Rank Is Not Capacity: Spectral Occupancy for Latent Graph Models

Nikolaos Nakis[1,*] **Panagiotis Promponas**[2,*] **Konstantinos Tsirkas**[3] **Katerina Mamali**[4]
Eftychia Makri[2] **Leandros Tassiulas**[2] **Nicholas A. Christakis**[1]

[1]Human Nature Lab, Yale University
[2]Department of Electrical and Computer Engineering, Yale University
[3]Department of Statistics and Data Science, Yale University
[4]Department of Computer Science, Yale University
New Haven, CT 06511
nicolaos.nakis@gmail.com, panagiotis.promponas@yale.edu

Abstract

Graph representation learning has become a standard approach for analyzing networked data, with latent embeddings widely used for link prediction, community detection, and related tasks. Yet a basic design choice, the latent dimension, is still treated as a brittle hyperparameter, fixed before training and tuned by held-out performance. Learned factors are also identifiable only up to rotation and rescaling, so the nominal rank rarely coincides with the quantity that governs model behavior. We propose Spectral Prefix Extraction and Capacity-Targeted Representation Analysis (Spectra), which replaces rank as the unit of analysis with the spectrum of a learned positive semidefinite kernel, trace-normalized so that spectra are comparable across fits. The normalized eigenvalues form a distribution on the simplex, and their Shannon effective rank acts both as a summary of learned capacity and as a controllable training-time coordinate: a single scalar shapes this realized dimension during training, and bisection targets any desired value within the rank cap. To theoretically support that, we show local regularity and monotonicity of the realized-dimension profile. Across collaboration, social, biological, and infrastructure networks, Spectra traces performance–capacity frontiers that make the trade-off between predictive accuracy and realized dimension visible. It performs competitively with strong link-prediction baselines, yields aligned lower-capacity views of the same fitted model through spectral prefixes, and provides a principled handle on capacity in the overparameterized regime. Capacity thus becomes a property of the fitted model rather than a hyperparameter of the training

1 Introduction

A central design choice in graph representation learning is the dimension of the latent space. Despite its influence on model behavior, this quantity is usually treated as a fixed hyperparameter, selected before training and tuned by held-out performance. Modern graph representation methods, including random-walk embeddings [Grover and Leskovec, 2016, Ahmed et al., 2018], matrix-factorization approaches [Qiu et al., 2018, Cao et al., 2015], mixed-membership models [Airoldi et al., 2007, Wang et al., 2017], latent space and inner-product models [Hoff et al., 2002, Hoff, 2007, Athreya et al., 2018, Rubin-Delanchy et al., 2022, Sussman et al., 2013], and simplex-volume extensions such as HM-LDM [Nakis et al., 2022, 2023a], all rely, explicitly or implicitly, on such a choice. Across these families, a basic question persists: how many latent dimensions do the data actually require?

*Shared first authorship

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

Existing selection procedures, including profile-likelihood elbows [Zhu and Ghodsi, 2006], Bayesian latent-dimension selection [Passino and Heard, 2020], network cross-validation [Chen and Lei, 2018, Li et al., 2020], structural criteria [Gu et al., 2021], input-graph entropy methods [Luo et al., 2021], and spectral parallel analysis such as NETFLIPPA [Hong and Cape, 2025], largely select a single dimension before or after fitting. They do not expose *representational capacity*, i.e., how many latent modes the fitted model effectively uses, as a quantity that can be measured, controlled, or audited in the fitted model itself. Also, learned latent factors are identifiable only up to rotation and global rescaling [Hoff et al., 2002, Airoldi et al., 2007, Athreya et al., 2018], and dimension misspecification has measurable statistical costs [Taing and Levin, 2026]. The operative quantity is therefore not only the declared rank, but how representational mass is distributed across latent modes.

We propose the Spectral Prefix Extraction and Capacity-Targeted Representation Analysis (SPECTRA), to replace rank as the unit of analysis with the spectrum of a learned positive semidefinite (PSD) kernel. The kernel is invariant under the gauge symmetries of the factorization, trace normalization fixes total spectral mass, and the normalized eigenvalues form a probability distribution on latent modes, which we call the *spectral occupancy distribution*. We summarize this distribution by the *effective spectral dimension* d_{spec}, the exponential of its Shannon entropy [Roy and Vetterli, 2007, Friedman and Dieng, 2022]. Spectral entropy has been used as a diagnostic in language models [Wei et al., 2024, Jha and Reagen, 2025], training dynamics [Yang et al., 2024], and adaptive-rank compression [Cherukuri and Lala, 2025]. Here, it becomes a controllable, end-to-end training-time capacity coordinate for latent graph models. This positioning connects to two views of overparameterized models: effective complexity can govern generalization even when parameter counts are large [Belkin et al., 2019, Bartlett et al., 2020], and gradient descent on factorized matrix models is implicitly biased toward low-rank solutions [Gunasekar et al., 2017, Arora et al., 2019]. Our approach, makes this bias explicit and parameterizable: an entropy weight η shapes the learned spectrum, bisection targets a desired d_{spec}, and the rank cap becomes a parameterization ceiling. Trace normalization separates scale from rank, avoiding the scale coupling of nuclear-norm penalties [Srebro et al., 2004], while spectral prefixes provide optimal low-rank PSD summaries of the fitted kernel [Eckart and Young, 1936].

Our contributions are: (i) **Spectral occupancy as realized capacity.** We replace nominal rank with the trace-normalized spectrum of the learned PSD kernel, and use its Shannon effective rank, d_{spec}, as a smooth measure of realized latent dimension. (ii) **Training-time capacity control.** We introduce an entropy-regularized objective in which a scalar η shapes the realized spectral dimension, and use bisection to target a prescribed d_{spec} within tolerance. (iii) **Spectral identifiability and summaries.** We establish local $\mathcal{C}^1$ behavior of the realized-dimension profile under regularity conditions, and use classical spectral identifiability under a simple-spectrum condition and optimal spectral approximation results to justify canonical low-rank PSD summaries of the learned kernel. (iv) **Empirical capacity frontiers.** Across eight benchmark networks and three rank caps, AUC aligns by achieved d_{spec} where rank caps overlap. The resulting frontiers distinguish saturated datasets, whose best operating points lie below the cap, from rank-cap-binding datasets, whose best operating points increase with r. (v) **Single-fit prefix families.** A fitted kernel yields a nested family of aligned spectral prefixes, providing lower-capacity views of the same representation without retraining. (vi) **Spectrally controlled overparameterization.** At matched effective capacity, η-targeted overparameterized fits improve test log-likelihood and train–test gap over rank-cap-only baselines in paired experiments. Code available at: `https://github.com/Nicknakis/Spectra`.

2 Related Work

Latent-distance and inner-product models [Hoff et al., 2002, Hoff, 2007] alongside random dot product graph (RDPG) frameworks [Athreya et al., 2018, Sussman et al., 2013] treat spectra of latent positions or adjacency matrices as canonical inferential objects, often with consistency guarantees, but do not provide explicit training-time capacity control. Gradient-based RDPG inference [Fiori et al., 2023] similarly infers realized rank from the optimizer rather than controlling it, while simplex-constrained and latent-distance extensions [Nakis et al., 2022, 2023b] add geometric interpretability but retain a fixed nominal dimension. SPECTRA works in this lineage, but uses an explicit kernel parameterization to expose capacity as a continuous training-time coordinate. Profile-likelihood elbow rules [Zhu and Ghodsi, 2006], Bayesian latent-dimension selection [Passino and Heard, 2020], network cross-validation [Chen and Lei, 2018, Li et al., 2020], structural information criteria [Gu et al., 2021, Luo et al., 2021], spectral parallel analysis [Hong and Cape, 2025], and logarithmic

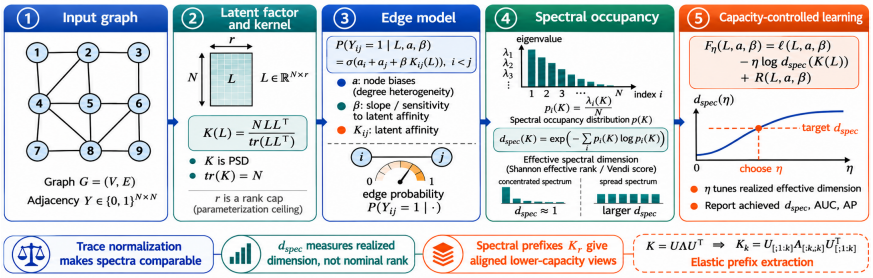


Figure 1: Overview of SPECTRA. A rank-capped factor L induces a trace-normalized PSD kernel $K(L)$ used in the edge model. The normalized spectrum of $K(L)$ defines spectral occupancy.

search under metric latent-distance models [Nakis et al., 2025] select a dimension from spectral, posterior, or held-out diagnostics in a one-shot procedure. The operative coordinate is fixed before or after training. SPECTRA instead shapes realized capacity during training and measures it on the learned kernel spectrum. Factorized PSD models exhibit gradient-descent bias toward low rank [Gunasekar et al., 2017, Arora et al., 2019], within a broader literature in which effective complexity governs overparameterized generalization [Belkin et al., 2019, Bartlett et al., 2020]. Nuclear-norm relaxations [Candes and Recht, 2012, Recht et al., 2010, Srebro et al., 2004] provide explicit but scale-coupled regularization, and stable-rank normalization [Sanyal et al., 2020] controls a Frobenius-to-spectral ratio. By trace-normalizing the learned PSD kernel, SPECTRA separates scale from rank, exposes capacity on the simplex of normalized eigenvalues, and yields Eckart–Young–Mirsky-optimal low-rank PSD summaries [Eckart and Young, 1936]. Shannon effective rank has appeared as a signal-processing spectral-entropy measure [Roy and Vetterli, 2007] and as the Vendi score for diversity [Friedman and Dieng, 2022]; spectral entropy also diagnoses capacity in language models [Wei et al., 2024, Jha and Reagen, 2025], characterizes neural-network training dynamics [Yang et al., 2024], and guides adaptive-rank compression [Cherukuri and Lala, 2025]. Related alternatives include the participation ratio, intrinsic dimension [Li et al., 2018], stable rank [Ipsen and Saibaba, 2025], and entropy functionals in determinantal point processes [Kulesza and Taskar, 2012]. In these settings the spectrum is typically read post hoc; SPECTRA uses Shannon effective rank as a controllable training-time coordinate.

3 Proposed Method

Preliminaries. Let $\mathcal{G} = (V, E)$ be a simple undirected graph with $N = |V|$ nodes and adjacency matrix $\boldsymbol{Y} \in \{0,1\}^{N\times N}$, where $\boldsymbol{Y}_{ij} = \boldsymbol{Y}_{ji}$ and $\boldsymbol{Y}_{ii} = 0$. Let $\mathcal{E}^+$ denote the observed training edges and $\widetilde{\mathcal{E}}^-$ the sampled training non-edges. Our goal is to learn a latent graph representation whose realized dimension can be measured from the fitted model, rather than fixed in advance by a chosen factorization rank. We represent latent affinities between nodes by a positive semidefinite (PSD) kernel $K \in \mathbb{R}^{N\times N}$. To compare the spectra of learned kernels across runs, rank caps, and regularization strengths, we fix the total spectral mass:

$$K \succeq 0, \qquad \operatorname{tr}(K) = N. \tag{1}$$

This trace normalization removes arbitrary global rescaling of the kernel and places all learned spectra on the same scale. As a result, differences in the eigenvalue distribution of K reflect how the representation allocates its fixed spectral budget across latent modes, rather than differences in overall magnitude. We parameterize the kernel by a rank-capped factor $L \in \mathbb{R}^{N\times r}$:

$$K(L) = N\frac{LL^\top}{\operatorname{tr}(LL^\top)}, \qquad L \neq 0. \tag{2}$$

The integer r is therefore a parameterization ceiling: it limits the maximum rank available to the optimizer, but it does not define the dimension actually used by the learned representation. The realized dimension will be measured from the normalized spectrum of $K(L)$.

Spectral Prefix Extraction and Capacity-Targeted Representation Analysis (SPECTRA). Given the trace-normalized kernel $K(L)$, node-specific offsets $\boldsymbol{a} \in \mathbb{R}^N$, and a slope parameter $\beta > 0$, we

model each undirected edge as

$$\mathbb{P}(\boldsymbol{Y}_{ij} = 1 \mid L, \boldsymbol{a}, \beta) = \phi(a_i + a_j + \beta K_{ij}(L)), \qquad i < j, \tag{3}$$

where $\phi(t) = (1 + \exp(-t))^{-1}$. The offsets $\boldsymbol{a}$ capture degree heterogeneity: nodes with larger offsets have higher baseline propensity to form edges. The kernel entry $K_{ij}(L)$ captures latent affinity between nodes i and j. The slope β controls the strength of this latent-affinity signal on the log-odds scale: when β is small, edge probabilities are driven mostly by the node offsets, whereas larger β makes differences in $K_{ij}(L)$ more influential. Because the kernel trace is fixed, β cannot be absorbed into an arbitrary rescaling of K and it has a well-defined role as the global sensitivity to latent affinity.

Spectral occupancy and effective spectral dimension. Let $\lambda_1(K) \geq \lambda_2(K) \geq \cdots \geq \lambda_N(K) \geq 0$ be the eigenvalues of the trace-normalized kernel K. Since $\mathrm{tr}(K) = N$, the normalized spectrum

$$p_i(K) = \frac{\lambda_i(K)}{N} \tag{4}$$

defines a probability distribution over spectral modes. We call $p(K) = (p_1(K), \ldots, p_N(K))$ the *spectral occupancy distribution*: it records how the fixed spectral budget of K is allocated across latent directions.

