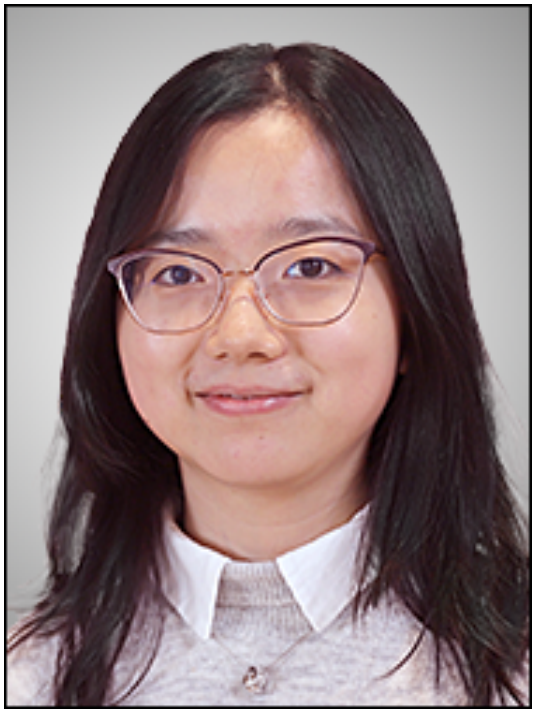
Community Detection via Curvature Gaps

Zachary Lubberts

IMSI Recent Advances in Random Networks Workshop
January 15, 2026

Collaborators

Linda Li
University of Virginia

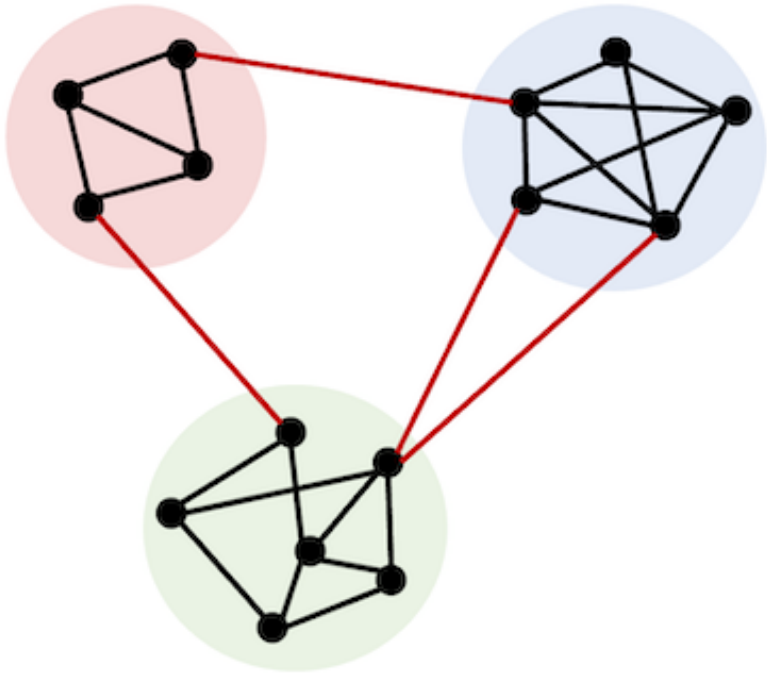


Yu Tian
Center for Systems Biology Dresden

Melanie Weber
Harvard University

Overview

Graph Clustering



- Graph clustering is a common inference problem on networks, to assign community labels to the vertices in a network
- Typically, we assume that vertices in a cluster are more likely to connect to one another than to other vertices

