Graph Anomaly Detection as Finite-Horizon Control: Training-Free Scoring via Empirical Bayes

Fred Xu*
University of California, Los Angeles
fredxu@cs.ucla.edu

Thomas Markovich
Block
tmarkovich@squareup.com

Florence Regol
Mila – Quebec AI Institute
florence.regol@mila.quebec

Yizhou Sun
University of California, Los Angeles
yzsun@cs.ucla.edu

Abstract

Node-level graph anomaly detection (GAD) identifies nodes whose attributes and interactions deviate from dominant graph regularities. Existing GAD models encode normality and anomaly scoring indirectly through architectures, message passing, reconstruction or contrastive objectives, and tuned score families. This entangles graph trust (how strongly graph structure should define normality), graph-spectral weighting, and anomaly-score choice, yielding scores that are costly, opaque, and unstable across graph regimes. We propose EB-GAD (Empirical-Bayes GAD), a training-free framework that models normality as graph-aware generalized Ornstein–Uhlenbeck (GOU) relaxation toward a graph-filtered template. Empirical Bayes fits the graph precision from the residual-field likelihood; the GOU then turns scoring into a closed-form finite-horizon control energy, the minimum effort to steer a feature-neutral node to its observed endpoint along graph-spectral relaxation. Sweeping relaxation horizon and endpoint tolerance yields a bank of scores that share one fitted prior: equilibrium Mahalanobis scoring is one limit, while finite-horizon control-energy and scale-normalized ratio scores reveal anomalies that static equilibrium scoring can mask. A label-free selector chooses the score family from feature homophily, edge density, and feature dimension, then ranks candidates by fitted-null deviation and rank stability. On 11 benchmarks and without labels at any step, EB-GAD has the best or tied-best AUROC on 9: the four financial fraud networks (up to 3.7M nodes), the YelpChi and Amazon review graphs, Weibo, Reddit and Facebook, with margins of up to 21.7 points. It is second on BlogCatalog and ACM.

1 Introduction

Node-level graph anomaly detection (GAD) asks which nodes look abnormal relative to both their attributes and graph neighborhoods, with applications in fraud detection, cybersecurity, and scientific networks [1, 23]. This is difficult without labels because edges are not always equally trustworthy: in homophilic graphs they reveal normal communities, while in fraud, camouflage, or heterophilic settings they can hide abnormal behavior [22, 7, 19, 34, 36, 37]. A detector must therefore decide how much graph structure should define normality (graph trust), which graph-scale variations matter, and how to turn deviations into an anomaly score.

Most GAD methods make these choices indirectly through architectures, message passing, reconstruction or contrastive losses, and hyperparameter sweeps [6, 17, 10, 21, 13, 43, 16, 25]. This

*Work done during an internship at Block, Inc.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

makes scores hard to interpret and unstable across graph regimes: in native-score evaluation, several baselines fall below chance on Elliptic and Facebook (Table 1). It also leaves no way to ask whether a node is anomalous because it is graph-rough, slow to match a normal template, or large in residual magnitude.

EB-GAD makes these choices explicit. We fit a structured prior in which normal features relax toward a graph-filtered template, while empirical Bayes (EB) selects graph trust and length-scale from the residual-field likelihood [27, 8]; since anomalies are rare, this likelihood is dominated by normal nodes. The prior is the stationary law of a graph-aware generalized Ornstein–Uhlenbeck (GOU) process [41] with graph Matérn precision [3, 38]. This connects modeling to scoring: starting from neutral features, how much control effort is needed to reach the observed node as graph modes relax? The answer is a *finite-horizon control energy*, available in closed form. Varying the relaxation horizon Γ and endpoint tolerance λ_c gives a *bank* of such scores. The bank keeps one fitted prior but tests different views: short-horizon effort, equilibrium roughness, and scale-normalized roughness.

We call the resulting framework EB-GAD (Empirical-Bayes GAD) and describe it as *training-free*. By this we mean that no neural parameters are learned, no gradient descent or epochs are run, and no labels or validation sets are used. EB-GAD does fit three scalars (ρ, κ, γ), by maximizing a closed-form likelihood on the unlabeled graph (a clamped Newton iteration in ρ, small grids over κ and γ), and it ranks a fixed bank of closed-form scores with a label-free selector whose grids and constants are fixed in advance (Appendix B.4). Training-free does not mean hyperparameter-free; a longer name is closed-form scoring with empirical-Bayes calibration. Each ingredient (graph Gaussian priors, likelihood-based fitting, OU processes, Mahalanobis scores) is established; our contribution is their calibrated composition and the evidence that it works without training:

1. **Anomaly score as finite-horizon control energy.** We extend equilibrium graph-Mahalanobis scoring to a finite-horizon bank using the GOU's closed-form Gaussian transitions. The scores (C_{Γ,λ_c}, CR_{Γ,λ_c}, $J^\star$, R) share one EB-fitted graph precision, hence one plug-in null; $J^\star$ is a limiting case, not the whole model.
2. **Label-free selection over the bank.** A predeclared selector chooses among equilibrium, control-energy, control-ratio, and profile scores without labels. Graph statistics pick the family; within it, fitted-null deviation and neighboring-rank stability pick a stable score rather than an isolated AUROC-favored configuration.
3. **Training-free empirical Bayes prior selection.** The linear-Gaussian structure makes the residual-space likelihood closed-form. Algorithm 1 fits the prior by scalar optimizations; its parameters are interpretable (ρ: graph trust; κ: inverse length-scale; γ: template bandwidth).

On 11 benchmarks and without labels at any step, EB-GAD has the best or tied-best AUROC on 9: the four financial fraud networks (up to 3.7M nodes), the YelpChi and Amazon review graphs, Weibo, Reddit and Facebook, with margins of up to 21.7 points. It is second on BlogCatalog and ACM. Section 4 and Section 5 state where it trails, and which entries depend on a design choice made for this version.

2 Background

2.1 Graph Anomaly Detection

Graph anomaly detection (GAD) identifies atypical nodes whose attributes and structural interactions deviate from graph regularities. We focus on *node-level* GAD on attributed graphs [1, 23]. Formally, given $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ with $|\mathcal{V}| = n$, adjacency A, and node features $X \in \mathbb{R}^{n \times D}$, unsupervised node-level GAD outputs scores $s_i = f(i, A, X)$ whose large values rank nodes $i \in \mathcal{V}$ as anomalous. Existing methods span reconstruction/generative approaches [6, 17, 10, 16], self-supervised/contrastive methods [21, 13, 43], and one-class methods [25]. In fraud settings, they must also handle class imbalance, camouflage, and heterophily [22, 7, 19, 34, 36, 37]. Across these families, graph trust, graph-spectral weighting, and anomaly scoring are often tied to architectures and hyperparameter sweeps. EB-GAD instead casts these decisions as parameters of a single linear-Gaussian model and selects them from data without labels.

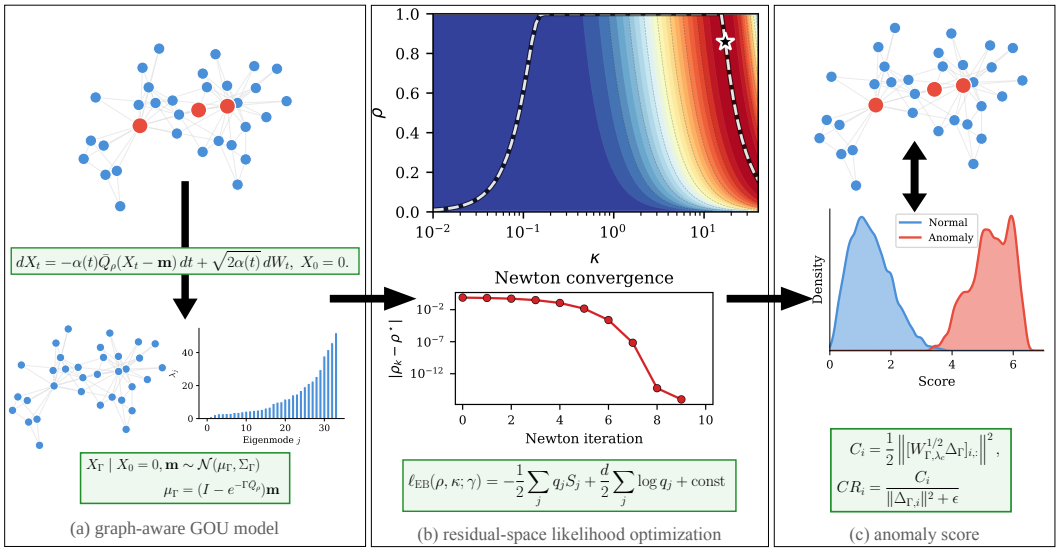

Figure 1: **EB-GAD overview. (a)** A graph-aware GOU around template $\mathbf{m}$ defines finite-horizon Gaussian transitions from a feature-neutral start. **(b)** Residual-space empirical Bayes fits (ρ, κ) for each template bandwidth γ without labels (d in (b) is the feature dimension D). **(c)** The fitted dynamics produce control-energy scores C_{Γ,λ_c}, CR_{Γ,λ_c} and equilibrium scores $J^\star$, R; a fixed selector routes each dataset to one of four score families.

2.2 Empirical Bayes for Anomaly Detection

Unsupervised anomaly detection can be cast as evaluating *Shannon surprise* under a normal-data model, using $s(x_i) = -\log p(x_i \mid \varphi)$. We use φ for prior or marginal-model hyperparameters, not arbitrary trained weights; in EB-GAD these are graph trust, graph length-scale, and template bandwidth. Many methods approximate this score through density estimation [4], reconstruction error [6, 10], or representation-space distance [25, 21]; in each case, the fitted normality notion determines the anomaly score. In *empirical Bayes* [27, 8], these hyperparameters are selected from data $\mathcal{D}$ by maximizing the marginal likelihood:

$$\varphi^\star = \arg\max_{\varphi}\ \log p(\mathcal{D} \mid \varphi), \tag{1}$$

which integrates out latent variables when present and balances fit against complexity (Occam's razor) [8]. Since anomalies are rare, this criterion is dominated by normal nodes and $\varphi^\star$ approximately recovers a normal-data prior. When the model is *conjugate* (Gaussian prior with linear-Gaussian likelihood), the marginal likelihood is available in closed form; otherwise one typically optimizes a variational lower bound (ELBO). In our setting, empirical Bayes is applied to the residual field induced by a low-pass graph-filtered template rather than directly to the raw features. The likelihood is a graph-coupled density over the full residual field, while the node score is the nodewise contribution to the same joint Gaussian energy. Bayesian outlier detection has been explored in non-graph settings [28, 32]; in graph anomaly detection, prior structure is usually fixed rather than selected from data.

2.3 Generalized Ornstein–Uhlenbeck Dynamics

We need a normal-data model that can encode prior knowledge about where node features should concentrate, while remaining tractable enough for label-free fitting. The Ornstein–Uhlenbeck process provides exactly this: a linear Gaussian model whose stationary mean can be set to a template $\mathbf{m}$. We use its matrix-drift generalization,

$$dX_t = -\alpha(t)\,\bar{Q}_\rho\,(X_t - \mathbf{m})\,dt + \sqrt{2\alpha(t)}\,dW_t, \tag{2}$$

where $\alpha(t) > 0$ is a scalar schedule, W_t is Brownian motion, and $\bar{Q}_\rho$ is a symmetric positive definite (SPD) drift operator. Coupling the noise scale to $\alpha(t)$ gives stationary distribution $\mathcal{N}(\mathbf{m}, \bar{Q}_\rho^{-1})$ and

closed-form Gaussian transitions. Section 3.1 instantiates $\mathbf{m}$ as a graph-filtered template and $\bar{Q}_\rho$ as a graph-aware precision, injecting graph prior knowledge into both the stationary mean and covariance while keeping EB fitting and node scoring closed form.

3 Graph Anomaly Detection via Finite-Horizon Control

The components of EB-GAD are standard: a graph Gaussian prior, the OU dynamics attached to it, and Gaussian quadratic scores. The contribution is their composition, calibrated by EB and selected without labels. We therefore state the model first and its control reading afterwards: §3.1–3.3 define a graph-aware GOU and derive its bank of finite-horizon scores; §3.4 fits the prior and selects a score.

3.1 Graph-Aware GOU Process

Notation. We use three forms of the same node-feature data: input matrix $X \in \mathbb{R}^{n\times D}$, its row-major vectorization $\bar{x} = \mathrm{vec}(X) \in \mathbb{R}^{nD}$, and the GOU state $X_t \in \mathbb{R}^{nD}$ (with $X_0 = 0$ and $X_T = \bar{x}$ at scoring time; lowercase x_T in SOC formulae denotes the same realization). The objects $X, M, \Delta, \Delta_\Gamma$ are used in either $n \times D$ matrix or $\mathbb{R}^{nD}$ vec form depending on context (covariance equations use the vec form; nodewise scoring uses the matrix); $[A]_{i,:} \in \mathbb{R}^D$ denotes the i-th row of an $n \times D$ matrix.

Reference dynamics. Let $W_t \in \mathbb{R}^{nD}$ be a standard Brownian motion. We define a graph-aware GOU reference process around a template $\mathbf{m} \in \mathbb{R}^{nD}$:

$$dX_t = -\alpha(t)\,\bar{Q}_\rho\,\big(X_t - \mathbf{m}\big)\,dt + \sqrt{2\alpha(t)}\,dW_t, \qquad t \in [0, T], \tag{3}$$

where $\alpha(t) > 0$ is a deterministic scalar schedule and $\bar{Q}_\rho$ is a graph-aware precision operator. The matrix drift $\bar{Q}_\rho$ embeds the graph into the dynamics: in eigencoordinates of $\bar{Q}_\rho$, mode j relaxes at rate $\alpha(t)q_j$, so smooth (low Laplacian frequency, small q_j) modes drift slowly toward the template while rough modes equilibrate fast. The closed-form Gaussian transition that follows therefore inherits a graph-spectral schedule of relaxation times.

Role of the dynamics for static data. We do not claim that the data evolved by OU dynamics, and static data could not support such a claim. The choice is representational. Among linear diffusions with isotropic noise, (3) is the unique one, up to time scale, that is reversible with stationary law $\mathcal{N}(\mathbf{m}, \bar{Q}_\rho^{-1})$; it is the canonical dynamics of the graph Gaussian prior that EB fits and adds no modeling freedom at equilibrium. What it adds is a horizon-indexed family of closed-form quadratic scores that share this one prior, so that every horizon is calibrated by the same fitted law; a heat or polynomial filter bank changes the implied prior at each scale. The Brownian increments are independent of the current state by definition of the linear-Gaussian transition, not as an assumption on the data: a state-dependent innovation would be absorbed into a different drift and stationary precision, which is what (ρ, κ) fit, and since only the stationary law and its finite-horizon interpolations enter the scores, such dependence is not identifiable from a static graph.

Precision blend: graph trust and inverse length-scale. We parameterize the precision operator with a graph trust parameter $\rho \in [0, 1]$ and Matérn inverse length-scale $\kappa \geq 0$:

$$\bar{Q}_\rho := Q_\rho \otimes I_D, \qquad Q_\rho := \rho\, Q_{\mathrm{prior}} + (1-\rho)\, I_n, \qquad Q_{\mathrm{prior}} = (\kappa^2 I_n + L)^\nu, \tag{4}$$

where $\otimes$ denotes the Kronecker product, $\nu > 0$ controls the spectral roll-off, i.e., how quickly precision grows with graph frequency (fixed to $\nu = 1$ throughout), and L is the Laplacian of the selected graph operator (the input graph by default, or a predeclared feature-affinity reweighting that keeps the same edge set; Appendix B.4). In eigenspace, the precision on mode j is $q_j = \rho(\kappa^2 + \lambda_j)^\nu + (1-\rho)$, where λ_j is the j-th selected-operator eigenvalue. The two parameters play distinct roles: ρ controls the overall *graph trust*: $\rho = 1$ gives full graph-spectral precision, while $\rho = 0$ gives isotropic precision. κ is the Matérn inverse length-scale: modes with $\lambda_j \gg \kappa^2$ are dominated by graph-frequency weighting, while modes with $\lambda_j \ll \kappa^2$ share the approximately constant precision level $(1-\rho) + \rho\kappa^{2\nu}$. Thus κ controls where the transition from low-frequency floor to graph-spectral growth occurs, rather than forcing the low-frequency precision to equal the isotropic value one. This follows from Matérn kernels on graphs [3, 38].

Prior template. The template $\mathbf{m} \in \mathbb{R}^{nD}$ encodes a prior notion of normal node attributes, defined as a low-pass graph-filtered version of the features:

$$\mathbf{m} := \bar{S}_{\gamma,\nu}\,\bar{x}, \qquad S_{\gamma,\nu} := (\gamma^2 I_n + L)^{-\nu}, \qquad \gamma > 0,\ \nu > 0, \tag{5}$$

where $\bar{S}_{\gamma,\nu} := S_{\gamma,\nu} \otimes I_D$ is the unnormalized Matérn resolvent used in our implementation [3, 38]. It is low-pass in graph frequency; because it is unnormalized, the profiled γ controls both bandwidth and low-frequency scale, and eigenvalues $1/(\gamma^2 + \lambda_j)^\nu$ can exceed one when $\gamma^2 + \lambda_j < 1$ (e.g., DC gain $\gamma^{-2\nu}$ on a Laplacian zero mode). We treat this as a fixed template family rather than a post-hoc normalization: EB fitting and scoring always use the same chosen residual operator, while a unit-DC-gain variant would simply define a different predeclared template family. The low-pass construction is canonical, while heterophilic high-dimensional regimes may use the same Matérn construction with a predeclared affinity-weighted template without changing the Gaussian or control-energy derivation. Define $T_{\gamma,\nu} := I_n - S_{\gamma,\nu}$ and $\Delta = T_{\gamma,\nu} X$. All likelihoods and scores are computed on this residual field rather than on raw features; truncated eigenspace and null-mode details are given in Appendix A.1.

Transition kernel. For $0 \le s < t \le T$, define the integrated schedule $\Gamma_{s:t} := \int_s^t \alpha(r)\,dr$ and the propagator $\Phi_{s:t} := \exp(-\Gamma_{s:t}\bar{Q}_\rho)$.

Proposition 1. *Under* (3)*, the transition kernel* $p(X_t \mid X_s)$ *is Gaussian with mean* $\mu_{t|s}(x_s) = \mathbf{m} + \Phi_{s:t}(x_s - \mathbf{m})$ *and covariance* $\Sigma_{t|s} = \bar{Q}_\rho^{-1}(I_{nD} - \exp(-2\Gamma_{s:t}\bar{Q}_\rho))$ *whenever* $q_j > 0$ *on all modeled modes (equivalently,* $\bar{Q}_\rho$ *is SPD on the modeled subspace).*

In eigenspace, the covariance on mode j at time Γ is $\sigma_j^2(\Gamma) = (1 - e^{-2\Gamma q_j})/q_j$, which grows from 0 (at $\Gamma = 0$) toward the equilibrium $1/q_j$ (as $\Gamma \to \infty$). The rate of growth is governed by q_j: high-frequency modes (q_j large) reach equilibrium quickly, while low-frequency modes relax slowly. The proof is given in Appendix A.

