WHY DON'T YOU TAKE A WALK? RANKING AND RISK WITH WALK AVERAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Random walks can improve link prediction without adding information to the observed graph. We study when this occurs at fixed embedding width. In a degree-corrected graphon model, we derive conditions for walk averaging to improve ranking or probability loss, first for a fixed spectral embedding. The conditions depend on the interaction between retained and omitted structure, with different weights for the two metrics. Under weak contrast and a separated rank cutoff, we identify the bounded, rank-constrained skip-gram with negative sampling (SGNS) optimum and apply the metric criteria to its decoded predictions. We also characterize the probability error removable by monotone calibration. Population calculations recover the predicted regimes. On sampled graphs, optimization of the expected SGNS objective reproduces opposite ranking responses in a controlled setting, but the smaller gain is not recovered with a finite sampled corpus. Real-network results motivate the analysis but do not establish a window-selection rule.

1 INTRODUCTION

Network embeddings learn vectors for nodes such that inner products encode their similarity. LINE and DeepWalk begin with the same graph and end with low-dimensional node vectors, but they train those vectors on different pairs. LINE samples observed edges, whereas DeepWalk samples nodes that occur near one another on random walks (Tang et al., 2015; Perozzi et al., 2014). A walk reveals no new edges and can pair nodes that are not adjacent. Nevertheless, replacing one-step pairs by walk co-occurrences can improve the ranking of missing edges. We study when this change helps at a fixed embedding width, and when it hurts.

These methods differ in two ways that must be analyzed separately. A walk changes which node pairs the model is trained to represent. The width constraint then limits how well the model can represent those pairs. Averaging powers of a transition matrix describes the first change: it preserves eigenvectors and changes their weights. It does not resolve the second, because minimizing a rank-constrained logistic loss need not give a spectral truncation of the walk matrix or its entrywise logarithm.

Without a width constraint or sampling error, one-step training already suffices in our population model: its corrected skip-gram with negative sampling (SGNS) prediction is the true edge probability. No score can improve on the ranking of those probabilities. Walking therefore adds no information; whatever it does for prediction, it does through what the width constraint leaves out. A fixed-width model may leave out some edge structure. Walking can improve prediction if its change to the represented part compensates for what was left out. It can also make that error worse.

1.1 OUR CONTRIBUTIONS

In this paper we give conditions for both outcomes in a degree-corrected graphon model, starting by studying spectral embeddings. Theorem 1 compares AUC and negative log-likelihood (NLL) as walk weights change. To leading order, walking shrinks the represented component. This can help when it offsets the error from omitted structure. The two metrics weight that error differently and can therefore favor different walk distributions. Under the graph's own measure the retained and omitted components are orthogonal, so any benefit exists only through the metric's weighting of pairs.

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

Proposition 1 then identifies the probability error that a monotone scalar correction removes without changing AUC.

Theorem 2 shows that these criteria apply to the global bounded, fixed-rank SGNS optimum, not just a spectral approximation. It also identifies the next-order change in the directions represented by the embedding. Theorem 3 gives conditions under which fits on sampled graphs recover this population score. Our experiments test the formulas and their transfer separately. The positive ranking effect survives in some expected-objective fits, but not at every tested graph size, contrast, or corpus budget. We do not claim to predict the preferred window on an arbitrary observed network. Our study of SGNS and optimization is simplified, in that we do not consider design choices such as degree-adjusted sampling.

1.2 Related work

Prior work studies both the design of walks and the matrices they represent. Node2vec changes the transition law, Walklets separates path lengths, and Watch Your Step learns weights across lengths (Grover and Leskovec, 2016; Perozzi et al., 2017; Abu-El-Haija et al., 2018). Spectral transformations have also been learned directly for link prediction (Kunegis and Lommatzsch, 2009). Matrix-factorization analyses identify the targets of negative-sampling embeddings (Levy and Goldberg, 2014; Landgraf and Bellay, 2017; Qiu et al., 2018), and InfiniteWalk studies the large-window limit (Chanpuriya and Musco, 2020). Walk-based similarity scores were compared for link prediction by Liben-Nowell and Kleinberg (2007). Recovery and optimization guarantees for DeepWalk and node2vec address block models (Zhang and Tang, 2024; Harker and Bhaskara, 2024); Zhang and Tang (2024) also show that longer windows aid estimation on sparse graphs. Our question is the separate one of approximation at fixed width, and Section 5.2 shows both effects. We compare ranking and probability loss after imposing the width constraint on the SGNS objective itself.

Fixed-rank SGNS optimization and graphon embedding asymptotics provide related methods (Fonarev et al., 2017; Davison and Austern, 2023). Learning a ranking and then calibrating probabilities is established as well (Menon et al., 2012; Bröcker, 2009; Guo et al., 2017). Our calibration results describe which walk-dependent errors remain after correction; calibration itself is not a new algorithm here. Appendix G gives a fuller comparison, including graph-specific calibration.

2 Model and evaluation

A graphon specifies edge probabilities before a finite graph is drawn. This lets us separate an error in representing those probabilities from an error caused by sampling. We first define the graphon, the walk pairs, and the evaluation task. The additional assumptions needed for SGNS fitting and sampling appear with those results.

Definition 1 (Graphon and sampled graph)**.** On a probability space $(\mathcal{X}, \zeta)$, a graphon is a symmetric measurable function $B : \mathcal{X}^2 \to [0,1]$. Draw $X_1, \dots, X_n \overset{\text{iid}}{\sim} \zeta$ and, conditional on these positions, independent edges $A_{ij} \sim \mathrm{Bern}(B(X_i, X_j))$ for $i < j$, with $A_{ji} = A_{ij}$ and $A_{ii} = 0$.

Assumption 1 (Degree-corrected family)**.** Let $0 < w_{\min} \le w(x) \le w_{\max} < \infty$, with $\int w \, d\zeta = 1$, and set $d\mu = w \, d\zeta$. For fixed $v > 0$, let $B_0(x,y) = vw(x)w(y)$ lie in a fixed interval inside $(0,1)$. Let k be a fixed bounded symmetric kernel satisfying $\int k(x,y) \, d\mu(y) = 0$ almost everywhere. We study

$$B_\eta(x,y) = B_0(x,y)\{1 + \eta k(x,y)\}, \qquad \eta \downarrow 0. \tag{1}$$

The function w describes degree variation, while k describes the remaining edge structure. Its centering condition gives $d_\eta(x) = vw(x)$ and mean edge probability v. Thus μ is the stationary node distribution, and a vertex at x has conditional expected degree $(n-1)vw(x)$. Training and future-edge prediction use the same B_η.

The decomposition alone is not restrictive. Any graphon with positive degree function d can be written as $B = (d \otimes d)/v \, (1 + \eta k)$ by taking $v = \int d \, d\zeta$ and $w = d/v$. Assumption 1 requires more: the degree baseline must lie inside $(0,1)$ and the normalized residual must be small.

Definition 2 (Walk pairs)**.** The *lag* of a pair is the number of transitions separating its endpoints on a random walk. A lag distribution $\omega = (\omega_\ell)_{\ell \ge 1}$ assigns their sampling weights. We fix a finite family

of finitely supported distributions with $\omega_1 > 0$, including the one-step law. A uniform window T has $\omega_\ell = 1/T$ for $1 \le \ell \le T$. Write $t = 1 - \omega_1$ for the mass assigned to longer lags; the subscript 1 denotes one step alone.

The operator K acts by $(Kh)(x) = \int k(x,y)h(y)\,d\mu(y)$; let $J = \mathbf{1} \otimes \mathbf{1}$. Relative to the stationary measure μ, a single transition has density $1 + \eta k$. Since $JK = KJ = 0$, averaging walk lengths gives the stationary co-occurrence ratio

$$R_{\omega,\eta} = J + \sum_{\ell \ge 1} \omega_\ell \eta^\ell K^\ell. \tag{2}$$

We use the same notation for an operator and its kernel, and write $(a \otimes c)(x,y) = a(x)c(y)$. Equation (2) changes each residual eigenvalue by a polynomial; scaling by ω_1 is only the first-order effect. The notation $O_\infty(\eta^j)$ means a remainder bounded uniformly by a constant times η^j.

Definition 3 (Prediction and evaluation)**.** Let $e = (x,y) \sim \zeta \otimes \zeta$, $p(e) = B_\eta(e)$, and $Y \mid e \sim \mathrm{Bern}(p(e))$. For independent $e^+ \sim (e \mid Y = 1)$ and $e^- \sim (e \mid Y = 0)$, define

$$\mathrm{AUC}_\eta(q) = \Pr\{q(e^+) > q(e^-)\} + \tfrac{1}{2}\Pr\{q(e^+) = q(e^-)\}, \tag{3}$$

$$\mathcal{R}_\eta(q) = \mathrm{NLL}_\eta(q) - \mathrm{NLL}_\eta(p) = \mathbb{E}_{\zeta^2}\,\mathrm{KL}\big(\mathrm{Bern}(p) \| \,\mathrm{Bern}(q)\big), \tag{4}$$

where $\mathrm{NLL}_\eta(q) = \mathbb{E}[-p \log q - (1-p)\log(1-q)]$. Both metrics evaluate the same probability output $q \in (0,1)$.

AUC measures ordering; NLL also measures the numerical probabilities. We call a prediction *native* when it uses the specified decoder without a fitted calibration map. Degree correction belongs to the decoder and can change rankings because it depends on the endpoints. Calibration applies the same scalar map to every pair. A strictly increasing map preserves AUC, including ties.

3 Walk averaging and probability correction

We begin with a representation whose spectral directions are fixed. Walking changes their weights, but does not change the edge probabilities to be predicted. We first compare the resulting ranking and probability errors, then determine which part of the probability error can be removed by calibration.

3.1 A fixed spectral representation

For a finite graph with adjacency A and degree matrix D, a spectral embedding retains selected eigenpairs of a symmetric matrix such as $D^{-1/2}AD^{-1/2}$. We use the analogous construction for the graphon operator. Separate input and context vectors allow negative eigenvalues, as in the generalised random dot product graph (Rubin-Delanchy et al., 2022); a single Euclidean Gram matrix would not.

Definition 4 (Retained and omitted structure)**.** Choose $m \ge 1$ nonzero eigenvalues λ_j of K with orthonormal eigenfunctions $u_j \in L^2(\mu)$. Put

$$f = \sum_{j=1}^{m} \lambda_j u_j \otimes u_j, \qquad r = k - f, \qquad \mathsf{R} = K - \mathsf{F}, \tag{5}$$

where F is the operator with kernel f. The width $d = m + 1$ includes the constant direction. The spectral walk prediction is

$$s_{\omega,\eta} = B_0\left(1 + \sum_{j=1}^{m}\sum_{\ell \ge 1} \omega_\ell(\eta\lambda_j)^\ell u_j \otimes u_j\right) = B_0(1 + \eta\omega_1 f) + O_\infty(\eta^2). \tag{6}$$

The omitted kernel r is still part of the true edge probability. Nothing in the model declares it to be noise or label-irrelevant structure. No gap between retained and omitted eigenvalues is needed for (6).

3.2 RANKING AND PROBABILITY LOSS

The following comparison uses only the first-order prediction in (6). We state this as an assumption for now. Theorem 2 will show that the constrained SGNS optimum satisfies it.

Assumption 2 (Leading prediction)**.** For the retained component in Definition 4, the probability predictions satisfy $q_{\omega,\eta} = B_0(1+\eta\omega_1 f)+O_\infty(\eta^2)$ uniformly over the fixed lag family in Definition 2.

Assumption 3 (Regularity for ranking)**.** For $e \sim \zeta^2$, the baseline $Z = B_0(e)$ has density $h \in L^2$. The conditional law of $(f(e), r(e))$ given $Z = z$ has a version that is weakly continuous for almost every z.

The four coefficients below measure retained structure and its interaction with omitted structure. NLL uses a weighted product over candidates; AUC uses variation among candidates with the same degree baseline:

$$A_R = \mathbb{E}\left[\frac{Z}{1-Z} f^2\right], \qquad C_R = \mathbb{E}\left[\frac{Z}{1-Z} fr\right], \tag{7}$$

$$A_A = 2\int h(z)^2 z^2 \operatorname{Var}(f \mid Z = z)\, dz, \qquad C_A = 2\int h(z)^2 z^2 \operatorname{Cov}(f, r \mid Z = z)\, dz. \tag{8}$$

Only candidates with nearly equal baseline probabilities can reverse order under a small perturbation. This is why the AUC coefficients use conditional variances and covariances. Assumption 3 makes this local comparison well defined.

Theorem 1 (Ranking and probability loss)**.** *Under Assumptions 1 and 2, for fixed $t = 1 - \omega_1$ as $\eta \downarrow 0$,*

$$\mathcal{R}_\eta(q_{\omega,\eta}) - \mathcal{R}_\eta(q_{1,\eta}) = \frac{\eta^2}{2}(t^2 A_R + 2tC_R) + O(\eta^3). \tag{9}$$

If Assumption 3 also holds, then

$$\mathrm{AUC}_\eta(q_{\omega,\eta}) - \mathrm{AUC}_\eta(q_{1,\eta}) = -\frac{\eta^2}{4v(1-v)}(t^2 A_A + 2tC_A) + o(\eta^2). \tag{10}$$

Theorem 1 gives a sign criterion for each metric. The coefficients A_R, A_A are nonnegative, while C_R, C_A may have either sign. In both formulas, a negative quadratic means improvement for sufficiently small contrast. For $A > 0$ and $t > 0$, this is equivalent to $C < -tA/2$. When $C < 0$, the leading benefit ends at $t = -2C/A$ if that point lies in $0 \le t < 1$. If the leading coefficient is zero, the theorem does not decide the sign.

The reason is visible in the error itself:

$$q_{\omega,\eta} - B_\eta = -\eta B_0(r + tf) + O_\infty(\eta^2). \tag{11}$$

At one step, the error comes from r. Walking adds tf. If f points against the omitted error under the metric's weighting, some shrinkage helps; if it points in the same direction, shrinkage hurts. For $A > 0$, completing the square gives

$$t^2 A + 2tC = A\left(t + \frac{C}{A}\right)^2 - \frac{C^2}{A}. \tag{12}$$

This quadratic is minimized at $t = -C/A$, subject to the allowed walk distributions. AUC and NLL can prefer different amounts of walking because they weight the same prediction error differently.

The NLL formula follows by expanding Bernoulli divergence. For AUC, we integrate the comparisons whose baseline scores differ by only $O(\eta)$. Appendix A gives both calculations.

3.3 THE COST OF CALIBRATION

A score can order pairs well and still assign poor probabilities. We now ask which errors can be corrected by applying the same increasing scalar map to every prediction. Such a map preserves rankings, but cannot distinguish pairs given the same score. Write $q_{t,\eta} = q_{\omega,\eta}$ for the chosen lag law with $t = 1 - \omega_1$.

Assumption 4 (Regularity for correction)**.** For each fixed t, $m_t(z) = \mathbb{E}[r + tf \mid Z = z]$ admits a bounded Lipschitz extension to a fixed interior probability interval containing the predictions.

Proposition 1 (A smooth scalar correction)**.** *Under Assumptions 1, 2, and 4, define*

$$g_{\eta,a}(q) = \sigma\{\mathrm{logit}(q) + \eta a(q)\}, \quad a_t^*(z) = \frac{m_t(z)}{1-z}, \quad c_{t,\eta} = g_{\eta,a_t^*}(q_{t,\eta}), \tag{13}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$. For sufficiently small η, this map is strictly increasing and preserves AUC. It minimizes the leading NLL coefficient among fixed bounded Lipschitz corrections a, with

$$\mathcal{R}_\eta(c_{t,\eta}) = \frac{\eta^2}{2}\mathbb{E}\left[\frac{Z}{1-Z}\,\mathrm{Var}(r + tf \mid Z)\right] + O(\eta^3), \tag{14}$$

$$\mathcal{R}_\eta(q_{t,\eta}) - \mathcal{R}_\eta(c_{t,\eta}) = \frac{\eta^2}{2}\mathbb{E}\left[\frac{Z}{1-Z}m_t(Z)^2\right] + O(\eta^3). \tag{15}$$

Proposition 1 removes the average error at each baseline probability. Differences between candidates with that baseline remain. Applying the correction to the one-step prediction as well gives

$$\mathcal{R}_\eta(c_{t,\eta}) - \mathcal{R}_\eta(c_{0,\eta}) = \frac{\eta^2}{2}(t^2\widetilde{A}_R + 2t\widetilde{C}_R) + O(\eta^3), \tag{16}$$

where

$$\widetilde{A}_R = \mathbb{E}\left[\frac{Z}{1-Z}\,\mathrm{Var}(f \mid Z)\right], \qquad \widetilde{C}_R = \mathbb{E}\left[\frac{Z}{1-Z}\,\mathrm{Cov}(f, r \mid Z)\right]. \tag{17}$$

Calibration therefore need not make walking better in NLL. Corrected NLL and AUC use the same conditional covariance, but weight it differently; Figure 1 shows their different break-even points. A restricted calibration family can remove less error. Appendix B.5 derives its coefficients. These comparisons concern population corrections, not maps estimated from finitely many labels.

4 From spectral predictions to SGNS

The spectral calculation does not yet tell us what SGNS learns. We now show that its constrained population optimum has the first-order prediction required by Theorem 1. We then connect this optimum to the factor constraints used in the experiments and to fits on sampled graphs.

4.1 Population objective

SGNS learns to distinguish walk pairs from negative samples. We define that loss first, then specify how its score is converted to an edge probability.

Definition 5 (Fixed-width SGNS)**.** Skip-gram with negative sampling assigns input and context vectors $U(x), V(y) \in \mathbb{R}^d$ and a score $F(x, y) = U(x)^\top V(y)$. With $b > 1$ negative contexts drawn from μ for each stationary positive pair, the normalized population loss is

$$\mathcal{L}_{\omega,\eta}(F) = \frac{1}{1+b}\int\int \{R_{\omega,\eta}\,\mathrm{softplus}(-F) + b\,\mathrm{softplus}(F)\}\,d\mu\,d\mu, \tag{18}$$

where $\mathrm{softplus}(z) = \log(1 + e^z)$. Its native probability decoder is $q = B_0 b \exp(F)$.

The labels in (18) distinguish walk pairs from negative samples, not edges from nonedges. Without a rank constraint, the loss is minimized entrywise by $\log(R_{\omega,\eta}/b)$ (Levy and Goldberg, 2014; Landgraf and Bellay, 2017; Qiu et al., 2018), giving native prediction B_η at one step. This is also the LINE second-order objective when the negative distribution is matched to μ. Original DeepWalk uses hierarchical softmax; the results here concern the negative-sampling variant.

Assumption 5 (Stable rank-constrained objective)**.** Use the objective of Definition 5 and the retained modes of Definition 4. Set $d = m + 1$, $L = \log(3b/2)$, and minimize over all kernels with rank at most d and $|F| \le L$. Assume

$$\|\mathsf{R}\|_{\mathrm{op}} < \tfrac{1}{2}\min_{1\le j\le m}|\lambda_j|. \tag{19}$$

This separates the retained spectrum by absolute magnitude. It allows signed eigenvalues, repeated retained eigenvalues, and an infinite omitted spectrum. The score bound is not a bound on each factor coordinate.

4.2 MINIMIZING SCORE

The minimizing directions need not stay fixed as walk weights change. Let P project onto $\operatorname{span}\{\mathbf{1}, u_1, \dots, u_m\}$ and $Q = I - P$ onto its orthogonal complement. Define the kernel M_b below and, for any bounded symmetric kernel M, the operator $\mathcal{S}_K(M)$:

$$M_b = \tfrac{1}{2} f^2 + \frac{rf}{1+b}, \qquad z_j = \lambda_j(\lambda_j I - \mathsf{R})^{-1} Q M u_j,$$

$$\mathcal{S}_K(M) = PMP + (QM\mathbf{1}) \otimes \mathbf{1} + \mathbf{1} \otimes (QM\mathbf{1}) + \sum_{j=1}^{m} (z_j \otimes u_j + u_j \otimes z_j). \tag{20}$$

Products in M_b are pointwise; F^2 below denotes kernel composition. The first term of $\mathcal{S}_K$ changes coefficients within the retained space, and its other terms change that space itself.