Local vs Global Information

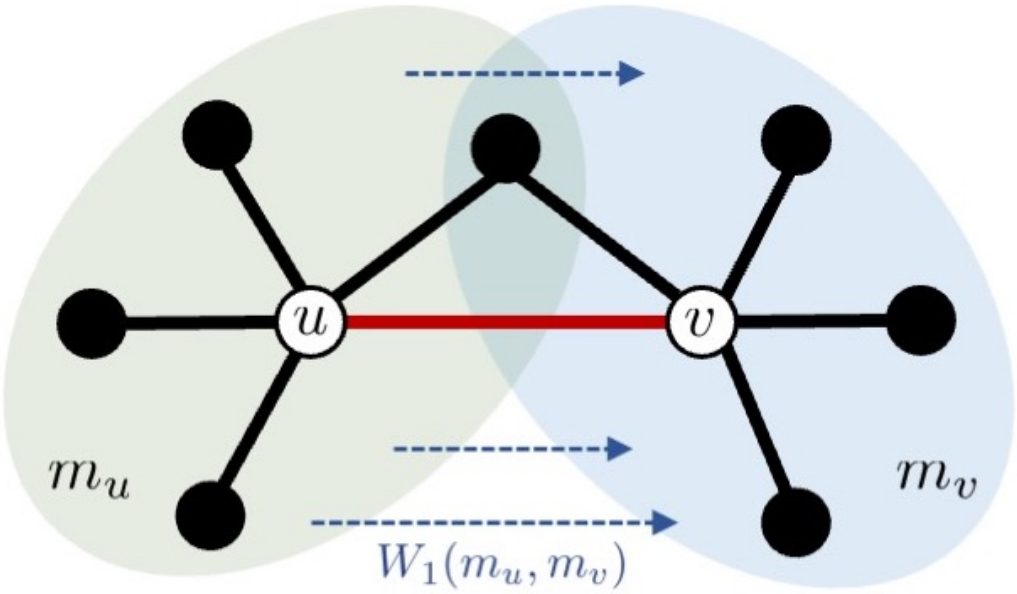
All of the popular methods for network inference make use of *global* information about the network:

- Spectral methods embed the vertices using the eigenvectors of the adjacency matrix
- Graph neural networks construct node embeddings through compositions of nonlinear transformations, but as the number of layers increase, the node embeddings quickly depend on large portions of the graph.

Today I'll talk about an alternative approach to network understanding, using *graph curvatures*. They're based on discrete analogues of curvature on manifolds: hence, they provide very local information.

Ollivier's Ricci Curvature

Just as in a continuous setting, Ollivier's Ricci Curvature, ORC, measures the cost of transporting a neighborhood of u to a neighborhood of v, relative to the distance between u and v.



Lazy random walk

To define this formally, we need a probability measure starting from a vertex u.

$$m_u^\alpha(v) := \begin{cases} \alpha & \text{if } v = u \\ \frac{1-\alpha}{C_u} \exp(-w_{u,v}) & \text{if } \{u, v\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

Here $w_{u,v}$ is the weight of the edge $\{u, v\}$, and C_u is a normalizing constant.

Optimal transport

To compute ORC, we compare the transportation cost of m_u to m_v compared to the length of the $\{u, v\}$ edge:

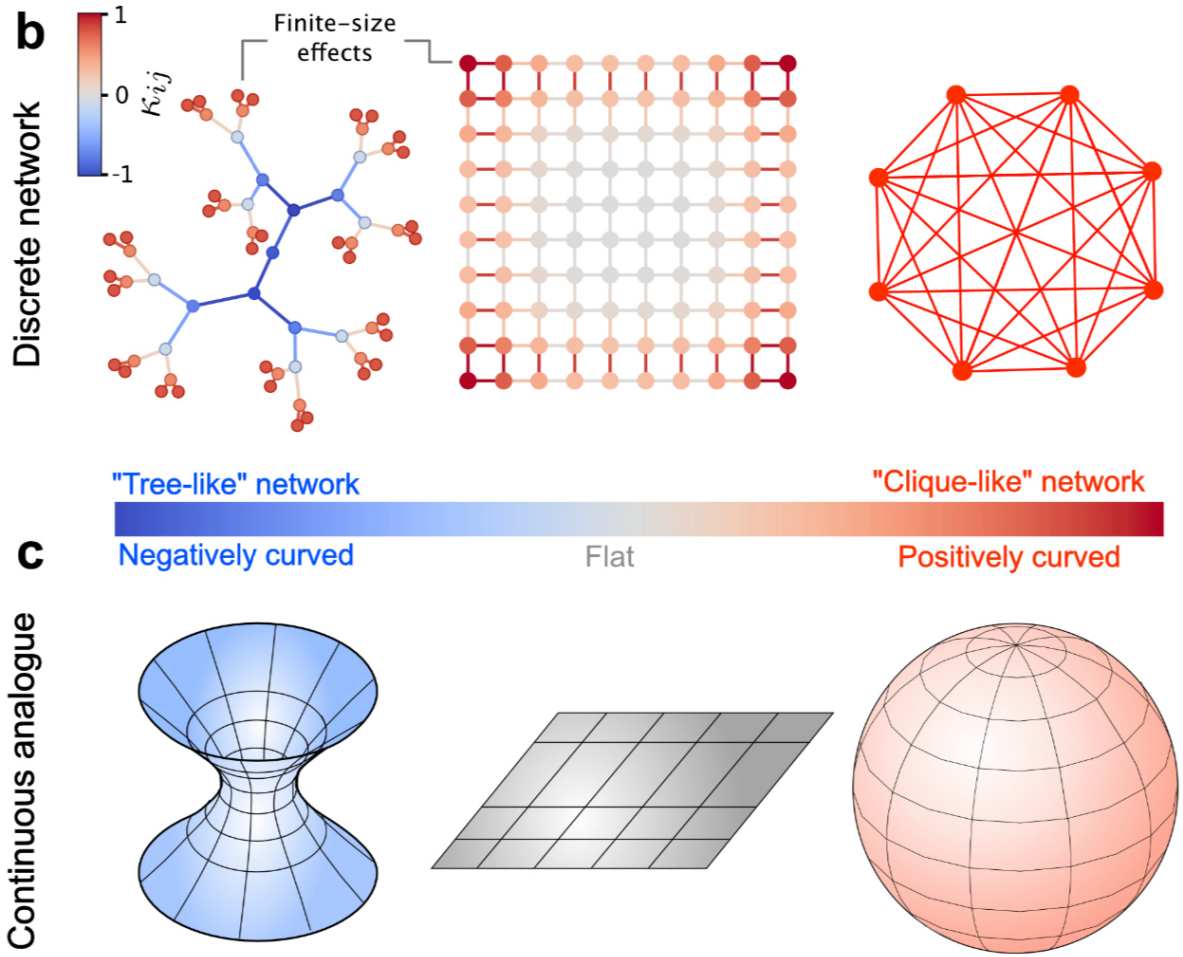
$$\kappa_{uv} := 1 - \frac{W_1(m_u, m_v)}{w_{u,v}}.$$

