

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Quantifying and Optimizing Graph Neural Network Sparsification with an Approximation Framework

Ruixiao Zeng Ningyi Liao Siqiang Luo

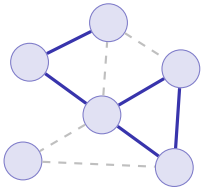
College of Computing and Data Science
Nanyang Technological University

CIKM 2026 Rome, Italy

Background: Pruning and Sampling

Repeated graph aggregation is expensive. Sparsification uses fewer edges.

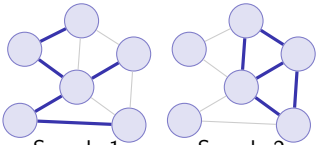
Deterministic pruning



Remove selected edges

A systematic operator shift

Stochastic sampling



Randomly retain edges

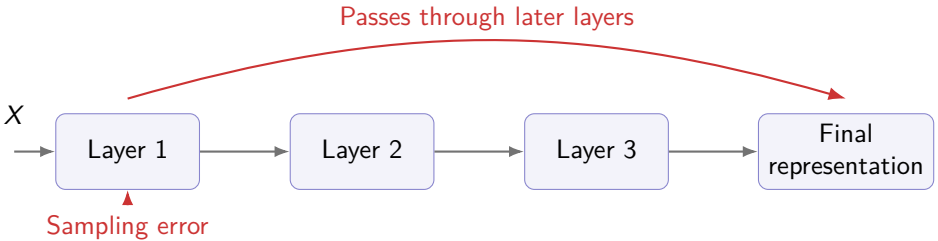
Sample-to-sample fluctuation

Both save graph operations. Both change the representations.

Paper Sections 1–2. Dashed / pale edges denote omitted edges; sketches are illustrative.

Motivation: Bounding Stochastic Sparsification Error

Sampling noise interacts with nonlinear layers and network depth.



Our contribution

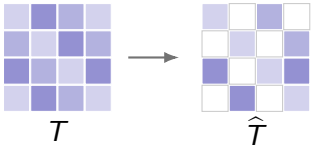
To our knowledge, the **first unified approximation framework** for deterministic pruning and **stochastic sampling**, with **layer-wise bounds on GNN representation error**.

Novelty scope follows the paper Abstract and Section 1: a common framework for pruning and stochastic sampling.

A GNN layer

$$H^{(\ell+1)} = \sigma\big(T\, H^{(\ell)} W\big)$$

T: propagation W: transformation



Prune entries or sample and reweight.

Approximate only the graph operator

$$\widehat{H}^{(\ell+1)} = \sigma\big(\widehat{T}\, \widehat{H}^{(\ell)} W\big)$$

Same input features and model weights.

$$\widehat{T} - T = \underbrace{\mathbb{E}[\widehat{T}] - T}_{\text{bias } \Delta_b} + \underbrace{\widehat{T} - \mathbb{E}[\widehat{T}]}_{\text{fluctuation } \Delta_s}$$

Deterministic pruning: $\Delta_s = 0$ Unbiased sampling: $\Delta_b = 0$

Paper Section 3.1. Bias scale $\varepsilon_b \geq \|\Delta_b\|_2$; fluctuation scale $\varepsilon_v^2 \geq \mathbb{E}\,\|\Delta_s\|_2^2$. Matrices are schematic.

Method: Measure Sensitivity, Then Allocate

1. Perturb one layer at a time



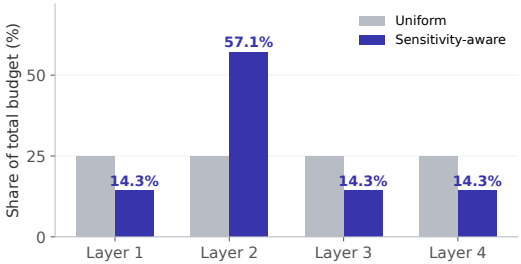
Sample only layer 2
others use the exact operator

2. Measure the final output change

Repeat at a common reference budget.
Larger change $\Rightarrow$ higher layer sensitivity s_ℓ.

3. Redistribute the same total budget

Give more samples to more sensitive layers.



The total stays fixed; its distribution changes.

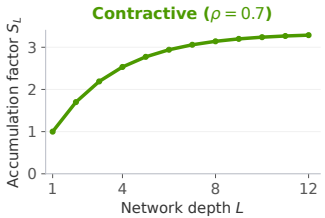
Measure each layer by its effect on the final representation.

Paper Section 4.3. Illustration uses $s = (1, 8, 1, 1)$ and $\gamma = 1/2$; it is not an experimental result.