Theorem 2 (The fixed-width SGNS minimum)**.** *Under Assumptions 1 and 5, for the fixed lag family in Definition 2 and all sufficiently small $\eta > 0$, the bounded rank-d problem has a unique minimizing score, up to equality almost everywhere. Competitors may be nonsymmetric. In uniform norm,*

$$F_{\omega,\eta} = -\log b\, J + \eta\omega_1 f + \eta^2\{\omega_2 \mathsf{F}^2 - \omega_1^2 \mathcal{S}_K(M_b)\} + O(\eta^3). \tag{21}$$

There is $c > 0$, independent of sufficiently small η, such that every feasible H satisfies $\mathcal{L}(H) - \mathcal{L}(F_{\omega,\eta}) \geq c\|H - F_{\omega,\eta}\|^2_{L^2(\mu^2)}$.

Theorem 2 supplies the connection to prediction: exponentiating (21) gives $q_{\omega,\eta} = B_0(1 + \eta\omega_1 f) + O_\infty(\eta^2)$, which verifies Assumption 2. The second-order terms separate walk composition, $\omega_2\mathsf{F}^2$, from constrained logistic fitting, $\mathcal{S}_K(M_b)$. Through the product rf, omitted structure can change both the weights and the directions of the fitted representation. Under Assumption 3, the first-order prediction already determines the leading metric comparison. With constant degrees, the second-order change can determine the leading ranking comparison instead.

Appendix C constructs the stationary score with a moving subspace and certifies it as the global bounded rank-d minimum.

Corollary 1 (Factor-norm constraints)**.** *Under Assumptions 1 and 5, the minimizing score admits width-d factors with $\sup_x \|U_{\omega,\eta}(x)\|^2, \sup_y \|V_{\omega,\eta}(y)\|^2 \leq \log b + C\eta$. For sufficiently small contrast it therefore also minimizes over the row-norm class $\|U(x)\|^2, \|V(y)\|^2 \leq L$.*

Corollary 1 covers the factor constraint used in the experiments. The constant component needs squared norm $\log b$, leaving a margin $L - \log b = \log(3/2)$ for the other components at small contrast. The corollary does not assert that the constraints are inactive on a sampled graph or at arbitrary fixed contrast. Appendix C.4 gives the factorization.

The calibration criterion in Proposition 1 applies to this SGNS optimum. At constant degrees an affine rescaling of the spectral prediction, or a power rescaling of the SGNS prediction, restores the retained coefficient and leaves only an $O(\eta^4)$ difference between calibrated windows (Propositions B.1 and B.2). The cancellation uses orthogonality of the second-order correction to r, including when the subspace moves; it does not fix the sign of that remainder.

4.3 SAMPLING THE GRAPH AND CORPUS

We finally ask when a finite fitted score approaches the population optimum. The next theorem requires its training loss to approach the best loss in the constrained class.

Assumption 6 (Sampled training)**.** Let $\mathcal{X} = [0,1]^p$, let ζ have a density bounded above and below, and let w, k be continuous. Fix a sufficiently small contrast, the width, cap, and lag family. Sample the graph as in Definition 1. Conditional on it, use N independent stationary training groups: each is a walk pair with negatives, or a whole trajectory with normalized context weights inducing ω. Use the observed stationary law for negatives and degree correction. Assume $N/[n\log(nN)] \to \infty$ and empirical SGNS loss within $o_P(1)$ of the best score in the bounded rank-d class.

Theorem 3 (Transfer to sampled fits)**.** *Under Assumptions 1, 5, and 6,*

$$\frac{1}{n^2}\sum_{i,j}|\widehat{F}_{ij} - F_{\omega,\eta}(X_i, X_j)|^2 \longrightarrow 0 \quad \textit{in probability}. \tag{22}$$

Degree-corrected predictions, clipped to a fixed interval whose interior contains the population range, and their native risks for independent future edges converge. AUC also converges if population-score ties have equal target probabilities almost surely. Every strict comparison in the fixed lag family then retains its sign with probability tending to one.

Theorem 3 transfers every strict population comparison once sampling and fitting errors vanish. The proof first bounds the difference between sampled and population losses, then uses the quadratic gap in Theorem 2 to control the scores. It counts independent trajectories, not their overlapping pairs. The result neither covers sparse graphs nor specifies a finite sample size that resolves a small gain. Appendix D also treats descent from a consistent reference fit.

5 Experiments and scope

Population fits check the metric formulas, sampled graphs test whether the population directions survive sampling, and real networks show how walking behaves outside the proved regime. More detail and complete tables are in Appendix E.

5.1 Population metric comparisons

Setup. We use two graphons that share the degree function, the retained mode, and the spectrum, and differ only in the sign of the omitted part. Let μ be uniform on $[-1, 1]$, $d\zeta = (1 + .6x)d\mu$, $w = (1 + .6x)^{-1}$, and $v = .12$. For normalized Legendre functions ϕ_j, set

$$f = \phi_1 \otimes \phi_1, \qquad r_- = .2(\phi_3 \otimes \phi_3 - \phi_2 \otimes \phi_2), \qquad r_+ = -r_-.$$

The residual eigenvalues are $1, .2, -.2$, so the width-two model satisfies the separation condition. We solve the population SGNS problem at one step and at lag laws $(1 - t, t)$ for $t \in \{.01, .04, .10, .18\}$, at contrast $\eta = 5 \cdot 10^{-4}$ on a 128-node quadrature grid, and evaluate native NLL, NLL after the 12-knot monotone calibration fitted at every window, and AUC by a continuous conditional integral. Two larger contrasts and a coarser grid are in Appendix E.1.

Results. Table 1 compares the measured changes with the quadratics of Theorem 1 and Proposition 1. All 72 sign comparisons at the finer grid agree with the predictions, and the magnitudes agree to a few percent; these are checks within two graphon families, not 72 independent experiments. In the r_- family, increasing t produces the four regimes in order: all metrics improve; calibration repairs an increase in native NLL; AUC improves while both losses worsen; and all metrics worsen. The r_+ control favors one step at every mass. Figure 1 shows the three break-even points: about .018 for native NLL, .051 after calibration, and .139 for AUC. Two further population checks, of the optimizer expansion and of the fourth-order calibrated difference at constant degrees, are in Appendix F.

5.2 SGNS on sampled graphs

Setup. Here we ask whether an SGNS fitted to a sampled graph reproduces population signs. We use two four-block graphons with varying degrees and mean edge probability .06, again differing only in the sign of the omitted modes, and call them D and C. At contrast .3, the population optimum, computed with the same solver rather than the expansion, loses AUC under walking for D and gains it for C; both losses worsen for both. We sample graphs of 800, 1600, and 3200 nodes, eight per setting, fit width-two norm-constrained SGNS to the expected objective of each graph at $t \in \{0, .05, .10, .20\}$, and compare each fit with the one-step fit on the same graph against the true edge probabilities. A control replaces the expected objective with a frozen corpus of two million walk pairs.

Results. Table 2 gives the AUC changes at $n = 3200$. Both population directions are recovered in all eight draws at $t = .05$ and .10: D at 58% of the population magnitude, C at 11%. For C, agreement rises with graph size, from $5/8$ at $n = 800$ and 1600 to $8/8$ at 3200 for $t = .05$, and from

Table 1: Matched graphon SGNS comparisons at contrast $\eta = 5 \cdot 10^{-4}$ and grid size 128. Entries are measured changes divided by η^2, multiplied by 10^4; predictions are in parentheses. AUC is better when positive; losses are better when negative. The same 12-feature monotone calibration family is fitted at both windows. These are population calculations, not sampled-graph fits.

Omitted part	t	ΔAUC	ΔNLL	Δcal. NLL
r_-	0.01	+1.063 (+1.056)	-0.066 (-0.065)	-0.094 (-0.093)
r_-	0.04	+3.265 (+3.237)	+0.687 (+0.691)	-0.104 (-0.098)
r_-	0.10	+3.230 (+3.165)	+6.469 (+6.482)	+1.108 (+1.123)
r_-	0.18	-6.023 (-6.132)	+23.046 (+23.079)	+5.278 (+5.304)
r_+	0.01	-1.226 (-1.220)	+0.224 (+0.223)	+0.141 (+0.138)
r_+	0.04	-5.888 (-5.866)	+1.845 (+1.845)	+0.836 (+0.827)
r_+	0.10	-19.649 (-19.592)	+9.363 (+9.366)	+3.456 (+3.436)
r_+	0.18	-47.194 (-47.094)	+28.254 (+28.269)	+9.499 (+9.468)

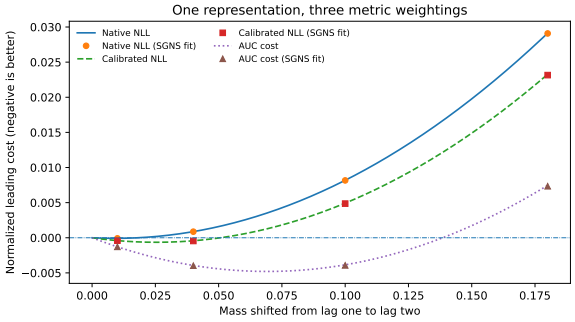


Figure 1: The r_- graphon has different break-even points for AUC, native NLL, and calibrated NLL. Curves show $t^2 + 2(C/A)t$; negative cost means improvement. Markers are optimized population SGNS results at $\eta = 5 \cdot 10^{-4}$. The calibrated curve uses the same finite feature family as the fits.

$3/8, 3/8$ to $8/8$ for $t = .10$. The gain for C is not recovered at $t = .20$, under the sampled corpus, or at contrast $.4$, where the population optimum itself sits on the score cap, outside the interior regime of Corollary 1. All three loss signs agree with the population from $n = 1600$ for both graphons. At $n = 800$, native NLL often improves where the population worsens, consistent with the estimation benefit of longer walks (Zhang and Tang, 2024) that the population comparison lacks. The graphs are dense and the fits sit on the norm bound; the study tests the population signs, not the expansions themselves.

5.3 Real networks

Setup. We fit norm-constrained SGNS on nine real networks: Karate, Dolphins, Les Misérables, Polbooks, and Football (34 to 115 nodes), and `email-Eu-core`, `ca-GrQc`, `p2p-Gnutella08`, and `ca-HepTh` (986 to 8,638 nodes). Node pairs are split, independently of their labels, into training, diagnostic, calibration, and test sets; the score has width sixteen, five stationary negatives, a frozen walk corpus, and factor-row radius $\sqrt{\log(7.5)}$, and a projected solver fits it to a stationary point. The small networks use three splits and 100,000 pairs; the larger ones five splits, 500,000 pairs, and the best of three restarts. Table 3 reports paired changes from one step on the test pairs, pooling the small networks and listing the larger ones separately.

Results. On the small networks AUC and native NLL improve on nearly every split at every lag law, and the gain per unit lag mass decreases as the mass grows, the diminishing-returns shape of the quadratic in Theorem 1; calibrated NLL is mixed, but the calibration sets there are small (Karate has about nine positive labels). On the larger networks the sign depends on the graph and on the metric. The collaboration graphs improve on everything at window five, but at window ten AUC and native NLL still improve on every split while calibrated NLL worsens on every split, and the AUC gain is halved. `email-Eu-core` worsens on average on every metric at every

Table 2: Sampled-graph SGNS: AUC change from one step at $\eta = .3$, $n = 3200$, and width two. Values are multiplied by 10^4; parentheses count draws (of eight) whose sign matches the population reference, which is computed with the same solver. Expected weights and two-million-pair corpora are separate experiments.

Graphon	t	Population	Expected (agree)	Corpus (agree)
D	0.05	-4.884	-2.839 (8/8)	-2.388 (8/8)
D	0.10	-10.325	-5.997 (8/8)	-6.314 (8/8)
D	0.20	-22.403	-13.336 (8/8)	-13.543 (8/8)
C	0.05	+3.966	+0.452 (8/8)	-0.250 (3/8)
C	0.10	+6.656	+0.591 (8/8)	-1.071 (1/8)
C	0.20	+7.864	-0.209 (3/8)	-2.743 (0/8)

Table 3: Norm-constrained SGNS on real networks: paired changes from one step on test pairs. The first block pools the five small networks over their fifteen graph–split pairs; the others are single graphs over five splits, best of three restarts. "Splits improving" counts splits where AUC rises or a loss falls. Means are multiplied by 10^4. Per-graph results and the remaining larger-network windows are in Appendix E.3.

		Splits improving			Mean change ($\times 10^4$)		
Graph	Window	AUC	Native	Cal.	ΔAUC	Δnative	Δcal.
Five small (34–115)	$(.96, .04)$	15/15	15/15	12/15	+83	-49	+21
Five small	$(.90, .10)$	15/15	14/15	11/15	+166	-86	+143
Five small	$T = 2$	14/15	14/15	10/15	+241	-132	+83
Five small	$T = 5$	14/15	14/15	11/15	+261	-202	-18
`ca-HepTh` (8,638)	$T = 5$	5/5	5/5	5/5	+46.6	-0.29	-0.12
`ca-HepTh`	$T = 10$	5/5	5/5	0/5	+23.2	-0.27	+0.07
`ca-GrQc` (4,158)	$T = 10$	5/5	5/5	0/5	+18.5	-0.33	+0.11
`email-Eu-core` (986)	$T = 2$	0/5	2/5	0/5	-9.11	+3.93	+9.56
`p2p-Gnutella08` (6,299)	$(.9, .1)$	3/5	0/5	0/5	+4.01	+0.08	+0.15

window. `p2p-Gnutella08` at the $(.9, .1)$ mixture worsens in native NLL on every split while AUC improves on average. The larger-network windows were not fixed in advance and the effects are small, so we read this as showing that both directions and metric disagreement occur on real networks, not that they can be predicted.

Limitations and conclusion. Our results describe bounded fixed-width population SGNS at weak contrast. They do not prove convergence of a default trainer or predict the preferred window on an arbitrary sparse graph. Calibration removes probability error only within the stated family. The experiments support the population formulas, recover opposite AUC responses on some sampled graphs while also showing failures of transfer, and find both directions on real networks. Walking helps when the change in the retained representation reduces the error left by the width constraint, as measured by the evaluation criterion.

REPRODUCIBILITY STATEMENT

Complete derivations are in Appendices A–D. Appendix E specifies every experiment in Section 5. Code for the experiments accompanies the submission.

AI USE STATEMENT

In this work, we used generative AI tools for proving some theorems, writing parts of code, and drafting sections. We have reviewed all AI-assisted work by checking proofs and code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

S. Abu-El-Haija, B. Perozzi, R. Al-Rfou, and A. A. Alemi. Watch Your Step: Learning node embeddings via graph attention. In *Advances in Neural Information Processing Systems*, volume 31, pages 9198–9208, 2018.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, and D. L. Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

A. Barot, S. Bhamidi, and S. Dhara. Community detection using low-dimensional network embedding algorithms. *arXiv:2111.05267*, 2021.

J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519, 2009.

S. Chanpuriya and C. Musco. InfiniteWalk: Deep network embeddings as Laplacian embeddings with a nonlinearity. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1325–1333, 2020.

A. Davison and M. Austern. Asymptotics of network embeddings learned via subsampling. *Journal of Machine Learning Research*, 24(138):1–120, 2023.

A. Fonarev, O. Grinchuk, G. Gusev, P. Serdyukov, and I. Oseledets. Riemannian optimization for skip-gram negative sampling. In *ACL*, 2017.

A. Grover and J. 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.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, PMLR 70:1321–1330, 2017.

C. Harker and A. Bhaskara. Convergence guarantees for the DeepWalk embedding on block models. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:17636–17656, 2024.

A. Jaffe, Y. Kluger, O. Lindenbaum, J. Patsenker, E. Peterfreund, and S. Steinerberger. The spectral underpinning of word2vec. *Frontiers in Applied Mathematics and Statistics*, 6:593406, 2020.

D. Karkada, D. J. Korchinski, A. Nava, M. Wyart, and Y. Bahri. Symmetry in language statistics shapes the geometry of model representations. *arXiv:2602.15029v3*, 2026.

J. Kunegis and A. Lommatzsch. Learning spectral graph transformations for link prediction. In *ICML*, pp. 561–568, 2009.

A. J. Landgraf and J. Bellay. word2vec skip-gram with negative sampling is a weighted logistic PCA. *arXiv:1705.09755*, 2017.

J. Lei. Network representation using graph root distributions. *Annals of Statistics*, 49(2):745–768, 2021.

O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *NIPS*, 2014.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007.

C. Lin, D. L. Sussman, and P. Ishwar. Ergodic limits, relaxations, and geometric properties of random walk node embeddings. *arXiv:2109.04526*, 2021.

A. K. Menon and R. C. Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

A. K. Menon, X. Jiang, S. Vembu, C. Elkan, and L. Ohno-Machado. Predicting accurate probabilities with a ranking loss. In *ICML*, 2012.

E. Nascimento, D. Mesquita, S. Kaski, and A. H. Souza. In-n-Out: calibrating graph neural networks for link prediction. *arXiv:2403.04605*, 2024.

B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: online learning of social representations. In *KDD*, 2014.

B. Perozzi, V. Kulkarni, H. Chen, and S. Skiena. Don't Walk, Skip!: Online learning of multi-scale network embeddings. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 258–265, 2017.

J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. Network embedding as matrix factorization: unifying DeepWalk, LINE, PTE, and node2vec. In *WSDM*, 2018.

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

T. Safavi, D. Koutra, and E. Meij. Evaluating the calibration of knowledge graph embeddings for trustworthy link prediction. In *EMNLP*, 2020.

J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: large-scale information network embedding. In *WWW*, 2015.

J. Xu. Rates of convergence of spectral methods for graphon estimation. In *ICML*, PMLR 80:5433–5442, 2018.

Z. Yang, M. Ding, C. Zhou, H. Yang, J. Zhou, and J. Tang. Understanding negative sampling in graph representation learning. In *KDD*, 2020.

Y. Zhang and M. Tang. A theoretical analysis of DeepWalk and node2vec for exact recovery of community structures in stochastic blockmodels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1065–1078, 2024.

A RANKING AND PROBABILITY COMPARISONS

We prove the two parts of Theorem 1 separately. The probability-loss calculation uses a Taylor expansion at each candidate pair. The AUC calculation instead integrates comparisons between two candidates whose baseline probabilities are close. All expectations use independent candidates from ζ^2 unless another measure is displayed.