The Wasserstein-1 distance is defined as

$$W_1(m_u, m_v) = \min_{\pi \in \mathcal{C}(m_u, m_v)} \sum_{\substack{x \in \mathcal{N}(u) \cup \{u\} \\ y \in \mathcal{N}(v) \cup \{v\}}} \pi(x, y) d_G(x, y).$$

π is a *coupling*, a probability distribution with marginals m_u, m_v, and we use shortest path distance d_G on the weighted graph.

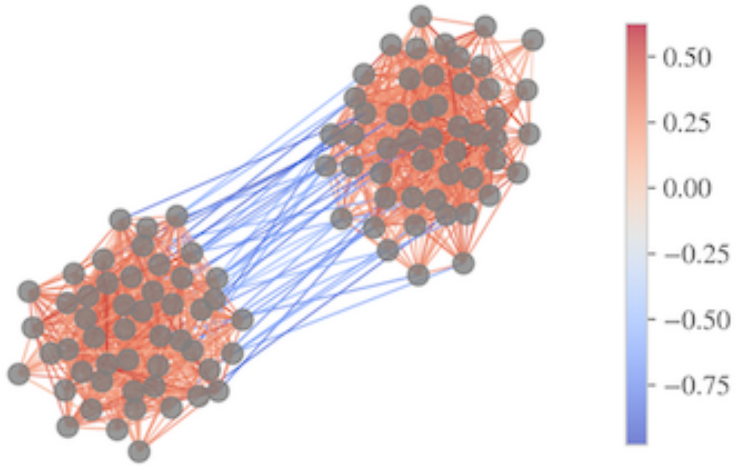
Intuition behind graph curvature



(Figure from Gosztolai & Arnaudon, 2021)

