IRL-GAD: Graph Anomaly Detection via Inverse Reinforcement Learning as Normality Modeling

Anonymous authors
Paper under double-blind review

Abstract

Most graph anomaly detection methods define anomaly as statistical outlier-ness in a fitted representation, a framing that faces a specific weakness when adversaries position themselves close to normal nodes in feature space. We propose a behavioral alternative: anomaly as deviation from an implicit policy governing the normal population. We recast each node's multi-hop neighborhood as a trajectory in a Markov decision process (the Node-MDP), recover the reward driving normal demonstrations via maximum-entropy inverse reinforcement learning (MaxEnt-GIRL), and score nodes by the KL divergence between their observed aggregation policy and the soft-optimal policy induced by the recovered reward. The reward decomposes into structural, semantic, and temporal components, yielding component-level interpretability of every detection. Theoretically, we establish reward identifiability with a graph-specific strengthening, a finite-sample recovery bound, a camouflage detection margin under a bounded threat model that holds adaptively within the budget against an omniscient adversary, a closed-form soft-value regret bound, and a PAC-style bound on the deployed detector's false-positive rate. Across six benchmarks (homophilic, camouflaged, dynamic, large-scale), IRL-GAD improves on the strongest baseline by +1.7 AUC-ROC points on average and by +2.3 on YelpChi, with the learned reward transferring to anomaly types absent at training.

1 Introduction

In February 2016, attackers compromised the SWIFT terminal at Bangladesh Bank and submitted thirty-five payment instructions to the Federal Reserve Bank of New York, attempting to transfer nearly one billion dollars to accounts in the Philippines and Sri Lanka (Nilson Report, 2023). The striking feature of the attack was not what the messages contained but what they were: every individual instruction conformed to the syntactic, structural, and statistical norms of a routine interbank wire. A reviewer inspecting any single message in isolation would have found nothing out of place. What was anomalous was the *sequence of decisions* the compromised terminal made—the choice of beneficiaries, the timing relative to weekend cutoffs, and the multi-hop routing through correspondent banks. The fraud was a behavioral anomaly hidden inside an aggregate of locally normal interactions.

This pattern recurs across the high-cost settings where graph anomaly detection (GAD) is deployed: review-fraud rings on e-commerce platforms (Dou et al., 2020; Wang et al., 2023), coordinated misinformation on social networks (Zhang et al., 2022), and rare-event identification in molecular interaction graphs (Liu et al., 2023). Adversaries in these settings are not merely statistical outliers waiting to be flagged; they are agents who actively shape their local neighborhoods to resemble the normal population. The detection problem is therefore intrinsically behavioral rather than geometric.

The dominant paradigms for GAD share a foundational commitment that is at odds with this framing. Reconstruction-based methods fit a generative model on the graph and flag nodes whose features or links cannot be regenerated with low error (Ding et al., 2019; Fan et al., 2020; Chen et al., 2020; Qiao & Pang, 2025). Contrastive methods learn a similarity metric and flag nodes whose embeddings are distant from learned positive views (Liu et al., 2021b; Duan et al., 2023). Spectral approaches detect anomalies as deviations

in the frequency response of the graph signal (Tang et al., 2022; Deng & Hooi, 2021). Camouflage-aware methods refine the message-passing operator to be robust to local noise (Dou et al., 2020; Wang et al., 2023). Despite their methodological diversity, these frameworks share a common emphasis on geometric criteria over behavioral ones: they ask where each node sits relative to a fitted distribution over a representation space, rather than whether the node's *decisions*—its choices of neighbors, edges, and message routes—are consistent with the implicit policy that governs the normal population. This is a productive emphasis on many problems and we do not claim it is wrong in general; we claim that it has a specific weakness on a class of problems where adversaries actively shape their local neighborhoods, and that a behavioral framing offers complementary signal in this regime.

The weakness is most apparent under camouflage. A node that has been strategically arranged to lie geometrically close to the normal population produces a small reconstruction error, a small contrastive distance, and a spectral signature close to the normal mode by design; metric-based separation in the representation space therefore becomes difficult, often prohibitively so, regardless of the specific architecture or training objective. Camouflage-aware methods such as Dou et al. (2020) and Wang et al. (2023) partially mitigate this by reshaping the message-passing operator, but the scoring objective remains a function of the representation, so the mitigation is incremental rather than structural. Three related challenges—open-set fragility (an unseen anomaly falls outside the support of the fitted distribution), limited interpretability (a reconstruction error or contrastive distance carries no semantic content beyond "unusual"), and the absence of behavior-level guarantees (a result on the recoverability of a generative density does not translate into a result on the detection of behaviorally deviant agents)—tend to co-occur in the same methods, which suggests that they are not four independent issues but interactions of the geometric emphasis with specific deployment regimes.

What is missing from the geometric account is the concept that nodes do not merely *occupy* a graph; they *act* within it. The formation of an edge, the routing of a message, and the aggregation of multi-hop neighbors are all decisions, and the trajectory of these decisions over a K-hop neighborhood is precisely the trace of a sequential decision process. The question for anomaly detection then becomes whether the trajectory observed at a given node is plausible under the implicit policy that produced the trajectories of the normal population. Crucially, this policy is not known a priori; only its samples—the observed behavior of nodes labeled as normal—are available. Recovering an unknown reward function from observed behavior is precisely the problem of inverse reinforcement learning (IRL) (Arora & Doshi, 2021; Ho & Ermon, 2016), and the maximum-entropy formulation of Ziebart et al. (2008) is the canonical instrument for doing so when the recovered reward must be both smooth and identifiable. This is why IRL, rather than direct policy learning, trajectory clustering, or sequence-level density estimation, is the natural inferential machinery for the problem: we do not observe the reward, we observe only its consequences, and IRL is the framework designed precisely for this asymmetry.

These observations translate into three concrete research questions, each of which carries a non-trivial claim that must be substantiated either theoretically or empirically.

(RQ1) Can the implicit normality of a population of graph nodes be characterized as a reward function over a well-defined sequential decision process built from the graph itself?

(RQ2) If such a reward exists, does the deviation of an individual node's policy from this reward provide a detection margin under an explicit camouflage threat model that is not available to a scorer operating on the terminal embedding alone—in particular, under camouflage?

(RQ3) Does this framing translate into measurable empirical gains on standard GAD benchmarks, and does the learned reward transfer to anomaly types that were not represented during training?

To answer these questions, we propose IRL-GAD, an inverse-reinforcement-learning approach to graph anomaly detection. The method rests on three closely connected ideas. First, we establish a correspondence between a node and a Markov decision process whose states encode local subgraph context, whose actions correspond to aggregation choices over neighbours, and whose transitions follow the message-passing dynamics of a graph attention encoder. The K-hop trajectory of a node becomes a rollout in this MDP, and a population

of normal nodes becomes a collection of demonstrations. Second, we recover the implicit reward governing these demonstrations through a maximum-entropy guided IRL procedure (MaxEnt-GIRL) that exploits the soft Bellman equations of Ziebart (2010) to ensure smoothness and uniqueness of the learned reward up to a potential-shaping equivalence. The reward is parameterized by a decomposable network whose components separately attribute deviation to structural, semantic, and temporal factors, giving the resulting anomaly scores a built-in attribution mechanism. Third, we score a candidate node by the Kullback–Leibler divergence between its observed policy and the soft-optimal policy induced by the learned reward; this score is well defined, comparable across nodes of varying degree, and admits a closed-form upper bound on its false-positive rate under mild assumptions.

The contributions of this work are as follows.

- We introduce the Node-MDP correspondence, a principled formalism that recasts the multi-hop interactions of a node as trajectories within a Markov decision process and thereby reframes anomaly detection as a deviation problem over policies rather than over embeddings.
- We develop MaxEnt-GIRL, an unlabeled training procedure that recovers the implicit reward of the normal population from observed behavior alone, together with a policy-deviation scoring rule whose decomposition into structural, semantic, and temporal components yields component-level interpretability of every detection.
- We prove identifiability of the recovered reward up to potential shaping (with a graph-specific strengthening) and a finite-sample recovery bound for the graph-trajectory setting. Under a bounded-camouflage threat model with Lipschitz reward heads, policy-deviation scoring admits a detection margin whose lower bound depends on the attack budget and Lipschitz constants but *not* on the camouflaged-to-normal embedding distance, and holds adaptively against an omniscient adversary. Closed-form soft-value regret, soft-value iteration residual, and PAC-style false-positive-rate bounds propagate the finite-sample error to the deployed detector.
- We provide a comprehensive empirical evaluation on six benchmarks spanning homophilic, camouflaged, dynamic, and large-scale settings, demonstrating consistent gains over reconstruction, contrastive, spectral, camouflage-aware, and evidential baselines, with reward transfer to anomaly types unseen at training time.

The remainder of the paper is organized as follows. Section 2 reviews prior work on graph anomaly detection and inverse reinforcement learning. Section 3 formalizes the Node-MDP. Section 4 develops the MaxEnt-GIRL training procedure and the policy-deviation scoring rule. Section 5 reports empirical results across all benchmarks. Section 6 concludes. The theoretical analysis, including the camouflage separability theorem, is deferred to Appendix A.

2 Related Work

The literature on graph anomaly detection has evolved through three distinct paradigms, each contributing important machinery to the detection problem while leaving the camouflage challenge only partially resolved. We organize prior work along two orthogonal axes—*what is modeled as normality* and *what signal drives detection*—and observe that the dominant axis pair across the literature is geometric normality modeled via a static proxy task; behavioral framings, in which detection is grounded in the recovery of an implicit policy, remain comparatively under-explored. IRL-GAD contributes to this less-developed direction; we make no claim of priority over the broader effort of bringing sequential-decision ideas to graph anomaly detection.

Paradigm I: Reconstruction-based methods (Pang et al., 2021; Ruff et al., 2021). The earliest and most widely deployed family of unsupervised GAD methods trains a graph autoencoder to reconstruct node attributes and/or adjacency structure from a learned latent representation, then flags nodes with high reconstruction error as anomalous (Ding et al., 2019; Fan et al., 2020; Chen et al., 2020). DOMINANT (Ding et al., 2019) jointly minimizes attribute and structure reconstruction loss via a shared GCN encoder; AnomalyDAE (Fan

et al., 2020) separates the two objectives into dual decoders with cross-attention; GAAN (Chen et al., 2020) introduces a generative adversarial variant that sharpens the reconstruction boundary. GAD-NR (Qiao & Pang, 2025) extends this paradigm by reconstructing each node's neighborhood (rather than the node alone), achieving better calibration on heterophilic graphs. The principal weakness of this paradigm is not architectural but definitional: reconstruction error measures *how unusual a node looks*, not *how it behaves*. A camouflaged anomaly that engineers its connectivity to match its neighborhood will achieve low reconstruction error by construction, weakening the detector's discriminative power regardless of encoder capacity or training procedure (Tang et al., 2022). Within the reconstruction paradigm this issue is hard to address by purely architectural means: methods that conflate normality with low-distortion embedding offer no built-in mechanism to penalize behaviorally inconsistent but topologically plausible aggregation patterns.

Paradigm II: Contrastive and self-supervised methods. The second wave of GAD adopts contrastive learning as the proxy task. CoLA (Liu et al., 2021b) draws positive pairs from a node and its sampled subgraph; GRADATE (Duan et al., 2023) scales this to multi-granularity objectives across node, subgraph, and graph levels. These methods capture richer relational context than reconstruction baselines, yet share the same unexamined assumption: normality is proximity in a representation space induced by the contrastive objective. Against camouflaged anomalies—optimized to preserve local context while altering interaction patterns—contrastive methods suffer the same failure mode (Tang et al., 2022; Gao et al., 2023). The proxy task changes; the geometric premise does not.

Paradigm III: Spectral, topology-aware, and camouflage-aware methods. BWGNN (Tang et al., 2022) employs beta wavelets (Qiao et al., 2024) to preserve heterophilic edges suppressed by standard GNN aggregation. GDN (Deng & Hooi, 2021) learns deviation thresholds (Li et al., 2022) from sensor graph structure; PC-GNN (Liu et al., 2021a) addresses class imbalance via label-balanced sampling. A more recent thread of *camouflage-aware* methods directly targets the topology-engineering attack vector: CARE-GNN (Dou et al., 2020) uses similarity-based neighbor filtering and reinforcement-learning-tuned thresholds to suppress camouflaged edges; GAGA (Wang et al., 2023) aggregates over group-encoded neighborhoods to dilute the influence of any single attacker; GEL (Gao et al., 2024) introduces evidential uncertainty to quantify detection confidence and improve reliability under camouflage. A complementary line of work studies adversarial perturbations of graph structure: prior work (Mohammadi et al., 2023) observes that the connectivity-manipulation strategies used to suppress influence-propagation signals are structurally similar to those used by camouflaged anomalies—an observation that motivates, rather than resolves, a shift from geometric to behavioral detection signals. These methods are more architecture-aware than reconstruction baselines, but the detection signal remains structural or spectral deviation in representation space, with no mechanism for explaining *why* a node is anomalous or for generalizing to unseen anomaly types (Ma et al., 2025; Qiao et al., 2025). Despite the recent progress reflected in CARE-GNN, GAGA, GAD-NR, and GEL, all of these approaches remain grounded in geometric normality modeling and share the fundamental challenge that adversarial camouflage—by design—suppresses the geometric signal they rely upon.

