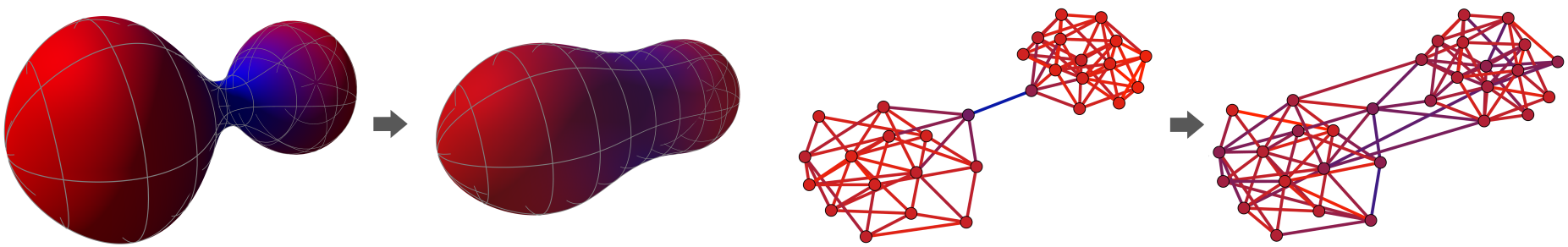
Understanding over-squashing and bottlenecks on graphs via curvature

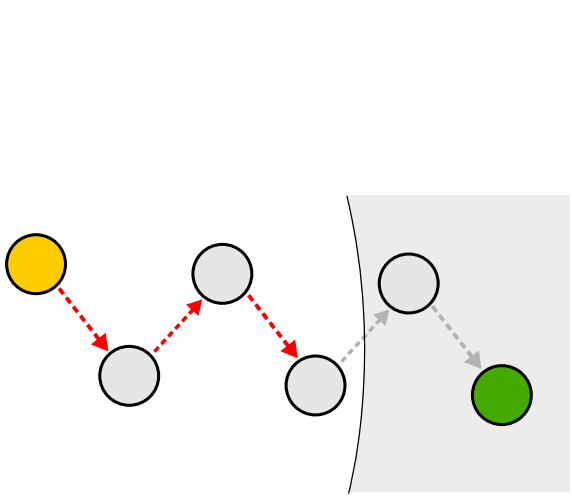
Jake Topping[13†] Francesco Di Giovanni[2†] Benjamin P. Chamberlain[2]
Xiaowen Dong[1] Michael M. Bronstein[12]

[1]University of Oxford [2]Twitter [3]Imperial College London
[†]Equal contribution

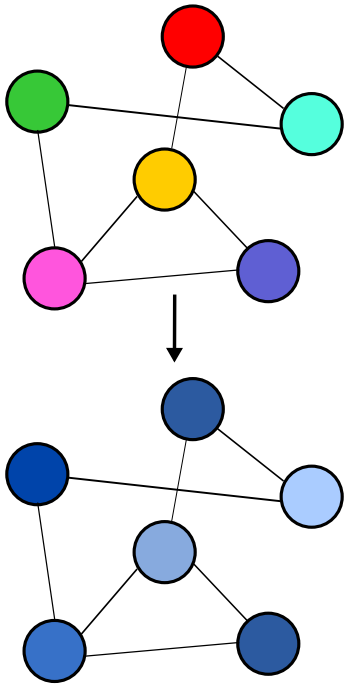


1. The problem: Over-squashing
2. Through the lens of curvature
3. A solution: Curvature-based rewiring

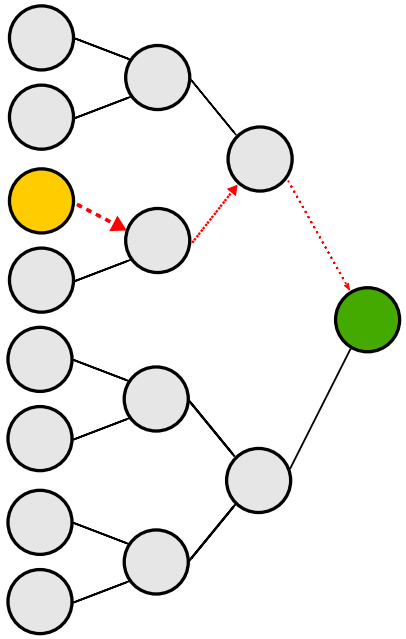
Potential problems with message-passing NNs



Under-reaching

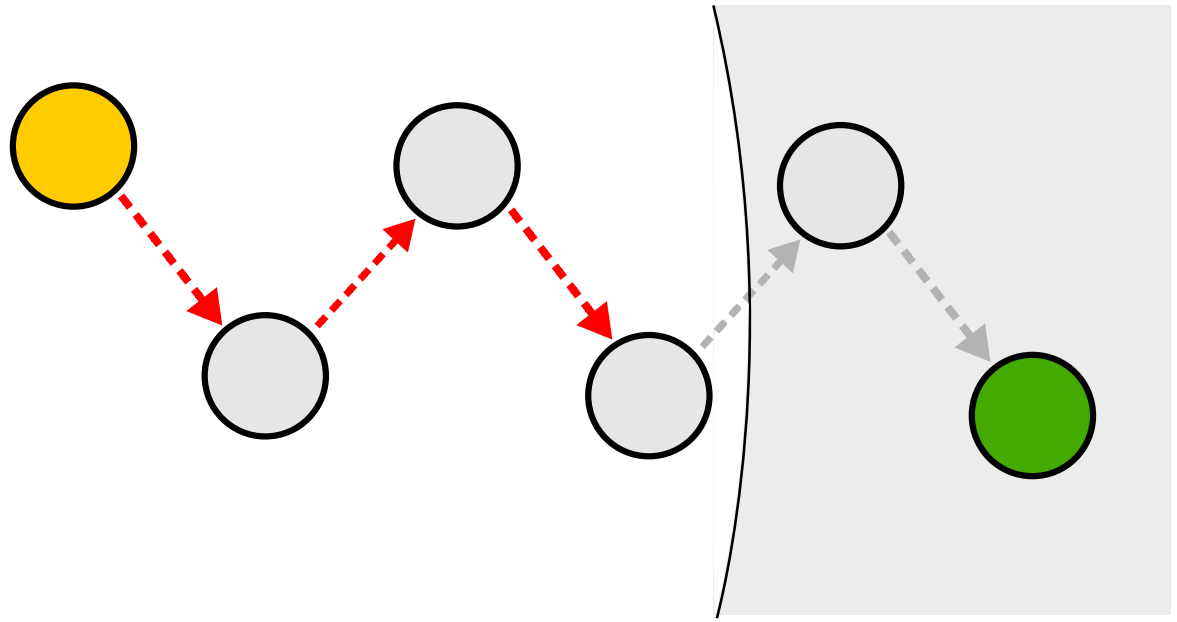


Over-smoothing



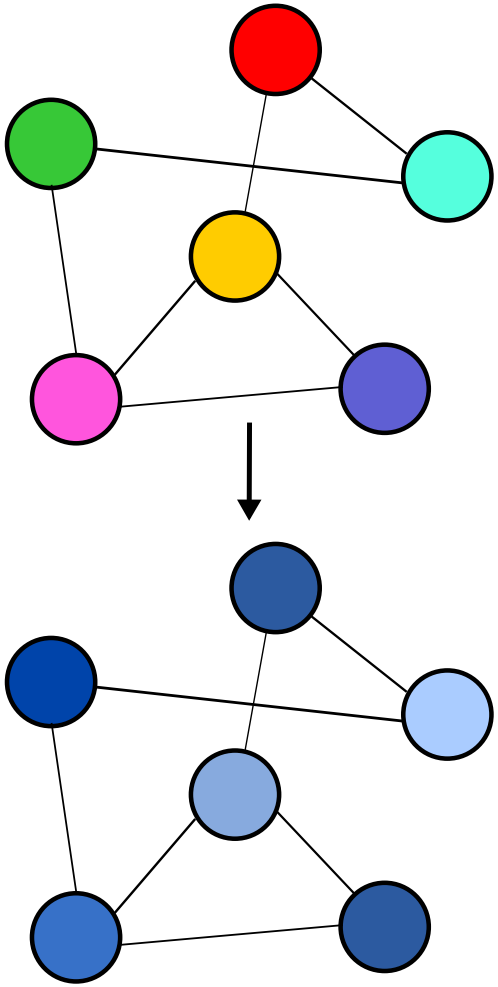
Over-squashing

Under-reaching



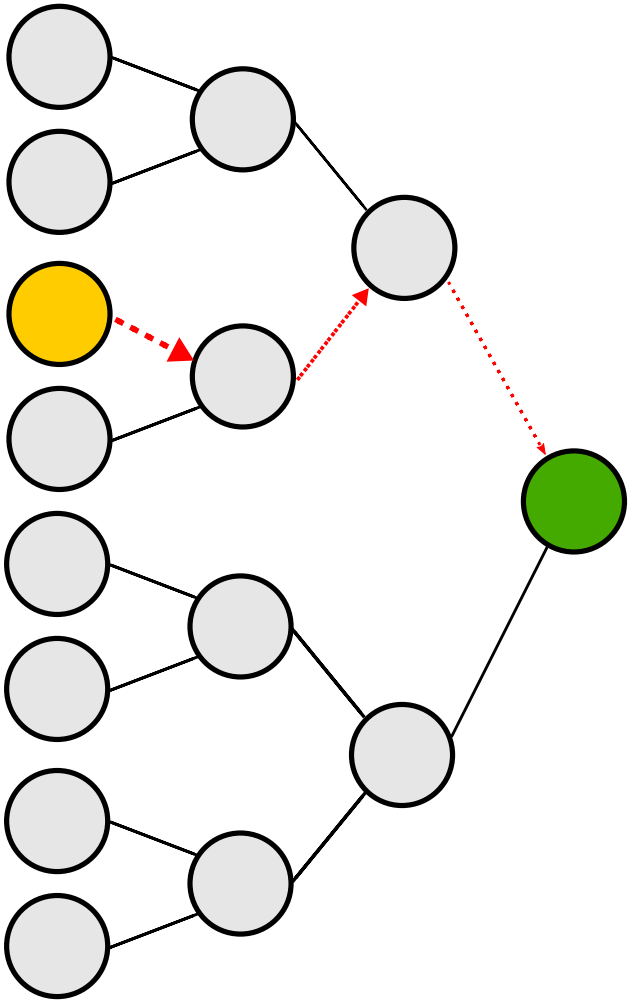
- Information cannot propagate further than there are layers in the MPNN (shown above with three layers)

Over-smoothing



- Over an MPNN with many layers, node representations can become similar (smoothed out) and weaken influence of graph structure

Over-squashing



- As exponentially-many[1] nodes' features are passed through a bottleneck, information transfer can be harmed by squashing/compression
- Present in RNN models, but here the problem can be exponentially worse!

[1] Alon and Yahav [2021]

Quantifying over-squashing by the Jacobian

The Jacobian[2] $\left|\frac{\partial h_i^{(\ell)}}{\partial x_s}\right|$ tells us how much our ℓth MPNN layer $h^{(\ell)}$ at node i varies w.r.t the feature x at node s

[2] Xu et al. [2018]

Quantifying over-squashing by the Jacobian

The Jacobian[2] $\left|\frac{\partial h_i^{(\ell)}}{\partial x_s}\right|$ tells us how much our ℓth MPNN layer $h^{(\ell)}$ at node i varies w.r.t the feature x at node s

Lemma

If $i, s \in V$ with $d_G(i,s) = r+1$ and

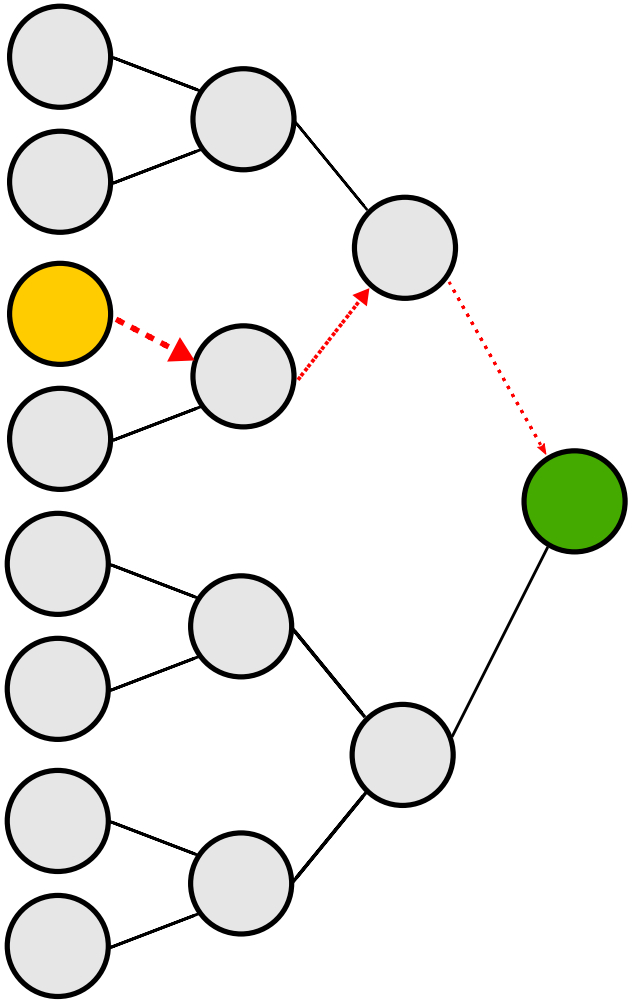
- *update functions ϕ_ℓ satisfy $|\nabla \phi_\ell| \leq \alpha$,*
- *and aggregation functions ψ_ℓ satisfy $|\nabla \psi_\ell| \leq \beta$*

for layers $\ell = 0, \ldots, r$ then

$$\left|\frac{\partial h_i^{(r+1)}}{\partial x_s}\right| \leq (\alpha\beta)^{r+1}(\hat{A}^{r+1})_{is}.$$