3.2 Finite-Horizon Generative Model and Control Energy

We score a node by the dynamical effort needed to generate its observed features from a feature-neutral start under the graph-aware GOU. Concretely: imagine steering the GOU from the origin to the observed endpoint $\bar{x}$ at time Γ; the minimum steering energy (under hard endpoint matching) is exactly the negative log-density of the GOU's unconditioned endpoint distribution, up to a constant. This identification turns the closed-form Gaussian transition kernel into an anomaly score, the *finite-horizon control energy*, and it lets a single fitted GOU yield a bank of scores by varying the horizon Γ and the endpoint tolerance λ_c. It is the minimum control energy of a GOU bridge to one observed endpoint: no coupling or transport plan is optimized, so it is not an optimal-transport cost.

Generative model. We model normal data as endpoints of an unconditioned GOU process mean-reverting toward $\mathbf{m}$ after effective duration $\Gamma := \int_0^T \alpha(r)\,dr$, started at the feature-neutral origin $X_0 = 0$ (justified because features are z-scored). This gives predictive mean $\mu_\Gamma = (I_{nD} - e^{-\Gamma\bar{Q}_\rho})\mathbf{m}$, which converges to $\mathbf{m}$ as $\Gamma \to \infty$. Throughout this subsection, we assume $q_j > 0$ on every modeled mode, equivalently that $\bar{Q}_\rho$ is SPD on the modeled subspace. For a fixed template (equivalently, fixed γ), Proposition 1 gives the residual-space law

$$p(\Delta_\Gamma \mid \varphi, \Gamma) = \mathcal{N}\big(\Delta_\Gamma;\ 0,\ \Sigma_\Gamma\big), \qquad \Delta_\Gamma := X_\Gamma - \mu_\Gamma, \qquad \Sigma_\Gamma = \bar{Q}_\rho^{-1}(I_{nD} - e^{-2\Gamma\bar{Q}_\rho}). \tag{6}$$

Equivalently, conditional on the template and neutral start, $X_\Gamma \sim \mathcal{N}(\mu_\Gamma, \Sigma_\Gamma)$; at scoring time this endpoint is the observed $\bar{x}$, so the realized residual is $\Delta_\Gamma = \bar{x} - \mu_\Gamma$. Normal residuals are plausible samples from this process; anomalies are not.

Time-dependent precision. The precision of the finite-horizon distribution in eigenspace is

$$\tilde{q}_j(\Gamma) := \Sigma_{\Gamma,j}^{-1} = \frac{q_j}{1 - e^{-2\Gamma q_j}} \ \ge\ q_j, \tag{7}$$

with equality in the limit $\Gamma \to \infty$.

Proposition 2. ***(Equilibrium convergence.)*** *For any fixed prior parameters (ρ, κ, γ) such that $q_j > 0$ on all modeled modes, the time-dependent precision satisfies $\tilde{q}_j(\Gamma) \geq q_j$ for all j and all finite $\Gamma > 0$, with exponential convergence:*

$$\frac{\tilde{q}_j(\Gamma)}{q_j} = \frac{1}{1 - e^{-2\Gamma q_j}} = 1 + O(e^{-2\Gamma q_j}). \tag{8}$$

Consequently, for any fixed residual object Y, the finite-horizon and equilibrium precision-weighted residual-space likelihood objectives $\ell_\Gamma(Y; \varphi)$ and $\ell_\infty(Y; \varphi)$ satisfy $\ell_\Gamma(Y; \varphi) \to \ell_\infty(Y; \varphi)$ as $\Gamma \to \infty$. Moreover, for $\Gamma q_{\min} \geq \frac{1}{2} \log 2$ with $q_{\min} = \min_{1 \leq j \leq k} q_j$, the objective difference is bounded by

$$\left|\ell_\Gamma(Y; \varphi) - \ell_\infty(Y; \varphi)\right| \leq C(Y, \varphi)\, e^{-2\Gamma q_{\min}}, \qquad C(Y, \varphi) := \sum_{j=1}^{k} (q_j S_j(Y) + D).$$

Here $S_j(Y)$ is the spectral energy of Y on retained mode j.

At the optimized $(\rho^\star, \kappa^\star)$, $q_{\min} > 0$ on every retained mode (because $\rho^\star < 1$ or $\kappa^\star > 0$ in every selected configuration), so the bound is exponentially tight; the endpoint residual used for scoring satisfies the same $O(e^{-\Gamma q_{\min}})$ convergence (Step 4 of the proof in Appendix A). The final method deliberately retains finite-horizon scores before this collapse, because different graph regimes expose their anomalous signal at different relaxation horizons.

Control energy interpretation. For a fixed template $\mathbf{m}$, the conditional Gaussian score induced by (6) equals, up to an x_T-independent constant, the minimum control energy in a stochastic optimal control (SOC) formulation, since both reduce to the quadratic form $\frac{1}{2}(x_T - \mu_\Gamma)^\top \Sigma_\Gamma^{-1}(x_T - \mu_\Gamma)$. Consider the GOU with steering input:

$$dX_t = -\alpha(t)\, \bar{Q}_\rho (X_t - \mathbf{m})\, dt + g(t)\, u_t(X_t)\, dt + g(t)\, dW_t, \qquad X_0 = 0, \tag{9}$$

with $g(t) = \sqrt{2\alpha(t)}$, where u_t steers the neutral-start process toward the observed data x_T while the drift mean-reverts toward the template $\mathbf{m}$. The uncontrolled endpoint mean is $\mu_\Gamma = (I_{nD} - e^{-\Gamma \bar{Q}_\rho})\mathbf{m}$. The control energy under hard endpoint matching ($\lambda_c \to \infty$) is

$$\mathcal{E}(0 \to x_T; \Gamma) = \frac{1}{2}(x_T - \mu_\Gamma)^\top \Sigma_\Gamma^{-1}(x_T - \mu_\Gamma) = -\log p(x_T \mid X_0 = 0, \mathbf{m}, \varphi, \Gamma) + \text{const}. \tag{10}$$

This gives the score a dynamical interpretation: it measures the effort needed to steer the GOU to the observed features at horizon Γ from a neutral start under attraction to the template. The graph-aware precision operator makes spectrally rough deviations expensive and spectrally smooth ones cheap.

3.3 Per-Node Anomaly Scoring

We now decompose the global Gaussian energy of (6) into per-node contributions, then generalize from hard endpoint matching to a finite-horizon bank parameterized by an endpoint tolerance.

Graph signal model. The Kronecker structure $\bar{Q}_\rho = Q_\rho \otimes I_D$ means the D feature dimensions are conditionally independent given the graph. At scoring time the GOU endpoint equals the observation, $X_T = \bar{x}$, so the realized finite-horizon residual is the matrix $\Delta_\Gamma \in \mathbb{R}^{n \times D}$ obtained by reshaping $\bar{x} - \mu_\Gamma$, with rows $\delta_{\Gamma,i} := [\Delta_\Gamma]_{i,:} = [X]_{i,:} - [\mu_\Gamma]_{i,:}$, representing D conditionally independent graph signals. Its hard-endpoint Gaussian log-likelihood, up to the additive constant $-\frac{kD}{2} \log 2\pi$, is

$$\log p(\mathcal{D} \mid \varphi, \Gamma) = -\frac{1}{2} \sum_{j=1}^{k} \tilde{q}_j(\Gamma)\, S_j(\Gamma) + \frac{D}{2} \sum_{j=1}^{k} \log \tilde{q}_j(\Gamma) \tag{11}$$

where $S_j(\Gamma) := \|\Delta_\Gamma^\top v_j\|^2$ and $k \leq n$ is the number of computed eigenpairs. Equation (11) is the hard-endpoint case ($\lambda_c \to \infty$) of a more general soft-endpoint Gaussian. We retain the finite-horizon GOU transition instead of taking $\Gamma \to \infty$, because short horizons ask whether the node is surprising before the template has fully relaxed, while long horizons approach the equilibrium residual $\Delta = \bar{x} - \mathbf{m}$.

Finite-horizon score bank. Hard endpoint matching is a knife-edge: it asserts that the GOU sample at time Γ exactly equals the observation $\bar{x}$. Allowing endpoint uncertainty $\lambda_c < \infty$ tests whether the anomaly evidence persists under softer matching, where small deviations from the noiseless endpoint are tolerated. We model this by adding observation noise: if $Z_\Gamma \sim \mathcal{N}(\mu_\Gamma, \Sigma_\Gamma)$ and $\bar{x} = Z_\Gamma + \eta$ with $\eta \sim \mathcal{N}(0, \lambda_c^{-1} I)$, then marginalizing Z_Γ gives $\bar{x} \sim \mathcal{N}(\mu_\Gamma, \Sigma_\Gamma + \lambda_c^{-1} I)$. The effective finite-horizon precision on spectral mode j is therefore

$$c_j(\Gamma, \lambda_c) := \left(\lambda_c^{-1} + \frac{1 - e^{-2\Gamma q_j}}{q_j} \right)^{-1}. \tag{12}$$

Thus c_j is not $\tilde{q}_j + \lambda_c$, which would correspond to adding a second precision penalty after evaluating the endpoint density. The two limits recover familiar special cases: $\lambda_c \to \infty$ gives hard endpoint matching ($c_j = \tilde{q}_j(\Gamma)$), and the joint $\Gamma, \lambda_c \to \infty$ limit gives equilibrium ($c_j = q_j$). The corresponding soft-endpoint log-likelihood is

$$\log p(\mathcal{D} \mid \varphi, \Gamma, \lambda_c) = -\frac{1}{2} \sum_{j=1}^{k} c_j(\Gamma, \lambda_c)\, S_j(\Gamma) + \frac{D}{2} \sum_{j=1}^{k} \log c_j(\Gamma, \lambda_c) + \text{const}. \tag{13}$$

The nodewise control-energy score uses the quadratic data-fit part of (13); the log-determinant term is global and does not decompose over nodes. The corresponding nodewise contribution to the control energy is

$$C_i(\Gamma, \lambda_c) := \frac{1}{2} \Big\| \sum_{j=1}^{k} \sqrt{c_j(\Gamma, \lambda_c)}\, [v_j]_i \left(v_j^\top \Delta_\Gamma \right) \Big\|^2, \tag{14}$$

The nodewise scores sum to the projected global quadratic energy, $\sum_i C_i = \frac{1}{2} \sum_{j=1}^{k} c_j S_j$. Because Q_ρ couples nodes, C_i is a nodewise contribution to the global control energy, not an independent nodewise energy.

In some regimes the absolute scale of C_i is confounded by overall residual magnitude, so a high-magnitude normal node can look anomalous under raw C_i even when its spectral profile is consistent with the template. We address this with a scale-normalized control ratio,

$$CR_i(\Gamma, \lambda_c) := \frac{C_i(\Gamma, \lambda_c)}{\|[\Delta_\Gamma]_{i,:}\|^2 + \epsilon}, \tag{15}$$

where $\epsilon > 0$ is a small numerical-stability constant (set to 10^{-8} in practice). This isolates graph-spectral roughness from raw residual scale. The equilibrium scores are recovered as the joint $\Gamma, \lambda_c \to \infty$ limit of the finite-horizon scores:

$$J_i^\star := \lim_{\Gamma, \lambda_c \to \infty} C_i(\Gamma, \lambda_c) = \frac{1}{2} \Big\| \sum_{j=1}^{k} \sqrt{q_j}\, [v_j]_i \left(v_j^\top \Delta \right) \Big\|^2, \qquad R_i := \lim_{\Gamma, \lambda_c \to \infty} CR_i(\Gamma, \lambda_c). \tag{16}$$

When $k = n$, $J_i^\star = \frac{1}{2} \|[Q_\rho^{1/2} \Delta]_{i,:}\|^2$ and the total score is the graph-Gaussian Mahalanobis energy. When $k < n$, the control-energy numerator operates on the retained spectral subspace while the ratio denominator uses the full nodewise residual norm: the latter is intentional, because anomalies may place substantial residual energy in truncated rough modes, so the ratio asks how much of a node's total residual is explained by the retained spectral geometry.

Together, the family of C_{Γ,λ_c} and CR_{Γ,λ_c} at varying (Γ, λ_c), with their $\Gamma, \lambda_c \to \infty$ limits $J^\star$ and R, constitutes the ***finite-horizon score bank*** (the bank). By Proposition 2 the soft-endpoint likelihood collapses to its equilibrium limit (Step 4 of the proof handles the endpoint-residual variant), so Γ and λ_c are predeclared resolutions rather than fitted parameters; their usefulness is judged downstream by unlabeled tail and stability diagnostics (§3.4).

3.4 Empirical Bayes Prior Selection

For each candidate template bandwidth γ, empirical Bayes selects (ρ, κ) from the exact equilibrium residual-space likelihood, and the fitted precision is reused by every finite-horizon score in (14)–(15). EB estimates the graph-Gaussian geometry, while the finite-horizon bank probes how anomaly evidence appears along the GOU relaxation path. Because $\Delta = T_{\gamma,\nu} X$ is γ-dependent, the outer γ

Algorithm 1 Empirical Bayes fit and finite-horizon score selection

Require: Graph (A, X)
Ensure: Optimized prior $\varphi^\star = (\rho^\star, \kappa^\star, \gamma^\star)$; final anomaly score
Compute/cache Laplacian eigenpairs $\{(\lambda_j, v_j)\}$ for each predeclared graph operator (input graph or deterministic affinity-refined operator)
Set candidate template, γ, PCA, horizon Γ, and endpoint-tolerance λ_c grids from graph statistics (Appendix B.4)
for each candidate γ in range **do**
  Compute template $\mathbf{m}(\gamma)$ (Eq. 5), residuals δ_i, spectral energies S_j
  for each candidate $\kappa \in [0, \kappa_{\max}]$ *(profile $\rho^\star(\kappa)$)* **do**
    $\rho^\star(\kappa) \leftarrow$ maximize over $\rho \in [0, 1]$ by clamped Newton updates with endpoint check
  end for
  $\kappa^\star(\gamma) \leftarrow \arg\max_\kappa \log p(\mathcal{D} \mid \rho^\star(\kappa), \kappa, \gamma)$
  $\rho^\star(\gamma) \leftarrow \rho^\star(\kappa^\star(\gamma))$; store $\mathcal{L}^\star(\gamma) = \log p(\mathcal{D} \mid \rho^\star(\gamma), \kappa^\star(\gamma), \gamma)$
end for
$\gamma^\star \leftarrow \arg\max_\gamma \mathcal{L}^\star(\gamma)$; set $\rho^\star \leftarrow \rho^\star(\gamma^\star)$ and $\kappa^\star \leftarrow \kappa^\star(\gamma^\star)$
Construct finite-horizon scores C_{Γ,λ_c}, CR_{Γ,λ_c} at $\gamma^\star$; anchor $J^\star$, R at γ_a with $(\rho^\star(\gamma_a), \kappa^\star(\gamma_a))$
Select score family from graph statistics (Appendix B.4)
Select the final score or rank-fused profile by fitted-null deviation and neighboring-score stability
return anomaly scores

step is *residual-space profile selection*: it optimizes the Gaussian fit in the transformed residual field and omits the raw-feature Jacobian $D \sum_{j=1}^{k} \log|1 - (\gamma^2 + \lambda_j)^{-\nu}|$. The selected $\gamma^\star$ is used by the finite-horizon families, whose p-values are computed under the fitted law of that residual; it is also the bandwidth that removes the most feature energy (Table 14). The equilibrium anchor instead uses the grid value γ_a nearest to the graph-statistic center γ_0, with its own EB fit (Appendix C.1).

Theorem 1. *At equilibrium ($\Gamma \to \infty$), assuming $q_j > 0$ on every retained mode, the residual-space log-likelihood projected onto the k retained modes (full likelihood when $k = n$) reduces, for fixed γ and up to an additive (ρ, κ)-independent constant, to*

$$\log p(\mathcal{D} \mid \varphi) = -\underbrace{\frac{1}{2}\sum_{j=1}^{k} q_j\, S_j}_{\text{data fit (control energy)}} + \underbrace{\frac{D}{2}\sum_{j=1}^{k} \log q_j}_{\text{Gaussian normalizer}} \tag{17}$$

where $S_j = \|\Delta^\top v_j\|^2$ is the spectral energy on eigenmode j.

The first term rewards high precision on low-energy modes; the second discourages making the covariance uniformly too diffuse. When $k = n$, the data-fit term equals the total equilibrium control energy $\sum_i J_i^\star$.

Four score families. The bank is used through four families:

- *Equilibrium anchor* ($J^\star$ or R): the stationary graph-Mahalanobis energy or its scale-normalized ratio. It is the default, used when the finite-horizon diagnostics are not stable enough to justify leaving the stationary score, and on sparse, heavily truncated graphs.
- *Control energy* (C_{Γ,λ_c}): the absolute effort to reach a node at a finite horizon. It is used when residual magnitude is itself informative: on dense graphs with homophilic low-dimensional features, or with high-dimensional features of near-zero homophily (then on the affinity template).
- *Control ratio* (CR_{Γ,λ_c}): the share of a node's residual that the fitted graph geometry makes expensive. It is used when residual magnitude would otherwise dominate: sparse graphs with high-dimensional, weakly homophilic features.
- *Profile aggregation*: per-node p-values of the bank members under the shared plug-in null, combined across horizons and tolerances. It is used when evidence is spread over horizons: very dense graphs, or graphs whose profile diagnostics pass.

Label-free selector. After the EB fit, a fixed rule chooses one family using only label-free statistics: cosine feature homophily h, edge density d_e, feature dimension D_x, and the predeclared truncation k.

Within the chosen family, scores in the admissible pool $\mathcal{B}$ are ranked by percentile-normalized quality plus neighboring-rank stability, $A(r) = \mathrm{pct}_{\mathcal{B}}(Q(r)) + \mathrm{pct}_{\mathcal{B}}(S(r))$, where $Q(r)$ is a family-specific fitted-null diagnostic and $S(r) = \mathrm{median}_{r' \in N(r)} \mathrm{corr}(\mathrm{rank}(s_r), \mathrm{rank}(s_{r'}))$ is local rank-stability over predeclared neighbors $N(r)$ of the node-score vector $s_r \in \mathbb{R}^n$; ranks average ties, so the selection does not depend on the order of the nodes. The selector is label-free model selection: the graph statistics decide which family's null model fits the regime, and the within-family diagnostics are goodness-of-fit statistics against that family's plug-in null. It is sound in this sense, but it is not derived from a single principle: its cutoffs are implementation defaults, identical on all datasets, whose provenance and tested sensitivity are given in Appendix B.4 with the ordered rules.