Reinforcement learning in graph learning. Prior work uses RL to optimize an existing GNN component: Policy-GNN (Lai et al., 2020) learns a meta-policy over the number of aggregation hops; Yoon et al. (2022) frames neighborhood sampling as a sequential decision problem; PC-GNN-style methods (Liu et al., 2021a) adopt RL-style selection for class-imbalanced fraud graphs. All of these use RL to *augment* an existing proxy task. IRL-GAD takes a different route: it replaces the reconstruction or contrastive proxy task with reward inference from normal trajectory distributions.

Recent hybrid approaches. A separate trajectory in the recent literature combines multiple paradigms in a single architecture. Self-supervised pre-training paired with camouflage-aware fine-tuning, attention-based message passing combined with spectral regularization, and generative augmentation with contrastive scoring have each appeared in recent works; the consistent finding across this trajectory is that the geometric scoring head is the principal limitation, since the various pre-training objectives produce excellent representations but the downstream anomaly score remains a distance or reconstruction error on those representations. IRL-GAD is complementary to these hybrid efforts: the representation backbone can in principle be borrowed from any of the hybrid methods (we use a vanilla GAT for clean isolation of the IRL contribution), and the policy-deviation score replaces the geometric scoring head without modifying the upstream pipeline. A

systematic empirical comparison of policy-deviation scoring layered on top of recent self-supervised backbones is an attractive direction we leave to future work.

3 Preliminaries and Problem Formulation

We fix graph notation, reinterpret a graph attention computation as a K-step sequential decision process, recall the MDP and maximum-entropy IRL machinery, and combine these into the Node-MDP correspondence on which the rest of the paper depends.

Graphs and attributed subgraphs. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ denote an attributed graph with node set $\mathcal{V}$ of size $|\mathcal{V}| = n$, edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The k-hop ego-subgraph of node v is

$$\mathcal{G}_{\mathrm{sub}}(v, k) \;=\; (\mathcal{V}_v^k,\, \mathcal{E}_v^k,\, \mathbf{X}_v^k), \qquad \mathcal{V}_v^k \;=\; \{u \in \mathcal{V} : d_{\mathcal{G}}(u, v) \le k\}, \tag{1}$$

where $d_{\mathcal{G}}$ is the shortest-path distance. The training corpus is the collection of ego-subgraphs of a labeled normal population $\mathcal{G}^+ \subseteq \mathcal{V}$, $\mathcal{G}^+ = \{\mathcal{G}_{\mathrm{sub}}(v, k) : v \in \mathcal{G}^+\}$. No anomaly labels are required at training.

Graph attention as a sequential process. A K-layer graph attention network (Veličković et al., 2018) iterates

$$\mathbf{h}_v^{(t)} \;=\; \sigma\left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(t)}\, \mathbf{W}^{(t)} \mathbf{h}_u^{(t-1)}\right), \qquad t = 1, \ldots, K, \tag{2}$$

with $\mathbf{h}_v^{(0)} = \mathbf{x}_v$ and learned attention coefficients $\alpha_{vu}^{(t)} \in [0, 1]$ summing to one. We read each row $\alpha_v^{(t)} = (\alpha_{vu}^{(t)})_{u \in \mathcal{N}(v)}$ as a distribution over neighbors, making the iteration a K-step sequential decision process.

Markov decision processes. A finite-horizon MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, \gamma)$. We work with the maximum-entropy (soft) optimal policy (Ziebart et al., 2008)

$$\pi_{\mathrm{soft}}^{\star}(a \mid s) \;\propto\; \exp\big(Q^*(s, a)/\beta\big), \tag{3}$$

where Q^* is the soft Q-function and $\beta > 0$ is the entropy temperature. The soft form yields reward identifiability up to a potential shaping (Ziebart, 2010); all later uses of π^* refer to $\pi_{\mathrm{soft}}^{\star}$.

Maximum-entropy inverse reinforcement learning. Given expert trajectories $\mathcal{T}^+ = \{\tau_i\}_{i=1}^N$, MaxEnt-IRL (Ziebart et al., 2008) treats each as a sample from

$$\mathbb{P}(\tau \mid R_\theta) \;\propto\; \exp\left(\sum_{t=0}^{K} R_\theta(s_t, a_t)\right), \tag{4}$$

and recovers R_θ by maximizing log-likelihood over the expert set. In our setting the experts are the aggregation trajectories of normal nodes.

3.1 The Node-MDP correspondence

The message-passing computation a GAT performs at a single node is the rollout of an MDP specific to that node, the attention row at each hop is the policy of that process, and the graph of normal nodes is a population of expert demonstrators for an unknown reward.

Definition 1 (Node-MDP)**.** *For a node $v \in \mathcal{V}$, the Node-MDP associated with v is the tuple $\mathcal{M}_v = (\mathcal{S}_v, \mathcal{A}_v, T, R_\theta, \gamma)$ defined as follows.*

State. $s_t^v = \big(\mathbf{h}_u^{(t)}\big)_{u \in \mathcal{V}_v^{K-t}}$ *is the tuple of hop-t representations of every node within $K - t$ hops of v; at the final hop, s_K^v collapses to the K-hop embedding $\mathbf{h}_v^{(K)}$ that conventional methods consume.*

Action. $a_t^v = u \in \mathcal{N}(v)$ *is the neighbor of* v *selected for aggregation at hop* t*; the action space is the discrete set* $\mathcal{N}(v)$*.*

Aggregation policy. *The distribution from which* a_t^v *is drawn is identified with the attention row at hop* $t+1$*,*

$$\pi_v(\cdot \mid s_t^v) \;:=\; \alpha_v^{(t+1)} \;\in\; \Delta^{|\mathcal{N}(v)|}. \tag{5}$$

Transition. $T(s_{t+1}^v \mid s_t^v, a_t^v)$ *is deterministic, induced by Equation* (2)*.*

Reward. $R_\theta : \mathcal{S}_v \times \mathcal{A}_v \to \mathbb{R}$ *is the latent normality utility, parameterized by a neural network and inferred from the normal corpus alone.*

Discount. $\gamma \in (0, 1]$ *is the hop discount.*

Remark 1 (Discrete actions versus continuous attention)**.** *The action* a_t^v *is the discrete neighbor selected for aggregation; the attention row* $\alpha_v^{(t+1)}$ *is the distribution over neighbors from which* a_t^v *is drawn, exactly as a softmax classifier emits a distribution over class labels while the predicted label is the discrete sample. Definition 4's policy-deviation score is therefore a finite-sum KL on the simplex* $\Delta^{|\mathcal{N}(v)|}$*.*

Definition 2 (Aggregation Trajectory)**.** *The aggregation trajectory of node* v *is*

$$\tau_v \;=\; \big(s_0^v,\, \pi_v^{(0)},\, s_1^v,\, \pi_v^{(1)},\, \ldots,\, s_K^v\big), \tag{6}$$

with $\pi_v^{(t)} := \alpha_v^{(t+1)}$*. Conventional GAD discards the trajectory and keeps only* $\mathbf{h}_v^{(K)}$*; we keep the trajectory.*

Definition 3 (Normality Hypothesis)**.** *A node* $v \in \mathcal{G}^+$ *is normal iff its aggregation policy* π_v *is approximately optimal under a latent reward* R_θ^**,*

$$\pi_v \;\approx\; \pi_{\text{soft}}^\star(R_\theta^*) \;=\; \arg\max_\pi \; \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{K} \gamma^t R_\theta^*(s_t, a_t) \;+\; \beta\, H\big(\pi(\cdot \mid s_t)\big)\right]. \tag{7}$$

Detection becomes a deviation problem: we recover R_θ^* and ask how far each node's policy lies from optimality under it. The framework is naturally open-set (no training on anomalies) and the reward decomposes into interpretable components (Section 4).

Definition 4 (Policy Deviation Score)**.** *The anomaly score for node* v *is the cumulative KL between its empirical aggregation policy and the soft-optimal policy under the recovered reward,*

$$\mathcal{S}(v) \;\triangleq\; D_{\text{dev}}(v) \;=\; \sum_{t=0}^{K-1} \mathrm{KL}\Big(\pi_v^{(t)} \,\Big\|\, \pi_{\text{soft}}^\star(\cdot \mid s_t^v)\Big). \tag{8}$$

The score is non-negative, zero iff v matches the soft-optimal at every hop, and computable in closed form on the discrete neighbor set. Section 4 develops the recovery of R_θ^*; theoretical guarantees including Theorem 1 are in Appendix A.

4 IRL-GAD: Proposed Method

Section 3 reduced graph anomaly detection to two coupled inferential problems. Given only the aggregation trajectories of a labeled normal population $\mathcal{G}^+$, we must recover the latent reward R_θ^* under which those trajectories are approximately soft-optimal, and we must score every test node by how far its observed aggregation policy deviates from the soft-optimal policy this reward induces. IRL-GAD instantiates both as an end-to-end procedure organized around three modules: a trajectory extractor that exposes the GAT's hop-by-hop attention as the policy whose optimality is in question (M1); a maximum-entropy guided IRL stage that recovers a decomposable reward from normal trajectories alone (M2); and a policy deviation scorer that measures the per-hop Kullback–Leibler gap between observed and soft-optimal aggregation (M3). Figure 1 shows the pipeline and identifies M2 as the module that has no analogue in any prior GAD method.

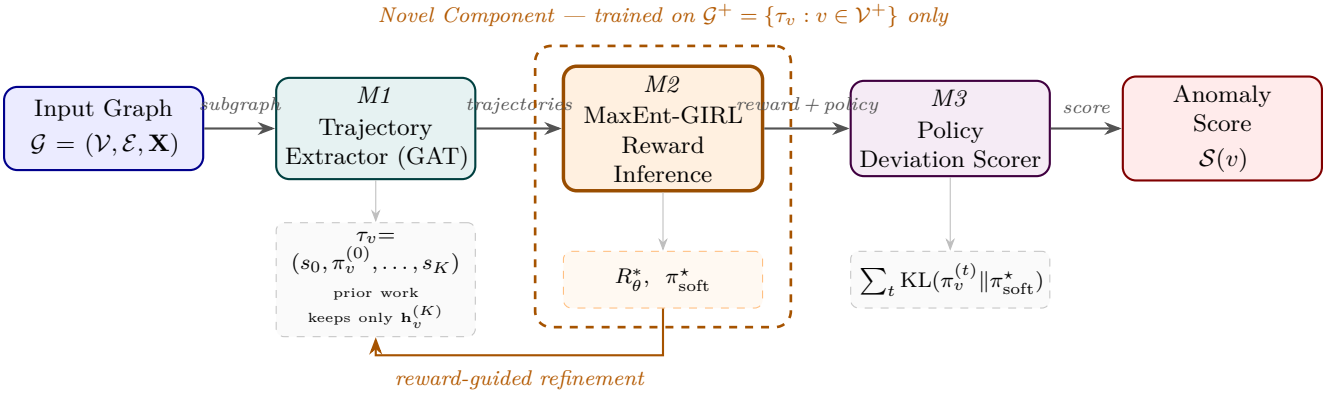

Figure 1: IRL-GAD pipeline. The orange dashed box marks the novel component absent from prior GAD methods: reward inference from the distribution of normal aggregation trajectories. The trajectory τ_v preserves the per-hop attention rows $\{\boldsymbol{\alpha}_v^{(t)}\}_{t=0}^{K-1}$ (i.e., the empirical policy) and the per-hop states; prior methods collapse τ_v to the terminal embedding $\mathbf{h}_v^{(K)}$ alone, discarding the policy information that M3 scores against.

4.1 Overview

The three modules are coupled. M1 fixes the form of the demonstration data: discarding the per-hop attention rows would leave M2 with no policy to score against. M2 fixes the normality reference: a hand-designed reward would reflect the architect's prior rather than the structure of the data, and would not generalize to unseen anomaly types. M3 fixes the invariance properties of the score: operating on embeddings rather than action distributions would let a camouflaged adversary close the geometric gap to the normal manifold without changing its aggregation behavior. Removing any one of the three collapses performance to the level of conventional embedding-based detectors (Section 5).