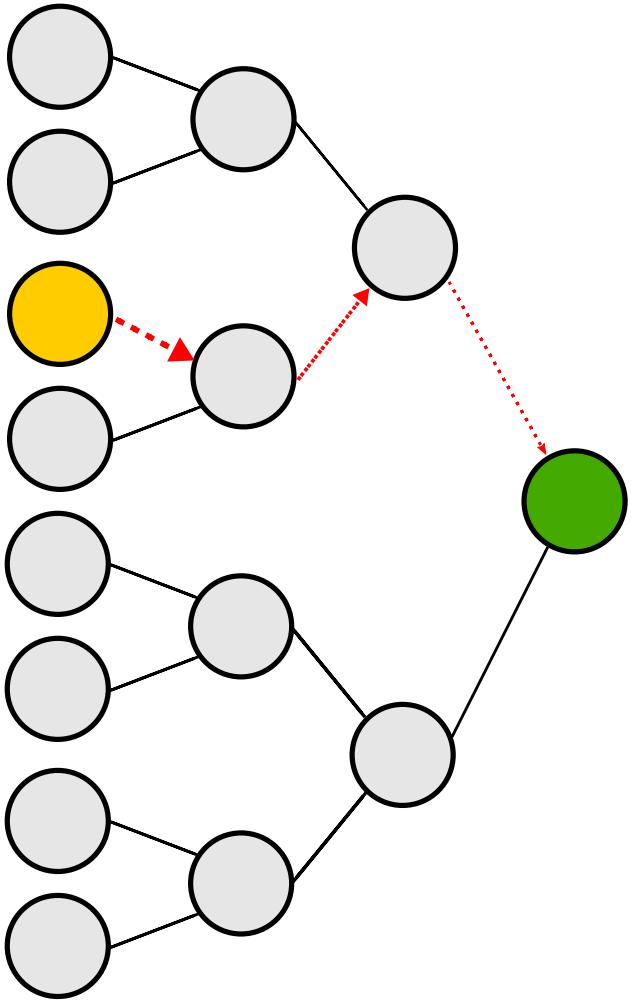
[2] Xu et al. [2018]

Over-squashing example: binary tree



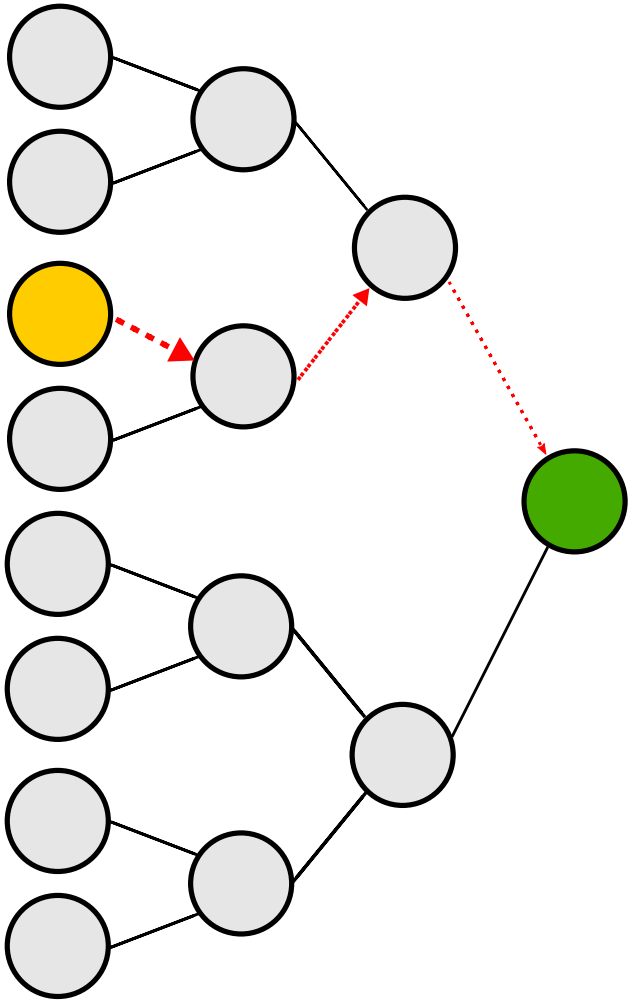
- Node s is one node in i's exponentially-growing receptive field
- $(\hat{\boldsymbol{A}}^{r+1})_{is} = \frac{1}{2} \cdot 3^{-r}$

Over-squashing example: binary tree



- Node s is one node in i's exponentially-growing receptive field
- $(\hat{\boldsymbol{A}}^{r+1})_{is} = \frac{1}{2} \cdot 3^{-r}$
- Demonstrated in Tree-NeighborsMatch experiment in Alon and Yahav [2021]

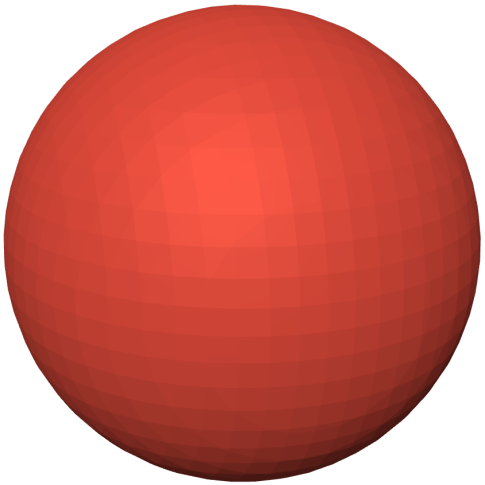
Over-squashing example: binary tree



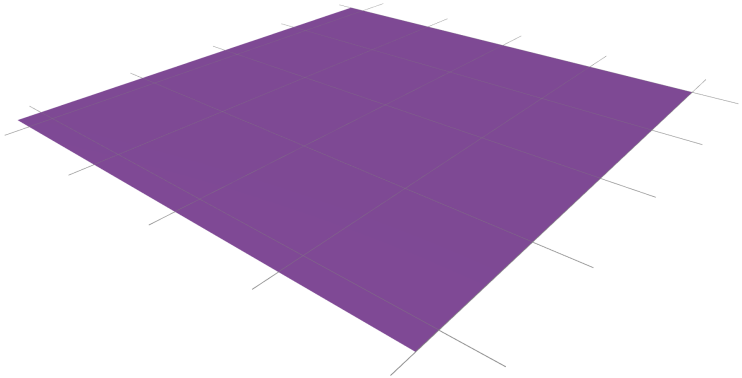
- Node s is one node in i's exponentially-growing receptive field
- $(\hat{\boldsymbol{A}}^{r+1})_{is} = \frac{1}{2} \cdot 3^{-r}$
- Demonstrated in Tree-NeighborsMatch experiment in Alon and Yahav [2021]
- If the graph topology induces over-squashing, can we identify the edges responsible for bottlenecks?

Curvature

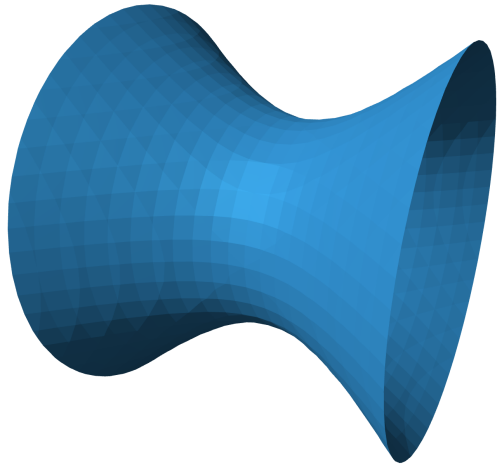
- In differential geometry, *Ricci curvature* on manifolds relates to information spreading



Spherical (> 0)



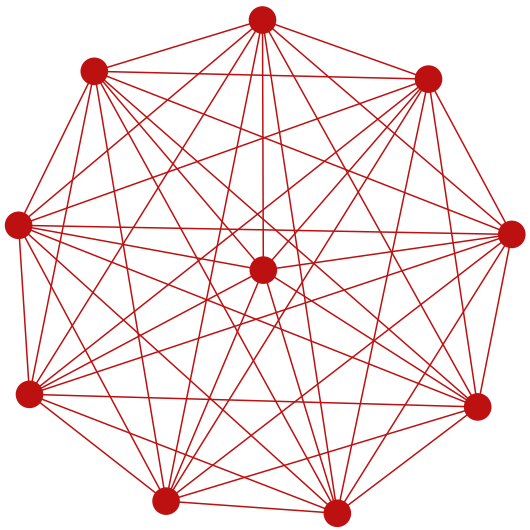
Euclidean ($= 0$)



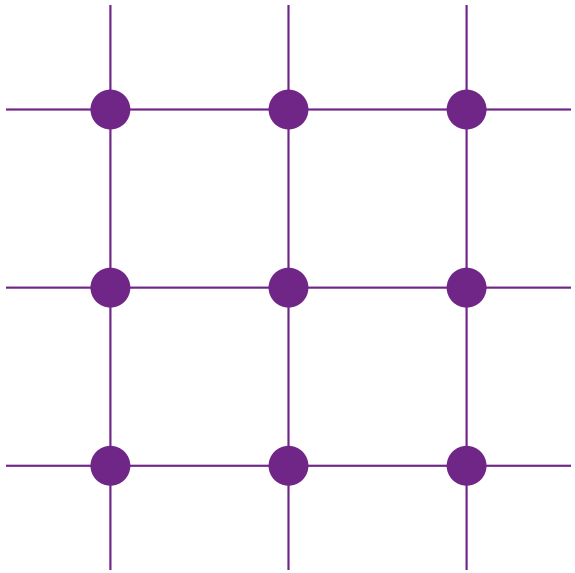
Hyperbolic (< 0)

Curvature

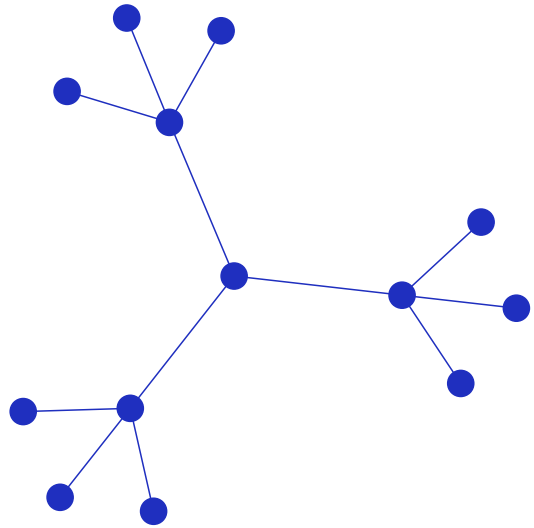
- Swapping geodesics for edges, we can take discrete analogues on graphs



Clique (> 0)



Grid ($= 0$)



Tree (< 0)

But then which curvature?

A general curvature quantity is a map $\mathrm{Ric} : E \to \mathbb{R}$ capturing some meaningful property

- **Forman curvature**[3] F:
  - Combinatorial so computationally cheap
  - but limited power (can only distinguish triangles, gives grids negative curvature)
- **Ollivier curvature**[4] κ:
  - More powerful and has nice theoretical results
  - but expensive and harder to control analytically (uses Wasserstein/earth-mover distances)
- Can we strike a balance?

[3] Forman [2003]
[4] Ollivier [2007]

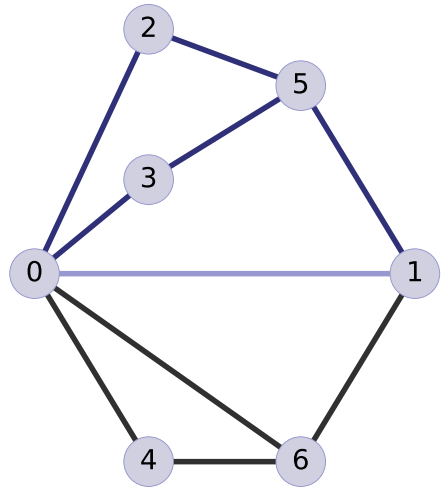
Balanced Forman curvature: preliminaries

(i) $\sharp_\Delta(i,j) := S_1(i) \cap S_1(j)$ are the triangles based at $i \sim j$.

(ii) $\sharp^i_\square(i,j)$ are the neighbors of i forming a 4-cycle based at $i \sim j$ *without* diagonals inside.

