

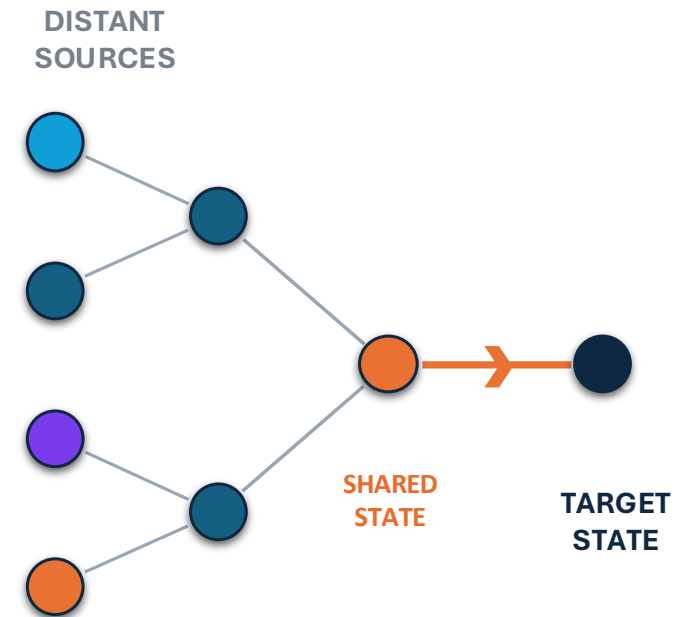
Oversquashing in graph neural networks

Lecture 12

What information fails to arrive?

How can we measure the loss?

Which changes address the cause?



Definition, measurements, and mitigation

DEFINITION

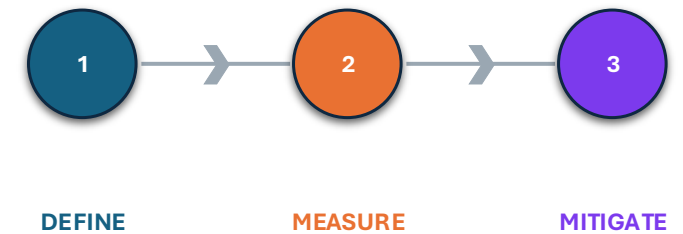
Reach, limited representations, and structural bottlenecks.

MEASUREMENTS

Actual model influence and graph-based warning signs.

MITIGATION

Changes to routes, propagation, or representation capacity.



Long-range tasks on graphs

The prediction depends on information from nodes several hops away, or on interactions between distant regions of the graph.

SYSTEMS AND NETWORKS

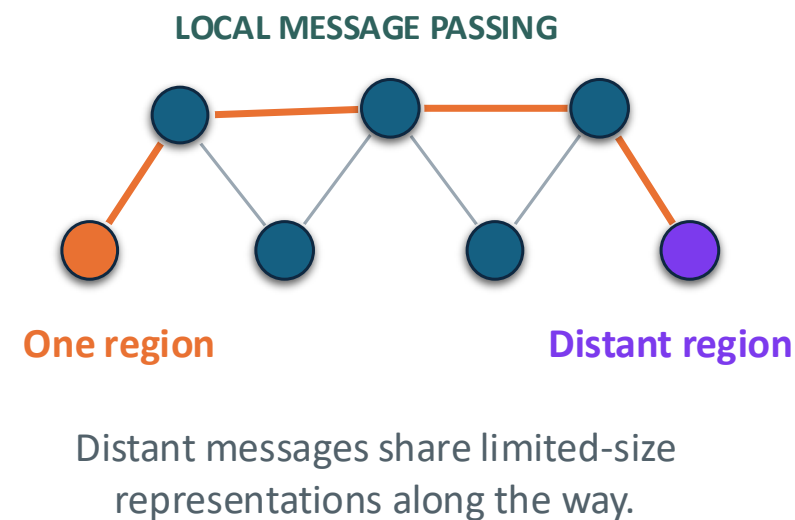
Local traffic may depend on congestion on roads farther upstream.

IMAGING

A cell's interpretation may depend on tissue context beyond its immediate neighbors.

Over-squashing is especially consequential for long-range tasks.

Losing distant information can remove evidence that the prediction needs.



Oversmoothing versus oversquashing

	Oversmoothing	Oversquashing
What happens	Nodes become indistinguishable	Distant signals fail to arrive
Where it is worst	Dense, well-connected regions	Bottlenecks and tree-like cuts
Typical fixes	Residual connections, normalisation	Rewiring, fully-adjacent layer, transformers

Both grow with depth, but their causes are opposite: fixes for one worsen the other.