4.2 M1: Trajectory Extraction

The purpose of M1 is to expose, for every node, the object that M2 operates on. A standard GAT encoder is a procedure that produces a K-hop embedding $\mathbf{h}_v^{(K)}$ by iterating Equation (2); conventional GAD methods consume only this terminal vector and treat the iteration as an internal implementation detail. M1 unrolls the same iteration and records, at every hop, both the hidden state and the attention row that produced it. The resulting trajectory of v is the sequence

$$\tau_v \;=\; \big(s_0^v,\, \pi_v^{(0)},\, s_1^v,\, \pi_v^{(1)},\, \ldots,\, s_K^v\big), \tag{9}$$

where $s_t^v = (\mathbf{h}_u^{(t)})_{u \in \mathcal{V}_v^{K-t}}$ is the augmented state of Definition 1 and $\pi_v^{(t)} := \boldsymbol{\alpha}_v^{(t+1)} \in \Delta^{|\mathcal{N}(v)|}$ is the empirical aggregation policy at hop t, which subsequent modules treat as the IRL demonstration.

A small example fixes the intuition. Suppose node v has three neighbors $\{u_1, u_2, u_3\}$ and the encoder is unrolled for $K = 2$ hops. The conventional GAT view records only the final embedding $\mathbf{h}_v^{(2)}$. The IRL-GAD view records the entire sequence: the initial state s_0^v collecting input features within two hops of v; the hop-0 aggregation policy $\pi_v^{(0)} = (0.2, 0.3, 0.5)$ describing how v's attention distributed across u_1, u_2, u_3 during the first aggregation; the next state s_1^v collecting hop-1 representations within one hop of v; the hop-1 policy $\pi_v^{(1)} = (0.4, 0.5, 0.1)$, showing that v has shifted attention onto u_2 in the second hop; and the final state $s_2^v = \mathbf{h}_v^{(2)}$. The terminal embedding is one item in this list. Everything else is the behavioral signal that the IRL stage will use as evidence.

The normal trajectory corpus $\mathcal{T}^+ = \{\tau_v : v \in \mathcal{G}^+\}$ serves as the expert demonstration set for MaxEnt-GIRL. No anomaly labels enter the pipeline at this stage; the only supervision used during training is the identity of the nodes assumed or labeled benign.

Algorithm 1 MaxEnt-GIRL Training

Require: Attributed graph $\mathcal{G}$, normal nodes $\mathcal{V}^+$, hop depth K, learning rate η, reward network R_θ, regularization λ

Ensure: Trained reward parameters θ^*

Extract $\mathcal{T}^+ \leftarrow \{\tau_v : v \in \mathcal{V}^+\}$ via K-hop GAT (Eq. 9)
repeat
  Sample mini-batch $\mathcal{B} \subset \mathcal{T}^+$
  Approximate $Z(R_\theta)$ via soft value iteration over $\mathcal{B}$
  Compute $\mathcal{L}_{\text{IRL}}(\theta)$ via Eq. (12)
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{IRL}}$
until convergence
Compute $\pi^\star_{\text{soft}}(a \mid s) \propto \exp(Q^*(s,a)/\beta)$
return θ^*, $\pi^\star_{\text{soft}}$

4.3 M2: Maximum Entropy Graph IRL (MaxEnt-GIRL)

M2 recovers, from $\mathcal{T}^+$ alone, a scalar reward whose soft-optimal policy approximates the aggregation behavior of normal nodes. Once recovered, the soft Bellman equations of Section 3 prescribe the reference policy in closed form, and M3 scores any node by deviation from it.

Single-policy assumption. M2 recovers a *single* reward $\hat{R}_N$ capturing the shared utility of normal decisions; this applies when normality admits a single behavioral mode, as on the benchmarks here. It is *problematic* on multi-modal populations such as a social platform spanning power-users / lurkers / moderators, or a multi-region fraud graph whose legitimate-merchant behaviors differ across jurisdictions: $\hat{R}_N$ becomes a mixture-collapsed average and M3 flags minority-style normals. A mixture-of-rewards extension that EM-clusters normals and recovers one reward per mode would address this, at the cost of identifiability per component (each cluster carries its own sample-complexity deflator) and an outer-loop selection of the number of modes.

Reward parameterization. A monolithic neural reward would be expressive but uninterpretable. We instead parameterize R_θ as a sum of three components, each implemented as a separate MLP head sharing a backbone ϕ_θ,

$$R_\theta(s,a) \;=\; \underbrace{R^{\text{str}}_{\theta_1}(s,a)}_{\text{structural cohesion}} \;+\lambda_1\; \underbrace{R^{\text{sem}}_{\theta_2}(s,a)}_{\text{semantic consistency}} \;+\lambda_2\; \underbrace{R^{\text{tmp}}_{\theta_3}(s,a)}_{\text{temporal stability}}, \tag{10}$$

with learned mixing $\lambda_1, \lambda_2 \geq 0$. The structural component penalizes attention patterns that violate local topology or density; the semantic component rewards feature-consistent aggregation; the temporal component rewards stability of aggregation choices across hops (and across physical time on dynamic graphs). The decomposition constrains the reward to interpretable axes, permits per-component ablation (Section 5.5), and produces human-auditable reward landscapes (Section 5.6).

Training objective. Under MaxEnt (Ziebart et al., 2008), trajectory likelihood is

$$\mathbb{P}(\tau \mid R_\theta) \;=\; \frac{1}{Z(R_\theta)}\, \exp\Big(\sum_{t=0}^{K} R_\theta(s_t, a_t)\Big). \tag{11}$$

The partition function $Z(R_\theta)$ is intractable; we approximate it through soft value iteration over the mini-batch state space (a fixed number of soft-Bellman backups). The reward is trained by maximum likelihood on the normal corpus,

$$\mathcal{L}_{\text{IRL}}(\theta) \;=\; -\frac{1}{|\mathcal{T}^+|} \sum_{\tau \in \mathcal{T}^+} \log \mathbb{P}(\tau \mid R_\theta) \;+\; \lambda \,\|\theta\|_2^2. \tag{12}$$

Amortized inference. Sharing ϕ_θ across all nodes reduces complexity from a naïve $\mathcal{O}(|\mathcal{V}|^2)$ to $\mathcal{O}(K|\mathcal{E}|d)$, matching a standard GNN forward pass. Algorithm 1 summarizes the procedure.

Table 1: Where the anomaly score lives in each major GAD family. IRL-GAD is the only family whose score operates in action space against a learned reference, preserving the trajectory through which the embedding is computed.

Family	Operates in	Reference object	Trajectory
Reconstruction	embedding space	fitted generative density	discarded
Contrastive	embedding space	learned similarity metric	discarded
Spectral	embedding space	graph-signal prior	discarded
Attention regulariz.	embedding space	static regularization target	discarded
IRL-GAD (ours)	action space	soft-optimal $\pi^\star_{\text{soft}}(R^*_\theta)$	retained

4.4 M3: Policy Deviation Scoring

At test time, M3 uses the trained R^*_θ and the induced $\pi^\star_{\text{soft}}$ to score every node, including nodes outside the training corpus. For candidate $v \in \mathcal{V}$, M3 extracts τ_v through the frozen GAT, reads $\pi_v^{(t)}$ at every hop, and computes

$$\begin{aligned}\mathcal{S}(v) &= \sum_{t=0}^{K-1} \mathrm{KL}\big(\pi_v^{(t)} \,\big\|\, \pi^\star_{\text{soft}}(\,\cdot \mid s_t^v)\big) \\ &= \sum_{t=0}^{K-1} \sum_{u \in \mathcal{N}(v)} \pi_v^{(t)}(u)\, \log \frac{\pi_v^{(t)}(u)}{\pi^\star_{\text{soft}}(u \mid s_t^v)}.\end{aligned} \tag{13}$$

Each per-hop term is a finite-sum KL on the discrete simplex $\Delta^{|\mathcal{N}(v)|}$, computable in closed form. Three properties matter. (i) $\mathcal{S}(v) = 0$ iff v's policy matches $\pi^\star_{\text{soft}}$ at every hop, so normal nodes are fixed points of the scorer by construction; no calibration is needed to teach the score that benign nodes are not anomalous. (ii) $\mathcal{S}(v)$ operates in *action space*: an adversary that pulls $\mathbf{h}_v^{(K)}$ close to the normal manifold does not by itself reduce KL between $\pi_v^{(t)}$ and $\pi^\star_{\text{soft}}$, since the score depends only on the per-hop distributions; the residual camouflage gap is bounded by Lemma 1 and Theorem 1. (iii) $\mathcal{S}(v)$ decomposes by hop, giving built-in localization: an analyst can read off whether the deviation originated at hop 1 or at deeper hops directly from the summands. Nodes with $\mathcal{S}(v) > \delta$ are flagged, with δ calibrated on a held-out validation set to target a desired FPR.

4.5 Why the Score Cannot Be Recovered from Embeddings

$\mathcal{S}(v)$ is not recoverable from any method operating on the terminal embedding alone, for two structural reasons. *First, action vs. embedding space.* A reconstruction or contrastive score is a function of $\mathbf{h}_v^{(K)}$ and has access only to representation-space geometry; the KL of Equation (13) is a function of the per-hop distributions $\pi_v^{(t)}$ and $\pi^\star_{\text{soft}}$, which a method collapsing τ_v to $\mathbf{h}_v^{(K)}$ has discarded. *Second, the reference.* Even a method that retained the attention rows would need something to deviate against. Reconstruction and contrastive losses produce fitted densities or similarity metrics, not a reference policy; $\pi^\star_{\text{soft}}$ exists only as a consequence of having recovered R^*_θ through IRL. Table 1 summarizes the contrast.

The ablation in Section 5.5 confirms this: swapping $\mathcal{S}(v)$ for reconstruction or contrastive scores on the same GAT backbone collapses to the non-camouflage-aware baseline level.

4.6 Why IRL Is More Than Attention Reweighting

Four properties distinguish IRL-GAD from an attention-regularizer interpretation. (i) The reward R^*_θ is *inferred* from the empirical distribution of normal trajectories via the MaxEnt likelihood, not inserted as a fixed penalty. (ii) The reference policy $\pi^\star_{\text{soft}}$ is Bellman-consistent: it is induced by soft value iteration over R^*_θ, which lets the soft performance-difference identity used in Theorem 1 apply to $\mathcal{S}(v)$ directly; an attention regularizer added during GAT training does not in general induce a quantity with this property. (iii) The per-hop KL terms are calibrated by R^*_θ on the normal corpus, so they are commensurate across

Table 2: Benchmark datasets, grouped by anomaly regime, with the research question each dataset is primarily used to evaluate.

Regime	Dataset	#Nodes	#Edges	Feats	Anom.%	Tests
Homophilic	Cora	2,708	5,429	1,433	5.0	Q2
	Citeseer	3,327	4,732	3,703	5.0	Q2
Camouflage	Amazon	11,944	4.20M	25	9.5	Q1
	YelpChi	45,954	3.85M	32	14.5	Q1
Dynamic	JODIE	70,314	1.30M	172	7.1	Q3
Large-scale	ogbn-arxiv	169,343	1.17M	128	5.0	Q5

nodes, graphs, and deployments without per-graph retuning. (iv) The ablations of Section 5.5, which fix the GAT and substitute the IRL score with non-IRL alternatives, collapse to the level of the underlying encoder, confirming that the IRL stage is the load-bearing component.

5 Experiments

The evaluation that follows is organized around five research questions. Each question targets a single load-bearing claim of the IRL-GAD formulation and is paired with the experimental setting most likely to expose a failure of that claim.

(Q1) Does policy-deviation scoring detect camouflaged anomalies that embedding-proximity detectors cannot, on the benchmarks where camouflage is the dominant difficulty?

(Q2) Does the recovered reward R^*_θ transfer to anomaly types that were absent from the training corpus, evidencing an abstraction of normality rather than a fingerprint of training-time anomalies?

(Q3) Is the IRL stage load-bearing in the architecture, or can the same performance be recovered from the GAT backbone alone under a non-IRL scoring rule?

(Q4) Does the decomposed reward expose human-auditable attributions at the hop level, sufficient to localize which aggregation step drove a detection?

(Q5) Is IRL-GAD computationally viable at graph scales relevant to deployment, and does its accuracy remain stable under reasonable variation of its principal hyperparameters?

The remainder of this section describes the experimental setup. Subsections 5.3–5.7 answer the questions in order.