(iii) $\gamma_{\max}(i,j)$ is the maximal number of 4 cycles based at $i \sim j$ traversing a common node.

The degeneracy factor $\gamma_{\max}(0,1) = 2$ there exist two 4 cycles passing the same node (5)



Balance Forman curvature: the definition

Consider an edge $i \sim j$ with d_i, d_j degrees of i and j
We define $\mathrm{Ric}(i,j)$ to be zero if $\min\{d_i, d_j\} = 1$ and otherwise

$$\mathrm{Ric}(i,j) = \frac{2}{d_i} + \frac{2}{d_j} - 2 + 2\frac{|\sharp_\Delta(i,j)|}{\max\{d_i, d_j\}} + \frac{|\sharp_\Delta(i,j)|}{\min\{d_i, d_j\}} + \frac{(\gamma_{\max})^{-1}}{\max\{d_i, d_j\}}(|\sharp^i_\square| + |\sharp^j_\square|).$$

Balance Forman curvature: the definition

Consider an edge $i \sim j$ with d_i, d_j degrees of i and j
We define $\mathrm{Ric}(i,j)$ to be zero if $\min\{d_i, d_j\} = 1$ and otherwise

$$\mathrm{Ric}(i,j) = \frac{2}{d_i} + \frac{2}{d_j} - 2 + 2\frac{|\sharp_\Delta(i,j)|}{\max\{d_i, d_j\}} + \frac{|\sharp_\Delta(i,j)|}{\min\{d_i, d_j\}} + \frac{(\gamma_{\max})^{-1}}{\max\{d_i, d_j\}}(|\sharp^i_\square| + |\sharp^j_\square|).$$

G	Ric_G
Cycle $C_{n\geq 5}$	0
Grid G_n	0
Clique K_n	$\frac{n}{n-1}$
Tree T_r	$\frac{4}{r+1} - 2$

Table: Ric can distinguish 4-cycles along with triangles

Does it make sense?

Theorem

Given an unweighted graph G, for any edge $i \sim j$ we have $\kappa(i,j) \geq \mathrm{Ric}(i,j)$.

[5] Paeng [2012]

Does it make sense?

Theorem

Given an unweighted graph G, for any edge $i \sim j$ we have $\kappa(i,j) \geq \mathrm{Ric}(i,j)$.

Intuition: Up to 4-cycles the two curvature notions are roughly the same.
Consequences: Connection between Wasserstein-based κ and Ric
$\rightarrow$ We can borrow several results[5] known for Ollivier curvature, e.g.

Corollary

If $\mathrm{Ric}(i,j) \geq k > 0$ for any edge $i \sim j$, then there exists a polynomial P s.t.

$$|B_r(i)| \leq P(r), \quad \forall i \in V.$$

[5] Paeng [2012]

Negatively curved edges cause over-squashing

If $\mathrm{Ric} > 0$ everywhere $\Rightarrow \sharp$ nodes in r-hops is polynomial $\Rightarrow$ *no over-squashing*

Are negatively curved edges responsible for the over-squashing?

Negatively curved edges cause over-squashing

If $\mathrm{Ric} > 0$ everywhere $\Rightarrow \sharp$ nodes in r-hops is polynomial $\Rightarrow$ *no over-squashing*

Are negatively curved edges responsible for the over-squashing?

Convention: We say that $\mathrm{Ric}(i,j)$ is *very* negative if there exists $\delta > 0$ s.t. $0 < \delta < (\max\{d_i, d_j\})^{-\frac{1}{2}}$, $\delta < \gamma_{\max}^{-1}$ and $\mathrm{Ric}(i,j) \leq -2 + \delta$.

If $\mathrm{Ric}(i,j)$ is very negative $\rightarrow$ exclude pathological cases with many 4-cycles traversing the same node

Negatively curved edges cause over-squashing

Informal statement: *If* Ric *is very negative along* $i \sim j$*, then there exist many nodes at hop-distance 2 from* i *s.t. MPNN struggle to send messages from* i *to such nodes in 2 layers.*

Negatively curved edges cause over-squashing

Informal statement: *If* Ric *is very negative along* $i \sim j$*, then there exist many nodes at hop-distance 2 from* i *s.t. MPNN struggle to send messages from* i *to such nodes in 2 layers.*

Theorem

Let $i \sim j$ *with* $d_i \leq d_j$ *and assume that* $\mathrm{Ric}(i,j)$ *is very negative. Then there exists* $Q_j \subset S_2(i)$ *satisfying* $|Q_j| > \delta^{-1}$ *and for* $0 \leq \ell_0 \leq L-2$ *we have*

$$\frac{1}{|Q_j|} \sum_{k \in Q_j} \left| \frac{\partial h_k^{(\ell_0+2)}}{\partial h_i^{(\ell_0)}} \right| < (\alpha\beta)^2 \delta^{\frac{1}{4}}.$$

Some comments

Recall that we know that graph topology leads to over-squashing:

$$\left| \frac{\partial h_i^{(r+1)}}{\partial x_s} \right| \leq (\alpha\beta)^{r+1} (\hat{A}^{r+1})_{is}.$$

Some comments

Recall that we know that graph topology leads to over-squashing:

$$\left|\frac{\partial h_i^{(r+1)}}{\partial x_s}\right| \leq (\alpha\beta)^{r+1}(\hat{A}^{r+1})_{is}.$$

We have provided a more local and refined 'expansion' of the right hand side:

Negatively curved edges $\rightarrow$ *bottlenecks* $\rightarrow$ *over-squashing.*

A benefit of this analysis $\rightarrow$ we can surgically identify bottlenecks by studying the curvature.

Bottlenecks from a spectral point of view

Suppose G has two communities separated by few edges $\rightarrow$ bottleneck

[6] Cheeger [2015], Chung and Graham [1997]

Bottlenecks from a spectral point of view

Suppose G has two communities separated by few edges $\rightarrow$ bottleneck
This property is captured by the **Cheeger constant**

$$h_G := \min_{S \subset V} h_S, \qquad h_S := \min_{S \subset V} \frac{|\partial S|}{\min\{\mathrm{vol}(S), \mathrm{vol}(V \setminus S)\}} \tag{1}$$

where $\partial S = \{(i,j) : i \in S,\ j \in V \setminus S\}$ and $\mathrm{vol}(S) = \sum_{i \in S} d_i$.

[6] Cheeger [2015], Chung and Graham [1997]

Bottlenecks from a spectral point of view

Suppose G has two communities separated by few edges $\rightarrow$ bottleneck
This property is captured by the **Cheeger constant**

$$h_G := \min_{S \subset V} h_S, \qquad h_S := \min_{S \subset V} \frac{|\partial S|}{\min\{\mathrm{vol}(S), \mathrm{vol}(V \setminus S)\}} \tag{1}$$

where $\partial S = \{(i,j) : i \in S,\ j \in V \setminus S\}$ and $\mathrm{vol}(S) = \sum_{i \in S} d_i$.

The main result about the Cheeger constant is the *Cheeger inequality* [6]

$$2h_G \geq \lambda_1 \geq \frac{h_G^2}{2}. \tag{2}$$

[6] Cheeger [2015], Chung and Graham [1997]

Curvature and spectral properties

Messages from different communities go through bridges $\rightarrow$ over-squashing

h_G is a rough measure of graph 'bottleneckedness'

Negatively curved edges induce bottlenecks: relation between Ric and h_G?

Curvature and spectral properties

Messages from different communities go through bridges $\rightarrow$ over-squashing

h_G is a rough measure of graph 'bottleneckedness'

Negatively curved edges induce bottlenecks: relation between Ric and h_G?

Corollary

If $\mathrm{Ric}(i,j) \geq k > 0$ *for all* $i \sim j$, *then* $\lambda_1 \geq 2h_G \geq k$.

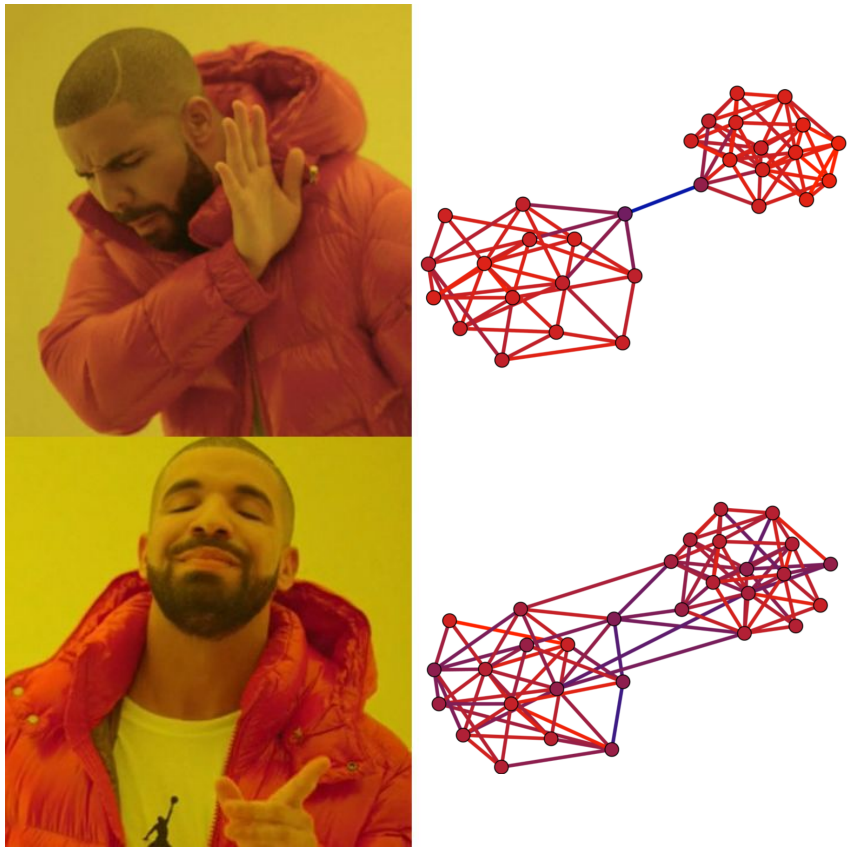
Follows from comparison theorem $\kappa \geq \mathrm{Ric}$ and Lin et al. [2011]

Graph rewiring as preprocessing

- Graph structure causing problems with MPNN learning?

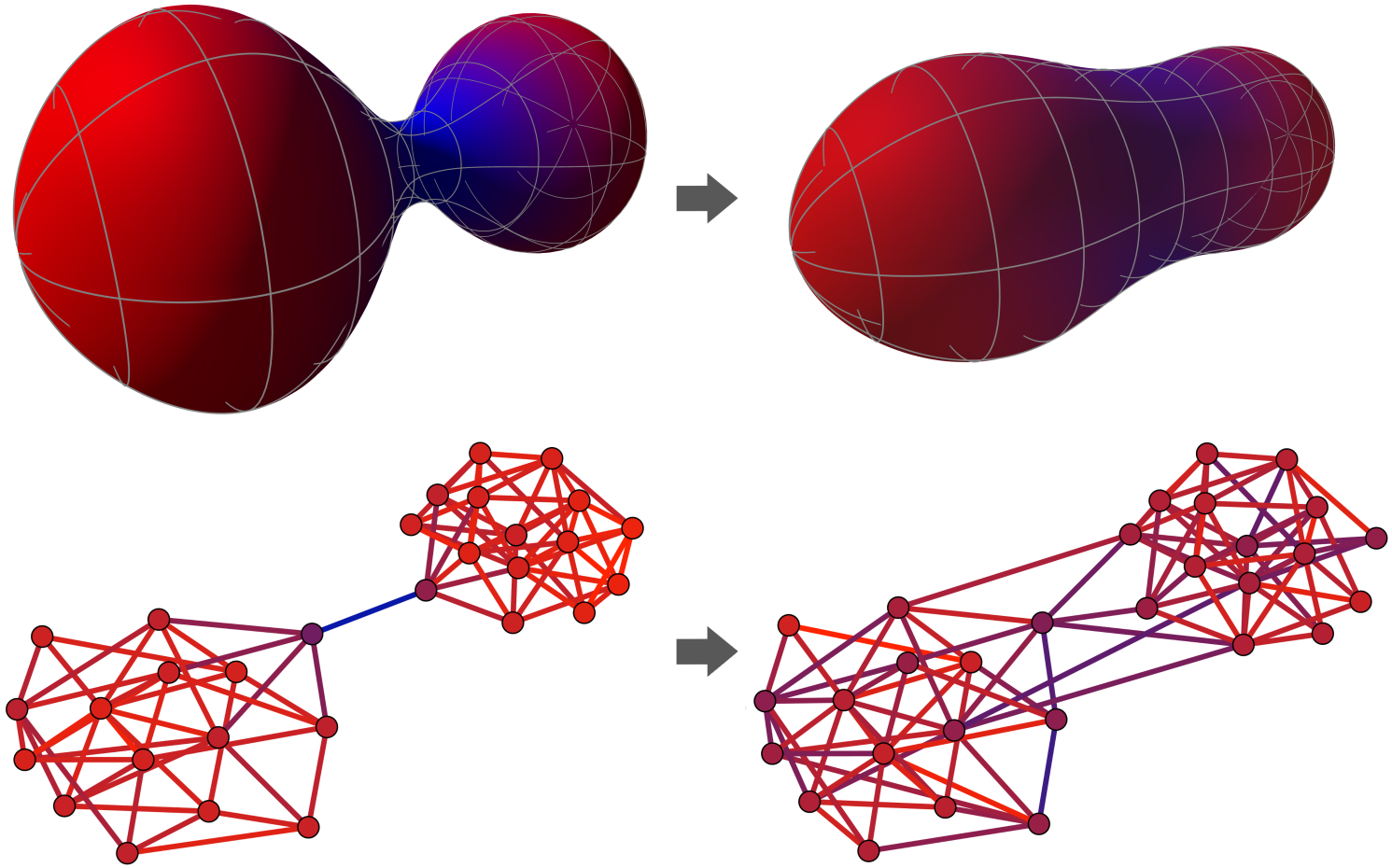
Graph rewiring as preprocessing

- Graph structure causing problems with MPNN learning?
- Change it! (and use the resulting topology for GNN input)