Definition 1 (Effective spectral dimension)**.** The *effective spectral dimension* of K is

$$d_{\mathrm{spec}}(K) := \exp\left(- \sum_{i : p_i(K) > 0} p_i(K) \log p_i(K) \right). \tag{5}$$

Equivalently, $d_{\mathrm{spec}}(K) = \exp(H(p(K)))$, where H is the Shannon entropy. This quantity is the Shannon effective rank of the trace-normalized kernel.

The value $d_{\mathrm{spec}}(K)$ can be interpreted as the number of equal-mass spectral modes that would have the same Shannon entropy as the observed occupancy distribution $p(K)$. It equals one when the entire trace budget is concentrated in a single mode. It equals s when the trace budget is spread uniformly over s nonzero modes, in which case $p_1(K) = \cdots = p_s(K) = 1/s$ and $p_i(K) = 0$ for $i > s$. Between these extremes, $d_{\mathrm{spec}}(K)$ varies continuously with the spectrum and provides a smooth effective-dimension coordinate for the learned representation. In contrast, the hard rank $\mathrm{rank}(K) = |\{i : p_i(K) > 0\}|$ only counts the support size of the spectrum and is sensitive to modes with arbitrarily small spectral mass.

Capacity-control objective. The effective spectral dimension $d_{\mathrm{spec}}(K)$ is bounded above by the hard rank of K, with equality attained when the spectral mass is uniformly distributed. This maximum-entropy reference case motivates controlling realized dimension through the entropy of the spectral occupancy distribution. Higher entropy corresponds to a more diffuse allocation of the trace budget across latent modes, while lower entropy corresponds to a more concentrated allocation.

Let $\ell(L, \boldsymbol{a}, \beta)$ denote the sampled link-prediction loss induced by the latent-kernel edge model, and let $R(L, \boldsymbol{a}, \beta)$ denote quadratic regularization. Since $\log d_{\mathrm{spec}}(K(L)) = H(p(K(L)))$, we estimate the model by minimizing the entropy-regularized objective

$$F_\eta(L, \boldsymbol{a}, \beta) = \ell(L, \boldsymbol{a}, \beta) - \eta \log d_{\mathrm{spec}}(K(L)) + R_\lambda(\boldsymbol{a}, \beta), \tag{6}$$

where the class-balanced sampled negative log-likelihood is $\ell(L, \boldsymbol{a}, \beta) = -\frac{1}{|\mathcal{E}^+|} \sum_{\mathcal{E}^+} \log \mathbb{P}_{ij} - \frac{1}{|\widetilde{\mathcal{E}}^-|} \sum_{\widetilde{\mathcal{E}}^-} \log(1 - \mathbb{P}_{ij})$ and $R_\lambda(\boldsymbol{a}, \beta) = \frac{\lambda}{2} \left(\beta^2 + \frac{1}{N} \sum_{i=1}^N a_i^2 \right)$, and $\lambda = 10^{-4}$. The node-offset penalty is averaged over N so that its scale does not grow with graph size. Because $K(L)$ is trace-normalized, this regularizer penalizes the global slope and node offsets without directly penalizing the shape of the normalized kernel spectrum. The scalar η serves as a Lagrange-style control parameter governing the tradeoff between predictive fit and spectral spread. Larger positive values of η place greater reward on spectral entropy and favor representations with larger effective spectral dimension, while smaller or negative values favor more concentrated spectra and lower effective spectral dimension.

For each value of η, optimization yields a fitted factor $L^\star(\eta)$ and a corresponding trace-normalized kernel $K^\star(\eta) = K(L^\star(\eta))$. This induces the realized-dimension profile

$$d_{\mathrm{spec}}(\eta) = d_{\mathrm{spec}}(K^\star(\eta)). \tag{7}$$

We report fitted models in the interpretable coordinate $d_{\mathrm{spec}}(K^\star(\eta))$, which gives the effective number of spectral modes realized by the learned representation.

Interpretation. Two kernels with the same hard rank can have very different d_{spec}, and conversely. The quantity summarizes the shape of the occupancy distribution rather than only its support size. Among spectral summaries, the participation ratio $\mathrm{PR}(K) = 1/\sum_i p_i(K)^2$ is the exponential of the second-order Rényi entropy and emphasizes high-mass modes through a second-order moment, while thresholded ranks $k_\tau(K) = |\{i : \lambda_i(K) \geq \tau\lambda_1(K)\}|$ depend on a discontinuous cutoff. The Shannon effective rank in d_{spec} is the unique choice that is continuous in the occupancy distribution, mass-proportional, and recovers the hard rank in the equimass limit, making it a smooth coordinate for controlling and reporting realized dimension.

Identifiability and optimality of spectral summaries. We record two properties of the spectrum of the trace-normalized kernel. The first shows that spectral prefixes are optimal low-rank summaries of a learned kernel and follows from Eckart-Young-Mirsky Eckart and Young [1936]. The second shows that, under a simple spectrum, the spectral modes are identified up to sign and permutation. We present the proofs in Appendices C and D for completeness.

Proposition 2 (Optimal spectral prefix)**.** *Let $K \succeq 0$ be a symmetric matrix with rank r and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and eigenvectors $u_1, \ldots, u_r$ such that $K = \sum_{j=1}^r \lambda_j u_j u_j^\top$. For any positive integer $k \leq r$, define $K_k := \sum_{j=1}^k \lambda_j u_j u_j^\top$. Then K_k is a best rank-k positive semidefinite approximation of K in Frobenius norm, i.e., $K_k \in \arg\min_{M \succeq 0, \mathrm{rank}(M) \leq k} \|K - M\|_F^2$. Moreover, if $\lambda_k > \lambda_{k+1}$, the optimizer is unique.*

Proposition 3 (Identifiability of simple spectral modes)**.** *Let $K \succeq 0$ be a symmetric matrix of rank r with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ and eigenvectors $u_1, \ldots, u_r$ such that $K = \sum_{j=1}^r \lambda_j u_j u_j^\top$. Then, if*

$$K = U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_r) U^\top = U' \,\mathrm{diag}(\lambda'_1, \ldots, \lambda'_r) {U'}^\top,$$

with $U^\top U = {U'}^\top U' = I_r$, there exist a permutation matrix Π and a diagonal matrix S, with diagonal entries in $\{\pm 1\}$, such that $U' = U\Pi S,\ \lambda' = \Pi^\top \lambda$.

Implications for representation summaries. Proposition 2 shows that the leading spectral modes provide canonical low-rank summaries of the fitted kernel: truncating to the first k eigenmodes is the optimal rank-k PSD approximation in Frobenius norm. Thus, spectral prefixes are not arbitrary coordinate projections of the factor L, but intrinsic summaries of the kernel itself. Proposition 3 further shows that, when the active spectrum is simple, these spectral modes are stable representation-level objects, identifiable up to sign and permutation. Together, these results separate the parameterization from the representation: the rank cap r controls the largest available kernel rank, while $d_{\mathrm{spec}}(K)$ measures the effective spectral dimension realized by the learned kernel.

Elastic prefix extraction. A fitted kernel also induces a nested family of aligned lower-dimensional representations, in the spirit of *Matryoshka Representation Learning* [Kusupati et al., 2024], but obtained here without any multi-scale training objective. If $K = U\Lambda U^\top$, define

$$K_k = U_{1:k} \Lambda_{1:k} U_{1:k}^\top, \qquad k = 1, \ldots, r. \tag{8}$$

The family $\{K_k\}_{k=1}^r$ is strictly nested, since $\mathrm{range}(K_k) \subset \mathrm{range}(K_{k+1})$ because K_{k+1} extends K_k by one additional eigenpair without modifying the previous k. Independently trained rank-k models do not satisfy this property: their eigenbases are unrelated across k, blocking direct comparison of low- and high-capacity summaries of the same data. Each K_k is the optimal rank-k PSD approximation to K in Frobenius norm by Theorem 2, in contrast to MRL, where nested representations are enforced through a joint loss at training time, our nestedness and optimality at every k are deterministic consequences of the Eckart–Young–Mirsky theorem applied to a single-objective fit. A single learned kernel thus yields an auditable sequence of deployment-time operating points indexed by k and summarized by $d_{\mathrm{spec}}(K_k)$.

Local behavior as η varies. We next investigate the one-dimensional capacity path induced by varying the entropy weight η. Because the objective may have multiple critical points, there is no globally single-valued optimizer map $\eta \mapsto L^\star(\eta)$ in general. The result below is therefore local and branch-wise: it establishes a path of nondegenerate restricted local minima corresponding to values η in the neighborhood of η_0. Along such a branch, the realized dimension

$$d_{\mathrm{spec}}(\eta) = d_{\mathrm{spec}}\big(K(L^\star(\eta))\big)$$

varies smoothly and is locally nondecreasing, provided no spectral or second-order degeneracy occurs.

The statement must quotient out the orthogonal gauge of the factorization. Indeed, $K(LQ) = K(L)$ for every orthogonal $Q \in \mathbb{R}^{r\times r}$, so the full parameterization contains flat directions unrelated to the represented kernel. Fix a rank $s > 0$. After a right-orthogonal reparameterization, work on the slice

$$\widetilde{\Theta} \coloneqq \big\{([L_1 \mid 0], a, \beta) : L_1 \in \mathbb{R}^{N\times s}, \operatorname{rank}(L_1) = s,\ L_1^\top L_1 \text{ is diagonal},\ a \in \mathbb{R}^N,\ \beta > 0\big\} \subset \Theta.$$

This slice is a smooth embedded submanifold of the parameter space, and all derivatives below are restricted to its tangent spaces.[2] Its tangent space at $\theta \in \widetilde{\Theta}$ is denoted by $T_\theta\widetilde{\Theta}$. The restricted objective $F_\eta|_{\widetilde{\Theta}}$ is $\mathcal{C}^2$; see Appendix E. We denote the restricted gradient by $\nabla_\theta F_\eta(\theta)\big|_{T_\theta\widetilde{\Theta}}$ and the Hessian of the restricted function $F_\eta|_{\widetilde{\Theta}}$, as $\nabla^2_\theta F_\eta(\theta)\big|_{T_\theta\widetilde{\Theta}}$ evaluated on tangent directions.

Fix $\eta_0 \in \mathbb{R}$. We take as reference an arbitrary restricted critical point $\theta^\star = (L^\star, a^\star, \beta^\star) \in \widetilde{\Theta}$ of $F_{\eta_0}|_{\widetilde{\Theta}}$, meaning that $\nabla_\theta F_{\eta_0}(\theta^\star)\big|_{T_{\theta^\star}\widetilde{\Theta}} = 0$. The following assumptions are local regularity conditions for the chosen restricted critical point $\theta^\star$. If a rank transition, eigenvalue crossing, or Hessian degeneracy occurs, this branch-wise smoothness guarantee no longer applies.

Assumption 4 (Simple active spectrum)**.** *The positive eigenvalues of $K(L^*)$ are pairwise distinct.*

Assumption 5 (Nondegenerate second-order condition)**.** *The restricted Hessian is positive definite:*

$$v^\top \left(\nabla^2_\theta F_{\eta_0}(\theta^*)\big|_{T_{\theta^*}\widetilde{\Theta}}\right) v > 0 \quad \textit{for every } 0 \neq v \in T_{\theta^*}\widetilde{\Theta}.$$

The proof of the following Theorem is in Appendix F and an accompanying illustration in Figure 4.

Theorem 6 (Local smoothness and monotonicity of the effective spectral dimension)**.** *Let $\eta_0 \in \mathbb{R}$, and let $\theta^\star = (L^\star, a^\star, \beta^\star)$ satisfy Assumptions 4–5. Then, there exists an open interval $U \ni \eta_0$ and a $\mathcal{C}^1$ map $\theta^\star\colon \mathbb{R} \to \widetilde{\Theta}$ such that $\theta^\star(\eta_0) = \theta^\star$ and the following three claims hold: **(i)** for every $\eta \in U$, it holds $\nabla_\theta F_\eta(\theta^\star(\eta))\big|_{T_{\theta^\star(\eta)}\widetilde{\Theta}} = 0$ and $\theta^\star(\eta)$ is a strict local minimum of $F_\eta|_{\widetilde{\Theta}}$, **(ii)** the capacity profile $d_{\mathrm{spec}}(\eta) \coloneqq d_{\mathrm{spec}}\big(K(L^\star(\eta))\big)$ is $\mathcal{C}^1$ on U and, **(iii)** for every $\eta \in U$,*

$$\frac{d}{d\eta} \log d_{\mathrm{spec}}(\eta) \geq 0$$

The theorem is intentionally local. Its hypotheses can fail when - as η varies - an eigenvalue enters or leaves the active support, when positive eigenvalues collide, or when the restricted Hessian becomes degenerate. These events can produce nonsmooth spectrum changes (see Appendix F.1, Fig. 5). The appropriate interpretation of Theorem 6 is a sensitivity statement for a chosen nondegenerate local optimum, rather than a global statement about the optimizer set. In nonconvex problems the argmin correspondence $\eta \mapsto \operatorname{argmin} F_\eta$ can be set-valued and need not vary continuously. The theorem therefore proves the property that is actually needed locally: whenever a fitted solution is nondegenerate and its active spectrum remains simple, the implicit-function theorem selects a unique nearby branch of local minima, and along this branch increasing the entropy weight cannot decrease $\log d_{\mathrm{spec}}$. Thus the result justifies local continuation and local one-dimensional calibration of realized capacity, but it does not assert that all local minima, or the global minimizer, lie on a single smooth monotone path.