Theory: How Error Accumulates with Depth

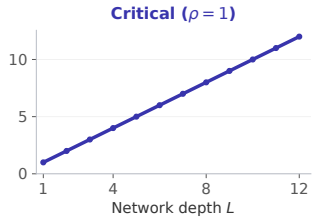
$$e_{\ell+1} \le \underbrace{\rho e_\ell}_{\text{old error}} + \underbrace{q}_{\text{new error}} \quad \Longrightarrow \quad e_L \le q S_L, \qquad S_L = \sum_{k=0}^{L-1} \rho^k$$

q: error from bias and sampling fluctuation | ρ: layer stability



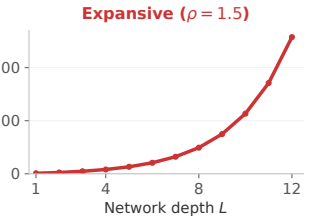
Bounded in depth

Earlier errors attenuate.



Linear in depth

Errors accumulate steadily.



Exponential in depth

Earlier errors can amplify.

Theorem 2: conditional expected error at fixed weights. Analytic factor S_L; different vertical scales. Assumptions and the local error term: backup A1.

Theory: An Optimal Split of a Fixed Budget

A layer's output contribution

$$\frac{s_\ell}{(g^{(\ell)})^\gamma}$$

s_ℓ: output sensitivity
$g^{(\ell)}$: layer budget

$$\min_{g^{(\ell)}>0} \sum_{\ell=1}^{L} \frac{s_\ell}{(g^{(\ell)})^\gamma} \quad \text{s.t.} \quad \sum_{\ell=1}^{L} g^{(\ell)} = g$$

Closed-form allocation

$$g^{(\ell)*} = g \, \frac{s_\ell^{1/(\gamma+1)}}{\sum_j s_j^{1/(\gamma+1)}}$$

Square-root sampling law

In a variance-dominated regime, $\gamma = 1/2$ gives

$$\boxed{g^{(\ell)*} \propto s_\ell^{2/3}}.$$

An $8\times$ more sensitive layer receives $4\times$ the budget.

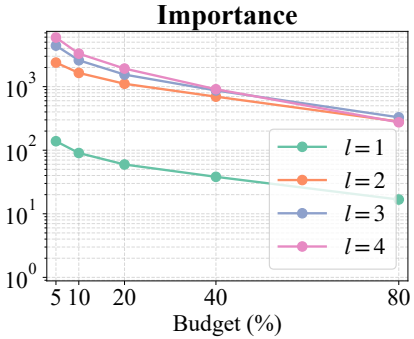
Uniform is optimal only when sensitivities are equal.

Balance the marginal error reduction across layers.

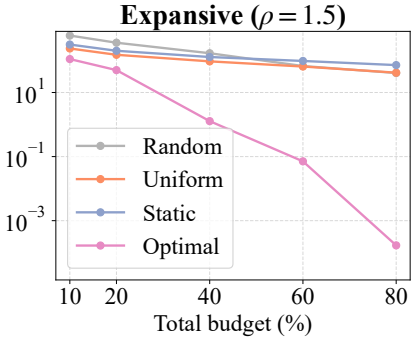
Proposition 4. Continuous first-order surrogate; fixed sensitivities, no caps. Derivation and scope: backup A3.

Evaluation: More Samples and Better Budget Allocation

More samples reduce e_ℓ



Better allocation reduces e_L



Static: geometric sensitivity

Optimal: calibrated sensitivity

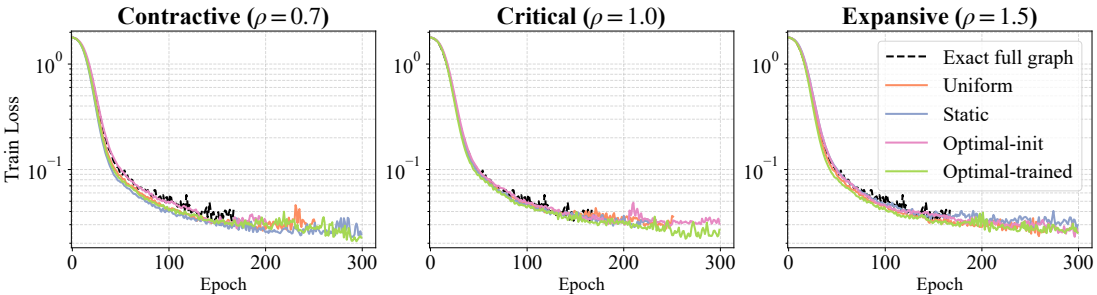
Up to 10× lower representation error in the reported evaluations.

GCN / PubMed / importance sampling; log-error axes. Selected panels: Figures 1–2 (full figures: A6–A7). The 10× claim summarizes the paper's evaluations.

Evaluation: How Does Allocation Behave in Training?

Calibrated allocations track the full-graph training loss closely.

CiteSeer | 80% total budget | Calibration at initialization / trained weights



The allocation remains useful as a practical training strategy in this setting.