The likelihood is concave in ρ, while κ and γ are searched on bounded grids (Remark 1); Remark 2 discusses the sensitivity of the fit to anomalies.

4 Experiments

Setup. We evaluate on 11 benchmarks (7 standard GAD graphs and 4 financial networks up to 3.7M nodes; Table 1) against LOF, DIF, and seven graph-based baselines [4, 39, 5, 24, 6, 10, 40, 21, 18, 25, 16], including TAM, a strong local-affinity baseline. On Amazon and YelpChi we follow the TAM/DiffGAD evaluation setup [25, 16]; all AUROCs are computed on the same node set across methods. EB-GAD is run by one fixed procedure on every dataset: the grids, the ordered rules and their cutoffs are those of Appendix B.4, and no label enters any step before evaluation. Every EB-GAD number was regenerated for this version and passes a *relabeling gate*: a score must not change when the nodes are renumbered. Appendix C.1 gives the protocol and lists what changed from the submitted version.

Main results. EB-GAD has the best or tied-best AUROC on 9 of 11 datasets. It leads on the four financial networks (T-Finance, Elliptic, Elliptic++, DGraph) and on YelpChi by 2.6 to 21.7 points over the best baseline, leads on Amazon by 0.9 (within the standard deviation of DIF, 77.1 ± 2.5) and on Weibo by 0.6, and ties TAM on Facebook (91.4). On Reddit its 60.6 is level with CoLA (60.3 ± 0.7) and TAM (60.2); it is second to TAM on BlogCatalog and ACM. The Weibo and Reddit entries depend on the anchor bandwidth γ_a (§3.4), a rule we adopted after evaluating the anchor at the likelihood-selected bandwidth, which gives 92.5 and 48.5 and a count of 7 of 11 (Appendix C.1). AUPRC (Appendix Table 15) confirms the five fraud graphs, with the largest margin on T-Finance (53.0 against 6.6 for the best baseline), and is less favorable elsewhere: EB-GAD is ahead on Weibo by 0.3, level with the best on Reddit (4.5), second on Amazon, third on BlogCatalog and ACM, and fourth on Facebook, where it trails TAM, CoLA and ANOMALOUS despite the tied AUROC. BlogCatalog ($h = 0.009$) and ACM are routed to the feature-affinity template; high-pass and band-pass variants of it do not improve BlogCatalog (Appendix C.8). Appendix Tables 5–6 give the configuration, rule and statistics behind each cell.

What the finite horizon contributes. Table 2 compares each selected score with the equilibrium score of the same configuration, so the fitted model and everything else are held fixed. Six datasets are routed to an equilibrium anchor. Of the five routed to a finite-horizon family, three gain (+3.6 on Facebook and +4.7 on T-Finance by profile aggregation, +4.7 on ACM by the control ratio) and the two control-energy fusions are at parity (+0.1, +0.2). A narrower comparison agrees: a single finite-horizon score, with no tolerance grid, fusion or aggregation, is within 0.3 points of its equilibrium limit or below it (Appendix C.4). The gain does not come from one fortunate horizon but from aggregating and normalizing a family of scores that share one fitted prior and one plug-in null.

Isolation, selector and robustness studies. The appendix reports the studies behind these claims, all rerun for this version. *Isolation* (Appendix C.3): with residuals and selection fixed and only the spectral precision changed, the GOU bank leads heat-kernel, polynomial and identity precisions by 10.2 to 12.6 points on Elliptic, Elliptic++ and T-Finance; elsewhere no family is consistently better, and on Facebook a heat-kernel bank is ahead. *Selector* (Appendix C.5): perturbing each of its 21 cutoffs by $\pm 25\%$ leaves the selection unchanged in 99.4% of the evaluations (97.4% at $\pm 50\%$); quality or stability alone fail on Facebook (17.6 and 25.5 AUROC) where their combination gives 91.4; a supervised combination of the same scores would gain less than one point on four datasets and 8 to 17 points on four. *Robustness* (Appendix C.6): 24% contamination costs 1.1 points on Weibo, heavy-tailed residuals at most 0.7, a multimodal normal population 9 to 19, and binarized features defeat the Gaussian residual model.

Table 1: AUROC (%) on 11 GAD benchmarks. **Bold**: best; underline: second; OOM = out of memory. Baselines use their native anomaly-score direction (Appendix Table 4). Cell sources: * published in the TAM benchmark [25]; † published by the baseline's authors (out of memory on our hardware); unmarked baseline cells are our runs with published hyperparameters. For truncated spectra (YelpChi, Elliptic, Elliptic++, DGraph, T-Finance) EB-GAD is the mean ± standard deviation over five node orders (Appendix C.1).

Dataset	LOF	DIF	ANOM.	DOMINANT	AnomDAE	CONAD	CoLA	DiffGAD	TAM	**EB-GAD**
Weibo	57.9	38.2 ± 1.8	94.5 ± 0.0	85.0 ± 14.6	91.5 ± 1.2	85.4 ± 14.3	28.1 ± 0.0	93.4 ± 0.3	70.9 ± 0.0	**95.1**
Reddit	57.2	53.5 ± 1.4	54.9 ± 5.6	56.0 ± 0.2	55.7 ± 0.4	56.1 ± 0.1	60.3 ± 0.7*	56.3 ± 0.1	60.2 ± 0.4*	**60.6**
Amazon	47.4	77.1 ± 2.5	44.6 ± 0.3*	60.0 ± 0.4*	64.6 ± 7.1	55.5 ± 1.6	59.0 ± 0.8*	55.9 ± 0.1	70.6 ± 1.0*	**78.0**
YelpChi	57.2	56.0 ± 0.9	49.6 ± 0.3*	41.3 ± 1.0*	56.6 ± 2.6	28.7 ± 0.2	46.4 ± 0.1*	29.1 ± 0.0	56.4 ± 0.7*	**71.7** ± 0.6
BlogCatalog	68.6	57.4 ± 2.1	56.5 ± 2.5*	75.9 ± 1.0*	74.7 ± 0.0	68.7 ± 0.0	77.5 ± 0.9*	76.6 ± 0.0	**82.5** ± 0.3*	78.6
Facebook	47.7	37.8 ± 2.8	90.2 ± 0.5*	56.8 ± 0.2*	54.7 ± 0.6	14.2 ± 1.5	84.3 ± 1.1*	28.0 ± 4.7	**91.4** ± 0.8*	**91.4**
ACM	72.0	49.2 ± 1.2	68.6 ± 6.3*	85.7 ± 2.0*	74.8 ± 0.0	80.0 ± 0.1	82.3 ± 0.1*	81.8 ± 1.9	**88.8** ± 2.4*	86.0
Elliptic	45.1	21.8 ± 0.5	OOM	11.2 ± 3.5	14.2 ± 1.2	11.3 ± 0.2	65.3 ± 1.3	26.6 ± 1.0	36.0 ± 0.2	**74.5** ± 0.3
Elliptic++	45.9	22.5 ± 0.7	OOM	39.7 ± 0.0	23.7 ± 3.9	39.3 ± 0.9	62.9 ± 6.7	18.4 ± 2.3	34.4 ± 0.2	**72.5** ± 0.0
DGraph	40.4	64.2 ± 1.8	OOM	OOM	OOM	34.7 ± 1.2†	54.8 ± 0.7	52.4 ± 0.0†	35.7 ± 0.8	**66.8** ± 0.1
T-Finance	34.5	35.5 ± 1.5	55.4 ± 0.0	53.8 ± 0.0	32.1 ± 0.0	63.1 ± 0.1	48.3 ± 0.0	53.6 ± 0.1	61.8 ± 0.0†	**84.8** ± 0.0

Table 2: Contribution of the finite horizon: AUROC (%) of the selected score against the equilibrium score ($J^\star$ or R, by NullKS) of the *same* configuration (operator, template, bandwidth, PCA, truncation, EB fit). *Family*: the family the selector routes to; for an anchor the two rows coincide.

	Weibo	Reddit	Amazon	YelpChi	BlogCat.	Facebook	ACM	Elliptic	Ell.++	DGraph	T-Fin.
Family	anchor	anchor	energy	anchor	energy	profile	ratio	anchor	anchor	anchor	profile
Equilibrium	95.1	60.6	77.9	71.7	78.4	87.8	81.3	74.5	72.5	66.8	80.1
Selected	95.1	60.6	**78.0**	71.7	**78.6**	**91.4**	**86.0**	74.5	72.5	66.8	**84.8**
Δ	0.0	0.0	+0.1	0.0	+0.2	+3.6	+4.7	0.0	0.0	0.0	+4.7

Encoders and scalability. Replacing PCA by a trained MLP or GraphSAGE encoder [11] breaks the linear-Gaussian residual model that EB exploits in closed form: the training-free pipeline is ahead on 8 of 11 datasets, and a trained encoder on Weibo, Facebook and, by 0.1, T-Finance (Appendix B.8). On 16 CPU threads the complete run behind a cell of Table 1 takes from 8 seconds (Facebook) to 33 minutes (DGraph, of which 30 are the eigendecomposition); the two trained graph baselines that fit in memory on DGraph need 65 and 69 minutes on a GPU (Appendix Table 9).

5 Scope, Limitations and Conclusion

EB-GAD is a training-free empirical-Bayes model for node-level graph anomaly detection: a graph-aware GOU defines a bank of finite-horizon control-energy scores, the residual-space likelihood selects the graph precision without labels, and a fixed selector chooses one of four score families. The equilibrium Mahalanobis score is one limiting case; aggregating or normalizing the finite-horizon family under its shared null adds 3.6 to 4.7 AUROC points on three datasets and is at parity elsewhere. Run without labels, EB-GAD has the best or tied-best AUROC on 9 of 11 benchmarks.

Scope and limitations. (i) *Selection.* The model does not identify the template bandwidth: the likelihood sets it for the finite-horizon families and graph statistics set it for the anchor, a rule adopted after the first had been evaluated and without which Weibo and Reddit are 92.5 and 48.5. The selector is label-free and stable under perturbation of its cutoffs, but it is not derived from one principle: its cutoffs are implementation defaults developed on the benchmarks of this paper, and we do not claim that they transfer to other graph families. (ii) *Gaussian model.* In a controlled test heavy-tailed residuals cost at most 0.7 points and a two-component normal population 9 to 19; binarized features fall outside the Gaussian residual model (AUROC below 61) and need a categorical likelihood. (iii) *Contamination.* With up to 24% contamination the fitted prior moves little and detection of the original anomalies loses 1.1 points on Weibo; large contiguous anomalous regions, which can become their own context, were not studied. (iv) *Graph operators.* BlogCatalog and ACM trail TAM, and high-pass or band-pass templates do not close the gap; heterophily is handled only through feature-affinity reweighting. (v) *Truncated spectra.* With many small components the truncated eigenbasis is not unique; on YelpChi the AUROC ranges over two points across node orders. (vi) *Fairness.* Scores on user graphs can encode degree and activity biases and should be audited across user segments before deployment.

Acknowledgments and Disclosure of Funding

We thank Block for supporting this work; part of it was carried out while Fred Xu was an intern at Block. This work was partially supported by NSF 2303037, NSF 2312501, NSF 2531008, NSF 2553619, SRC JUMP 2.0 Center, CDSC Industry Partners, Amazon Research Awards, Snapchat, and Google Gifts.

References

[1] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

[2] Robert H. Berk and Douglas H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(1):47–59, 1979.

[3] Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Deisenroth, and Nicolas Durrande. Matérn Gaussian processes on graphs. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2593–2601. PMLR, 2021.

[4] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD/PODS00, pages 93–104, New York, NY, USA, 2000. Association for Computing Machinery.

[5] Sihan Chen, Zhuangzhuang Qian, Wingchun Siu, Xingcan Hu, Jiaqi Li, Shawn Li, Yuehan Qin, Tiankai Yang, Zhuo Xiao, Wanghao Ye, Yichi Zhang, Yushun Dong, and Yue Zhao. PyOD 2: A Python library for outlier detection with LLM-powered model selection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2807–2810, 2025.

[6] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 594–602. Society for Industrial and Applied Mathematics, 2019.

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

[8] Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

[9] Youssef Elmougy and Ling Liu. Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3979–3990, 2023.

[10] Haoyi Fan, Fengbin Zhang, and Zuoyong Li. AnomalyDAE: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689. IEEE, 2020.

[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. DGraph: A large-scale financial dataset for graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 22765–22777. Curran Associates, Inc., 2022. Datasets and Benchmarks Track.

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

[14] Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[15] Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6(60):1783–1816, 2005.

[16] Jinghan Li, Yuan Gao, Jinda Lu, Junfeng Fang, Congcong Wen, Hui Lin, and Xiang Wang. Diff-GAD: A diffusion-based unsupervised graph anomaly detector. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Yuening Li, Xiao Huang, Jundong Li, Mengnan Du, and Na Zou. SpecAE: Spectral AutoEncoder for anomaly detection in attributed networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2233–2236, 2019.

[18] Kay Liu, Yingtong Dou, Xueying Ding, Xiyang Hu, Ruitong Zhang, Hao Peng, Lichao Sun, and Philip S. Yu. PyGOD: A Python library for graph outlier detection. *Journal of Machine Learning Research*, 25(141):1–9, 2024.

[19] Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, pages 3168–3177, 2021.

[20] Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[21] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2378–2392, June 2022.

[22] Zhiwei Liu, Yingtong Dou, Philip S. Yu, Yutong Deng, and Hao Peng. Alleviating the inconsistency problem of applying graph neural network to fraud detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1569–1572, 2020.

[23] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12012–12038, 2023.

[24] Zhen Peng, Minnan Luo, Jundong Li, Huan Liu, and Qinghua Zheng. ANOMALOUS: A joint modeling approach for anomaly detection on attributed networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3513–3519. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

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

[26] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[27] Herbert Robbins. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability (1954/1955), Volume I: Contributions to the Theory of Statistics*, 1:157–163, 1956.

[28] Matthew S. Shotwell and Elizabeth H. Slate. Bayesian outlier detection with Dirichlet process mixtures. *Bayesian Analysis*, 6(4):665–690, 2011.

[29] Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21076–21089. PMLR, 2022.

[30] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. Association for Computing Machinery.

[31] Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 817–826, New York, NY, USA, 2009. Association for Computing Machinery.

[32] Jo-Anne Ting, Aaron D'Souza, and Stefan Schaal. Automatic outlier detection: A Bayesian approach. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2489–2494. IEEE, 2007.

[33] Yanling Wang, Jing Zhang, Shasha Guo, Hongzhi Yin, Cuiping Li, and Hong Chen. Decoupling representation learning and classification for GNN-based anomaly detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1239–1248, New York, NY, USA, 2021. Association for Computing Machinery.

[34] Yuchen Wang, Jinghui Zhang, Zhengjie Huang, Weibin Li, Shikun Feng, Ziheng Ma, Yu Sun, Dianhai Yu, Fang Dong, Jiahui Jin, Beilun Wang, and Junzhou Luo. Label information enhanced fraud detection against low homophily in graphs. In *Proceedings of the ACM Web Conference 2023*, pages 406–416, 2023.

[35] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019. KDD 2019 Workshop on Anomaly Detection in Finance.

[36] Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. Semi-supervised credit card fraud detection via attribute-driven graph representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14557–14565, 2023.

[37] Fan Xu, Nan Wang, Hao Wu, Xuezhi Wen, Xibin Zhao, and Hai Wan. Revisiting graph-based fraud detection in sight of heterophily and spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):9214–9222, 2024.

[38] Fred Xu and Thomas Markovich. Uncertainty estimation on graphs with structure informed stochastic partial differential equations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. NeurIPS 2025 poster.

[39] Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604, 2023.

[40] Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Advances in Knowledge Discovery and Data Mining*, pages 444–457, Cham, 2022. Springer International Publishing.

[41] Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58068–58089. PMLR, 2024.

[42] Tong Zhao, Chuchen Deng, Kaifeng Yu, Tianwen Jiang, Daheng Wang, and Meng Jiang. Error-bounded graph anomaly loss for GNNs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 1873–1882, New York, NY, USA, 2020. Association for Computing Machinery.

[43] Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T. Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12220–12233, 2023.

[44] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations*, 2024.

[45] Kaizhen Zhu, Mokai Pan, Yuexin Ma, Yanwei Fu, Jingyi Yu, Jingya Wang, and Ye Shi. UniDB: A unified diffusion bridge framework via stochastic optimal control. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 79985–80012. PMLR, 2025. Spotlight.

A Derivation of Technical Results

We provide proofs of the propositions and theorem stated in Section 3.1–3.4.

A.1 Modeling Details

Notation: vec convention. We use the row-major vectorization throughout, so $\bar{Q}_\rho = Q_\rho \otimes I_D$ acts as $\bar{Q}_\rho \operatorname{vec}(X) = \operatorname{vec}(Q_\rho X)$ for any $X \in \mathbb{R}^{n\times D}$. Equivalently, Q_ρ couples nodes while I_D leaves the feature direction untouched.

Residual template and modeled subspace. The template operator $S_{\gamma,\nu}$ in (5) is low-pass but is not normalized to have unit gain on the Laplacian nullspace; consequently γ controls both bandwidth and low-frequency scale. This scale is part of the profiled template family, and the same residual operator is used consistently in likelihood evaluation and scoring. For fixed γ, each feature column of the residual field $\Delta = T_{\gamma,\nu}X$ has the modeled full-spectrum law $\mathcal{N}(0, Q_\rho^{-1})$; equivalently, the vectorized residual has covariance $\bar{Q}_\rho^{-1} = Q_\rho^{-1} \otimes I_D$. The Laplacian supplies graph-aware anisotropy and the $(1-\rho)I_n$ term supplies an isotropic background. When $k < n$, the likelihood and the energy score $J^\star$ are computed on the projected residual $P_k\Delta$, while the ratio score R normalizes this projected energy by the full nodewise residual norm. Feature dimensions are z-score standardized before template construction and scoring. Disconnected graphs may still have additional Laplacian null modes; we do not remove them explicitly, and instead require only $q_j > 0$ on every retained mode so graph-uninformative directions are handled by the isotropic part of Q_ρ.

