldsymbol{P}}$ beyond row-stochasticity, so the guarantee holds for any transition, learned or uniform. □

SCOPE OF PROPOSITION 2

Severity of Dilution. The one-step dilution of part (i) is most severe where the neighborhood of a node disagrees with it, that is, on isolated anomalies and along heterophilous or mixed boundaries. Inside a homophilous out-of-distribution cluster the neighbors reinforce one another, so averaging, teleport smoothing, and even the raw local score behave similarly, and a shallow average can help. The guaranteed self-weight α^{ℓ} vanishes only as the depth grows, and only then does the ranking collapse toward the stationary distribution. The deployed energy-propagation baselines run two-step averages, so the collapse clause describes the depth limit of their operator and not their deployed configurations, which Section 4.2 compares with REND at a shared input score.

From Scores to Rankings. The floor bounds the *score* of each node, whereas detection metrics are functionals of the *ranking*. The two bounds of Proposition 2 together also yield a sufficient condition for the ranking. Writing $q_i = 1 - c_i$ and $M_j = \max_{k \neq j} q_k$, the order $u_i \geq u_j$ is guaranteed whenever

$$q_i - q_j \;\geq\; \frac{m}{1-m}\, M_j, \tag{30}$$

because then

$$u_i \;\geq\; (1-m)\, q_i \;\geq\; (1-m)\, q_j + m\, M_j \;\geq\; u_j. \tag{31}$$

The last step is the sharpened bound of Proposition 2, whose hypothesis holds automatically here, since the margin forces $M_j \geq q_i \geq q_j$. A pair whose local propensities differ by at least this margin therefore keeps its order under diffusion. Closer pairs are ordered by the data rather than by the bound, and Table 1 together with the exposure analysis of Appendix .8 measures that ordering directly.

.2 TRAINING ALGORITHM

Algorithm 1 lists the construction of Section 3 as a single training step. Every classifier quantity that feeds the diffusion enters through a stop-gradient $\mathrm{sg}[\cdot]$, and the classifier and reject gradients are clipped separately rather than under one joint norm, so the reject branch neither sends a gradient to the classifier nor rescales its step. The teleport smoothing is evaluated by the truncated Neumann recurrence of Lemma 1, so that each of the T steps is a single sparse message pass and the whole procedure is differentiable end to end. One epoch on the full graph is shown, and the optimizer together with every hyperparameter appearing in the algorithm is listed in Appendix .4.

.3 DATASETS AND SPLITS

The suite spans three graph families, whose statistics Table 2 reports. Wiki, BlogCatalog, Facebook, and Flickr are obtained through the packaged attributed-graph collection of Yang et al. (2020), which redistributes them in a common format.

Homophilous Graphs. Seven graphs form the homophilous family.

- **full-CiteSeer**, **full-PubMed**, **DBLP**: citation graphs (Bojchevski & Günnemann, 2018) in which papers are nodes, citations are edges, and research topics are labels.
- **Photo**: an Amazon co-purchase graph (Shchur et al., 2018) whose edges join products that are bought together.
- **Wiki**: a Wikipedia article graph (Yang et al., 2015) with dense word-based node features.
- **BlogCatalog**: a blogger social network (Huang et al., 2017) that is only mildly homophilous ($h = 0.40$) and therefore already stresses the propagation assumption from within the homophilous group.

Algorithm 1 REND Training Step

Input: graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, adjacency $\boldsymbol{A}$, features $\boldsymbol{X}$, labels y on $\mathcal{V}_L$; horizon T, outlier fraction π, weights $\lambda_{\mathrm{id}}, \lambda_{\mathrm{pse}}$, clip κ
Parameters: classifier $f_\theta, \boldsymbol{W}_c$; reject encoder ψ_ξ; novelty head g_ϕ; scalars τ, b, η
Classifier forward
$\boldsymbol{H} \leftarrow f_\theta(\boldsymbol{X}, \boldsymbol{A}),\ \boldsymbol{Z} \leftarrow \boldsymbol{H}\boldsymbol{W}_c,\ \boldsymbol{g}_i \leftarrow \mathrm{softmax}(\boldsymbol{z}_i)$
$\epsilon_i \leftarrow -\log \sum_k \exp(z_{ik})$ ▷ free energy
Reject branch: affinity and propensity (Eqs. 1 and 2)
$\boldsymbol{e} \leftarrow \psi_\xi(\boldsymbol{X}, \boldsymbol{A}),\ \nu_i \leftarrow g_\phi(\boldsymbol{e}_i)$
$a_i \leftarrow \zeta(-\mathrm{sg}[\epsilon_i]) + \zeta(-\nu_i)$, standardized over $\mathcal{V}_L$
$m \leftarrow \sigma(\eta),\ c_i \leftarrow \sigma(\tau(a_i - \mathrm{med}(a) - b))$
Teleport smoothing: truncated Neumann iteration (Eq. 3)
$\tilde{\boldsymbol{P}} \leftarrow \tilde{\boldsymbol{D}}^{-1}(\boldsymbol{A} + \boldsymbol{I}),\ r_i \leftarrow (1-m)(1-c_i)$ ▷ uniform walk, fixed
$\boldsymbol{u} \leftarrow \boldsymbol{r}$
for $t = 1$ **to** T **do**
    $\boldsymbol{u} \leftarrow \boldsymbol{r} + m\tilde{\boldsymbol{P}}\boldsymbol{u}$
end for
Self-supervised outlier exposure (Eq. 4)
sample $\mathcal{C} \subset \mathcal{V}$, $|\mathcal{C}| = \lfloor \pi N \rfloor$; form $\boldsymbol{X}'$ by swapping features of $\mathcal{C}$ with random nodes
recompute $\boldsymbol{u}'$ by a full forward pass on $\boldsymbol{X}'$ ▷ ϵ_i' recomputed under $\boldsymbol{X}'$, entering through $\mathrm{sg}[\cdot]$; gradient reaches the reject branch only
$\mathcal{L}_{\mathrm{pse}} \leftarrow -\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \log u_i' - \frac{1}{|\mathcal{V}_L \setminus \mathcal{C}|}\sum_{i\in\mathcal{V}_L\setminus\mathcal{C}} \log(1-u_i')$
Objective and decoupled update (Eq. 5)
$\mathcal{L} \leftarrow \mathrm{CE}(\boldsymbol{Z}_{\mathcal{V}_L}, y) + \lambda_{\mathrm{id}} \frac{1}{|\mathcal{V}_L|}\sum_{i\in\mathcal{V}_L} u_i + \lambda_{\mathrm{pse}}\mathcal{L}_{\mathrm{pse}}$
backpropagate $\nabla\mathcal{L}$
clip $\{\theta, \boldsymbol{W}_c\}$ and reject parameters to norm κ *separately*; optimizer step
Inference
return OOD score u_i; label $\hat{y}_i = \arg\max_k g_{ik}$

