Revisiting Over-smoothing and Over-squashing Using Ollivier-Ricci Curvature

Presenter: Khang Nguyen
FPT Software AI Center, Vietnam
VNUHCM - University of Science, Vietnam

Khang Nguyen, Hieu Nong, Vinh Nguyen, Nhat Ho, Stanley J. Osher, Tan Nguyen

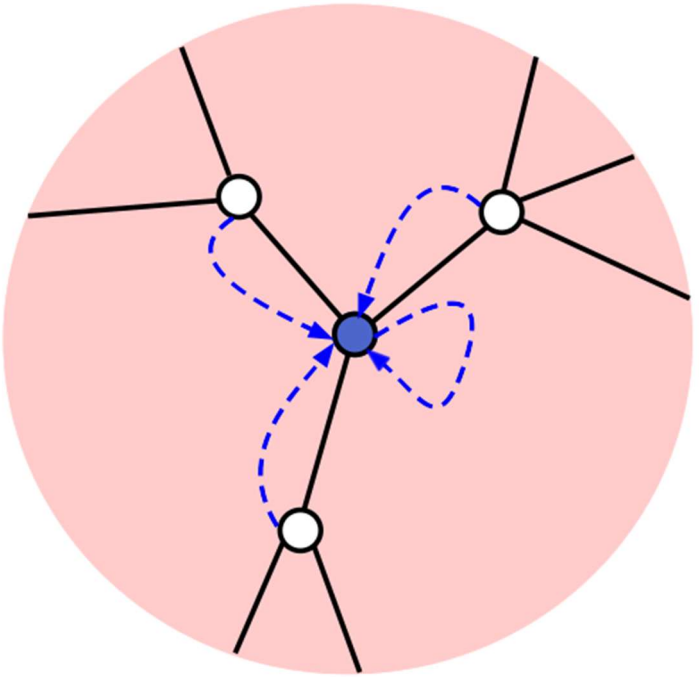
Graph Neural Networks

Popular graph neural networks (GNNs) are message passing neural networks.

Adjacent nodes send messages to each other, which are then aggregated and used to update node features.

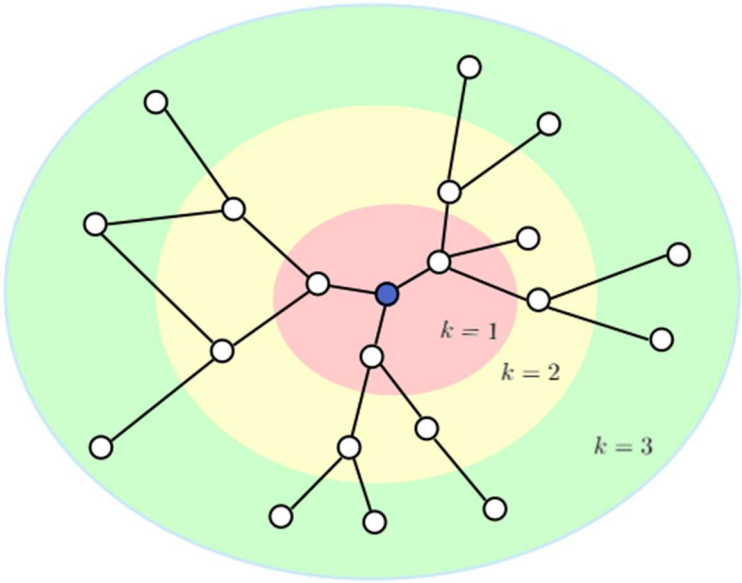
$$\boldsymbol{X}_u^{k+1} = \phi_k \left(\bigoplus_{p \in \tilde{\mathcal{N}}_u} \psi_k(\boldsymbol{X}_p^k) \right)$$

ψ_k is the message function
$\bigoplus$ is an aggregating function
ϕ_k is an update function



The Depth Limitation of GNNs

- To capture long range information, we need GNNs with sufficient depth.
- Current GNNs suffer from performance degradation at higher depth, which limits their application.
- This degradation is attributed to two problems:
  over-smoothing and over-squashing