Remark 1 (Optimization landscape)**.** *Writing $a_j := (\kappa^2 + \lambda_j)^\nu - 1$ and $q_j = \rho a_j + 1$, the residual-space log-likelihood has gradient $\partial \log p/\partial\rho = \sum_j a_j(-S_j/2 + D/(2q_j))$ and Hessian $\partial^2 \log p/\partial\rho^2 = -\sum_j Da_j^2/(2q_j^2) \le 0$, so it is concave in ρ with at most one interior root; Algorithm 1 locates it by clamped Newton steps and otherwise compares the boundary values $\rho \in \{0, 1\}$. The profile likelihood is non-concave in κ and the template-induced objective is non-convex in γ, so both are searched on bounded grids.*

Remark 2 (Outlier sensitivity)**.** *The squared-residual data-fit $\frac{1}{2}\sum_j q_j S_j$ is in principle sensitive to extreme anomalies, but three properties mitigate this here: anomaly fractions are 1–10% so the bulk of the likelihood is set by normal nodes; rough-mode energy that anomalies inflate is dropped by the truncated eigenspace on large graphs; and the $\frac{D}{2}\sum_j \log q_j$ normalizer prevents a small high-residual group from collapsing ρ to zero. Appendix C.6 measures the effect of added contamination; robust pseudolikelihoods are a natural extension.*

Connection to stochastic optimal control. We use λ_c as an inverse endpoint-tolerance variance: the latent GOU endpoint has covariance Σ_Γ, the observation tolerance has covariance $\lambda_c^{-1}I$, and marginalizing the latent endpoint gives the quadratic form induced by $(\lambda_c^{-1}I + \Sigma_\Gamma)^{-1}$. Under hard endpoint matching, $\lambda_c \to \infty$, this reduces to $\frac{1}{2}(x_T - \mu_\Gamma)^\top \Sigma_\Gamma^{-1}(x_T - \mu_\Gamma)$. Equation (14) is the per-node *contribution* to this total control energy, not an independent nodewise control energy because Q_ρ couples nodes. The energy is that of steering one process to one observed endpoint; no coupling between distributions is optimized. For general nonlinear bridge models, computing the analogous endpoint control energy usually requires a learned score, controller, or reverse transition model [44, 41, 45]; here the Gaussian transition kernel gives it directly, making such training unnecessary.

Proposition 1. *Under* (3)*, the transition kernel $p(X_t \mid X_s)$ is Gaussian with mean $\mu_{t|s}(x_s) = \mathbf{m} + \Phi_{s:t}(x_s - \mathbf{m})$ and covariance $\Sigma_{t|s} = \bar{Q}_\rho^{-1}(I_{nD} - \exp(-2\Gamma_{s:t}\bar{Q}_\rho))$ whenever $q_j > 0$ on all modeled modes (equivalently, $\bar{Q}_\rho$ is SPD on the modeled subspace).*

Proof. Fix $0 \le s < t \le T$. Define the centered process $Y_t := X_t - \mathbf{m}$. Then (3) becomes

$$dY_t = -\alpha(t)\bar{Q}_\rho\, Y_t\, dt + \sqrt{2\alpha(t)}\, dW_t, \qquad Y_s = x_s - \mathbf{m}. \tag{18}$$

This is a linear SDE with time-dependent scalar coefficient $\alpha(t)$ multiplying the constant symmetric positive-definite matrix $\bar{Q}_\rho$ on the modeled subspace.

Step 1: Variation of constants. Let $\Phi_{s:t} := \exp(-\Gamma_{s:t}\bar{Q}_\rho)$ with $\Gamma_{s:t} := \int_s^t \alpha(r)\,dr$ be the fundamental solution, satisfying $\partial_t \Phi_{s:t} = -\alpha(t)\bar{Q}_\rho\,\Phi_{s:t}$. Applying Itô's product rule to $\Phi_{s:t}^{-1}Y_t$:

$$d(\Phi_{s:t}^{-1}Y_t) = \alpha(t)\bar{Q}_\rho\,\Phi_{s:t}^{-1}Y_t\,dt + \Phi_{s:t}^{-1}\big(-\alpha(t)\bar{Q}_\rho\,Y_t\,dt + \sqrt{2\alpha(t)}\,dW_t\big) = \Phi_{s:t}^{-1}\sqrt{2\alpha(t)}\,dW_t.$$

Integrating and left-multiplying by $\Phi_{s:t}$:

$$Y_t = \Phi_{s:t}Y_s + \int_s^t \Phi_{\tau:t}\sqrt{2\alpha(\tau)}\,dW_\tau.$$

Step 2: Conditional mean and covariance. Taking expectation: $\mathbb{E}[Y_t \mid Y_s] = \Phi_{s:t}Y_s$, giving $\mu_{t|s}(x_s) = \mathbf{m} + \Phi_{s:t}(x_s - \mathbf{m})$. By the Itô isometry (using $\bar{Q}_\rho$ symmetric, so $\Phi_{\tau:t}^\top = \Phi_{\tau:t}$):

$$\Sigma_{t|s} = \int_s^t 2\alpha(\tau)\,\Phi_{\tau:t}^2\,d\tau = \int_s^t 2\alpha(\tau)\,\exp(-2\Gamma_{\tau:t}\bar{Q}_\rho)\,d\tau.$$

Step 3: Closed form via substitution. Substituting $u = \Gamma_{\tau:t} = \int_\tau^t \alpha(r)\,dr$ gives $du = -\alpha(\tau)\,d\tau$. The limits are $u = \Gamma_{s:t}$ when $\tau = s$ and $u = 0$ when $\tau = t$, so:

$$\Sigma_{t|s} = \int_0^{\Gamma_{s:t}} 2\,e^{-2u\bar{Q}_\rho}\,du = \bar{Q}_\rho^{-1}\big(I_{nD} - e^{-2\Gamma_{s:t}\bar{Q}_\rho}\big),$$

where the last equality uses $\int_0^a 2e^{-2uA}\,du = A^{-1}(I - e^{-2aA})$ for any SPD matrix A. □

Stationary distribution. When $\bar{Q}_\rho$ is SPD (or after restricting to the modeled subspace), the GOU process (3) has stationary distribution $\mathcal{N}(\mathbf{m}, \bar{Q}_\rho^{-1})$. This follows from Proposition 1: if $X_0 \sim \mathcal{N}(\mathbf{m}, \bar{Q}_\rho^{-1})$, then in each retained mode

$$\mathrm{Var}(X_T)_j = e^{-2\Gamma q_j}/q_j + (1 - e^{-2\Gamma q_j})/q_j = 1/q_j,$$

recovering the initial covariance. The mean is also preserved because $\mathbb{E}[X_T] = \mathbf{m} + e^{-\Gamma\bar{Q}_\rho}(\mathbb{E}[X_0] - \mathbf{m}) = \mathbf{m}$ whenever $\mathbb{E}[X_0] = \mathbf{m}$. Because every retained mode has $q_j > 0$, the mean term $e^{-\Gamma q_j}(x_0 - \mathbf{m})_j$ and the initial-covariance contribution $e^{-2\Gamma q_j}\,\mathrm{Var}(X_0)_j$ vanish as $\Gamma \to \infty$ for any initial law with finite second moment, so this fixed point is also the limiting stationary distribution. The equilibrium covariance $\bar{Q}_\rho^{-1}$ is the $\Gamma \to \infty$ limit of the finite-horizon covariance Σ_Γ used in the generative model (6).

Theorem 1. *At equilibrium ($\Gamma \to \infty$), assuming $q_j > 0$ on every retained mode, the residual-space log-likelihood projected onto the k retained modes (full likelihood when $k = n$) reduces, for fixed γ and up to an additive (ρ, κ)-independent constant, to*

$$\log p(\mathcal{D} \mid \varphi) = -\underbrace{\frac{1}{2}\sum_{j=1}^{k} q_j\,S_j}_{\text{data fit (control energy)}} + \underbrace{\frac{D}{2}\sum_{j=1}^{k}\log q_j}_{\text{Gaussian normalizer}} \tag{17}$$

where $S_j = \|\Delta^\top v_j\|^2$ is the spectral energy on eigenmode j.

Proof. Fix γ and let $\Delta = T_{\gamma,\nu}X$ with $T_{\gamma,\nu} = I_n - S_{\gamma,\nu}$ as in Section 3.1. The Kronecker structure $\bar{Q}_\rho = Q_\rho \otimes I_D$ means the D feature dimensions are conditionally independent given the graph. Let $V_k = [v_1, \ldots, v_k] \in \mathbb{R}^{n\times k}$ and $P_k = V_k V_k^\top$ be the projection onto the k computed eigenmodes. For each feature dimension ℓ, define the retained spectral coordinates

$$z_\ell := V_k^\top \Delta_{:,\ell} \in \mathbb{R}^k, \qquad \Delta_{:,\ell} \in \mathbb{R}^n \text{ the } \ell\text{-th column of } \Delta \in \mathbb{R}^{n\times D}.$$

Under the modeled residual law, these coordinates are Gaussian with diagonal covariance

$$z_\ell \sim \mathcal{N}\big(0, \mathrm{diag}(1/q_1, \ldots, 1/q_k)\big),$$

since $Q_\rho v_j = q_j v_j$ on the retained subspace. Therefore the log-density factorizes across feature dimensions:

$$\log p(P_k\Delta \mid \varphi) = \sum_{\ell=1}^{D}\log p(z_\ell \mid \varphi) = -\tfrac{1}{2}\sum_{j=1}^{k} q_j S_j + \tfrac{D}{2}\sum_{j=1}^{k}\log q_j - \tfrac{kD}{2}\log 2\pi,$$

where

$$S_j = \sum_{\ell=1}^{D} (v_j^\top \Delta_{:,\ell})^2 = \|\Delta^\top v_j\|^2.$$

When $k = n$, $P_n = I_n$ and the modeled residual object is Δ itself, yielding (17) up to the explicit additive constant $-\frac{nD}{2}\log 2\pi$. When $k < n$, the projected likelihood is the marginal likelihood of the data restricted to the computed spectral subspace; the additive constant becomes $-\frac{kD}{2}\log 2\pi$, which is independent of (ρ, κ) and therefore irrelevant for model selection. This is consistent with the energy part of (16), which uses the same k-mode projection; ratio scores then normalize that projected energy by the full nodewise residual norm as described in Section 3.3. □

Proposition 2. ***(Equilibrium convergence.)*** *For any fixed prior parameters (ρ, κ, γ) such that $q_j > 0$ on all modeled modes, the time-dependent precision satisfies $\tilde{q}_j(\Gamma) \geq q_j$ for all j and all finite $\Gamma > 0$, with exponential convergence:*

$$\frac{\tilde{q}_j(\Gamma)}{q_j} = \frac{1}{1 - e^{-2\Gamma q_j}} = 1 + O(e^{-2\Gamma q_j}). \tag{8}$$

Consequently, for any fixed residual object Y, the finite-horizon and equilibrium precision-weighted residual-space likelihood objectives $\ell_\Gamma(Y;\varphi)$ and $\ell_\infty(Y;\varphi)$ satisfy $\ell_\Gamma(Y;\varphi) \to \ell_\infty(Y;\varphi)$ as $\Gamma \to \infty$. Moreover, for $\Gamma q_{\min} \geq \frac{1}{2}\log 2$ with $q_{\min} = \min_{1\leq j\leq k} q_j$, the objective difference is bounded by

$$\big|\ell_\Gamma(Y;\varphi) - \ell_\infty(Y;\varphi)\big| \leq C(Y,\varphi)\, e^{-2\Gamma q_{\min}}, \qquad C(Y,\varphi) := \sum_{j=1}^{k} (q_j S_j(Y) + D).$$

Here $S_j(Y)$ is the spectral energy of Y on retained mode j.

Proof. **Step 1: Precision bound.** Since $0 < 1 - e^{-2\Gamma q_j} < 1$ for all finite $\Gamma > 0$ and $q_j > 0$:

$$\tilde{q}_j(\Gamma) = \frac{q_j}{1 - e^{-2\Gamma q_j}} > q_j.$$

Equality holds only in the limit $\Gamma \to \infty$.

Step 2: Convergence rate. Let $r_j(\Gamma) := \tilde{q}_j(\Gamma)/q_j = 1/(1 - e^{-2\Gamma q_j})$. For $\Gamma q_j \geq 1$, $e^{-2\Gamma q_j} \leq e^{-2}$, so $r_j \leq 1/(1 - e^{-2}) \approx 1.16$. More generally, $r_j - 1 = e^{-2\Gamma q_j}/(1 - e^{-2\Gamma q_j}) \leq 2e^{-2\Gamma q_j}$ for $\Gamma q_j \geq \frac{1}{2}\log 2$, establishing the exponential bound (8).

Step 3: Objective convergence for a fixed residual object. Let Y be any fixed residual object with retained spectral energies $S_j(Y)$. Write $\tilde{q}_j = q_j \cdot r_j$ and expand the log-likelihood difference:

$$\ell_\Gamma(Y;\varphi) - \ell_\infty(Y;\varphi) = \sum_j \Big[-\frac{q_j S_j(Y)}{2}(r_j - 1) + \frac{D}{2}\log r_j \Big].$$

Since $r_j \to 1$ exponentially as $\Gamma \to \infty$ (Step 2), both $(r_j - 1)$ and $\log r_j$ vanish, giving pointwise convergence. Using $\log r_j \leq r_j - 1$ and the bound from Step 2, if $\Gamma q_{\min} \geq \frac{1}{2}\log 2$ then

$$\big|\ell_\Gamma(Y;\varphi) - \ell_\infty(Y;\varphi)\big| \leq \frac{1}{2}\sum_{j=1}^{k} (q_j S_j(Y) + D)(r_j - 1) \leq C(Y,\varphi) e^{-2\Gamma q_{\min}},$$

with $C(Y,\varphi) = \sum_{j=1}^{k}(q_j S_j(Y) + D)$, matching the bound in Proposition 2. Moreover, the convergence is uniform over modes: for $\Gamma \geq c/q_{\min}$, $\max_j |r_j - 1| \leq 2e^{-2c}$.

Step 4: Endpoint-residual convergence. The proposition above fixes the residual object because this is the quantity optimized by the residual-space likelihood. For the endpoint score, the residual also depends on Γ:

$$\Delta_\Gamma = \Delta + e^{-\Gamma Q_\rho} M, \qquad v_j^\top \Delta_\Gamma = v_j^\top \Delta + e^{-\Gamma q_j} v_j^\top M,$$

where M is the template reshaped as an $n \times D$ matrix; this is matrix form, so $\bar{Q}_\rho = Q_\rho \otimes I_D$ acts trivially on the feature direction. On the retained subspace,

$$\|P_k(\Delta_\Gamma - \Delta)\|_F \leq e^{-\Gamma q_{\min}} \|P_k M\|_F.$$

Thus the endpoint residual converges to the equilibrium residual, and the full endpoint score converges to the equilibrium score with an $O(e^{-\Gamma q_{\min}})$ contribution from the mean and an $O(e^{-2\Gamma q_{\min}})$ contribution from the precision.

Step 5: Practical indistinguishability. At the selected $(\rho^\star, \kappa^\star)$ and $\nu = 1$, the minimum precision is bounded below by

$$q_{\min} = (1 - \rho^\star) + \rho^\star(\kappa^{\star\,2} + \lambda_{\min}^{\text{ret}}) \geq (1 - \rho^\star) + \rho^\star \kappa^{\star\,2},$$

where $\lambda_{\min}^{\text{ret}}$ is the smallest retained eigenvalue; equality holds when a zero Laplacian mode is retained. This quantity is strictly positive whenever $\rho^\star < 1$ or $\kappa^\star > 0$. Hence for any $c \geq \frac{1}{2}\log 2$ and any horizon satisfying $\Gamma \geq c/q_{\min}$,

$$\max_j (r_j - 1) \leq 2e^{-2c}.$$

For example, $c = 3$ gives a uniform relative precision error below 0.5% on every retained mode. Thus the finite-horizon and equilibrium objectives become practically indistinguishable once Γ is a few multiples of $1/q_{\min}$, exactly as stated in Proposition 2. □

B Experimental Details

B.1 Dataset Specifications

Here we provide the details of each dataset. The detailed dataset statistics can be found in Table 3.

- **Weibo** [42]: Social network from the Tencent-Weibo platform. Nodes are users and edges represent follower/followee relationships. Labels mark suspicious or abusive accounts.
- **Reddit** [33]: Reddit interaction graph used in GAD benchmarks, with text-derived node attributes and labels associated with banned or anomalous users.
- **Amazon** [7, 25]: Review-derived user graph from Amazon. Nodes are reviewers; relations connect reviewers through shared review behavior such as reviewing the same product. Labels indicate suspicious or fraudulent review behavior.
- **Yelp-Chi** [7, 25]: Review graph from YelpChi. Nodes are reviews; relations connect reviews through shared user/product/rating/time information, and our selected operator uses the review–user–review relation. The task is spam review detection.
- **BlogCatalog** [31, 25]: Blog author social network with nodes as bloggers and edges as friendships. Node features are derived from blog content; the GAD benchmark uses injected contextual/structural anomalies.
- **Facebook** [40, 25]: Facebook social network benchmark with users as nodes, friendship edges, and user attributes. Labels mark anomalous users in the benchmark.
- **ACM** [30, 25]: Academic network derived from ArnetMiner/ACM data, with papers as nodes and content/metadata-derived attributes. The GAD benchmark uses injected contextual/structural anomalies.
- **T-Finance** [29]: Financial transaction graph for fraud detection. Nodes are accounts/entities, edges represent transaction relationships, and anomalies correspond to fraudulent actors.
- **Elliptic** [35]: Bitcoin transaction graph with ∼200K nodes and ∼234K edges. Nodes are transactions; edges are payment flows. Features include local and aggregated transaction attributes. Labels indicate licit vs. illicit transactions.
- **Elliptic++** [9]: Extension of Elliptic that augments Bitcoin transaction data with wallet/address graph information for AML research; we use the transaction-level graph/features in the benchmark.
- **DGraph** [12]: Large-scale financial graph with ∼3.7M nodes and ∼4.3M edges from a real-world lending platform. Nodes are users, directed edges encode emergency-contact relationships, and labels identify users with fraudulent borrowing behavior.

Table 3: Raw benchmark statistics for graph datasets used in Table 1. #Feat = feature dimension; Anom.% = anomaly rate (%).