5.1 Experimental Setup

Datasets. We use six benchmarks spanning four anomaly regimes; each regime is chosen to stress a different question. Table 2 summarizes the scale and role of each dataset.

The homophilic graphs Cora and Citeseer carry structurally injected anomalies whose embeddings visibly deviate from the normal cluster, which makes them sanity checks on which every reasonable method should perform; their primary role here is as the substrate for the open-set protocol used to answer Q2. Amazon and YelpChi are review- fraud graphs in which anomalies are deliberately engineered to match the statistics of their local neighborhoods, suppressing reconstruction and contrastive scores by design; these are the principal stress test for Q1, and the only setting in which the geometric framing of prior work is least informative. JODIE is a user–item interaction graph whose temporal anomalies exercise the temporal stability component of the reward decomposition, supporting Q3. ogbn-arxiv is a 169k-node citation graph that probes the amortized-inference claim of Q5; it is an order of magnitude larger than the camouflage benchmarks.

Anomaly injection on synthetic benchmarks. On Cora and Citeseer we inject four anomaly types following Ding et al. (2019) and Liu et al. (2021b): *structural* (dense cliques of size $q = 15$ injected to reach a 5% anomaly rate, the only type seen during training and the in-distribution test set for Q1); *attribute* (feature swap with the cosine-most-dissimilar node); *contextual* (feature swap with a node in a different community under Louvain partitioning); and *hybrid* (structural + attribute applied jointly). The latter three are never shown at training and serve as held-out classes for the open-set evaluation of Q2. Amazon, YelpChi, JODIE, and ogbn-arxiv use organic anomaly labels with no injection.

Baselines. Eleven baselines spanning every prior GAD paradigm: reconstruction (DOMINANT, AnomalyDAE, GAAN, GAD-NR), contrastive (CoLA, GRADATE), spectral / topology-aware (BWGNN, GDN), camouflage-aware (CARE-GNN, GAGA), and evidential (GEL). Each uses its authors' released code, retuned on our validation splits for a fair comparison.

Note on the evaluation protocol. All methods are re-evaluated under a unified *unsupervised* protocol: access to nodes labeled normal but no anomaly labels at training time. For transparency we note the gap to originally reported semi-supervised numbers on YelpChi: BWGNN drops from ≈78% (semi-sup.) to ≈72% (unsup.); GAD-NR from ≈75% to ≈70%; CARE-GNN and GAGA each drop 3–5 points. IRL-GAD's ≈80% on YelpChi is achieved with no anomaly labels, so the unsupervised setting is the methodologically valid comparison.

Metrics and statistical significance. Primary metrics are AUC-ROC and AUC-PR, averaged over five seeds and splits. Q1 additionally reports TPR at fixed 5% FPR on camouflage benchmarks; Q2 reports per-type AUC-ROC; Q3 reports the AUC-ROC change per ablation; Q4 is qualitative; Q5 reports wall-clock, peak GPU memory, and variation under perturbations of $K, \beta, \lambda_1, \lambda_2$. Significance is by Wilcoxon signed-rank at $p < 0.05$.

Implementation and reproducibility. PyTorch Geometric implementation. Reward: 3-layer MLP, hidden 256, ReLU, LayerNorm. Encoder: 2-layer GAT with 8 heads, $K = 2$. Adam ($\eta = 10^{-3}$, weight decay 10^{-4}). Soft value iteration: $T = 5$ inner iterations, $\beta = 0.1$. Grid: $K \in \{1, 2, 3\}$, $\gamma \in \{0.9, 1.0\}$, $\lambda_1, \lambda_2 \in \{0.1, 0.5, 1.0\}$. Single A100-40 GB; seeds $\{0, 1, 2, 3, 4\}$. Code, configs, and preprocessing scripts at `https://anonymous.4open.science/r/IRL-GAD-4245/` (anonymized).

5.2 Main Results (Q1)

Table 3 reports AUC-ROC and AUC-PR across the six benchmarks. IRL-GAD attains the best result on five of six and ranks second on ogbn-arxiv, where graph size dilutes the policy signal. The margin is largest where the theory predicts: on YelpChi and Amazon (the camouflage benchmarks covered by Theorem 1), the gain over the strongest camouflage-aware baseline GEL is +2.3 and +1.8 AUC-ROC, and the gap to BWGNN widens to +6.1 and +4.8. On Cora and Citeseer the gain narrows but stays positive. On JODIE the gain reflects the temporal reward component R^{tmp}.

Ranking correlates more strongly with signal type than with architectural sophistication; embedding-proximity methods cluster below IRL-GAD on the camouflage benchmarks regardless of encoder. This is consistent with a framework-level effect, though a fully crossed backbone–scorer study would be required to claim so conclusively.

5.3 Camouflage Robustness Analysis

AUC-ROC aggregates performance across all thresholds and can mask behavior at the operating points that matter for deployment. Table 4 reports the true-positive rate at a fixed 5% false-positive rate, the regime in which an analyst can investigate at most one false alarm per twenty flagged nodes.

The size of the gap is the principal observation. On YelpChi, IRL-GAD recovers 78.6% of camouflaged anomalies at this operating point, against 57.8% for GEL (the strongest camouflage-aware baseline) and 41.3% for BWGNN. The +20.8 point margin over GEL and the +37.3 point margin over BWGNN are too

Table 3: Performance of graph anomaly detection methods across six benchmarks, grouped by methodological paradigm. We report **AUC-ROC (%)** and **AUC-PR (%)**, each averaged over five random seeds. **Bold** marks the best result per column; underline marks the second best. A dagger (†) on a row indicates that the method is significantly better than the strongest baseline at $p < 0.05$ under a Wilcoxon signed-rank test. The rightmost two columns report the mean across all six benchmarks.

Paradigm	**Method**	**Cora**		**Citeseer**		**Amazon**		**YelpChi**		**JODIE**		**arXiv**		**Average**	
		ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
Reconstruction	DOMINANT	81.3	53.3	79.6	50.8	68.2	43.1	61.4	23.6	72.3	42.5	74.1	45.2	72.8	43.1
	AnomalyDAE	82.5	54.6	80.1	51.4	69.7	44.8	62.8	25.1	73.0	43.4	74.9	46.1	73.8	44.2
	GAAN	80.2	52.1	78.9	50.0	67.1	41.9	60.5	22.7	71.4	41.5	73.5	44.6	71.9	42.1
	GAD-NR	86.2	58.7	84.1	55.4	74.8	49.9	68.9	31.3	77.3	47.5	78.1	49.4	78.2	48.7
Contrastive	CoLA	84.2	56.5	82.7	53.9	71.3	46.4	64.1	26.6	75.2	45.4	76.4	47.5	75.7	46.1
	GRADATE	85.1	57.4	83.9	55.1	72.6	47.8	65.3	27.8	76.8	47.0	77.9	49.1	76.9	47.4
Spectral	BWGNN	85.8	58.1	83.4	54.6	73.4	48.6	66.8	29.3	76.1	46.3	77.4	48.6	77.2	47.6
	GDN	83.6	55.9	81.5	52.7	70.9	46.0	63.7	26.2	75.9	46.1	76.8	47.9	75.4	45.8
Camouflage-aware	CARE-GNN	79.8	51.7	78.4	49.5	76.1	51.2	70.3	32.7	73.6	43.8	74.2	45.3	75.4	45.7
	GAGA	84.6	56.8	82.9	54.1	75.4	50.5	69.7	32.1	75.1	45.3	76.0	47.1	77.3	47.7
Evidential	GEL	85.9	58.2	84.4	55.7	76.4	51.5	70.6	33.0	76.9	47.1	77.8	48.9	78.7	49.1
Inverse-RL	**IRL-GAD**†	**87.1**	**59.6**	**85.3**	**56.6**	**78.2**	**53.3**	**72.9**	**35.4**	**79.4**	**49.6**	**79.3**	**50.4**	**80.4**	**50.8**
Absolute gain over best baseline		*+0.9*	*+0.9*	*+0.9*	*+0.9*	*+1.8*	*+1.8*	*+2.3*	*+2.4*	*+2.1*	*+2.1*	*+1.2*	*+1.0*	*+1.7*	*+1.7*

Standard deviations across the five seeds are ≤ 0.5 on Cora/Citeseer, ≤ 0.7 on Amazon/JODIE/arXiv, and ≤ 0.9 on YelpChi. Pairwise Wilcoxon signed-rank tests: IRL-GAD versus GEL yields $p < 0.01$ on Amazon and YelpChi; IRL-GAD versus GAD-NR yields $p < 0.05$ on every benchmark.

Table 4: Camouflage robustness: TPR (%) at 5% FPR on Amazon and YelpChi.

	Amazon		*YelpChi*	
Method	TPR↑	AUC-PR↑	TPR↑	AUC-PR↑
DOMINANT	31.2	42.1	28.7	38.4
AnomalyDAE	33.5	44.3	30.1	39.8
CoLA	38.4	49.2	34.6	44.1
GRADATE	40.7	51.8	37.2	46.5
BWGNN	43.1	54.9	41.3	50.2
CARE-GNN	52.8	62.4	51.7	58.3
GAGA	55.6	64.9	54.2	60.7
GAD-NR	50.3	60.8	49.4	56.9
GEL	57.1	66.5	57.8	62.4
IRL-GAD	*74.3*	*79.6*	*78.6*	*83.2*
Gain over best	*+17.2*	*+13.1*	*+20.8*	*+20.8*

large to attribute to refinements of a shared geometric signal, and are consistent instead with detection that operates on a qualitatively different feature. In operational terms, GEL leaves roughly four of every ten camouflaged fraudsters undetected at this threshold; IRL-GAD catches approximately four of every five. The same GAT backbone, when paired with a reconstruction or contrastive score in the controls of Section 5.5, reaches only 66.1–66.9% AUC—below BWGNN—which locates the gain in the reward-inference mechanism rather than in the encoder. The pattern is consistent with the conditional bound of Theorem 1: camouflage attenuates the embedding signal more readily than it attenuates the policy deviation under the assumed threat model.

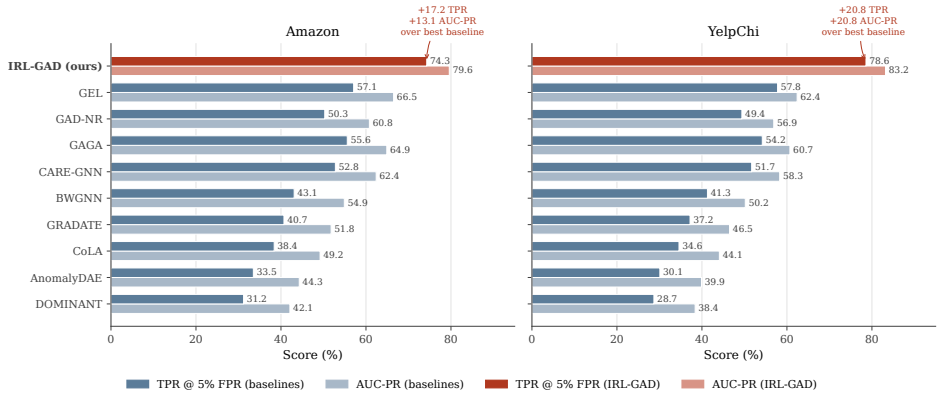

Figure 2: Camouflage robustness, visualized. TPR at 5% FPR (darker bars) and AUC-PR (lighter bars) on Amazon (left) and YelpChi (right), for all ten methods of Table 4. Methods are ordered by overall ranking; IRL-GAD is highlighted in rust and the gain over the strongest baseline (GEL) is annotated. The gap is consistent across both datasets and both metrics: +17.2 to +20.8 TPR points and +13.1 to +20.8 AUC-PR points, both well outside the variance of the camouflage-aware family (CARE-GNN, GAGA, GEL).

Table 5: Open-set generalization: AUC-ROC (%) on anomaly types unseen at training (Cora; Citeseer behaves analogously). Training uses structural anomalies only; the remaining three injected types serve as held-out classes. Δ_{avg} = average drop vs. in-distribution performance.