Targeting a desired effective dimension. Motivated by the local monotonicity in Theorem 6, we target a desired effective dimension by treating $d_{\mathrm{spec}}(\eta)$ as a one-dimensional response curve on empirically identified monotone intervals. The entropy weight η parameterizes a one-dimensional family of fitted kernels: $\eta > 0$ raises d_{spec}, $\eta < 0$ lowers it. Given a target $d^\star_{\mathrm{spec}}$, we solve $d_{\mathrm{spec}}(\eta) \approx d^\star_{\mathrm{spec}}$ as one-dimensional root finding for $g(\eta) \coloneqq d_{\mathrm{spec}}(\eta) - d^\star_{\mathrm{spec}}$. Starting from an anchor fit at $\eta = 0$, the sign of $g(0)$ selects the search direction; geometric expansion brackets the target and bisection refines η until $|g(\eta)|/d^\star_{\mathrm{spec}} \leq \tau$ (we use $\tau = 0.02$). A monotone guard during expansion falls back to bidirectional search if local monotonicity fails, and probe fits may early-stop once d_{spec} stabilizes, since exact convergence is unnecessary for bracketing. The selected $\eta^\star$ is retrained from a fresh initialization at the full training budget. By Theorem 6, bisection reaches η-precision δ in $O(\log(1/\delta))$ probe evaluations on intervals where the branch is monotone, so calibration adds at most a logarithmic multiplicative factor over a single training run.

[2] Basic differential-geometric definitions are recalled in Appendix B.

Validation-only capacity selection. To choose the target effective dimension without accessing the test set, we reserve validation edges from the non-test training data and fit an uncontrolled anchor at $\eta = 0$ (see supplementary Algorithm 1). For each candidate target $d^{\star}_{\mathrm{spec}}$, we calibrate η by bracketing and bisection, verify the achieved d_{spec}, and record the validation negative log-likelihood (NLL) across seeds. We then select the smallest target whose mean validation NLL lies within one standard error of the minimum (the one-standard-error rule). Finally, we refit at the selected target using the non-test training data and evaluate once on the held-out test set.

Complexity and parametrization. We parametrize the kernel through its SVD, $L = Q\,\mathrm{diag}(\sigma)$, with Q column-orthonormal and σ positive via softplus. This is equivalent to a free factor under $K(L) = N\,LL^{\top}/\mathrm{tr}(LL^{\top})$ but reads the eigendecomposition of K directly off the parameters ($\lambda_j(K) = N\sigma_j^2/\sum_l \sigma_l^2$, eigenvectors Q), making d_{spec} and prefixes K_k available without an eigensolve and exposing the signed-permutation gauge of Proposition 3 at the parameter level. Each training step costs $\mathcal{O}(|E_{\mathrm{batch}}|\,r + Nr)$ with $\mathcal{O}(Nr)$ memory; total cost is $\mathcal{O}(T(|E_{\mathrm{batch}}|\,r + Nr))$, linear in N and r and tractable for $r \ll N$. Calibration adds a logarithmic factor.

4 Experiments & Results

We evaluate SPECTRA against representative latent representation methods spanning Euclidean embeddings, matrix factorization, mixed-membership models, and latent-distance simplex-based representations. SPECTRA is optimized with Adam [Kingma and Ba, 2014] at 5:1 negatives-to-positives sampling per step. For embedding baselines, dyadic edge features use the binary operator (average, Hadamard, weighted-L_1, weighted-L_2) maximizing each baseline's test AUC, giving each baseline its strongest configuration; likelihood-based latent graph models, including SPECTRA, are evaluated directly from the learned log-odds. SPECTRA is reported in a strictly test-blind regime: no hyperparameter, η value, d_{spec} target, or checkpoint is selected with reference to any test quantity at any stage. Hyperparameters, hardware, and implementation pointers are in Appendix G.

Datasets and Baselines. All experiments use featureless undirected graphs spanning five structural regimes: collaboration networks *ca-GrQc* and *ca-HepTh* [Leskovec et al., 2007]; social networks *socfb-American75* and *socfb-Amherst41* [Traud et al., 2012]; biological protein-protein interaction networks *bio-grid-human* and *bio-grid-worm* [Stark et al., 2006]; and infrastructure networks *inf-power* [Watts and Strogatz, 1998] and *inf-openflights* [Opsahl, 2011] (see appendix Table 4 for details). Datasets are obtained from SNAP [Leskovec and Krevl, 2014] and the Network Repository [Rossi and Ahmed, 2015]. We compare against four families of latent representation methods: (i) *Euclidean vector-space embeddings* that learn node representations in $\mathbb{R}^d$ under inner-product geometry, represented by NODE2VEC [Grover and Leskovec, 2016] and ROLE2VEC [Ahmed et al., 2018]; (ii) *matrix-factorization methods*, represented by NETMF [Qiu et al., 2018] and GRAREP [Cao et al., 2015]; (iii) *mixed-membership models*, represented by MNMF [Wang et al., 2017]; and (iv) *latent-distance simplex-based models*, represented by HM-LDM [Nakis et al., 2022]. We focus on featureless graphs in an unsupervised link-prediction setting; the comparison class is therefore the latent-representation methods that operate in this regime, rather than message-passing GNNs, which usually underperform under this regime Nikolentzos et al. [2024].

Empirical tracking protocol. Unlike the local branches considered in Theorem 6, each η is trained from independent random initializations. Consistency of achieved d_{spec} across seeds therefore provides an empirical check of capacity-control stability.

Capacity Frontiers. We test whether predictive performance is better organized by realized spectral capacity $d_{\mathrm{spec}}(\eta)$ than by the nominal rank cap. For each dataset, we sweep η at rank caps $r \in \{64, 128, 256\}$ over three seeds, using an adaptive cold-start grid that refines regions where $d_{\mathrm{spec}}(\eta)$ changes rapidly (Appendix section H). Figure 2 shows the resulting frontiers. Across all eight datasets, the map $\eta \mapsto d_{\mathrm{spec}}$ is monotone with narrow cross-seed variation, spanning roughly two orders of magnitude in realized dimension before saturating near the rank cap. When plotted against achieved d_{spec}, AUC curves from different rank caps align wherever their realized-capacity ranges overlap, indicating that performance is organized by realized spectral capacity rather than by r alone. At very large d_{spec}, performance can decline because positive η pushes the spectrum toward equal mass over all available modes, including modes that carry little signal. The peak locations reveal two regimes. Infrastructure and social networks (*inf-power*, *inf-openflights*, *socfb-American75*, *socfb-Amherst41*) are saturated: their best operating points occur at similar d_{spec} across rank caps and lie well below

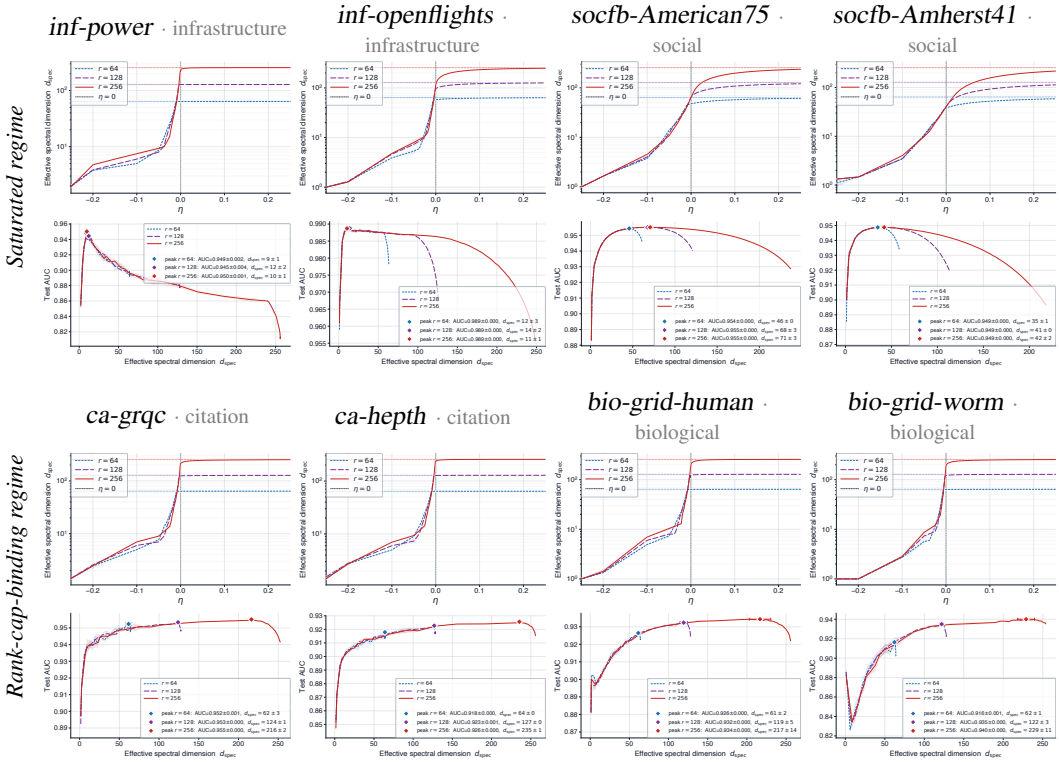


Figure 2: **Capacity frontiers across rank caps.** Each dataset is swept over $\eta \in [-0.25, 0.25]$ at $r \in \{64, 128, 256\}$; curves show mean$\pm 1\sigma$ across three seeds. Upper panels show achieved d_{spec} versus η, lower panels show test AUC versus achieved d_{spec}. Top: saturated datasets. Bottom: rank-cap-binding datasets. Legends report per-rank peak coordinates.

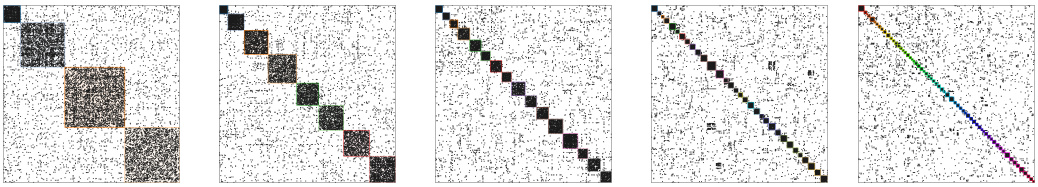
the smallest cap tested. Citation and biological networks (*ca-grqc*, *ca-hepth*, *bio-grid-human*, *bio-grid-worm*) are rank-cap-binding: their best achieved d_{spec} increases with r, and AUC continues to improve as the cap is enlarged. Thus the multi-r sweep distinguishes datasets whose capacity demand is already saturated from those for which the parameterization remains constraining through $r = 256$.

Link prediction. We follow the standard unsupervised protocol [Perozzi et al., 2014, Nakis et al., 2023b, Liben-Nowell and Kleinberg, 2003]: remove 50% of edges while keeping the training graph connected, and sample an equal number of non-edges as negatives. Table 1 reports mean AUC-ROC over $S = 5$ seeds for $d \in \{16, 32, 64, 128\}$; AUC-PR gives qualitatively identical conclusions. Baselines are trained at native dimension d, while SPECTRA uses rank cap $r = d$, so comparisons are matched in parameter count. We also report each method's achieved $d_{\text{spec}} = \exp(H(\sigma^2/\sum_i \sigma_i^2))$ from embedding singular values. SPECTRA is evaluated at $\eta = 0$ and at two bisection-selected settings targeting the median and minimum baseline Shannon effective ranks at the same d; these targets are fixed from baseline embeddings, with no validation or test AUC used to choose η. At matched parameter count, SPECTRA ($\eta = 0$) is top or second on $7/8$ datasets at every d; the exceptions are *inf-power*, where GRAREP dominates, and *inf-openflights*, where HM-LDM leads narrowly. Capacity-targeted SPECTRA remains competitive at much smaller realized dimensions: the median-target setting leads on *ca-HepTh*, *socfb-American75*, and *bio-grid-human* at multiple d, while the minimum-target setting often matches full-rank baselines at $5-10\times$ smaller d_{spec}. Baselines also realize Shannon effective ranks well below nominal d on socfb and bio-grid graphs (e.g., HM-LDM at $d = 128$ has $d_{\text{spec}} \approx 21$ on *socfb-American75* and ≈ 36 on *bio-grid-worm*), reinforcing realized spectral capacity as the operative comparison coordinate.

Visualizing the prefix family. Figure 3 visualizes spectral prefixes $d_{\text{pre}} \in \{4, 8, 16, 32, 64\}$ from a *single* SPECTRA fit on *grqc* with rank cap $r = 128$ and target $d^{\star}_{\text{spec}} = 64$. Each panel uses the rank-d_{pre} prefix $K_{d_{\text{pre}}} = U_{1:d_{\text{pre}}} \Lambda_{1:d_{\text{pre}}} U_{1:d_{\text{pre}}}^{\top}$, which is the optimal rank-d_{pre} PSD approximation of the fitted kernel in Frobenius norm by Theorem 2. Nodes are reordered by the dominant retained

Table 1: **AUC-ROC scores** on eight benchmark graphs, over $S = 5$ seeds for various d. Each method block reports AUC and achieved d_{spec}. **Bold**: best AUC for each (dataset, d); underline: second best.