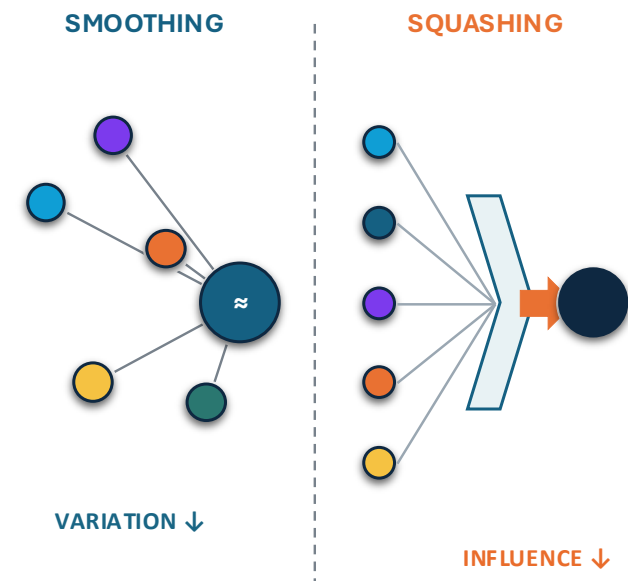
Oversmoothing and oversquashing

OVERSMOOTHING

Node representations lose useful variation under repeated mixing.

OVERSQUASHING

A target loses relevant information from particular sources.



Distinct communities can remain visible while a distant source has little influence.

Oversmoothing and oversquashing

Oono et al. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. ICLR 2020.

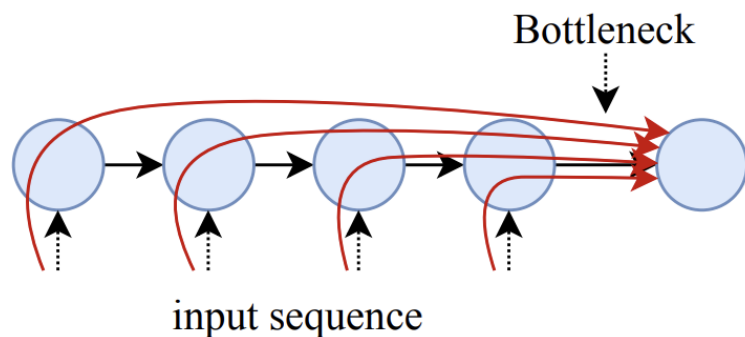
Di Giovanni et al. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. ICML 2023.

What oversquashing means

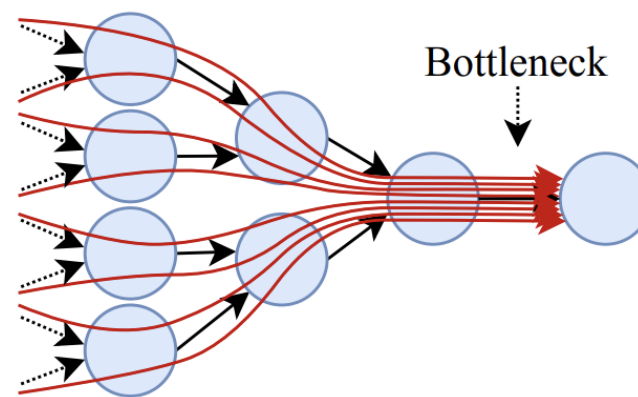
Task-relevant information weakens or disappears as many messages share limited intermediate states.

$$h_v^{(K)} \in \mathbb{R}^d$$

Underreaching: the source cannot arrive yet.
Oversquashing: useful detail fails to survive.