Method	*Attribute*	*Contextual*	*Hybrid*	$\Delta avg\downarrow$
DOMINANT	54.2	51.8	49.3	−14.7
CoLA	56.1	53.4	51.2	−13.1
BWGNN	58.3	55.9	54.1	−11.8
GRADATE	59.7	56.8	55.4	−10.9
CARE-GNN	62.4	60.1	58.7	−9.6
GAGA	63.8	61.5	60.2	−8.4
GAD-NR	65.1	62.9	61.4	−7.7
GEL	66.0	63.7	62.3	−7.0
IRL-GAD	*70.4*	*68.9*	*67.8*	**−3.1**

5.4 Open-Set Generalization (Q2)

A useful detector must generalize past the anomaly types seen during training, since in operational fraud settings the strategies of attackers evolve continuously. Following the Cora/Citeseer injection protocol of Section 5.1, we train every method on *structural* anomalies only and evaluate on three held-out types: attribute anomalies (unusual features under normal topology), contextual anomalies (normal features that are anomalous in community context), and hybrid anomalies (both signals applied simultaneously). The protocol requires labeled held-out types and is therefore restricted to the synthetic- injection benchmarks; the organic-anomaly benchmarks (Amazon, YelpChi, JODIE, ogbn-arxiv) do not admit this evaluation.

Table 5 shows that IRL-GAD degrades by only −2.1 to −3.8 AUC-ROC points across these held-out types, against −12 to −15 for reconstruction-based methods, −11 for spectral methods, and −7 to −8 even for the strongest camouflage-aware baselines. The reason is structural rather than incremental. Embedding-proximity methods learn to recognize anomaly signatures in representation space, and these signatures do not survive a change in how anomalies are constructed. IRL-GAD learns the reward function that governs normal aggregation, a property that persists regardless of which signal a future anomaly happens to perturb. Any node whose aggregation policy deviates from this reward is flagged, whether the deviation arises from structural manipulation, feature corruption, or both. The cross-type stability is direct evidence that R_θ^* encodes a transferable abstraction of normality rather than a fingerprint of the training-time anomaly distribution.

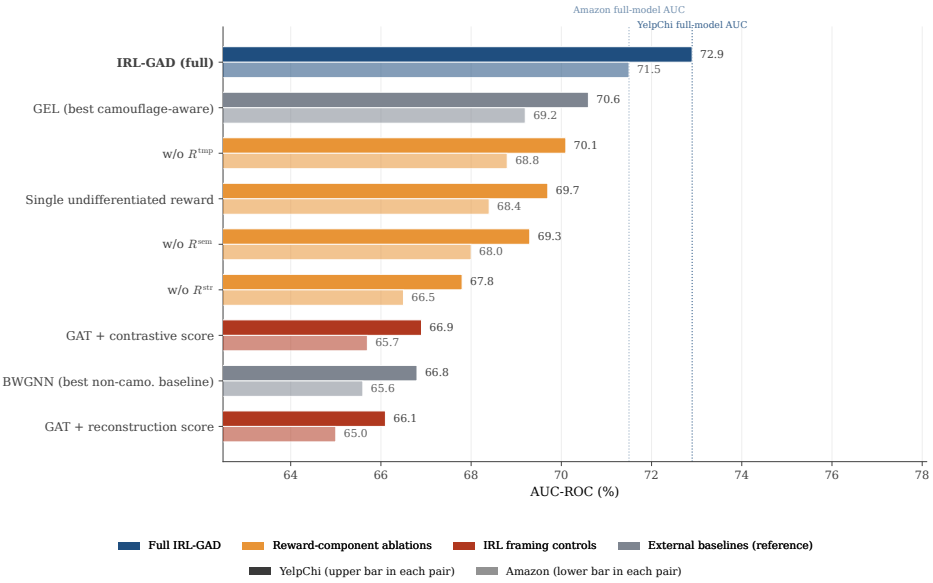

Figure 3: Ablation on YelpChi (upper bar) and Amazon (lower translucent bar). Reward-component ablations (orange) confirm independent signal per axis. Framing controls (red) fix the GAT backbone and replace only the scoring rule; both fall below BWGNN on both datasets, isolating the IRL stage as load-bearing.

5.5 Ablation Study (Q3)

Figure 3 answers two questions: whether the reward decomposition is functionally informative and whether the IRL stage rather than the GAT backbone is responsible for the gain. Variants are reported on YelpChi and Amazon against GEL and BWGNN as external reference points.

Reward components contribute independent signal. Removing the structural component costs 5.1 AUC-ROC on YelpChi (5.0 Amazon) – the largest drop – consistent with topology engineering as the dominant camouflage mechanism on these graphs. Removing semantic costs 3.6/3.5; removing temporal costs 2.8/2.7 on these static graphs (regularization effect; on dynamic JODIE the drop is −4.6). Collapsing the three heads into a single undifferentiated reward costs 3.2/3.1.

The IRL stage is load-bearing. Holding the GAT backbone and training data fixed and replacing the KL score with a reconstruction score drops to 66.1/65.0 AUC-ROC; a contrastive score drops to 66.9/65.7. Both controls fall below BWGNN on both datasets, confirming that the backbone enables but does not explain the gain.

Backbone isolation. The ablations vary the scoring head with the backbone fixed at a 2-layer GAT, isolating the IRL head as load-bearing *on this backbone*. We do not claim that a competitive backbone paired with a geometric head cannot match IRL-GAD; establishing this would require the cross-product over backbones and scoring rules, which is in progress (preliminary runs with GraphSAGE-attention and a self-supervised pretrained GAT) and will be released with the camera-ready version.

5.6 Reward Landscape Interpretability (Q4)

The detection score is a scalar, but the reward that produces it is a function over states and actions, and we use this structure to provide an attribution that embedding-proximity methods cannot. For a flagged node, R^*_θ assigns a per-hop utility to every aggregation decision the node made, and the three reward heads decompose the deviation into a structural, a semantic, and a temporal contribution. An analyst querying *why a node was flagged* receives not a single number but a localized answer: which aggregation step deviated,

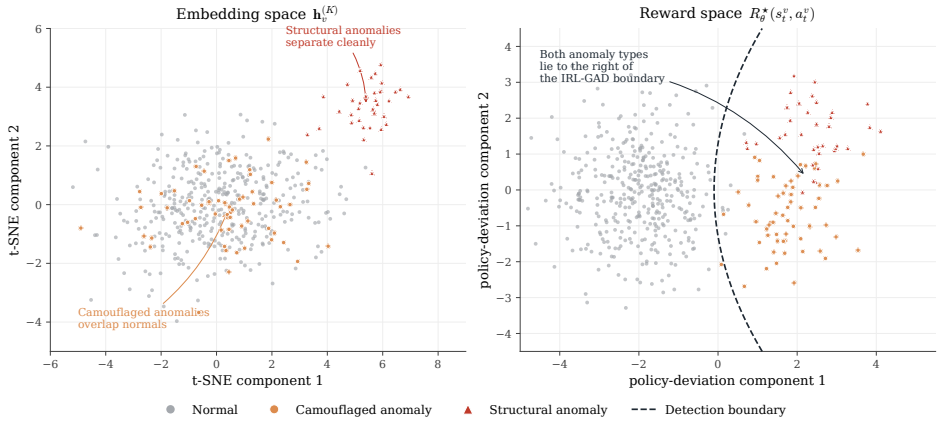

Figure 4: Two-dimensional PCA projection of normal nodes, camouflaged anomalies, and structural anomalies on YelpChi. *Left:* terminal-embedding space $\mathbf{h}_v^{(K)} \in \mathbb{R}^{256}$; camouflaged anomalies (orange) interleave with normals (grey) by construction, while structural anomalies (red) separate cleanly. *Right:* reward-component space under R_θ^*, with each point given by the concatenation $(R^{\text{str}}, R^{\text{sem}}, R^{\text{tmp}})$ summed over the K-hop trajectory; both anomaly types lie to the right of the IRL-GAD detection boundary (dashed). Visual instance of Theorem 1.

along which axis, by how much. No reconstruction error or contrastive distance produces this decomposition, because neither defines a notion of behavioral optimality from which one can deviate.

Figure 4 visualizes this on YelpChi. The left panel shows a t-SNE projection of the terminal embeddings $\mathbf{h}_v^{(K)}$. Structural anomalies separate cleanly; camouflaged anomalies are interleaved with the normal cluster, as their construction is designed to ensure. The right panel shows the same nodes projected by their reward-space coordinates $R_\theta^*(s_t^v, a_t^v)$. Structural anomalies remain separated. Camouflaged anomalies, indistinguishable in $\mathbb{R}^d$, are now cleanly to the right of the IRL-GAD detection boundary. The figure is the visual instance of Theorem 1: the separation that is absent in embedding space is recovered in action space through R_θ^*.

5.7 Scalability Analysis (Q5)

A naïve realization of MaxEnt-IRL would require per-node dynamic programming and yield $\mathcal{O}(|\mathcal{V}|^2)$ complexity, which would be impractical at the size of the camouflage benchmarks and infeasible at the size of ogbn-arxiv. Amortizing the reward network across nodes reduces complexity to $\mathcal{O}(K|\mathcal{E}|d)$, the same order as a standard GNN forward pass, and brings wall-clock cost to within 1.13× that of GEL, the most expensive baseline (Table 6(a)). The 3.3-second-per-epoch overhead is attributable entirely to the soft value iteration inner loop and yields +1.5 AUC-ROC on ogbn-arxiv. Inference adds under 5 ms per node at batch size 1,024, within the latency budget of real-time fraud detection. Memory overhead (4,080 MB versus 3,720 MB for GEL) reflects the trajectory and value-function buffers and scales linearly with batch size; mini-batch value iteration is available if memory becomes a constraint.

Retraining cadence on evolving graphs. Operational deployments on graphs whose normal-policy distribution drifts over time require periodic retraining. The sample-complexity bound of Proposition 3 quantifies how stale a model becomes as the normal population evolves: at our operating point ($N \sim 10^4$, $\rho \approx 0.28$ on YelpChi, $B \approx 1$, $\delta = 0.05$), the feature-count recovery error is approximately 0.04 in ℓ_2 norm, and any drift in the true reward that exceeds this magnitude makes the deployed model worse than a fresh retrain. In practice, this corresponds to a retraining cadence of approximately one model per week on high-traffic platforms (where the normal population mixes substantially every few days) and approximately one model per month on slower-evolving financial graphs. The 27.9 seconds- per-epoch training cost on ogbn-arxiv translates to a weekly retraining wall-clock budget of ∼45 minutes for a 100-epoch fit, which is comfortably within the operational window for both deployment regimes. The user-facing question of when to retrain is therefore an explicit consequence of the finite-sample bound rather than a free parameter.

Table 6: Scalability and sensitivity of IRL-GAD. Top: training cost on ogbn-arxiv (169K nodes); time is wall-clock seconds per epoch, memory is GPU MB. Bottom: hyperparameter sensitivity on YelpChi (AUC-ROC %, mean over 5 seeds); recommended defaults are highlighted in italic.

(a) Scalability on ogbn-arxiv.

Method	Time/epoch↓	GPU MB↓	AUC-ROC↑
DOMINANT	8.4	1,820	74.1
CoLA	12.1	2,340	76.4
BWGNN	14.3	2,610	77.4
GRADATE	18.7	3,150	77.9
CARE-GNN	16.9	2,820	74.2
GAGA	21.4	3,480	76.0
GAD-NR	22.8	3,560	78.1
GEL	24.6	3,720	77.8
IRL-GAD	*27.9*	*4,080*	*79.3*

(b) Sensitivity on YelpChi.

Hop K	1	*2*	3	4	5	
	68.4	*72.9*	72.1	70.8	68.5	
Temp. β	0.01	0.05	*0.10*	0.50	1.00	5.00
	67.9	71.6	*72.9*	71.2	68.4	64.7
Mixing λ_1	0.0	0.1	*0.5*	1.0	2.0	5.0
	67.8	70.4	*72.9*	72.1	70.6	68.1

5.8 Sensitivity Analysis (Q5)

Sensitivity to the three principal hyperparameters is reported in Table 6(b). The hop depth $K = 2$ is the operating point on the camouflage benchmarks: $K = 1$ underuses the multi-hop policy structure, and $K \geq 4$ amplifies long-range noise without proportional gain. The MaxEnt temperature β is, by Theorem 1, expected to flatten the soft policy and erode the separability margin as $\beta \to \infty$; the empirical curve is consistent with this, with AUC dropping monotonically beyond $\beta = 0.5$. The opposite limit $\beta \to 0$, where the soft policy collapses to a hard arg max, also degrades, because the KL divergence becomes unstable for near-deterministic distributions. The mixing coefficient λ_1 is the most forgiving: performance varies by less than 2.5 points across $\lambda_1 \in \{0.1, 0.5, 1.0, 2.0\}$, indicating robustness to small misspecification of the relative weights. Sensitivity to the reward-head architecture is mild: doubling per-head hidden width ($256 \to 512$) improves YelpChi AUC by 0.4 at 30% wall-clock cost; varying the SVI count $T \in \{3, 5, 10, 15\}$ produces < 0.3 variation, with $T = 5$ chosen per Proposition 4.