		NODE2VEC		ROLE2VEC		NETMF		GRAREP		MNMF		HM-LDM		SPECTRA ($\eta=0$)		SPECTRA (med d_{spec})		SPECTRA (min d_{spec})	
Dataset	d	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}	AUC	d_{spec}
ca-GrQc	16	.931	15.91	.936	15.36	.887	15.58	.906	8.45	.905	14.99	.937	14.72	**.940**	15.99	.938	15.18	.932	8.45
	32	.936	31.62	.936	30.47	.899	30.71	.909	13.7	.918	30.94	.942	28.15	**.944**	31.99	.942	30.59	.940	13.7
	64	.937	59.64	.931	55.88	.897	60.38	.897	22.45	.924	62.68	.937	48.40	.949	63.74	**.950**	57.76	.944	22.45
	128	.934	86.41	.924	125.58	.888	116.33	.893	35.08	.918	125.58	.949	90.85	.952	124.51	**.953**	103.59	.946	35.08
ca-HepTh	16	.883	15.95	**.907**	15.12	.831	15.75	.846	5.47	.845	14.99	.876	14.70	.902	15.99	.905	15.06	.891	5.47
	32	.888	31.80	.902	30.13	.836	31.41	.842	8.24	.864	30.97	.877	29.32	.910	31.99	**.911**	30.55	.895	8.24
	64	.892	63.36	.896	58.86	.827	62.37	.832	13.11	.875	62.91	.874	56.30	.915	63.94	**.917**	60.61	.905	13.11
	128	.892	109.57	.882	108.52	.810	122.42	.825	21.59	.876	126.70	.891	95.85	**.921**	127.00	.920	109.04	.908	21.59
socfb-Am.75	16	.923	14.35	.910	14.47	.879	13.02	.932	8.05	.857	14.58	.938	11.03	.944	14.82	**.945**	13.68	.938	8.05
	32	.934	28.75	.914	28.81	.896	26.35	.938	11.47	.876	30.21	.938	15.66	**.951**	27.61	**.951**	27.55	.944	11.47
	64	.930	57.64	.910	60.97	.905	52.22	.942	16.95	.895	60.97	.938	17.24	**.954**	47.65	.952	54.93	.947	16.95
	128	.914	114.33	.896	120.43	.906	100.69	.943	26.44	.912	120.43	.939	20.99	**.955**	64.58	.949	107.51	.949	20.99
socfb-Amh.41	16	.926	13.18	.903	13.51	.892	11.11	.930	7.81	.884	14.07	.939	10.12	**.943**	14.23	.901	12.14	.938	7.81
	32	.926	26.36	.899	26.87	.901	21.80	.933	11.47	.899	29.19	.941	12.35	**.947**	25.00	.915	24.08	.942	11.47
	64	.909	52.58	.882	53.59	.899	42.35	.934	17.95	.909	58.34	.941	13.11	.949	38.01	**.924**	47.46	.944	13.11
	128	.890	102.40	.862	107.40	.885	79.84	.929	29.64	.913	115.79	.942	13.87	**.949**	40.82	.932	91.12	.944	13.87
bio-grid-h.	16	.902	15.75	.878	15.45	.867	14.62	.910	8.66	.840	14.97	.919	14.29	.902	15.99	**.943**	14.80	.899	8.66
	32	.900	31.51	.876	30.69	.851	29.09	.917	14.34	.858	30.89	.917	27.61	.915	31.97	**.947**	29.89	.903	14.34
	64	.892	62.43	.875	60.36	.847	56.72	.918	23.99	.867	62.71	.919	45.53	.925	63.68	**.946**	58.54	.910	23.99
	128	.883	119.15	.877	118.62	.854	109.69	.913	38.77	.872	125.91	.923	76.82	.932	122.87	**.938**	114.16	.919	38.77
bio-grid-w.	16	.786	15.52	.744	15.21	.769	15.57	.900	11.95	.769	14.79	**.913**	13.32	.852	15.99	.852	15.00	.835	11.95
	32	.775	30.58	.748	29.00	.821	30.10	.911	17.67	.780	30.64	**.914**	23.06	.886	31.93	.882	29.55	.860	17.67
	64	.754	57.84	.772	54.80	.852	56.28	.916	24.17	.792	62.35	**.920**	32.03	.915	63.33	.913	55.54	.877	24.17
	128	.747	95.37	.812	103.96	.878	98.86	.917	30.30	.785	124.80	.920	36.19	**.934**	121.33	.931	97.12	.889	30.30
inf-power	16	.892	15.95	.942	15.25	.899	15.73	**.956**	13.01	.816	14.99	.885	14.67	.936	15.99	.939	15.12	.949	13.01
	32	.919	31.9	.942	30.92	.920	31.44	**.956**	20.64	.841	30.97	.881	28.75	.912	31.99	.916	30.94	.925	20.64
	64	.914	60.28	.940	55.54	.892	62.45	**.953**	31.07	.841	62.89	.847	50.53	.895	63.98	.896	57.91	.916	31.07
	128	.901	80.41	.935	63.67	.845	122.82	**.945**	45.15	.813	126.61	.889	95.65	.880	127.89	.884	88.03	.910	45.15
inf-openfl.	16	.974	15.41	.962	13.56	.948	15.45	.984	11.59	.944	14.63	**.992**	12.88	.987	15.83	.988	14.10	.989	11.59
	32	.974	30.16	.963	25.26	.952	29.95	.986	16.45	.959	19.93	**.990**	21.43	.987	30.87	.988	23.34	.988	16.45
	64	.974	55.31	.966	48.12	.949	57.18	.985	21.90	.968	60.95	**.991**	26.67	.987	57.73	.987	51.72	.988	21.90
	128	.973	87.94	.966	94.68	.935	103.84	.982	26.60	.972	122.08	**.991**	38.86	.987	96.43	.987	91.31	.988	26.60



(a) d_{pre}=4 (b) d_{pre}=8 (c) d_{pre}=16 (d) d_{pre}=32 (e) d_{pre}=64

Figure 3: Adjacency matrices reordered by spectral assignment $c_i(d_{\text{pre}}) = \arg\max_{j \le d_{\text{pre}}} U_{ij}^2$ for $d_{\text{pre}} \in \{4, 8, 16, 32, 64\}$. Panels come from a single fitted SPECTRA kernel via the rank-d_{pre} prefix.

spectral coordinate, $c_i(d_{\text{pre}}) = \arg\max_{1 \le j \le d_{\text{pre}}} U_{ij}^2$, and blocks are shaded by this assignment. This is a standard spectral way to read block structure from leading eigenspaces [Ng et al., 2001, von Luxburg, 2007], but here the eigenspaces come from the learned trace-normalized kernel rather than directly from the adjacency. As d_{pre} increases, the reordered adjacency resolves progressively finer structure: broad macro-groups at $d_{\text{pre}} = 4$ split into smaller, denser blocks by $d_{\text{pre}} = 64$, while off-diagonal density decreases. No panel is a refit. The role of η is to set the spectral operating point *before* prefix extraction: targeting $d_{\text{spec}}^{\star}$ determines how spectral mass is distributed across modes, and hence how far the prefix family adds meaningful structure. This gives a graph analogue of Matryoshka-style multi-resolution representations [Kusupati et al., 2024], but with alignment and optimality following from Eckart–Young–Mirsky for a single fitted kernel rather than from a multi-scale training loss.

Overparameterized target-$d_{\text{spec}}^{\star}$ acquisition. We test whether overparameterization, used through spectral control, finds a better-calibrated solution than the rank-cap-only practice. We sweep target Shannon effective ranks $d_{\text{spec}}^{\star} \in \{16, 32, 64\}$ and rank caps $r \in \{d_{\text{spec}}^{\star}, 64, 128, 256\}$ on *grqc* and *bio-grid-human*, with $n = 10$ seeds per cell. Each cell trains two models on identical data: a Phase-0 anchor at $\eta = 0$ (d_{spec} saturates near r) and a Phase-2 retrain at $\eta^{\star}$ obtained by bidirectional bisection of fully-trained probes ($\tau = 0.05$). Metrics target the optimization objective: test log-likelihood TLL and the train–test generalization gap $\text{GenGap} = \text{trNLL} + \text{TLL}$ (lower is better). To assess whether observed differences between the two conditions reflect a systematic effect, we report paired Wilcoxon signed-rank tests on the per-seed differences, paired because the two conditions in each cell share data, negatives, and seed, and signed-rank rather than t-test because paired-difference distributions

Table 2: Community detection on Cora, Citeseer, and LastFM using a rank cap of $r = 64$, across five model seeds; each score is averaged over ten k-means restarts. K: ground-truth communities.

	CORA ($K = 7$)			CITESEER ($K = 6$)			LASTFM ($K = 18$)		
η	d_{spec}	ARI	NMI	d_{spec}	ARI	NMI	d_{spec}	ARI	NMI
0.00	62.6	$.237 \pm .005$	$.396 \pm .004$	62.8	$.031 \pm .002$	$.195 \pm .008$	60.0	$.252 \pm .001$	$.551 \pm .001$
-0.02	17.6	$.337 \pm .005$	$.417 \pm .007$	17.6	$.130 \pm .004$	$.185 \pm .003$	15.9	$.372 \pm .002$	$.558 \pm .003$
-0.04	10.4	$.362 \pm .005$	$.439 \pm .003$	9.2	$.149 \pm .002$	$.183 \pm .002$	8.4	$.357 \pm .011$	$.535 \pm .010$
-0.08	6.6	$.334 \pm .010$	$.420 \pm .006$	6.4	$.157 \pm .012$	$.188 \pm .011$	6.2	$.301 \pm .016$	$.484 \pm .015$

across seeds are heavy-tailed and not well-modeled as Gaussian. *Comparison A* (within-cell, isolating η at fixed r): without spectral control, overparameterization is calibration-disastrous, across 240 paired comparisons, the retrain improves GenGap ($p < 10^{-30}$) and TLL ($p < 10^{-25}$), while the anchor fits training strictly harder (lower trNLL, $p < 10^{-25}$). *Comparison B* (matched $d^{\star}_{\text{spec}}$, against the rank-cap-only baseline at $r = d^{\star}_{\text{spec}}$): η-targeted overparameterization finds a calibration-superior minimum at matched effective capacity, across 180 paired comparisons, both GenGap ($p \approx 10^{-11}$) and TLL ($p \approx 10^{-12}$) favor the retrain, with larger effect on *bio-grid-human* (each $p < 10^{-10}$) than on the more-saturated *grqc* (each $p \approx 0.015$). Since both conditions reach the same realized d_{spec}, the gain comes not from additional representational capacity but from the optimizer landscape that the larger parameterization opens up, a setting consistent with the broader literature on overparameterized models that generalize well when their effective complexity is controlled [Belkin et al., 2019, Bartlett et al., 2020]. Finally, we applied the validation-only selection procedure with rank cap $r = 64$ to three representative graphs. The achieved effective spectral dimension decreased from 58.8 to 4.9 on *inf-openflights*, from 64.0 to 19.0 on *inf-power*, and from 48.6 to 27.7 on *socfb-American75*. The corresponding test AUC changed from 0.973 to 0.974, 0.749 to 0.767, and 0.9505 to 0.9479, respectively. Test NLL improved on all three graphs. Validation-only capacity selection reduced realized spectral dimension and improved test NLL across all three datasets, and comparable AUC.

Community detection. We further examine the role of realized spectral capacity in community detection on networks with known labels Cora and Citeseer [Sen et al., 2008], and LastFM [Rozemberczki and Sarkar, 2020]. With a fixed rank cap $r = 64$, we sweep $\eta \in \{0, -0.02, -0.04, -0.08\}$ and apply k-means to embeddings formed from the leading $\lceil d_{\text{spec}} \rceil$ spectral modes. We report adjusted Rand index (ARI) and normalized mutual information (NMI) against the community labels over five model seeds, averaging each clustering score over ten k-means restarts. As shown in Table 2, the highest ARI observed in the sweep increases from 0.237 to 0.362 on Cora, from 0.031 to 0.157 on Citeseer, and from 0.252 to 0.372 on LastFM, relative to the uncontrolled $\eta = 0$ fit. NMI also improves on Cora and slightly on LastFM, but not on Citeseer. These results show that clustering quality varies with realized spectral capacity, with the effects depending on the dataset and evaluation metric.

5 Conclusion & Limitations

We introduced SPECTRA, which treats capacity in latent graph models as a property of the fitted representation rather than as a nominal rank chosen before training. Its key object is the trace-normalized learned PSD kernel: the Shannon effective rank d_{spec} gives a smooth realized-dimension coordinate, shaped by an entropy weight η and targeted by bisection. The framework provides identifiability of active spectral modes, local C^1 regularity away from structural events, and Eckart–Young–Mirsky-optimal spectral prefixes. Empirically, SPECTRA is competitive with strong link-prediction baselines at matched parameter count, remains competitive at substantially smaller realized dimensions, reveals saturated and rank-cap-binding regimes through capacity frontiers, and yields aligned Matryoshka-style prefix views without a multi-scale objective. Beyond link prediction, SPECTRA suggests a broader use for comparing ensembles of networks through realized representational complexity. For example, d_{spec} could be used to compare structural complexity across village social networks in Honduras [Airoldi and Christakis, 2024], school friendship networks [Harris et al., 2019], and protein interaction networks [Jeong et al., 2001], potentially serving as an outcome variable for studying variation across social and biological settings. The main limitation is geometric: we work with Euclidean latent spaces, where capacity is read from an $O(r)$-invariant Gram spectrum. Simplex-valued and non-Euclidean latent models require the corresponding invariant kernel, e.g., a log-ratio Gram under Aitchison geometry for compositional memberships [Nakis et al., 2026] or a geometry-native kernel for hyperbolic embeddings. Cross-network comparisons would also require common preprocessing and fitting protocols, with controls for graph size and density. We leave these extensions to future work.