(a) The bottleneck of RNN seq2seq models



(b) The bottleneck of graph neural networks

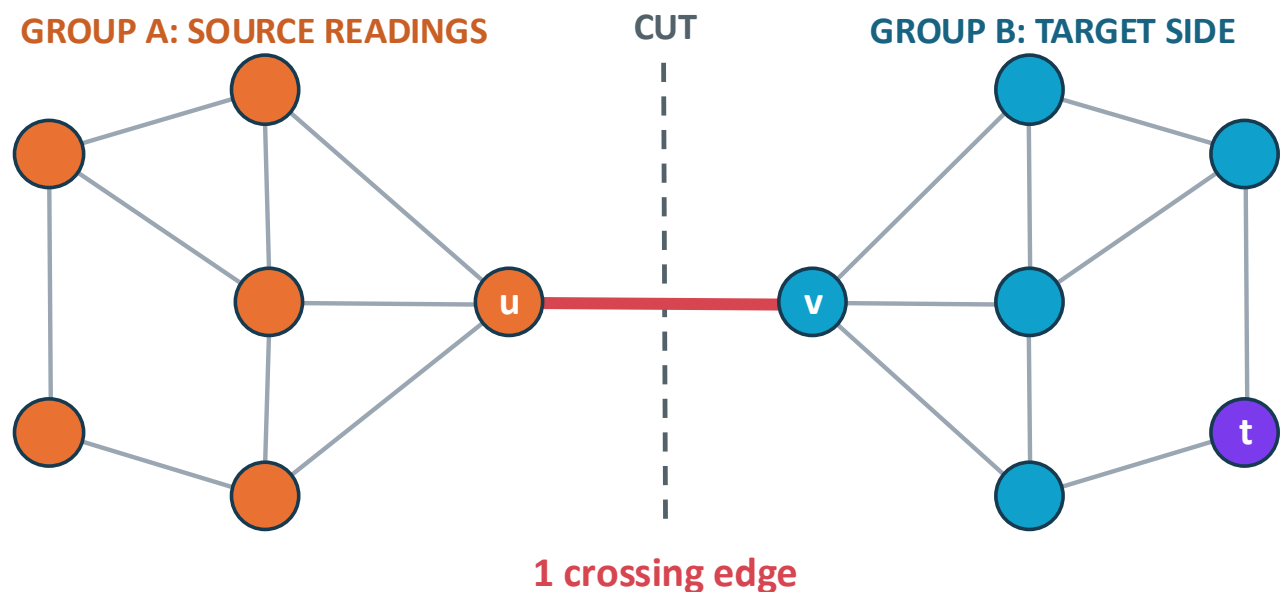
Distant sources and narrow cuts often increase the communication challenge.

Source

Alon et al. On the Bottleneck of Graph Neural Networks and its Practical Implications. ICLR 2021.

Small cuts and task demand

Partition the nodes into groups A and B. The cut consists of the edges crossing between them.



WHAT MUST TARGET RECOVER?

The task sets how much must cross the cut.
Regional mean: one number crosses; one edge carries it.
Any individual sensor: all node values must cross; one state cannot hold them.

A small cut limits routes. The task determines how much detail must cross.

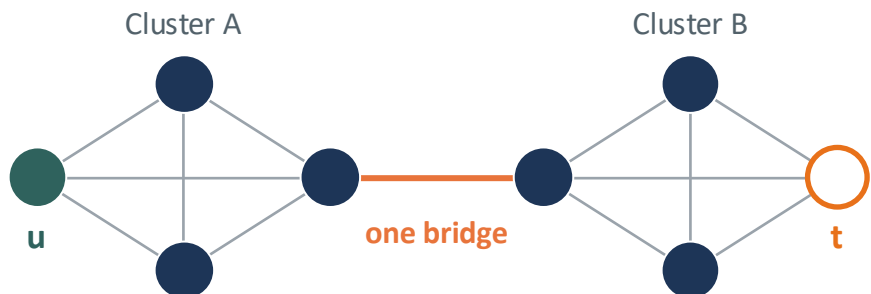
Bottlenecks and capacity

Alon et al. On the Bottleneck of Graph Neural Networks and its Practical Implications. ICLR 2021.

Di Giovanni et al. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. ICML 2023.

Oversquashing as a matrix calculation

THE GRAPH



Linear GNN, no activation:

$$H^{(K)} = \hat{A}^K X W, \quad h_t = \sum_u (\hat{A}^K)_{tu} \cdot x_u W$$

COLUMN u OF $\hat{A}^K$, HOW MUCH OF u REACHES EACH NODE

	u	A	A	A	B	B	B	t
K = 0	1.00	.00	.00	.00	.00	.00	.00	.00
K = 2	.24	.24	.24	.19	.04	.00	.00	.00
K = 4	.22	.22	.22	.18	.05	.02	.02	.02
K = 8	.19	.19	.19	.16	.07	.05	.05	.05

cluster A keeps .16–.24 each

reaches t: .00 → .02 → .05

Only $(\hat{A}^K)_{tu} = .05$ of u reaches t : every A node arrives this faintly, all summed into one h_t .