Dataset	#Nodes	#Edges	#Feat	Anom.%
Standard benchmarks (Table 1)				
Weibo	8,405	407,963	400	4.1
Reddit	10,984	168,016	64	3.3
Amazon	11,944	4,398,392	25	6.9
YelpChi	45,954	3,846,979	32	4.9
BlogCatalog	5,196	171,743	8,189	5.8
Facebook	1,081	55,104	576	2.5
ACM	16,484	71,980	8,337	3.6
Fraud benchmarks (Table 1)				
T-Finance	39,357	1,222,543	10	4.6
Elliptic	203,769	234,355	166	$\sim$2.2
Elliptic++	203,769	234,355	182	$\sim$2.2
DGraph	3,700,550	4,300,999	17	1.3

Note. This table reports the edge count supplied by the benchmark source. The selector statistics in Table 6 are computed on the graph operator actually used by EB-GAD after deterministic symmetrization, relation/operator selection, affinity reweighting, or transaction aggregation. Therefore $\#\mathrm{Edges}/(2n)$ here need not equal the operator-level d_e used by the fixed selector rule; the mismatched cases are due to the selected operators spelled out in Table 6. Anomaly rates are computed from the label files as loaded by our pipeline; for Weibo the released PyGOD label file contains 347 anomalous accounts of 8,405 (4.1%), not the 868 (10.3%) listed in the source documentation, and all methods in Table 1 are evaluated against the same released labels. Evaluation sets: Amazon and YelpChi are scored on the graphs of the TAM benchmark (10,224 nodes, 6.8% anomalous; 23,831 nodes, 5.1%); Elliptic and Elliptic++ on the 46,564 labeled transactions (9.8% illicit); DGraph on its 1,225,601 labeled nodes (1.3%).

B.2 Baseline Models

Here we provide more details on the baseline models covered in Table 1.

- **LOF** [4]: Local Outlier Factor, a density-based method that assigns each point a degree of outlierness based on the local density of its neighborhood. Implemented via PYOD [5].
- **DIF** [39]: Deep Isolation Forest, a deep learning extension of Isolation Forest that learns representations for anomaly detection. Implemented via PYOD.
- **ANOMALOUS** [24]: A graph-based method that models anomalies via residual analysis of the adjacency matrix. From the PYGOD library [18].
- **DOMINANT** [6]: Deep Anomaly Detection on Attributed Networks using a graph autoencoder with structure and attribute reconstruction. From PYGOD.
- **AnomalyDAE** [10]: Dual autoencoder for attributed-network anomaly detection, jointly modeling structure and attribute reconstruction. From PYGOD.
- **CONAD** [40]: Contrastive self-supervised learning for graph anomaly detection. From PYGOD.
- **CoLA** [21]: Contrastive Learning for Anomaly detection on graphs; leverages self-supervised contrastive objectives. From PYGOD.
- **TAM** [25]: Truncated Affinity Maximization, a one-class homophily method that learns local-affinity-based representations for graph anomaly detection.
- **DiffGAD** [16]: Diffusion-based graph anomaly detector that injects discriminative content into latent representations and uses diffusion/reconstruction signals for anomaly scoring.

B.3 Native-Score Evaluation and Orientation Audit

Table 1 evaluates every unsupervised baseline in its native anomaly-score direction, as fixed by the official implementation or paper: PyOD scores use `decision_scores_`, PyGOD scores use `decision_score_`, TAM uses its published local-affinity anomaly score, and DiffGAD uses its reconstruction/diffusion anomaly score. This convention is important for a label-free comparison. Reversing a score after observing test labels is an oracle operation, so it is not used in the main table even when a native score is anti-correlated with the benchmark labels.

To check whether the sub-50 entries in Table 1 reflect missing data or merely native-score anti-correlation, Table 4 reruns the threatening PyOD/DiffGAD fraud cases and reports both the native

Table 4: Score-orientation audit for sub-50 baselines whose oracle sign reversal could affect the fraud-dataset comparison. Native AUROC uses the method's fixed label-free anomaly-score direction and is the only value used in Table 1. Oracle-flipped AUROC is computed from the negated score after labels are known and is diagnostic only.

Dataset	Baseline	Native AUROC	Oracle-flipped AUROC	Mean score normal/anomaly
Elliptic	DIF	22.1±0.9	77.9±0.9	0.335/0.303
Elliptic++	DIF	22.6±0.5	77.4±0.5	0.334/0.303
Elliptic	DiffGAD	27.1±0.4	72.9±0.4	20.90/13.31
Elliptic++	DiffGAD	27.9±7.8	72.1±7.8	967.09/417.67

AUROC and the diagnostic oracle value obtained by negating the scores. The flipped column is not a competing baseline; it only shows how much a label-based score reversal would change the ranking. In every audited case, normal nodes receive larger native scores on average than anomalous nodes, so the sub-50 AUROCs are genuine anti-correlations of the native baseline scores.

B.4 Selector and Hyperparameter Specification

Scope. Main §3.4 introduces the selector at the level of principles: four score families tied to graph statistics, a quality plus stability score $A(r)$, and the role of EB in fitting (ρ, κ). This section states the rule exactly as it is run; the released code contains the same implementation and the script that regenerates Table 1. No rule takes a label as input. The cutoffs are implementation defaults, identical on all 11 datasets, not derived quantities: they were fixed during the development of the method on the benchmark families studied here, so they are part of its engineering. Table 7 classifies every constant, and Appendix C.5 reports what happens when the cutoffs are perturbed.

Preprocessing, spectrum and grids. Features are z-scored. For the equilibrium and profile families, features with $D_x > 100$ are projected on their first 64 principal components; PCA is a deterministic unsupervised linear map, so the closed-form results apply in the projected space. The full Laplacian spectrum is used when $n < 20{,}000$; larger graphs use the k smallest eigenpairs with $k = 500$ ($n < 10^5$), 300 ($n < 5{\cdot}10^5$), 200 ($n < 10^6$) or 128. Dense eigendecompositions run in double precision, and LOBPCG bases are orthonormalized and Ritz-refined in double precision. All computed modes, including Laplacian null modes, are kept in the modeled subspace (Appendix A.1). The bandwidth grid holds the three values of $\{0.1, 0.2, 0.5, 0.7, 1, 2, 5\}$ that are closest on a log scale to $\gamma_0 = \mathrm{clip}((0.5 + h)/\sqrt{\max(d_e, 10)/10}, 0.05, 10)$ among those not larger than $2\gamma_0$ (DGraph, whose spectrum is computed once and cached, uses $\{0.5, 1, 2\}$). The run and its finite-horizon families are built at $\gamma^\star$, the grid value with the largest residual-space profile likelihood. The equilibrium anchor is computed at γ_a, the grid value closest to γ_0 on the log scale, with the EB fit of (ρ, κ) at γ_a (Table 14; Appendix C.1 gives the reason and the alternatives). The other grids are $\kappa \in \{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 0.5, 1, 2, 3, 5, 7, 10, 15, 20\}$, $\rho \in [0, 1]$ by a bounded one-dimensional search (clamped Newton or Brent's method) with an endpoint check (Remark 1), and horizons $\Gamma \in \{0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 50, \infty\}$; the control-energy and control-ratio pools use $\Gamma \in \{0.25, 0.5, 1, 2, 3, 4, 8, 12\}$ and $\lambda_c \in \{0.3, 0.5, 1, 2, 5, 10, 50, 100, 500\}$.

Affinity template. The "affinity" template in Table 5 keeps the edge set of the graph operator fixed but multiplies features by deterministic node reliability weights before smoothing,

$$m_{\mathrm{aff}} = S_{\gamma,\nu} W X, \qquad W = \mathrm{diag}(w_1, \ldots, w_n),$$

where $w_i \in [0, 1]$ is the min–max normalized local affinity, $a_i = \deg(i)^{-1} \sum_{j \in N(i)} \cos(z_i, z_j)$. When PCA is enabled, it is applied right after z-scoring and before the affinity weights. The Laplacian is that of the input graph in every run of Table 1.

Selector statistics. The graph statistics (h, d_e, D_x) are computed from the graph operator and the node features as provided (no class labels enter h or any selector diagnostic):

$$h = \frac{1}{|E_s|} \sum_{(i,j) \in E_s} \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2 + \epsilon}, \qquad d_e = |E_{\mathrm{dir}}|/(2n),$$

where E_s is a fixed sample of up to 10^5 edges. The operator-level d_e in Table 6 is not recomputed from the raw edge column in Table 3: Amazon uses the filtered U–P–U user–product–user relation (n=10,224, $|E_{\mathrm{dir}}|$=351,216); YelpChi uses the filtered R–U–R review–user–review relation

(n=23,831, $|E_{\mathrm{dir}}|$=98,630); BlogCatalog reports undirected raw edges but the operator is bidirectional ($|E_{\mathrm{dir}}|$=345,566); ACM uses the symmetrized input graph ($|E_{\mathrm{dir}}|$=164,350); T-Finance uses the BWGNN-released homogenized transaction graph [29] ($|E_{\mathrm{dir}}|$=42,445,086); and Weibo, Reddit, Facebook, Elliptic, Elliptic++, DGraph use raw conventions. For DGraph, h and d_e are computed on the full graph and the spectrum on the subgraph of its labeled nodes ($k = 128$, $n = 1{,}225{,}601$). h saturates near one for correlated non-negative features (e.g., Reddit), so it enters only through coarse cutoffs.

Ordered rules. The selector chooses among four score families: equilibrium anchors ($J^\star$ or R), control-energy scores (C_{Γ,λ_c}), control-ratio scores (CR_{Γ,λ_c}), and profile aggregation over the finite-horizon bank. The rules are evaluated in order and the first that applies is used:

1. If $d_e < 1$, $D_x \le 256$, and $k/n \le 5 \cdot 10^{-3}$, use the equilibrium ratio anchor R.
2. Else if $h < 0.05$, $D_x > 1000$, and $d_e \ge 20$, use the control-energy family on the affinity template.
3. Else if $0.05 \le h < 0.25$, $D_x > 1000$, and $d_e < 10$, use the control-ratio family on the affinity template.
4. Else if $d_e \ge 100$, use profile aggregation (trajectory branch); in the intermediate dense case $10 \le d_e < 100$, this family is used only if the dense profile diagnostic passes and the anchor guard is inactive.
5. Else if $h \ge 0.50$ and $D_x \le 128$: if the sparse-tail diagnostic passes, use profile aggregation (sparse-tail branch); otherwise use the control-energy family when $d_e \ge 10$ and the equilibrium anchor when $d_e < 10$.
6. Else if $0.25 \le h < 0.50$, $128 < D_x \le 1000$, $d_e \ge 20$, and the scale-entropy diagnostic passes, use profile aggregation (scale-entropy branch).
7. Otherwise use the equilibrium anchor.

The diagnostics are unlabeled. The *dense profile* diagnostic requires $\mathrm{NullKS}(J^\star) \ge 0.50$ and a Berk–Jones statistic [2] of the profile p-values in $[20, 120]$. The *anchor guard* is active when NullKS selects $J^\star$ with $\mathrm{NullKS}(J^\star) \ge 0.60$. The *sparse-tail* diagnostic requires Berk–Jones in $[20, 100]$, equilibrium NullKS below 0.20, and profile NullKS below 0.35. The *scale-entropy* diagnostic requires a two-groups mixture mass $\pi \in [0.50, 0.95]$ and that the tail-stability and graph-tail criteria both pick the component-entropy score.

Within-family score. *Equilibrium anchor*: $J^\star$ or R at the anchor bandwidth γ_a, whichever has the larger NullKS there (ties go to $J^\star$; rule 1 fixes R). The diagnostics that the ordered rules read are those of the EB-fitted run at $\gamma^\star$. *Profile aggregation*: the scale-entropy branch uses the entropy of the two-groups component responsibilities of the path p-values; the sparse-tail branch uses the posterior of a two-groups mixture over the p-value bank; the trajectory branch uses the path band with the largest NullKS. *Control energy and control ratio*: the admissible pool is the set of candidates of the family on the rule's template (rule 2: affinity template, $\gamma \in [0.12, 0.25]$, PCA dimension $\lceil 256 \cdot 5/d_e \rceil$ rounded up to the grid $\{16, 24, 32, 48, 64\}$, $\Gamma \ge 1$, $\lambda_c \le 1$, NullKS in $[0.55, 0.85]$, Gini coefficient of the score in $[0.70, 0.90]$; rule 3: affinity template, $\gamma \in [0.075, 0.20]$, PCA 256, $\Gamma \le 0.5$, $\lambda_c \le 0.5$, Gini at most 0.05; rule 5: low-pass template, $\gamma = 5$, no PCA, NullKS in $[0.20, 0.50]$, Gini in $[0.25, 0.70]$). In rule 5 the two pool members with the largest NullKS are rank-averaged. In rules 2 and 3 the members are ranked by $A(r)$ (stability plus a 0.1-weighted quality term) and the top three are rank-averaged; if the pool is empty, all control-energy candidates are used.

Null model and diagnostic aggregation. Under the full-spectrum Gaussian model with fixed precision Q_ρ, $[Q_\rho^{1/2}\Delta]_{i,:} \sim \mathcal{N}(0, I_D)$ and so $2J_i^\star \sim \chi^2_D$ exactly. A finite-horizon member is a quadratic form with weights $c_j(\Gamma, \lambda_c)$ in place of q_j; under the same fitted law its null is a weighted chi-squared with known weights c_j/q_j. After EB fitting or spectral truncation these laws are used as plug-in nulls. For selection we summarize each empirical score distribution by a moment-matched scaled chi-squared $a_S \cdot \chi^2_{\hat{k}_S}$ and compute

$$\mathrm{NullKS}(S) := \mathrm{KS}\big(S/\hat{a}_S,\ \chi^2_{\hat{k}_S}\big), \qquad \hat{k}_S = 2\hat{\mu}_S^2/\hat{\sigma}_S^2, \quad \hat{a}_S = \hat{\sigma}_S^2/(2\hat{\mu}_S). \tag{19}$$

Larger NullKS indicates stronger unlabeled deviation from a chi-squared bulk. Each scalar diagnostic is converted to a percentile rank within the admissible pool before being combined with stability

terms, making KS distances and Spearman-rank stability commensurate; percentile ranks average ties. For profile aggregation, scores produce nodewise p-values p_{ir} that are combined by ACAT [20], minimum-p, mean $-\log p_{ir}$, or a two-groups mixture over the p-value bank. When κ is fitted at the upper end of its grid the plug-in p-values can saturate to a constant; constant members carry no ranking information and, with averaged ties, cannot contribute any.

Table 5: Label-free EB-GAD configuration behind each cell of Table 1. *Template*: low-pass Matérn or feature-affinity-weighted (§B.4). k: number of retained low-frequency Laplacian eigenpairs ("all" = full spectrum). ρ, κ, γ: EB-fitted graph trust and Matérn inverse length-scale, and the template bandwidth of the selected score (the anchor bandwidth γ_a for an anchor, the likelihood-selected bandwidth for a profile, the pool rule for control energy and ratio). *PCA*: retained PCA dimensions. *Rule*: the ordered rule of §B.4 that fires.

Dataset	Template	k	ρ	κ	γ	PCA	Rule	Selected family (score)
Weibo	low-pass	all	1.000	1	0.5	64	7	equilibrium anchor ($J^\star$)
Reddit	low-pass	all	1.000	1	2	no	5	equilibrium anchor (R)
Amazon	low-pass	all	0.458	0.5	5	no	5	control energy (top-2 C_{Γ,λ_c} fusion)
YelpChi	low-pass	500	1.000	20	1	no	5	equilibrium anchor ($J^\star$)
BlogCatalog	affinity	all	1.000	1	0.18	48	2	control energy (stability-ranked C_{Γ,λ_c} fusion)
Facebook	low-pass	all	0.727	3	0.7	64	6	profile aggregation (scale-entropy)
ACM	affinity	all	1.000	0.01	0.15	256	3	control ratio (stability-ranked CR_{Γ,λ_c} fusion)
Elliptic	low-pass	300	1.000	20	1	64	1	equilibrium anchor (R)
Elliptic++	low-pass	300	0.933	0.1	2	64	1	equilibrium anchor (R)
DGraph	low-pass	128	1.000	20	1	no	1	equilibrium anchor (R)
T-Finance	low-pass	500	1.000	0.003	0.2	no	4	profile aggregation (trajectory)

Table 6: Unlabeled quantities read by the selector. h: cosine feature homophily over a fixed sample of edges, d_e: edges per node of the graph operator, D_x: raw feature dimension, k/n: truncation ratio. Diagnostics of the EB-fitted run: NullKS of its two equilibrium scores, Berk–Jones statistic (BJ) and NullKS of the profile score, and the two-groups mixture mass π.

Dataset	h	d_e	D_x	k/n	NullKS($J^\star$)	NullKS(R)	BJ	NullKS(profile)	π
Weibo	0.068	24.27	400	all	0.663	0.156	78.5	0.258	0.786
Reddit	0.993	7.65	64	all	0.945	0.138	57.6	0.206	0.878
Amazon	0.644	17.18	25	all	0.277	0.071	236.2	0.208	0.740
YelpChi	0.876	2.07	32	2.10×10^{-2}	0.000	0.000	1225.2	0.334	0.024
BlogCatalog	0.009	33.25	8189	all	0.673	0.089	169.9	0.246	0.077
Facebook	0.375	25.49	576	all	0.036	0.044	8.3	0.227	0.811
ACM	0.143	4.99	8337	all	0.492	0.074	24.4	0.253	0.968
Elliptic	0.751	0.58	166	1.47×10^{-3}	0.853	0.909	1810.7	0.289	0.024
Elliptic++	0.978	0.58	182	1.47×10^{-3}	0.874	0.905	1987.2	0.289	0.024
DGraph	0.408	0.58	17	1.04×10^{-4}	0.000	0.000	0.0	0.000	0.024
T-Finance	0.817	539.23	10	1.27×10^{-2}	0.203	0.557	578.5	0.205	0.996

Table 7: Provenance of every quantity that enters a score or the selection. *Model-derived*: follows from the fitted model. *Canonical*: a standard statistic or combiner, with citation. *Default*: an implementation default, fixed once and identical on all datasets, with the sensitivity study that covers it.