A.1 PROBABILITY LOSS

Proof of the probability-loss part of Theorem 1. Write $p_\eta = B_\eta$ and $t = 1 - \omega_1$. Assumptions 1 and 2 give

$$p_\eta = B_0\{1 + \eta(f + r)\}, \qquad q_{\omega,\eta} = B_0(1 + \eta\omega_1 f) + O_\infty(\eta^2),$$

so

$$q_{\omega,\eta} - p_\eta = -\eta B_0(r + tf) + O_\infty(\eta^2).$$

For small contrast, both probabilities lie in one interval $[\epsilon, 1 - \epsilon]$ by Assumption 1. On this interval the cross-entropy $\ell(p, q) = -p \log q - (1 - p) \log(1 - q)$ has derivatives

$$\partial_q \ell(p, q) = \frac{q - p}{q(1 - q)}, \qquad \partial_q^2 \ell(p, q) = \frac{p}{q^2} + \frac{1 - p}{(1 - q)^2}.$$

At $q = p$, the first derivative vanishes and the second is $1/[p(1-p)]$. The third derivative is uniformly bounded on the interval. Taylor's theorem therefore gives

$$\begin{aligned} \mathrm{KL}(\mathrm{Bern}(p_\eta) \| \mathrm{Bern}(q_{\omega,\eta})) &= \frac{(q_{\omega,\eta} - p_\eta)^2}{2p_\eta(1 - p_\eta)} + O(|q_{\omega,\eta} - p_\eta|^3) \\ &= \frac{\eta^2 B_0}{2(1 - B_0)}(r + tf)^2 + O_\infty(\eta^3). \end{aligned}$$

In the second line, replacing p_η in the denominator by B_0 costs $O_\infty(\eta^3)$, since $p_\eta - B_0 = O_\infty(\eta)$. The second-order prediction remainder also contributes only $O_\infty(\eta^3)$. Integrating yields

$$\mathcal{R}_\eta(q_{\omega,\eta}) = \frac{\eta^2}{2} \mathbb{E}\left[\frac{B_0}{1 - B_0}(tf + r)^2\right] + O(\eta^3). \tag{23}$$

At one step $t = 0$. Subtracting that value and using $(r + tf)^2 - r^2 = t^2 f^2 + 2tfr$ proves (9). □

The weighting explains why spectral orthogonality does not force C_R to vanish. Under μ^2, the retained and omitted kernels are orthogonal, but NLL is evaluated under ζ^2. Changing measure and expanding the denominator gives

$$\begin{aligned} C_R &= v \int\int \frac{f(x, y)r(x, y)}{1 - vw(x)w(y)} \, d\mu(x)d\mu(y) \\ &= v \sum_{a \geq 0} v^a \int\int w(x)^a w(y)^a f(x, y)r(x, y) \, d\mu(x)d\mu(y). \end{aligned}$$

For finite-rank r, substitute the two spectral sums and separate the x and y integrals. This gives

$$C_R = v \sum_{a \geq 0} v^a \sum_{i \text{ retained}} \sum_{j \text{ omitted}} \lambda_i \lambda_j \left(\int w(x)^a u_i(x) u_j(x) \, d\mu(x) \right)^2. \tag{24}$$

For the infinite-rank case, Parseval's inequality gives, for each retained i,

$$\sum_{j \text{ omitted}} |\lambda_j| |\langle w^a u_i, u_j \rangle|^2 \leq \|\mathsf{R}\|_{\mathrm{op}} \|w^a u_i\|_2^2 \leq \|\mathsf{R}\|_{\mathrm{op}} w_{\max}^{2a}.$$

There are finitely many retained indices, and $vw_{\max}^2 < 1$. The sum is therefore absolutely convergent. Approximate r in Hilbert–Schmidt norm by its finite spectral sums to obtain the same identity.

At $a = 0$, each cross inner product in (24) is zero. Constant degree weights make every term zero. If all nonzero eigenvalues have the same sign, every summand is nonnegative. Opposite-signed modes can instead contribute negatively when they overlap after degree weighting. A positive-semidefinite residual therefore excludes this particular native-loss improvement mechanism. Positive semidefiniteness is not equivalent to homophily, which is not assumed here.

A.2 A LOCAL RANKING IDENTITY

For ranking, we first prove a local statement that does not refer to a graph. Let a candidate have baseline probability Z and bounded marks a, c. Assume that Z takes values in a compact subinterval of $(0, 1)$, has density $h \in L^2$, and that the conditional law of (a, c) given $Z = z$ is weakly continuous for almost every z. The baseline has no atoms, but the conditional marks may be discrete. In the limiting coefficient below, primed and unprimed marks are independent draws conditional on the same z.

Lemma 1 (Regret near a nondegenerate baseline)**.** *Let the true probability be $p_\eta = Z + \eta a$ and let the evaluated score be $s_\eta = Z + \eta c + O_\infty(\eta^2)$. Under the preceding conditions, with $\rho_0 = \mathbb{E}Z$,*

$$\mathrm{AUC}_\eta(p_\eta) - \mathrm{AUC}_\eta(s_\eta) = \frac{\eta^2}{4\rho_0(1-\rho_0)} \int h(z)^2 \mathbb{E}[((a-c) - (a'-c'))^2 \mid z] dz + o(\eta^2). \quad (25)$$

Proof. **Expressing AUC over candidate pairs.** Let e, e' be independent candidates, set $p = p_\eta(e)$ and $p' = p_\eta(e')$, and write $\rho_\eta = \mathbb{E}p$. Define $H_s(e, e')$ to be one, one half, or zero according as $s(e)$ is greater than, equal to, or less than $s(e')$. The class-conditional definition gives

$$\mathrm{AUC}_\eta(s) = \frac{\mathbb{E}[p(1-p')H_s(e,e')]}{\rho_\eta(1-\rho_\eta)}, \qquad H_s = \frac{1 + \operatorname{sgn}(s - s')}{2}.$$

Swapping e and e' in the signed term gives

$$\mathbb{E}[p(1-p')\operatorname{sgn}(s-s')] = \tfrac{1}{2}\mathbb{E}[(p-p')\operatorname{sgn}(s-s')].$$

We obtain the standard weighted ranking identity (Menon and Williamson, 2016)

$$\mathrm{AUC}_\eta(s) = \frac{1}{2} + \frac{\mathbb{E}[(p_\eta - p'_\eta)\operatorname{sgn}(s-s')]}{4\rho_\eta(1-\rho_\eta)}, \qquad \operatorname{sgn}(0) = 0. \quad (26)$$

Subtracting the score s from the Bayes score p_η expresses the regret as

$$\mathrm{AUC}_\eta(p_\eta) - \mathrm{AUC}_\eta(s) = \frac{\mathbb{E}[|p_\eta - p'_\eta|\ell_\eta]}{2\rho_\eta(1-\rho_\eta)}.$$

Here ℓ_η is one for a reversed comparison, one half for a score tie with unequal target probabilities, and zero otherwise.

Integrating the comparisons that can change. Take $s = s_\eta$ in this identity. If $|a|, |c| \le M$ and the score remainder is bounded by $M\eta^2$, a disagreement requires $|Z - Z'| \le 4M\eta + 2M\eta^2$. Outside this band, the baseline difference fixes both orderings. Inside it, write $Z = z$ and $Z' = z - \eta u$. With $A = a - a'$ and $C = c - c'$, the true and scored differences, divided by η, are

$$u + A, \qquad u + C + O(\eta).$$

For fixed marks and with the remainder removed, disagreement occurs between the roots $-A$ and $-C$. If $A > C$, its contribution is

$$\int_{-A}^{-C} |u + A| \, du = \int_0^{A-C} v \, dv = \frac{(A-C)^2}{2}.$$

Interchanging the roots gives the same expression for $A < C$. Therefore

$$\int_{\mathbb{R}} |u + A| \, \mathbf{1}\{(u + A)(u + C) < 0\} \, du = \frac{(A-C)^2}{2}.$$

The probability difference contributes one factor η and the change of variables contributes another. The regret is consequently

$$\frac{\eta^2}{2\rho_\eta(1-\rho_\eta)} \int\int h(z)h(z-\eta u)\mathbb{E}[|u + A|\ell_\eta \mid z, z - \eta u] \, du \, dz.$$

Its limit after division by η^2 is the coefficient in (25), once the density product is replaced by $h(z)^2$ and the root integral above is used.

Justifying the limit. Only a fixed bounded interval of u contributes. Extend h by zero outside its support. Cauchy–Schwarz and translation continuity in L^2 give

$$\int |h(z)h(z-\eta u) - h(z)^2|\, dz \le \|h\|_2 \|h(\cdot - \eta u) - h\|_2 \longrightarrow 0,$$

uniformly for u in that interval. After this replacement, a bound of the form $C'h(z)^2\mathbf{1}\{|u| \le C\}$ is integrable. For almost every z, conditional weak continuity gives convergence of the remaining bounded integrand except at values of u where a comparison boundary has positive conditional mass. The boundaries are $u = -A$ and $u = -C$, and each conditional law has at most countably many atoms, so there are at most countably many such values for each z and Fubini's theorem removes them from the integral. The $O_\infty(\eta^2)$ score remainder changes the rescaled difference by $O(\eta)$ and has the same zero contribution at all other boundaries. Dominated convergence and $\rho_\eta \to \rho_0$ now prove the lemma. The half-credit rule handles score ties throughout. □

Proof of the AUC part of Theorem 1. Apply Lemma 1 with

$$Z = B_0(e), \qquad a = Z(f + r), \qquad c = Z(1 - t)f.$$

Assumption 3 supplies the conditional-law condition after multiplication by z. Centering gives $\rho_\eta = \mathbb{E}B_\eta = v$. Conditional on $Z = z$, the leading error is $z(r + tf)$. For independent conditional copies,

$$\mathbb{E}[((r + tf) - (r' + tf'))^2 \mid Z = z] = 2\,\mathrm{Var}(r + tf \mid Z = z).$$

The regret formula is therefore

$$\mathrm{AUC}_\eta(B_\eta) - \mathrm{AUC}_\eta(q_{\omega,\eta}) = \frac{\eta^2}{2v(1-v)} \int h(z)^2 z^2\, \mathrm{Var}(r + tf \mid Z = z)\, dz + o(\eta^2).$$

Subtract the corresponding regret at $t = 0$ in the opposite order. Since

$$\mathrm{Var}(r + tf \mid Z) - \mathrm{Var}(r \mid Z) = t^2\, \mathrm{Var}(f \mid Z) + 2t\, \mathrm{Cov}(f, r \mid Z),$$

we obtain

$$\mathrm{AUC}_\eta(q_{\omega,\eta}) - \mathrm{AUC}_\eta(q_{1,\eta}) = -\frac{\eta^2}{4v(1-v)}(t^2 A_A + 2tC_A) + o(\eta^2).$$

This proves (10). It also explains the conditional covariance: AUC changes through comparisons whose baseline probabilities approach the same value. □

A.3 When the degree baseline is constant

When $w = 1$, every candidate has the same baseline probability v, so the previous density assumption fails. Ranking can instead be expanded around the retained score f. Subtract the constant from the fitted SGNS logit and divide by $\eta\omega_1 > 0$. By Theorem 2, and because at constant degrees $q = vb\,e^F$ is monotone in F, this leaves the ranking unchanged and gives

$$f + \eta a_\omega + O_\infty(\eta^2), \qquad a_\omega = \frac{\omega_2}{\omega_1}\mathsf{F}^2 - \omega_1 \mathcal{S}_K(M_b).$$

The next statement specifies the regularity needed for this different expansion.

Proposition A.1 (Regular-baseline comparison)**.** *Suppose $w = 1$, $S = f(e)$ has a density $g \in L^2$, and the conditional law of $(k(e), a_\omega(e), a_1(e))$ given $S = s$ is weakly continuous for almost every s. Define*

$$\mathcal{B}(h) = \frac{1}{2(1-v)} \int g(s)^2 \mathbb{E}[(k - k')(h - h') \mid S = S' = s] ds.$$

Then the constrained minima of Theorem 2 satisfy

$$\mathrm{AUC}_\eta(q_{\omega,\eta}) - \mathrm{AUC}_\eta(q_{1,\eta}) = \eta^2 \mathcal{B}\left(\frac{\omega_2}{\omega_1}\mathsf{F}^2 + (1 - \omega_1)\mathcal{S}_K(M_b)\right) + o(\eta^2). \tag{27}$$

Proof. The target difference is $v\eta(k-k')$, and its prevalence is v. Using (26) for the two normalized scores gives

$$\Delta\,\mathrm{AUC} = \frac{\eta}{4(1-v)}\mathbb{E}\big[(k-k')\,\mathrm{sgn}(S-S'+\eta(a_\omega-a'_\omega)+O(\eta^2))\big] - \frac{\eta}{4(1-v)}\mathbb{E}\big[(k-k')\,\mathrm{sgn}(S-S'+\eta(a_1-a'_1)+O(\eta^2))\big].$$

Only $S-S'=O(\eta)$ contributes. Write $S=s$, $S'=s-\eta u$. For fixed marks,

$$\int_{\mathbb{R}}\{\mathrm{sgn}(u+a_\omega-a'_\omega)-\mathrm{sgn}(u+a_1-a'_1)\}\,du = 2\{(a_\omega-a_1)-(a'_\omega-a'_1)\}.$$

The Jacobian is η. Applying the same density-replacement and conditional-continuity argument as in Lemma 1 gives

$$\Delta\,\mathrm{AUC} = \eta^2\mathcal{B}(a_\omega-a_1)+o(\eta^2).$$

At one step $\omega_1=1$ and $\omega_2=0$, so

$$a_\omega-a_1 = \frac{\omega_2}{\omega_1}\mathsf{F}^2+(1-\omega_1)\mathcal{S}_K(M_b).$$

Substitution proves (27). □

On a level set of f, the target difference is $k-k'=r-r'$. The functional $\mathcal{B}$ therefore measures how the second-order fitted correction aligns with omitted target structure. This is why that correction can determine the leading ranking change when degrees are constant. The additional density condition belongs to this ranking calculation, not to the population optimizer theorem.

A.4 A SMOOTH EXAMPLE WITH BOTH SIGNS

We give a smooth model in which a small lag-two perturbation improves both metrics, then reverse the omitted component to obtain the other sign. This construction establishes that the two regimes exist; it is not fitted to a real network. Let μ be uniform on $[-1,1]$ and choose

$$d\zeta(x)=(1+\delta x)d\mu(x),\qquad w(x)=(1+\delta x)^{-1},\qquad 0<\delta<1,\quad 0<v<(1-\delta)^2.$$

For $\phi(x)=\sqrt{3}x$ and $\psi(x)=\sqrt{5}(3x^2-1)/2$, set

$$f=\phi\otimes\phi,\qquad r=-\epsilon\psi\otimes\psi,\qquad k=f+r,\qquad 0<\epsilon<1/2.$$

These two functions are centered and orthonormal under μ, so the width-two spectral gap holds. The model is smooth and has varying degrees. A change of variable by the CDF of ζ gives a uniform-latent representation.

For NLL, (24) becomes

$$C_R=-v\epsilon\sum_{q\ge1}v^q\left(\int\frac{\phi\psi}{(1+\delta x)^q}\,d\mu\right)^2.$$

Every term is nonpositive. To show strictness, consider $q=1$. The series $(1+\delta x)^{-1}=\sum_{j\ge0}(-\delta x)^j$ converges uniformly because $\delta<1$. Since $\phi\psi$ is odd, only odd $j=2h+1$ survive, with negative coefficients. Direct integration gives

$$\int\phi\psi\,x^{2h+1}\,d\mu=\frac{\sqrt{15}}{2}\left(\frac{3}{2h+5}-\frac{1}{2h+3}\right)=\frac{2\sqrt{15}(h+1)}{(2h+5)(2h+3)}>0.$$

Thus the $q=1$ square is nonzero and $C_R<0$.

For AUC, we compute the conditional moments on baseline level sets. Put $P_0=(1+\delta x)(1+\delta y)$, $a=1-\delta$, $b_0=1+\delta$, and

$$l(p)=\max(a,p/b_0),\qquad h_0(p)=\min(b_0,p/a).$$

Under ζ, the variable $\xi=1+\delta x$ has density $\xi/(2\delta)$ on $[a,b_0]$. Changing variables from (ξ,v) to $(\xi,p=\xi v)$ gives joint density $p/(4\delta^2\xi)$. Hence

$$f_{P_0}(p)=\frac{p\log(h_0(p)/l(p))}{4\delta^2},\qquad a^2\le p\le b_0^2,$$

and $\log \xi$ conditional on $P_0 = p$ is uniform on $[\log l(p), \log h_0(p)]$. Since $Z = B_0 = v/P_0$, this also gives its density by the change of variables $p = v/z$. For each fixed $\delta > 0$, the resulting densities and conditional moments are bounded and piecewise continuous, as required by the ranking calculation.

To determine the sign, fix $v < 1$ and let $\delta \downarrow 0$. Rescaling $P_0 = 1 + \delta s$ makes its level sets tend to $x + y = s$. For $0 \le s \le 2$, let $L_s = 2 - s$ and write $x = s/2 + z$, $y = s/2 - z$, where z is uniform on $[-L_s/2, L_s/2]$. The even moments are

$$\mathbb{E}[z^{2j} \mid s] = \frac{L_s^{2j}}{2^{2j}(2j+1)}.$$

In these coordinates,

$$f = 3(s^2/4 - z^2), \qquad r = -\frac{5\epsilon}{4}\left\{9z^4 - (9s^2/2 + 6)z^2 + (3s^2/4 - 1)^2\right\}.$$

Using $\mathrm{Var}(z^2) = L_s^4/180$ and $\mathrm{Cov}(z^2, z^4) = L_s^6/840$, we obtain

$$\begin{aligned}
\mathrm{Var}(f \mid s) &= 9\,\mathrm{Var}(z^2) = \frac{L_s^4}{20}, \\
\mathrm{Cov}(f, r \mid s) &= \frac{15\epsilon}{4}\left\{9\frac{L_s^6}{840} - (9s^2/2 + 6)\frac{L_s^4}{180}\right\} = -\frac{\epsilon L_s^4(3s^2 + 9s - 2)}{56}.
\end{aligned}$$

The density of $x + y$ is $(2 - |s|)/4$ on $[-2, 2]$. Changing variables from Z to s in the coefficient integrals and using symmetry gives

$$\begin{aligned}
\lim_{\delta \downarrow 0} \frac{\delta A_A}{2v} &= 2\int_0^2 \frac{L_s^2}{16}\frac{L_s^4}{20}\,ds = \frac{4}{35}, \\
\lim_{\delta \downarrow 0} \frac{\delta C_A}{2v} &= -2\int_0^2 \frac{L_s^2}{16}\frac{\epsilon L_s^4(3s^2 + 9s - 2)}{56}\,ds = -\frac{\epsilon}{42}.
\end{aligned}$$

The finite-δ conditional parametrization above bounds the rescaled integrands and justifies dominated convergence. Thus $C_A < 0$ for sufficiently small fixed δ.

After fixing such a δ, choose

$$0 < t < \min\{1, -2C_R/A_R, -2C_A/A_A\}.$$

Theorem 1 and the SGNS expansion then give improvement in both AUC and native NLL for sufficiently small η. The order of choices is δ, then t, then η; the ranking calculation does not substitute a point mass for the baseline density. Replacing $-\epsilon$ by $+\epsilon$ reverses both cross terms without changing the degree baseline, retained component, or spectral separation. That control favors one step at every fixed positive t. The improving interval need not include $t = 1/2$, the uniform two-step window.

A.5 Smoothness and scope

Smoothness can bound approximation error, but does not determine a window preference. Suppose an unrestricted SGNS logit has an s-Hölder representation on $[0, 1]^p$, with $0 < s \le 1$. Partition the node space into m^p cubes and replace the logit by its value at representatives of each pair of cubes. The uniform error is $O(m^{-s})$. The resulting kernel has at most m^p distinct row functions, hence rank at most $d = m^p$, and squared L^2 error $O(d^{-2s/p})$. This is an elementary approximation argument of the kind used in smooth-graphon estimation (Xu, 2018). It assumes that a smooth latent representation exists, not that the graph is smooth in an arbitrary vertex ordering.

A.6 Further interpretations of the cross terms

The error-path interpretation of Theorem 1 can also be written as a regression coefficient. Define the two quadratic forms

$$\langle a, c\rangle_R = \mathbb{E}\left[\frac{B_0}{1 - B_0}ac\right], \qquad \langle a, c\rangle_A = 2\int h(z)^2 z^2\,\mathrm{Cov}(a, c \mid Z = z)\,dz.$$

The second is a positive semidefinite form: it ignores components that are constant on each baseline level. For either form, $A_M = \langle f, f \rangle_M$ and $C_M = \langle f, r \rangle_M$. If $A_M > 0$, the best scalar multiple of f for approximating k in that form has coefficient

$$\beta_M = \arg\min_{\beta \in \mathbb{R}} \|k - \beta f\|_M^2 = 1 + \frac{C_M}{A_M}. \tag{28}$$

To verify the coefficient, expand and complete the square:

$$\begin{aligned} \|k - \beta f\|_M^2 &= \|r + (1-\beta) f\|_M^2 \\ &= \|r\|_M^2 + 2(1-\beta) C_M + (1-\beta)^2 A_M \\ &= \|r\|_M^2 - \frac{C_M^2}{A_M} + A_M \left(\beta - 1 - \frac{C_M}{A_M} \right)^2 . \end{aligned}$$

If $A_M = 0$, Cauchy–Schwarz for the positive semidefinite form gives $C_M = 0$; there is no strict leading comparison. One step uses coefficient one, whereas walking uses $1 - t$ to first order. Thus $\beta_M < 1$ means that the one-step representation gives too much weight to f for this metric. The unconstrained best shrinkage is $t = -C_M / A_M$; the improvement over one step ends at $t = -2C_M / A_M$. Only values in the allowed lag family are candidates.