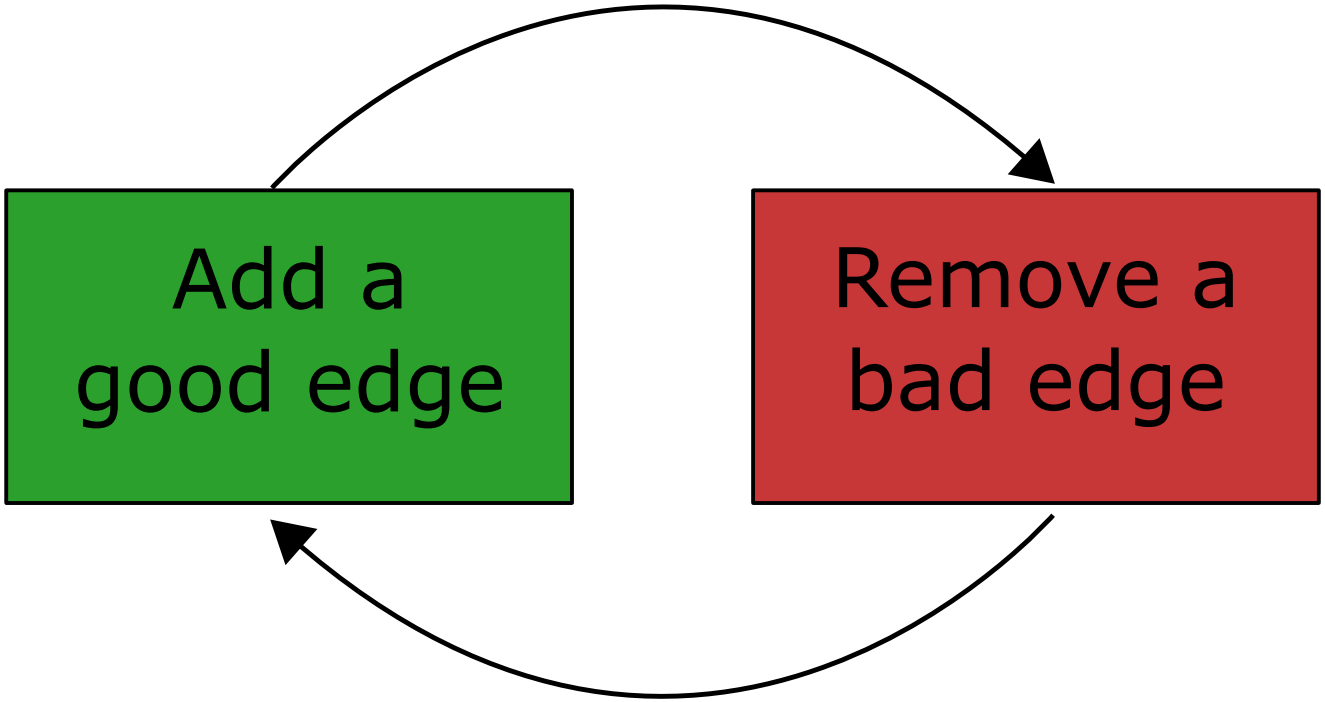
Curvature-based rewiring

- Can we take a process on the graph to normalize curvature and help alleviate bottlenecks?



Stochastic Discrete Ricci Flow

- High-level:



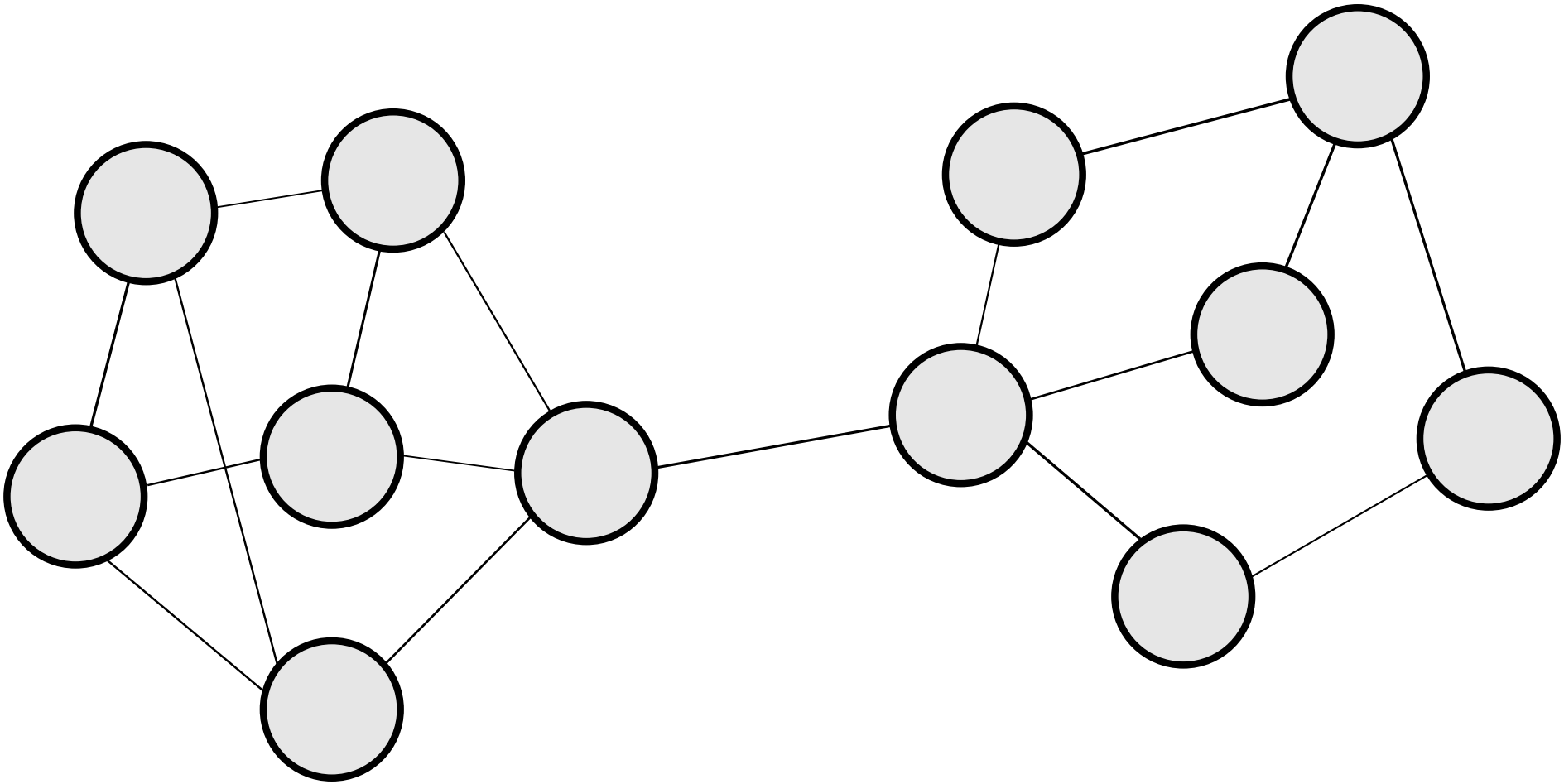
- What's good and what's bad? Decide with curvature

Stochastic Discrete Ricci Flow

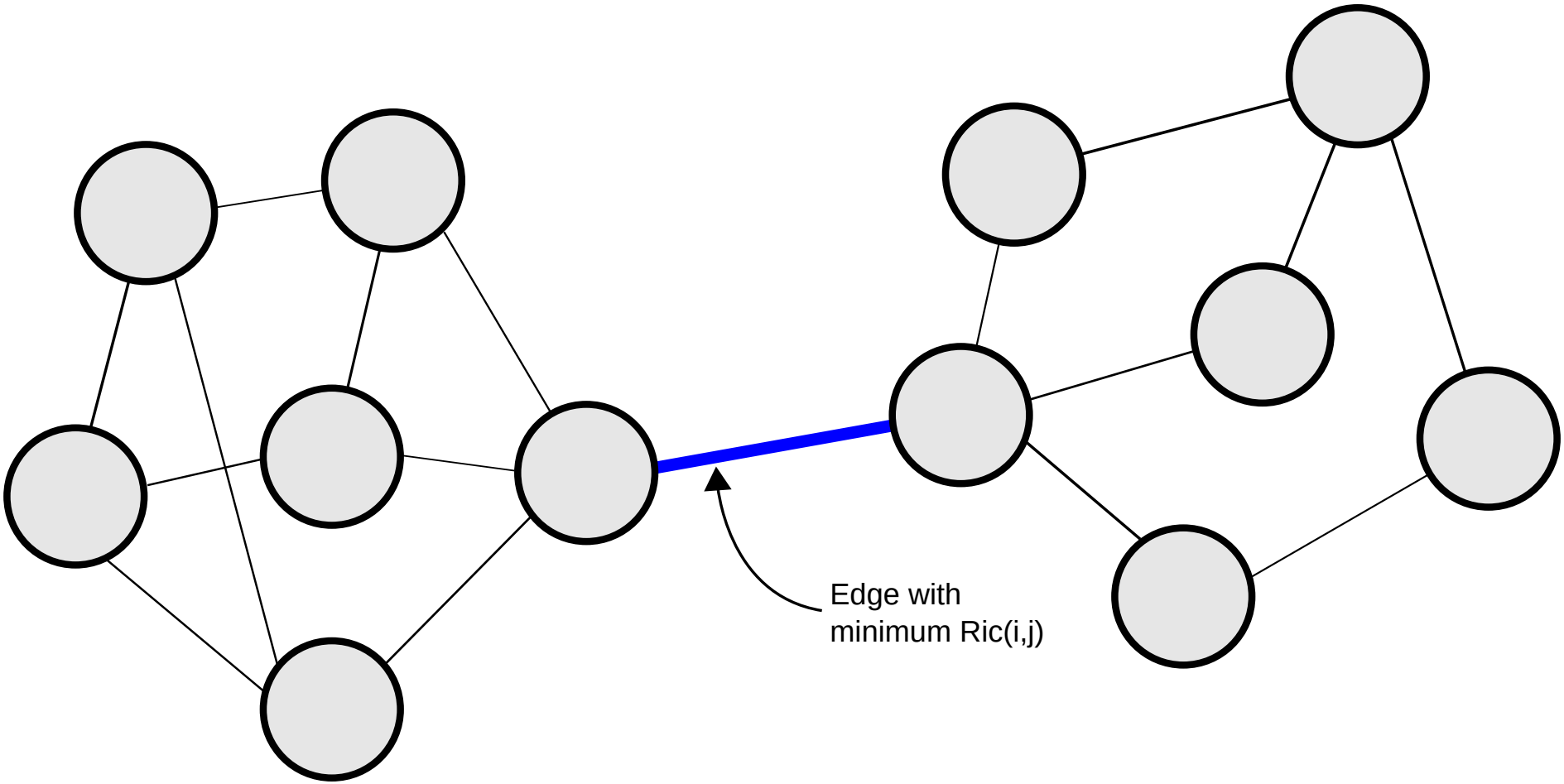
Algorithm 1: Stochastic Discrete Ricci Flow (SDRF)

Input: graph G, temperature $\tau > 0$, max number of iterations, optional Ric upper-bound C^+
Repeat
  1) For edge $i \sim j$ with minimal Ricci curvature $\text{Ric}(i, j)$:
    Calculate vector $\boldsymbol{x}$ where $x_{kl} = \text{Ric}_{kl}(i, j) - \text{Ric}(i, j)$, the improvement to $\text{Ric}(i, j)$ from adding edge $k \sim l$ where $k \in B_1(i), l \in B_1(j)$;
    Sample index k, l with probability $\text{softmax}(\tau \boldsymbol{x})_{kl}$ and add edge $k \sim l$ to G.
  2) Remove edge $i \sim j$ with maximal Ricci curvature $\text{Ric}(i, j)$ if $\text{Ric}(i, j) > C^+$.
Until convergence, or max iterations reached;