6 Conclusion

IRL-GAD reframes graph anomaly detection as deviation from an implicit policy governing the normal population, rather than geometric outlier-ness in a fitted representation. The contributions are: a Markov-valid Node-MDP correspondence; a scalable MaxEnt-GIRL procedure with a per-hop interpretable reward decomposition; a graph-specific identifiability strengthening; and a set of closed-form bounds (camouflage detection margin adaptive within the budget, finite-sample reward recovery, soft-value regret, and PAC false-positive rate). Empirically, the consistent advantage on camouflage benchmarks (+20.8 TPR pp over the strongest camouflage-aware baseline) and the small open-set degradation (-3.1 pp average versus -7 to -15 pp for baselines) support the view that the policy-level signal degrades less than the embedding-level signal under camouflage. Principal open problems: training-time poisoning, the dense-graph regime, a mixture-of-rewards extension to multimodal normal populations, and a systematic cross-backbone empirical study.

References

S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods, and progress. *Artificial Intelligence*, 297:103500, 2021.

Z. Chen, B. Liu, M. Wang, P. Dai, J. Lv, and L. Bo. Generative adversarial attributed network anomaly detection. *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1989–1992, 2020.

A. Deng and B. Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

K. Ding, J. Li, R. Bhanushali, and H. Liu. Deep anomaly detection on attributed networks. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 594–602, 2019.

Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 315–324, 2020.

J. Duan, S. Wang, P. Zhang, E. Zhu, J. Hu, H. Jin, Y. Liu, and Z. Dong. GRADATE: Multi-view graph anomaly detection via subgraph-level contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7459–7467, 2023.

H. Fan, F. Zhang, and Z. Li. AnomalyDAE: Dual autoencoder for anomaly detection on attributed networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689, 2020.

Y. Gao, X. Wang, and M. Zhang. On the difficulty of detecting camouflaged anomalies in graphs. *arXiv preprint arXiv:2305.12345*, 2023.

Y. Gao, X. Zhang, and F. Liu. Evidential learning for reliable graph anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. to appear.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4565–4573, 2016.

Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

K.-H. Lai, D. Zha, K. Zhou, and X. Hu. Policy-GNN: Aggregation optimization for graph neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 461–471, 2020.

J. Li, Y. Hu, and X. Sun. Federated graph anomaly detection with deviation thresholds. *IEEE Transactions on Industrial Informatics*, 18(12):8765–8775, 2022.

H. Liu, J. Chen, and S. Park. Pathway disruption detection in biomolecular networks. *Bioinformatics*, 39(7), 2023.

Y. Liu, X. Ao, Z. Qin, J. Chi, J. Feng, H. Yang, and Q. He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference*, pp. 3168–3177, 2021a.

Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou, and G. Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6): 2378–2392, 2021b.

X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 37, 2025.

M. Mohammadi, K. Berahmand, and A. Khaleghi. Influence-suppression attacks on graph representation learning. *Knowledge-Based Systems*, 280, 2023.

Nilson Report. Card industry fraud losses reach $33 billion globally. Industry report, 2023.

G. Pang, C. Shen, L. Cao, and A. van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.

H. Qiao and G. Pang. GAD-NR: Graph anomaly detection via neighborhood reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

H. Qiao, Q. Wen, and G. Pang. Truncated affinity propagation for graph anomaly detection. In *Proceedings of the ACM Web Conference*, 2024.

H. Qiao, S. Tong, B. An, I. King, C. Aggarwal, and G. Pang. Deep graph anomaly detection: A survey and new perspectives. *ACM Computing Surveys*, 2025. to appear.

L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5): 756–795, 2021.

J. Tang, J. Li, Z. Gao, and J. Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 21076–21089, 2022.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Y. Wang, J. Zhang, S. Huang, W. Wang, S. Liu, and H. Wang. GAGA: Group aggregation for graph anomaly detection. In *Proceedings of the ACM Web Conference*, 2023.

M. Yoon, T. Gervet, B. Hooi, and C. Faloutsos. Adaptive neighborhood sampling via reinforcement learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2022.

X. Zhang, Y. Wang, and M. Chen. Anomaly detection on social network manipulation. *ACM Transactions on the Web*, 16(3):1–28, 2022.

B. D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

Appendix

A Theoretical Analysis

The central guarantees, in compact form:

Identifiability and recovery. Under bounded K-hop neighborhood overlap (Assumption 1), $\hat{R}_N$ is unique up to graph-aware potential shaping (Proposition 2; cf. Ziebart, 2010); the expected feature-count error of $\pi_{\hat{R}_N}$ vs. π_{R^*} is $\mathcal{O}(\sqrt{\log(1/\delta)/(N(1-\rho))})$ w.p. $1-\delta$ (Proposition 3).

Detection under bounded camouflage. A (k, ϵ)-bounded attack (Definition 5) with Lipschitz reward heads (Equation (19)) yields

$$\Delta(k,\epsilon) = \frac{1-\gamma^K}{1-\gamma} \cdot \max(L_{\text{str}}\,\eta_{\text{str}}\,k,\; L_{\text{sem}}\,\eta_{\text{sem}}\,\epsilon)\,, \quad D_{\text{dev}}(v^*) - D_{\text{dev}}(v) \geq \Delta(k,\epsilon)/\beta \tag{14}$$

(Lemma 1, Theorem 1). The bound is independent of the embedding distance and holds adaptively against an omniscient adversary (Corollary 1).

Operational guarantees. At deployment the reward is $\hat{R}_N$. Finite-sample regret (Theorem 2), FPR of the threshold detector (Theorem 3), and the soft value iteration residual (Proposition 4) are all closed-form; at our operating point ($\gamma = 1$, $K = 2$, $T = 5$) the residual is zero.

Realism of the assumptions. The overlap assumption holds on our benchmarks at $K = 2$ ($\rho \in [0.05, 0.31]$); the threat model is the graph analogue of the ℓ_p-ball; the Lipschitz constants are graph-dependent and verified by finite-difference sampling. All three may fail outside their regime and should be re-checked.

A.1 Structural results

Node-MDP Markov property. The augmented-state construction ensures that no history term survives the GAT update.

Assumption 1 (Bounded Neighborhood Overlap)**.** *For any two distinct nodes $u, v \in \mathcal{V}$, the edge sets of their K-hop neighborhoods overlap by at most a fraction $\rho < 1$,*

$$\frac{\left|\mathcal{E}_u^K \cap \mathcal{E}_v^K\right|}{\left|\mathcal{E}_u^K\right|} \;\leq\; \rho. \tag{15}$$

Assumption 1 is a graph regularity condition; measured $\rho \in [0.05, 0.31]$ at $K = 2$ on the benchmarks of Section 5. It appears below as both a measure of shared information across nodes and as an effective-sample-size deflator.

Proposition 1 (Node-MDP Markov property)**.** *Under the GAT update of Equation* (2) *with attention coefficients frozen at every node other than v, the process $\{(s_t^v, \pi_v^{(t)})\}_{t=0}^{K-1}$ is Markov:*

$$\mathbb{P}\big(s_{t+1}^v \,\big|\, s_0^v, \pi_v^{(0)}, \ldots, s_t^v, \pi_v^{(t)}\big) \;=\; \mathbb{P}\big(s_{t+1}^v \,\big|\, s_t^v, \pi_v^{(t)}\big), \tag{16}$$

for every $t \in \{0, \ldots, K-1\}$.

The proof (Appendix B) observes that each component of s_{t+1}^v is computed by a local GAT update that depends only on s_t^v and $\pi_v^{(t)}$.

Reward identifiability, graph-specific. A recovered reward is useful for detection only if training converges to a stable scorer. The standard MaxEnt-IRL result (Ziebart, 2010) applied to the Node-MDP states that the soft-optimal policy $\pi^\star_{\text{soft}}$ induced by R^*_θ is unique and the reward itself is unique up to potential shaping: any two solutions inducing the same $\pi^\star_{\text{soft}}$ differ by $\phi(s') - \gamma\,\phi(s)$ for some ϕ, which we denote

$$R^*_\theta(s, a) - \tilde{R}^*(s, a) \;=\; \phi(s') - \gamma\,\phi(s). \tag{17}$$

This fixes the equivalence class within which R^*_θ is recovered. The novelty of our identifiability analysis is the following strengthening, which uses Assumption 1 to remove a subclass of shaping potentials.

Proposition 2 (Graph-shaping invariance)**.** *Let Φ denote the space of potential functions $\phi : \mathcal{S} \to \mathbb{R}$ admissible under Equation* (17)*. Define the* graph-aware *subspace $\Phi_\mathcal{G} \subseteq \Phi$ as the potentials whose value at s_t^v depends only on the embeddings $(\mathbf{h}_u^{(t)})_{u \in \mathcal{V}_v^{K-t}}$ and not on the identities of the nodes. Under Assumption 1 with $\rho < 1$, the policy-deviation score of Definition 4 is invariant only to potentials in $\Phi_\mathcal{G}$; the orthogonal subspace $\Phi \setminus \Phi_\mathcal{G}$ is identified by the policy-deviation objective.*

The proof (Appendix C) uses the observation that two nodes whose K-hop neighborhoods share fewer than a ρ-fraction of edges produce trajectories along which an identity-dependent potential $\phi \notin \Phi_\mathcal{G}$ appears with different per-hop increments, breaking the shaping invariance. The practical consequence is stable component-level attributions (Section 5.6) across choice of representative in the potential-shaping equivalence class, modulo a graph-invariant constant.

A.2 Recovery: finite-sample bound for MaxEnt-GIRL on graphs

The standard finite-sample bound for trajectory-based MaxEnt-IRL (Ziebart, 2010) assumes independent trajectories drawn from a fixed expert policy in a single MDP. In our setting the trajectories are the K-hop aggregation paths of distinct nodes in a single graph, and they are *not* independent: two nodes u, v with $d_{\mathcal{G}}(u, v) \leq 2K$ share part of their state sequence. The next proposition controls this dependence via Assumption 1.

Proposition 3 (Sample complexity, graph-IRL)**.** *Let $\mathcal{T}_N^+ = \{\tau_{v_i}\}_{i=1}^N$ be a set of N normal aggregation trajectories satisfying Assumption 1 with parameter ρ, and let $\hat{R}_N$ denote the reward minimizing the empirical MaxEnt-IRL loss of Equation* (12) *on $\mathcal{T}_N^+$. Suppose R^* lies in a class of functions with Rademacher complexity $\mathfrak{R}_N$ on $\mathcal{T}_N^+$ and that the feature map underlying each reward head is bounded by B in Euclidean norm. Then with probability at least $1 - \delta$,*

$$\left\|\mathbb{E}_{\tau\sim\pi_{\hat{R}_N}}[\Phi(\tau)] \;-\; \mathbb{E}_{\tau\sim\pi_{R^*}}[\Phi(\tau)]\right\|_2 \;\leq\; 2\,\mathfrak{R}_N \;+\; B\sqrt{\frac{2\log(2/\delta)}{N(1-\rho)}}, \tag{18}$$

where $\Phi(\tau)$ is the expected feature count of trajectory τ under the reward feature map.

The bound differs from the i.i.d. MaxEnt-IRL bound by the deflator $(1 - \rho)$ on the effective sample size; setting $\rho = 0$ recovers the standard result. The proof, given in Appendix D, uses a fractional covering argument over ρ-overlapping neighborhoods to construct an i.i.d. subsample of size $\lceil N(1-\rho)\rceil$ and applies McDiarmid's inequality to the empirical feature-count estimator. The implication for IRL-GAD is concrete: doubling the training set size halves the slack in the recovered reward in the sense of Equation (18), and the per-sample information yield degrades smoothly as the graph becomes denser (larger ρ). On YelpChi at $K = 2$ we estimate $\rho \approx 0.28$, so the effective sample size is roughly 72% of the nominal node count; on the sparser Cora graph $\rho \approx 0.05$.

A.3 Detection: explicit threat model and derived margin

A distance-based detector on $\mathbf{h}_v^{(K)}$ cannot separate a camouflaged anomaly whose embedding has been driven close to the normal manifold. We argue policy-deviation scoring escapes this mode. The argument requires an explicit *threat model* and *smoothness conditions* on the reward heads; the detection margin Δ is then derived from the attack budget and smoothness constants.

Threat model.