For AUC, this regression compares pairs with similar degree-based probabilities. A coarse community component might assign high scores to pairs whose finer structure suppresses their edge probabilities, producing a negative conditional covariance. The reverse is also possible. Shrinking f does not reorder pairs with exactly the same B_0; it changes the balance between residual structure and the slightly different baselines of nearly tied pairs. The covariance on an equal-baseline level is the limit of those comparisons, not a claim that shrinkage changes every within-level ordering.

For NLL, the relation to degree variation is visible in the signed expansion

$$C_R = v \sum_{a \geq 1} v^a \sum_{i \text{ retained}} \sum_{j \text{ omitted}} \lambda_i \lambda_j \left(\int w(x)^a u_i(x) u_j(x) \, d\mu(x) \right)^2 . \tag{29}$$

The term at $a = 0$ vanishes by spectral orthogonality. Constant degree weights make every term vanish. Mixed-sign modes can instead give a negative cross term when they overlap after degree weighting. Thus neither degree variation nor negative eigenvalues alone predicts a benefit: their relation to the retained and omitted eigenfunctions matters. The derivation is the series calculation in (24), with its zero $a = 0$ term removed.

B What calibration removes

This appendix proves Proposition 1 and the constant-degree statements below. The proofs separate two questions: whether a correction preserves ordering, and how much probability error it removes. All expectations use ζ^2. Predictions and targets lie in a common interval $[\epsilon, 1 - \epsilon]$ unless stated otherwise. Write $\ell(p, q) = -p \log q - (1 - p) \log(1 - q)$; differences of its expectation equal differences of excess NLL. The Taylor bounds in Appendix A are uniform on this interval.

B.1 Regular-baseline statements

When degrees are constant, write $p_0 = v$ for their common probability baseline. Changing a calibration slope by order one can then change probabilities by only order η. This requires a separate calculation from the near-identity correction of Proposition 1. We state the spectral and SGNS results first, then prove each below.

Proposition B.1 (Spectral scale restoration)**.** *Under Assumption 1, use the spectral prediction of Definition 4, assume $w = 1$, and put $s = \omega_1 > 0$. Define*

$$\bar{s}_{\omega,\eta} = p_0 + \frac{s_{\omega,\eta} - p_0}{s}. \tag{30}$$

For a fixed finite lag family and sufficiently small η, this is a probability in $(0, 1)$ and has the same AUC as $s_{\omega,\eta}$. Moreover,

$$\mathcal{R}_\eta(\bar{s}_{\omega,\eta}) - \mathcal{R}_\eta(s_{1,\eta}) = O(\eta^4), \tag{31}$$
$$\mathcal{R}_\eta(\bar{s}_{\omega,\eta}) = \frac{p_0\eta^2}{2(1-p_0)}\|r\|_2^2 + O(\eta^3).$$

Proposition B.2 (SGNS scale restoration and logistic calibration)**.** *Under Assumptions 1 and 5, assume $w = 1$, put $p_0 = v$ and $s = \omega_1 > 0$, and define*

$$\bar{q}_{\omega,\eta} = p_0\left(\frac{q_{\omega,\eta}}{p_0}\right)^{1/s} = p_0\exp\left(\frac{F_{\omega,\eta} + \log b}{s}\right). \tag{32}$$

For sufficiently small η, this probability preserves AUC and satisfies

$$\mathcal{R}_\eta(\bar{q}_{\omega,\eta}) - \mathcal{R}_\eta(q_{1,\eta}) = O(\eta^4).$$

If $\|f\|_2 > 0$, population logistic calibration

$$c_{\omega,\eta} = \sigma(\alpha_{\omega,\eta} + \beta_{\omega,\eta}F_{\omega,\eta})$$

has a unique minimum, with

$$\beta_{\omega,\eta} = \frac{1}{s(1-p_0)} + O(\eta) > 0, \qquad \mathcal{R}_\eta(c_{\omega,\eta}) - \mathcal{R}_\eta(c_{1,\eta}) = O(\eta^4). \tag{33}$$

The one-step score is calibrated as well. Both corrected risks have leading term $p_0\eta^2\|r\|_2^2/[2(1-p_0)]$.

The power correction uses the known baseline. Logistic calibration instead fits an intercept and slope to F, or equivalently to $\log q$ when degrees are constant. Calibration on $\mathrm{logit}(q)$ is a different family. Both corrections preserve AUC at sufficiently small contrast; neither determines the sign of the remaining fourth-order NLL difference.

B.2 Order preservation and the proper-loss decomposition

Order preservation follows directly from the definition of AUC. For a strictly increasing g,

$$\mathrm{sgn}(g(q(e)) - g(q(e'))) = \mathrm{sgn}(q(e) - q(e')).$$

Substitution in (26) proves AUC preservation, including ties. A nondecreasing map can merge unequal scores and is not covered by this argument. For the local correction in (13), if $q_2 > q_1$ then

$$\begin{aligned}&\mathrm{logit}(q_2) + \eta a(q_2) - \mathrm{logit}(q_1) - \eta a(q_1)\\ &\quad\geq \{4 - \eta\,\mathrm{Lip}(a)\}(q_2 - q_1) > 0\end{aligned} \tag{34}$$

whenever $\eta\,\mathrm{Lip}(a) < 4$. Here we used $\mathrm{logit}'(q) = 1/[q(1-q)] \geq 4$ and the Lipschitz bound on a. The outer sigmoid is also strictly increasing.

A separate calculation identifies the loss that any decoder of a fixed score can remove (Bröcker, 2009). Put $m(S) = \mathbb{E}[p \mid S]$. Cross-entropy is affine in its first argument, so conditional expectation gives

$$\begin{aligned}\mathbb{E}[\ell(p, g(S)) \mid S] &= \ell(m(S), g(S)),\\ \mathbb{E}[\ell(p, m(S)) \mid S] &= \ell(m(S), m(S)).\end{aligned}$$

Subtracting and then adding the Bayes loss yields

$$\mathbb{E}\,\mathrm{KL}(p\|g(S)) = \mathbb{E}\,\mathrm{KL}(p\|m(S)) + \mathbb{E}\,\mathrm{KL}(m(S)\|g(S)). \tag{35}$$

Here KL is Bernoulli divergence. The first term is the error left by the score; only the second changes with its decoder. Although m minimizes loss among all decoders of S, it need not be increasing. This identity therefore does not promise that the minimum can be attained without changing AUC.

B.3 REGULAR SPECTRAL PREDICTIONS

The affine correction restores the retained first-order term. To obtain the stronger fourth-order loss comparison, we also use orthogonality to the omitted kernel.

Proof of Proposition B.1. Set $s = \omega_1$, $w = 1$, and $p_0 = v$. The spectral prediction and its one-step reference are

$$s_{\omega,\eta} = p_0 \left\{ 1 + \eta s f + \sum_{\ell \geq 2} \omega_\ell \eta^\ell \mathsf{F}^\ell \right\}, \qquad s_{1,\eta} = p_0(1 + \eta f),$$

where F^ℓ denotes its kernel. Therefore

$$D_{\omega,\eta} := \bar{s}_{\omega,\eta} - s_{1,\eta} = \frac{p_0}{s} \sum_{\ell \geq 2} \omega_\ell \eta^\ell \mathsf{F}^\ell = O_\infty(\eta^2). \tag{36}$$

Every term is retained, so $\langle r, D_{\omega,\eta} \rangle = 0$. The affine correction has slope $1/s > 0$, and uniform convergence to p_0 puts all its values in the probability interval for sufficiently small contrast.

For the target $p_\eta = p_0(1 + \eta(f + r))$, the NLL derivative at the one-step prediction is

$$\partial_q \ell(p_\eta, s_{1,\eta}) = -\frac{\eta r}{(1 + \eta f)(1 - p_0 - p_0 \eta f)} = -\frac{\eta r}{1 - p_0} + O_\infty(\eta^2).$$

Expanding in the prediction, not the target, gives

$$\begin{aligned} \mathcal{R}_\eta(\bar{s}_{\omega,\eta}) - \mathcal{R}_\eta(s_{1,\eta}) &= \mathbb{E}[\partial_q \ell(p_\eta, s_{1,\eta}) D_{\omega,\eta}] + O(\|D_{\omega,\eta}\|_2^2) \\ &= -\frac{\eta}{1 - p_0} \langle r, D_{\omega,\eta} \rangle + O(\eta^2 \|D_{\omega,\eta}\|_1) + O(\|D_{\omega,\eta}\|_2^2) \\ &= O(\eta^4). \end{aligned}$$

Finally, $\bar{s}_{\omega,\eta} - p_\eta = -p_0 \eta r + O_\infty(\eta^2)$ gives its leading risk by the Bernoulli Taylor formula. If $r = 0$, its entire excess risk is $O(\eta^4)$. □

B.4 A SMOOTH CORRECTION WITH VARYING DEGREES

We first expand the corrected prediction for a fixed function a. The coefficient of its loss then becomes a conditional least-squares problem, solved separately at each baseline probability.

Proof of Proposition 1. Write $q = q_{t,\eta}$, $Z = B_0$, and $d_t = r + tf$. Uniformly on the prediction interval,

$$q = Z + \eta Z(1 - t) f + O_\infty(\eta^2), \qquad a(q) = a(Z) + O_\infty(\eta),$$

where the second estimate uses the fixed Lipschitz constant of a. Taylor expansion of the sigmoid at logit q gives

$$\begin{aligned} g_{\eta,a}(q) &= q + \eta q(1 - q) a(q) + O_\infty(\eta^2) \\ &= Z + \eta Z\{(1 - t) f + (1 - Z) a(Z)\} + O_\infty(\eta^2), \\ g_{\eta,a}(q) - p_\eta &= -\eta Z\{d_t - (1 - Z) a(Z)\} + O_\infty(\eta^2). \end{aligned} \tag{37}$$

Hence

$$\mathcal{R}_\eta(g_{\eta,a}(q)) = \frac{\eta^2}{2} \mathbb{E}\left[\frac{Z}{1 - Z} \{d_t - (1 - Z) a(Z)\}^2 \right] + O(\eta^3). \tag{38}$$

The Bernoulli Taylor formula supplies the factor $Z/(1 - Z)$: the squared prediction error contributes Z^2, and the curvature contributes $1/[Z(1 - Z)]$. The remainder is uniform when a has common sup-norm and Lipschitz bounds.

Conditional on Z, expand the square around $m_t(Z) = \mathbb{E}[d_t \mid Z]$:

$$\begin{aligned} &\mathbb{E}[\{d_t - (1 - Z) a(Z)\}^2 \mid Z] \\ &\quad = \mathrm{Var}(d_t \mid Z) + \{m_t(Z) - (1 - Z) a(Z)\}^2 \\ &\qquad + 2\{m_t(Z) - (1 - Z) a(Z)\} \mathbb{E}[d_t - m_t(Z) \mid Z] \\ &\quad = \mathrm{Var}(d_t \mid Z) + \{m_t(Z) - (1 - Z) a(Z)\}^2. \end{aligned}$$

This is minimized pointwise by $a_t^* = m_t/(1-Z)$. The interior bound and the assumed extension of m_t make a_t^* bounded and Lipschitz. Formula (34) therefore proves that the map is strictly increasing for small contrast. Setting $a = a_t^*$ in (38) proves (14). Setting $a = 0$ and subtracting proves (15).

The window comparison uses the correction at both windows. Since

$$\mathrm{Var}(r + tf \mid Z) - \mathrm{Var}(r \mid Z) = t^2 \,\mathrm{Var}(f \mid Z) + 2t \,\mathrm{Cov}(f, r \mid Z),$$

integration proves (16). A nonzero quadratic coefficient determines the sign for each fixed lag law and sufficiently small contrast. At a zero coefficient, the remainder may determine the sign. Also,

$$A_R - \widetilde{A}_R = \mathbb{E}\left[\frac{Z}{1-Z}\mathbb{E}[f \mid Z]^2\right] \geq 0,$$
$$C_R - \widetilde{C}_R = \mathbb{E}\left[\frac{Z}{1-Z}\mathbb{E}[f \mid Z]\mathbb{E}[r \mid Z]\right].$$

The second difference has no fixed sign. This is why calibration can change a window preference without making every walk beneficial. □

The correction leaves AUC unchanged, so its coefficients still use the original ranking weights. Since Z has density h, they can be written under the same candidate distribution as

$$A_A = 2\mathbb{E}[h(Z)Z^2 \,\mathrm{Var}(f \mid Z)], \qquad C_A = 2\mathbb{E}[h(Z)Z^2 \,\mathrm{Cov}(f, r \mid Z)].$$

The corrected-NLL weight is $Z/(1-Z)$ instead. For example, a pointwise condition

$$\mathrm{Cov}(f, r \mid Z = z) < -\frac{t}{2}\,\mathrm{Var}(f \mid Z = z)$$

on all contributing levels, with strictness on a set of positive weighted measure, makes both criteria improve. Aggregate signs suffice for each individual metric and can disagree when the covariance varies across levels. The constant-degree case is a different limit: a slope change of order one can change its probabilities by only order η, and is not represented uniformly by a near-identity correction with fixed Lipschitz bounds as degree variation vanishes.

B.5 The calibration family actually fitted

A finite-dimensional calibrator need not represent the best function a_t^*. We derive the loss of the family actually fitted, rather than substitute the unrestricted conditional-variance formula. Fix a bounded Lipschitz feature vector ψ and put

$$G = \mathbb{E}[Z(1-Z)\psi(Z)\psi(Z)^\top], \qquad u_f = \mathbb{E}[Z\psi(Z)f], \qquad u_r = \mathbb{E}[Z\psi(Z)r].$$

Assume G is positive definite. The prediction family and its population objective are

$$g_\theta(q) = \sigma\{\mathrm{logit}\, q + \theta^\top \psi(q)\}, \qquad J_{t,\eta}(\theta) = \mathbb{E}\ell(p_\eta, g_\theta(q_{t,\eta})).$$

The identity map is $\theta = 0$. We first treat a parameter set that contains a neighborhood of zero; a boundary constraint requires minimizing the corresponding constrained quadratic.

Writing $\theta = \eta a$ in (38) gives

$$\frac{\eta^2}{2}\left\{\mathbb{E}\left[\frac{Z}{1-Z}d_t^2\right] - 2a^\top u_t + a^\top G a\right\}, \qquad u_t = u_r + t u_f. \tag{39}$$

The quadratic has minimizer $a_t = G^{-1}u_t$. After subtraction of the one-step value, the proposed coefficients are

$$A_R^\psi = A_R - u_f^\top G^{-1} u_f, \qquad C_R^\psi = C_R - u_f^\top G^{-1} u_r. \tag{40}$$

The next calculation checks that fitting the actual logistic objective gives these coefficients, not merely that they minimize a Taylor polynomial.

Differentiating the actual calibration loss gives

$$\nabla J_{t,\eta}(\theta) = \mathbb{E}[(g_\theta(q) - p_\eta)\psi(q)], \tag{41}$$
$$\nabla^2 J_{t,\eta}(\theta) = \mathbb{E}[g_\theta(q)(1 - g_\theta(q))\psi(q)\psi(q)^\top].$$

Boundedness and Lipschitz continuity imply

$$\nabla J_{t,\eta}(0) = -\eta u_t + O(\eta^2), \qquad \nabla^2 J_{t,\eta}(\theta) = G + O(\eta + \|\theta\|)$$

uniformly for θ near zero. In a small fixed ball the Hessian is at least $\lambda_{\min}(G)/2$. If $\|\nabla J(0)\| \le C\eta$, then

$$\langle \nabla J(\theta), \theta \rangle \ge -C\eta\|\theta\| + \frac{\lambda_{\min}(G)}{2}\|\theta\|^2 > 0 \quad \text{when } \|\theta\| = \frac{4C\eta}{\lambda_{\min}(G)}.$$

The minimum over this closed ball is therefore interior. In particular it is stationary and $\theta = O(\eta)$. Integrating the Hessian along the segment from zero gives

$$0 = \nabla J(0) + \left(\int_0^1 \nabla^2 J(u\theta)\, du \right) \theta = -\eta u_t + G\theta + O(\eta^2).$$

Multiplication by G^{-1} yields

$$\theta_{t,\eta} = \eta G^{-1} u_t + O(\eta^2). \tag{42}$$

For small contrast the features $\psi(q)$ span the parameter space in L^2, and the target stays strictly between zero and one. The displayed Hessian is then positive definite at every finite parameter, so the logistic objective is strictly convex. The stationary point is its unique global minimum. A monotonicity constraint with the identity in its interior is inactive here, because the Lipschitz constant of $\theta_{t,\eta}^\top \psi$ is $O(\eta)$ and (34) applies.

Substitution of (42) in (39) gives

$$\begin{aligned} \mathcal{R}_\eta(g_{\theta_{t,\eta}}(q_{t,\eta})) = {} & \frac{\eta^2}{2}\mathbb{E}\left[\frac{Z}{1-Z}(r + tf)^2 \right] \\ & - \frac{\eta^2}{2}(u_r + tu_f)^\top G^{-1}(u_r + tu_f) + O(\eta^3). \end{aligned}$$

Subtract the expression at $t = 0$ and expand

$$(u_r + tu_f)^\top G^{-1}(u_r + tu_f) - u_r^\top G^{-1} u_r = 2t\, u_f^\top G^{-1} u_r + t^2 u_f^\top G^{-1} u_f.$$

The resulting coefficients are those in (40). Also $A_R^\psi \ge 0$: it is the minimum of the nonnegative quadratic obtained from (39) with $d_t = f$. Choosing $\psi(z) = (1, \operatorname{logit} z)$ gives ordinary affine-logit calibration near the identity. A piecewise-linear basis with fixed knots gives a richer Lipschitz class. The appropriate coefficient is C_R^ψ, not $\widetilde{C}_R$, unless the finite family represents the optimal conditional correction at this order.

B.6 The regular fixed-rank SGNS score

For SGNS, scale restoration also leaves a fourth-order window difference. The fitted subspace may move; what matters is that the second-order change remains orthogonal to the omitted operator.

Proof of the restoration part of Proposition B.2. Set $w = 1$, $p_0 = v$, and $s = \omega_1$. Write H_ω for the second-order term in Theorem 2. Then

$$F_{\omega,\eta} = -\log b + \eta s f + \eta^2 H_\omega + O_\infty(\eta^3), \qquad \mathsf{R} = Q\mathsf{R}Q, \qquad QH_\omega Q = 0.$$

The last two identities imply

$$\langle r, H_\omega \rangle_{\mathrm{HS}} = \operatorname{tr}(\mathsf{R}^* Q H_\omega Q) = 0, \qquad \langle r, \mathbf{1} \rangle = \langle r, f \rangle = 0. \tag{43}$$

No symmetry of the graphon beyond its being undirected is used here. In particular, retained–omitted blocks in H_ω do not change the inner product.

Expand the corrected prediction and its one-step reference:

$$\begin{aligned} \bar{q}_{\omega,\eta} &= p_0\{1 + \eta f + \eta^2(H_\omega/s + f^2/2)\} + O_\infty(\eta^3), \\ q_{1,\eta} &= p_0\{1 + \eta f + \eta^2(H_1 + f^2/2)\} + O_\infty(\eta^3). \end{aligned}$$

Thus $D = \bar{q}_{\omega,\eta} - q_{1,\eta} = p_0\eta^2(H_\omega/s - H_1) + O_\infty(\eta^3)$. Also

$$\partial_q \ell(p_\eta, q_{1,\eta}) = -\frac{\eta r}{1 - p_0} + O_\infty(\eta^2).$$

The prediction Taylor formula now gives

$$\mathcal{R}_\eta(\bar{q}_{\omega,\eta}) - \mathcal{R}_\eta(q_{1,\eta}) = -\frac{p_0\eta^3}{1-p_0}\langle r, H_\omega/s - H_1\rangle + O(\eta^4)$$
$$= O(\eta^4).$$

The map in (32) is strictly increasing. Its output tends uniformly to p_0, so it is a probability for small contrast and preserves AUC. Its error is $-p_0\eta r + O_\infty(\eta^2)$, which gives the leading risk. If $r = 0$ the whole excess risk is fourth order. □

B.7 Population logistic calibration of SGNS

The power map above uses the population baseline. We now fit an intercept and slope instead. Rescaling the nearly constant SGNS feature makes the limiting score equations nonsingular; orthogonality then gives the same fourth-order comparison.