- **Facebook**: the union of ten ego-networks whose nodes are user profiles joined by friendship (McAuley & Leskovec, 2012) and labeled by hand-annotated social circles. The circles overlap in the original release, so each node takes the lowest-indexed circle to which it belongs, and the 28% of nodes that belong to no circle stay unlabeled rather than forming a class of their own. This leaves 191 circle indices, of which 151 are populated. The split keeps the 96 lowest indices as in-distribution and withholds the 95 highest, and 77 of the kept and 74 of the withheld indices carry nodes. Facebook is thus by far the most class-rich benchmark of the suite on either count, and the only one whose classifier head is wider than the 64-dimensional embedding that it reads.

Heterophilous Graphs. Two graphs probe the regime in which neighbors mostly disagree.

- **Chameleon**: a Wikipedia page graph (Pei et al., 2020) with edge homophily $h = 0.23$, so that most neighbors carry a different label and any detector that trusts neighborhood agreement is misled. We use the standard preprocessed version. The benchmark has been criticized for duplicate nodes shared between the train and test folds (Platonov et al., 2023). We retain it for comparability with the graph out-of-distribution literature that reports on it, and every heterophilous conclusion is also read off Flickr, which carries no such defect.
- **Flickr**: an image-sharing social graph (Huang et al., 2017) with edge homophily $h = 0.24$. Its neighborhoods are as label-mixed as those of Chameleon, but its dense 12,047-dimensional attributes make it a distinct regime, in which the local features are informative while most edges join nodes of different classes.

Heterogeneous Graphs. A metapath collapses several edge types into one homogeneous graph, which tests a different kind of edge semantics without changing the model. We homogenize three graphs through the metapaths of heterogeneous graph attention networks (Wang et al., 2019), taking IMDB and DBLP-H as released by metapath aggregated graph neural networks (Fu et al., 2020) and ACM from the heterogeneous graph benchmark (Lv et al., 2021). IMDB uses the full standard pair,

Dataset	Nodes	Edges	Feat.	Cls.	h
full-CiteSeer	4,230	5,337	602	6	0.95
full-PubMed	19,717	44,324	500	3	0.80
DBLP	17,716	52,867	1,639	4	0.83
Photo	7,650	119,081	745	8	0.83
Wiki	2,405	11,596	4,973	17	0.61
BlogCatalog	5,196	171,743	8,189	6	0.40
Facebook	4,039	88,234	1,283	191	0.65
Chameleon	2,277	31,371	2,325	5	0.23
Flickr	7,575	239,738	12,047	9	0.24
ACM	3,025	13,374	1,902	3	0.79
IMDB	4,278	46,293	3,066	3	0.47
DBLP-H	4,057	3,528	334	4	0.80

Table 2: The twelve node-classification benchmarks. Edges are undirected without self-loops, and h is the edge homophily measured on the complete label set before the leave-out split.

movie-actor-movie and movie-director-movie. ACM and DBLP-H keep only the sparse metapath, paper-author-paper and author-paper-author respectively, because the remaining standard metapaths pass through a low-cardinality node type, 56 subjects on ACM and 20 conferences on DBLP-H, and would join every paper or author sharing that attribute into a clique. Admitting them would raise both graphs above one million edges and turn the benchmark into a test of dense propagation rather than of edge semantics.

- **ACM**: a paper graph whose metapath restores high label agreement ($h = 0.79$).
- **IMDB**: a movie graph in which the metapath leaves label agreement low ($h = 0.47$), which probes robustness to weaker propagation.
- **DBLP-H**: an author graph whose metapath again yields high homophily ($h = 0.80$).

Held-Out Classes. The leave-out-classes rule of Section 4 is more uniform than either released convention. Zhao et al. (2020) draw the held-out categories at random and vary their number by dataset, and the graph posterior network withholds the last classes by index but hard-codes how many for each dataset (Stadler et al., 2021), a count running from roughly a quarter to a half of the label set. We keep the deterministic last-by-index choice and fix the count at $\lfloor K/2 \rfloor$ everywhere, so that no graph receives a tailored split. The held-out classes leave the training and validation masks, while their nodes stay in the graph and form the out-of-distribution test pool. Table 3 lists the resulting composition of every split.

Labeled Budget. The homophilous graphs and Flickr carry no canonical split, so we follow common practice (Shchur et al., 2018) and sample 20 labeled nodes per class for training, or the whole class when it holds fewer, together with 500 validation and 1,000 test nodes. This split is drawn afresh for every run, so the reported spread includes split variance as well as initialization variance. Chameleon uses its released split, and the metapath graphs use the fixed splits released with their preprocessing (Fu et al., 2020; Lv et al., 2021).[1] The leave-out shift is then imposed on top of whichever split a dataset uses. The fallback to the whole class matters most on Facebook, whose median populated circle holds six nodes, so its labeled set is far smaller than 20 nodes for each of its 96 in-distribution classes. The resulting budget is small and identical for every method, ranging from a few tenths of a percent of the in-distribution nodes on the large randomly split graphs up to the larger native fixed splits.

.4 Implementation Details and Hyperparameters

Shared Configuration. Table 4 lists the complete REND configuration, which is held fixed across all twelve datasets, so that the reported numbers reflect a single configuration rather than a per-dataset sweep. The values it fixes for the surrogate fraction and the two loss weights are not delicate, since

[1]The heterogeneous benchmark withholds the true test labels behind its leaderboard for official submissions. The files that it distributes for local reproducibility carry the real labels, and those are the ones used here, as in the prior work reporting on these graphs.