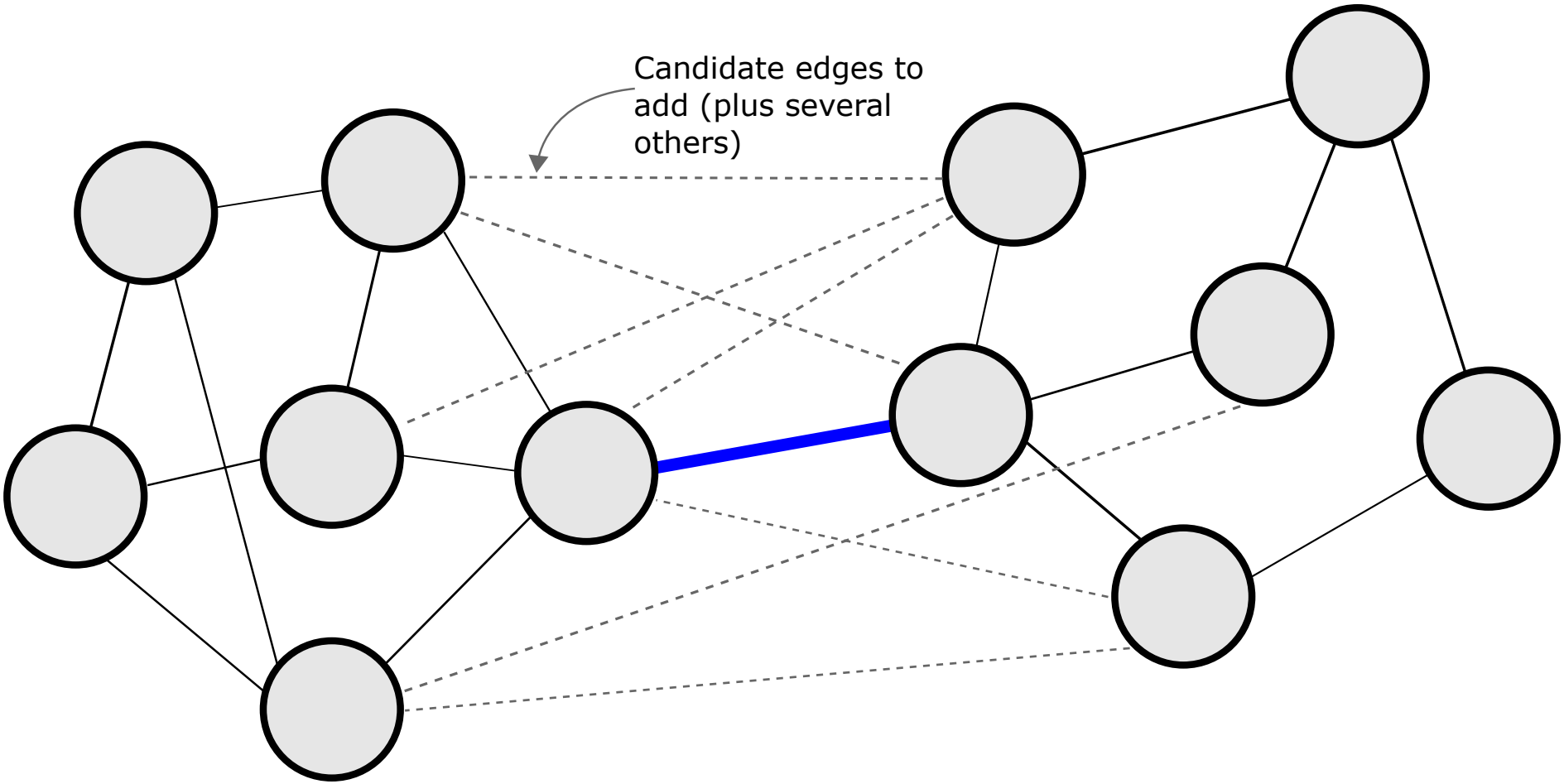
SDRF: Example



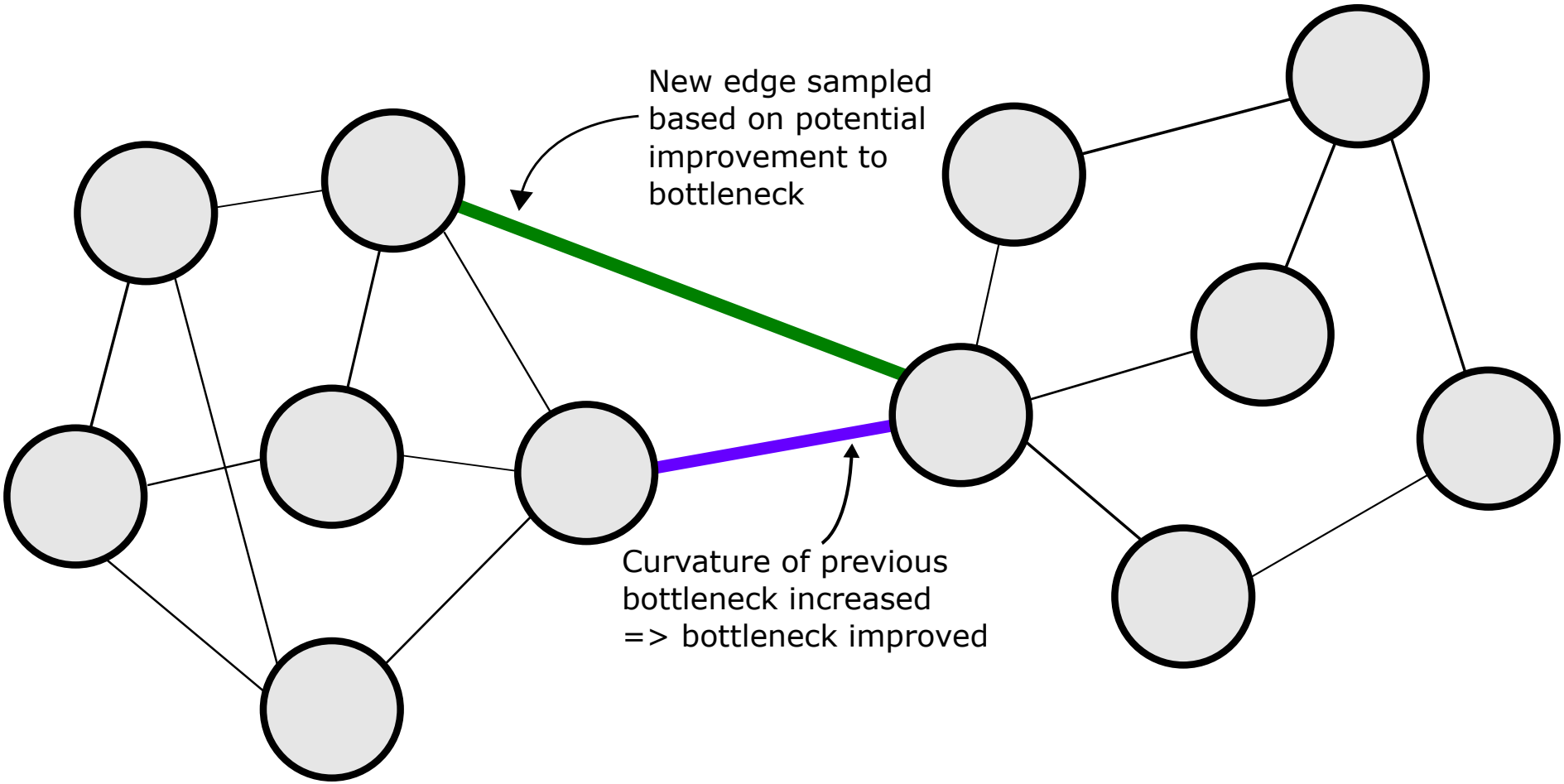
SDRF: Example



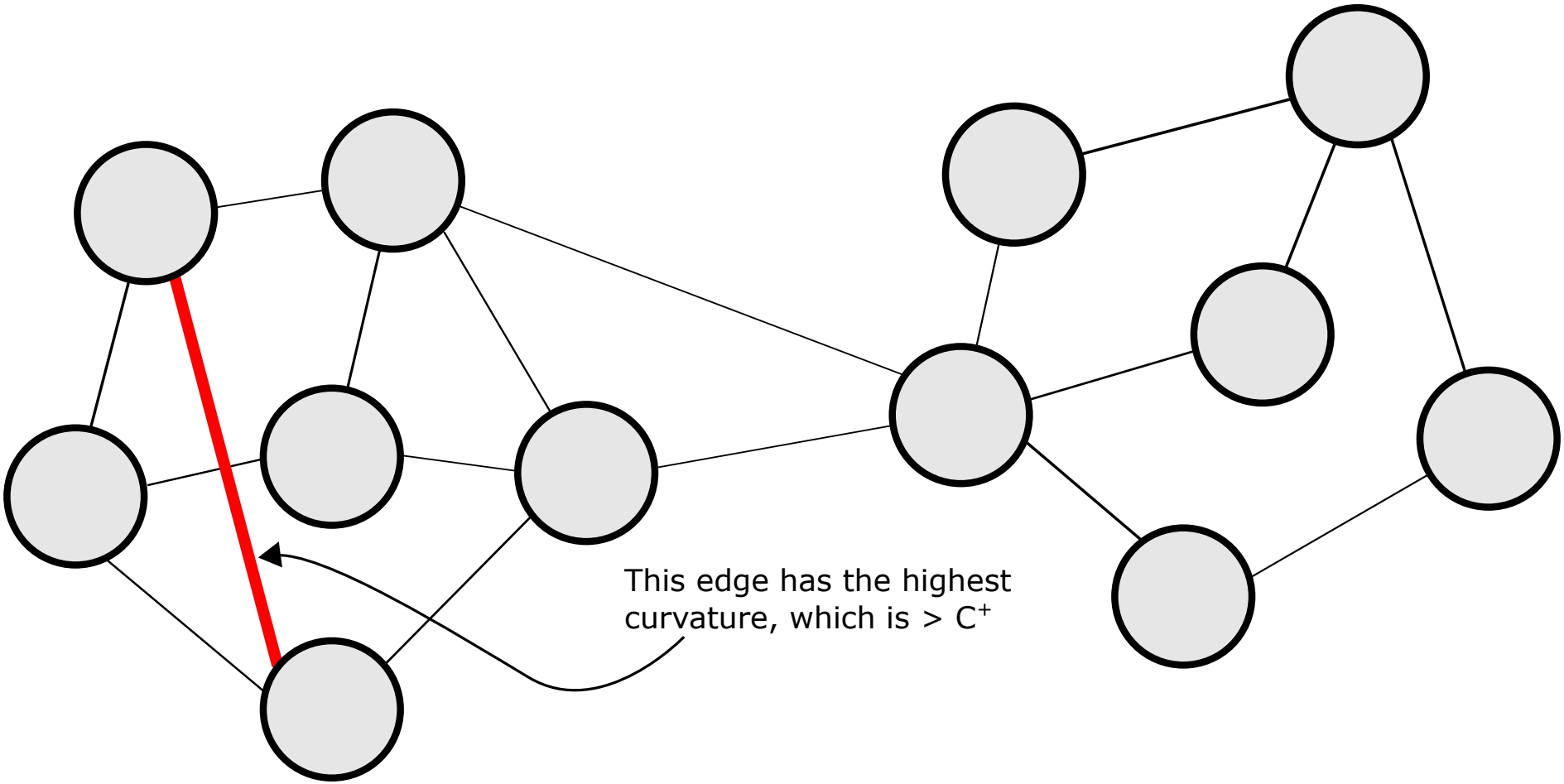
SDRF: Example



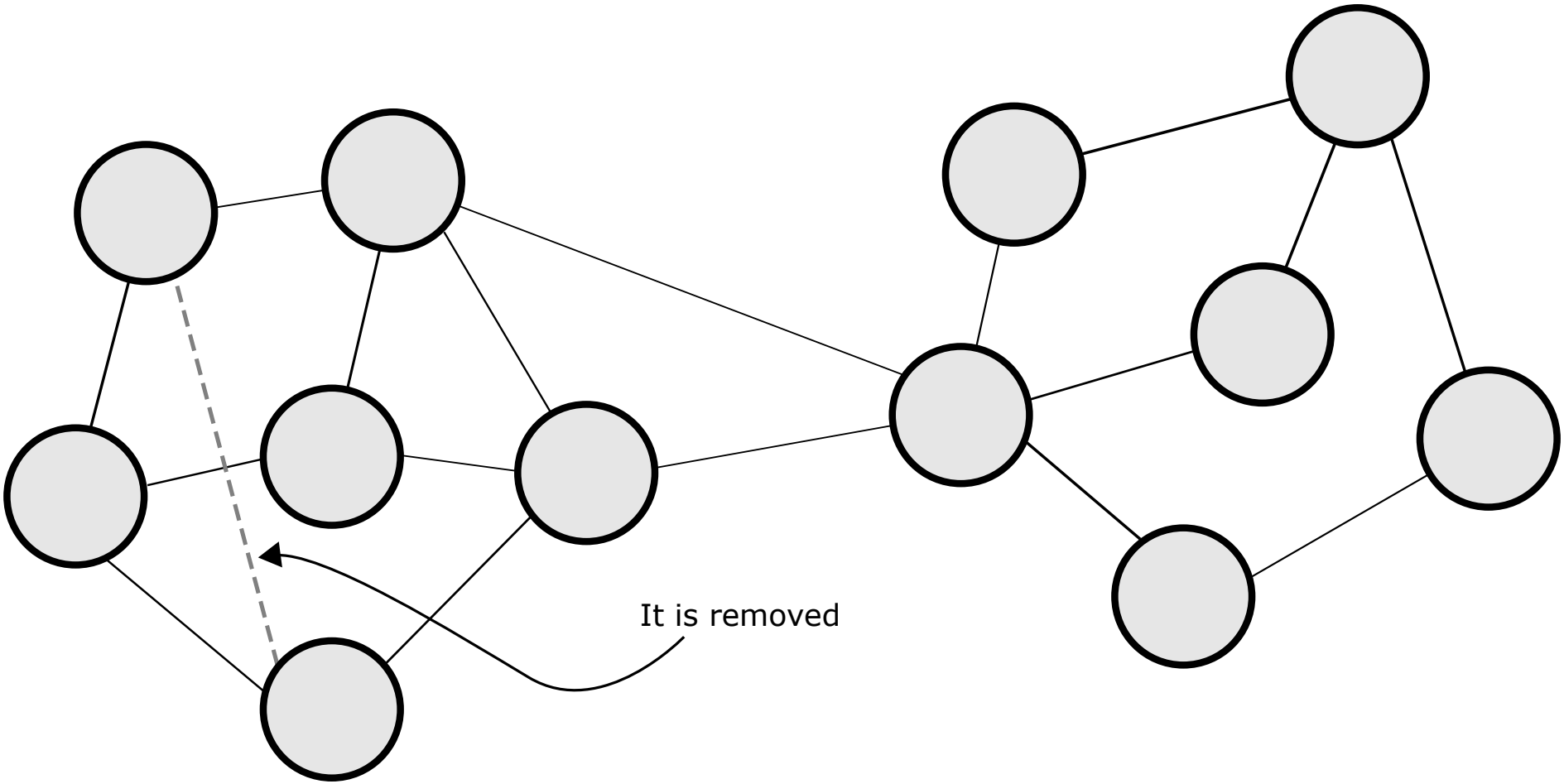
SDRF: Example



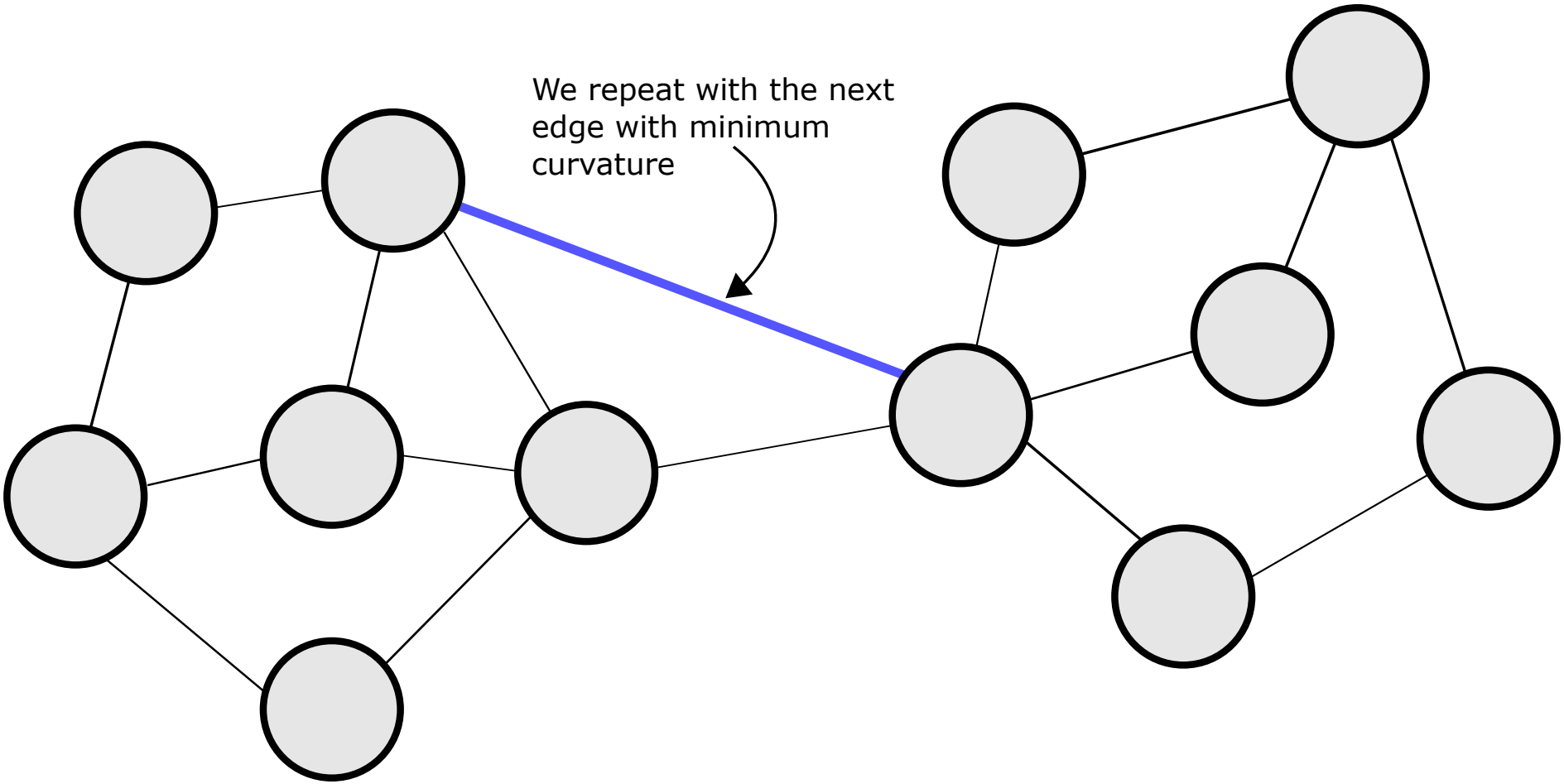
SDRF: Example



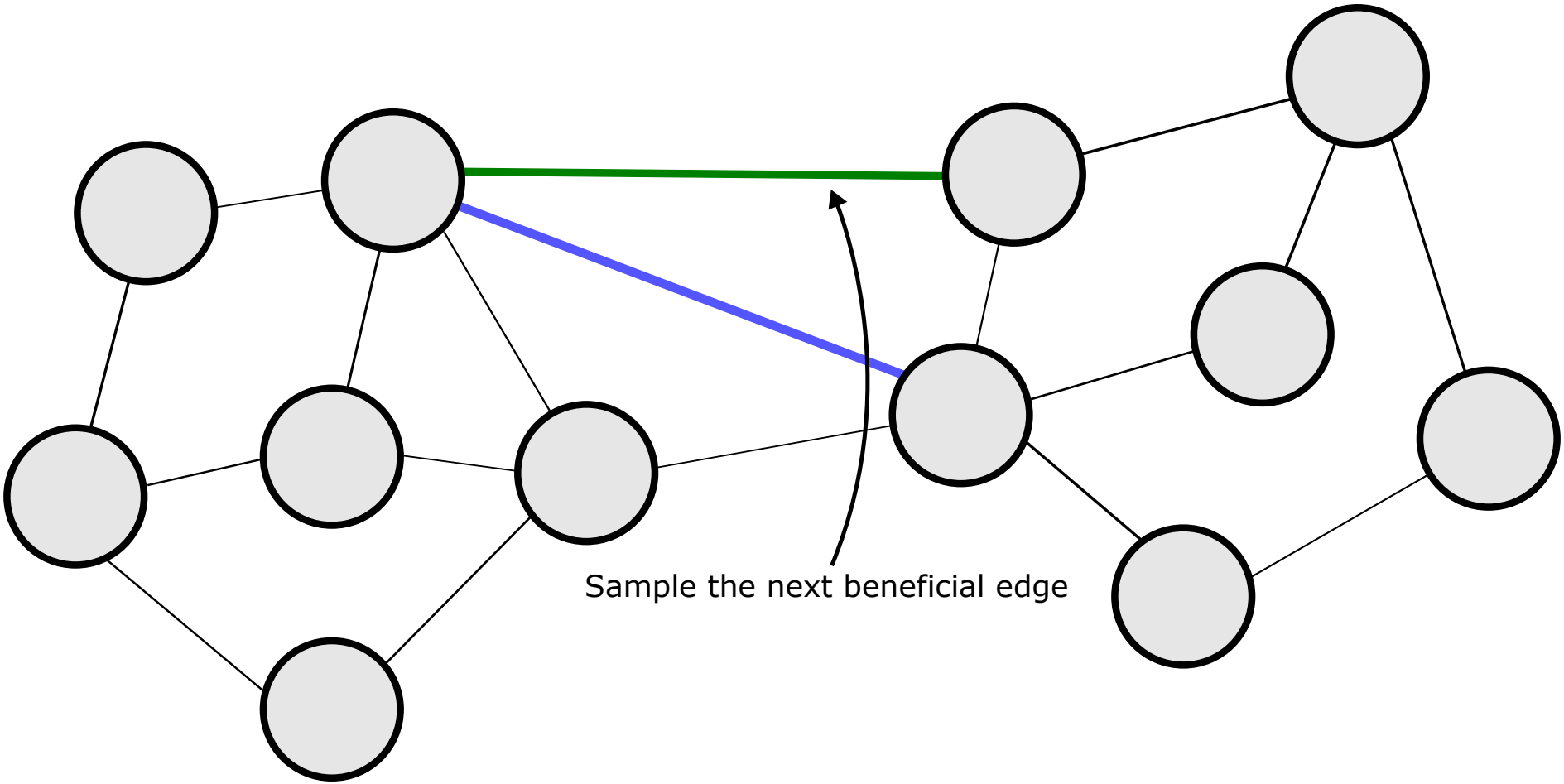
SDRF: Example



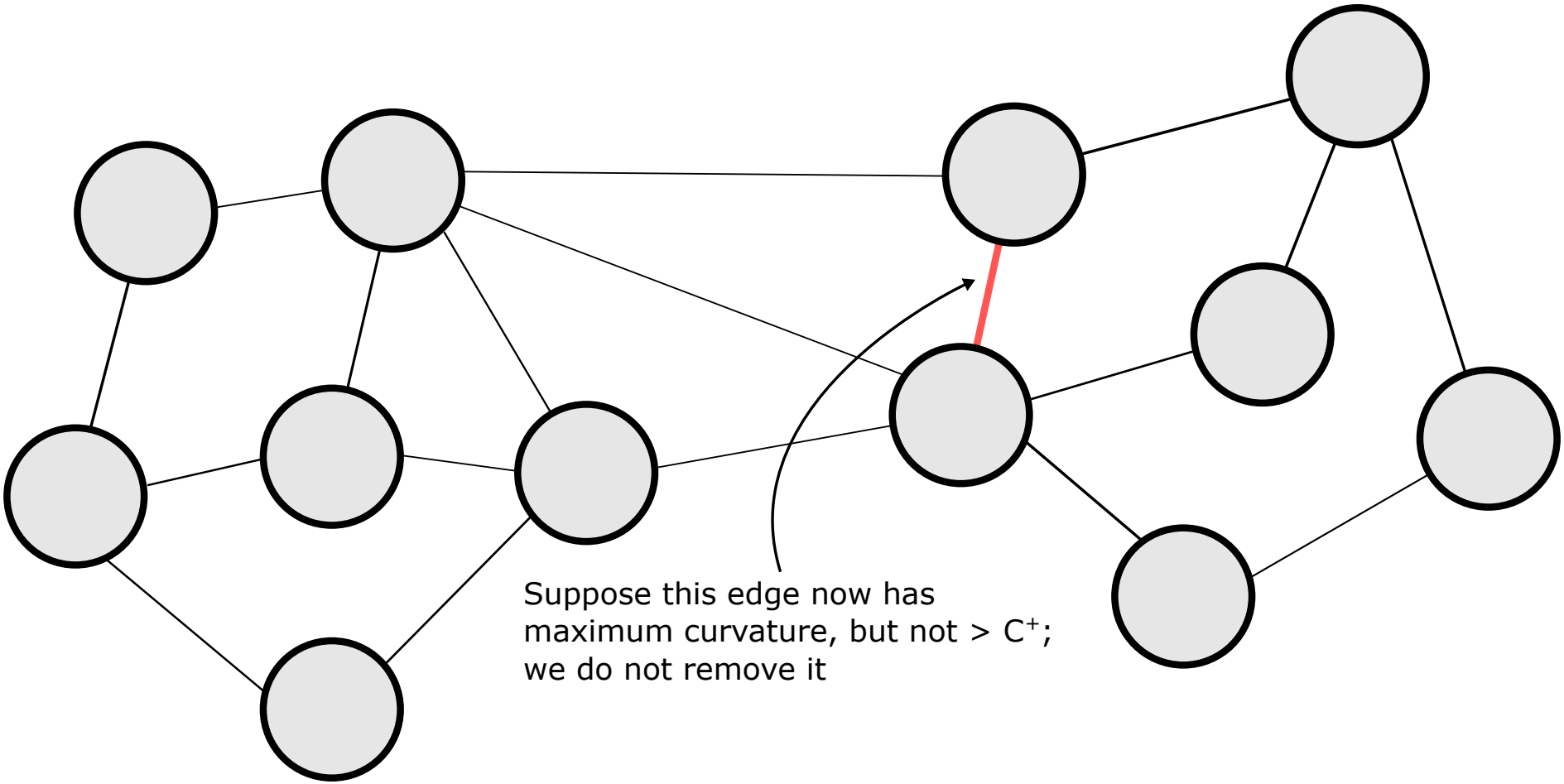
SDRF: Example



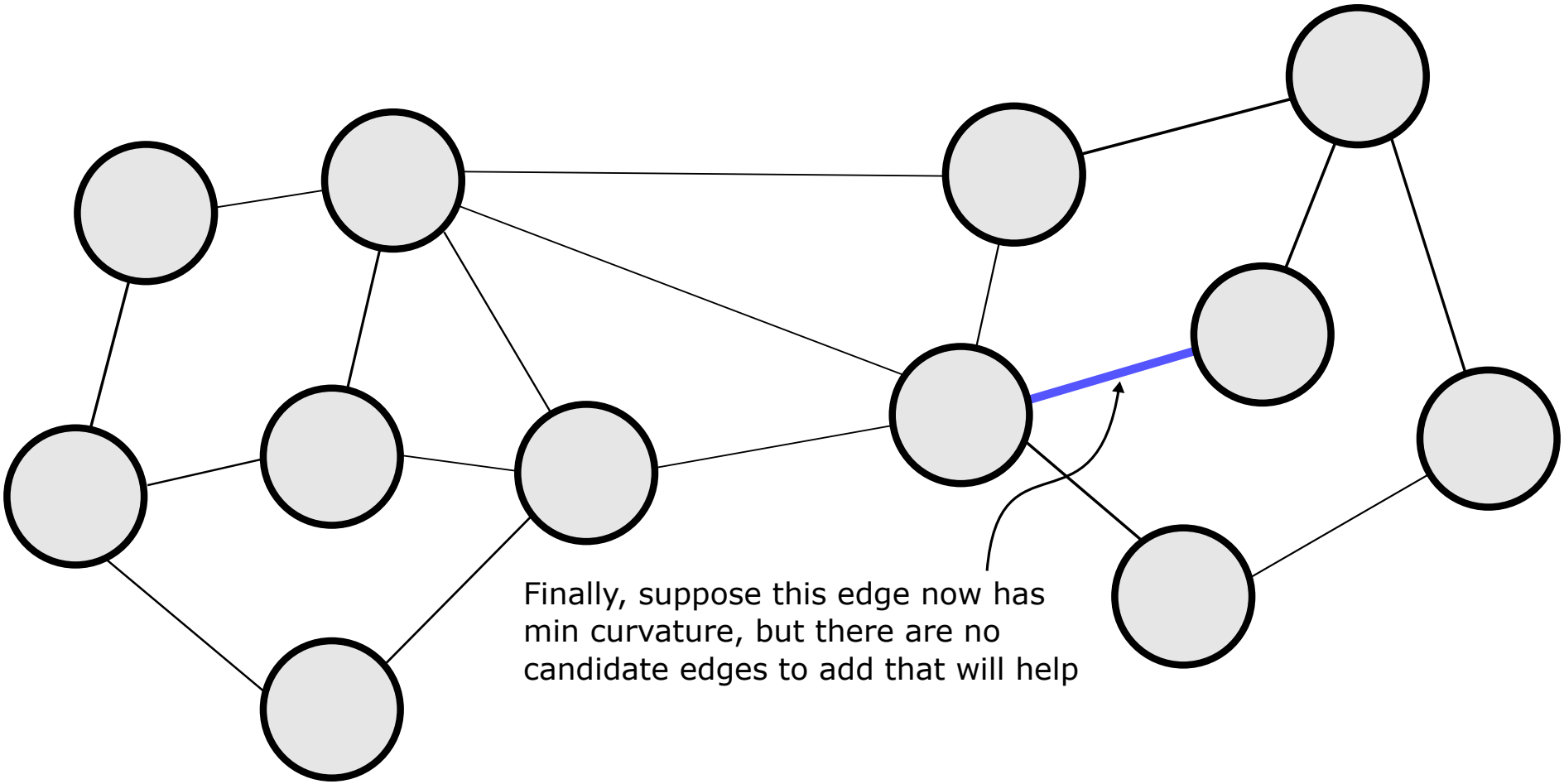
SDRF: Example