Proof of the logistic part of Proposition B.2. For fixed $\eta > 0$, replace the nearly constant feature $F_{\omega,\eta}$ by

$$x_{\omega,\eta} = \frac{F_{\omega,\eta} + \log b}{\eta s} = f + \eta H_\omega/s + O_\infty(\eta^2).$$

This affine change of feature leaves the two-parameter calibration class unchanged. Write its prediction as

$$c_{\omega,\eta} = \sigma\{\operatorname{logit} p_0 + \eta(a + \gamma x_{\omega,\eta})\}.$$

The exact moment equations, divided by η, are

$$\mathbb{E}\left[\frac{c_{\omega,\eta} - p_\eta}{\eta}\begin{pmatrix}1\\ x_{\omega,\eta}\end{pmatrix}\right] = 0. \tag{44}$$

At $\eta = 0$ they reduce, after division by p_0, to

$$\mathbb{E}\left[\{(1-p_0)(a+\gamma f) - (f+r)\}\begin{pmatrix}1\\ f\end{pmatrix}\right] = 0.$$

Centering and orthogonality give $\mathbb{E}f = \mathbb{E}r = \mathbb{E}fr = 0$. The two equations therefore reduce to

$$(1-p_0)a = 0, \qquad \{(1-p_0)\gamma - 1\}\mathbb{E}f^2 = 0.$$

Since $\mathbb{E}f^2 > 0$, their unique solution is $a = 0$ and $\gamma = (1-p_0)^{-1}$.

The derivative of the divided equations in (a, γ) is

$$p_0(1-p_0)\begin{pmatrix}1 & 0\\ 0 & \mathbb{E}f^2\end{pmatrix},$$

which is invertible. The analytic SGNS branch and the implicit-function theorem give

$$a = \eta a_{\omega,1} + O(\eta^2), \qquad \gamma = \frac{1}{1-p_0} + \eta\gamma_{\omega,1} + O(\eta^2).$$

For each positive η, the two-parameter logistic objective has positive-definite Hessian, because its feature is nonconstant and all fitted probabilities are interior. This stationary branch is therefore its unique global minimum. Returning to the original feature,

$$\beta_{\omega,\eta} = \frac{\gamma}{s} = \frac{1}{s(1-p_0)} + O(\eta) > 0, \qquad \alpha_{\omega,\eta} = \operatorname{logit} p_0 + \beta_{\omega,\eta}\log b + \eta a.$$

This proves both the slope formula and AUC preservation.

For the risk comparison expand the probabilities to second order:

$$c_{\omega,\eta} = p_0\left[1 + \eta f + \eta^2\left\{\frac{H_\omega}{s} + (1-p_0)(a_{\omega,1} + \gamma_{\omega,1}f) + \frac{1-2p_0}{2(1-p_0)}f^2\right\}\right] + O_\infty(\eta^3).$$

The f^2 term is the same at every window. Hence

$$c_{\omega,\eta} - c_{1,\eta} = p_0\eta^2\left\{\frac{H_\omega}{s} - H_1 + (1-p_0)\big[(a_{\omega,1} - a_{1,1}) + (\gamma_{\omega,1} - \gamma_{1,1})f\big]\right\} + O_\infty(\eta^3).$$

Every term in braces is orthogonal to r by (43). Since $\partial_q \ell(p_\eta, c_{1,\eta}) = -\eta r/(1-p_0) + O_\infty(\eta^2)$, the cubic contribution to the NLL difference vanishes. More explicitly, if $D = c_{\omega,\eta} - c_{1,\eta}$ and $D = p_0\eta^2 D_2 + O_\infty(\eta^3)$, then

$$\begin{aligned}\mathcal{R}_\eta(c_{\omega,\eta}) - \mathcal{R}_\eta(c_{1,\eta}) &= -\frac{p_0\eta^3}{1-p_0}\langle r, D_2\rangle + O(\eta^4)\\ &= O(\eta^4).\end{aligned}$$

Here the Taylor remainder is $O(\|D\|_2^2) = O(\eta^4)$, and the first inner product is zero. This proves (33). □

These conclusions use a specified calibration feature. At constant degrees, F and $\log q$ differ by a constant, but $\operatorname{logit} q$ is not affine in F. Changing to that feature changes its second-order expansion. Similarly, with varying degrees the map $B_0(q/B_0)^{1/s}$ is not scalar in q: at $s = 1/2$, two pairs with $q = 0.2$ and baselines 0.1 and 0.2 are mapped to 0.4 and 0.2. It changes a tie and can reverse nearby orderings. No AUC-preservation statement is made for this pair-dependent map.

C THE CONSTRAINED SGNS MINIMUM

We prove Theorem 2 under Assumptions 1 and 5. The proof has two tasks: construct a stationary rank-d score, allowing its represented directions to move, and show that this score minimizes the loss over the whole bounded rank-d class. We first give the comparison used for the second task.

All function inner products use $L^2(\mu)$; kernel inner products use $L^2(\mu \otimes \mu)$. Thus

$$\|H\|_{\mathrm{HS}}^2 = \int |H(x,y)|^2 \, d\mu(x) d\mu(y), \qquad \langle H_1, H_2\rangle_{\mathrm{HS}} = \operatorname{tr}(H_1^* H_2),$$

$$(a \otimes c)h = a\langle c, h\rangle, \qquad \langle a \otimes c, H\rangle_{\mathrm{HS}} = \langle a, Hc\rangle.$$

A rank-d kernel is a sum of at most d terms $a_j(x)b_j(y)$. Changing between μ and ζ does not change that rank, since w is bounded above and below. A bound on the kernel does not bound every factorization of it. We write F for a candidate score and F for the operator with retained kernel f.

C.1 A GLOBAL COMPARISON FROM A NORMAL GRADIENT

A stationary point under a rank constraint can have a nonzero loss gradient. Its gradient must be orthogonal to the allowed first-order changes. The lemma below gives a sufficient bound on that remaining gradient for global optimality.

Lemma 2 (Normal-gradient certificate)**.** *Let $\mathcal{L}$ be κ-strongly convex on a convex set of Hilbert–Schmidt kernels. Suppose a feasible kernel F of rank exactly d has smallest nonzero singular value $s > 0$ and gradient G. If the row and column spaces of G are orthogonal to those of F, every feasible H of rank at most d satisfies*

$$\mathcal{L}(H) - \mathcal{L}(F) \geq \frac{1}{2}\left(\kappa - \frac{\|G\|_{\mathrm{op}}}{s}\right)\|H - F\|_2^2. \tag{45}$$

Proof. Put $\Delta = H - F$ and choose $a > \|G\|_{\mathrm{op}}/s$. The operators F and $-G/a$ have orthogonal row and column spaces. Their singular values therefore combine without mixing. Every singular value of G/a is smaller than the smallest nonzero singular value of F, so F is a best rank-d approximation to $F - G/a$. In particular,

$$\begin{aligned}\|G/a\|_{\mathrm{HS}}^2 &\leq \|H - (F - G/a)\|_{\mathrm{HS}}^2\\ &= \|\Delta\|_{\mathrm{HS}}^2 + \frac{2}{a}\langle G, \Delta\rangle_{\mathrm{HS}} + \frac{\|G\|_{\mathrm{HS}}^2}{a^2}.\end{aligned}$$

Canceling the last term and multiplying by $a/2$ gives

$$\langle G, \Delta\rangle_{\mathrm{HS}} \geq -\frac{a}{2}\|\Delta\|_{\mathrm{HS}}^2.$$

Strong convexity now yields

$$\begin{aligned}\mathcal{L}(H) - \mathcal{L}(F) &\geq \langle G, \Delta\rangle_{\mathrm{HS}} + \frac{\kappa}{2}\|\Delta\|_{\mathrm{HS}}^2 \\ &\geq \frac{\kappa - a}{2}\|\Delta\|_{\mathrm{HS}}^2.\end{aligned}$$

Let $a \downarrow \|G\|_{\mathrm{op}}/s$ to obtain (45). The segment between F and H need not have low rank: strong convexity is used on the convex score-bounded set containing that segment. The approximation step allows nonsymmetric H. □

C.2 Constructing the moving subspace

We now construct the score to which the certificate will apply. Put $c_0 = -\log b$ and $c_b = b/(1+b)^2$. Use the orthonormal retained basis $e_0 = \mathbf{1}$, $e_j = u_j$ for $1 \leq j \leq m$, with projection P, and set $Q = I - P$. The omitted operator R acts only on $QL^2(\mu)$:

$$PK = KP = \mathsf{F}, \qquad QKQ = \mathsf{R}, \qquad \mathsf{R}e_j = 0 \quad (0 \leq j \leq m), \qquad K^2 = \mathsf{F}^2 + \mathsf{R}^2.$$

The last identity uses $\mathsf{FR} = \mathsf{RF} = 0$.

The construction will use the uniform norm. Each retained eigenfunction is bounded because

$$|u_j(x)| = \frac{1}{|\lambda_j|}\left|\int k(x,y)u_j(y)\,d\mu(y)\right| \leq \frac{\|k\|_\infty\|u_j\|_1}{|\lambda_j|} \leq \frac{\|k\|_\infty}{|\lambda_j|}.$$

Consequently P and Q are bounded on L^∞ as well as on L^2. For example,

$$\|Ph\|_\infty \leq \sum_{j=0}^{m}\|e_j\|_\infty\|e_j\|_1\|h\|_\infty.$$

Coordinates for the score. Let $X : PL^2 \to QL^2$ describe a change of the retained space, and let A be a symmetric matrix on PL^2. A symmetric candidate is

$$W = P + X, \qquad F = WAW^*.$$

For positive η, the candidates below have invertible A and hence rank $d = m + 1$. We use the coordinates

$$\begin{aligned}A &= c_0 e_0 e_0^* + \eta\omega_1\,\mathrm{diag}(0, \lambda_1, \ldots, \lambda_m) + \eta^2 V, \\ Xe_0 &= \eta^2 a_0, \qquad Xe_j = \eta a_j \quad (1 \leq j \leq m),\end{aligned} \tag{46}$$

where V is symmetric and $a_j \in QL^\infty$. The constant singular value is of order one and the other retained singular values are of order η. These different scales explain the different powers of η in X.

Expanding the product gives

$$WAW^* = A + XA + AX^* + XAX^* = c_0 J + \eta\omega_1 f + \eta^2 H + O_\infty(\eta^3),$$

where

$$H = V + c_0(a_0 \otimes e_0 + e_0 \otimes a_0) + \omega_1\sum_{j=1}^{m}\lambda_j(a_j \otimes u_j + u_j \otimes a_j).$$

The constant part of XAX^* is $O_\infty(\eta^4)$ and its nonconstant part is $O_\infty(\eta^3)$. It contributes nothing to H. Thus

$$PHP = V, \qquad QH\mathbf{1} = c_0 a_0, \qquad QHu_j = \omega_1\lambda_j a_j, \qquad QHQ = 0.$$

These four blocks will determine the second-order correction.

Expanding the loss gradient. The gradient of (18), taken in the kernel inner product, is

$$G = \frac{(R_{\omega,\eta} + b)\sigma(F) - R_{\omega,\eta}}{1+b}, \qquad \sigma(z) = (1+e^{-z})^{-1}.$$

To expand it before imposing the first-order score, write

$$R_{\omega,\eta} = 1 + \eta a + \eta^2 c + O_\infty(\eta^3), \qquad F = c_0 + \eta u + \eta^2 H + O_\infty(\eta^3),$$

with $a = \omega_1 k$ and $c = \omega_2 K^2$. Here K^2 is operator composition; the scalar products in the following Taylor expansion are pointwise.

At the constant optimum,

$$\sigma(c_0) = \frac{1}{1+b}, \qquad \sigma'(c_0) = c_b, \qquad \sigma''(c_0) = c_b \frac{b-1}{b+1}.$$

Substitution into the numerator of G gives

$$\begin{aligned}(R_{\omega,\eta} + b)\sigma(F) - R_{\omega,\eta} = \eta &\left((1+b)c_b u - \frac{b}{1+b}a\right) \\ &+ \eta^2 \left((1+b)c_b H + \frac{c_b(b-1)}{2}u^2 + c_b a u - \frac{b}{1+b}c\right) + O_\infty(\eta^3).\end{aligned}$$

Divide by $(1+b)c_b$ and set $u = \omega_1 f$. The first-order term is $-\eta\omega_1 r$. The nonlinear part of the second-order term is

$$\begin{aligned}\frac{b-1}{2(b+1)}u^2 + \frac{au}{1+b} &= \omega_1^2 \left\{ \frac{b-1}{2(b+1)} f^2 + \frac{(f+r)f}{1+b} \right\} \\ &= \omega_1^2 \left\{ \frac{1}{2} f^2 + \frac{rf}{1+b} \right\} = \omega_1^2 M_b.\end{aligned}$$

We have therefore obtained

$$\frac{G}{c_b} = -\eta\omega_1 r + \eta^2 \{H + \omega_1^2 M_b - \omega_2 K^2\} + O_\infty(\eta^3). \tag{47}$$

Solving the stationarity equations. Differentiation with respect to the two coordinates gives, for symmetric $\dot{A}$ and $\dot{X} : PL^2 \to QL^2$,

$$D_A \mathcal{L}[\dot{A}] = \langle W^* G W, \dot{A} \rangle, \qquad D_X \mathcal{L}[\dot{X}] = 2\langle QGWA, \dot{X} \rangle.$$

Since A is invertible for small positive η, stationarity is equivalent to

$$W^* G W = 0, \qquad QGW = 0.$$

The second equality gives $GW = PGW$. Since $X^* P = 0$, the first then gives

$$0 = W^* G W = (P+X)^* P G W = PGW.$$

Hence $GW = 0$. Symmetry also gives $W^* G = 0$. If $P_\eta = W(W^*W)^{-1}W^*$ is the orthogonal projection onto the candidate range, then

$$GP_\eta = P_\eta G = 0, \qquad G = (I - P_\eta) G (I - P_\eta).$$

Thus a solution of the chart equations has the normal gradient required by Lemma 2 in the full kernel space, not only among symmetric candidates.

We solve these equations after dividing by $c_b\eta^2$. Let $D = H + \omega_1^2 M_b - \omega_2 K^2$. Because $rP = Pr = 0$ and $QK^2P = 0$, the leading equations in the retained block and the individual columns are

$$\begin{aligned}W^* G W/(c_b\eta^2) &= PDP + O(\eta), \\ QG(e_0 + \eta^2 a_0)/(c_b\eta^2) &= QDe_0 + O_\infty(\eta), \\ QG(u_j + \eta a_j)/(c_b\eta^2) &= QDu_j - \omega_1 \mathsf{R} a_j + O_\infty(\eta).\end{aligned}$$

Using the four blocks of H, their values at $\eta = 0$ are

$$\begin{aligned} V + \omega_1^2 P M_b P - \omega_2 \mathsf{F}^2 &= 0, \\ c_0 a_0 + \omega_1^2 Q M_b \mathbf{1} &= 0, \\ \omega_1(\lambda_j I - \mathsf{R}) a_j + \omega_1^2 Q M_b u_j &= 0 \quad (1 \le j \le m). \end{aligned} \tag{48}$$

The divided equations extend analytically to zero on $\mathrm{Sym}_d \times (QL^\infty)^d$. Their derivative with respect to $V, a_0, a_1, \ldots, a_m$ is block diagonal, with blocks $I, c_0 I$, and $\omega_1(\lambda_j I - \mathsf{R})$.

The gap in Assumption 5 makes the last blocks invertible on QL^2:

$$\|(\lambda_j I - \mathsf{R})^{-1}\|_{\mathrm{op}} \le \frac{1}{|\lambda_j| - \|\mathsf{R}\|_{\mathrm{op}}}.$$

To use the same inverse in the uniform norm, take $h \in QL^\infty$ and let $z = (\lambda_j I - \mathsf{R})^{-1} h$. The identity $\lambda_j z = h + \mathsf{R} z$ gives

$$\begin{aligned} \|z\|_\infty &\le \frac{1}{|\lambda_j|}\big(\|h\|_\infty + \|r\|_\infty \|z\|_2\big) \\ &\le \frac{1}{|\lambda_j|}\left(1 + \frac{\|r\|_\infty}{|\lambda_j| - \|\mathsf{R}\|_{\mathrm{op}}}\right)\|h\|_\infty. \end{aligned}$$

Here we used $\|h\|_2 \le \|h\|_\infty$. The other two derivative blocks are invertible because $c_0 \neq 0$. The full derivative is therefore boundedly invertible in the Banach space used by the chart.

The analytic implicit-function theorem now gives bounded analytic coordinates $V(\eta), a_j(\eta)$. Solving (48) at zero and substituting into the blocks of H yields

$$\begin{aligned} PHP &= \omega_2 \mathsf{F}^2 - \omega_1^2 P M_b P, \\ QH\mathbf{1} &= -\omega_1^2 Q M_b \mathbf{1}, \\ QHu_j &= -\omega_1^2 \lambda_j (\lambda_j I - \mathsf{R})^{-1} Q M_b u_j, \qquad QHQ = 0. \end{aligned}$$

These are precisely the blocks of $H = \omega_2 \mathsf{F}^2 - \omega_1^2 \mathcal{S}_K(M_b)$ in (20). They prove the expansion (21). Analyticity in the uniform norm supplies its $O_\infty(\eta^3)$ remainder. Since the lag family is finite, the same sufficiently small contrast and remainder bound can be used for all its members.

C.3 Certifying the minimum

The constructed score is stationary. To prove that it is the unique global minimum, we compare its remaining gradient with the curvature of the loss. The score lies strictly inside $[-L, L]$ for small contrast because $|c_0| = \log b < L$, and it has rank exactly d because A is invertible for small positive η. Its smallest nonzero singular value is

$$s = \eta\omega_1 \min_j |\lambda_j| + O(\eta^2).$$

Indeed, $W^*W = I + X^*X = I + O(\eta^2)$, so the nonzero eigenvalues of WAW^* are those of $(W^*W)^{1/2} A (W^*W)^{1/2}$, whose leading terms are c_0 and $\eta\omega_1\lambda_j$.

The gradient expansion gives $G = -c_b \eta \omega_1 r + O_\infty(\eta^2)$, and therefore

$$\frac{\|G\|_{\mathrm{op}}}{s} \longrightarrow c_b \frac{\|\mathsf{R}\|_{\mathrm{op}}}{\min_j |\lambda_j|} < \frac{c_b}{2}.$$

On the entire score cap, the pointwise second derivative of the loss is at least

$$\kappa_\eta = \inf_{x,y} \frac{R_{\omega,\eta}(x,y) + b}{1+b} \frac{e^L}{(1+e^L)^2} \longrightarrow \kappa_0 = \frac{3b/2}{(1+3b/2)^2}.$$

The inequality

$$\frac{\kappa_0}{c_b} = \frac{3}{2}\frac{(1+b)^2}{(1+3b/2)^2} > \frac{2}{3}$$

leaves a positive gap between this curvature and the limiting gradient-to-singular-value ratio.

To make the uniform constant explicit, set $\gamma = \|\mathsf{R}\|_{\mathrm{op}}/\min_j |\lambda_j| < 1/2$ and $\Delta_0 = \kappa_0 - c_b\gamma > 0$. Choose $\eta_0 > 0$ so that, for every lag law in the fixed family,

$$\kappa_\eta \geq \kappa_0 - \Delta_0/4, \qquad \|G\|_{\mathrm{op}}/s \leq c_b\gamma + \Delta_0/4 \quad (0 < \eta < \eta_0).$$

Lemma 2 then gives

$$\mathcal{L}(H) - \mathcal{L}(F_{\omega,\eta}) \geq \frac{\Delta_0}{4}\|H - F_{\omega,\eta}\|_{\mathrm{HS}}^2$$

for every bounded score H of rank at most d, including nonsymmetric scores. This proves the global claim and uniqueness up to equality almost everywhere.

The expansion also gives the off-subspace coefficient stated in the main text. For an omitted eigenfunction u_ℓ,

$$\langle u_\ell, Hu_j\rangle = -\omega_1^2 \frac{\lambda_j}{\lambda_j - \lambda_\ell}\langle u_\ell, M_b u_j\rangle.$$

The denominator is nonzero because the inverse acts on the omitted space. Repeated retained eigenvalues therefore cause no problem. The formula includes nullspace directions by taking $\lambda_\ell = 0$; the constant-column correction and the zero omitted–omitted block are given above.