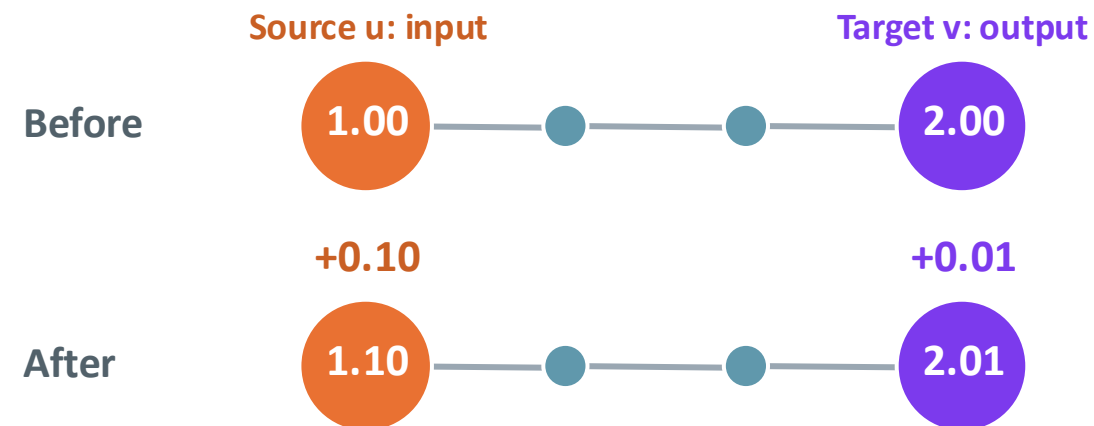
Pairwise influence through a Jacobian

How much can source u change the representation at target v ?

$$J_{vu}^{(K)} = \frac{\partial h_v^{(K)}}{\partial x_u} \quad I_{vu}^{(K)} = \left\| J_{vu}^{(K)} \right\|_2$$

Small norm: every local perturbation direction has a weak first-order effect.

TWO RUNS OF THE SAME GNN



$$\frac{\Delta h_v}{\Delta x_u} = \frac{0.01}{0.10} = 0.1$$

Control feature units, activation saturation, and whether the source is relevant.

Jacobian approach (Section 2)

Topping et al. Understanding Over-Squashing and Bottlenecks on Graphs via Curvature. ICLR 2022.

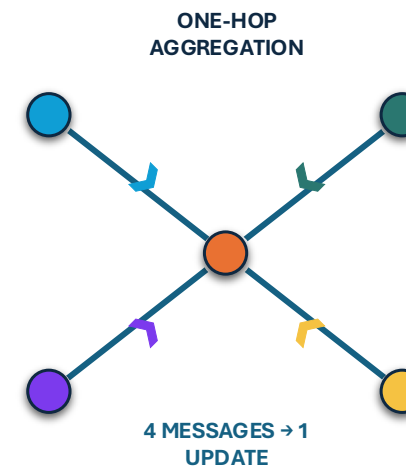
How topology enters a Jacobian bound

$$H^{(\ell+1)} = \sigma(SH^{(\ell)}W_{\ell}), \quad H^{(0)} = X$$

$$\|J_{vu}^{(K)}\|_2 \leq \left(\prod_{\ell=0}^{K-1} \|W_{\ell}\|_2 \right) (S^K)_{vu}$$

MODEL FACTOR × WEIGHTED WALK SUM

For a fixed model factor, weaker weighted routes give a tighter influence bound.



A small upper bound certifies weak sensitivity. A large bound is inconclusive.

Topping et al. Understanding Over-Squashing and Bottlenecks on Graphs via Curvature. ICLR 2022.

Di Giovanni et al. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. ICML 2023.

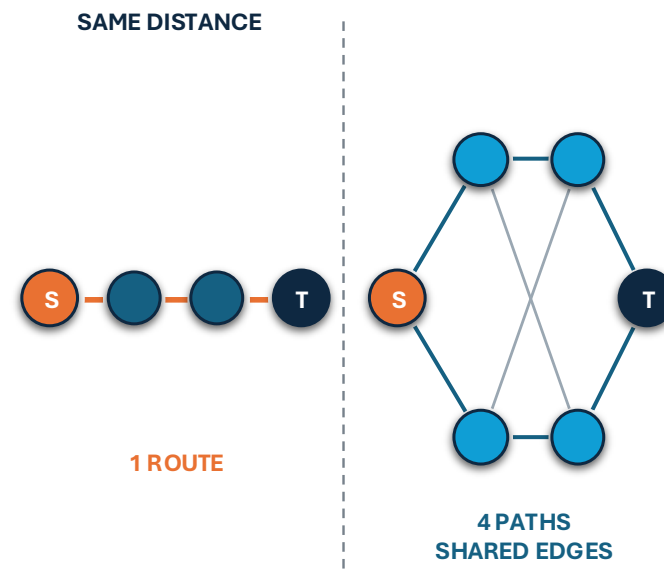
Effective resistance and parallel routes

Both distances are three hops. The right graph has four shortest paths with shared edges.