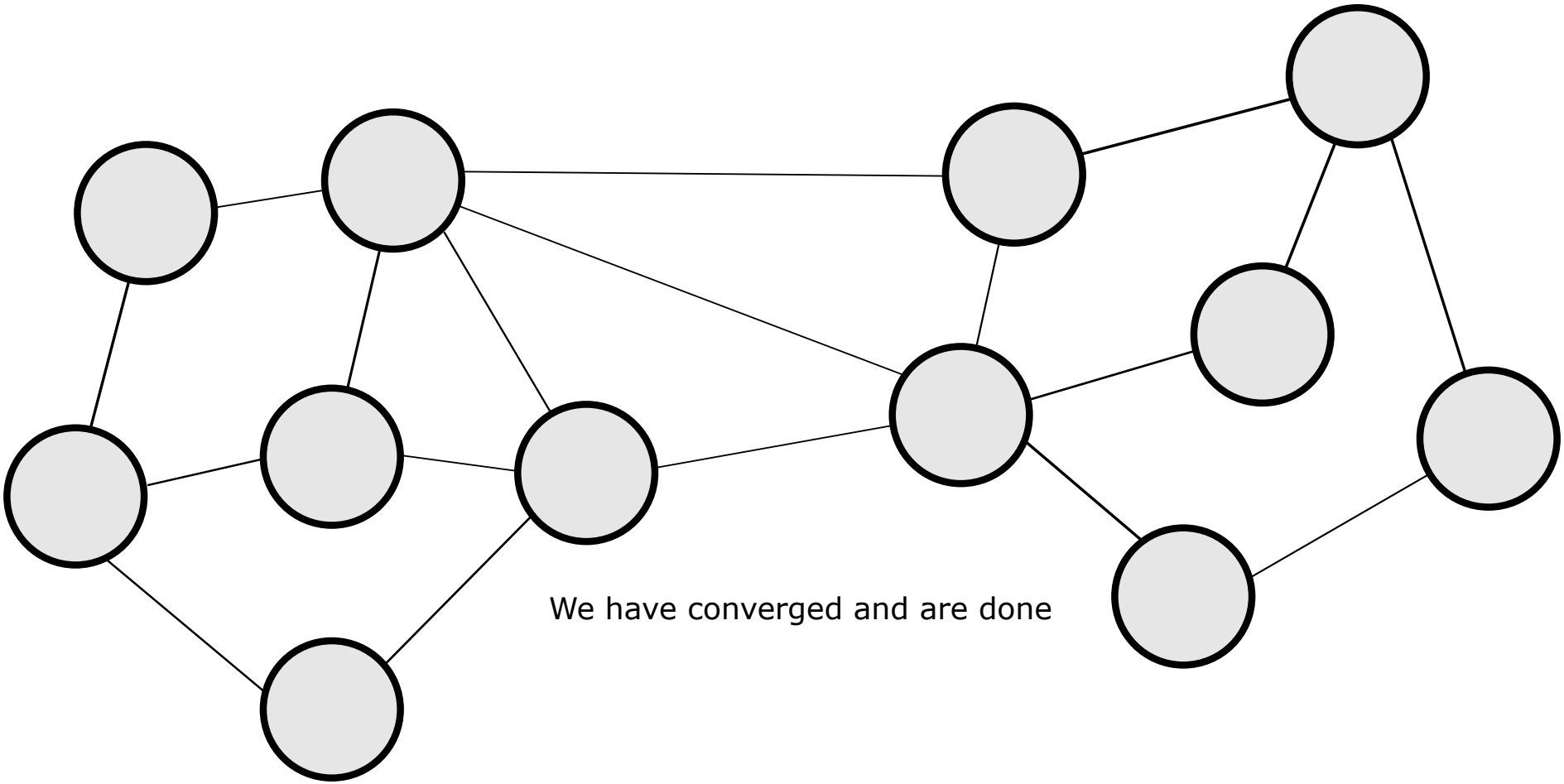
SDRF: Example



SDRF: Example

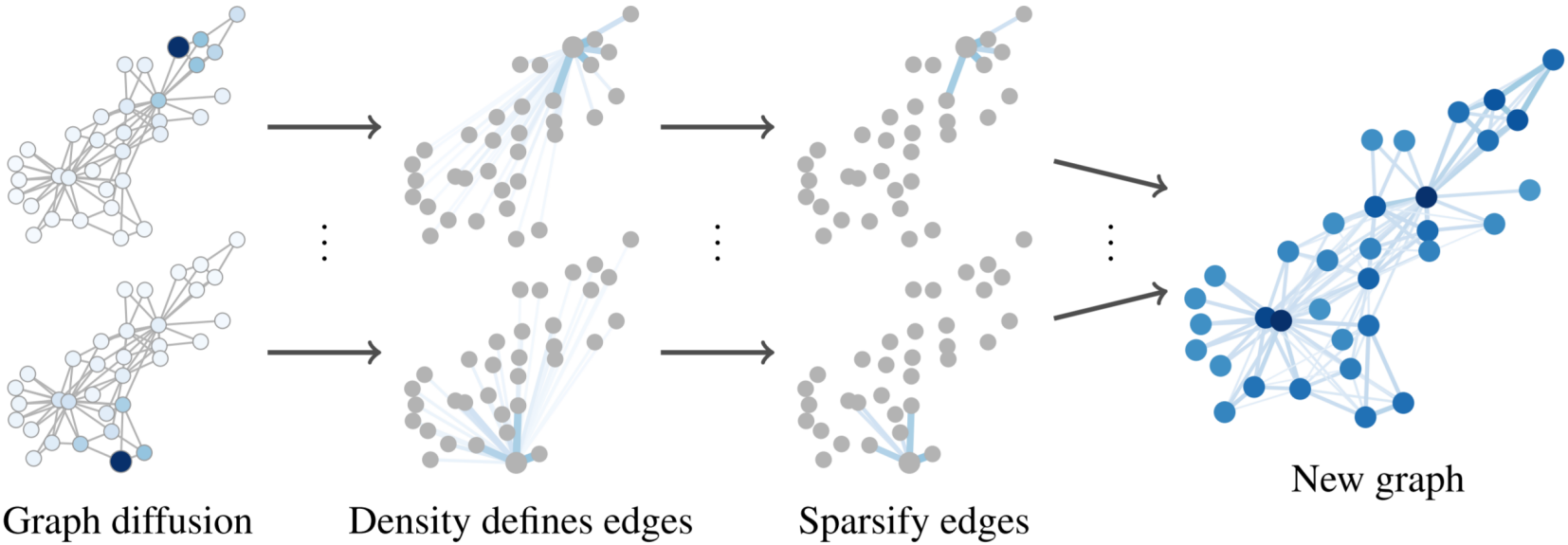


SDRF: Example



What about DIGL / Graph Diffusion Convolution?

- DIGL[7] rewires the graph by graph diffusion
- Leads to significant improvements in performance on a range of models and datasets



[7] Klicpera et al. [2019]

What about DIGL / Graph Diffusion Convolution?

- Based on an assumption of homophily - common but not guaranteed
- Considering GDC with the PPR (Personalized Page Rank) kernel

$$R_\alpha := \sum_{k=0}^{\infty} \theta_k^{PPR} (D^{-1}A)^k = \alpha \sum_{k=0}^{\infty} \left((1-\alpha)(D^{-1}A)\right)^k$$

we obtain a constraint on the Cheeger constant of the rewired graph:

Theorem

Let $S \subset V$ with $vol(S) \leq vol(G)/2$. Then $h_{S,\alpha} \leq \left(\frac{1-\alpha}{\alpha}\right) \frac{d_{avg}(S)}{d_{min}(S)} h_S$, where $d_{avg}(S)$ and $d_{min}(S)$ are the average and minimum degree on S, respectively.

- Try both!