Quantity	Class	Source or tested range
(ρ, κ); γ within its grid	model-derived	maximum of the residual-space likelihood (Alg. 1)
$J^\star$, R, C_{Γ,λ_c}, CR_{Γ,λ_c}	model-derived	closed forms of §3.3
plug-in nulls of the scores	model-derived	chi-squared laws under the fitted prior
Kolmogorov–Smirnov, Berk–Jones	canonical	[2]
ACAT, minimum-p, mean $-\log p$ combiners	canonical	[20]; Fisher's method
two-groups mixture of p-values	canonical	[8]
rank correlation as stability $S(r)$	canonical	Spearman correlation, average ties
grids for γ, κ, Γ, λ_c	default	listed in this section
PCA dimension, truncation k by graph size	default	listed in this section
graph-statistic cutoffs on h, d_e, D_x, k/n	default	$\pm25\%$ and $\pm50\%$ (Table 18)
diagnostic windows (NullKS, Berk–Jones, π)	default	$\pm25\%$ and $\pm50\%$ (Table 18)
pool windows of rules 2, 3, 5 (NullKS, Gini)	default	listed in this section; not perturbed
weights inside $A(r)$	default	ingredient ablation (Table 19)

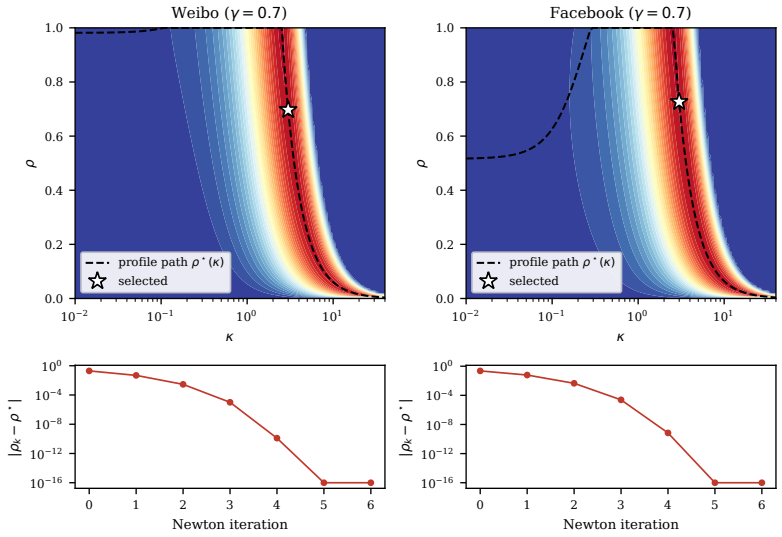

Figure 2: **Top:** residual-space log-likelihood over (ρ, κ) of the EB-fitted run (likelihood-selected bandwidth $\gamma^\star = 0.7$) for Weibo and Facebook (values below the 35th percentile are clipped for display). Dashed curve: profile path $\rho^\star(\kappa)$; star: selected optimum on the κ grid. The likelihood is concave in ρ (Remark 1) with dataset-dependent structure in κ. **Bottom:** Newton iterates for ρ at the selected $\kappa^\star$.

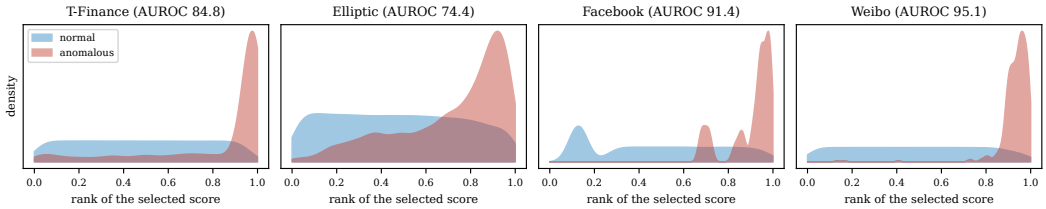

Figure 3: Density of the selected EB-GAD score for normal and anomalous nodes on four datasets (original node order). Scores are shown on their rank scale, so a method at chance would give two flat densities.

B.5 Equilibrium Scores

Table 8 reports the two equilibrium scores at the anchor bandwidth and the one that NullKS selects without labels. For the six datasets routed to an anchor this is the entry of Table 1; for the other five, the comparator of main Table 2 is the equilibrium score at the configuration of the selected score. NullKS picks the better of the two on Weibo, Amazon, BlogCatalog and ACM, and the worse on Reddit, Facebook and T-Finance; on Facebook the difference is large (47.4 against 89.7) and is the cost that the cutoff study of Appendix C.5 reports. On YelpChi and DGraph the fitted κ is at the upper end of its grid, both NullKS values are zero, and the tie goes to $J^\star$ (DGraph, like Elliptic and Elliptic++, is assigned R by rule 1).

Table 8: Equilibrium anchor: AUROC (%) of $J^\star$ and R at the anchor bandwidth γ_a (the grid value nearest to γ_0) with the EB fit of (ρ, κ) at that bandwidth. **Bold**: the score chosen without labels (larger NullKS; ties go to $J^\star$; rule 1 of §B.4 fixes R on sparse truncated graphs). Last row: the score selected in Table 1.

	Weibo	Reddit	Amazon	YelpChi	BlogCat.	Facebook	ACM	Elliptic	Ell.++	DGraph	T-Fin.
γ_a	0.5	2	1	1	0.2	0.5	0.7	1	2	1	0.2
$J^\star$	**95.1**	62.0	**75.8**	**71.7**	54.0	**47.4**	**75.6**	58.3	30.2	60.5	81.6
R	47.3	**60.6**	48.2	64.6	**61.0**	89.7	72.1	**74.5**	**72.5**	**66.8**	**80.1**
Table 1	95.1	60.6	78.0	71.7	78.6	91.4	86.0	74.5	72.5	66.8	84.8

B.6 Resource Utility and Scalability Measures

Baselines were trained and evaluated with their reported hyperparameters on NVIDIA A100 40GB GPUs. The EB-GAD results of Table 1 were regenerated on CPU. For medium and large graphs we use a truncated eigendecomposition; the retained k for each dataset is reported in Table 5.

To scale EB-GAD to large graphs, we adopt the following efficiency measures:

- **Truncated eigendecomposition:** For graphs with $n \geq 20$K nodes we use the Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG [14]) method, with a seeded initial block, to compute the k smallest eigenpairs of the sparse Laplacian; the basis is then orthonormalized and Ritz-refined in double precision. These are the smooth graph modes used by the low-pass template and the Matérn precision; unretained rough modes are omitted from the likelihood and the control-energy numerator, while ratio scores still normalize by the full nodewise residual norm. The eigenvector matrix is $(n \times k)$ rather than $(n \times n)$, reducing memory from $O(n^2)$ to $O(nk)$; k is scaled by graph size (Table 5).
- **Eigenspace operations:** All graph filters and prior operations use products $V(\mathbf{f}(\lambda) \odot (V^\top \mathbf{x}))$ in $O(n \cdot k \cdot D)$ per step, avoiding dense matrix exponentials.
- **Sparse Laplacian:** For graphs with $n \geq 8$K nodes, we use a sparse construction to avoid materializing dense $n \times n$ matrices.
- **Eigen cache:** The dense precision matrix Q is skipped for large graphs; all scoring operations use the precomputed eigen cache instead.

B.7 Computation Time

Table 9 reports wall-clock times. The baseline times are training plus inference on one NVIDIA A100 GPU, as measured for the submitted version. The EB-GAD times were measured for this version on 16 CPU threads of a shared server and cover the complete run behind a cell of Table 1: loading the data, one double-precision eigendecomposition per template bandwidth of the grid, the EB fits, every score family that the run computes, and the label-free selection. They were measured before the anchor bandwidth γ_a was introduced; the anchor reuses the residual that the run computes at γ_a, a value of its grid, and adds one EB fit and two scores. The times are upper bounds for a deployment, which would cache one eigendecomposition and compute one family. The two sides use different hardware, so the table indicates orders of magnitude and is not a controlled comparison. EB-GAD takes 8 seconds on Facebook and 33 minutes on DGraph, where the eigendecomposition of the labeled subgraph accounts for 30 minutes. The candidate-pool datasets (Amazon, BlogCatalog, ACM) take 11 to 26 minutes because every candidate configuration recomputes the eigendecomposition; the implementation also recomputes it for every template bandwidth although the Laplacian does not depend on the bandwidth, so caching it would reduce all times.

Table 9: Wall-clock time (seconds). Baselines: training and inference on one NVIDIA A100 GPU, as measured for the submitted version. EB-GAD: the complete run behind its cell of Table 1 on 16 CPU threads, from loading the data to the selected score: eigendecompositions in double precision for every template bandwidth of the grid, EB fits, all score families of the run and the label-free selection. The two columns use different hardware and are not a controlled comparison. OOM = out of memory.

Dataset	ANOM.	DOMINANT	AnomDAE	CONAD	CoLA	DiffGAD	TAM	EB-GAD (CPU)
Weibo	78	63	1	95	130	134	858	84
Reddit	38	35	2	57	8	110	5,467	136
Amazon	25	37	153	308	14	106	10,739	708
YelpChi	170	185	1,152	1,359	5	118	5,106	27
BlogCatalog	447	430	92	3,047	18	3,272	8,764	675
Facebook	3	20	9	56	3	118	479	8
ACM	6,503	290	580	2,293	26	39,209	14,446	1,575
Elliptic	OOM	19,802	OOM	OOM	263	294	1,176	1,101
Elliptic++	OOM	16,962	OOM	OOM	288	292	1,186	1,429
DGraph	OOM	OOM	OOM	OOM	4,117	OOM	3,929	1,956
T-Finance	963	3,425	9,761	5,028	1,107	384	OOM	1,471

B.8 Dimensionality Reduction and Nonlinear Extensions

The main text uses PCA as a training-free dimensionality reduction that preserves the linear-Gaussian structure. Here we discuss the theoretical landscape of dimensionality reduction for EB-GAD and report results with trained nonlinear encoders.

Linear reductions preserve the Gaussian structure. Feature-space linear maps preserve Gaussianity. If $z = (I_n \otimes W)x$ and $x \sim \mathcal{N}(\mathbf{m}, Q_\rho^{-1} \otimes I_D)$, then $z \sim \mathcal{N}((I_n \otimes W)\mathbf{m}, Q_\rho^{-1} \otimes WW^\top)$. For PCA with orthonormal retained components (and after standardization), $WW^\top = I_p$, so the same closed-form anomaly scores and residual-space likelihood apply exactly in the projected space. PCA is deterministic, unsupervised, and maximizes retained variance among linear projections.

Nonlinear reductions break the Gaussian structure. When a nonlinear encoder $f_\theta : \mathbb{R}^D \to \mathbb{R}^p$ is applied, the latent features $z = f_\theta(x)$ are no longer Gaussian even if x is. The residual-space likelihood objective is no longer available in closed form, and a variational bound is needed for prior selection.

Variational formulation (ELBO). For this nonlinear-encoder extension only, one could construct a terminally conditioned GOU bridge via Doob's h-transform on a discrete time grid and optimize a reconstruction-plus-KL variational objective in the spirit of learned diffusion-bridge/reverse-transition models [44, 41, 45]. The bound would combine a per-step KL between the analytic bridge transition q_φ and the learned reverse transition p_θ with an endpoint reconstruction term; in the linear-Gaussian case (linear encoder or no encoder), the bridge transitions match the true posterior exactly, the variational gap vanishes, and the bound reduces to the exact residual-space objective (17). This bridge is not used in the main EB-GAD pipeline, which relies on the unconditioned GOU transition law; we mention the ELBO only to indicate how a future nonlinear-encoder extension would inherit the empirical-Bayes structure.

Non-neural nonlinear options. Beyond neural encoders, several non-neural nonlinear methods could extend the framework:

- *Kernel PCA*: maps features to a reproducing kernel Hilbert space via a kernel function (e.g., RBF). Non-parametric and training-free, but the infinite-dimensional feature space requires truncation, and the Gaussian assumption is approximate.
- *Random Fourier Features* [26]: approximates kernel PCA with explicit features $z(x) = \cos(Wx + b)$ for random W. Deterministic, finite-dimensional, and training-free, but the nonlinear map makes the likelihood objective approximate.
- *GPLVM* [15]: a principled Bayesian nonlinear latent variable model with a natural variational formulation. Requires optimization of variational parameters, making it no longer training-free.

These represent a spectrum from fully closed-form (PCA) to fully variational (GPLVM), trading tractability for expressive power. We leave exploration of these extensions to future work.

Empirical comparison: training-free vs. trained encoders. Table 10 compares the training-free results of Table 1 against the better of two trained nonlinear encoders (MLP or GraphSAGE) for each dataset; choosing the better encoder by AUROC favors the trained side. Trained encoders are applied as a preprocessing step: the encoder is trained separately, and the same score formula is then applied to the latent features. *Training setup:* Adam optimizer (learning rate 10^{-3}, weight decay $5 \cdot 10^{-4}$), 200 epochs with early stopping on the unsupervised reconstruction loss, batch size 1024 (full-batch on small graphs), 2 hidden layers of width 64 with ReLU; for GraphSAGE, mean-aggregator with neighbor sampling sizes $[25, 10]$; for the small-encoder retries, hidden width 16 and 1 hidden layer. For T-Finance and DGraph, we use a small feature-only retry to avoid dense structure reconstruction; the DGraph row reports the best of seven completed small-encoder configurations. The trained-encoder numbers are those of the submitted version; they were not rerun.

Reading. The training-free pipeline is ahead on 8 of 11 datasets, by 1.0 to 22.3 points (Reddit, Amazon, YelpChi, BlogCatalog, ACM, Elliptic, Elliptic++, DGraph). A trained encoder is ahead on Weibo and Facebook, by 0.6 and 4.0 points, and on T-Finance by 0.1. The encoder helps on two social

Table 10: Training-free EB-GAD (Table 1) against the better of two trained encoders (MLP, GraphSAGE), chosen per dataset by AUROC, followed by the same score formula. AUROC (%); gap = training-free minus trained.

Dataset	Training-free	Best trained encoder	Gap
Weibo	95.1	**95.7** (MLP)	-0.6
Reddit	**60.6**	58.9 (MLP)	$+1.7$
Amazon	**78.0**	60.3 (MLP)	$+17.7$
YelpChi	**71.7**	68.0 (SAGE)	$+3.7$
BlogCatalog	**78.6**	77.6 (MLP)	$+1.0$
Facebook	91.4	**95.4** (MLP)	-4.0
ACM	**86.0**	63.7 (MLP)	$+22.3$
Elliptic	**74.5**	56.4 (SAGE)	$+18.1$
Elliptic++	**72.5**	57.7 (SAGE)	$+14.8$
DGraph	**66.8**	58.8 (SAGE-small)	$+8.0$
T-Finance	84.8	**84.9** (SAGE-small)	-0.1

graphs and hurts on five of the six fraud graphs, on ACM, on BlogCatalog and on Reddit. Where it hurts, two causes are plausible: a nonlinear encoder breaks the Gaussian structure, so the exact residual-space objective (17) no longer applies to the latent features, and it can distort the relation between graph frequencies and feature variation on which the scores operate. The training-free pipeline is also deterministic given the eigenbasis, whereas trained encoders add seed-dependent variance.

C Regenerated Results and Additional Studies

Every EB-GAD number in this paper, in the main tables and in the studies below, was produced for this version by one pipeline. Wherever a score is selected, it is selected without labels; "best member" columns report the best score of a pool by AUROC and are diagnostics, not results. The studies of the finite-horizon bank (isolation, contamination, misspecification, neighboring anomalies) run at the EB-fitted configuration, that is, at the likelihood-selected bandwidth $\gamma^\star$; where the anchor bandwidth γ_a differs, the text also quotes the study at γ_a.

C.1 Protocol, Relabeling Gate and Changes from the Submitted Version

Protocol. The procedure of Appendix B.4 follows Algorithm 1 and the modeling details of Appendix A.1 as submitted, and no label enters any step before evaluation. Four points are made explicit. (i) The equilibrium anchor is computed by Algorithm 1 on every dataset, at the anchor bandwidth γ_a; no per-dataset configuration is used. (ii) All computed Laplacian modes, null modes included, are kept in the modeled subspace. (iii) Eigendecompositions run in double precision. (iv) Ranks average ties. The protocol file is part of the released code, together with the scripts that regenerate every table.

Relabeling gate. The model is equivariant to node relabeling, so every reported score must be unchanged, up to the relabeling, when the nodes are randomly renumbered. Table 11 reports this check for the scores of Table 1. On the six datasets with a full spectrum the selected score is the same under relabeling: the rank correlation is at least 0.9994 and the AUROC differs by at most 0.03. With a truncated spectrum the computed basis depends on the node order through the iterative eigensolver. The effect is absent on T-Finance and small on Elliptic, Elliptic++ and DGraph (standard deviation at most 0.3 over five node orders). It is largest on YelpChi, whose graph has 7,308 connected components with at most 47 nodes: all 500 retained modes lie in the Laplacian nullspace, the basis is not unique, and the AUROC ranges from 70.7 to 72.7. Table 1 reports the mean over the five node orders for the truncated datasets.

Changes from the submitted version. Four EB-GAD entries of Table 1 are unchanged up to rounding (Amazon, BlogCatalog, Facebook, ACM), together with their fitted configurations. The others changed; we list every change and its cause.

- *YelpChi (*$95.3 \to 71.7$*).* The submitted entry was an artifact of the node order. At the fitted κ, the upper end of its grid, the plug-in p-values of all path bands are constant; the rank transform then broke ties by node index, and the released node order of YelpChi is correlated with the labels (the node index alone has AUROC 94.4). With the tie-breaking of the submitted version the profile

Table 11: Relabeling gate for the scores of Table 1. The model is equivariant to node relabeling, so the selected score must not depend on the node order. *Orders*: number of node orders run (the original and random relabelings). AUROC (%): smallest, largest and standard deviation over the orders. Spearman: rank correlation of the selected score between the original order and one relabeling.

Dataset	Spectrum	Orders	min	max	std	Spearman
Weibo	full	2	95.12	95.12	0.00	1.0000
Reddit	full	2	60.56	60.56	0.00	0.9995
Amazon	full	2	78.05	78.05	0.00	1.0000
YelpChi	truncated	5	70.72	72.65	0.62	0.8368
BlogCatalog	full	2	78.62	78.62	0.00	1.0000
Facebook	full	2	91.40	91.40	0.00	1.0000
ACM	full	2	86.02	86.02	0.00	1.0000
Elliptic	truncated	5	74.01	74.82	0.27	0.9956
Elliptic++	truncated	5	72.45	72.55	0.03	0.9990
DGraph	truncated	5	66.63	66.93	0.10	–
T-Finance	truncated	5	84.75	84.75	0.00	1.0000

score has AUROC 93.7 in the released order and 60.3 after a random relabeling; with averaged ties it has 67.3 and 66.8 (Table 12). The sparse-tail diagnostic no longer passes, and the rules route YelpChi to the equilibrium anchor. The gain of 24.6 points attributed to the finite horizon on YelpChi in the submitted ablation was the same artifact. Facebook, whose node order is also correlated with the labels (index AUROC 97.9), is not affected: its score involves no rank step and is identical under relabeling.