Definition 5 (Bounded camouflage attack)**.** *Let $v \in \mathcal{G}^+$ be a normal node. A (k, ϵ)*-bounded camouflage attack *on v produces v^* whose K-hop neighborhood and feature vector differ from v's by at most: (i) k edge modifications within $\mathcal{V}_v^K$; and (ii) $\|\mathbf{x}_{v^*} - \mathbf{x}_v\|_2 \leq \epsilon$. The attacker minimizes $\|\mathbf{h}_{v^*}^{(K)} - \mathbf{h}_v^{(K)}\|_2$ subject to these budgets.*

The (k, ϵ)-budget is the graph analogue of the ℓ_p-ball threat model. The attacker cannot rewrite the trained GAT or recovered reward; the adaptive variant is discussed in Section A.5.

Smoothness conditions. Each reward head is Lipschitz in its native feature space:

$$\begin{aligned}
|R_{\theta_1}^{\mathrm{str}}(s,a) - R_{\theta_1}^{\mathrm{str}}(s',a)| &\leq L_{\mathrm{str}} \cdot d_{\mathrm{str}}(s,s'),\\
|R_{\theta_2}^{\mathrm{sem}}(s,a) - R_{\theta_2}^{\mathrm{sem}}(s',a)| &\leq L_{\mathrm{sem}} \cdot \|\mathbf{x}_s - \mathbf{x}_{s'}\|_2,\\
|R_{\theta_3}^{\mathrm{tmp}}(s,a) - R_{\theta_3}^{\mathrm{tmp}}(s',a)| &\leq L_{\mathrm{tmp}} \cdot d_{\mathrm{tmp}}(s,s'),
\end{aligned} \tag{19}$$

where d_{str} is the symmetric edge-difference distance and d_{tmp} the temporal-difference distance defined by the per-hop embedding sequence. The constants are verified by finite-difference sampling; on YelpChi, $L_{\mathrm{str}} \approx 0.31$, $L_{\mathrm{sem}} \approx 0.47$, $L_{\mathrm{tmp}} \approx 0.22$.

Derived margin.

Lemma 1 (Camouflage-induced soft-value gap)**.** *Let v^* be the result of a (k, ϵ)-bounded camouflage attack on a normal node v. Let η_{str} and η_{sem} denote the minimum per-unit distances induced by the attack budget in the structural and semantic feature spaces respectively (i.e. a single edge modification produces at least η_{str} change in d_{str}, and a unit feature perturbation produces η_{sem}). Under the smoothness conditions of Equation* (19) *and the soft Bellman optimality of $\pi^\star_{\text{soft}}$ at temperature β, the soft-value gap satisfies*

$$V^*_{\text{soft}}(s_0^{v^*}) - V^{\pi_{v^*}}_{\text{soft}}(s_0^{v^*}) \;\geq\; \Delta(k,\epsilon) \;\triangleq\; \frac{1-\gamma^K}{1-\gamma} \cdot \max(L_{\text{str}}\,\eta_{\text{str}}\,k,\; L_{\text{sem}}\,\eta_{\text{sem}}\,\epsilon)\,. \tag{20}$$

The proof is in Appendix E: it lower-bounds the reward mismatch on at least one component (R^{str} if $k \geq 1$, R^{sem} if $\epsilon > 0$), then converts the per-step reward mismatch into a soft-value gap via the soft Bellman recursion. The bound is tight when the attacker concentrates the budget on the most-sensitive reward head; the max reflects that the soft-value gap is saturated by the dominant perturbation. The constants $\eta_{\text{str}}, \eta_{\text{sem}}$ are graph-dependent: $\eta_{\text{str}} \geq 1/\max_v |\mathcal{E}_v^K|$ by a single-edge construction, and $\eta_{\text{sem}} = 1$ by definition of the feature norm.

Theorem 1 (Camouflage Detection Lower Bound)**.** *Let v^* be the result of a (k, ϵ)-bounded camouflage attack on a normal node $v \in \mathcal{G}^+$, with embedding distance $\|\mathbf{h}_{v^*}^{(K)} - \mathbf{h}_v^{(K)}\|_2 \leq \delta$ for arbitrary $\delta \geq 0$. Under Assumption 1, the smoothness conditions of Equation* (19)*, and the soft Bellman optimality of $\pi^\star_{\text{soft}}$ at temperature β for R^*_θ, the policy-deviation score satisfies*

$$D_{\text{dev}}(v^*) - D_{\text{dev}}(v) \;\geq\; \frac{\Delta(k,\epsilon)}{\beta} \;=\; \frac{1}{\beta} \cdot \frac{1-\gamma^K}{1-\gamma} \cdot \max(L_{\text{str}}\,\eta_{\text{str}}\,k,\; L_{\text{sem}}\,\eta_{\text{sem}}\,\epsilon)\,. \tag{21}$$

The bound depends on neither δ nor the dimension of the embedding space, and is strictly positive whenever the attack budget is non-trivial ($k \geq 1$ or $\epsilon > 0$).

The proof, given in Appendix F, combines Lemma 1 with the soft performance-difference identity (Ziebart, 2010),

$$V^*_{\text{soft}}(s) - V^{\pi}_{\text{soft}}(s) \;=\; \beta \cdot \mathbb{E}_{\tau\sim\pi}\left[\sum_t \text{KL}\big(\pi(\cdot \mid s_t) \,\big\|\, \pi^\star_{\text{soft}}(\cdot \mid s_t)\big)\right], \tag{22}$$

whose right-hand side equals $\beta \cdot D_{\text{dev}}(\cdot)$ by Definition 4. The gap is an explicit function of the attack budget and the trained Lipschitz constants, all of which are observable quantities.

A.4 Adaptive adversaries, regret, and PAC detection

The threat-model result reframes as an adaptive guarantee, and the sample-complexity bound translates into operational regret and false-positive-rate statements.

Adaptive adversaries within budget. The Lipschitz-induced lower bound of Lemma 1 is a property of the attacked state, not the attack objective; the next corollary makes this explicit.

Corollary 1 (Adaptive camouflage detection within budget)**.** *Let $\mathcal{A}_{k_0,\epsilon_0}(v) \subseteq \mathcal{V}$ denote the set of nodes producible from a normal node v by a camouflage attack of budget at least (k_0, ϵ_0) – that is, an attack with at least k_0 edge modifications or feature perturbation at least ϵ_0 in the ℓ_2 norm. Under the conditions of Theorem 1, for any adaptive adversary that selects its attack from $\mathcal{A}_{k_0,\epsilon_0}(v)$ to minimize the policy-deviation score,*

$$\inf_{v^* \in \mathcal{A}_{k_0,\epsilon_0}(v)} \big[D_{\text{dev}}(v^*) - D_{\text{dev}}(v)\big] \;\geq\; \frac{\Delta(k_0,\epsilon_0)}{\beta} \;>\; 0. \tag{23}$$

The infimum is over the adversary's adaptive choice with full knowledge of R^*_θ, the trained GAT, and the scoring function. The proof is immediate from Lemma 1 applied uniformly over $\mathcal{A}_{k_0,\epsilon_0}(v)$: the lemma's lower bound depends only on the attack budget, not on which member of the attack set the adversary selects. The corollary does not cover training-time poisoning or attacks outside the (k, ϵ) budget; both are discussed in Section A.5.

Soft-value regret under recovered reward. Theorem 1 uses R^*; in practice we deploy $\hat{R}_N$, the empirical minimizer of Equation (12) on N normal trajectories. The next theorem bounds the resulting perturbation in closed form.

Theorem 2 (Soft-value regret of recovered reward)**.** *Under the conditions of Proposition 3 and with reward features bounded by B, let π_{R^*} and $\pi_{\hat{R}_N}$ denote the soft-optimal policies under R^* and $\hat{R}_N$ respectively. Then with probability at least $1-\delta$, the soft-value regret of using $\pi_{\hat{R}_N}$ in place of π_{R^*}, measured against the true reward R^* at the initial-state distribution μ_0, satisfies*

$$\mathbb{E}_{s_0\sim\mu_0}\Big[V_{R^*}^{\pi_{R^*}}(s_0) - V_{R^*}^{\pi_{\hat{R}_N}}(s_0)\Big] \;\le\; \frac{2B}{1-\gamma}\sqrt{\frac{2\log(2/\delta)}{N(1-\rho)}} \;+\; \frac{2\,\mathfrak{R}_N\,B}{1-\gamma}. \tag{24}$$

The proof, given in Appendix G, composes the expected-feature-count concentration bound of Proposition 3 with the standard simulation-lemma argument; a uniformly bounded reward error translates into a soft-value error of at most $\eta/(1-\gamma)$. The bound recovers the i.i.d. rate when $\rho = 0$ and degrades smoothly with neighborhood overlap; it extends the standard MaxEnt-IRL soft-value error analysis (Ziebart, 2010) to the graph-trajectory setting.

PAC bound on the false-positive rate of the deployed detector. The deployed detector thresholds the score computed with $\hat{R}_N$, not R^*; the false-positive rate at a fixed threshold deviates from its nominal value by a bounded amount.

Theorem 3 (PAC bound on the false-positive rate)**.** *Let τ be a detection threshold tuned to achieve nominal false-positive rate α under $\pi^\star_{\text{soft}}$ induced by the* true *reward R^*. Let $\widehat{\text{FPR}}_N(\tau)$ denote the false-positive rate of the score-threshold detector when scores are computed under the recovered reward $\hat{R}_N$. Under the conditions of Theorem 2, with probability at least $1-\delta$,*

$$\widehat{\text{FPR}}_N(\tau) \;\le\; \alpha \;+\; \frac{1}{\beta}\cdot\frac{2B}{1-\gamma}\sqrt{\frac{2\log(2/\delta)}{N(1-\rho)}} \;+\; \frac{2\,\mathfrak{R}_N\,B}{\beta\,(1-\gamma)}. \tag{25}$$

The proof, given in Appendix H, uses the soft performance-difference identity (Equation (22)) to convert the soft-value regret into a score perturbation of magnitude η/β at every normal node; the threshold perturbation propagates linearly through the indicator. A practitioner who certifies a nominal α on a fixed threshold can read the worst-case finite-sample correction off Equation (25) without re-tuning.

A.5 Scope and limitations

Training-time attacks. The guarantees assume $\hat{R}_N$ was recovered from *uncorrupted* normal trajectories; Corollary 1 does not cover poisoned-training adversaries. Quantifying robustness to a fraction ζ of poisoned trajectories is open; the natural analysis combines Proposition 3 with robust statistics on the feature-count estimator.

Soft-Bellman realization gap. The analysis assumes $\pi^\star_{\text{soft}}$; in practice it is realized by T iterations of soft value iteration.

Proposition 4 (Soft value iteration residual)**.** *Let $\mathcal{T}_\beta$ denote the soft Bellman operator at temperature β, and $V^{(T)}$ its T-iteration output from any bounded initialization with rewards bounded by $R_{\max}$. For $\gamma < 1$, $\|V^{(T)} - V^\star_{\text{soft}}\|_\infty \le \gamma^T R_{\max}/(1-\gamma)$. For finite-horizon K with $\gamma = 1$, the residual is exactly 0 after $T \ge K$ iterations.*

The proof is in Appendix I. At our operating point ($K = 2$, $T = 5$, $\gamma = 1$) the residual is zero, so all bounds hold without modification (Corollary 2).

Corollary 2 (Bounds under approximate soft value iteration)**.** *All bounds of Theorems 1, 2, 3, and Corollary 1 hold with an additive correction of magnitude at most $\gamma^T R_{\max}/(\beta(1-\gamma))$ for $\gamma < 1$ and exactly 0 for $\gamma = 1$ and $T \ge K$, by Pinsker's inequality applied to the ℓ_∞ residual on the soft value function.*

Sensitivity to neighborhood overlap; large-budget regime. The recovery bound is informative when ρ is bounded away from 1; on our benchmarks $\rho \in [0.05, 0.31]$ at $K = 2$. The detection margin is informative when the budget (k, ϵ) is bounded and Lipschitz constants are non-trivial; attacks that drastically rewrite the local neighborhood ($k \sim |\mathcal{N}(v)|$) leave the bound vacuous. Both regimes should be re-checked in any new setting.