Paper Figure 3. These are empirical training-loss results; the theorem itself concerns forward representation error at fixed weights.

Conclusion: A Unified Theory for Pruning and Sampling

Our main contribution

To our knowledge, the **first unified approximation framework for pruning and stochastic sampling**, with **layer-wise GNN representation-error bounds**.

Explain error propagation

Layer stability determines how approximation error accumulates with depth.

Optimize the budget split

Calibrated layer sensitivity yields better approximation with the same total budget.

Thank you!

github.com/Rechard16/QuOS DOI: 10.1145/3799682.3840751

Backup A1: Assumptions and the Conditional Theorem

- σ is L_σ-Lipschitz; the Taylor version also has M_σ-Lipschitz σ'.
- Fixed model weights; $\sup_{0\le\ell\le L} \left\|H^{(\ell)}\right\|_{\mathrm{F}} \le B$.
- $\Pr(\mathcal{E}_\delta) \ge 1-\alpha > 0$, with $\mathcal{E}_\delta = \{\left\|\widehat{T} - T\right\|_2 \le \delta\}$.

$$e_\ell = \mathbb{E}[\left\|\widehat{H}^{(\ell)} - H^{(\ell)}\right\|_{\mathrm{F}} \mid \mathcal{E}_\delta], \qquad a = \frac{L_\sigma \left\|W\right\|_2 B}{1-\alpha}, \quad b = \frac{M_\sigma \left\|W\right\|_2^2 B^2}{2(1-\alpha)}.$$

Theorem 2

$$e_\ell \le S_\ell\{a(\varepsilon_b + \varepsilon_v) + b(\varepsilon_b^2 + \varepsilon_v^2)\}, \quad S_\ell = \textstyle\sum_{k=0}^{\ell-1} \rho^k.$$

A Lipschitz-only bound uses $b = 0$, including for ReLU.

The result bounds a conditional expectation, not every individual sampled graph.

Paper Section 3. A sufficient failure bound is $\alpha \le \varepsilon_v^2/(\delta - \varepsilon_b)^2$ for $\delta > \varepsilon_b$.

Backup A2: Unbiased Sampling and Concentration

Independent edge indicators $\xi_{ij} \sim \mathrm{Bernoulli}(p_{ij})$ give

$$\widehat{A}_{ij} = \frac{\xi_{ij}}{p_{ij}} A_{ij}, \qquad \widehat{T} = D^{-1}\widehat{A}, \qquad \mathbb{E}\widehat{T} = T.$$

The normalization D comes from the original graph.

$$\Pr\left(\left\|\widehat{T} - T\right\|_2 \geq \delta\right) \leq 2n \exp\left[-\frac{\delta^2/2}{v + \mathcal{L}\delta/3}\right].$$

$$\delta_\beta = O\left(\sqrt{v \log(2n/\beta)} + \mathcal{L}\log(2n/\beta)\right).$$

- v: matrix variance proxy; $\mathcal{L}$: maximum summand norm.
- δ_β is a tail threshold; ε_v is a second-moment scale.
- The square-root budget law uses the variance-dominated regime.

Paper Section 4.1 and Appendix A.3. Recomputing sampled degrees can introduce bias.

Backup A3: Allocation Derivation and Scope

Fixed $s_\ell > 0$, $\gamma > 0$: the surrogate is strictly convex in positive budgets.

$$\mathcal{J} = \sum_\ell s_\ell (g^{(\ell)})^{-\gamma} + \lambda \left(\sum_\ell g^{(\ell)} - g \right).$$

$$\frac{\partial \mathcal{J}}{\partial g^{(\ell)}} = 0 \quad \Longrightarrow \quad g^{(\ell)} = \left(\frac{\gamma s_\ell}{\lambda} \right)^{1/(\gamma+1)}.$$

- Enforce the sum constraint to obtain the closed-form allocation.
- Equal budgets are optimal iff all s_ℓ are equal.
- Geometric proxy: $s_\ell = \rho^{L-\ell}$. Calibrated sensitivity also captures local injection and representation scale.
- Finite edge capacities require caps; integer counts require rounding.

Paper Proposition 4 and Appendix A.6. The guarantee is for the continuous first-order surrogate, not global optimality of the original GNN learning objective.