- *Weibo (*$95.1 \to 95.1$*) and Reddit (*$61.8 \to 60.6$*).* The submitted equilibrium anchors used fixed per-dataset configurations (template bandwidth, (ρ, κ) and score type) carried over from an earlier version of the pipeline, not a configuration that the algorithm selects. They are replaced by the anchor of this version: bandwidth γ_a, EB fit of (ρ, κ) at that bandwidth, score type by NullKS. On Weibo $\gamma_a = 0.5$ is also the submitted bandwidth; on Reddit $\gamma_a = 2$ (submitted: 1) and NullKS selects R. With the bandwidth that the residual likelihood selects, $\gamma^\star = 0.7$ on both datasets, the entries are 92.5 and 48.5; the paragraph on the bandwidth of the anchor below explains the choice and how it was made.
- *T-Finance (*$86.8 \to 84.8$*).* The submitted value was the best path band by AUROC; the band with the largest NullKS gives 84.8.
- *Elliptic (*$73.0 \to 74.5$*), Elliptic++ (*$72.4 \to 72.5$*), DGraph (*$66.9 \to 66.8$*).* Same selection (R by rule 1); the values changed with the numerical corrections below and are now means over five node orders. The anchor bandwidth is $\gamma_a = 1$ on Elliptic and DGraph and $\gamma_a = 2$ on Elliptic++, where $\gamma = 1$ gives 73.7.
- *Numerics.* Dense eigendecompositions ran in single precision, in which the zero eigenvalues of a Laplacian are of order 10^{-7} with either sign. A flag of the submitted runs removed modes with eigenvalue below 10^{-8}, so whether a null mode was removed depended on rounding, and with it on the node order and the platform. Eigendecompositions now run in double precision, LOBPCG uses a seeded initial block, and its basis is Ritz-refined in double precision. All modes are kept, as the modeling details state; with this convention the four entries above and their configurations are reproduced, and Table 13 reports the other convention.
- *Ablation table.* The submitted bank ablation compared scores at different configurations, so its differences mixed the effect of the horizon with the effect of the configuration. Table 2 now compares at the same configuration.
- *Text and tables.* The grids, the definition of h and the statistics table of Appendix B.4 now state what the code runs. The studies of this appendix were rerun with the same pipeline, so some of their numbers differ from those quoted during the review.

Bandwidth of the anchor. The submitted Algorithm 1 selects one bandwidth, by the residual-space likelihood, for every score. In the first regeneration for this version the anchor used that bandwidth, which gave 92.5 on Weibo and 48.5 on Reddit. The residual $\Delta = T_{\gamma,\nu} X$ is a different function of the data at each bandwidth, and its likelihood is largest at the grid value whose residual map removes the largest share of the feature energy, on all eleven datasets (79 to 95% on the six full-spectrum datasets; Table 14). The criterion therefore measures how much of the features the template explains. The p-values of the finite-horizon families are computed under the fitted law of that residual and keep $\gamma^\star$;

Table 12: The ordering artifact. AUROC (%) of the profile scores and of the equilibrium scores on YelpChi and Facebook, in the released node order and after a random relabeling, with the tie-breaking of the submitted version (ties ordered by node index) and with averaged ties. The equilibrium scores involve no rank step. Last column: AUROC of the node index itself.

		ties by node index		averaged ties		
Dataset	Score	released	relabeled	released	relabeled	node index
YelpChi	profile, mixture posterior	93.7	60.3	67.3	66.8	94.4
	profile, mean $-\log p$	93.5	60.6	68.7	68.3	
	profile, ACAT	92.8	61.9	69.3	69.8	
	$J^\star$	70.9	70.7	70.9	70.7	
	R	64.2	64.0	64.2	64.0	
Facebook	profile, component entropy	91.4	91.4	91.4	91.4	97.9
	$J^\star$	63.7	63.7	63.7	63.7	
	R	87.8	87.8	87.8	87.8	

how well a single energy of the residual ranks anomalies is a different question, which the likelihood does not address.

In this version the anchor uses γ_a, the grid value nearest to γ_0. The center $\gamma_0 = (0.5 + h)/\sqrt{\max(d_e, 10)/10}$ is small for dense graphs with low feature homophily and large for sparse homophilic ones. For $\gamma < 1$ the residual filter $1 - (\gamma^2 + \lambda)^{-1}$ is negative and large in magnitude on low graph frequencies, so the anchor reads the smoothed signal of a node's neighborhood (clustered anomalies, as on Weibo); for $\gamma \geq 1$ it is close to the deviation of the node itself (Reddit and the fraud graphs). The center and the grid are those of the submitted code, the rule is the same on every dataset, and no label enters it.

The decision to adopt the rule, however, was taken after the anchors at both bandwidths had been evaluated on the labels, and its effect is not uniform. Against the equilibrium score at $\gamma^\star$, the one at γ_a is higher on Weibo, Reddit and Amazon, lower on BlogCatalog, Facebook and Elliptic++, and identical on the five datasets where the two bandwidths coincide (Table 14). Of the six datasets where they differ, three are routed to an anchor: Table 1 gains 2.6 points on Weibo and 12.1 on Reddit and loses 1.2 on Elliptic++. The rule is therefore not a blind choice, and Table 13 reports the alternatives next to it: the anchor at the likelihood-selected bandwidth, and the anchor averaged over the three bandwidths of the grid (mean percentile rank), which gives 95.3 on Weibo and 60.7 on Reddit, raises Elliptic and Elliptic++ (to 75.1 and 75.9) and lowers YelpChi and DGraph (to 64.0 and 66.1). Using γ_0 itself instead of the nearest grid value gives 95.5 on Weibo and 59.8 on Reddit and lowers YelpChi to 62.9 and Elliptic to 70.8. The rule is restricted to the anchor: applied to every family it would set $\gamma = 0.5$ on Facebook, where the scale-entropy diagnostic then fails and the rules fall back to $J^\star$ (47.4 instead of 91.4). The choice of the score type by NullKS is the same in every node order at every anchor bandwidth; among the other grid values it changes with the node order on Reddit at $\gamma = 1$ (NullKS of R: 0.45 and 0.24) and on YelpChi at $\gamma = 2$. For the anchor the bandwidth is thus a declared part of the prior and not a fitted quantity, and the Weibo and Reddit entries should be read with this in mind.

Sensitivity to four conventions. Table 13 reruns Table 1 under four alternative conventions, with the rules and the within-family scores unchanged. The first two are the anchor variants of the previous paragraph. If the Laplacian null modes are excluded from the modeled subspace, the entries of the six full-spectrum datasets change by -6.6 to $+0.1$ points; the largest changes are BlogCatalog ($78.6 \to 72.0$) and ACM ($86.0 \to 82.1$). If every bandwidth, including the anchor's, is chosen by the likelihood with the Jacobian term of the residual map, which Section 3.4 omits, Reddit is at 61.9 while Weibo, YelpChi, Facebook and Elliptic fall by 1.6 to 5.6 points.

C.2 AUPRC

Table 15 reports AUPRC for the comparison of Table 1. EB-GAD is best on five fraud columns: YelpChi (9.7 against 7.8 for the best baseline), Elliptic (22.1 against 15.3), Elliptic++ (19.1 against 12.2), DGraph (2.3 against 1.9) and T-Finance (53.0 against 6.6). With the anchor bandwidth of this version it is also best or tied-best on Weibo (34.9 against 34.6 for the best baseline) and Reddit (4.5, the value of TAM and CoLA). It is second on Amazon, third on BlogCatalog and ACM, and fourth on Facebook. On Facebook its AUROC ties the best (91.4) while its AUPRC (17.2) trails TAM (22.3),

Table 13: Sensitivity of Table 1 to four conventions (original node order, AUROC %; "=": same score and value; "–": not applicable, truncated spectra keep all computed modes; n/a: not run). *Anchor, likelihood*: the anchor uses the bandwidth with the largest residual-space likelihood, as in the submitted Algorithm 1. *Anchor, averaged*: the anchor is the mean percentile rank of the equilibrium score over the three bandwidths of the grid. *Nullspace dropped*: Laplacian null modes are excluded from the modeled subspace on the full-spectrum datasets. *Jacobian*: every bandwidth, including the anchor's, maximizes the likelihood with the term $D \sum_j \log \max(t_j, 10^{-12})$ of the residual map $t_j = 1 - (\gamma^2 + \lambda_j)^{-1}$, which excludes bandwidths whose map is negative on a computed mode. The ordered rules and the within-family scores are unchanged.

Dataset	Protocol	Anchor, likelihood	Anchor, averaged	Nullspace dropped	Jacobian
Weibo	95.1	92.5	95.3	95.2	92.5
Reddit	60.6	48.5	60.7	60.4	61.9
Amazon	78.0	=	=	76.5	=
YelpChi	70.7	=	64.0	–	69.1
BlogCatalog	78.6	=	=	72.0	=
Facebook	91.4	=	=	90.8	85.8
ACM	86.0	=	=	82.1	=
Elliptic	74.4	=	75.1	–	70.9
Elliptic++	72.5	73.6	75.9	–	=
DGraph	66.8	=	66.1	–	n/a
T-Finance	84.8	=	=	–	=

Table 14: Template bandwidth. γ_0: graph-statistic center of the grid. For every bandwidth γ of a dataset's grid: AUROC (%) of the equilibrium score chosen by NullKS after the EB fit at that bandwidth (low-pass template; R on Elliptic, Elliptic++ and DGraph, where rule 1 fixes it) and, in parentheses, the share (%) of the feature energy that the residual map removes (0 when the residual has more energy than the features; the shares are small on truncated spectra, where the map acts on the computed modes only). **Bold**: the anchor bandwidth γ_a (grid value nearest to γ_0 on the log scale). $^{\ell}$: the bandwidth with the largest residual-space likelihood, used by the finite-horizon families. The AUROC values are diagnostics; no rule reads them. Original node order.

Dataset	γ_0	γ: AUROC (removed share), in increasing γ		
Weibo	0.36	0.2: 95.6 (0)	**0.5: 95.1** (59)	0.7: 92.5$^{\ell}$ (86)
Reddit	1.49	0.7: 48.5$^{\ell}$ (89)	1: 51.8 (87)	**2: 60.6** (39)
Amazon	0.87	0.5: 53.5 (12)	0.7: 63.8$^{\ell}$ (79)	**1: 75.8** (77)
YelpChi	1.38	0.7: 63.3 (0)	**1: 70.9**$^{\ell}$ (3)	2: 69.1 (1)
BlogCatalog	0.28	0.1: 44.9 (22)	**0.2: 61.0** (92)	0.5: 74.8$^{\ell}$ (95)
Facebook	0.55	**0.5: 47.4** (67)	0.7: 87.8$^{\ell}$ (86)	1: 85.9 (77)
ACM	0.64	0.5: 60.8 (26)	**0.7: 75.6**$^{\ell}$ (85)	1: 78.4 (83)
Elliptic	1.25	0.7: 70.2 (0)	**1: 74.4**$^{\ell}$ (3)	2: 70.9 (1)
Elliptic++	1.48	0.7: 71.2 (0)	1: 73.7$^{\ell}$ (3)	**2: 72.5** (1)
DGraph	0.91	0.5: 64.4 (0)	**1: 66.8**$^{\ell}$ (0)	2: 64.4 (0)
T-Finance	0.18	0.1: 82.5 (0)	**0.2: 80.1**$^{\ell}$ (0)	

Table 15: AUPRC (%) for the comparison of Table 1. **Bold**: best; underline: second. *: published in the TAM benchmark [25], used exactly where Table 1 follows that benchmark; other baseline cells are our runs. OOM = out of memory; n/a = out of memory and the published report gives AUROC only. EB-GAD cells are the AUPRC of the score selected without labels in Table 1.

Method	Weibo	Reddit	Amazon	YelpChi	BlogCat.	Facebook	ACM	Elliptic	Ell.++	DGraph	T-Fin.
LOF	7.0	4.1	6.7	7.0	24.7	2.1	14.6	10.2	8.5	0.9	3.2
DIF	3.9	3.7	15.4	6.4	7.4	1.8	3.5	6.0	6.1	1.9	3.2
ANOMALOUS	34.6	3.9	5.6*	5.2*	6.5*	19.0*	6.4*	OOM	OOM	OOM	4.8
DOMINANT	22.0	3.7	14.2*	4.0*	31.0*	3.1*	44.0*	5.3	7.4	OOM	6.6
AnomalyDAE	30.9	3.1	10.5	6.7	36.7	2.7	27.0	5.4	6.0	OOM	3.1
CONAD	22.8	3.7	10.8	3.3	11.1	1.3	11.3	5.3	7.3	n/a	6.6
CoLA	4.3	**4.5***	6.8*	4.5*	32.7*	21.1*	32.4*	15.3	12.2	1.4	3.4
DiffGAD	32.3	3.7	12.3	3.3	30.4	1.5	17.1	6.2	5.7	n/a	4.6
TAM	6.8	**4.5***	**26.3***	7.8*	**41.8***	**22.3***	**51.2***	6.9	6.8	0.8	n/a
EB-GAD	**34.9**	**4.5**	15.5	**9.7**	34.2	17.2	35.6	**22.1**	**19.1**	**2.3**	**53.0**

CoLA (21.1) and ANOMALOUS (19.0): the selected score ranks most anomalies above the normal nodes but is less precise at the very top of the ranking. On Reddit (3.3% anomalies) no method is

more than 1.2 points above the base rate, EB-GAD included (4.5); we read that column as unsolved. The T-Finance AUPRC of DOMINANT, CONAD and CoLA come from reruns made for this table, not from the runs behind Table 1.

C.3 Isolation of the Spectral Precision Family

A reviewer asked whether the gains come from the OU formulation or would be obtained by any spectral scoring of the same residuals. For each dataset we freeze the operator, the template, PCA, the truncation, the residuals and the label-free selection, and change only the spectral precision: the EB-fitted GOU precision with its finite-horizon members, the GOU precision at fixed $(\rho, \kappa) = (0.5, 1)$, heat-kernel precisions, polynomial precisions, and the identity (Table 16). One configuration per dataset is shared by the five families, so the columns are compared with each other and not with Table 1, which selects over families and pools.

On Elliptic, Elliptic++ and T-Finance the GOU bank leads the best generic family by 12.6, 11.6 and 10.2 points under identical label-free selection, and on T-Finance the EB fit alone adds 23.6 points over the fixed GOU precision. The mechanism is selectability rather than the quality of the best score: on Elliptic and Elliptic++ the best members of all five families are within 1.2 points of each other (74.4 to 75.2, 73.7 to 74.9), so the generic banks contain an equally good score, but their label-free selection falls to 58 to 62. Elsewhere no family is consistently better. The GOU bank is within 1.1 points of the best generic selection on Weibo, YelpChi, BlogCatalog, ACM and DGraph; a polynomial bank is ahead on Reddit ($+3.3$, all families near chance) and Amazon ($+6.8$), and a heat-kernel bank is ahead on Facebook by 11.6 points and reaches 91.6, the level of Table 1. We therefore claim no superiority of the OU structure on those datasets; the claim is specific to the three financial graphs.

Table 16: Isolation of the spectral precision family. Operator, template, PCA, truncation, residuals and the label-free selection (largest NullKS) are fixed per dataset; only the precision changes. AUROC (%) of the selected score, with the best member of the family by AUROC in parentheses (a diagnostic, not a result). **Bold**: best selected score of the row. GOU-EB: EB-fitted GOU precision with its finite-horizon members; GOU fixed: $\rho = 0.5$, $\kappa = 1$; heat: $\exp(t\lambda/\lambda_{\max})$, 8 scales; polynomial: $(c + \lambda)^p$, 6 configurations; identity: graph-free residual energy.

Dataset	GOU-EB	GOU fixed	heat	polynomial	identity
Weibo	92.5 (92.7)	92.0 (92.7)	85.9 (92.5)	80.1 (91.4)	**92.7** (92.7)
Reddit	48.2 (55.8)	48.3 (55.9)	48.6 (57.4)	**51.5** (56.3)	48.2 (48.2)
Amazon	62.9 (67.7)	62.9 (67.8)	50.8 (67.4)	**69.7** (70.1)	62.8 (62.8)
YelpChi	**69.1** (71.3)	**69.1** (71.3)	66.1 (70.5)	68.7 (70.5)	68.7 (70.5)
BlogCatalog	78.2 (78.5)	78.2 (78.5)	78.3 (78.3)	77.6 (77.6)	**78.5** (78.5)
Facebook	80.0 (88.0)	79.8 (88.0)	**91.6** (92.8)	80.9 (87.7)	48.1 (62.0)
ACM	79.6 (81.9)	79.6 (81.3)	79.8 (81.5)	**80.4** (83.8)	79.6 (79.6)
Elliptic	**74.4** (74.4)	**74.4** (74.4)	61.8 (75.2)	58.5 (74.5)	58.5 (74.4)
Elliptic++	**73.7** (73.7)	**73.7** (73.7)	62.1 (74.9)	58.1 (73.8)	58.1 (73.7)
DGraph	62.2 (66.8)	62.2 (66.8)	63.2 (66.8)	**63.3** (66.8)	**63.3** (66.8)
T-Finance	**82.5** (82.5)	58.9 (60.8)	41.6 (60.3)	72.3 (76.6)	59.0 (60.7)

C.4 Horizon Axis: Matched Single-Score Comparison

A reviewer asked for a direct comparison of equilibrium-only scoring with finite-horizon scoring, with the fitted model and all other components fixed. Two comparisons answer it. Table 2 in the main text compares the score that the selector uses with the equilibrium score of the same configuration. Table 17 is narrower: one score type on both sides, a single finite horizon, no endpoint-tolerance grid, no fusion and no aggregation. DGraph is not included (its spectrum is computed on the labeled subgraph by a separate script).

At the granularity of single scores the answer is parity. The best finite-horizon member by AUROC is within 0.3 points of its equilibrium limit on nine of ten datasets and 1.4 below it on T-Finance; when NullKS picks the horizon, the finite score is within 0.2 points of the limit on eight datasets and below it on ACM (-1.6) and T-Finance (-1.4). No single finite-horizon score reproduces the gains of Table 2 ($+3.6$ on Facebook, $+4.7$ on ACM and T-Finance). Those gains come from what is done with the family: aggregation of p-values across horizons (Facebook, T-Finance) and scale normalization with stability-ranked fusion (ACM). The finite-horizon contribution is therefore not one fortunate horizon. It is that the fitted GOU attaches one prior, hence one plug-in null, to every

horizon, which is what makes the family comparable and aggregatable without labels; the equilibrium limit is a single member and cannot be aggregated.

Table 17: Matched single-score comparison at the configuration and EB fit of Table 5, shared by both sides. *Equilibrium*: $J^\star$ or R, whichever has the larger NullKS (R where rule 1 fixes it). *Finite*: a single hard-endpoint score of the same type ($C_{\Gamma,\infty}$ against $J^\star$, $CR_{\Gamma,\infty}$ against R) at one finite horizon $\Gamma \in \{0.02, \ldots, 50\}$, with no tolerance grid, fusion or aggregation; NullKS picks the horizon, and the best horizon by AUROC is reported as a diagnostic. AUROC (%).