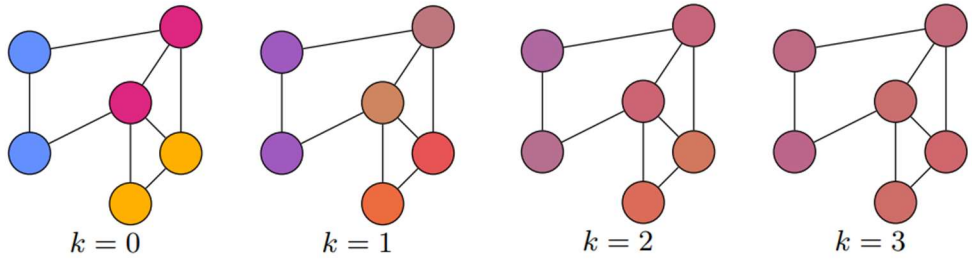


Over-smoothing and Over-squashing

Over-smoothing: nodes' features become similar to each other as the number of layers increases.

$$\sum_{(u,v)\in\mathcal{E}} \left|\boldsymbol{X}_u^k - \boldsymbol{X}_v^k\right| \to 0 \text{ as } k \to \infty.$$

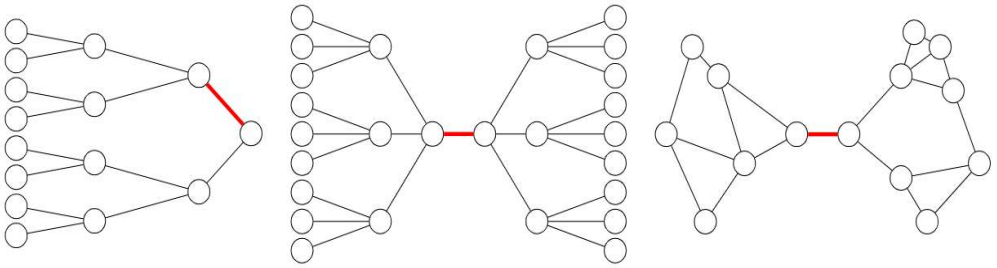
➡ Degrade classification accuracy



Over-smoothing causes nodes' features to become indistinguishable

Over-squashing: graph bottlenecks cause messages between distant nodes to become overly-squashed.

➡ Unable to capture long range interactions



Over-squashing impedes the GNN's ability to capture long range interactions

Ollivier-Ricci Curvature

In Riemannian geometry, curvature is a local measure of geodesic dispersion.

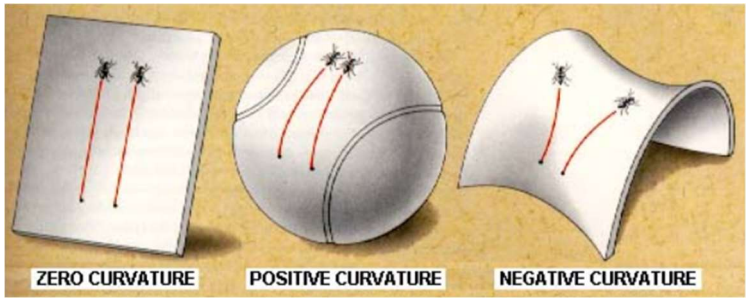


image source:
https://starchild.gsfc.nasa.gov/docs/StarChild/questions/question35.html

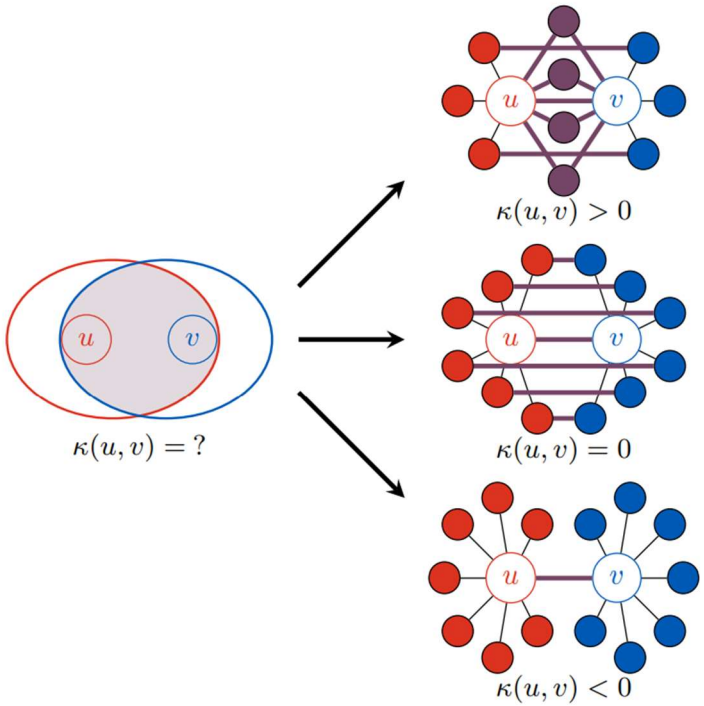
The Ollivier-Ricci curvature captures the essence of geodesic dispersion using optimal transport.