Dataset	ID Cls.	OOD Cls.	Train	Val	Test (ID)	OOD nodes	Train %ID
Homophilous							
full-CiteSeer	3	3	60	261	492	2,084	2.8
full-PubMed	2	1	40	291	593	7,875	0.3
DBLP	2	2	40	382	773	4,151	0.3
Photo	4	4	80	234	501	3,977	2.2
Wiki	9	8	170	299	635	911	11.4
BlogCatalog	3	3	60	240	491	2,613	2.3
Facebook	96	95	612	166	295	1,795	27.3
Heterophilous							
Chameleon	3	2	663	439	267	908	48.4
Flickr	5	4	100	263	517	3,490	2.4
Heterogeneous (metapath)							
ACM	2	1	452	121	1,391	1,061	23.0
IMDB	2	1	259	242	2,218	1,559	9.5
DBLP-H	2	2	186	195	1,561	2,115	9.6

Table 3: Composition of the leave-out-classes split (first run). The highest-indexed half of the classes is held out; the last column gives the labeled budget as a fraction of the ID nodes.

Hyperparameter	Value
Architecture	
classifier encoder f_θ	GraphSAGE, 2 layers
reject encoder ψ_ξ	mean aggregator, $L_r = 3$ layers
hidden width (both branches)	64
dropout	0.5
transition $\tilde{\boldsymbol{P}}$	uniform walk $\tilde{\boldsymbol{D}}^{-1}(\boldsymbol{A} + \boldsymbol{I})$, fixed
Diffusion and objective	
teleport horizon T	5
diffusion rate $m = \sigma(\eta)$	learned; init $m_0 = 0.5$ ($\eta_0 = 0$)
affinity steepness τ	learned; init $\tau_0 = 2$
affinity offset b	learned; init $b_0 = 0$
surrogate outlier fraction π	0.25
loss weights λ_{id}, λ_{pse}	1, 1
per-branch gradient clip κ	5
Optimization	
optimizer	Adam
learning rate	0.01
weight decay	5×10^{-4}
epochs (early-stopping patience)	300 (50)
independent runs per result	10

Table 4: The single REND configuration used on all twelve datasets. No hyperparameter is tuned per dataset.

varying them changes mean detection by at most 1.1 points (Appendix .7). Both branches are trained jointly with Adam at a learning rate of 0.01 and a weight decay of 5×10^{-4} for up to 300 epochs, with early stopping on validation accuracy at a patience of 50 epochs. Since the reject branch is decoupled from the classifier down to the optimizer, the gradient norm of each branch is clipped separately to five instead of clipping their combined gradient, and Appendix .7 measures what a shared clip would cost.

Learned Diffusion Scalars. The diffusion rate and the affinity steepness start from the same values on every dataset, and Table 5 reports where they settle. The rate remains within $[0.44, 0.53]$, which keeps the floor $1 - m$ of Proposition 2 between 0.47 and 0.56 on every graph and far from the degenerate regime $m \to 1$ in which both of its guarantees would vanish. The steepness changes only slightly from its initial value and remains within $[1.73, 2.08]$.

Dataset	Move rate m	Floor $1-m$	Steepness τ
full-CiteSeer	0.49 ± 0.04	0.51	1.90 ± 0.09
full-PubMed	0.44 ± 0.09	0.56	2.04 ± 0.21
DBLP	0.48 ± 0.04	0.52	1.90 ± 0.10
Photo	0.52 ± 0.04	0.48	1.79 ± 0.10
Wiki	0.48 ± 0.08	0.52	1.96 ± 0.22
BlogCatalog	0.51 ± 0.02	0.49	1.92 ± 0.05
Facebook	0.53 ± 0.02	0.47	1.73 ± 0.10
Chameleon	0.48 ± 0.06	0.52	2.08 ± 0.23
Flickr	0.50 ± 0.03	0.50	1.92 ± 0.04
ACM	0.53 ± 0.01	0.47	1.87 ± 0.03
IMDB	0.52 ± 0.00	0.48	1.90 ± 0.02
DBLP-H	0.53 ± 0.00	0.47	1.85 ± 0.02

Table 5: Learned walk scalars (mean $\pm$ standard deviation over ten runs), initialized identically on every dataset ($m_0 = 0.5$, $\tau_0 = 2$).

Baseline Implementations. The energy-scored softmax classifier, its Mahalanobis post-hoc variant, and the evidential network share the identical two-layer GraphSAGE encoder and the optimization settings given above, so they differ from the classifier branch of REND only in how the uncertainty is read out. GNNSafe, NODESAFE, GPN, GOLD, and OODGAT instead keep the method-defining hyperparameters of their released code. These are the per-class radial flows and personalized-PageRank diffusion of GPN at its released teleport of 0.1 over ten power iterations; the bounded-and-uniform regularizer of NODESAFE, run in its exposure-free form at a single weight across the suite because no method here is given real out-of-distribution data; the adversarial latent generator and energy-margin detector of GOLD under one fixed schedule for every dataset; and the attention layer and objective of OODGAT at its released width and edge dropout, scored by its outlierness head. The ports follow the code released at https://github.com/qitianwu/GraphOOD-GNNSafe, https://github.com/ShenzhiYang2000/NODESAFE, https://github.com/stadlmax/Graph-Posterior-Network, https://github.com/DannyW618/GOLD, and https://github.com/SongYYYY/KDD22-OODGAT. The shared protocol forbids any test-set model selection, so every trained baseline selects its epoch by in-distribution validation accuracy, exactly as REND does. The evidential row uses the Bayes-risk cross-entropy of Sensoy et al. (2018), which scored 2.7 AUROC above the sum-of-squares form recommended by those authors, so the stronger of the two variants is the one reported.