Backup A4: Practical Sensitivity Calibration

Only layer ℓ is approximated; all others use the exact operator. At reference budget g_{ref} and M trials:

$$\widehat{s}_\ell = g_{\mathrm{ref}}^{\gamma} \frac{1}{M} \sum_{m=1}^{M} \left\| \widehat{H}_{\ell,m}^{(L)} - H^{(L)} \right\|_{\mathrm{F}}.$$

Why the final output? A first-order perturbation has the form

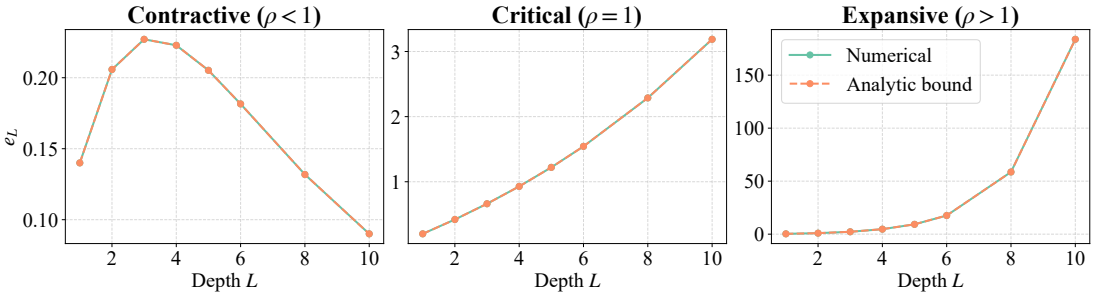
$$\widehat{H}^{(L)} - H^{(L)} \approx J_{\ell \to L}\, J_\sigma[\Delta T\, H^{(\ell-1)} W].$$

- Sensitivity includes downstream response, local features, and sampling noise.
- Calibration uses LM single-layer perturbation evaluations plus a reference.
- Feature-dependent operators need extra boundedness / Lipschitz assumptions.

Paper Section 4.3, Remark 3 and Appendix A.5. The paper does not quantify calibration runtime or an amortization break-even point.

Backup A5: Tightness in an Aligned Construction

Synthetic linear propagation with aligned rank-one perturbations.



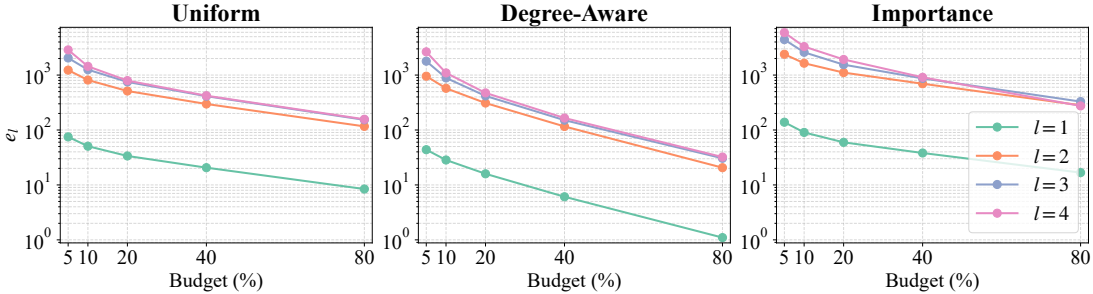
Numerical error meets the analytic bound in this construction.

General networks can have cross-layer cancellation and looser bounds.

Paper Figure 5 and Section 5.4. Tightness is shown for this construction, not every GNN.

Representation error decreases as the sampling budget grows.

GCN on PubMed | Three samplers | Layers 1–4



Within each panel

More retained edges → smaller error.

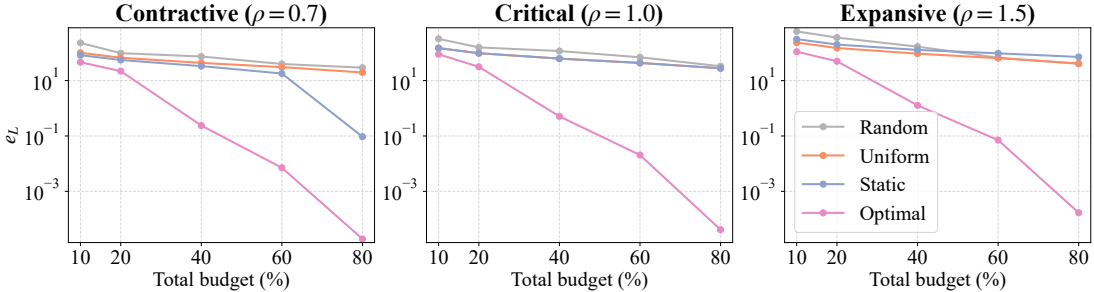
Across layers

Later layers → larger error here.

Paper Figure 1. Logarithmic vertical axes. Other evaluated graphs: Cora, CiteSeer, BA, ER, CSBM, GenCAT.

Calibrated allocation gives smaller final error at the same total budget.

PubMed | Same within-layer sampler | Three stability regimes



Static: geometric sensitivity **Optimal**: calibrated sensitivity

Up to 10× lower representation error in the reported evaluations.

Paper Figure 2. The 10× statement is the paper's aggregate result, not a pointwise ratio from this figure.