$$W_1(\mu_u, \mu_v) = \inf_{\pi \in \Pi(\mu_u, \mu_v)} \left(\sum_{(p,q) \in \mathcal{V}^2} \pi(p,q) d(p,q) \right)$$

$$\kappa(u,v) = 1 - \frac{W_1(\mu_u, \mu_v)}{d(u,v)}$$

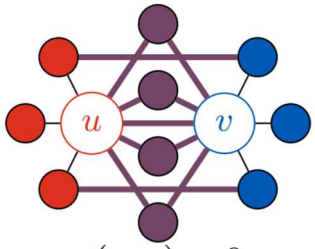
➡ Not dependent on the differential structure

The Ollivier-Ricci curvature on graph characterizes the well-connectedness of local graph neighborhoods.



Ollivier-Ricci curvature values on a local graph neighborhood

Positive Curvature and Over-smoothing



$\kappa(u, v) > 0$

well-connected neighborhoods aggregate the same information

Theorem 4.2. *Consider the update rule given by Equation (1). Suppose the edge curvature $\kappa(u, v) > 0$. For some k, assume the update function ϕ_k is L-Lipschitz, $\left|\boldsymbol{X}_p^k\right| \leq C$ for all $p \in \mathcal{N}(u) \cup \mathcal{N}(v)$, and the message function ψ_k is bounded, i.e. $|\psi_k(\boldsymbol{x})| \leq M|\boldsymbol{x}|, \forall \boldsymbol{x}$. There exists a positive function $h : (0, 1) \to \mathbb{R}^+$ dependent on the constants L, M, C, n satisfying*

- *if $\bigoplus$ is the sum operation then h is constant;*
- *if $\bigoplus$ is the mean operation then h is decreasing;*

such that

$$\left|\boldsymbol{X}_u^{k+1} - \boldsymbol{X}_v^{k+1}\right| \leq (1 - \kappa(u, v))h(\kappa(u, v)). \quad (6)$$

In both cases, we clearly have

$$\lim_{x \to 1}(1 - x)h(x) = 0. \quad (7)$$

A positively curved edge causes its nodes to have similar representation

Proposition 4.3. *Assume the graph is regular. Suppose there exists a constant $\delta > 0$ such that for all edges $(u, v) \in \mathcal{E}$, the curvature is bounded by $\kappa(u, v) \geq \delta > 0$. Consider the update rule given by equation 1. For all $k \geq 1$, assume the functions ϕ_k are L-Lipschitz, $\bigoplus$ is realised as the mean operation, $\left|\boldsymbol{X}_p^0\right| \leq C$ for all $p \in \mathcal{V}$, and the functions ψ_k are bounded linear operators, i.e. $|\psi_k(\boldsymbol{x})| \leq M|\boldsymbol{x}|, \forall \boldsymbol{x}$. The following inequality holds for $k \geq 1$ and any neighboring vertices $u \sim v$*

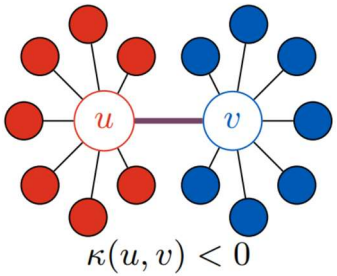
$$\left|\boldsymbol{X}_u^k - \boldsymbol{X}_v^k\right| \leq \frac{2}{3}C\left(\frac{3LM\lfloor(1-\delta)n\rfloor}{n+1}\right)^k. \quad (8)$$

Furthermore, for any $u, v \in \mathcal{V}$ that are not necessarily neighbors, the following inequality holds

$$\left|\boldsymbol{X}_u^k - \boldsymbol{X}_v^k\right| \leq \frac{2}{3}\left\lfloor\frac{2}{\delta}\right\rfloor C\left(\frac{3LM\lfloor(1-\delta)n\rfloor}{n+1}\right)^k. \quad (9)$$

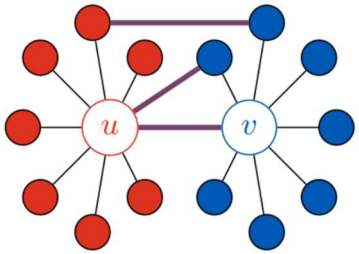
A sufficiently positively curved graph causes its node features to exponentially converge