Mahalanobis Covariance. Each node is scored by its squared distance to the nearest class Gaussian fitted on the labeled nodes of the final embedding, under a covariance tied across classes of which only the diagonal is retained. Released graph ports of the method of Lee et al. (2018) instead invert the full tied covariance. We ran that variant as well, and it is weaker on ten of the twelve graphs and by 4.9 AUROC on average. Part of that gap is the sample-size effect described by the classical analysis of Bickel & Levina (2004), since four of the splits label fewer than 65 nodes against a 64-dimensional embedding and the pooled scatter is then singular. The gap is widest, however, on Chameleon, where 663 labeled nodes leave the covariance well determined and the full-covariance score nonetheless falls from 40.7 to 23.4, which suggests that once the ordering of a density score has inverted, fitting the off-diagonal structure more faithfully only sharpens the inversion. We therefore report the diagonal variant, the stronger of the two. As in every graph port of this score, both variants read the penultimate layer alone, without the layer-wise ensemble and input perturbation of the original method, whose combination weights are fitted on outlier validation samples that the present protocol withholds.

Computational Cost. Training takes a few seconds per run on a single GPU for most of the suite, a small constant factor over the plain softmax network that is dominated by the extra forward passes of the reject encoder and by the $T = 5$ diffusion steps. Inference is a single forward pass, since the truncated Neumann iteration is unrolled inside the model. All reported numbers are means and standard deviations over ten independent runs with different random initializations, and the code and configuration files will be released.

Dataset	Softmax	REND	Δ
full-CiteSeer	85.2 ± 1.4	85.5 ± 1.6	$+0.3$
full-PubMed	93.1 ± 1.4	92.9 ± 1.3	-0.3
DBLP	86.0 ± 2.0	86.0 ± 1.9	$+0.0$
Photo	93.8 ± 1.0	93.7 ± 1.0	-0.1
Wiki	74.8 ± 2.2	75.2 ± 1.4	$+0.3$
BlogCatalog	94.1 ± 1.0	93.6 ± 1.4	-0.6
Facebook	66.8 ± 5.5	70.0 ± 6.3	$+3.1$
Chameleon	53.4 ± 2.0	53.2 ± 2.1	-0.2
Flickr	66.5 ± 3.2	66.0 ± 4.1	-0.5
ACM	96.6 ± 0.6	96.0 ± 0.5	-0.6
IMDB	79.6 ± 0.5	79.4 ± 0.8	-0.2
DBLP-H	90.9 ± 0.1	91.0 ± 0.2	$+0.1$
Mean	81.7	81.9	$+0.1$

Table 6: In-distribution test accuracy (%, mean $\pm$ standard deviation over ten runs) of the plain softmax network and of REND. The mean gap of $+0.1$ is not significant ($p = 0.70$, paired t-test).

Dataset	GCN	Softmax	EDL	GPN	NODESAFE	GOLD	REND
full-CiteSeer	82.9 ± 3.6	85.9 ± 2.8	**86.0** ± 2.3	85.6 ± 2.8	84.5 ± 2.1	84.7 ± 3.4	85.5 ± 2.7
full-PubMed	86.8 ± 2.8	89.7 ± 2.0	**90.3** ± 2.1	89.0 ± 2.4	88.1 ± 2.8	82.8 ± 4.5	89.1 ± 2.6
DBLP	79.5 ± 4.2	76.7 ± 3.2	78.8 ± 3.3	**80.8** ± 4.0	78.4 ± 2.5	78.1 ± 3.0	77.7 ± 2.2
Photo	77.4 ± 5.6	84.0 ± 9.3	83.6 ± 11.8	83.9 ± 3.9	86.4 ± 3.9	81.1 ± 3.9	**86.6** ± 4.4
Wiki	76.0 ± 1.9	80.4 ± 2.9	78.5 ± 4.3	79.9 ± 3.4	81.3 ± 2.0	75.0 ± 4.2	**83.2** ± 2.1
BlogCatalog	73.2 ± 3.9	83.0 ± 3.9	**88.2** ± 3.8	73.6 ± 10.1	80.3 ± 5.1	72.9 ± 5.5	87.6 ± 3.6
Facebook	73.3 ± 10.6	77.9 ± 8.6	54.7 ± 17.0	79.2 ± 3.7	71.3 ± 13.6	**84.7** ± 2.5	73.2 ± 13.8
Chameleon	56.8 ± 3.3	56.4 ± 3.4	56.6 ± 3.0	57.2 ± 2.8	56.4 ± 3.0	54.9 ± 2.1	**57.5** ± 3.5
Flickr	63.3 ± 5.4	71.6 ± 2.5	72.9 ± 5.1	67.9 ± 5.4	73.8 ± 2.8	73.1 ± 3.2	**75.2** ± 2.9
ACM	88.3 ± 0.5	**92.8** ± 0.7	92.5 ± 1.0	87.9 ± 3.4	92.4 ± 0.8	86.4 ± 2.2	92.8 ± 1.0
IMDB	68.5 ± 0.6	71.0 ± 1.8	**73.4** ± 0.5	67.6 ± 1.4	72.4 ± 0.7	66.4 ± 1.5	71.3 ± 1.3
DBLP-H	76.0 ± 1.3	80.1 ± 1.4	78.3 ± 1.9	73.8 ± 3.5	**81.9** ± 1.1	74.5 ± 1.2	79.4 ± 1.1
Mean	75.2	79.1	77.8	77.2	78.9	76.2	**79.9**

Table 7: Misclassification-detection AUROC (%, mean $\pm$ standard deviation over ten runs) for every method that trains its own classifier; the best per dataset is in bold.

.5 Accuracy and Misclassification Detection per Dataset

Accuracy. Table 6 breaks the accuracy claim of Section 4.1 down by benchmark, comparing the in-distribution test accuracy of REND with that of the plain softmax GraphSAGE network. Since the classifier of REND receives gradients from the cross-entropy alone and is clipped separately, this comparison is between two trainings of the same network on the same objective. The reject branch can influence the classifier only through the random stream that its dropout and corrupted forward passes consume, so the per-dataset gaps are run-to-run noise rather than a systematic cost. Consistent with this, they scatter in both directions with a mean of $+0.1$ against a mean run-to-run spread of 1.9. The one large entry, the $+3.1$ on Facebook, also carries the widest run-to-run spread in the table, five to six points in either column, on a 96-way task whose classes average six labeled examples. On that graph the particular examples that a split supplies matter more than the choice between the two networks, and both readouts, far above the 1% chance rate, agree to within that spread.