Dataset (type)	Equilibrium	Finite, selected	Finite, best	Δ selected	Δ best
Weibo ($J^\star$)	95.1	95.1	95.1	$+0.0$	$+0.0$
Reddit (R)	60.6	60.6	60.6	$+0.0$	$+0.0$
Amazon ($J^\star$)	77.9	78.0	78.1	$+0.1$	$+0.2$
YelpChi ($J^\star$)	70.7	70.7	70.7	$+0.0$	$+0.0$
BlogCatalog ($J^\star$)	78.4	78.4	78.6	$+0.0$	$+0.2$
Facebook (R)	87.8	87.8	88.0	$+0.0$	$+0.2$
ACM ($J^\star$)	81.3	79.7	81.6	-1.6	$+0.3$
Elliptic (R)	74.4	74.4	74.4	$+0.0$	$+0.0$
Elliptic++ (R)	72.5	72.5	72.5	$+0.0$	$+0.0$
T-Finance (R)	80.1	78.7	78.7	-1.4	-1.4

C.5 Selector Studies

Cutoff sensitivity. The ordered rules of Appendix B.4 contain 21 cutoffs. We multiply each one by 0.75 and 1.25, and by 0.5 and 1.5, one at a time, and re-evaluate the rules on the 11 datasets with the diagnostics of the runs of Table 1. At $\pm 25\%$ the selected family and branch are unchanged in 459 of 462 evaluations (99.4%), at $\pm 50\%$ in 450 of 462 (97.4%). Scaling all cutoffs jointly by 0.75 changes the selection on 2 of 11 datasets (Reddit and Facebook); scaling them by 1.25 changes it on one (Weibo). Table 18 lists the fifteen changes. Every one of them lowers the AUROC, by 2.2 to 44.0 points. Eleven select an equilibrium anchor instead of the score of Table 1, and the cost is largest where NullKS then selects $J^\star$ at the anchor bandwidth: on Facebook (six changes, 47.4 instead of 91.4), whose routing to profile aggregation depends on four graph-statistic cutoffs and on the upper end of the mixture-mass window ($\pi = 0.81$), and on Elliptic++ (two changes, 30.3 instead of 72.5), where rule 1 otherwise fixes R. Two reroute Weibo from the anchor to profile aggregation (-2.6) and two reroute Reddit from the anchor to the control-energy family (-15.7). The selections are therefore stable under sizable perturbations of single cutoffs, but where a selection changes the cost can be large, and the choice of the score type by NullKS is the fragile step.

Table 18: Cutoff sensitivity of the ordered rules. Each of the 21 cutoffs is multiplied by 0.5, 0.75, 1.25 and 1.5, one at a time, and the rules are re-evaluated on the 11 datasets (231 evaluations per factor). The table lists every evaluation in which the selected family or branch changes, with the AUROC (%) before and after (original node order).

Cutoff	Factor	Dataset	Selection	before	after	change
$d_e < 1$ (rule 1)	$\times 0.5$	Elliptic++	anchor (R) $\to$ anchor ($J^\star$)	72.5	30.3	-42.2
$d_e < 1$ (rule 1)	$\times 0.5$	DGraph	anchor (R) $\to$ anchor ($J^\star$)	66.8	60.5	-6.3
$D_x \leq 256$ (rule 1)	$\times 0.5$	Elliptic++	anchor (R) $\to$ anchor ($J^\star$)	72.5	30.3	-42.2
$D_x > 1000$	$\times 0.5$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0
$d_e \geq 20$	$\times 1.5$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0
$h < 0.25$	$\times 0.5$	ACM	ratio (aff.) $\to$ anchor ($J^\star$)	86.0	75.6	-10.4
$h < 0.25$	$\times 1.5$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0
$d_e \geq 10$	$\times 0.5$	Reddit	anchor (R) $\to$ energy (low)	60.6	44.9	-15.7
$d_e \geq 10$	$\times 0.75$	Reddit	anchor (R) $\to$ energy (low)	60.6	44.9	-15.7
guard 0.60	$\times 1.25$	Weibo	anchor ($J^\star$) $\to$ profile (traj.)	95.1	92.5	-2.6
guard 0.60	$\times 1.5$	Weibo	anchor ($J^\star$) $\to$ profile (traj.)	95.1	92.5	-2.6
$h \geq 0.50$	$\times 0.5$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0
$h \geq 0.50$	$\times 1.5$	Amazon	energy (low) $\to$ anchor ($J^\star$)	78.0	75.8	-2.2
$\pi \leq 0.95$	$\times 0.5$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0
$\pi \leq 0.95$	$\times 0.75$	Facebook	profile (entropy) $\to$ anchor ($J^\star$)	91.4	47.4	-44.0

Ingredients of $A(r)$. Table 19 separates the two ingredients of the within-family rule on the eight datasets routed to an equilibrium or profile bank. Quality alone (NullKS) picks the score of the rule on four of them. It fails on Facebook (17.6), where the two-groups posterior has the largest NullKS but ranks anomalies last, and on Elliptic++ (30.3), where rule 1 fixes R instead. Stability alone never picks the score of the rule on these datasets and ranges from 25.5 to 85.4. The combined tail-stability

variant picks the score of the rule on Facebook (91.4) and YelpChi and is worse on the ratio datasets, which is why the rule weights stability by family. The rule is within 2.0 points of the best member of its bank on seven of the eight datasets; on Reddit it picks R (60.6) while the best member of the bank has 62.9.

Table 19: Ingredients of the within-family rule $A(r)$ on the datasets routed to an equilibrium or profile bank (original node order). AUROC (%) of the member picked by quality only (NullKS), by stability only, by the two combined variants (tail-stability, correlation-stability), by the rule used in Table 1, and the best member of the bank by AUROC (a diagnostic). The equilibrium bank contains $J^\star$, R and their rank aggregates; the rule restricts it to $J^\star$ and R.

Dataset	Bank	NullKS	stability	tail-stab.	corr-stab.	rule	best
Weibo	equilibrium	95.1	74.7	90.8	89.6	95.1	95.1
Reddit	equilibrium	60.6	62.9	62.8	62.9	60.6	62.9
YelpChi	equilibrium	68.7	67.9	70.7	70.7	70.7	70.7
Facebook	two-groups	17.6	25.5	91.4	69.4	91.4	91.4
Elliptic	equilibrium	74.4	67.2	58.4	67.4	74.4	74.4
Elliptic++	equilibrium	30.3	52.3	30.3	54.5	72.5	72.5
DGraph	equilibrium	63.3	64.1	64.1	64.1	66.8	66.8
T-Finance	path bands	84.8	85.4	61.3	83.4	84.8	86.8

Supervised bound. A reviewer asked whether weights for the scores could be learned. Table 20 bounds what any label-driven linear combination of our scores could gain: a cross-validated logistic regression over each dataset's pool, which uses the labels, against the label-free pick. The gap is below one point on Weibo, Amazon, BlogCatalog and Facebook, 3.1 to 4.3 points on DGraph, YelpChi and Reddit, and 8.2 to 17.1 points on T-Finance, ACM, Elliptic and Elliptic++. The pools therefore contain more information than the label-free selector extracts, most of all on the financial graphs and ACM; exploiting it requires labels and leaves the training-free setting. On Reddit the best single member of the pool reaches 62.0 and the selector picks 60.6.

Table 20: Supervised bound on combining the scores. Every score of a dataset's pool is rank-transformed and a logistic regression is fitted on the labels with 5-fold cross-validation; the out-of-fold AUROC (%) bounds what a label-driven linear combination of our scores could reach. *Selector*: the label-free pick of Table 1 (original node order). *Best single*: best member of the pool by AUROC.

Dataset	Scores	Selector	Best single	Stacked (CV)	Gap
Weibo	224	95.1	95.1	95.2	+0.1
Reddit	224	60.6	62.0	64.9	+4.3
Amazon	887	78.0	78.1	78.3	+0.3
YelpChi	158	70.7	73.0	74.9	+4.2
BlogCatalog	4391	78.6	78.7	79.5	+0.9
Facebook	224	91.4	91.4	92.3	+0.9
ACM	741	86.0	86.1	96.4	+10.4
Elliptic	166	74.4	74.4	88.3	+13.9
Elliptic++	168	72.5	73.6	89.6	+17.1
DGraph	118	66.8	66.8	69.9	+3.1
T-Finance	232	84.8	86.8	93.0	+8.2

C.6 Robustness of the Fit and Scope of the Gaussian Model

Contamination. A reviewer asked how the fit behaves when anomalies are not rare. We inject contextual anomalies (feature swaps with random nodes) on top of the real ones, refit EB on the contaminated graph and repeat the label-free selection (Table 21). On Weibo the fitted prior moves from $(\rho, \kappa) = (0.70, 3)$ to $(0.74, 3)$ between 4.1% and 24.1% total contamination, and the AUROC on the original anomalies falls from 92.5 to 91.4. The injected swaps themselves are barely separable on Weibo (56.1 to 57.2), so there they stress the fit and not the detector. On Reddit the original anomalies stay at chance at every level; the injected ones are ranked at 78.3 to 80.2, and the fit stays at $\kappa = 3$ up to 13.3% contamination and moves to $(0.43, 5)$ at 23.3%. The likelihood is a sum over all nodes, so a contamination ϵ perturbs the estimating equations of (ρ, κ) by a term of order ϵ times the anomalous residual energy: the prior degrades gradually. The benchmarks themselves span anomaly rates from 1.3% (DGraph) to 9.8% (Elliptic, evaluated nodes). At the anchor bandwidth of Table 1 ($\gamma_a = 0.5$ on Weibo, 2 on Reddit) the fit does not move, and the original anomalies stay at 95.0 to 95.2 on Weibo and 60.1 to 60.6 on Reddit at every contamination level.

Table 21: Contamination. Contextual anomalies (feature swaps with random nodes) are injected on top of the real anomalies, EB is refitted on the contaminated graph and the score is selected without labels (largest NullKS over the finite-horizon bank). AUROC (%) on the original anomalies and on the injected ones (against untouched normal nodes).

Dataset	Injected	Total contamination	Fitted (ρ, κ)	Original anomalies	Injected anomalies
Weibo	$+0\%$	4.1%	(0.697, 3)	92.5	–
Weibo	$+5\%$	9.1%	(0.710, 3)	92.0	57.1
Weibo	$+10\%$	14.1%	(0.720, 3)	92.1	57.2
Weibo	$+20\%$	24.1%	(0.737, 3)	91.4	56.1
Reddit	$+0\%$	3.3%	(0.910, 3)	48.3	–
Reddit	$+5\%$	8.3%	(0.997, 3)	48.3	79.9
Reddit	$+10\%$	13.3%	(1.000, 3)	49.2	80.2
Reddit	$+20\%$	23.3%	(0.431, 5)	50.2	78.3

Misspecification. Table 22 is a controlled stress test of the Gaussian residual model on stochastic block model graphs, with the same pipeline in every column and seed-matched conditions. Heavy tails cost at most 0.7 points against the Gaussian control. A two-component normal population costs 9.0 to 19.3 points: the quadratic score measures the distance to one template, and a second mode of normal nodes is far from it. Binarized features fall outside the model: the selected score is at 56 to 59 (best member of the bank: 58 to 61), and at 59 to 61 (best member: 61 to 64) when the anomalies are bit flips. The failure is that of a Gaussian residual on quantized data; a categorical likelihood for the residual is the natural replacement and is future work. One-hot features in the benchmarks pass through PCA first, which gives an approximately continuous input. In this test the fitted graph trust is small ($\rho = 0.05$) in every condition, because the noise added to the smooth signal is white; the test probes the residual distribution, not the graph prior. At the anchor bandwidth the first three rows change by at most 0.1 and the binarized rows are at 59.6 to 67.2.

Table 22: Misspecification stress test on stochastic block model graphs ($n = 3000$, four blocks, edge probabilities 0.02 within and 0.002 between blocks, $D = 16$, 5% injected attribute anomalies, three seeds). The normal signal is a low-pass graph signal plus noise whose distribution is varied; the pipeline, grids and label-free selection are identical in every column. AUROC (%) of the selected score and of the best member of the bank (a diagnostic), and the fitted graph trust, per seed.

Residual distribution	Selected	Best member	Fitted ρ
Gaussian (control)	96.3 / 95.7 / 95.5	98.8 / 98.1 / 97.6	0.05 / 0.05 / 0.05
Student-t, 3 d.o.f.	95.6 / 95.7 / 95.3	97.4 / 97.5 / 96.8	0.05 / 0.05 / 0.05
two-component mixture	80.2 / 86.7 / 76.2	87.0 / 91.6 / 82.7	0.05 / 0.05 / 0.05
binarized	59.3 / 56.1 / 56.4	60.8 / 58.1 / 61.1	0.05 / 0.05 / 0.05
binarized, bit-flip anomalies	59.1 / 60.5 / 59.4	63.4 / 60.9 / 63.7	0.05 / 0.05 / 0.05

C.7 Graph Structure, Edge Uncertainty and Neighboring Anomalies

A worked example. A reviewer asked how the network structure enters the score and how uncertainty in the edges is reflected. We use Zachary's karate club with the connectivity profiles (rows of the adjacency matrix) as features and run the pipeline of the paper (EB fit, equilibrium score $J^\star$). The three most anomalous profiles are those of the two faction leaders (individuals 34 and 1) and of individual 33. Zachary's published matrix is inconsistent about the edge between individuals 23 and 34; we score the graph with and without it. The three largest changes of the score are at individuals 33, 34 and 23, and 98.8% of the total absolute change falls on the two endpoints and their neighbors (18 of the 34 nodes). The ranking as a whole is stable (rank correlation 0.989); eight nodes outside that neighborhood change rank, by at most five positions. Structure enters the score through two channels: the smoothed template, which the edge changes directly, and the coupling in the precision, whose weight ρ is what EB fits. Because these features are the connectivity profiles themselves, EB assigns the graph's information to the template ($\rho = 0.008$). These numbers are for $\gamma^\star = 1$, the likelihood-selected bandwidth, at which the template has unit gain on the constant mode. At the bandwidth nearest to γ_0 (0.7) the template amplifies low frequencies and the change is less local: 74.7% of it falls on the same 18 nodes, and the rank correlation is 0.927.

Neighboring anomalies. A reviewer asked whether connected anomalies can be detected, since each could make its neighbor look normal. On Reddit we inject 5% feature-swap anomalies either scattered or as connected groups of at most 10 neighbors, refit EB on each graph, and select the

score without labels (Table 23). Connected groups are easier to detect, 92.0 to 93.6 against 79.9 to 83.0: a connected anomalous group shifts the smoothed template away from all of its members at once. A pairwise readout also works (Table 24): scoring an edge by the smaller of its two endpoint energies separates edges between two injected nodes from edges between two normal nodes at 94.4 to 96.7; the standardized cross-term of the quadratic form alone gives 77.5 to 84.0. We did not study large contiguous anomalous regions, which can become their own context. These results are for the EB-fitted configuration ($\gamma^\star = 0.7$), whose residual is contextual: most of the feature energy is removed and what remains is the deviation of a node from its smoothed neighborhood. At the anchor bandwidth of Reddit ($\gamma_a = 2$) the residual is close to the features themselves, and a feature swap between two nodes is harder to see: the label-free choice gives 55.0 to 57.4 on connected groups (best member of the bank: 89.7 to 91.4) and 52.1 to 77.6 on scattered swaps, the smaller endpoint energy separates edges at 87.5 to 91.1, and the standardized cross-term is inverted (8.4 to 9.6). The two bandwidths therefore see different anomalies: the labeled anomalies of Reddit are found at γ_a and not at $\gamma^\star$, injected swaps at $\gamma^\star$ and not at γ_a.

Table 23: Neighboring anomalies on Reddit. 5% of the nodes receive a feature swap, either scattered over the graph or as connected groups of at most 10 neighbors; EB is refitted on each contaminated graph and the score is selected without labels. AUROC (%) of the injected nodes against untouched normal nodes, per seed.

Placement	Selected, seeds 0–4	Range	Best member, range
scattered	79.9 / 81.3 / 82.7 / 80.5 / 83.0	79.9–83.0	79.9–83.0
connected groups of at most 10	92.8 / 92.2 / 93.6 / 92.9 / 92.0	92.0–93.6	92.0–93.6

Table 24: Pairwise readout on the connected-group runs of Table 23. Every undirected edge is scored and edges that join two injected nodes are ranked against edges that join two untouched normal nodes. AUROC (%) per seed.

Edge score	Seeds 0–4	Range
smaller endpoint energy	95.9 / 96.0 / 94.4 / 96.7 / 94.4	94.4–96.7
standardized cross-term	83.0 / 84.0 / 77.5 / 83.4 / 80.2	77.5–84.0

C.8 Template Variants on BlogCatalog

A reviewer asked whether the low-pass filter limits the method to homophilic graphs and explains the BlogCatalog result. BlogCatalog has the lowest feature homophily of the benchmarks ($h = 0.009$) and is routed to the feature-affinity template, which reweights features by node reliability before smoothing. We replaced that template by a high-pass and by a mixed (band-pass) affinity variant and ran the same pipeline. The equilibrium anchor chosen by NullKS gives 77.4 (affinity), 75.3 (high-pass) and 74.6 (mixed) AUROC, and the best member of each bank by AUROC is 78.6, 76.0 and 74.8. Richer filters do not help: the affinity template of the paper is the best of the three, and none approaches TAM (82.5), which optimizes a local-affinity objective that matches the injected anomalies of this graph.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the results of Table 1, including the datasets on which the method trails; Section 4 and Appendix C.1 state which entries depend on the anchor bandwidth rule, adopted after a first regeneration had been evaluated, and give the results without it.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 has a scope-and-limitations paragraph, and Appendix C reports the studies behind it.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: all the theoretical statements in the paper are stated with conditions and proved in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix B.4 states the procedure exactly as it is run (grids, ordered rules, cutoffs), and Appendix C.1 the protocol and the relabeling gate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, the protocol file and the scripts that regenerate every EB-GAD number are available at `https://github.com/heraclixus/EBGAD`; the datasets are public and their sources are listed there.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The implementation details are covered in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Baselines report mean and standard deviation over seeds. EB-GAD is deterministic given the eigenbasis; for truncated spectra we report mean and standard deviation over five random node relabelings (Table 11).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational resources are described in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: our work can help with ML system deployment safety and there are no ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: EB-GAD targets fraud detection, cybersecurity, and scientific networks (§4 evaluates on financial-fraud benchmarks T-Finance, Elliptic, Elliptic++, and DGraph with up to 3.7M nodes). Positive impacts include improved detection of illicit transactions and abusive accounts in production systems; the training-free, deterministic pipeline also reduces compute and energy footprint relative to GNN-based detectors. Negative impacts of any anomaly detector deployed in financial or social settings include downstream harms from false positives (e.g., unwarranted account suspensions, financial exclusion) and disparate impact across user subgroups; mitigations include human review of flagged cases, calibrated thresholds, and audit of subgroup error rates before deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no pretrained models and no new or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Dataset and baseline sources are cited in Appendix B and the references; release materials will include license notes where available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is documented (installation, data sources, protocol, and the script behind each table).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: no crowdsourcing is involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not a component of the method. An LLM assistant was used to polish the writing and to help orchestrate experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.