Acknowledgments

The authors gratefully acknowledge support from the NOMIS Foundation, the Stavros Niarchos Foundation, and the National Institutes of Health (R01AG081814).

References

Nesreen K Ahmed, Ryan Rossi, John Boaz Lee, Theodore L Willke, Rong Zhou, Xiangnan Kong, and Hoda Eldardiry. Learning role-based graph embeddings. *arXiv preprint arXiv:1802.02896*, 2018.

Edoardo Airoldi and Nicholas Christakis. Induction of social contagion for diverse outcomes in structured experiments in isolated villages. *Science (New York, N.Y.)*, 384:eadi5147, 05 2024. doi: 10.1126/science.adi5147.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels, 2007. URL `https://arxiv.org/abs/0705.4485`.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, April 2020. ISSN 1091-6490. doi: 10.1073/pnas.1907378117. URL `http://dx.doi.org/10.1073/pnas.1907378117`.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL `http://dx.doi.org/10.1073/pnas.1903070116`.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 891–900, 2015.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

Kalyan Cherukuri and Aarav Lala. Low-rank matrix approximation for neural network compression. In *2025 10th International Conference on Machine Learning Technologies (ICMLT)*, pages 166–170. IEEE, 2025.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. URL `https://api.semanticscholar.org/CorpusID:10163399`.

Marcelo Fiori, Bernardo Marenco, Federico Larroca, Paola Bermolen, and Gonzalo Mateos. Gradient-based spectral embeddings of random dot product graphs. *IEEE Transactions on Signal and Information Processing over Networks*, 10:1–16, 2023.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

Weiwei Gu, Aditya Tandon, Yong-Yeol Ahn, and Filippo Radicchi. Principled approach to the selection of the embedding dimension of networks. *Nature Communications*, 12(1):3772, 2021.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Kathleen Harris, Carolyn Halpern, Eric Whitsel, Jon Hussey, Ley Killeya-Jones, Joyce Tabor, and Sarah Dean. Cohort profile: The national longitudinal study of adolescent to adult health (add health). *International journal of epidemiology*, 48, 06 2019. doi: 10.1093/ije/dyz115.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in neural information processing systems*, 20, 2007.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

David Hong and Joshua Cape. Network signflip parallel analysis for selecting the embedding dimension. *arXiv preprint arXiv:2509.05722*, 2025.

Ilse CF Ipsen and Arvind K Saibaba. Stable rank and intrinsic dimension of real and complex matrices. *SIAM Journal on Matrix Analysis and Applications*, 46(3):1988–2007, 2025.

H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001. ISSN 1476-4687. doi: 10.1038/35075138. URL `http://dx.doi.org/10.1038/35075138`.

Nandan Kumar Jha and Brandon Reagen. Spectral scaling laws in language models: emphhow effectively do feed-forward networks use their latent space? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35047–35058, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, December 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL `http://dx.doi.org/10.1561/2200000044`.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2024. URL `https://arxiv.org/abs/2205.13147`.

John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2 edition, 2013.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection, June 2014.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes, 2018. URL `https://arxiv.org/abs/1804.08838`.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM*, page 556–559, 2003.

Gongxu Luo, Jianxin Li, Jianlin Su, Hao Peng, Carl Yang, Lichao Sun, Philip S Yu, and Lifang He. Graph entropy guided node embedding dimension selection for graph neural networks. *arXiv preprint arXiv:2105.03178*, 2021.

Nikolaos Nakis, Abdulkadir Çelikkanat, and Morten Mørup. Hm-ldm: A hybrid-membership latent distance model, 2022. URL `https://arxiv.org/abs/2206.03463`.

Nikolaos Nakis, Abdulkadir Celikkanat, Louis Boucherie, Christian Djurhuus, Felix Burmester, Daniel Mathias Holmelund, Monika Frolcová, and Morten Mørup. Characterizing polarization in social networks using the signed relational latent distance model. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11489–11505. PMLR, 25–27 Apr 2023a. URL https://proceedings.mlr.press/v206/nakis23a.html.

Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann Jørgensen, and Morten Mørup. A hierarchical block distance model for ultra low-dimensional graph representations, 2023b. URL https://arxiv.org/abs/2204.05885.

Nikolaos Nakis, Niels Raunkjær Holm, Andreas Lyhne Fiehn, and Morten Mørup. How low can you go? searching for the intrinsic dimensionality of complex networks using metric node embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=V71ITh2w40.

Nikolaos Nakis, Chrysoula Kosma, Panagiotis Promponas, Michail Chatzianastasis, and Giannis Nikolentzos. Aitchison embeddings for learning compositional graph representations, 2026. URL https://arxiv.org/abs/2605.00716.

A. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems*, 2001. URL https://api.semanticscholar.org/CorpusID:18764978.

Giannis Nikolentzos, Michail Chatzianastasis, and Michalis Vazirgiannis. What do gnns actually learn? towards understanding their representations, 2024. URL https://arxiv.org/abs/2304.10851.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Tore Opsahl. Why anchorage is not (that) important: Binary ties and sample selection. 2011. https://toreopsahl.com/datasets/.

Francesco Sanna Passino and Nicholas A Heard. Bayesian estimation of the latent dimension and communities in stochastic blockmodels. *Statistics and Computing*, 30(5):1291–1307, 2020.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 459–467, 2018.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Ryan A. Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI Conference on Artificial Intelligence*, 2015. URL https://api.semanticscholar.org/CorpusID:10141934.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models, 2020. URL https://arxiv.org/abs/2005.07959.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1446–1473, 2022.

Amartya Sanyal, Philip H. S. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans, 2020. URL https://arxiv.org/abs/1906.04659.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 2008.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34, 01 2006. doi: 10.1093/nar/gkj109. URL https://doi.org/10.1093/nar/gkj109.

Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2013.

Roddy Taing and Keith Levin. On the effect of misspecifying the embedding dimension in low-rank network models. *arXiv preprint arXiv:2601.06014*, 2026.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, August 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2011.12.021. URL http://dx.doi.org/10.1016/j.physa.2011.12.021.

Loring W. Tu. *An Introduction to Manifolds*. Springer, 2 edition, 2011.

Ulrike von Luxburg. A tutorial on spectral clustering, 2007. URL https://arxiv.org/abs/0711.0189.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. URL https://api.semanticscholar.org/CorpusID:3034643.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:39501–39521, 2024.

Jiang Yang, Yuxiang Zhao, and Quanhui Zhu. Effective rank and the staircase phenomenon: New insights into neural network training dynamics. *arXiv e-prints*, pages arXiv–2412, 2024.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

A Dataset statistics

B Extra Preliminaries

We give some basic definitions used in the proofs. We only consider smooth submanifolds of the Euclidean space.

The definitions in this appendix are standard. See, for example, Lee [2013] and Tu [2011].

Definition 7 (Smooth embedded submanifold)**.** Let $M \subset \mathbb{R}^D$. We say that M is a smooth embedded submanifold of dimension m if, around every point $p \in M$, there exists an open set $W \subset \mathbb{R}^m$ and a smooth parametrization

$$\psi : W \to \mathbb{R}^D$$

such that $\psi(W)$ is a neighborhood of p in M, the map $\psi : W \to \psi(W)$ is one-to-one, with smooth inverse, and the Jacobian matrix $J_\psi(x)$ has rank m for every $x \in W$.

Definition 8 (Local chart)**.** Let $M \subset \mathbb{R}^D$ be a smooth embedded submanifold. A local chart around $p \in M$ is a smooth parametrization

$$\psi : W \subset \mathbb{R}^m \to M$$

whose image is a neighborhood of p in M. If $p = \psi(x)$, then $x \in \mathbb{R}^m$ is the local coordinate representation of p. Thus, a function $f : M \to \mathbb{R}$ can locally be written as the ordinary Euclidean function $\widehat{f}(x) := f(\psi(x))$.

Definition 9 (Tangent vector)**.** Let $M \subset \mathbb{R}^D$ be a smooth embedded submanifold and let $p \in M$. A vector $v \in \mathbb{R}^D$ is tangent to M at p if there exists a smooth curve

$$\gamma : (-\varepsilon, \varepsilon) \to M$$

such that $\gamma(0) = p,\ \gamma'(0) = v$.

Definition 10 (Tangent space)**.** The tangent space of M at p, denoted T_pM, is the set of all tangent vectors to M at p:

$$T_pM := \{\gamma'(0) \mid \gamma : (-\varepsilon, \varepsilon) \to M \text{ is smooth and } \gamma(0) = p\}\,.$$

It is a vector subspace of the ambient space $\mathbb{R}^D$.

Definition 11 (Tangent space in local coordinates)**.** Let

$$\psi : W \subset \mathbb{R}^m \to M$$

be a local chart and let $p = \psi(x)$. Then $T_pM = \operatorname{Im} J_\psi(x)$. Equivalently, every tangent vector $v \in T_pM$ can be written as $v = J_\psi(x)u$ for some $u \in \mathbb{R}^m$.

Definition 12 (Tangent space of a constraint set)**.** Suppose

$$M = \{x \in \mathbb{R}^D : c(x) = 0\},$$

where $c : \mathbb{R}^D \to \mathbb{R}^q$ is smooth. Suppose also that the Jacobian matrix $J_c(p)$ has full rank for every $p \in M$. Then M is a smooth embedded submanifold, and its tangent space is

$$T_pM = \ker J_c(p).$$

Equivalently, $v \in T_pM \ \Leftrightarrow\ J_c(p)v = 0$.

C Proof of Proposition 2

Since K is symmetric and PSD, the spectral theorem gives an orthonormal eigenbasis for its image $\operatorname{Im}(K) = \{Kx : x \in \mathbb{R}^N\}$. Thus, we may write

$$K = \sum_{j=1}^{r} \lambda_j u_j u_j^\top,$$

where $r = \operatorname{rank}(K)$, the vectors $u_1, \ldots, u_r$ are orthonormal, and the nonzero eigenvalues satisfy $\lambda_j > 0$. Equivalently,

$$K = U \operatorname{diag}(\lambda_1, \ldots, \lambda_r) U^\top,$$

where $U = (u_1, \dots, u_r)$ and $U^\top U = I_r$. By the Eckart–Young–Mirsky Theorem Eckart and Young [1936], the best rank-k approximation to K in Frobenius norm, among all matrices of rank at most k, is obtained by keeping the k largest singular values and their corresponding singular vectors. Since $K \succeq 0$, its singular values are exactly its eigenvalues. Therefore the unconstrained best rank-k approximation is

$$K_k = \sum_{j=1}^{k} \lambda_j u_j u_j^\top.$$

Next, we want to check that the positive semidefinite constraint does not change the optimizer. The optimization over all rank-k matrices is a relaxation of the optimization over rank-k positive semidefinite matrices:

$$\min_{\mathrm{rank}(M) \le k} \|K - M\|_F^2 \le \min_{\substack{M \succeq 0 \\ \mathrm{rank}(M) \le k}} \|K - M\|_F^2.$$

But the unconstrained minimizer K_k is itself trivially PSD, because it is a nonnegative linear combination of rank-one positive semidefinite matrices. Therefore, it is also optimal among positive semidefinite rank-k matrices.

It remains to prove the uniqueness statement. Assume that $\lambda_k > \lambda_{k+1}$, and let $M \succeq 0$ with $\mathrm{rank}(M) \le k$ be any minimizer. Write the spectral decomposition of M as

$$M = \sum_{i=1}^{k} \mu_i v_i v_i^\top,$$

where $\mu_1 \ge \cdots \ge \mu_k \ge 0$ and $v_1, \dots, v_k$ are orthonormal. Using the Frobenius inner product,

$$\|K - M\|_F^2 = \|K\|_F^2 + \|M\|_F^2 - 2\langle K, M \rangle.$$

Since $K \succeq 0$ and $M \succeq 0$, von Neumann's trace inequality gives

$$\langle K, M \rangle = \mathrm{tr}(KM) \le \sum_{i=1}^{k} \lambda_i \mu_i.$$

Therefore,

$$\begin{aligned} \|K - M\|_F^2 &\ge \sum_{j=1}^{r} \lambda_j^2 + \sum_{i=1}^{k} \mu_i^2 - 2 \sum_{i=1}^{k} \lambda_i \mu_i \\ &= \sum_{i=1}^{k} (\mu_i - \lambda_i)^2 + \sum_{j=k+1}^{r} \lambda_j^2. \end{aligned}$$

On the other hand, the truncated matrix K_k satisfies $\|K - K_k\|_F^2 = \sum_{j=k+1}^{r} \lambda_j^2$. Since M is also assumed to be optimal, we must have $\|K - M\|_F^2 = \sum_{j=k+1}^{r} \lambda_j^2$. Comparing this with the lower bound above, we get

$$\sum_{i=1}^{k} (\mu_i - \lambda_i)^2 = 0.$$

Hence $\mu_i = \lambda_i,\ i = 1, \dots, k$.