Negative Curvature and Over-quashing



$\kappa(u, v) < 0$

sparsely-connected neighborhoods do not effectively propagate information



Proposition 4.4. *Let $\tilde{\mathcal{E}}$ be union of the edge set $\mathcal{E}$ with the set of all possible self-loops. Let S be the subset of $\tilde{\mathcal{E}}$ containing edges of the form (p, q) with $p \in \tilde{\mathcal{N}}_u \backslash \{v\}$ and $q \in \tilde{\mathcal{N}}_v \backslash \{u\}$. Supposing each vertex w is a vertex of at most $\frac{n}{m}$ edges in S. The following inequality holds*

$$|S| \leq \frac{n(\kappa(u, v) + 2)}{2}. \tag{10}$$

A negatively curved edge indicates a lack of sufficient information pathways, i.e., it induces a bottleneck

Theorem 4.5. *Consider the update rule given by Equation (1). Suppose ψ_k, ϕ_k are linear operators for all k, and $\bigoplus$ is the sum operation. If u, v are neighboring vertices with neighborhoods as in Proposition 4.4 and S is defined similarly then for all $p \in \tilde{\mathcal{N}}_u \backslash \{v\}$, $q \in \tilde{\mathcal{N}}_v \backslash \{u\}$, we have*

$$\begin{aligned} \left[\frac{\partial \boldsymbol{X}_u^{k+2}}{\partial \boldsymbol{X}_q^k}\right] &= \alpha \sum_{w \in V} \left[\frac{\partial \boldsymbol{X}_u^{k+2}}{\partial \boldsymbol{X}_w^k}\right], \\ \left[\frac{\partial \boldsymbol{X}_v^{k+2}}{\partial \boldsymbol{X}_p^k}\right] &= \beta \sum_{w \in V} \left[\frac{\partial \boldsymbol{X}_v^{k+2}}{\partial \boldsymbol{X}_w^k}\right], \end{aligned} \tag{11}$$

where $\left[\frac{\boldsymbol{y}}{\boldsymbol{x}}\right]$ is used to denote the Jacobian of $\boldsymbol{y}$ with regard to $\boldsymbol{x}$, and α, β satisfy

$$\begin{aligned} \alpha &\leq \frac{|S| + 2}{\sum_{w \in \tilde{\mathcal{N}}_v} (\deg(w) + 1)}, \\ \beta &\leq \frac{|S| + 2}{\sum_{w \in \tilde{\mathcal{N}}_u} (\deg(w) + 1)}. \end{aligned} \tag{12}$$

A negatively curved edge inhibits effective message passing

Batch Ollivier-Ricci Flow (BORF)

Algorithm 1 Batch Ollivier-Ricci Flow (BORF)

Input: graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, # rewiring batches N, # edges added per batch h, # edges removed per batch k
for $i = 1$ **to** N **do**
  Find h edges $(u_1, v_1), \ldots, (u_h, v_h)$ with minimal Ollivier-Ricci curvature κ, along with each summand $\pi_j(p, q)d(p, q)$ in their optimal transportation cost sum for all $p, q \in \mathcal{V}$ and $j = \overline{1, h}$
  Find k edges $(u^1, v^1), \ldots, (u^k, v^k)$ with maximal Ollivier-Ricci curvature κ
  for $j = 1$ **to** h **do**
    Add to $\mathcal{G}$ the edge (p^*, q^*) given by

$$(p^*, q^*) = \operatorname{argmax} d(p, q)\pi_j(p, q)$$

  end for
  Remove edges $(u^1, v^1), \ldots, (u^k, v^k)$ from $\mathcal{G}$
end for

1. Find the most positively curved and most negatively curved edges

2. Add edges to promote message passing around bottlenecks, thus alleviating over-squashing

3. Remove the most positively curved edges to suppress over-smoothing

Using optimal transport

- Advantage: can utilize the optimal transport plan to find the best edges to add
- Disadvantage: computational cost, but this can be reduced by using approximations

Experiments

Table 2. Classification accuracies of GCN and GIN with None, SDRF, FoSR, and BORF rewiring on various node classification datasets. Best results are highlighted in bold.