C.4 Proof of Corollary 1

The theorem constrains scores; the corollary requires a factorization satisfying row-norm bounds. The chart above writes

$$F_{\omega,\eta}(x,y) = \psi_\eta(x)^\top A_\eta \psi_\eta(y), \qquad A_\eta = -\log b\, e_0 e_0^\top + O(\eta),$$

where $\psi_{\eta,0} = 1 + O_\infty(\eta^2)$ and $\psi_{\eta,j} = u_j + O_\infty(\eta)$ for $j \geq 1$. The eigenvalue of A_η near $-\log b$ is separated from zero, its unit eigenvector is $e_0 + O(\eta)$, and all other eigenvalues are $O(\eta)$. Hence

$$|A_\eta| = \log b\, e_0 e_0^\top + O(\eta).$$

For $A_\eta = Q_\eta \Lambda_\eta Q_\eta^\top$, define

$$U_\eta(x) = |\Lambda_\eta|^{1/2} Q_\eta^\top \psi_\eta(x), \qquad V_\eta(y) = \mathrm{sgn}(\Lambda_\eta)|\Lambda_\eta|^{1/2} Q_\eta^\top \psi_\eta(y).$$

Their product is the required score, and at each argument their squared norms are

$$\begin{aligned}\|U_\eta(x)\|^2 = \|V_\eta(x)\|^2 &= \psi_\eta(x)^\top |A_\eta| \psi_\eta(x) \\ &= \log b\, \psi_{\eta,0}(x)^2 + O_\infty(\eta) = \log b + O_\infty(\eta).\end{aligned}$$

The finite set of bounded basis functions makes the error uniform in x and over the fixed lag family. Choose η so that $C\eta < \log(3/2)$. These factors then lie strictly inside the row-norm bound $L = \log(3b/2)$.

Conversely, any width-d factors satisfying the row bounds produce a score with

$$\mathrm{rank}(UV^\top) \leq d, \qquad |U(x)^\top V(y)| \leq \|U(x)\|\,\|V(y)\| \leq L.$$

The row-norm class is therefore contained in the theorem's score-bounded class, and it contains the theorem's unique minimizing score. That score minimizes both classes. This proves the corollary without identifying the two feasible sets.

D From a population minimum to sampled fits

We prove Theorem 3 by comparing losses before comparing scores. First we bound the change from population walk weights to sampled graph and corpus weights. We then associate each fitted matrix with a feasible population kernel and apply the quadratic gap of Theorem 2. The graphon, sufficiently small contrast, width, score cap, and finite lag family are fixed throughout. In particular, η does not shrink with n. The continuity and density conditions are those of Assumption 6; latent positions are used only in the proof.

D.1 COMPARING TRAINING WEIGHTS

Write $B = B_\eta$, $X_i \sim \zeta$, and $M_{ij} = B(X_i, X_j)$. The design matrix M includes its diagonal, although the sampled adjacency A has zero diagonal. Set

$$m_i = \sum_j M_{ij}, \quad v_M = \sum_i m_i, \quad P_M = \mathrm{diag}(m)^{-1} M, \quad \pi_i^M = m_i / v_M.$$

For the sampled graph, use

$$\widehat{d}_i = \sum_j A_{ij}, \quad \widehat{v} = \sum_i \widehat{d}_i, \quad \widehat{D} = \mathrm{diag}(\widehat{d}), \quad P_A = \widehat{D}^{-1} A, \quad \widehat{\pi}_i = \widehat{d}_i / \widehat{v}.$$

Define $H_\omega^A = \sum_\ell \omega_\ell P_A^\ell$ and the stationary positive-pair weights $C_\omega^A = \mathrm{diag}(\widehat{\pi}) H_\omega^A$. Since $B \in [a, 1-a]$ for a fixed $a > 0$, all sampled degrees are positive with probability tending to one. We work on that event.

Lemma 3 (Sampled walk weights)**.** *Let $e_n = \sqrt{\log n / n}$. For the fixed finite lag family, uniformly off the diagonal,*

$$(\widehat{D} H_\omega^A)_{ij} = B_0(X_i, X_j) R_{\omega,\eta}(X_i, X_j) + \omega_1 (A_{ij} - M_{ij}) + O_P(e_n), \tag{49}$$

$$n^2 C_{\omega,ij}^A = w(X_i) w(X_j) R_{\omega,\eta}(X_i, X_j) + \frac{\omega_1}{v} (A_{ij} - M_{ij}) + O_P(e_n), \tag{50}$$

$$n^2 \widehat{\pi}_i \widehat{\pi}_j = w(X_i) w(X_j) + O_P(e_n). \tag{51}$$

The diagonals of the normalized training weights are bounded in probability. Constants may depend on the fixed lag family and the probability bounds.

Proof. **Degrees and stationary probabilities.** Conditional on the latent positions, Hoeffding's inequality and a union bound give

$$\Pr\left\{ \max_i |\widehat{d}_i - m_i| > nu + 1 \mid X \right\} \le 2n \exp(-2nu^2).$$

The extra 1 accounts for the absent diagonal. Conditional on each X_i, the other M_{ij} are independent bounded variables with mean $d(X_i)$. The same argument gives

$$\max_i |\widehat{d}_i - m_i| / n = O_P(e_n), \qquad \max_i |m_i/n - d(X_i)| = O_P(e_n).$$

Averaging these degree estimates and applying Hoeffding to $d(X_i)$ also gives $\widehat{v}/n^2 = v + O_P(e_n)$; the same holds for v_M. Since $d(X_i) = v w(X_i)$ is bounded below,

$$n \widehat{\pi}_i = \frac{\widehat{d}_i / n}{\widehat{v}/n^2} = \frac{d(X_i)}{v} + O_P(e_n) = w(X_i) + O_P(e_n),$$

$$n^2 \widehat{\pi}_i \widehat{\pi}_j = w(X_i) w(X_j) + O_P(e_n),$$

uniformly in the indices. This proves (51).

Two steps. Introduce stationary ratios

$$R_\ell^A = P_A^\ell \, \mathrm{diag}(\widehat{\pi})^{-1}, \qquad R_\ell^M = P_M^\ell \, \mathrm{diag}(\pi^M)^{-1}.$$

Their entries are uniformly bounded on the degree event: this holds at lag one by the degree bounds and remains true after left multiplication by a stochastic matrix. For $i \neq j$,

$$(R_2^A)_{ij} = \frac{\widehat{v}}{\widehat{d}_i \widehat{d}_j} \sum_k \frac{A_{ik} A_{kj}}{\widehat{d}_k}.$$

Replace each $\widehat{d}_k$ by m_k. Since $|\widehat{d}_k^{-1} - m_k^{-1}| = O_P(e_n/n)$, the sum changes by $O_P(e_n)$. The outside factor is bounded and differs from $v_M/(m_i m_j)$ by $O_P(e_n)$.

For $k \notin \{i, j\}$, the products $A_{ik} A_{kj}$ use disjoint pairs of edges as k varies. Conditional on X, they are independent and have means $M_{ik} M_{kj}$. Therefore, for

$$S_{ij} = \sum_{k \notin \{i,j\}} \frac{A_{ik} A_{kj} - M_{ik} M_{kj}}{m_k},$$

we have

$$\Pr\{|S_{ij}| > u \mid X\} \le 2\exp\left(-\frac{2u^2}{\sum_{k\notin\{i,j\}} m_k^{-2}}\right) \le 2\exp(-2a^2nu^2).$$

The two excluded design terms cost $O(1/n)$. A union bound over i, j now gives $\|R_2^A - R_2^M\|_{\max,\text{off}} = O_P(e_n)$. The diagonal difference is bounded, but need not vanish: a two-step return uses the same sampled edge twice.

Higher lags and the population kernel. Conditional on X, each R_ℓ^M is deterministic and bounded. Replacing row degrees and applying rowwise concentration gives $\|(P_A - P_M)R_\ell^M\|_{\max} = O_P(e_n)$ for each fixed ℓ. Use

$$R_{\ell+1}^A - R_{\ell+1}^M = P_A(R_\ell^A - R_\ell^M) + (P_A - P_M)R_\ell^M.$$

At lag three, the diagonal discrepancy D_2 from lag two satisfies

$$|(P_A D_2)_{ij}| = |(P_A)_{ij}(D_2)_{jj}| = O_P(1/n),$$

because $\max_{ij}(P_A)_{ij} = O_P(1/n)$. Stochastic multiplication does not enlarge a maximum-entry bound. Induction therefore gives $\|R_\ell^A - R_\ell^M\|_{\max} = O_P(e_n)$ for each fixed $\ell \ge 3$.

The same induction compares R_ℓ^M with the graphon's stationary ratio R_ℓ^B evaluated at (X_i, X_j). At lag one,

$$(R_1^M)_{ij} = \frac{v_M M_{ij}}{m_i m_j}, \qquad R_1^B(X_i, X_j) = \frac{vB(X_i, X_j)}{d(X_i)d(X_j)},$$

so the degree and volume bounds suffice. In the induction step, replace the design ratio by its population value and average $B(X_i, z)R_\ell^B(z, X_j)$ over X_k. Conditional on X_i, X_j, all but two summands are independent bounded variables. Hoeffding and the same denominator bounds give an $O_P(e_n)$ error uniformly over pairs and fixed lags.

For $\ell \ge 2$, multiplying stationary ratios by $\widehat{d}_i\widehat{\pi}_j = B_0(X_i, X_j) + O_P(e_n)$ gives the population walk prediction up to $O_P(e_n)$ off the diagonal. At lag one, use the identity $\widehat{D}P_A = A$ rather than a concentration approximation. Its difference from M is exactly $A - M$. Averaging lags proves (49). Finally,

$$C_\omega^A = \frac{\widehat{D}H_\omega^A}{\widehat{v}}, \qquad \frac{n^2}{\widehat{v}} = \frac{1}{v} + O_P(e_n),$$

which proves (50). Bounded stationary ratios give the claimed diagonal bounds. A finite union makes all statements simultaneous over the prescribed lag family. □

D.2 Uniform comparison of objectives

The fitted matrix depends on the graph, so a fixed-score concentration bound is not enough. We prove a uniform bound over $\mathcal{F}_{d,L}^{(n)} = \{F : \operatorname{rank}(F) \le d,\ |F_{ij}| \le L\}$. The proof reference is the latent-design loss

$$\mathcal{L}_{X,\omega}(F) = \frac{1}{(1+b)n^2}\sum_{ij} w_i w_j\{R_{\omega,\eta}(X_i, X_j)\operatorname{softplus}(-F_{ij}) + b\operatorname{softplus}(F_{ij})\},$$

where $w_i = w(X_i)$. The learner does not evaluate this loss. Its weights need not sum to one at a finite latent design.

Lemma 4 (Uniform objective error)**.** *For fixed d, b, L and the finite lag family,*

$$\sup_{F\in\mathcal{F}_{d,L}^{(n)}} |\widehat{\mathcal{L}}_{N,\omega}(F) - \mathcal{L}_{X,\omega}(F)| = O_P\left(\sqrt{\frac{d\log n}{n}} + \sqrt{\frac{nd\log(nN)}{N}} + \frac{1}{N}\right). \tag{52}$$

Proof. **Graph error for a fixed score.** Let $\mathcal{L}_A$ denote expected SGNS loss conditional on the graph, with positive weights C_ω^A and negative weights $b\widehat{\pi}\widehat{\pi}^\top$, normalized by $1 + b$. Lemma 3 shows that the

remainder and negative-weight discrepancies contribute $O_P(e_n)$ uniformly over bounded scores; the diagonal contributes $O_P(1/n)$. Apart from fixed constants, the remaining graph error is

$$Z_F = \frac{1}{n^2}\sum_{i<j}(A_{ij} - M_{ij})\{\mathrm{softplus}(-F_{ij}) + \mathrm{softplus}(-F_{ji})\}.$$

Put $B_L = \log(1 + e^L)$. For a deterministic F conditional on X, the summands are independent and centered, and their coefficients are at most $2B_L/n^2$. Thus

$$\Pr\{|Z_F| > u \mid X\} \le 2\exp(-c_L n^2 u^2)$$

for a constant $c_L > 0$.

A finite covering of the score class. A Frobenius-radius-u covering of $\mathcal{F}_{d,L}^{(n)}$ has logarithmic size at most $Cnd\log(CnL/u)$. To see the scale, write an SVD $F = U\Sigma V^\top$, padded to d columns. Since $\|F\|_F \le nL$, cover U, V in Frobenius norm to accuracy $u/(6nL)$ and Σ to accuracy $u/3$. The product error is at most u by

$$\begin{aligned}\|U\Sigma V^\top - \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top\|_F \le \|U - \widetilde{U}\|_F\|\Sigma\|_{\mathrm{op}} + \|\Sigma - \widetilde{\Sigma}\|_F \\ + \|\widetilde{\Sigma}\|_{\mathrm{op}}\|V - \widetilde{V}\|_F.\end{aligned}$$

The two factor matrices have nd entries each and the singular-value vector has d entries. Standard Euclidean ball coverings give the stated logarithmic bound, with constants depending on fixed d. A maximal packing replaces the centers by feasible scores without changing the order.

Take $u = 1/n$ and denote this net by $\mathcal{N}_n$. Hoeffding and a union bound give, with conditional probability at least $1 - \delta$,

$$\begin{aligned}\max_{F\in\mathcal{N}_n}|Z_F| &\le C_L\sqrt{\frac{\log(2|\mathcal{N}_n|/\delta)}{n^2}} \\ &\le C_{d,L}\sqrt{\frac{nd\log(Cn^2L) + \log(2/\delta)}{n^2}}.\end{aligned}$$

Softplus is one-Lipschitz, and $\|F - H\|_{\max} \le \|F - H\|_F$. Moving from a net center to any score therefore costs $O(1/n)$. Together with the weight remainders, this proves

$$\sup_{F\in\mathcal{F}_{d,L}^{(n)}}|\mathcal{L}_A(F) - \mathcal{L}_{X,\omega}(F)| = O_P\left(\sqrt{\frac{d\log n}{n}}\right).$$

Uniformity permits substitution of a fitted, graph-dependent score.

Corpus error. Conditional on the graph, each normalized training-group loss lies in $[0, B_L]$ and is one-Lipschitz in the maximum norm. Different groups are independent; pairs within a trajectory need not be. For a Frobenius-radius-$1/N$ net $\mathcal{N}_N$,

$$\Pr\left\{\max_{F\in\mathcal{N}_N}|\widehat{\mathcal{L}}_{N,\omega}(F) - \mathcal{L}_A(F)| > u \mid A\right\} \le 2|\mathcal{N}_N|\exp(-2Nu^2/B_L^2).$$

Since $\log|\mathcal{N}_N| \le Cnd\log(CnLN)$, extending from the net gives

$$\sup_F|\widehat{\mathcal{L}}_{N,\omega}(F) - \mathcal{L}_A(F)| = O_P\left(\sqrt{\frac{nd\log(nN)}{N}} + \frac{1}{N}\right).$$

Stationary starts and the prescribed normalized context weights make the expected group loss equal to $\mathcal{L}_A$. Normalization bounds the loss of a whole trajectory by B_L; it does not make its pairs independent. Adding the graph error and taking a finite union over lag laws proves (52). □

D.3 From the latent design to the graphon

We next turn a fitted matrix into a feasible population kernel without assuming its factors are continuous. Partition the latent space into measurable cells E_i, each of ζ-mass $1/n$, and associate E_i with X_i. The cells can be chosen so that

$$\sum_i\int_{E_i}\|x - X_i\|^2\,d\zeta(x) \longrightarrow 0 \quad \text{in probability.} \tag{53}$$

For completeness, partition the cube into finitely many boxes Q_j of diameter at most ε. Write $p_j = \zeta(Q_j)$ and $\widehat{p}_j = n^{-1}\sum_i \mathbf{1}\{X_i \in Q_j\}$. Match as much mass as possible within each box, assigning mass $1/n$ to each sampled point. The unmatched mass is at most $\frac{1}{2}\sum_j |p_j - \widehat{p}_j|$. Assign it arbitrarily; nonatomicity allows every cell to have exactly the required mass. If D is the domain diameter and $T_n(x) = X_i$ on E_i, then

$$\int \|x - T_n(x)\|^2 \, d\zeta(x) \le \varepsilon^2 + \frac{D^2}{2}\sum_j |p_j - \widehat{p}_j|.$$

For a fixed partition, the sum tends to zero by the law of large numbers. Refining the partition sufficiently slowly with n proves (53).

Define the lifted kernel $F^{\uparrow}(x,y) = F_{ij}$ on $E_i \times E_j$. A matrix factorization extends by making its factor rows constant on the cells, so rank and the score cap are preserved. Let $a(x,y)$ be either population weight $w(x)w(y)R_{\omega,\eta}(x,y)$ or $w(x)w(y)$. Both are continuous. If ω_a is a modulus of continuity, then for every $\delta > 0$,

$$\begin{aligned}\int |a(x,y) - a(T_n x, T_n y)| \, d\zeta(x)d\zeta(y) \\ \le \omega_a(2\delta) + \frac{4\|a\|_\infty}{\delta^2}\int \|x - T_n x\|^2 \, d\zeta(x).\end{aligned}$$

First let n grow, then let δ decrease. Multiplying by the bounded loss terms gives

$$\sup_{F \in \mathcal{F}_{d,L}^{(n)}} |\mathcal{L}_{X,\omega}(F) - \mathcal{L}_{\omega,\eta}(F^{\uparrow})| \longrightarrow 0 \quad \text{in probability.}$$

Only the weights must be continuous; the fitted factors need not be. Related rank-constrained population loss comparisons appear in the broader embedding analysis of Davison and Austern (2023).

The population optimum $F_{\omega,\eta}$ is continuous under Assumption 6. Nonzero eigenfunctions of a continuous kernel are continuous, and the construction in Appendix C also works in the Banach space of continuous functions. The resolvent is bounded in the uniform norm because the omitted kernel maps L^2 into continuous functions and obeys the same inverse bound. Hence

$$G_{n,ij} = F_{\omega,\eta}(X_i, X_j)$$

is a feasible matrix, its lift converges to $F_{\omega,\eta}$ in $L^2(\zeta^2)$, and its population loss converges as well. This supplies a comparison score; it is not an initialization available to the learner.

D.4 Loss, scores, and predictions

Proof of Theorem 3. Write $F_* = F_{\omega,\eta}$ and suppress the fixed lag law in the losses. Let

$$a_n = \sup_F |\widehat{\mathcal{L}}(F) - \mathcal{L}_X(F)|, \qquad b_n = \sup_F |\mathcal{L}_X(F) - \mathcal{L}(F^{\uparrow})|.$$

The corpus-size assumption and the preceding comparisons give $a_n, b_n = o_P(1)$. Empirical near-minimality supplies $\xi_n = o_P(1)$ such that $\widehat{\mathcal{L}}(\widehat{F}) \le \widehat{\mathcal{L}}(G_n) + \xi_n$. Therefore

$$\begin{aligned}\mathcal{L}(\widehat{F}^{\uparrow}) &\le \mathcal{L}_X(\widehat{F}) + b_n \\ &\le \widehat{\mathcal{L}}(\widehat{F}) + a_n + b_n \\ &\le \widehat{\mathcal{L}}(G_n) + \xi_n + a_n + b_n \\ &\le \mathcal{L}(G_n^{\uparrow}) + \xi_n + 2a_n + 2b_n.\end{aligned}$$

Since $\mathcal{L}(G_n^{\uparrow}) \to \mathcal{L}(F_*)$, the global gap of Theorem 2 gives

$$c\|\widehat{F}^{\uparrow} - F_*\|^2_{L^2(\mu^2)} \le \mathcal{L}(\widehat{F}^{\uparrow}) - \mathcal{L}(F_*) = o_P(1).$$

Changing measure costs at most $w_{\min}^{-2}$. Finally,

$$\frac{1}{n^2}\|\widehat{F} - G_n\|_F^2 = \|\widehat{F}^{\uparrow} - G_n^{\uparrow}\|_{L^2(\zeta^2)}^2 \\ \leq 2\|\widehat{F}^{\uparrow} - F_*\|_{L^2(\zeta^2)}^2 + 2\|G_n^{\uparrow} - F_*\|_{L^2(\zeta^2)}^2 = o_P(1).$$

This proves score convergence.