Moreover, equality must also hold in von Neumann's trace inequality. The equality implies that the image of M must be a top-k eigenspace of K. Because $\lambda_k > \lambda_{k+1}$, this top-$k$ eigenspace is uniquely determined, namely

$$\mathrm{span}\{u_1, \dots, u_k\}.$$

Therefore M has eigenvalues $\lambda_1, \dots, \lambda_k$ on this uniquely determined subspace and is zero on its orthogonal complement. Consequently,

$$M = \sum_{j=1}^{k} \lambda_j u_j u_j^\top = K_k.$$

Thus K_k is the unique minimizer.

D Proof of Propostion 3

Since K is symmetric and PSD, the spectral theorem gives an orthonormal basis of eigenvectors for its image. Thus K can be written as

$$K = \sum_{j=1}^{r} \lambda_j u_j u_j^\top,$$

where $u_1, \ldots, u_r$ are orthonormal eigenvectors and $\lambda_1, \ldots, \lambda_r$ are the positive eigenvalues.

For any eigenvalue λ_j, the associated eigenspace is

$$E_{\lambda_j} = \{v : Kv = \lambda_j v\}.$$

Since E_{λ_j} is one-dimensional, any unit eigenvector u'_j associated with λ_j satisfies $u'_j = c_j u_j$ for some scalar c_j. Because the vectors in the eigendecompositioon are orthonormal it holds that $\|u'_j\|_2 = \|u_j\|_2 = 1$, we have $|c_j| = 1$, and hence $c_j \in \{\pm 1\}$. Therefore

$$u'_j = \pm u_j. \tag{9}$$

Now suppose

$$K = U \operatorname{diag}(\lambda_1, \ldots, \lambda_r) U^\top = U' \operatorname{diag}(\lambda'_1, \ldots, \lambda'_r) {U'}^\top$$

are two spectral decompositions, with $U^\top U = {U'}^\top U' = I_r$.

For each $\ell \in \{1, \ldots, r\}$, the vector u'_ℓ, the ℓ-th column of U', is a unit eigenvector of K with eigenvalue λ'_ℓ. Since the positive eigenvalues of K are exactly $\lambda_1, \ldots, \lambda_r$, and these eigenvalues are pairwise distinct, there exists a unique index $\sigma(\ell) \in \{1, \ldots, r\}$ such that

$$\lambda'_\ell = \lambda_{\sigma(\ell)}.$$

The map $\sigma : \{1, \ldots, r\} \to \{1, \ldots, r\}$ is a permutation.

Since both u'_ℓ and $u_{\sigma(\ell)}$ are unit eigenvectors associated with the same one-dimensional eigenspace $E_{\lambda_{\sigma(\ell)}}$, using (9) there exists a sign $s_\ell \in \{\pm 1\}$ such that

$$u'_\ell = s_\ell u_{\sigma(\ell)}.$$

Let Π be the permutation matrix whose ℓ-th column is $e_{\sigma(\ell)}$, and let

$$S = \operatorname{diag}(s_1, \ldots, s_r).$$

Then the column-wise identities above imply

$$U' = U \Pi S.$$

Moreover, the identities $\lambda'_\ell = \lambda_{\sigma(\ell)}$ imply $\lambda' = \Pi^\top \lambda$ and the proof is complete.

E Smoothness of the restricted objective, $F_\eta|_{\widetilde{\Theta}}$.

Lemma 13. *For any η, the restricted objective $F_\eta|_{\widetilde{\Theta}}$ is $\mathcal{C}^2$.*

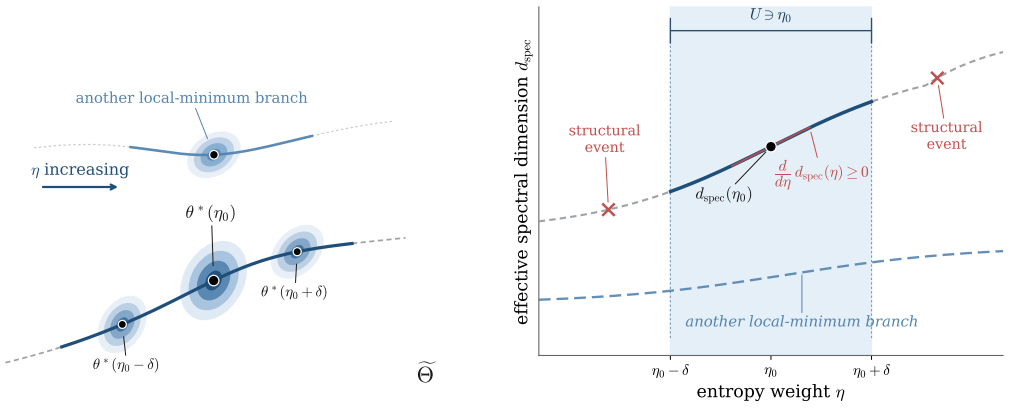
Proof. Recall that

$$F_\eta(\theta) = \Phi(\theta) - \eta h(\theta), \qquad \Phi(\theta) := \ell(\theta) + R(\theta), \qquad h(\theta) := \log d_{\mathrm{spec}}(K(L)).$$

The logistic negative log-likelihood ℓ is $\mathcal{C}^\infty$ in θ on $\{L \neq 0\}$ as it is a finite sum of terms $\log \phi(\pm(a_i + a_j + \beta K_{ij}(L)))$, where ϕ and $\log \phi$ are $\mathcal{C}^\infty$ on $\mathbb{R}$, and $K(L) = nLL^\top / \|L\|_F^2$ is rational in L with denominator nonzero on $\{L \neq 0\}$. The regularizer R is a quadratic polynomial and therefore $\mathcal{C}^\infty$. It remains to show that $h|_{\widetilde{\Theta}}$ is smooth.

Let

$$\theta = ([L_1 \mid 0], a, \beta) \in \widetilde{\Theta},$$



(a) (b)

Figure 4: Illustration of Theorem 6. **(a)** Solution branches in $\widetilde{\Theta}$. Near a reference point $\theta^*(\eta_0)$ satisfying the Assumptions 4 and 5, the optimum traces a smooth curve as η varies, with each $\theta^*(\eta)$ a strict local minimum of F_η on its slice. F_η may have additional local minima, each generating its own local branch under the same conditions. **(b)** Capacity profiles $d_{\text{spec}}(\eta)$ along the branches. The tracked branch is smooth and nondecreasing on U. Beyond U, structural events on the tracked branch can interrupt the local guarantee.

and write the columns of L_1 as $L_1 = (\ell_1, \dots, \ell_s)$. Since $L_1^\top L_1$ is diagonal and L_1 has full column rank, define

$$d_i := \|\ell_i\|_2^2 > 0, \qquad Q := \sum_{j=1}^{s} d_j = \|L_1\|_F^2.$$

Then $L_1^\top L_1 = \operatorname{diag}(d_1, \dots, d_s)$. The nonzero eigenvalues of $L_1 L_1^\top$ are the same as the eigenvalues of $L_1^\top L_1$. Therefore the nonzero eigenvalues of

$$K(L) = N \frac{L_1 L_1^\top}{\|L_1\|_F^2}$$

are

$$\lambda_i(K(L)) = N \frac{d_i}{Q}, \qquad i = 1, \dots, s.$$

Hence the active spectral occupancies are

$$p_i(L) = \frac{\lambda_i(K(L))}{N} = \frac{d_i}{Q}, \qquad i = 1, \dots, s.$$

Thus, on $\widetilde{\Theta}$,

$$h(\theta) = \log d_{\text{spec}}(K(L)) = -\sum_{i=1}^{s} \frac{d_i}{Q} \log\left(\frac{d_i}{Q}\right).$$

Each d_i is a polynomial function of the entries of L_1, and on $\widetilde{\Theta}$ we have $d_i > 0$ and $Q > 0$. Therefore the expression above is smooth in L_1. It does not depend on a or β. Hence $h|_{\widetilde{\Theta}}$ is smooth.

Consequently,

$$F_\eta|_{\widetilde{\Theta}} = \Phi|_{\widetilde{\Theta}} - \eta h|_{\widetilde{\Theta}}$$

is $\mathcal{C}^2$ for every fixed η. This proves the lemma. □

F Proof of Theorem 6

First, we show the set $\widetilde{\Theta}$ is a smooth submanifold and define the tangent space $T_\theta \widetilde{\Theta}$. Then we show the three parts of Theorem 6.

Smooth Submanifold and Tangent Space. Recall the definition of the set $\widetilde{\Theta}$, i.e.,

$$\widetilde{\Theta} := \left\{([L_1 \mid 0], a, \beta) : L_1 \in \mathbb{R}^{N\times s},\ \mathrm{rank}(L_1) = s,\ L_1^\top L_1 \text{ is diagonal},\ a \in \mathbb{R}^N,\ \beta > 0\right\}.$$

The set $\widetilde{\Theta}$ is a smooth submanifold as per Definition 7: indeed, the conditions $\mathrm{rank}(L_1) = s$ and $\beta > 0$ are open conditions, while the constraints $L_2 = 0$ and $\mathrm{offdiag}(L_1^\top L_1) = 0$ form a smooth submersion on the full-column-rank set; hence the claim follows from the regular level-set theorem (see Chapter 8, Lee [2013]).

Fix a point $\theta = ([L_1 \mid 0], \alpha, \beta) \in \widetilde{\Theta}$. Next we show the tangent space of $\widetilde{\Theta}$ is

$$T_\theta\widetilde{\Theta} := \left\{([\dot{L}_1 \mid 0], \dot{a}, \dot{\beta}) : \mathrm{offdiag}\left(L_1^\top \dot{L}_1 + \dot{L}_1^\top L_1\right) = 0\right\},$$

using Definition 12. We write a general factor matrix $L \in \mathbb{R}^{N\times r}$ as $L = [L_1 \mid L_2]$ where $L_1 \in \mathbb{R}^{N\times s}$, $L_2 \in \mathbb{R}^{N\times(r-s)}$. Consider the open set $\mathcal{U} := \{(L_1, L_2, a, \beta) : \mathrm{rank}(L_1) = s,\ \beta > 0\}$. Then $\widetilde{\Theta} \subseteq \mathcal{U}$ and is defined by the smooth equality constraints $L_2 = 0$ and $\mathrm{offdiag}(L_1^\top L_1) = 0$. Equivalently, define $c : \mathcal{U} \to \mathbb{R}^{N\times(r-s)} \times \mathbb{S}_0^s$ where $\mathbb{S}_0^s := \{B \in \mathbb{R}^{s\times s} : B^\top = B,\ B_{ii} = 0 \text{ for all } i\}$ by

$$c(L_1, L_2, a, \beta) := \left(L_2,\ \mathrm{offdiag}(L_1^\top L_1)\right).$$

Then $\widetilde{\Theta} = c^{-1}(0)$. After identifying matrix spaces with Euclidean spaces by vectorization, this is exactly of the form $M = \{x : c(x) = 0\}$ in Definition 12. Thus, we know $T_\theta\widetilde{\Theta} = \ker J_c(\theta)$ which we calculate next.

Let $\dot{\theta} = ([\dot{L}_1 \mid \dot{L}_2], \dot{a}, \dot{\beta})$ be an arbitrary direction in the ambient space. The first component of c is the map $(L_1, L_2, a, \beta) \mapsto L_2$, which is linear in L_2 and independent of L_1, a, β. Therefore the Jacobian of this component maps $([\dot{L}_1 \mid \dot{L}_2], \dot{a}, \dot{\beta}) \mapsto \dot{L}_2$.

It remains to compute the Jacobian of the second component, $L_1 \mapsto \mathrm{offdiag}(L_1^\top L_1)$. Write $L_1 = (\ell_1, \ldots, \ell_s)$ and $\dot{L}_1 = (\dot{\ell}_1, \ldots, \dot{\ell}_s)$, where $\ell_i, \dot{\ell}_i \in \mathbb{R}^n$. For $i \neq j$, define the scalar constraint

$$c_{ij}(L_1) := (L_1^\top L_1)_{ij} = \ell_i^\top \ell_j = \sum_{k=1}^n (L_1)_{ki}(L_1)_{kj}.$$

We identify L_1 with the vector $\mathrm{vec}(L_1)$. Hence $J_{c_{ij}}(L_1)$ denotes the ordinary Jacobian row of the scalar function c_{ij} with respect to the entries of L_1. For $a \in [N]$ and $b \in [s]$, we have

$$\frac{\partial c_{ij}}{\partial (L_1)_{ab}}(L_1) = \mathbf{1}_{\{b=i\}}(L_1)_{aj} + \mathbf{1}_{\{b=j\}}(L_1)_{ai}.$$

Therefore, applying this Jacobian row to the direction $\mathrm{vec}(\dot{L}_1)$, we obtain

$$\begin{aligned} J_{c_{ij}}(L_1)\,\mathrm{vec}(\dot{L}_1) &= \sum_{a=1}^n \sum_{b=1}^s \frac{\partial c_{ij}}{\partial (L_1)_{ab}}(L_1)(\dot{L}_1)_{ab} \\ &= \sum_{a=1}^n (L_1)_{aj}(\dot{L}_1)_{ai} + \sum_{a=1}^n (L_1)_{ai}(\dot{L}_1)_{aj} \\ &= \ell_j^\top \dot{\ell}_i + \ell_i^\top \dot{\ell}_j \\ &= \left(L_1^\top \dot{L}_1 + \dot{L}_1^\top L_1\right)_{ij}. \end{aligned}$$

Collecting these identities over all off-diagonal pairs $i \neq j$, we get $J_{\mathrm{offdiag}(L_1^\top L_1)}(L_1)\,\mathrm{vec}(\dot{L}_1) = \mathrm{offdiag}\left(L_1^\top \dot{L}_1 + \dot{L}_1^\top L_1\right)$.