	GCN				GIN			
DATA SET	NONE	SDRF	FOSR	BORF	NONE	SDRF	FOSR	BORF
CORA	86.7 ± 0.3	86.3 ± 0.3	85.9 ± 0.3	$\mathbf{87.5 \pm 0.2}$	76.0 ± 0.6	74.9 ± 0.1	75.1 ± 0.8	$\mathbf{78.4 \pm 0.4}$
CITESEER	72.3 ± 0.3	72.6 ± 0.3	72.3 ± 0.3	$\mathbf{73.8 \pm 0.2}$	59.3 ± 0.9	60.3 ± 0.8	61.7 ± 0.7	$\mathbf{63.1 \pm 0.8}$
TEXAS	44.2 ± 1.5	43.9 ± 1.6	46.0 ± 1.6	$\mathbf{49.4 \pm 1.2}$	53.5 ± 3.1	50.3 ± 3.7	47.0 ± 3.7	$\mathbf{63.1 \pm 1.7}$
CORNELL	41.5 ± 1.8	42.2 ± 1.5	40.2 ± 1.6	$\mathbf{50.8 \pm 1.1}$	36.5 ± 2.2	40.0 ± 2.1	35.6 ± 2.4	$\mathbf{48.6 \pm 1.2}$
WISCONSIN	44.6 ± 1.4	46.2 ± 1.2	48.3 ± 1.3	$\mathbf{50.3 \pm 0.9}$	48.5 ± 2.2	48.8 ± 1.9	48.5 ± 2.1	$\mathbf{54.9 \pm 1.2}$
CHAMELEON	59.2 ± 0.6	59.4 ± 0.5	59.3 ± 0.6	$\mathbf{61.5 \pm 0.4}$	58.1 ± 2.1	58.4 ± 2.1	56.3 ± 2.2	$\mathbf{65.3 \pm 0.8}$

Table 3. Classification accuracies of GCN and GIN with None, SDRF, FoSR, and BORF rewiring on various graph classification datasets. Best results are highlighted in bold.

	GCN				GIN			
DATA SET	NONE	SDRF	FOSR	BORF	NONE	SDRF	FOSR	BORF
ENZYMES	25.5 ± 1.3	26.1 ± 1.1	$\mathbf{27.4 \pm 1.1}$	24.7 ± 1.0	31.3 ± 1.2	33.5 ± 1.3	25.3 ± 1.2	$\mathbf{35.5 \pm 1.2}$
IMDB	49.3 ± 1.0	49.1 ± 0.9	49.6 ± 0.8	$\mathbf{50.1 \pm 0.9}$	69.0 ± 1.3	68.6 ± 1.2	69.5 ± 1.1	$\mathbf{71.3 \pm 1.5}$
MUTAG	68.8 ± 2.1	70.5 ± 2.1	75.6 ± 1.7	$\mathbf{75.8 \pm 1.9}$	75.5 ± 2.9	77.3 ± 2.3	75.2 ± 3.0	$\mathbf{80.8 \pm 2.5}$
PROTEINS	70.6 ± 1.0	71.4 ± 0.8	$\mathbf{72.3 \pm 0.9}$	71.0 ± 0.8	69.7 ± 1.0	72.2 ± 0.9	$\mathbf{74.2 \pm 0.8}$	71.3 ± 1.0

Experiments

Table 4. Classification accuracies of GCN at depths 5, 7, and 9 with different BORF rewiring options on Cornell and Mutag datasets.

DATA SET	# LAYERS	NONE	BEST SETTINGS	ONLY REMOVE	ONLY ADD	REMOVE & ADD EQUALLY
CORNELL	5	41.3 ± 1.4	45.5 ± 1.1	46.4 ± 1.2	44.7 ± 1.3	45.9 ± 1.2
	7	39.5 ± 1.7	41.5 ± 1.5	43.2 ± 1.3	42.8 ± 1.4	41.8 ± 1.3
	9	35.5 ± 1.4	40.9 ± 1.3	41.9 ± 1.6	40.3 ± 2.0	39.9 ± 1.6
MUTAG	5	67.7 ± 1.6	75.4 ± 2.1	68.5 ± 2.8	76.1 ± 2.2	71.8 ± 1.2
	7	64.1 ± 2.1	72.1 ± 1.3	65.1 ± 1.5	75.2 ± 2.4	66.2 ± 1.9
	9	63.1 ± 1.2	69.7 ± 1.5	60.7 ± 2.5	70.4 ± 1.7	61.3 ± 1.5

Conclusion

- Ollivier-Ricci graph curvature provides a unified framework to study both the Over-smoothing and Over-squashing problems.
- Rewiring is a natural way to better GNN performance.
- Future work: Explore ways we can utilize curvature to study and improve GNNs.