UNIT RESISTANCE ON EVERY EDGE

$$R_{left} = 1 + 1 + 1 = 3$$

$$R_{right} = \frac{1}{2} + \frac{1}{4} + \frac{1}{2} = \frac{5}{4}$$



Alternative paths reduce resistance without changing shortest-path distance.

Oversquashing connection (classroom circuit example)

Black et al. Understanding Oversquashing in GNNs through the Lens of Effective Resistance. ICML 2023.

Resistance and commute time

Connected, undirected, unweighted graph without self-loops.

$$L = D - A$$

$$R_{eff}(u, v) = (e_u - e_v)^T L^\dagger (e_u - e_v)$$

$$C_{uv} = t_{u \rightarrow v} + t_{v \rightarrow u} = 2|E| R_{eff}(u, v)$$

K GNN layers = a K-step walk.

If $K \ll t_{u \rightarrow v}$, the walk from u has not reached v yet, so little of u is in h_v .

High effective resistance between two nodes means little of one reaches the other at any practical depth.

Classical identities and oversquashing analysis

Black et al. Understanding Oversquashing in GNNs through the Lens of Effective Resistance. ICML 2023.

Di Giovanni et al. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. ICML 2023.

The spectral gap as a global proxy

SYMMETRIC NORMALIZED LAPLACIAN

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2}$$

$$0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$$

$$\frac{1}{n\lambda_2} \leq R_{max} \leq \frac{2}{\lambda_2}$$

Spectral gap is proportional to the size of the sparsest cut in the graph – Cheeger's inequality

Cheeger's bound uses the normalized Laplacian. A larger gap can also speed up mixing.

Cheeger inequality (Chung, Theorem 1) and its application

Chung. Four Cheeger-type Inequalities for Graph Partitioning Algorithms. ICCM 2007.

Jamadandi et al. Spectral Graph Pruning Against Over-Squashing and Over-Smoothing. NeurIPS 2024.

Curvature gives a local bottleneck score

Tree case: both endpoint degrees are greater than one

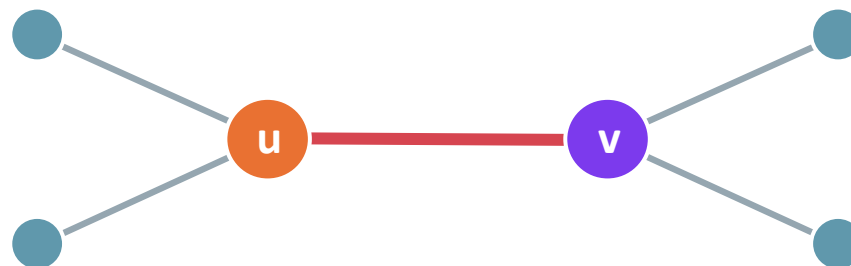
$$\kappa_{BF}(u, v) = \frac{2}{d_u} + \frac{2}{d_v} - 2$$

Path edge



$$d_u = d_v = 2, \quad \kappa_{BF} = 0$$

Branching edge



$$d_u = d_v = 3, \quad \kappa_{BF} = -\frac{2}{3}$$

More branching without local detours gives a more negative tree-edge score.

Balanced Forman curvature (Definition 1, tree case)

Topping et al. Understanding Over-Squashing and Bottlenecks on Graphs via Curvature. ICLR 2022.

A diagnostic workflow

REACH AND TASK RELEVANCE

Can the needed source reach the target in the available computation?