Combining the two components of c, the full Jacobian satisfies

$$J_c(L_1, L_2, a, \beta) \begin{pmatrix} \dot{L}_1 \\ \dot{L}_2 \\ \dot{a} \\ \dot{\beta} \end{pmatrix} = \left(\dot{L}_2,\ \mathrm{offdiag}\left(L_1^\top \dot{L}_1 + \dot{L}_1^\top L_1\right)\right).$$

By Definition 12, the tangent space is the kernel of the Jacobian, i.e, all $\dot{\theta}$ such that $J_c(\theta)\,\dot{\theta} = 0$. Therefore, using the formula above, we get the result

$$T_\theta\widetilde{\Theta} = \left\{([\dot{L}_1 \mid 0], \dot{a}, \dot{\beta}) : \mathrm{offdiag}\left(L_1^\top \dot{L}_1 + \dot{L}_1^\top L_1\right) = 0\right\}.$$

Claim (i) Recall that $\theta^\star \in \widetilde{\Theta}$ satisfies Assumptions 4 and 5 and

$$F_\eta(\theta) = \Phi(\theta) - \eta h(\theta), \qquad \Phi(\theta) := \ell(\theta) + R(\theta), \qquad h(\theta) := \log d_{\mathrm{spec}}(K(L)).$$

Since $\widetilde{\Theta}$ is a smooth manifold as per Definition 7 there exists a smooth local chart $\psi : W \subset \mathbb{R}^m \to \tilde{\Theta}$ with $\psi(0) = \theta^\star$. Define $\widehat{F}(x, \eta) := F_\eta(\psi(x))$, where $\widehat{h}(x) := h(\psi(x))$ and $\widehat{\Phi}(x) := \Phi(\psi(x))$. Then,

$$\widehat{F}(x, \eta) = \widehat{\Phi}(x) - \eta \widehat{h}(x).$$

By Lemma 13 in Appendix E, $\widehat{F}$ is C^2 in (x, η). Now define $G \colon \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ as

$$G(x, \eta) := \nabla_x \widehat{F}(x, \eta).$$

By Assumption 5, since $\psi(0) = \theta^\star$, it holds $G(0, \eta_0) = \nabla_x \widehat{F}(0, \eta_0) = 0$. Moreover, let $\nabla_x G(0, \eta_0) = \nabla_x^2 \widehat{F}(0, \eta_0)$. By Assumption 5 matrix $\nabla_x G(0, \eta_0)$ is positive definite. In particular, it is invertible. Hence, by the implicit function theorem [Lee, 2013, Ch. 5], there exist an open interval $U \ni \eta_0$ and a $\mathcal{C}^1$ map $x^\star : U \to W$ such that $x^\star(\eta_0) = 0$ and $G(x^\star(\eta), \eta) = 0$ for every $\eta \in U$. Define

$$\theta^\star(\eta) := \psi(x^\star(\eta)).$$

Then $\theta^\star : U \to \tilde{\Theta}$ is $\mathcal{C}^1$, $\theta^\star(\eta_0) = \theta^\star$, and

$$\nabla_\theta F_\eta(\theta^\star(\eta))\big|_{T_{\theta^\star(\eta)}\tilde{\Theta}} = 0 \qquad \text{for every } \eta \in U. \tag{10}$$

This proves the restricted criticality part of claim (i). It remains to show that these restricted critical points, i.e., the ones satisfying Equation 10, are strict local minima. Define

$$H(\eta) := \nabla_x^2 \widehat{F}(x^\star(\eta), \eta).$$

The map $\eta \mapsto H(\eta)$ is continuous, because $\widehat{F}$ is $\mathcal{C}^2$ and $x^\star$ is $\mathcal{C}^1$. Notice that, for $\eta = \eta_0$, we have $H(\eta_0) = \nabla_x^2 \widehat{F}(0, \eta_0) \succ 0$ by Assumption 5. Positive definiteness is an open condition. Therefore, after possibly shrinking U, we have $H(\eta) \succ 0 \qquad$ for every $\eta \in U$.

Thus for any $\eta \in U$ the function $x \mapsto \widehat{F}(x, \eta)$ satisfies

$$\nabla_x \widehat{F}(x^\star(\eta), \eta) = 0 \qquad \nabla_x^2 \widehat{F}(x^\star(\eta), \eta) = H(\eta) \succ 0\,.$$

By the standard second-order sufficient condition for an unconstrained local minimum [Nocedal and Wright, 2006, Ch. 2], $x^\star(\eta)$ is a strict local minimum of $x \mapsto \widehat{F}(x, \eta)$. Since ψ is a local coordinate chart for $\widetilde{\Theta}$, this is exactly the statement that $\theta^\star(\eta)$ is a strict local minimum of F_η restricted to $\widetilde{\Theta}$. Claim (i) follows.

Claim (ii) By claim (i), the map $\eta \mapsto \theta^\star(\eta)$ is $\mathcal{C}^1$ and takes values in $\tilde{\Theta}$. Therefore we may write $L^\star(\eta) = [L_1^\star(\eta) \mid 0]$, where $L_1^\star(\eta) = (\ell_1(\eta), \ldots, \ell_s(\eta))$ depends $\mathcal{C}^1$ on η. Define

$$d_i(\eta) := \|\ell_i(\eta)\|_2^2, \qquad Q(\eta) := \sum_{j=1}^{s} d_j(\eta).$$

Since $d_i(\eta_0) = d_i^\star > 0$, continuity implies that, after shrinking U if necessary, $d_i(\eta) > 0 \qquad$ for every $i = 1, \ldots, s$ and every $\eta \in U$. Therefore,

$$p_i(\eta) := \frac{d_i(\eta)}{Q(\eta)}$$

is $\mathcal{C}^1$ on U. Notice that $p_i(\eta) = \lambda_i(K(L^\star(\eta)))/N$ since $(L^\star)^T L$ is diagonal and $K(L^\star) = N \cdot L^\star(L^\star)^T / \operatorname{tr}(L^\star(L^\star)^T)$.[3] Hence

$$\log d_{\mathrm{spec}}(\eta) = \log d_{\mathrm{spec}}(K(L^\star(\eta))) = -\sum_{i=1}^{s} p_i(\eta) \log p_i(\eta).$$

The function $x \mapsto x \log x$ is $\mathcal{C}^\infty$ on $(0, \infty)$, and each $p_i(\eta)$ is positive and $\mathcal{C}^1$ (since it is a polynomial in the entries of $L^\star$ and $L^\star$ is $\mathcal{C}^1$ on η). Therefore $\log d_{\mathrm{spec}}(\eta)$ is $\mathcal{C}^1$. Since $d_{\mathrm{spec}}(\eta) = \exp(\log d_{\mathrm{spec}}(\eta))$, the capacity profile d_{spec} is also $\mathcal{C}^1$. This proves Claim (ii).

[3] We repeat this argument in more detail in the proof of Lemma 13.

Claim (iii) We work in the same local coordinates as in the proof of claim (i). Recall that

$$\widehat{F}(x,\eta) = \widehat{\Phi}(x) - \eta\widehat{h}(x).$$

The branch satisfies

$$\nabla_x \widehat{F}(x^\star(\eta),\eta) = 0 \qquad \text{for every } \eta \in U.$$

Differentiate this identity with respect to η. Here the first term comes from differentiating the map

$$x \mapsto \nabla_x \widehat{F}(x,\eta)$$

along the curve $x^\star(\eta)$, while the second term comes from the explicit dependence of $\widehat{F}$ on η. Since

$$\partial_\eta \nabla_x \widehat{F}(x,\eta) = -\nabla_x \widehat{h}(x),$$

we get

$$0 = \frac{d}{d\eta}\nabla_x \widehat{F}(x^\star(\eta),\eta) = H(\eta)\frac{dx^\star}{d\eta}(\eta) - \nabla_x \widehat{h}(x^\star(\eta)).$$

Therefore,

$$H(\eta)\frac{dx^\star}{d\eta}(\eta) = \nabla_x \widehat{h}(x^\star(\eta)).$$

By claim (i), $H(\eta) \succ 0$ for every $\eta \in U$. Thus $H(\eta)$ is invertible, and

$$\frac{dx^\star}{d\eta}(\eta) = H(\eta)^{-1}\nabla_x \widehat{h}(x^\star(\eta)).$$

Now

$$\log d_{\mathrm{spec}}(\eta) = h(\theta^\star(\eta)) = \widehat{h}(x^\star(\eta)).$$

Hence, by the chain rule,

$$\frac{d}{d\eta}\log d_{\mathrm{spec}}(\eta) = \nabla_x \widehat{h}(x^\star(\eta))^\top \frac{dx^\star}{d\eta}(\eta).$$

Substituting the expression for $dx^\star/d\eta$, we obtain

$$\frac{d}{d\eta}\log d_{\mathrm{spec}}(\eta) = \nabla_x \widehat{h}(x^\star(\eta))^\top H(\eta)^{-1}\nabla_x \widehat{h}(x^\star(\eta)).$$

Since $H(\eta) \succ 0$, also $H(\eta)^{-1} \succ 0$. Therefore

$$\frac{d}{d\eta}\log d_{\mathrm{spec}}(\eta) \geq 0.$$

F.1 Structural events along the η path

Theorem 6 guarantees $\mathcal{C}^1$ regularity and local monotonicity of the realized effective dimension $d_{\mathrm{spec}}(\eta)$ on a neighborhood of any reference η_0 where Assumptions 4–5 hold. Three failure modes can interrupt this regularity along an extended η path: an eigenvalue enters or leaves the active support; two adjacent positive eigenvalues approach equality (violating Assumption 4); or the restricted Hessian becomes degenerate (violating Assumption 5).

We monitor the first two directly during the η sweep. Support changes are flagged when $k_\tau(K) = |\{i : \lambda_i(K) \geq \tau\lambda_1(K)\}|$ at relative-mass threshold τ changes between consecutive probes. Near-degeneracy events are flagged when the minimum gap between adjacent positive eigenvalues drops below a relative tolerance of the largest eigenvalue. Hessian degeneracy is not directly observable from the spectrum and is treated as correlated with the first two.

Figure 5 shows the η-path of the realized effective dimension on four representative datasets, one from each structural family, at rank caps $r{=}128$ and $r{=}256$. Markers indicate detected structural events, without distinguishing type. Three observations are consistent across datasets: **(i)** the $\eta \mapsto d_{\mathrm{spec}}$ map is monotone in η across the entire sweep, including across detected events; **(ii)** events concentrate near the extremes of the η-range — at large positive η the spectrum is driven toward uniformity and adjacent eigenvalues collide, while at large negative η the spectrum compresses and modes drop

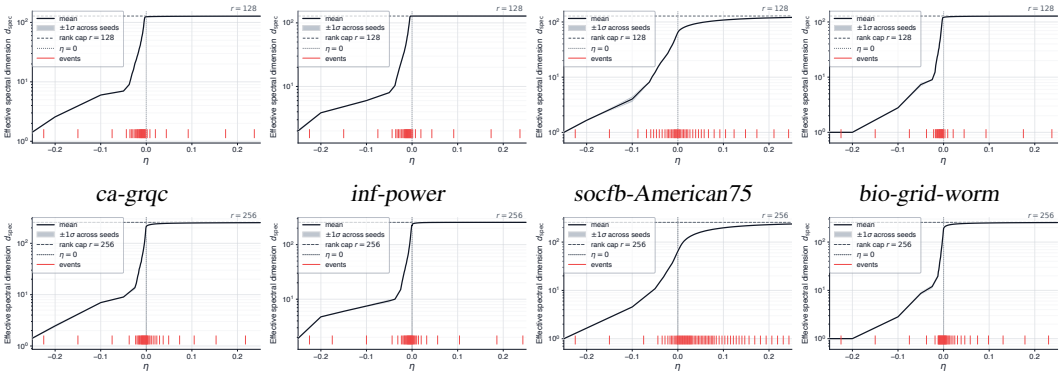


Figure 5: **Structural events along the η path at two rank caps**, on four representative datasets spanning the four structural families. *Top row:* r=128. *Bottom row:* r=256. Columns: citation (*ca-grqc*), infrastructure (*inf-power*), social (*socfb-American75*), biological (*bio-grid-worm*). Each panel shows $d_{\mathrm{spec}}(\eta)$ vs η on log scale, mean $\pm 1\sigma$ across three seeds. Vertical line: η=0. Event markers indicate probes where a structural event was detected along the fitted branch (eigenvalue entering or leaving the active support, or adjacent positive eigenvalues approaching equality; see text). Most η-paths are smooth across $\eta \in [-0.25, 0.25]$ with events concentrating near the extremes; the realized-dimension profile remains continuous and monotone in η on every dataset even where local $\mathcal{C}^1$ regularity is interrupted by an event.

below the support threshold; **(iii)** between events, the path is visibly smooth, consistent with the local $\mathcal{C}^1$ guarantee of Theorem 6.