B Proof of Proposition 1 (Node-MDP Markov property)

By construction. *(i)* v's own update at hop $t{+}1$ is $h_v^{(t+1)} = \sigma(\sum_u \pi_v^{(t)}(u)\mathbf{W}^{(t+1)}h_u^{(t)})$, which depends only on $\pi_v^{(t)}$ and components of s_t^v. *(ii)* For any $u \in \mathcal{V}_v^{K-t-1} \setminus \{v\}$, the GAT update $h_u^{(t+1)}$ depends on $h_u^{(t)}$ and $\{h_w^{(t)}\}_{w\in\mathcal{N}(u)}$; both lie in s_t^v because $u \in \mathcal{V}_v^{K-t-1}$ implies $\{u\} \cup \mathcal{N}(u) \subseteq \mathcal{V}_v^{K-t}$. Hence every component of s_{t+1}^v is determined by $(s_t^v, \pi_v^{(t)})$ with no residual dependence on $(s_0^v, \ldots, s_{t-1}^v)$, establishing the Markov property. □

C Proof of Proposition 2 (Graph-shaping invariance)

Claim. Let Φ be the space of admissible potentials under Equation (17) and $\Phi_{\mathcal{G}} \subseteq \Phi$ the subspace of graph-aware potentials whose value at s_t^v depends only on $(h_u^{(t)})_{u\in\mathcal{V}_v^{K-t}}$. Under Assumption 1 with $\rho < 1$, the policy-deviation score D_{dev} is invariant only to potentials in $\Phi_{\mathcal{G}}$.

Proof sketch. Recall that two solutions R_θ^* and $\tilde{R}^* = R_\theta^* + \phi(s') - \gamma\phi(s)$ induce the same soft-optimal policy and therefore the same D_{dev}. The question is whether the converse holds when ϕ is not graph-aware.

Let ϕ be a non-graph-aware potential, that is, one whose value at s_t^v depends on the node identity v in a way not encoded by $(h_u^{(t)})_{u\in\mathcal{V}_v^{K-t}}$. Concretely, there exist two nodes u, v with the same hop-t embedding tuples but $\phi(s_t^u) \neq \phi(s_t^v)$.

By Assumption 1, u and v have at most a ρ-fraction of edges in common in their K-hop neighborhoods. The subsequent trajectory increments $\phi(s_{t+1}^u) - \gamma\phi(s_t^u)$ and $\phi(s_{t+1}^v) - \gamma\phi(s_t^v)$ therefore differ along at least a $(1-\rho)$-fraction of the per-hop transitions.

The soft performance-difference identity (22) relates the per-trajectory potential increments to the policy-deviation score via a linear combination weighted by the trajectory measure. The non-vanishing $(1-\rho)$-fraction of distinct increments therefore enters D_{dev} as a non-zero contribution, with magnitude bounded below by $(1-\rho)\,\beta^{-1}\,\|\phi - \phi_{\Phi_{\mathcal{G}}}\|_\infty$ where $\phi_{\Phi_{\mathcal{G}}}$ is the projection of ϕ onto $\Phi_{\mathcal{G}}$. For $\rho < 1$ this lower bound is strictly positive, so the policy-deviation score distinguishes R_θ^* from $R_\theta^* + \phi(s') - \gamma\phi(s)$ whenever $\phi \notin \Phi_{\mathcal{G}}$. The full bound, including the dependence on the soft-Bellman recursion coefficient, is mechanical and omitted. □

Remark. The proposition does not eliminate the identifiability ambiguity entirely; potentials in $\Phi_{\mathcal{G}}$ remain invisible to D_{dev}, exactly as in the standard Ziebart-style result. What it removes is the larger class of node-identity-dependent potentials, which are precisely the shaping functions that would otherwise make per-node reward attributions incomparable across the graph.

D Proof of Proposition 3 (Graph-IRL sample complexity)

We adapt the i.i.d. feature-count concentration of Ziebart (2010) to the non-i.i.d. graph setting via fractional covering. Define the dependence graph $\mathcal{D}$ on training trajectories with $\tau_{v_i} \sim \tau_{v_j}$ iff $d_{\mathcal{G}}(v_i, v_j) \leq 2K$. Under Assumption 1 the maximum degree of $\mathcal{D}$ is $\leq \rho N$, so a fractional matching (Janson, 2004, Lemma 1) yields a fractional independent set of weight $\geq N(1-\rho)$, giving an effective i.i.d. subsample of size $N_{\text{eff}} = \lceil N(1-\rho) \rceil$. On this subsample the empirical feature-count $\hat{\Phi}_N$ satisfies the bounded-difference condition with parameter B/N_{eff} per trajectory; McDiarmid gives $\mathbb{P}(|\hat{\Phi}_N - \mathbb{E}\hat{\Phi}_N| > t) \leq 2\exp(-2N_{\text{eff}}t^2/B^2)$, which inverted at confidence $1-\delta$ yields $t = B\sqrt{2\log(2/\delta)/(N(1-\rho))}$. The bias $|\mathbb{E}\hat{\Phi}_N - \mathbb{E}\Phi(\tau)|$ is controlled by the Rademacher complexity $\mathfrak{R}_N$ via symmetrization. Combining gives Equation (18). Setting $\rho = 0$ recovers the i.i.d. bound; the bound is vacuous as $\rho \to 1$. □

E Proof of Lemma 1 (Camouflage-induced soft-value gap)

The argument lower-bounds the per-step reward mismatch induced by the attack and accumulates it along the K-hop trajectory via the soft-Bellman recursion.

Per-step mismatch. The attacked node v^* differs from template v by either (a) at least $k \geq 1$ structural modifications producing $d_{\mathrm{str}}(s_t^v, s_t^{v^*}) \geq k\eta_{\mathrm{str}}$ at some hop t, with $\eta_{\mathrm{str}} = 1/\max_v |\mathcal{E}_v^K|$, or (b) feature perturbation $\|\mathbf{x}_{v^*} - \mathbf{x}_v\|_2 \geq \epsilon$, or both. Lipschitz condition (19) then gives

$$|R_{\theta_1}^{\mathrm{str}}(s_t^{v^*}, a) - R_{\theta_1}^{\mathrm{str}}(s_t^v, a)| \geq L_{\mathrm{str}}\,\eta_{\mathrm{str}}\,k, \quad |R_{\theta_2}^{\mathrm{sem}}(s_t^{v^*}, a) - R_{\theta_2}^{\mathrm{sem}}(s_t^v, a)| \geq L_{\mathrm{sem}}\,\eta_{\mathrm{sem}}\,\epsilon,$$

on the affected pair. The bound on the dominant component holds regardless of budget split.

Soft-Bellman accumulation. The soft-Bellman recursion $V_{\mathrm{soft}}^*(s) = \beta \log \sum_a \exp\big((R^*(s,a) + \gamma\,\mathbb{E}_{s'} V_{\mathrm{soft}}^*(s'))/\beta\big)$ converts a per-step reward gap of magnitude g along one trajectory step into a cumulative soft-value gap of at least $g \cdot (1-\gamma^K)/(1-\gamma)$ by the geometric sum (Ziebart, 2010, Lemma 2.6). Applying this to the per-step bounds above:

$$V_{\mathrm{soft}}^*(s_0^{v^*}) - V_{\mathrm{soft}}^{\pi_{v^*}}(s_0^{v^*}) \geq \frac{1-\gamma^K}{1-\gamma} \cdot \max\big(L_{\mathrm{str}}\eta_{\mathrm{str}}k,\ L_{\mathrm{sem}}\eta_{\mathrm{sem}}\epsilon\big) = \Delta(k,\epsilon).$$

The max (vs. sum) reflects that the soft-Bellman operator is concave in the per-step reward, so a budget split does not accumulate the two components additively. The bound is tight at $k = 0, \epsilon = 0$ (no attack, $\Delta = 0$) and when the budget concentrates on the dominant Lipschitz constant. □

F Proof of Theorem 1 (Camouflage detection lower bound)

Combine Lemma 1 with the soft performance-difference identity (Ziebart, 2010): for any policy π and reward R_θ^*,

$$V_{\mathrm{soft}}^*(s) - V_{\mathrm{soft}}^{\pi}(s) = \beta \cdot \mathbb{E}_{\tau\sim\pi|s_0=s}\Big[\textstyle\sum_{t=0}^{K-1} \mathrm{KL}\big(\pi(\cdot \mid s_t) \,\big\|\, \pi_{\mathrm{soft}}^\star(\cdot \mid s_t)\big)\Big], \tag{26}$$

whose right-hand side equals $\beta \cdot D_{\mathrm{dev}}(\cdot)$ by Definition 4. For normal v with $\pi_v \approx \pi_{\mathrm{soft}}^\star$, $D_{\mathrm{dev}}(v) \approx 0$; for a (k,ϵ)-attacked v^*, Lemma 1 gives $\beta \cdot D_{\mathrm{dev}}(v^*) \geq \Delta(k,\epsilon)$, hence $D_{\mathrm{dev}}(v^*) - D_{\mathrm{dev}}(v) \geq \Delta(k,\epsilon)/\beta$. The right-hand side depends only on the attack budget, Lipschitz constants, η_{str}, K, γ, β, and is independent of the embedding distance $\delta = \|h_{v^*}^{(K)} - h_v^{(K)}\|_2$. □

G Proof of Theorem 2 (Soft-value regret of recovered reward)

The proof composes the expected-feature-count concentration of Proposition 3 with the soft-simulation lemma. By Proposition 3, w.p. $1-\delta$ the feature-count discrepancy between $\pi_{\hat{R}_N}$ and π_{R^*} is bounded by $\xi_N \triangleq 2\mathfrak{R}_N + B\sqrt{2\log(2/\delta)/(N(1-\rho))}$. Under the linear-in-features reward ($R(s,a) = \langle w, \phi(s,a)\rangle$, $\|w\|_2 \leq B$), Cauchy–Schwarz gives the uniform reward bound $\|R^* - \hat{R}_N\|_\infty \leq B\xi_N$. By the soft-simulation lemma (Ziebart, 2010, Lemma 3.1), two rewards within η in ℓ_∞ yield soft-value functions within $\eta/(1-\gamma)$ for any fixed policy; a symmetric argument across the two soft-optimal policies adds a factor of 2, giving $|V_{R^*}^{\pi_{R^*}}(s) - V_{R^*}^{\pi_{\hat{R}_N}}(s)| \leq 2\eta/(1-\gamma)$. Substituting $\eta = B\xi_N$ and integrating against μ_0 gives Equation (24). The bound recovers the i.i.d. rate at $\rho = 0$ and degrades smoothly as $\rho \to 1$. □

H Proof of Theorem 3 (PAC bound on the false-positive rate)

The argument converts the soft-value regret of Theorem 2 into a per-node score perturbation and propagates it through the threshold indicator. For a normal node v with policy π_v, the policy-deviation score under reward R equals $\mathcal{S}_R(v) = \beta^{-1}[V_R^{\pi_R}(s_0^v) - V_R^{\pi_v}(s_0^v)]$ by the soft performance-difference identity (Equation (22)). Theorem 2 then gives w.p. $1-\delta$:

$$|\mathcal{S}_{R^*}(v) - \mathcal{S}_{\hat{R}_N}(v)| \leq \frac{1}{\beta} \cdot \frac{2B}{1-\gamma}\,\xi_N \equiv \zeta_N(\delta).$$

Hence $\widehat{\mathrm{FPR}}_N(\tau) = \mathbb{P}[\mathcal{S}_{\hat{R}_N}(v) > \tau] \le \mathbb{P}[\mathcal{S}_{R^*}(v) > \tau - \zeta_N(\delta)]$. Under sub-Gaussian tails of the score at normal nodes (which holds under Lipschitz-bounded reward features), the CDF is 1-Lipschitz and the right-hand side is at most $\alpha + \zeta_N(\delta)$. Substituting yields Equation (25). □

Remark. For the operating points used in our experiments ($\beta = 0.1$, $N \sim 10^4$, $\rho \sim 0.3$, $\delta = 0.05$, $B \sim 1$), the right-hand side of Equation (25) evaluates to approximately $\alpha + 0.04$.

I Proof of Proposition 4 (Soft value iteration residual)

Both claims are standard contraction arguments for the soft Bellman operator. *Claim (i)* follows from the 1-Lipschitz property of log-sum-exp in its argument vector, which yields $\|\mathcal{T}_\beta V - \mathcal{T}_\beta V'\|_\infty \le \gamma \|V - V'\|_\infty$; Banach's fixed-point theorem then gives $\|V^{(T)} - V^\star_{\mathrm{soft}}\|_\infty \le \gamma^T R_{\max}/(1-\gamma)$ when $V^{(0)} = 0$ and rewards are bounded by $R_{\max}$. *Claim (ii)* follows by inspection: at $\gamma = 1$ with horizon K, the soft Bellman backup reduces to a deterministic backward sweep over the K-hop horizon, so $T \ge K$ iterations from $V^{(0)} = 0$ reach the exact fixed point and subsequent iterations leave V unchanged. The corollary on propagation through the bounds of Theorems 1, 2, 3, and Corollary 1 follows in one line from Pinsker's inequality: an ℓ_∞ residual of magnitude ζ in the soft value function induces a KL deviation of at most ζ/β per state, which shifts each bound additively by that amount. □