Experimental results

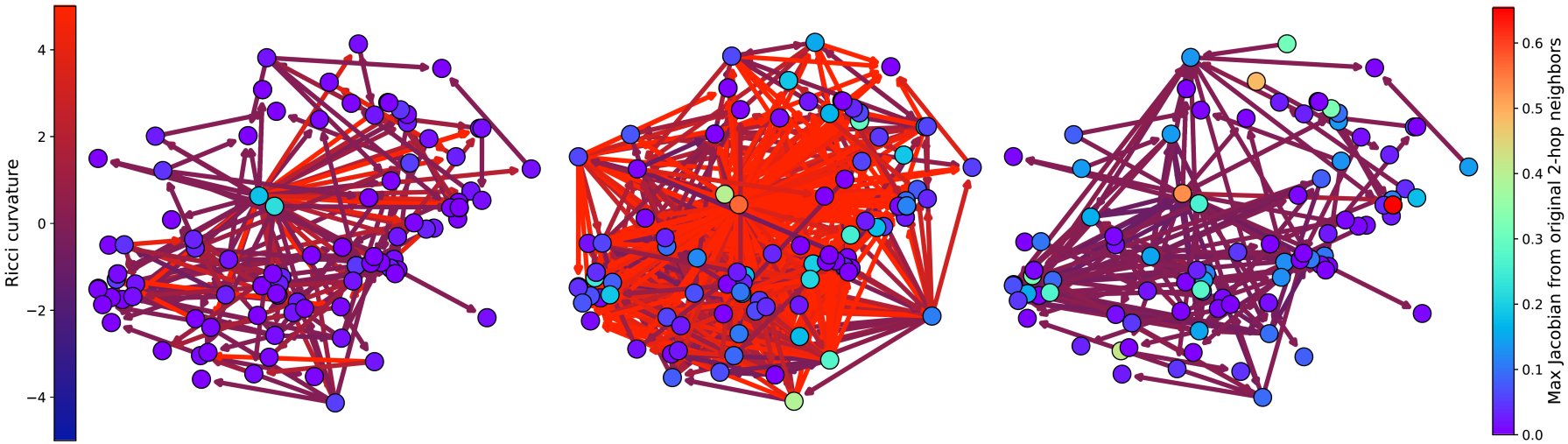
- Experiment: semi-supervised node classification with a simple GCN using the given rewiring for pre-processing
- +FA is making the last layer fully connected as in Alon and Yahav [2021]
- $\mathcal{H}(G)$ is a measure of homophily [Pei et al., 2019]

	Cornell	Texas	Wisconsin	Chameleon	Squirrel	Actor	Cora	Citeseer	Pubmed
$\mathcal{H}(G)$	0.11	0.06	0.16	0.25	0.22	0.24	0.83	0.71	0.79
None	52.69 ± 0.21	61.19 ± 0.49	54.60 ± 0.86	41.33 ± 0.18	30.32 ± 0.99	23.84 ± 0.43	81.89 ± 0.79	72.31 ± 0.17	78.16 ± 0.23
Undirected	53.20 ± 0.53	63.38 ± 0.87	51.37 ± 1.15	42.02 ± 0.30	35.53 ± 0.78	21.45 ± 0.47	-	-	-
+FA	$\mathbf{58.29 \pm 0.49}$	$\mathbf{64.82 \pm 0.29}$	55.48 ± 0.62	42.67 ± 0.17	36.86 ± 0.44	24.14 ± 0.43	81.65 ± 0.18	70.47 ± 0.18	$\mathbf{79.48 \pm 0.12}$
DIGL (PPR)	58.26 ± 0.50	62.03 ± 0.43	49.53 ± 0.27	42.02 ± 0.13	33.22 ± 0.14	24.77 ± 0.32	$\mathbf{83.21 \pm 0.27}$	$\mathbf{73.29 \pm 0.17}$	78.84 ± 0.08
DIGL + Undirected	$\mathbf{59.54 \pm 0.64}$	63.54 ± 0.38	52.23 ± 0.54	42.68 ± 0.12	32.48 ± 0.23	25.45 ± 0.30	-	-	-
SDRF	54.60 ± 0.39	64.46 ± 0.38	$\mathbf{55.51 \pm 0.27}$	$\mathbf{42.73 \pm 0.15}$	$\mathbf{37.05 \pm 0.17}$	$\mathbf{28.42 \pm 0.75}$	$\mathbf{82.76 \pm 0.23}$	$\mathbf{72.58 \pm 0.20}$	$\mathbf{79.10 \pm 0.11}$
SDRF + Undirected	57.54 ± 0.34	$\mathbf{70.35 \pm 0.60}$	$\mathbf{61.55 \pm 0.86}$	$\mathbf{44.46 \pm 0.17}$	$\mathbf{37.67 \pm 0.23}$	$\mathbf{28.35 \pm 0.06}$	-	-	-

- SDRF improves on the baseline in all cases; DIGL in all but one
- SDRF tends to perform better in more heterophilic settings, and DIGL in more homophilic settings

Graph structure preservation

- With SDRF, the graph-edit distance between the original and pre-processed graphs is upper-bounded by 2 $\times$ the max iterations

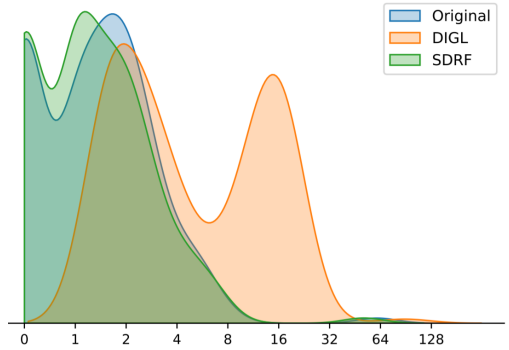


- Left, original Cornell; center, after DIGL; right, after SDRF
- SDRF changes the graph less while achieving similar improvements on the Jacobian

Distributions of node degree

	DIGL	SDRF
Cornell	351.1% / 0.0%	7.8% / 33.3%
Texas	483.3% / 0.0%	2.4% / 10.4%
Wisconsin	300.6% / 0.0%	1.4% / 7.5%
Chameleon	336.1% / 11.8%	5.6% / 5.6%
Squirrel	73.2% / 66.4%	4.2% / 4.2%
Actor	8331.3% / 0.0%	1.9% / 3.0%
Cora	3038.0% / 0.5%	1.0% / 1.0%
Citeseer	2568.3% / 0.0%	1.1% / 1.1%
Pubmed	2747.1% / 0.1%	0.2% / 0.2%

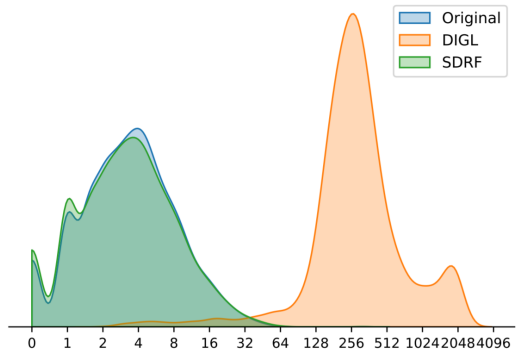
% edges added / removed by method.



(a) Wisconsin:

$W_1(\text{Original}, \text{DIGL}) = 11.83$

$W_1(\text{Original}, \text{SDRF}) = 0.28$



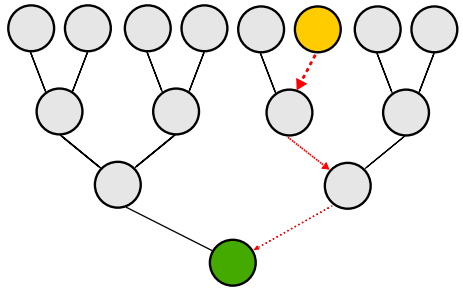
(b) Actor:

$W_1(\text{Original}, \text{DIGL}) = 831.88$

$W_1(\text{Original}, \text{SDRF}) = 0.28$

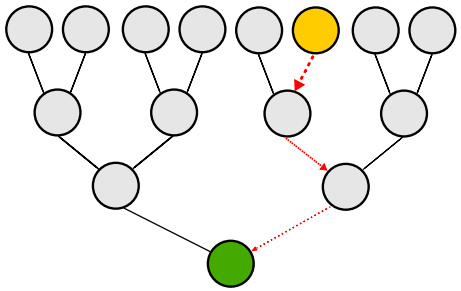
Wrap-up

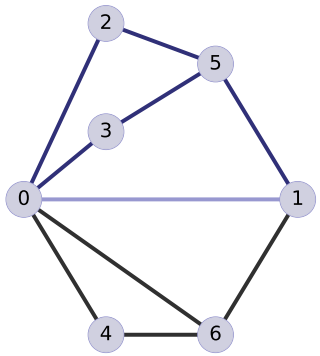
- Over-squashing, induced by topology, can be a problem for MPNNs



Wrap-up

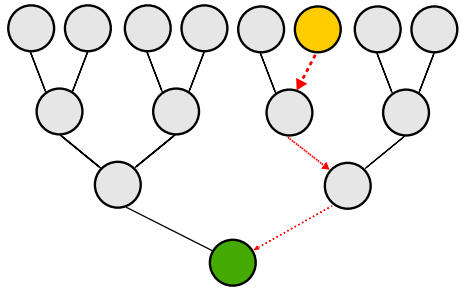
- Over-squashing, induced by topology, can be a problem for MPNNs
- We propose a discrete notion of curvature called Balanced Forman curvature that makes sense on standard graphs and is bounded above by Ollivier, giving us that
  1. Positive Ric means a polynomially-growing receptive field (tackling over-squashing),
  2. and negatively curved edges are those responsible for bottlenecks

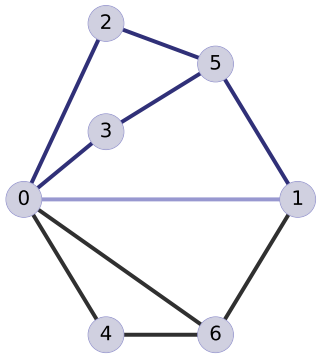


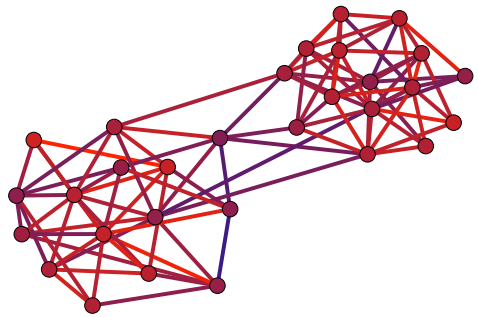


Wrap-up

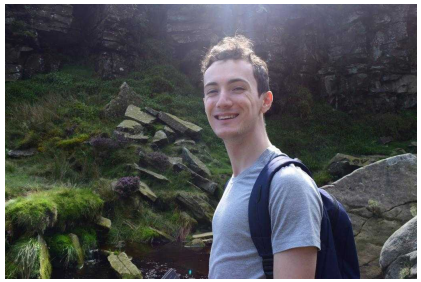
- Over-squashing, induced by topology, can be a problem for MPNNs
- We propose a discrete notion of curvature called Balanced Forman curvature that makes sense on standard graphs and is bounded above by Ollivier, giving us that
  1. Positive Ric means a polynomially-growing receptive field (tackling over-squashing),
  2. and negatively curved edges are those responsible for bottlenecks
- Curvature-based rewiring methods such as SDRF are valid methods for improving GNN performance, improving on the baselines with controlled changes to the graph topology







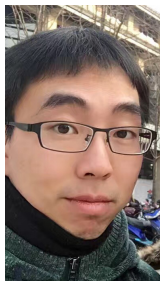
Thank you!



Jake Topping*
@jctopping

Francesco Di Giovanni*
@Francesco_dgv

Benjamin P. Chamberlain
@b_p_chamberlain



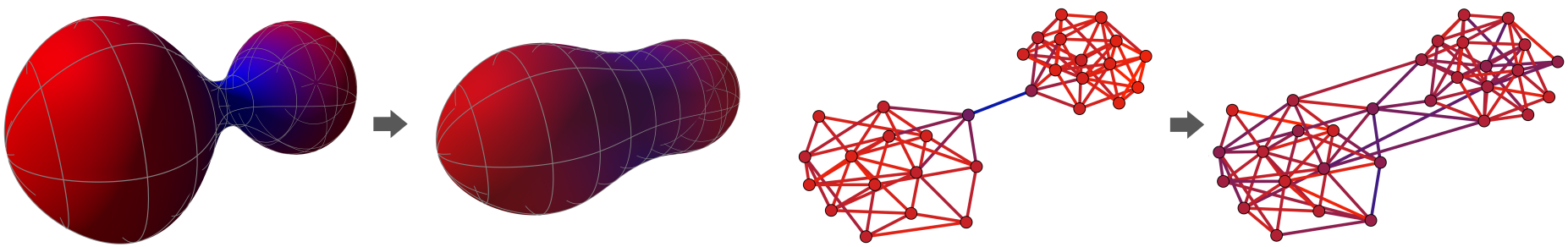
Xiaowen Dong
@epomqo

Michael M. Bronstein
@mmbronstein

Understanding over-squashing and bottlenecks on graphs via curvature

Jake Topping[13†] Francesco Di Giovanni[2†] Benjamin P. Chamberlain[2]
Xiaowen Dong[1] Michael M. Bronstein[12]

[1]University of Oxford [2]Twitter [3]Imperial College London
[†]Equal contribution



Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pages 5453–5462. PMLR, 2018.

Robin Forman. Discrete and computational geometry, 2003.

Yann Ollivier. Ricci curvature of metric spaces. *Comptes Rendus Mathematique*, 345(11):643–646, 2007.

Seong-Hun Paeng. Volume and diameter of a graph and ollivier's ricci curvature. *European Journal of Combinatorics*, 33(8):1808–1819, 2012.

Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in analysis*, pages 195–200. Princeton University Press, 2015.

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

Yong Lin, Linyuan Lu, and Shing-Tung Yau. Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627, 2011.

Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. 2019.