Misclassification Detection. Table 7 reports misclassification detection, in which each model ranks its own in-distribution test errors by its own confidence, for every method that trains its own classifier. GNNSafe and Mahalanobis are post-hoc scores on the softmax classifier, so the Softmax column already reports their confidence. REND attains the best suite mean, plotted in Fig. 2a, and also the flattest profile in the comparison. It comes within two points of the per-dataset winner on eight of the twelve graphs, and its average distance to that winner, 1.8 points, is the smallest in the comparison, against 2.6 for the plain softmax network and 2.8 for NODESAFE, the two methods that come closest.

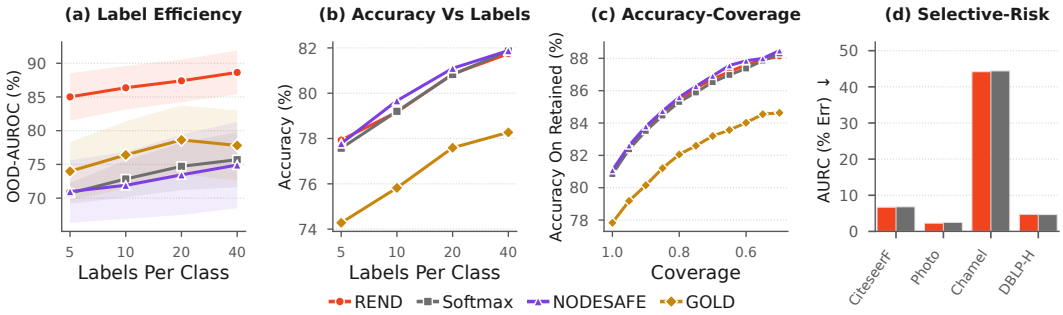

Figure 3: Practical robustness: OOD detection (a) and accuracy (b) versus labels per class, accuracy–coverage (c) averaged over the representative datasets, and per-dataset AURC (d).

	Labels per class			
Method	5	10	20	40
OOD-AUROC (%)				
Softmax	70.7	72.8	74.7	75.7
NODESAFE	71.0	71.9	73.4	74.9
GOLD	74.0	76.4	78.6	77.8
REND	**85.0**	**86.4**	**87.4**	**88.6**
Accuracy (%)				
Softmax	77.6	79.2	80.8	81.9
NODESAFE	77.8	79.7	81.1	81.9
GOLD	74.3	75.8	77.6	78.3
REND	77.9	79.2	80.8	81.7

Table 8: Label efficiency: out-of-distribution detection and in-distribution accuracy (%, mean over four datasets and ten runs) as the number of labeled examples per class varies.

.6 Label Efficiency and Selective Prediction

Label Efficiency. Fig. 3a,b shows the detection and the accuracy as the labeled set shrinks from 40 to 5 examples per class, and Table 8 lists the means behind the two panels. The detection score of REND remains at least eight points above the strongest baseline at every budget, and it is eleven points ahead at the smallest budget, where it still exceeds the score of every baseline at *any* budget. Its accuracy tracks the plain softmax network throughout, and only the adversarial training of GOLD visibly trades accuracy for its synthetic exposure. The surrogate corrupts mostly unlabeled nodes, so its training signal does not shrink with the label budget.

Selective Prediction. Detection asks whether a node is foreign, whereas selective prediction asks whether the answer of the model on a familiar node can be trusted, so it reveals whether a gain in AUROC was paid for with the confidence of the classifier. Abstaining on the least confident in-distribution test nodes traces a mean accuracy–coverage curve that lies above that of the softmax classifier and on a par with the strongest baselines, and it lowers the per-dataset area under the risk–coverage curve (Geifman & El-Yaniv, 2017) of the softmax classifier on three of the four datasets (Fig. 3c,d). The decoupling rules out such a trade-off, since the confidence that ranks familiar nodes comes from a classifier trained by the cross-entropy alone.

.7 Additional Ablations and Robustness Studies

Reject-Encoder Depth. Fig. 2c sweeps ψ_ξ from one to five layers on a five-dataset core suite (full-CiteSeer, full-PubMed, Wiki, Chameleon, and Flickr), which also carries the horizon and backbone studies below. Detection peaks at three layers and decreases beyond it. The gap between three and four layers lies well inside the per-dataset run-to-run spread of approximately 2.5 points, so the cheaper three-layer default is retained.

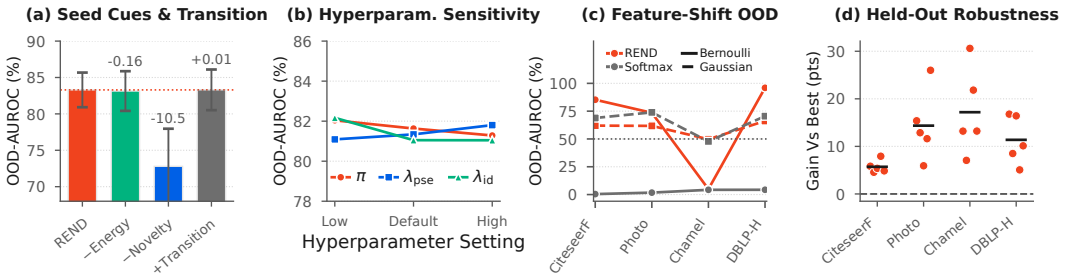

Figure 4: Further ablations: score cues and transition (a), hyperparameter sensitivity (b), feature-corruption shift (c), and held-out-class robustness (d).

Teleport Horizon. Retraining at horizons from one to ten steps on the same suite shows detection saturating at approximately five steps, with no change beyond run-to-run fluctuation at longer horizons (Fig. 2d). The saturation follows from the truncated tail, which is at most m^{T+1} (Lemma 1) and about 0.02 at $T = 5$ for the learned rates of Table 5, and longer horizons do not erode a node's own propensity, because the teleport self-weight V_{ii} stays above $1 - m$ at every depth. The horizon therefore needs no per-graph tuning, and the truncated Neumann iteration can stay shallow.