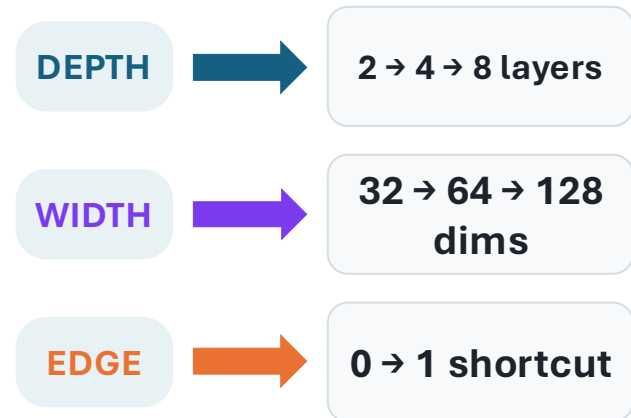
MODEL INFLUENCE

Measure Jacobians or controlled perturbations. Check units and saturation.

STRUCTURAL CONTEXT

Inspect cuts, resistance, and the normalized gap.

CONTROLLED SWEEP



**RECORD ACCURACY +
DISTANT-NODE SENSITIVITY**

Prediction error plus task-relevant influence gives a stronger diagnostic.

Three ways to change communication

ROUTES

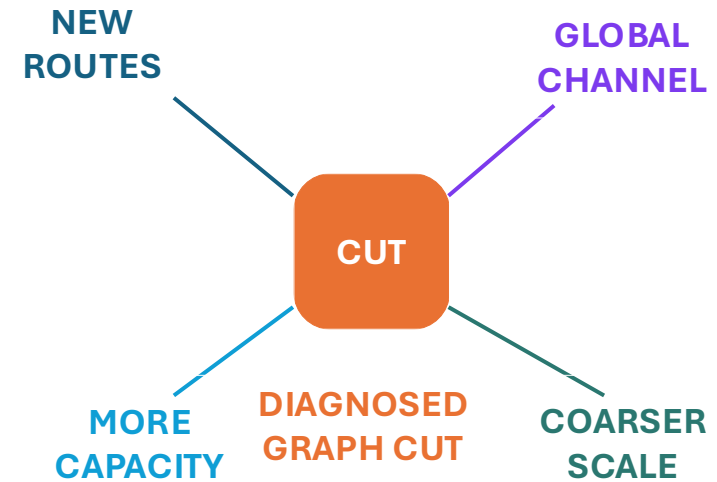
Rewire or introduce global communication.

PROPAGATION

Change the weights on path lengths and spectral modes.

REPRESENTATION

Keep enough detail in intermediate states.



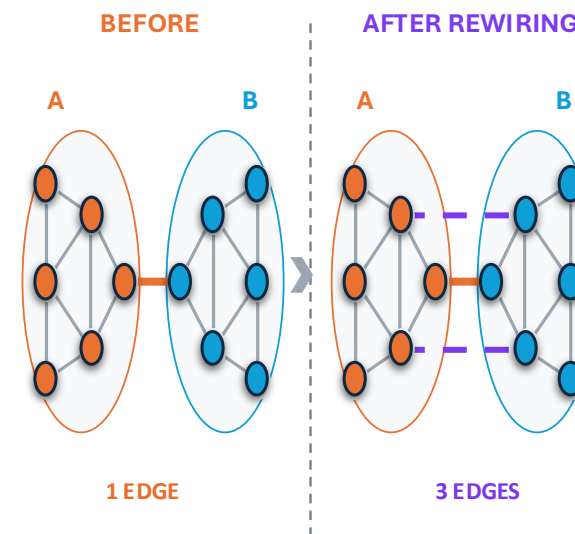
Rewiring changes the routes

TARGETED COMMUNICATION EDGES

Add useful cross-cut routes or connect relevant high-resistance pairs.

$$L_{new} = L + w b b^T, \quad b = e_u - e_v$$

Keep physical edges distinct from learned communication edges.



Resistance cannot increase after adding a positive edge. Learned sensitivity may decrease.

Rewiring and spectral-gap optimization

Black et al. Understanding Oversquashing in GNNs through the Lens of Effective Resistance. ICML 2023.

Jamadandi et al. Spectral Graph Pruning Against Over-Squashing and Over-Smoothing. NeurIPS 2024.

