Università Bocconi
MILANO

Topology and Machine Learning

Graph Convolutional Networks, Random Walks, and Curvature: Why Oversmoothing and Oversquashing Happen

Giorgio Micaletto

Bocconi University

September, 2025

Roadmap

Setup

Graph & Features. $G = (V, E)$, $|V| = n$, adjacency with self-loops $\widetilde{A} = A + I_n$, random-walk normalization $\widetilde{A}_{\mathrm{rw}} = \widetilde{D}^{-1}\widetilde{A}$.

Definition (Affine GCN Layer Kipf and Welling, 2017)

$$H^{(l+1)} \;=\; \sigma\big(\widetilde{A}_{\mathrm{rw}}\, H^{(l)} W^{(l)}\big), \qquad H^{(0)} \in \mathbb{R}^{n\times d_0}.$$

Why $\widetilde{A}_{\mathrm{rw}}$ matters. Iteration depth L induces multi-hop diffusion; ignoring nonlinearity and feature mixing,

$$\widetilde{A}_{\mathrm{rw}}^{L} H^{(0)} \xrightarrow[L\to\infty]{} \Pi H^{(0)}, \quad \Pi = \mathbf{1}\,\pi^{\top} \quad (\pi \text{ stationary}).$$

Two pathologies. *Oversmoothing* (fast mixing to $\Pi H^{(0)}$) and *oversquashing* (many sources, few edges, information has a bottleneck, hence $\mathrm{Inf}^{(L)}(u \to v)$ small).

Positive Curvature and Oversmoothing

Idea. Uniformly *positive* edge curvature ($\mathfrak{c}(e) \geq \varepsilon > 0$) yields a *spectral gap* for $\widetilde{A}_{\mathrm{rw}}$, so diffusion contracts everything orthogonal to π.

Claim (Informal Contraction Statement)

Let $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$ *for* $\widetilde{A}_{\mathrm{rw}}$. *If* $\mathfrak{c}(e) \geq \varepsilon \ \forall e$, *then* $\lambda_2 \leq 1 - \varepsilon$ *and*

$$\left\|\widetilde{A}_{\mathrm{rw}}^{L} - \Pi\right\|_2 = \lambda_2^L \;\Rightarrow\; \left\|\widetilde{A}_{\mathrm{rw}}^{L} H^{(0)} - \Pi H^{(0)}\right\|_\infty \leq (1-\varepsilon)^L \left\|H^{(0)} - \Pi H^{(0)}\right\|_\infty.$$

Takeaway. Positively curved regions $\Rightarrow$ quicker averaging $\Rightarrow$ earlier oversmoothing for deep stacks unless counteracted (e.g., residual/identity connections, normalization changes, decoupled propagation).

Negative Curvature and Oversquashing

Core phenomenon. The number of nodes at distance d from v can grow *exponentially* in d, but messages into v often cross only $\mathcal{O}(1)$ edges (a narrow cut).

Claim (Informal Oversquashing Statement)

If an edge (x, y) has negative curvature $\mathfrak{c}(x, y) < 0$ then for some function g depending on $\mathfrak{c}$ and $\Delta_{\max}$, given the funneling set on the y side $\mathfrak{Q}_y := \mathcal{N}(y) \setminus \big(\triangle(x, y) \cup \{x\} \cup \square_y(x, y) \big)$, for a 2-layer GCN,

$$\frac{1}{|\mathfrak{Q}_y|} \sum_{u \in \mathfrak{Q}_y} \mathrm{Inf}^{(2)}(u \to x) \;\leq\; g_{\mathfrak{c}}(\mathfrak{c}(x, y), \Delta_{\max})\, \alpha^{(2)}. \qquad \alpha^{(L)} := \prod_{l=0}^{L-1} \mathcal{L}_\sigma \, \|W^{(l)}\|_2.$$

Difference with oversmoothing. Oversmoothing is a *spectral mixing* effect; oversquashing is a *topological capacity* effect and appears even without strong mixing.

Curvature-Guided Editing: What It Can and Can't Do

Heuristic goal. Use curvature to *reduce bottlenecks* and *avoid too-fast mixing* via *rewiring* ($\mathcal{R}$), which adds (removes) edges where curvature is very negative (positive).

Claim (Informal Impossibility Statement)

For any non-absorbing $\mathcal{R}$, there is no strictly decreasing curvature-only functional $\mathcal{E} : \{graphs\} \to \mathbb{R}$ in expectation whenever edits occur:

$$\nexists \mathcal{E} : \quad \mathbb{E}\big[\mathcal{E}(\mathcal{R}(G)) \mid G\big] < \mathcal{E}(G) \text{ whenever } \mathcal{R}(G) \neq G.$$

Practical message. Curvature is an excellent *diagnostic* (where to look) and a strong *prior* for edits, but not a universal scalar objective guaranteeing monotone improvement.

Scoping and Planning

- **Follow your curiosity.** Read widely in areas you *enjoy*; sustained reading surfaces tensions and gaps to turn into a research question.
- **One-sentence → one-paragraph → one-page.** Maintain these three living summaries; they keep scope tight and alignment with your advisor easy.
- **Define your contribution type early:** new theorem, sharper bound, negative result, method, or empirical finding. What will a skeptical reader remember?
- **Use supervisors early and often.** Bring half-formed ideas; feedback on scope prevents weeks of rework.
- **Start earlier than you think.** Build slack; set weekly milestones; *write as you go* so the endgame is editing, not drafting.
- **Deliverables.** Decide what must exist at the end (artifact+result+figure set); plan backwards.

Reading, Experimenting And Final Writing

- **Targeted reading.** For each paper: problem, core idea in 2 lines, why prior art fails, what the bound/intuition actually depends on.
- **Replication first.** Stand up a minimal pipeline reproducing one credible baseline before any novelty.
- **Ablation plan.** Design ablations to *falsify* your favorite story.
- **Reproducibility.** Fix random seeds; report variance; release a small runnable artifact where possible.
- **Methods and Results $\rightarrow$ Intro.** Write the intro and abstract *last*.
- **Narrative spine.** Problem $\Rightarrow$ obstacle (over-smooth/squash) $\Rightarrow$ idea (curvature lens) $\Rightarrow$ what works/doesn't $\Rightarrow$ takeaways.
- **Notation hygiene.** One symbol per concept; add a small glossary; state assumptions.

Thank you!

Questions, comments, or thoughts?

Scan me to read my thesis

References I

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907.