Gradient Clipping. Clipping the two branches separately, rather than under a single global norm, raises detection by approximately 2.5 points on full-CiteSeer, where the surrogate is strongest, and by two points on Flickr, while also slightly improving the suite mean. The surrogate produces bursty gradients, and a shared norm lets those bursts shrink the classifier's step, which reintroduces at the optimizer the coupling that the stop-gradient removes from the computation graph.

Transition Matrix. Adding a learned attention transition on top of the uniform walk, with every other component held identical, changes mean detection by one hundredth of a point and raises the mean run-to-run spread from 2.4 to 2.8 points. The learned transition is the less stable choice on eight of the twelve graphs, and it inflates the spread on ACM from 0.8 to 2.6 (Fig. 4a). Since the floor of Proposition 2 holds for every row-stochastic transition, the fixed walk obtains the same guarantee with fewer parameters and more stable training.

Hyperparameter Sensitivity. Sweeping the surrogate fraction π over $\{0.1, 0.25, 0.5\}$ and either loss weight over a fourfold range changes mean detection by at most 1.1 points (Fig. 4b). The fraction π sets how many nodes are corrupted per step, not how strongly each one is perturbed, and the two weights scale reject terms that share no gradient with the cross-entropy, so no setting of them trades detection against classification.

Choice of Held-Out Classes. The identity of the held-out classes could favor one detector over another, so we repeat the evaluation on four datasets with five halves of the label set, namely the fixed split and four random draws. The detection score of REND remains above the softmax, evidential, and posterior-network baselines on every draw of every graph (Fig. 4d). Averaged over the four graphs, its margin on the fixed split (12.8 points) is close to its mean margin over the random draws (12.0), so the high-index split of Table 1 does not favor REND.

Feature-Corruption Shift. Every other experiment removes semantic classes, a shift that the surrogate already resembles, so we also corrupt the features of scattered nodes, which realizes the isolated-anomaly regime of Proposition 2. Under Bernoulli corruption, random binary features drive the energy score to overconfidence and softmax detection falls below chance on every graph, whereas the reject readouts of REND still isolate the corrupted nodes at a mean AUROC of 65 (Fig. 4c). Under Gaussian corruption, which leaves the energy score intact, the local reject readout reaches 63, close to the 65 of the energy score.

Classifier Backbone. On the core suite, GraphSAGE is the strongest classifier backbone, at 81.7 mean detection and 74.5 accuracy against 77.7 and 67.0 for GCN. Detection also depends much less on the backbone than accuracy does, since a graph-free MLP classifier or a fixed-propagation APPNP classifier (Gasteiger et al., 2019) gives up nearly six accuracy points yet at most one and a half points of detection. The reject branch reads the graph through ψ_ξ and not through the classifier, so detection inherits only part of the classifier's architectural weaknesses.

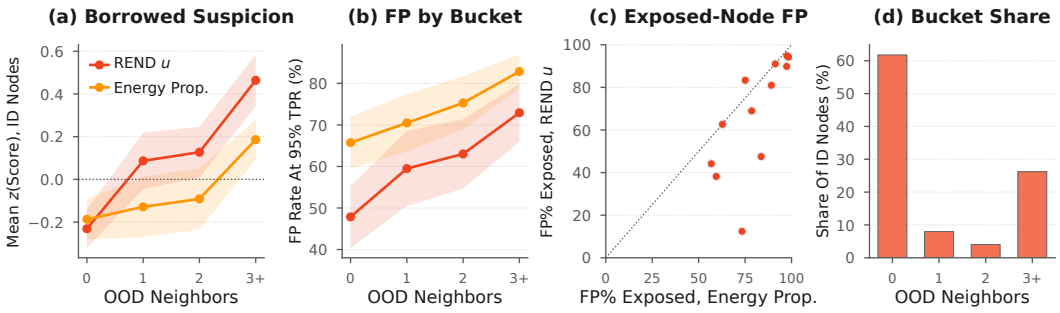

Figure 5: In-distribution test nodes stratified by out-of-distribution neighbor count: mean standardized score (a), FPR at 95% TPR (b), per-dataset FPR on exposed nodes (c), and bucket populations (d).

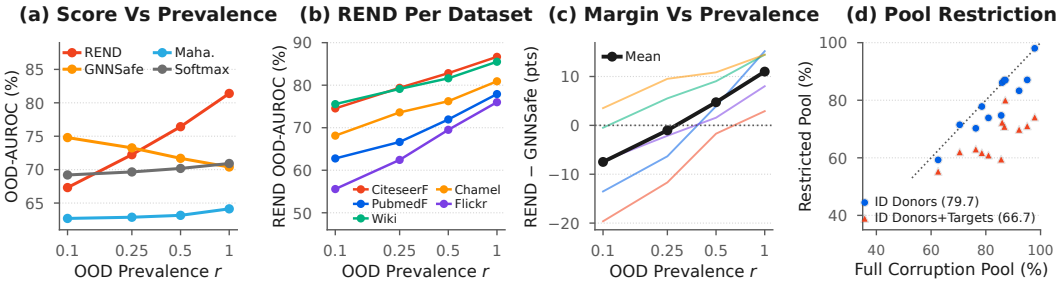

Figure 6: Transductive access on the core suite: OOD-AUROC versus the retained held-out-class fraction r, averaged (a) and per dataset (b); margin over GNNSafe (c); restricted surrogate pools (d).

.8 In-Distribution Nodes Exposed to Novelty

Effect of Novel Neighbors. Under the clustered shift, a quarter of the in-distribution test nodes have three or more out-of-distribution neighbors (Fig. 5d), so the borrowing that Proposition 2 bounds can be measured directly. The standardized reject score climbs from -0.23 in the unexposed bucket to $+0.46$ in the bucket with three or more out-of-distribution neighbors, a steeper rise than that of energy propagation, as expected from a diffusion that propagates the reject propensity itself (Fig. 5a).