Global communication and its costs

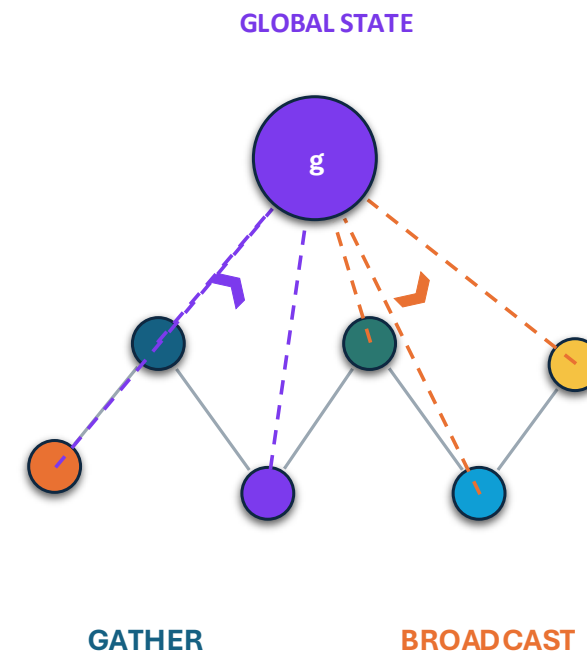
A virtual node gathers and broadcasts in two synchronous steps.

$$g^{(\ell+1)} = \text{AGG}_{i \in V} \varphi \left(h_i^{(\ell)} \right)$$

$$h_i^{(\ell+2)} = \psi \left(h_i^{(\ell+1)}, g^{(\ell+1)} \right)$$

Global attention offers target-specific summaries, at greater computation cost.

A shared global state can itself become a bottleneck.



A resolvent combines many path lengths

S is symmetric normalized adjacency. α is the continuation weight.

$$S = I - \mathcal{L}, \quad \lambda = 1 - \mu, \quad 0 < \alpha < 1$$

$$G_\alpha = (1 - \alpha)(I - \alpha S)^{-1} = (1 - \alpha) \sum_{k=0}^{\infty} \alpha^k S^k$$

$$h_S(\mu) = \frac{1 - \alpha}{1 - \alpha\mu}, \quad g_{\mathcal{L}}(\lambda) = \frac{1 - \alpha}{1 - \alpha + \alpha\lambda}$$

A nonlocal operator still requires a linear solve or propagation iterations.

Combining many path lengths does not automatically remove a structural bottleneck.

Adapted resolvent (α is continuation probability here)

Klicpera et al. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. ICLR 2019.

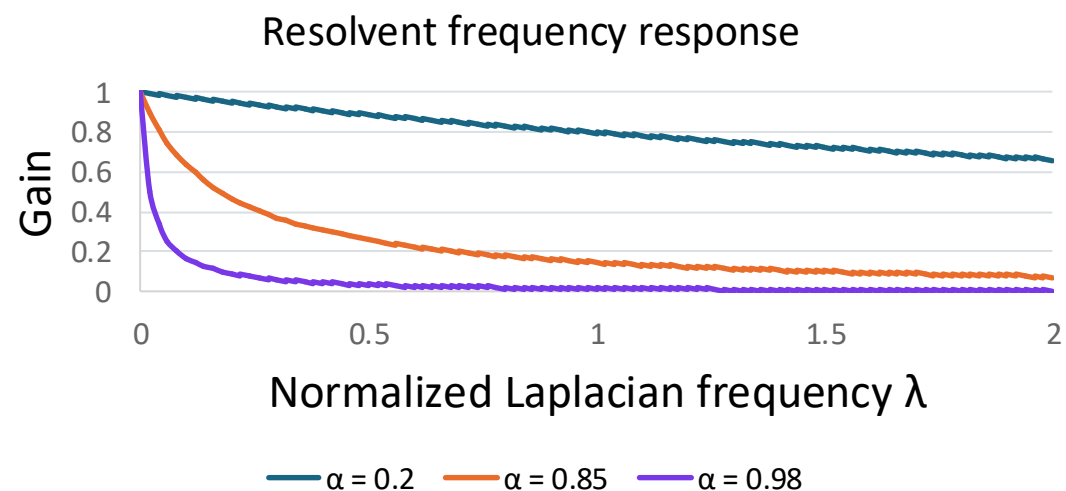
Filter response and useful information

THIS RESOLVENT IS LOW-PASS

$$g(\lambda) = \frac{1 - \alpha}{1 - \alpha + \alpha\lambda}$$

$$g(1) = 1 - \alpha$$

ARMA and Cayley offer flexible responses.
Useful passbands depend on the task.



Greater diffusion range does not imply stronger preservation of intermediate frequencies.

PPNP, ARMA, CayleyNets, and Stable-ChebNet

Klicpera et al. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. ICLR 2019.

Bianchi et al. Graph Neural Networks with Convolutional ARMA Filters. IEEE TPAMI 2022.