Probabilities and NLL. The finite degree correction is $b\widehat{d}_i\widehat{d}_j/\widehat{v}$. Degree concentration gives $\widehat{d}_i\widehat{d}_j/\widehat{v} = B_0(X_i, X_j) + o_P(1)$ uniformly. The exponential is Lipschitz on $[-L, L]$, and clipping is one-Lipschitz. The decoded predictions therefore converge in mean square to the clipped population predictions. Choose the clipping interval so that its interior contains the population range. It then leaves the limiting prediction unchanged, while making Bernoulli log-loss uniformly Lipschitz. Cauchy–Schwarz converts mean-square prediction convergence into convergence of the average losses.

The remaining passage from averages at (X_i, X_j) to population risk uses the law of large numbers for bounded functions of latent pairs. It does not assume that all pairs are independent. For a bounded function of k distinct latent positions, summands using disjoint vertex sets are independent, and only $O(n^{2k-1})$ pairs of summands share an index. The variance of the normalized average is therefore $O(1/n)$. Repeated-index terms also have mass $O(1/n)$. Taking $k = 2$ proves the required risk limit; the same argument with $k = 4$ will be used for AUC.

AUC. Let $s = q_{\omega,\eta}$ and $p = B_\eta$. For fixed $\epsilon > 0$, a comparison can change only if one candidate's prediction error exceeds ϵ or the two population scores differ by at most 2ϵ. The first event has probability bounded by the mean-square error divided by ϵ^2. By the weighted identity (26), the second event matters only through

$$M(\epsilon) = \mathbb{E}\big[|p(e) - p(e')|\mathbf{1}\{|s(e) - s(e')| \leq 2\epsilon\}\big].$$

The bounded four-position averaging argument above and convergence of the prevalence to $v \in (0, 1)$ give an AUC-error bound of the form

$$C\epsilon^{-2}\|\widehat{q}^{\uparrow} - s\|_2^2 + CM(\epsilon) + o_P(1).$$

At fixed ϵ, the first term vanishes. As $\epsilon \downarrow 0$, dominated convergence gives

$$M(\epsilon) \longrightarrow \mathbb{E}\big[|p(e) - p(e')|\mathbf{1}\{s(e) = s(e')\}\big] = 0$$

by the theorem's tie condition. This proves AUC convergence. There are only finitely many lag laws. For any strict population comparison, eventually both metric errors are smaller than half its gap with probability tending to one. A finite union proves sign preservation for all such comparisons. □

The same argument applies if the empirical loss is no greater, up to $o_P(1)$, than that of a feasible reference $\widetilde{G}_n$ converging in mean square to $F_*(X, X)$. Bounded-score Lipschitz continuity makes the reference's loss consistent, and the same chain applies. This does not establish consistency of a generic spectral initialization, nor does it change the stationary negative distribution required by the population theorem.

E Additional experimental details

This appendix specifies each experiment in Section 5: the generator or network, the training objective, the decoder, the calibration map, and the evaluation sample.

E.1 Population metric comparisons

The population study has 60 fitted cells: the two omitted signs $r_\pm$, three contrasts $\eta \in \{.002, .001, .0005\}$, quadrature sizes 64 and 128, and five lag laws including one step. Each cell solves the population SGNS objective of Definition 5 at width two on the Gauss–Legendre grid and fits the same 12-knot monotone calibration family by population NLL. Risk is integrated over the population probabilities; AUC uses the continuous conditional integral of Appendix A rather than a

grid comparison. The 72/72 sign count in Section 5.1 uses the finer grid, the four nonzero lag masses, both omitted signs, and all three metrics; the coarser grid and larger contrasts give the same signs, with quadratic coefficients within a few percent of the predictions. Agreement after grid refinement supports the calculations but is not a proof.

The AUC prediction is displayed in the convention of Theorem 1,

$$\widehat{\Delta\,\mathrm{AUC}}/\eta^2 = -\frac{t^2 A_A + 2tC_A}{4v(1-v)}, \qquad v = .12,$$

and the NLL predictions in the convention of (9) and (16). The 24-cell rotation check and the 12-cell fourth-order calibration check of Section 5.1 use different generators and are reported in Appendix F.

E.2 Sampled-graph comparison

The sampled-graph study separates numerical population optimization, graph sampling, and corpus sampling. Its generator, fitting configuration, evaluation, and sign-agreement rule were fixed before any sampled fit was run; the deviations from that plan are stated below.

Conventions for this appendix. Population references are numerical optima of the discretized objective at the stated finite contrast; they are not values of the weak-contrast expansion and the contrasts are not certified to lie in its regime. Ranges over draws are empirical 2.5/97.5 percentiles of eight draw-level differences, not confidence intervals for a mean. The sign-agreement rule (at least 7 of 8 draws with the population sign and a range on that side of zero) is the pre-registered protocol criterion, not a significance test.

Generator and names. Let μ be uniform on $[-1,1]$, $\delta = .6$, $v = .06$, and $\epsilon = .4$. Set

$$d\zeta(x) = (1+\delta x)d\mu(x), \qquad w(x) = (1+\delta x)^{-1},$$
$$u_1(x) = \begin{cases} -1 & x < 0, \\ 1 & x \geq 0, \end{cases} \qquad u_2(x) = \begin{cases} -1 & |x| < 1/2, \\ 1 & |x| \geq 1/2, \end{cases} \qquad u_3 = u_1 u_2.$$

These modes are centered and orthonormal under μ. For $s \in \{-1, 1\}$, define

$$B_{\eta,s}(x,y) = vw(x)w(y)\left[1 + \eta\left(u_1(x)u_1(y) + s\epsilon\{u_3(x)u_3(y) - u_2(x)u_2(y)\}\right)\right]. \tag{54}$$

Graphon D uses $s = +1$, and graphon C uses $s = -1$. The names refer to their measured regimes at $\eta = .3$, $t = .05, .10, .20$: walking lowers AUC for D and raises it for C, while increasing native and calibrated loss for both. They are not synonyms for a favorable or unfavorable sign of an asymptotic coefficient at every contrast. Both graphons have the same degree function vw and residual spectrum $1, .4, -.4$, and differ in the assignment of the omitted eigenvalues.

The residual is constant on four blocks, but the degree weights vary continuously within them. The baseline probability therefore has a continuous distribution. To verify this, put $a = 1-\delta$, $b_0 = 1+\delta$, and $P = (1+\delta X)(1+\delta Y)$. Under ζ^2, its density is

$$f_P(p) = \frac{p}{4\delta^2}\log\frac{h(p)}{l(p)}, \qquad l(p) = \max\{a, p/b_0\}, \quad h(p) = \min\{b_0, p/a\}, \qquad a^2 \leq p \leq b_0^2.$$

Thus $Z = B_0 = v/P$ has a continuous density. The residual kernel nevertheless has jumps at the block boundaries, whereas Assumption 6 asks for a continuous kernel.

Population reference. The main grid uses 96 quadrature nodes, with Gauss–Legendre quadrature on each of the four intervals determined by $-1, -1/2, 0, 1/2, 1$. With nodes x_i and μ-weights m_i, form $K_{ij} = k(x_i, x_j)$ and

$$R_t = J + (1-t)\eta K + t\eta^2 K\,\mathrm{diag}(m)K, \qquad C_{ij} = m_i m_j R_{t,ij}, \qquad N_{ij} = 5m_i m_j.$$

The objective is normalized by $\sum_{ij}(C_{ij} + N_{ij}) = 6$. The population and sampled-graph fits use the same projected SGNS solver, width two, five negatives, and squared factor-row bound $L = \log(7.5)$. The reference uses three restarts, up to 4000 projected iterations, requested tolerances 10^{-7} for

Table 4: Population-reference diagnostics on the 96-node quadrature design. The pair-score bound is $L = \log(7.5) \simeq 2.014903$; there are 192 factor rows. An active row constraint does not imply that any pair score attains L. "Polish" says whether the unconstrained refinement stayed in the row-norm class; rejected refinements are not used.

Graphon	η	t	Active rows	$\max \lvert F \rvert$	Polish
D	0.3	0.05	0	1.951224	accepted
D	0.4	0.00	192	2.014903	rejected
D	0.4	0.05	192	2.014903	rejected
D	0.4	0.10	192	2.014903	rejected
D	0.4	0.20	192	2.014903	rejected
C	0.3	0.05	0	1.951224	accepted
C	0.4	0.00	192	2.014903	rejected
C	0.4	0.05	192	2.014903	rejected
C	0.4	0.10	192	2.014903	rejected
C	0.4	0.20	192	2.014903	rejected

projected gradient and 10^{-12} for relative loss, followed by an unconstrained L-BFGS-B refinement. The refinement is retained only if its factor rows remain feasible.

The independent input and context factors give $F = UV^\top$. Evaluation uses $S = (F + F^\top)/2$, whose rank can be up to four, and $q_{ij} = 5B_0(x_i, x_j)e^{S_{ij}}$. Population metrics use the ζ^2 quadrature weights $m_i m_j(1 + \delta x_i)(1 + \delta x_j)$, including quadrature diagonal cells. Unlike the population comparisons of Section 5.1, this AUC is a grid calculation, not a separate continuous comparison-boundary integral. Scans at grid sizes 48 and 96 give the same regimes and magnitudes within ten percent, except the smallest $\eta = .4$ AUC cell.

Population constraint status. Table 4 distinguishes factor-row activity from pair-score saturation. At $\eta = .4$, $t \le .20$, both numerical references reach $\max |S| = L$ and retain the constrained solution, outside the interior conclusion of Corollary 1. At $\eta = .3$, $t \le .20$, no row constraint is active and the refinement is accepted for both graphons; at $t = .35$ the row bound is active for D at both contrasts while $\max |F|$ stays below L.

Sampled graphs and objective. For each sign and contrast $\eta \in \{.3, .4\}$, sample $n \in \{800, 1600, 3200\}$ latent positions independently from ζ and Bernoulli edges from (54). Eight seeds are used at each setting. The inverse-CDF sampler has

$$X = \frac{-1 + \sqrt{(1-\delta)^2 + 4\delta U}}{\delta}, \qquad U \sim \mathrm{Unif}[0, 1].$$

For adjacency A, degrees d_i, and volume $V_G = \sum_i d_i$, the expected training weights are

$$C_t = \frac{(1-t)A + tA\,\mathrm{diag}(d)^{-1}A}{V_G}, \qquad N = 5\pi\pi^\top, \qquad \pi_i = d_i/V_G.$$

These weights average over the choice of training pairs on a fixed graph. The graph itself is still sampled. We fit $t \in \{0, .05, .10, .20, .35\}$ from three restarts, with 4000 iterations at most and declared tolerances 10^{-6} and 10^{-9}. All 480 width-two expected-objective fits satisfy the stopping rule.

The decoder on distinct unordered pairs is

$$q_{ij} = \frac{5d_i d_j}{V_G} e^{S_{ij}}, \qquad S = (UV^\top + VU^\top)/2.$$

Both endpoints of every evaluated pair have positive training degree. Using $B_{\eta,s}(X_i, X_j)$ as the target averages AUC and excess NLL over possible test labels. A separate evaluation draws an independent Bernoulli graph on the same positions, and another independent graph supplies calibration labels. The main tables use the known probabilities. Metrics evaluated on the realized labels have the same mean sign as the true-probability metrics in every cell; individual draws differ only where the change is within label noise.

The reported native probabilities are clipped to $[10^{-6}, 1-10^{-6}]$. No native prediction hits those thresholds in any of the 480 expected-objective or 160 corpus-control fits. Sampled width-two fits reach the pair-score cap numerically. Their mean degree is close to $.06(n-1)$, about 192 at $n = 3200$; the first recorded draw at that size has mean degree 193.35. These are dense graphs with a deliberately small embedding width.

Calibration and aggregation. The dictionary calibrator fits

$$g_\theta(q) = \sigma\{\mathrm{logit}(q) + \psi(q)^\top \theta\},$$

where ψ contains linear hat functions at twelve geometrically spaced knots between $v/(1+\delta)^2$ and $v/(1-\delta)^2$, extended constantly outside this interval. Coefficients lie in $[-10, 10]$ and every offset slope is at least -3.9; consequently the unrounded map is strictly increasing. The optimizer is SLSQP with tolerance 10^{-12}, at most 2000 iterations, and no ridge penalty. The separate Platt map is $\sigma(a + s\,\mathrm{logit}(q))$, with $a \in [-15, 15]$ and $s \in [.05, 20]$, fitted by L-BFGS-B. Both classes contain the identity. This Platt feature differs from calibration directly on F in Proposition B.2.

At most one million uniformly subsampled calibration dyads are used, selected without replacement with a fixed seed; all eligible dyads are evaluated. Population reference calibrators use true probabilities as labels. Sampled-model calibrators use the independent calibration graph. All reported dictionary fits in the primary width-two runs return a successful status. AUC is computed on the native ordering; calibration is not credited with an AUC change. Output rounding and clipping are numerical conventions, not proofs of exact AUC invariance for the implemented probabilities.

For each draw and metric, subtract the $t = 0$ value fitted on the same graph. Tables report the mean, sign agreement, and draw-percentile range defined above. The population scan precedes the chosen transfer runs; the intended minimum population effect was $2 \cdot 10^{-4}$. The retained $\eta = .4$, C, $t = .05$ comparison is below that value ($1.09 \cdot 10^{-4}$) and remains reported.

Expected-objective results. Table 5 gives every metric at the three headline lag masses. Table 6 includes both contrasts and all graph sizes. At $\eta = .3$, $t = .05$, D has population AUC change $-4.88 \cdot 10^{-4}$ and sampled mean $-2.84 \cdot 10^{-4}$; C has population change $+3.97 \cdot 10^{-4}$ and sampled mean $+4.52 \cdot 10^{-5}$. Both directions agree in all eight largest graphs at $t = .05, .10$. For C at $t = .05$, the counts are $5/8, 5/8, 8/8$ across increasing sizes; at $t = .10$ they are $3/8, 3/8, 8/8$. At $t = .20$, the largest-size count falls to $3/8$ and the mean is negative. At $\eta = .4$, C's positive population direction is absent at every tested size. D's negative direction persists.

All three loss signs agree at $n \geq 1600$ in the expected-objective runs, but native NLL often improves at $n = 800$ when the population comparison worsens.

Table 5: Expected-objective comparisons at $\eta = .3$, $n = 3200$, width two. All changes are multiplied by 10^4 and evaluated against true edge probabilities. Ranges are the empirical 2.5th and 97.5th percentiles of eight draw-level changes, not confidence intervals for their mean. Calibration is fitted separately at each lag mass; both calibrators are applied to $\mathrm{logit}(q)$ as specified in the text.

Graphon	t	Metric	Population	Mean	Draw range	Agree
D	0.05	AUC	-4.884	-2.839	[-3.085, -2.685]	8/8
D	0.05	Native NLL	+0.037	+0.067	[+0.055, +0.073]	8/8
D	0.05	Cal. NLL	+0.049	+0.041	[+0.031, +0.050]	8/8
D	0.05	Platt NLL	+0.028	+0.113	[+0.106, +0.119]	8/8
D	0.10	AUC	-10.325	-5.997	[-6.521, -5.691]	8/8
D	0.10	Native NLL	+0.153	+0.179	[+0.157, +0.192]	8/8
D	0.10	Cal. NLL	+0.160	+0.109	[+0.089, +0.124]	8/8
D	0.10	Platt NLL	+0.120	+0.260	[+0.247, +0.272]	8/8
D	0.20	AUC	-22.403	-13.336	[-14.466, -12.651]	8/8
D	0.20	Native NLL	+0.613	+0.555	[+0.509, +0.579]	8/8
D	0.20	Cal. NLL	+0.539	+0.332	[+0.298, +0.359]	8/8
D	0.20	Platt NLL	+0.501	+0.673	[+0.647, +0.697]	8/8
C	0.05	AUC	+3.966	+0.452	[+0.274, +0.732]	8/8
C	0.05	Native NLL	+0.041	+0.068	[+0.055, +0.077]	8/8
C	0.05	Cal. NLL	+0.114	+0.127	[+0.113, +0.135]	8/8
C	0.05	Platt NLL	+0.034	+0.117	[+0.107, +0.121]	8/8
C	0.10	AUC	+6.656	+0.591	[+0.257, +1.147]	8/8
C	0.10	Native NLL	+0.160	+0.184	[+0.156, +0.201]	8/8
C	0.10	Cal. NLL	+0.312	+0.282	[+0.253, +0.297]	8/8
C	0.10	Platt NLL	+0.130	+0.269	[+0.247, +0.277]	8/8
C	0.20	AUC	+7.864	-0.209	[-0.827, +0.830]	3/8
C	0.20	Native NLL	+0.627	+0.565	[+0.511, +0.598]	8/8
C	0.20	Cal. NLL	+0.954	+0.681	[+0.615, +0.714]	8/8
C	0.20	Platt NLL	+0.521	+0.691	[+0.649, +0.708]	8/8

Table 6: Expected-objective AUC changes across graph sizes; values multiplied by 10^4. Each entry gives the mean and sign agreement over eight draws. This finite sequence is not a convergence-rate estimate.

Graphon	η	t	Population	$n = 800$	$n = 1600$	$n = 3200$
D	0.3	0.05	-4.884	-1.260 (8/8)	-2.107 (8/8)	-2.839 (8/8)
D	0.3	0.10	-10.325	-2.655 (8/8)	-4.439 (8/8)	-5.997 (8/8)
D	0.3	0.20	-22.403	-5.964 (8/8)	-9.866 (8/8)	-13.336 (8/8)
D	0.4	0.05	-3.886	-3.738 (8/8)	-4.275 (8/8)	-4.508 (8/8)
D	0.4	0.10	-7.871	-7.749 (8/8)	-8.989 (8/8)	-9.564 (8/8)
D	0.4	0.20	-16.512	-16.623 (8/8)	-19.873 (8/8)	-21.695 (8/8)
C	0.3	0.05	+3.966	+0.022 (5/8)	+0.078 (5/8)	+0.452 (8/8)
C	0.3	0.10	+6.656	-0.104 (3/8)	-0.064 (3/8)	+0.591 (8/8)
C	0.3	0.20	+7.864	-0.774 (2/8)	-1.103 (0/8)	-0.209 (3/8)
C	0.4	0.05	+1.091	-2.044 (0/8)	-1.535 (0/8)	-0.641 (0/8)
C	0.4	0.10	+2.010	-4.289 (0/8)	-3.338 (0/8)	-1.590 (0/8)
C	0.4	0.20	+3.592	-9.428 (0/8)	-7.914 (0/8)	-4.725 (0/8)

Finite-corpus control. At $n = 3200$, the control draws two million stationary positive pairs per fit, with lag law $(1 - t, t)$, and five negatives per positive. Corpora at different lag masses are independent. Duplicate ordered positive or negative examples are aggregated with multiplicities, preserving the frozen-corpus objective. The same row-norm solver and stopping rule produce 160 converged width-two fits. Table 8 reports their values and Table 7 the complete sign counts for both training objectives. At $\eta = .3$, C's AUC agreement is $3/8, 1/8, 0/8$ at $t = .05, .10, .20$; its mean at .05 is $-2.50 \cdot 10^{-5}$ rather than the population gain. D's negative direction persists. All loss signs agree from $t = .10$; at .05, D has seven of eight agreements for native and dictionary-calibrated NLL. Thus the finite corpus preserves the negative AUC response but does not consistently recover the positive one at this budget.

Table 7: Complete $n = 3200$ sign agreement with the corresponding finite-contrast population reference. Each count uses eight draws. "Dict." denotes the dictionary-calibrated NLL; the other loss columns are native and Platt-calibrated NLL. Expected weights and sampled corpora are separate training experiments.