False Positive Analysis. The m-cap of Proposition 2 bounds how much novelty such a node can import, and REND keeps its advantage even among the most exposed nodes. At a shared 95%-true-positive threshold, its false-positive rate remains below that of energy propagation in every bucket, and on eleven of the twelve graphs once the exposed nodes are pooled (Fig. 5b,c). Over the whole suite, REND also lowers the mean FPR95 of Table 1 from the 71.0% of the strongest baseline to 55.2%.

.9 Transductive and Inductive Settings

Transductive Setting. The protocol of Section 4 is transductive, because the held-out classes lose their labels rather than their nodes. The reject encoder, the corruption pool, the affinity midpoint $\mathrm{med}(a) + b$, and the teleport-anchored diffusion therefore all read a graph that already contains the ID–OOD test mixture. Every compared method shares this single-graph leave-out-classes protocol, which the literature treats as standard (Zhao et al., 2020; Stadler et al., 2021) and which inherits the unlabeled mixture of open-world semi-supervised learning (Cao et al., 2022). We measure what this access contributes to REND by withdrawing it in four stages. The surrogate generator is first denied genuine held-out-class feature donors and then unlabeled corruption sites; training is next confined to the ID-induced subgraph; and finally the novel population itself is thinned. The last two stages probe two deployment conditions, a mixture that arrives only after training and a novel population that is rare.

Restricted Surrogate Pools. The first restriction draws feature donors only from the labeled ID nodes, so that no genuine held-out-class vector can be copied into a neighborhood, while corruption

	Transductive	ID-Induced Subgraph Training				
Dataset	REND	REND	REND+adapt	Softmax	GNNSafe	Mahalanobis
full-CiteSeer	86.8 $\pm$ 1.4	67.4 $\pm$ 9.9	79.2 $\pm$ 4.8	80.2 $\pm$ 3.5	**82.7** $\pm$ 3.4	**82.7** $\pm$ 1.8
full-PubMed	78.5 $\pm$ 4.1	65.6 $\pm$ 6.8	**75.2** $\pm$ 4.2	60.6 $\pm$ 5.8	61.9 $\pm$ 6.0	62.9 $\pm$ 4.3
DBLP	70.5 $\pm$ 4.8	61.2 $\pm$ 4.4	64.9 $\pm$ 5.7	66.6 $\pm$ 2.0	**71.4** $\pm$ 2.8	66.2 $\pm$ 2.1
Photo	97.9 $\pm$ 0.4	78.7 $\pm$ 8.6	**97.7** $\pm$ 0.5	88.9 $\pm$ 3.1	90.8 $\pm$ 3.2	87.7 $\pm$ 4.4
Wiki	86.0 $\pm$ 3.5	69.2 $\pm$ 7.8	79.6 $\pm$ 3.6	**79.8** $\pm$ 2.5	73.9 $\pm$ 1.7	70.5 $\pm$ 4.3
BlogCatalog	92.2 $\pm$ 2.0	71.1 $\pm$ 7.3	**87.4** $\pm$ 1.7	71.7 $\pm$ 11.1	75.9 $\pm$ 11.8	83.3 $\pm$ 5.5
Facebook	87.1 $\pm$ 1.8	71.8 $\pm$ 7.4	**87.7** $\pm$ 2.2	71.8 $\pm$ 5.0	71.8 $\pm$ 7.5	73.4 $\pm$ 5.5
Chameleon	81.0 $\pm$ 2.0	61.2 $\pm$ 7.5	**80.3** $\pm$ 2.2	69.7 $\pm$ 5.8	66.4 $\pm$ 4.4	41.9 $\pm$ 6.4
Flickr	76.3 $\pm$ 2.1	46.6 $\pm$ 7.2	56.2 $\pm$ 4.1	**74.1** $\pm$ 4.8	72.3 $\pm$ 5.1	67.4 $\pm$ 6.4
ACM	95.1 $\pm$ 0.8	76.9 $\pm$ 8.9	**93.8** $\pm$ 1.7	73.9 $\pm$ 3.4	75.9 $\pm$ 3.7	72.3 $\pm$ 3.4
IMDB	62.6 $\pm$ 1.6	53.2 $\pm$ 2.0	57.2 $\pm$ 1.4	56.5 $\pm$ 2.4	57.8 $\pm$ 2.8	**59.5** $\pm$ 1.1
DBLP-H	85.6 $\pm$ 4.1	58.2 $\pm$ 4.4	**82.3** $\pm$ 1.6	67.5 $\pm$ 0.9	69.0 $\pm$ 1.1	52.8 $\pm$ 1.2
Mean	83.3	65.1	**78.5**	71.8	72.5	68.4

Table 9: OOD-detection AUROC (%, mean $\pm$ std over ten runs) under ID-induced subgraph training with full-graph evaluation; the transductive column repeats Table 1.

still occurs everywhere in the graph (Fig. 6d). REND then reaches a mean of 79.7, still 6.6 points above the strongest baseline of Table 1, so its margin does not rest on copying genuine OOD features, and the citation and co-purchase graphs lose less than one point against their unrestricted scores. The second restriction also confines the corruption *targets* to the labeled set, a quarter of which is only 15 positions per step on full-CiteSeer, and the resulting shortfall in volume and placement diversity lowers the mean to 66.7. The surrogate therefore draws its signal from corruption *sites* spread across the deployment graph far more than from genuine OOD *feature vectors*.

Training on the ID-Induced Subgraph. In the strict variant, every training forward pass, the classification loss, both reject losses, the surrogate-outlier generator, and the validation pass that drives early stopping run on the subgraph induced by the ID nodes, so that the held-out classes contribute neither features nor edges to any gradient. The trained model is then placed on the full graph and evaluated exactly as before. The OOD population thus arrives only after training ends, which is the single-graph analogue of the multi-graph control that GNNSafe runs beside its transductive benchmarks (Wu et al., 2023). Table 9 reports the outcome next to the transductive reference, together with a softmax classifier trained under the same restriction and scored post hoc with its free energy, with the propagated energy of GNNSafe, and with Mahalanobis distances. The scale, midpoint, and surrogate outliers of the strictly trained branch are calibrated on a smaller graph than the one on which it is evaluated, which leaves it at a mean of 65.1 as trained, and because the branch is decoupled from the classifier, it can be recalibrated on the grown graph without any OOD label, as described next.