Empirically, structural events do not invalidate the capacity frontier: $d_{\mathrm{spec}}(\eta)$ remains monotone and continuous in η on every dataset and rank cap we examined. The events appear as visible inflections rather than as discontinuities, suggesting that on these graphs the local $\mathcal{C}^1$ branches connect into a globally continuous (if not globally $\mathcal{C}^1$) profile. We report $d_{\mathrm{spec}}(K^\star)$ as the operating coordinate while tracking events as a diagnostic.

Comparing the two rank caps shows that the event landscape depends on whether r binds the representation. On rank-cap-binding graphs (*ca-grqc*, *bio-grid-worm*), increasing r from 128 to 256 extends the d_{spec} range and shifts events upward; on saturated-regime graphs (*inf-power*, *socfb-American75*), the path stays well below the cap at both r and the event landscape is sparser. This is consistent with the regime distinction in the main paper.

G Experimental setup

We extensively evaluate SPECTRA against prominent baseline graph representation learning methods, including unconstrained Euclidean embeddings, matrix factorization, mixed-membership models, and latent-distance simplex-based representations, across networks of varying sizes and structures. We ran all methods using publicly available implementations. When GPU support was available, experiments were executed on an NVIDIA A100 GPU; otherwise, we used an Apple M2 machine with 8 GB RAM.

Optimization. For SPECTRA we optimize the η-augmented objective F_η (Eq. (6)) with Adam [Kingma and Ba, 2014] at learning rate 10^{-2} for 6,000 iterations unless stated otherwise. At each step we sample five non-edges per observed edge uniformly from the complement of the training graph, and combine the positive and negative log-likelihood terms with equal weight.

Regularization. The spectral regularizer R in F_η is weighted by a fixed coefficient $\lambda = 10^{-4}$ in all experiments reported in the main text and supplement. We found results to be robust to this choice: varying λ over several orders of magnitude around 10^{-4} produced negligible changes in test AUC and in the achieved d_{spec}, so no per-dataset tuning was performed.

Algorithm 1 Validation-only selection of realized spectral capacity

Require: Training edges $\mathcal{E}_{\text{train}}$, candidate targets $\mathcal{D}$, tolerance τ, and random seeds $\mathcal{S}$
Ensure: Selected target $d^{\star}_{\text{spec}}$ and final fitted model

Split $\mathcal{E}_{\text{train}}$ into fitting and validation edges; keep the test set untouched.
Fit the anchor at $\eta = 0$ on the fitting edges and record its realized dimension d_0.
for each target $d \in \mathcal{D}$ **do**
    Choose the initial search direction from $\operatorname{sign}(d - d_0)$.
    Bracket and bisect η as described above; verify $|d_{\text{spec}}(\eta) - d|/d \leq \tau$ for each successful fit.
    Record the achieved dimension $\widehat{d}(d)$ and the mean validation NLL $\bar{\ell}_{\text{val}}(d)$ with standard error $\text{SE}_{\text{val}}(d)$ across seeds $\mathcal{S}$.
end for
Let $\mathcal{D}_{\text{ok}} \subseteq \mathcal{D}$ be targets reached within tolerance.
$d_{\text{best}} \leftarrow \arg\min_{d \in \mathcal{D}_{\text{ok}}} \bar{\ell}_{\text{val}}(d)$.
$\mathcal{D}_{1\text{SE}} \leftarrow \big\{ d \in \mathcal{D}_{\text{ok}} : \bar{\ell}_{\text{val}}(d) \leq \bar{\ell}_{\text{val}}(d_{\text{best}}) + \text{SE}_{\text{val}}(d_{\text{best}}) \big\}$.
$d^{\star}_{\text{spec}} \leftarrow \arg\min_{d \in \mathcal{D}_{1\text{SE}}} \widehat{d}(d)$.
Refit at $d^{\star}_{\text{spec}}$ using all training edges and the full training budget; evaluate once on the untouched test set.

Baselines. For embedding baselines, dyadic edge features are constructed using the standard binary operators (average, Hadamard, weighted-L_1, weighted-L_2), followed by an L_2-regularized logistic regression classifier; in each cell we report the operator that maximizes the baseline's test AUC, giving each baseline its strongest possible configuration. For likelihood-based latent graph models, including ours, we evaluate links directly from the learned log-odds, with no auxiliary classifier and no operator selection.

Test-blind protocol. SPECTRA is reported in a strictly test-blind regime: no hyperparameter, η value, d_{spec} target, or model checkpoint is selected with reference to any test quantity, on any dataset, at any stage.

H Frontier protocol

We test whether predictive performance is organized by realized spectral capacity $d_{\text{spec}}(\eta)$ rather than by the nominal rank cap r. If this is the case, curves obtained at different r should collapse onto a common frontier when plotted against d_{spec}, and any residual dependence on r should appear only once the rank cap binds the achievable capacity.

For each dataset and three seeds, we sweep the spectral-prior weight η at three rank caps $r \in \{64, 128, 256\}$. Each probe is trained from scratch (cold start, no warm-starting across η values) so that the realized capacity at a given η reflects optimization under that prior alone, rather than path-dependence from a neighboring configuration.

Adaptive grid. The η values are chosen *adaptively*: starting from the most negative η, the next probe is selected by halving or doubling the current η-step according to whether d_{spec} moved by more than an upper threshold or less than a lower threshold between consecutive probes. This refines the grid where $d_{\text{spec}}(\eta)$ changes rapidly while spending few probes on slowly-varying segments, so structural transitions are resolved without paying for a uniformly fine sweep.

Table 3: **Notation used throughout the paper.** Symbols appear in the order they are introduced. The two effective-dimension symbols d_{spec} and d_{pre} are conceptually distinct: d_{spec} is the continuous Shannon effective rank of the fitted kernel (controlled at training time by η), while d_{pre} is the integer prefix size chosen post-hoc when extracting a rank-d_{pre} prefix of the fitted kernel.

Symbol	Meaning
Graph and observations	
$\mathcal{G} = (V, E)$	Simple undirected graph with node set V and edge set E.
N	Number of nodes, $N = \lvert V\rvert$.
$\boldsymbol{Y} \in \{0,1\}^{N\times N}$	Symmetric adjacency matrix, $\boldsymbol{Y}_{ii} = 0$.
Latent representation	
$L \in \mathbb{R}^{N\times r}$	Latent factor; rank cap r.
r	Parameterization ceiling on $\mathrm{rank}(L)$.
$K(L) \in \mathbb{R}^{N\times N}$	Trace-normalized PSD kernel, $K(L) = N\,LL^\top / \mathrm{tr}(LL^\top)$.
$\boldsymbol{a} \in \mathbb{R}^N$	Node-specific log-odds offsets.
$\beta > 0$	Global slope on the latent affinity.
$\phi(t)$	Sigmoid, $\phi(t) = (1 + e^{-t})^{-1}$.
Spectrum and effective dimension	
$\lambda_j(K)$	The j-th eigenvalue of K, sorted descending.
u_j	Orthonormal eigenvector associated with $\lambda_j(K)$.
U, Λ	Orthonormal eigenvector matrix and diagonal eigenvalue matrix in $K = U\Lambda U^\top$.
$p_i(K)$	Spectral occupancy, $p_i(K) = \lambda_i(K)/N$.
$p(K)$	Spectral occupancy distribution $(p_1(K), \ldots, p_N(K))$ on the $(N-1)$-simplex.
$H(p)$	Shannon entropy of p, $H(p) = -\sum_i p_i \log p_i$.
$d_{\mathrm{spec}}(K)$	Effective spectral dimension, $d_{\mathrm{spec}}(K) = \exp(H(p(K)))$. Continuous; ranges over $[1, \mathrm{rank}(K)]$.
$d_{\mathrm{spec}}(\eta)$	Realized d_{spec} along the η-controlled solution branch, $d_{\mathrm{spec}}(\eta) = d_{\mathrm{spec}}(K(L^\star(\eta)))$.
$d^\star_{\mathrm{spec}}$	Target value of d_{spec} specified by the user; η is calibrated so that $d_{\mathrm{spec}}(\eta) \approx d^\star_{\mathrm{spec}}$.
$\mathrm{PR}(K)$	Participation ratio, $\mathrm{PR}(K) = 1/\sum_i p_i(K)^2$.
$k_\tau(K)$	Thresholded active rank, $k_\tau(K) = \lvert\{i : \lambda_i(K) \geq \tau\,\lambda_1(K)\}\rvert$.
Spectral prefixes	
d_{pre}	Integer prefix size used for prefix-based visualization and analysis; $1 \leq d_{\mathrm{pre}} \leq r$.
$K_{d_{\mathrm{pre}}}$	Rank-d_{pre} spectral prefix of K, $K_{d_{\mathrm{pre}}} = U_{1:d_{\mathrm{pre}}}\Lambda_{1:d_{\mathrm{pre}}}U^\top_{1:d_{\mathrm{pre}}}$. EYM-optimal among rank-d_{pre} PSD approximations of K (Thm. 2).
π_{ij}	Soft membership of node i in prefix mode j, $\pi_{ij} \propto q_j[i]^2$, normalized to a row-simplex.
Training objective and capacity control	
$\eta \in \mathbb{R}$	Spectral-entropy weight (Lagrange-style coefficient) controlling realized d_{spec}.
$\ell(L, \boldsymbol{a}, \beta)$	Sampled link-prediction negative log-likelihood.
$R(L, \boldsymbol{a}, \beta)$	Quadratic regularization on parameters.
$F_\eta(L, \boldsymbol{a}, \beta)$	Entropy-regularized objective, $F_\eta = \ell - \eta \log d_{\mathrm{spec}}(K(L)) + R$.
$L^\star(\eta)$	Local minimizer of F_η along a fitted branch.
$K^\star(\eta)$	Trace-normalized kernel at $L^\star(\eta)$.
$\widetilde{\Theta}$	Gauge-fixed slice of the parameter space used to state second-order conditions.
Calibration of η	
τ	Relative tolerance of the bisection stopping criterion, $\lvert d_{\mathrm{spec}}(\eta) - d^\star_{\mathrm{spec}}\rvert / d^\star_{\mathrm{spec}} \leq \tau$.
$\eta^\star$	Calibrated entropy weight selected by bisection.
$[\eta_{\min}, \eta_{\max}]$	Bracket interval used in bisection; $W = \eta_{\max} - \eta_{\min}$.
δ	Target η-precision in the complexity statement.
Evaluation	
d	Nominal embedding dimension at which baselines are trained (matched to SPECTRA's rank cap $r = d$ in Table 1).
TLL	Test log-likelihood.
trNLL	Training negative log-likelihood.
GenGap	Generalization gap, $\mathrm{GenGap} = \mathrm{trNLL} + \mathrm{TLL}$.
S	Number of random seeds per experimental cell.

Table 4: Dataset statistics for eight undirected graphs used in the experiments.

Statistic	ca-GrQc	ca-HepTh	socfb-American75	socfb-Amherst41	bio-grid-human	bio-grid-worm	inf-power	inf-openflights
Domain	Collaboration	Collaboration	Social	Social	Biological	Biological	Infrastructure	Infrastructure
Nodes	4,158	8,638	6,370	2,235	9,186	3,343	4,941	2,905
Edges	13,426	24,818	224,024	93,188	40,224	9,780	11,534	18,550

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 contributions enumerated in the abstract and introduction (the spectral-prior formulation, the capacity-frontier analysis, and the empirical evaluation against latent-graph baselines) correspond directly to the methodological and experimental sections of the paper.

   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: Limitations are discussed in the final paragraph of the manuscript (Conclusion section).

   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 theorems and proofs are stated in the main paper and reproduced with full assumptions and complete derivations in the supplementary material.

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: A dedicated experimental setup section is provided in the supplementary material, and the main paper specifies the optimization, regularization, and evaluation protocol used in all experiments.

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: Both the code and the datasets with the specific train/test splits are submitted as supplementary material, and they will be made publicly available once the anonymity period ends.

   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: Training and test details, including data splits, optimizer, learning rate, regularization weight, and number of iterations, are reported in the main paper and expanded in the supplementary experimental setup.

   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: All figures report mean $\pm$ one standard deviation across three random seeds, capturing variability over model initialization and negative-edge sampling.

   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 supplementary experimental setup specifies the hardware used (NVIDIA A100 GPU when available, otherwise an Apple M2 machine with 8 GB RAM) along with the optimization budget per run.

   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: We have reviewed the NeurIPS Code of Ethics and confirm that the research complies with it; the work uses only publicly available network datasets and involves no human subjects.

   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: Potential positive and negative societal impacts of the proposed graph representation framework are discussed in the Conclusion section.

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 a graph representation learning model trained on publicly available network datasets and poses no foreseeable misuse risk of the kind addressed by this question.

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: All datasets and baseline implementations used in the paper are cited to their original sources, and we use them in accordance with their terms of use.

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: We release our codebase together with the train/validation/test splits used in all experiments; the supplementary archive contains a README describing the environment, dependencies, and commands needed to reproduce the reported results.

    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: The paper does not involve crowdsourcing or research with human subjects; all experiments are conducted on publicly available network datasets.

    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: The paper does not involve human subjects research, so no IRB approval or equivalent review was required.

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: [Yes]

Justification: LLM usage was declared on the submission page. LLMs were used for writing, editing, coding, and formatting assistance, but were not important, original, or non-standard components of the core methodology. All scientific content was independently verified by the authors.

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.