			Expected objective				$2 \cdot 10^6$-pair corpus			
Graphon	η	t	AUC	Native	Dict.	Platt	AUC	Native	Dict.	Platt
D	0.3	0.05	8/8	8/8	8/8	8/8	8/8	7/8	7/8	8/8
D	0.3	0.10	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.3	0.20	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.3	0.35	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.4	0.05	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.4	0.10	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.4	0.20	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
D	0.4	0.35	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
C	0.3	0.05	8/8	8/8	8/8	8/8	3/8	8/8	8/8	8/8
C	0.3	0.10	8/8	8/8	8/8	8/8	1/8	8/8	8/8	8/8
C	0.3	0.20	3/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
C	0.3	0.35	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
C	0.4	0.05	0/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
C	0.4	0.10	0/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
C	0.4	0.20	0/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
C	0.4	0.35	0/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8

Table 8: Two-million-pair corpus comparisons at $\eta = .3$, $n = 3200$, width two. All changes are multiplied by 10^4 and evaluated against true edge probabilities. Ranges are the empirical 2.5th and 97.5th percentiles of eight draw-level changes, not confidence intervals for their mean. Calibration is fitted separately at each lag mass; both calibrators are applied to $\mathrm{logit}(q)$ as specified in the text.

Graphon	t	Metric	Population	Mean	Draw range	Agree
D	0.05	AUC	-4.884	-2.388	[-3.688, -1.189]	8/8
D	0.05	Native NLL	+0.037	+0.041	[-0.014, +0.086]	7/8
D	0.05	Cal. NLL	+0.049	+0.026	[-0.019, +0.055]	7/8
D	0.05	Platt NLL	+0.028	+0.094	[+0.030, +0.127]	8/8
D	0.10	AUC	-10.325	-6.314	[-8.058, -4.791]	8/8
D	0.10	Native NLL	+0.153	+0.220	[+0.137, +0.306]	8/8
D	0.10	Cal. NLL	+0.160	+0.161	[+0.110, +0.236]	8/8
D	0.10	Platt NLL	+0.120	+0.311	[+0.248, +0.380]	8/8
D	0.20	AUC	-22.403	-13.543	[-15.199, -11.613]	8/8
D	0.20	Native NLL	+0.613	+0.625	[+0.580, +0.750]	8/8
D	0.20	Cal. NLL	+0.539	+0.407	[+0.353, +0.503]	8/8
D	0.20	Platt NLL	+0.501	+0.766	[+0.720, +0.878]	8/8
C	0.05	AUC	+3.966	-0.250	[-1.264, +0.688]	3/8
C	0.05	Native NLL	+0.041	+0.112	[+0.060, +0.196]	8/8
C	0.05	Cal. NLL	+0.114	+0.150	[+0.087, +0.247]	8/8
C	0.05	Platt NLL	+0.034	+0.158	[+0.104, +0.248]	8/8
C	0.10	AUC	+6.656	-1.071	[-2.148, +0.683]	1/8
C	0.10	Native NLL	+0.160	+0.285	[+0.166, +0.375]	8/8
C	0.10	Cal. NLL	+0.312	+0.371	[+0.249, +0.436]	8/8
C	0.10	Platt NLL	+0.130	+0.374	[+0.246, +0.472]	8/8
C	0.20	AUC	+7.864	-2.743	[-3.752, -2.038]	0/8
C	0.20	Native NLL	+0.627	+0.683	[+0.622, +0.761]	8/8
C	0.20	Cal. NLL	+0.954	+0.772	[+0.700, +0.867]	8/8
C	0.20	Platt NLL	+0.521	+0.823	[+0.755, +0.915]	8/8

Scope. A partial width-four control ($n \leq 1600$) and an exploratory decomposition of the AUC changes under reversal of the omitted modes were also run; the decomposition was added after inspecting results, and neither is used here.

Table 9: Final sampled-corpus norm-constrained SGNS: mean paired test changes from one-step training across five fixed graphs and three splits per graph. Probability loss uses the clipped native decoder; calibration uses a bounded positive-slope logistic map.

Lag law	ΔAUC	Δnative NLL	Δcal. NLL
$(.90, .10)$	+0.0166	-0.0086	+0.0143
$(.96, .04)$	+0.0083	-0.0049	+0.0021
$T = 2$	+0.0241	-0.0132	+0.0083
$T = 5$	+0.0261	-0.0202	-0.0018

Table 10: All final sampled-corpus controls. Entries are paired mean test changes from the model's own one-step fit on the same five graphs and three splits. Native losses use clipping. Rows differ in representation, feasible class, or update rule and are not replicas of the same population optimum.

Model	Lag law	ΔAUC	ΔNLL	Δcal. NLL
Norm-constrained SGNS	$(.90, .10)$	+0.0166	-0.0086	+0.0143
Norm-constrained SGNS	$(.96, .04)$	+0.0083	-0.0049	+0.0021
Norm-constrained SGNS	$T = 2$	+0.0241	-0.0132	+0.0083
Norm-constrained SGNS	$T = 5$	+0.0261	-0.0202	-0.0018
Spectral	$(.90, .10)$	+0.0040	-0.0127	+0.0003
Spectral	$(.96, .04)$	+0.0002	-0.0032	-0.0001
Spectral	$T = 2$	+0.0310	-0.0992	-0.0074
Spectral	$T = 5$	+0.0710	-0.2394	-0.0170
Unregularized SGNS	$(.90, .10)$	+0.2731	-1.7018	-0.0423
Unregularized SGNS	$(.96, .04)$	+0.2643	-1.5592	-0.0460
Unregularized SGNS	$T = 2$	+0.2729	-1.7905	-0.0420
Unregularized SGNS	$T = 5$	+0.3168	-2.2917	-0.0466
Row-averaged control	$(.90, .10)$	+0.0004	+0.0005	+0.0000
Row-averaged control	$(.96, .04)$	-0.0000	-0.0000	-0.0000
Row-averaged control	$T = 2$	+0.0010	+0.0249	+0.0003
Row-averaged control	$T = 5$	+0.0028	+0.0027	+0.0000

E.3 Real-network training and evaluation

Each of the five networks uses three fixed seeds to split unordered node pairs. The mask selects 60% for training and partitions the rest into diagnostic, calibration, and test sets. Only observed edges within the training mask become adjacency entries. Selection is independent of the edge label, unlike retaining only positive edges. Pairs incident to isolated training nodes are excluded; the exclusion is recorded for every set because it changes the candidate population. Repeated splits of one network are not independent network samples.

Positive and negative examples are sampled once, and their counts remain fixed during fitting. The norm-constrained score is $F = UV^\top$ with width 16, five negatives and squared row bound $\log(7.5)$. Tables 9, 11, and 12 give the comparisons behind Table 3: every lag law improves mean AUC and native NLL on every graph, while the calibrated column changes sign across graphs and the within-window effect of calibration is a deterioration on every lag law, largest on the graphs with the fewest calibration labels. Table 10 adds three controls that change the feasible class or the update rule: a spectral embedding, an unregularized SGNS fit, and a row-averaged update. They need not fit the same population score, and large scores in the unregularized fit do not establish separability.

To compute native loss, the implementation first symmetrizes the score and then restores the degree and sampling factors:

$$S_{ij} = \tfrac{1}{2}(F_{ij} + F_{ji}), \quad q_{ij}^{\mathrm{raw}} = \frac{b\, d_i d_j}{.6\ \mathrm{vol}(G)} e^{S_{ij}}, \quad q_{ij}^{\varepsilon} = \min\{1-\varepsilon, \max\{\varepsilon, q_{ij}^{\mathrm{raw}}\}\}, \quad \varepsilon = 10^{-12}.$$

The factor .6 is the training-dyad inclusion probability. NLL is computed stably from the logit of q^{ε}. Ranking uses that logit, or its positive affine recalibration, rather than floating-point probabilities saturated by the sigmoid. Thus the same predictor is evaluated, with ties defined before numerical sigmoid saturation. The symmetrized readout can have rank $2d$ even though F has rank at most d.

Table 11: Final norm-constrained sampled-corpus SGNS, paired changes from one step averaged over three splits of each graph. The calibration comparison calibrates both windows.

Graph	Lag law	ΔAUC	ΔNLL	Δcal. NLL
karate	$(.90, .10)$	+0.0161	-0.0213	+0.0503
dolphins	$(.90, .10)$	+0.0277	-0.0084	-0.0027
lesmis	$(.90, .10)$	+0.0134	-0.0067	+0.0305
polbooks	$(.90, .10)$	+0.0125	-0.0041	-0.0035
football	$(.90, .10)$	+0.0134	-0.0026	-0.0032
karate	$(.96, .04)$	+0.0045	-0.0126	+0.0176
dolphins	$(.96, .04)$	+0.0133	-0.0039	-0.0011
lesmis	$(.96, .04)$	+0.0074	-0.0039	-0.0017
polbooks	$(.96, .04)$	+0.0072	-0.0024	-0.0021
football	$(.96, .04)$	+0.0090	-0.0016	-0.0021
karate	$T = 2$	+0.0263	-0.0278	+0.0317
dolphins	$T = 2$	+0.0483	-0.0129	-0.0038
lesmis	$T = 2$	+0.0159	-0.0159	+0.0210
polbooks	$T = 2$	+0.0168	-0.0063	-0.0047
football	$T = 2$	+0.0133	-0.0030	-0.0025
karate	$T = 5$	+0.0355	-0.0332	+0.0424
dolphins	$T = 5$	+0.0607	-0.0148	-0.0046
lesmis	$T = 5$	+0.0125	-0.0425	-0.0436
polbooks	$T = 5$	+0.0211	-0.0085	-0.0051
football	$T = 5$	+0.0005	-0.0018	+0.0017

Table 12: Within-window correction effect for final norm-constrained sampled-corpus SGNS: calibrated minus native test NLL. Positive differences are deterioration after calibration. This differs from comparing two calibrated windows and does not contradict minimization of calibration-sample loss.

Lag law	Mean correction effect
$(.90, .10)$	+0.03556
$(.96, .04)$	+0.01965
$T = 2$	+0.03419
$T = 5$	+0.03098

The real-network calibrator fits an intercept in $[-15, 15]$ and slope in $[.05, 20]$ to $\mathrm{logit}(q^{\varepsilon})$, using only calibration labels. The identity lies inside this box, so minimizing calibration loss cannot do worse than the identity on that sample. Test loss can still increase, and Table 12 shows that it does. This is neither the 12-feature family used for the synthetic graphons nor affine calibration of F in the constant-degree theorem.

E.4 Larger networks

The larger-network comparison of Section 5.3 runs the same pipeline on `email-Eu-core` (986 nodes), `ca-GrQc` (4,158), `p2p-Gnutella08` (6,299), and `ca-HepTh` (8,638), treated as undirected simple graphs. Each graph has five label-independent splits with the same proportions as above. The frozen corpus has 500,000 positive pairs, so these fits see far fewer pairs per node than the five small networks did. Width, negatives, row bound, decoder, and the bounded Platt map are unchanged. Because the initial scale that suits small graphs places a fit of this size at a saddle near $F = 0$, the initial factors are scaled so that $|F|$ starts near .5, the projected-gradient tolerance is 10^{-7}, and each fit is the best of three restarts with at most 3,000 iterations. Windows compared with window one are the mixture $(.9, .1)$ and uniform windows of two, three, five, and ten, chosen per graph after a two-split screen; they were not fixed in advance, which is why Section 5.3 reads the result as existence of both directions rather than as a test.

Table 13 lists every confirmation run. On the two collaboration graphs, window five improves all three metrics on all five splits; window ten still improves AUC and native NLL on every split, at roughly half the AUC gain, while calibrated NLL worsens on every split. On `email-Eu-core` the

Table 13: All confirmation runs on the larger networks: paired changes from window one over five splits, best of three restarts per fit. Means are multiplied by 10^4. Every fit converged except one baseline/window pair on `p2p-Gnutella08`.

		Splits improving			Mean change ($\times 10^4$)		
Graph	Window	AUC	Native	Cal.	ΔAUC	Δnative	Δcal.
`ca-HepTh` (8,638)	$T = 5$	5/5	5/5	5/5	+46.6	-0.29	-0.12
`ca-HepTh` (8,638)	$T = 10$	5/5	5/5	0/5	+23.2	-0.27	+0.07
`ca-GrQc` (4,158)	$T = 5$	5/5	5/5	5/5	+36.0	-0.40	-0.22
`ca-GrQc` (4,158)	$T = 10$	5/5	5/5	0/5	+18.5	-0.33	+0.11
`email-Eu-core` (986)	$(.9, .1)$	2/5	3/5	0/5	-0.41	+4.13	+5.43
`email-Eu-core` (986)	$T = 2$	0/5	2/5	0/5	-9.11	+3.93	+9.56
`email-Eu-core` (986)	$T = 3$	0/5	1/5	0/5	-24.4	+2.24	+10.9
`p2p-Gnutella08` (6,299)	$(.9, .1)$	3/5	0/5	0/5	+4.01	+0.08	+0.15
`p2p-Gnutella08` (6,299)	$T = 2$	4/5	2/5	0/5	+7.60	+0.00	+0.24

mean of every metric worsens at every window, and AUC worsens on every split from window two on. On `p2p-Gnutella08` native NLL worsens on all five splits at the $(.9, .1)$ mixture while AUC improves on three and on average; one of its five baseline/window pairs did not converge. In the two-split screen, `polblogs`, `facebook`, and `wiki-Vote` worsened on AUC and native NLL at every window, increasingly with the window; those runs used one restart and a lower iteration cap and are not tabulated.

F Additional population checks

Two population calculations test the approximation orders claimed in Sections 4 and 3.3, separately from the matched-family comparisons. The first checks that including the change of the fitted subspace leaves a remainder cubic in contrast, while holding the subspace fixed leaves a quadratic one (Table 14). The second checks that, at constant degrees, population logistic calibration reduces the window-two loss penalty from second to fourth order (Table 15, Figure 2).

The moving-subspace check evaluates polynomial eigenfunctions at sixteen midpoint nodes, orthonormalizes them, and applies a fixed orthogonal rotation to six nonconstant functions. The residual eigenvalues are $.45, -.30, .08, -.05, .025, -.015$; the score rank is three, retaining the two largest absolute eigenvalues. The 24 cells vary contrast, negative-sampling count, and lag distribution. Each solves the rank-manifold stationary equations and checks the finite-dimensional normal-gradient certificate in floating point.

Table 14: Population-grid checks of the general expansion. The grid has 16 nodes, with rotated polynomial eigenfunctions and signed residual spectrum; the score rank is three. This slice uses $b = 5$ and $\omega = (1/2, 1/2)$. Errors are RMS differences from a numerically certified minimum, not sampled-graph errors.

Contrast η	Moving-subspace error	Frozen-subspace error	Error$/\eta^3$
0.08	2.155×10^{-5}	2.849×10^{-4}	0.0421
0.04	2.644×10^{-6}	6.935×10^{-5}	0.0413
0.02	3.276×10^{-7}	1.711×10^{-5}	0.0410
0.01	4.077×10^{-8}	4.251×10^{-6}	0.0408

The constant-degree calibration check fixes $p_0 = .3$ and $b = 5$, uses the same residual construction, and solves twelve population SGNS problems: four contrasts and windows one, two, and five. At each solution it fits a sigmoid of an affine function of F by population NLL. Table 15 compares window two with one, calibrating both. The remaining calibrated difference is positive in this example; the check concerns its fourth-order size, not a universal calibrated advantage.

Table 15: Completed population-grid SGNS calculations. Differences are window two minus one. The last column divides the calibrated difference by η^4. No graph or calibration-label sample is used.

η	Native NLL difference	Calibrated NLL difference	Difference$/\eta^4$
0.08	$9.6860 \cdot 10^{-5}$	$2.5014 \cdot 10^{-7}$	0.006107
0.04	$2.4632 \cdot 10^{-5}$	$1.5978 \cdot 10^{-8}$	0.006241
0.02	$6.2123 \cdot 10^{-6}$	$1.0098 \cdot 10^{-9}$	0.006311
0.01	$1.5600 \cdot 10^{-6}$	$6.3471 \cdot 10^{-11}$	0.006347

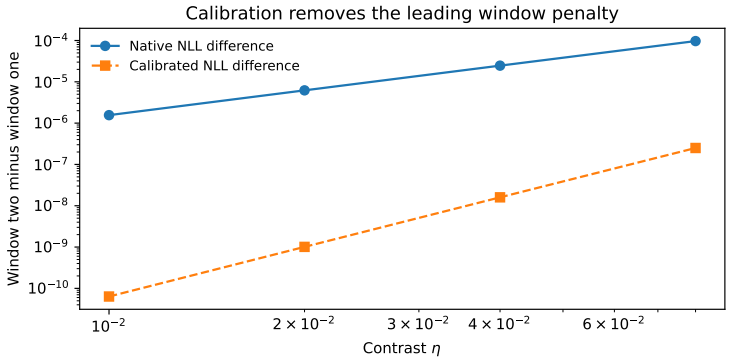


Figure 2: Additional loss at window two relative to one in the constant-degree population model. Both windows use the same decoder class. The post-calibration difference remains positive but is of fourth rather than second order.

G Related work in more detail

The closest prior analyses differ in what they characterize: the matrix induced by a walk, the embedding fitted to that matrix, or the probabilities obtained from a fitted score. These distinctions locate the results of this paper.

G.1 Walks as targets and filters

Walk co-occurrence matrices are polynomial transforms of the transition operator, and their SGNS targets are the corresponding shifted log-PMI matrices (Levy and Goldberg, 2014; Qiu et al., 2018); the objective itself can be viewed as weighted logistic principal component analysis (Landgraf and Bellay, 2017). Kunegis and Lommatzsch (2009) learn such spectral transformations directly for link prediction, and Chanpuriya and Musco (2020) and Lin et al. (2021) characterize the large-window and ergodic limits. Liben-Nowell and Kleinberg (2007) already compared walk-based similarity scores for link prediction. These results describe the target a walk induces. They do not say which target a fixed-width model should be fitted to: the matrix-factorization identity holds only at unconstrained rank (Harker and Bhaskara, 2024), and truncating the transformed matrix is a heuristic for the constrained logistic problem, not its solution. For node2vec, the second-order transition law needs an augmented-state description and is not a reweighting of powers of one node-transition matrix. Theorem 1 compares two targets under the width constraint on the two evaluation metrics; Theorem 2 shows that the truncation heuristic is exact for SGNS at first order and identifies its second-order failure.

G.2 Width-constrained fitting and random-graph theory

Fonarev et al. (2017) optimize fixed-rank SGNS on the Grassmannian, and Davison and Austern (2023) prove that rank-constrained population minimizers of subsampled embedding losses exist and that sampled fits converge to them, without identifying the minimizer. Theorem 2 supplies the minimizer for the degree-corrected weak-contrast family, and Theorem 3 uses their comparison scheme for the particular optimum it identifies. Barot et al. (2021) and Harker and Bhaskara (2024) analyze low-dimensional walk embeddings on block models for community recovery, the latter through gradient descent on a two-block model. Zhang and Tang (2024) show that on sparse block models longer windows improve convergence of the DeepWalk and node2vec embeddings to their population counterparts, an estimation effect; the present analysis fixes the contrast and asks how the population optimum itself changes with the walk, and Section 5.2 exhibits both effects at small and large graph sizes. The population spectral embedding with signed eigenvalues is the generalised random dot product graph of Rubin-Delanchy et al. (2022); see also Athreya et al. (2018) and Lei (2021). Negative-sampling choices also change the representation (Yang et al., 2020; Jaffe et al., 2020; Karkada et al., 2026); our population theorem fixes stationary negatives.

G.3 Ranking, probability loss, and calibration

Bipartite ranking has a risk-theoretic treatment (Menon and Williamson, 2016); probability estimation after ranking is studied directly by Menon et al. (2012). Proper-score decompositions separate information missing from the score from error in its probability scale (Bröcker, 2009). Logistic and temperature calibration (Lin et al., 2007; Guo et al., 2017), as well as calibration of knowledge-graph embeddings and graph neural-network link predictions (Safavi et al., 2020; Nascimento et al., 2024), are established methods. Our statements concern the walk-dependent error left by a specified correction. A common strictly increasing map preserves AUC, but a pair-dependent correction need not. Population fitting and calibration on finitely many labels are also distinct; the conditional-variance and finite-feature formulas describe the population comparison.