Label-Free Re-Anchoring. Every learned reject component is optimized by the reject losses alone and touches the classifier only through stop-gradients, so the reject branch can resume training on the deployment graph as it stands, with no OOD label and no retraining of the classifier. This variant, listed as REND+adapt in Table 9, runs 100 steps of the same label-free objective with the classifier frozen, the anchor on the unchanged labeled ID nodes, and surrogate outliers produced by corruption, and it retains the iterate of lowest smoothed adaptation loss. It is a test-time *transductive* adaptation that reads the deployment mixture, not ordinary inductive inference. The procedure recovers a mean of 78.5, which is ahead of every post-hoc reading of the strictly trained softmax model, within 4.8 points of the transductive result, and above the 73.1 that the strongest baseline of Table 1 attains with full transductive training. The largest recoveries come on DBLP-H, from 58.2 to 82.3, and on Chameleon, from 61.2 to 80.3, while on Facebook the adapted detector even exceeds its transductive score.

Varying the OOD Proportion. We keep only a fraction r of the held-out-class nodes in the graph, remove the features and edges of the dropped nodes entirely before training, and score the retained OOD nodes against the ID test set. This probes how the transductive scores, including the global midpoint $\mathrm{med}(a)$, depend on the prevalence of the very mixture on which they are computed. On the core-suite mean, REND leads the propagated-energy score of the identically trained softmax model

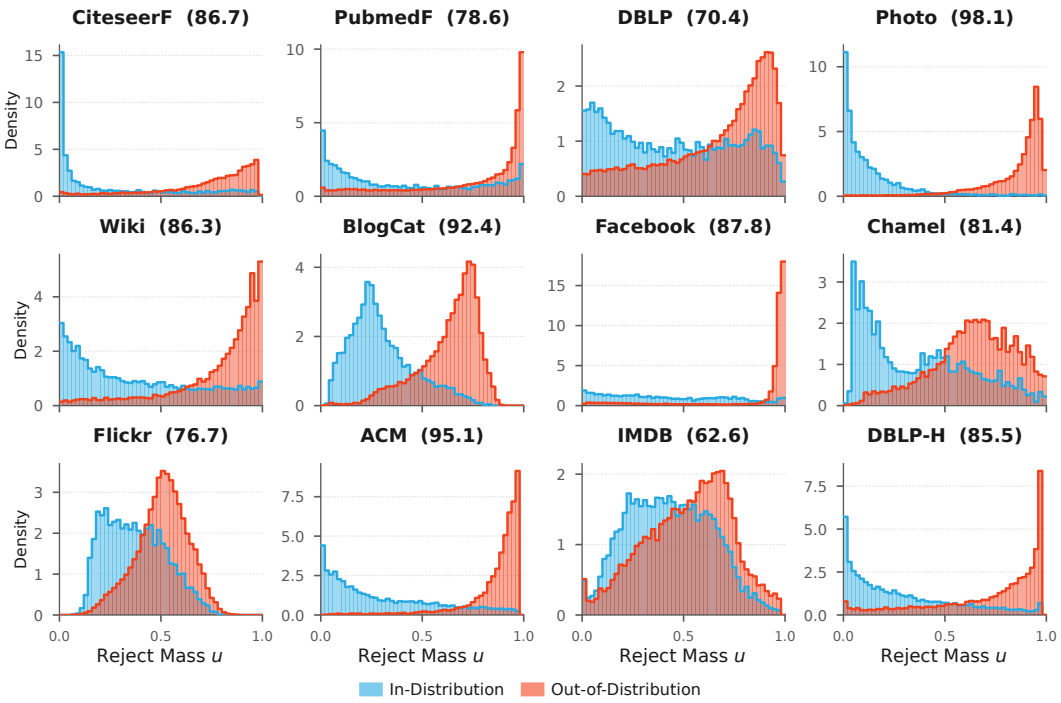

Figure 7: Distribution of the diffused reject mass u on the evaluation nodes of every benchmark, in-distribution against out-of-distribution, pooled over ten runs.

by eleven points at full prevalence and by nearly five at half prevalence, and on the heterophilous Chameleon graph it leads at every prevalence tested (Fig. 6a–c). The core-suite lead persists down to roughly three tenths of the held-out population, below which too few genuine donors remain and the affinity midpoint drifts toward a purely in-distribution median.

Discussion. Every method in Table 1 is granted the same transductive mixture, and the margin of REND survives the removal of every genuine OOD feature donor. When new nodes arrive after training, the decoupling that protects the classifier also lets the reject channel re-anchor itself on the grown graph without any new label. The same decoupling keeps the classifier of REND a plain cross-entropy network, so its propagated-energy score comes at no extra cost as a complementary readout whenever novelty is known to be rare.

.10 Distribution of Reject Scores

Score Distributions. The detection numbers of Table 1 rank nodes but say nothing about how the score is spread, so Fig. 7 shows the reject mass itself. This study retrains the model rather than reusing the predictions of the main runs, so its per-dataset AUROC matches Table 1 only to within run-to-run variance. The ordering that the reject channel predicts holds without exception, since on all twelve graphs the in-distribution nodes carry less reject mass on average than the out-of-distribution nodes. The in-distribution mode sits at the bottom of the range on nine benchmarks, and the out-of-distribution mode lies above 0.9 on eight. The population means separate most widely on Photo, at 0.14 against 0.84, on ACM, at 0.30 against 0.86, and on full-CiteSeer, at 0.26 against 0.73, and the width of this gap tracks the AUROC across the whole suite. Since the modes themselves move apart, the separation comes from a shift of the whole distribution and not from a few extreme scores.

Overlap Analysis. One source of overlap between the two histograms is the borrowing that Proposition 2 bounds. In-distribution nodes beside out-of-distribution neighbors import novelty through the same edges that carry the signal elsewhere, so their scores rise with the number of such neighbors (Fig. 5a), while the cap of the proposition keeps that borrowing bounded. Even among these exposed nodes, the false-positive rate of REND stays below that of energy propagation in every exposure bucket (Appendix .8).