Levie et al. CayleyNets: Graph Convolutional Neural Networks with Complex Rational Spectral Filters. IEEE TSP 2019.

Hariri et al. Return of ChebNet: Understanding and Improving an Overlooked GNN on Long Range Tasks. NeurIPS 2025.

Capacity, attention, and hierarchy

WIDTH AND SEPARATE STREAMS

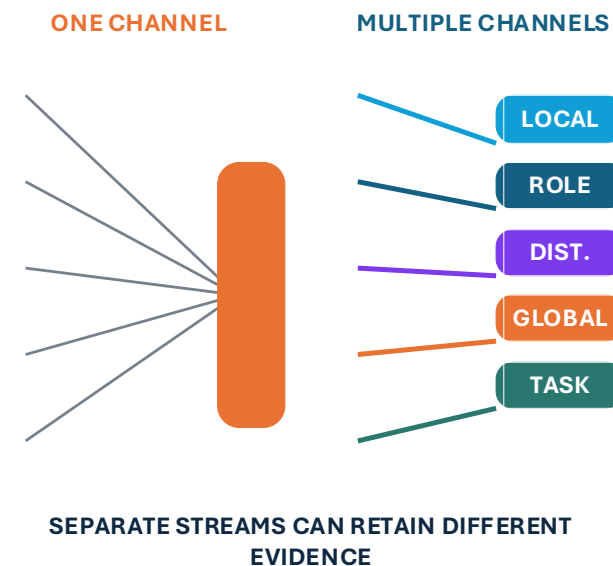
Retain several kinds of evidence before combining them.

LOCAL ATTENTION AND POSITIONS

Weight or identify sources without necessarily adding routes.

HIERARCHY

Communicate across regions while keeping access to fine detail.



Width analysis

Di Giovanni et al. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. ICML 2023.

A controlled test of the remedy

Intervention	Measure	Matched control	Main cost or risk
More depth	Reach and influence	Same width	Smoothing, optimization
Greater width	Retained detail	Parameter budget	Memory and computation
Rewiring	Relevant-pair influence	Same edge budget	Changed graph meaning
Virtual node	Global dependencies	Same tuning budget	Shared summary bottleneck
Global attention	Distant dependencies	Compute budget	Many pair scores
Spectral filter	Response and influence	Propagation cost	Attenuation, solver cost
Hierarchy	Regional dependencies	Comparable compute	Loss of fine detail

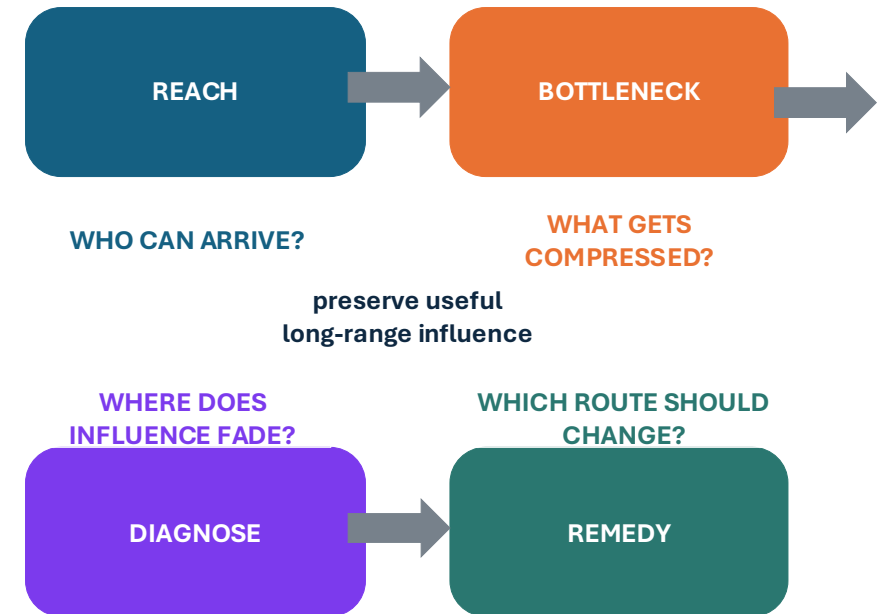
Fixed splits, multiple seeds, comparable tuning, and a compatible smoothing diagnostic.

Definition, evidence, and intervention

Task-relevant dependencies must survive

Model measurements and graph proxies differ

Test the effect of the chosen remedy



Which distant information does your research problem need?