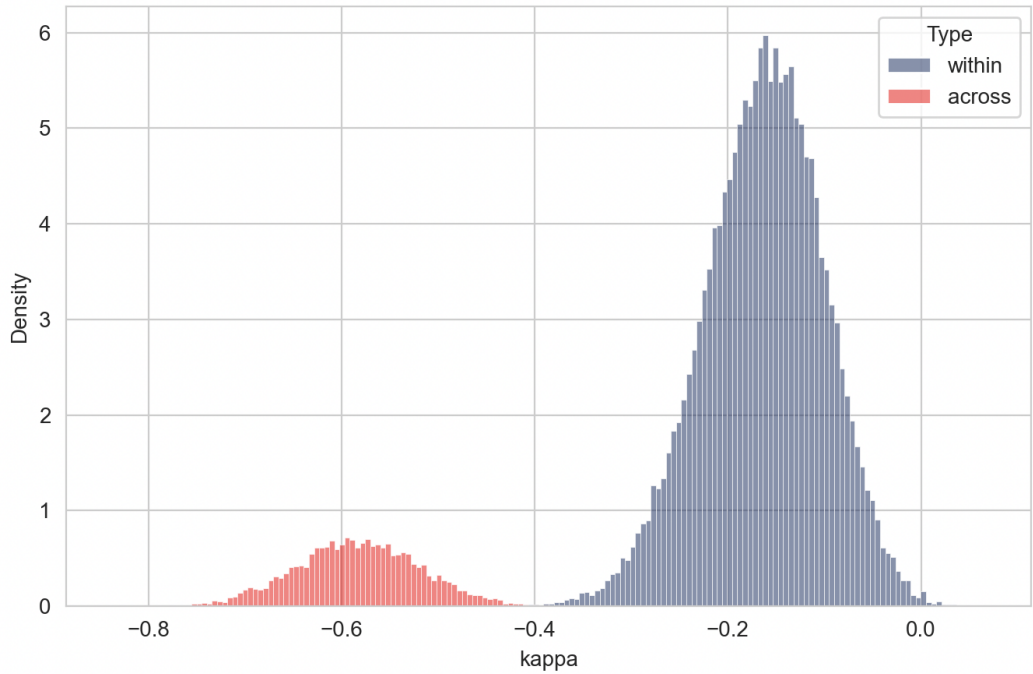
Identifying bridges

- Vertices within a community tend to have a cheap transportation cost relative to their distance (positively curved edges)
- Vertices in different communities tend to have a large transportation cost relative to their distance (negatively curved edges)
- Identifying bridge edges lets us separate communities!



Curvature Gap

We can easily identify bridge edges when there is a curvature gap.



Characterizing these curvature gaps requires understanding the distribution of ORC in large graphs!

Results

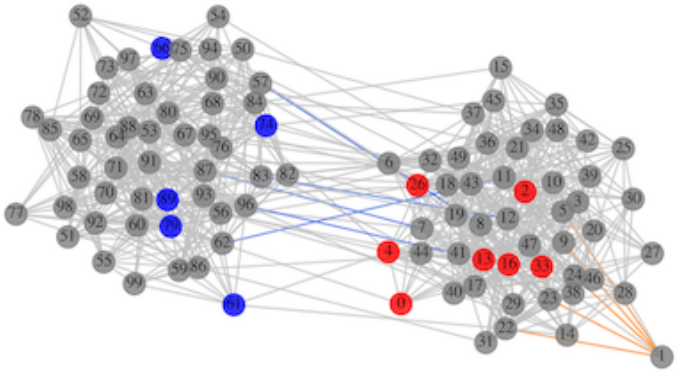
- A general, curvature-based algorithm, Ricci Flow (RF), which reinforces community structure and improves community recovery
- Fast, accurate approximations to Ollivier's Ricci Curvature (ORC)
- Experimental illustration of curvature gap in two block SBMs
- Theoretical distributional results for Ollivier's Ricci Curvature (ORC) under strong signal, showing a curvature gap

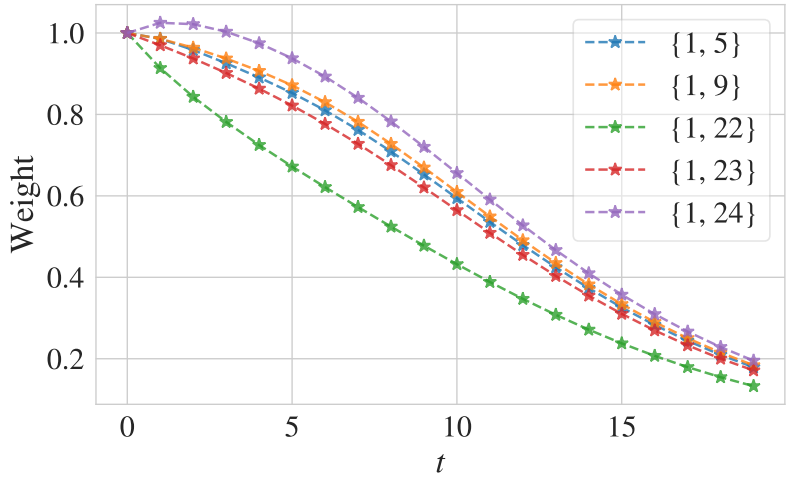
Ricci Flow

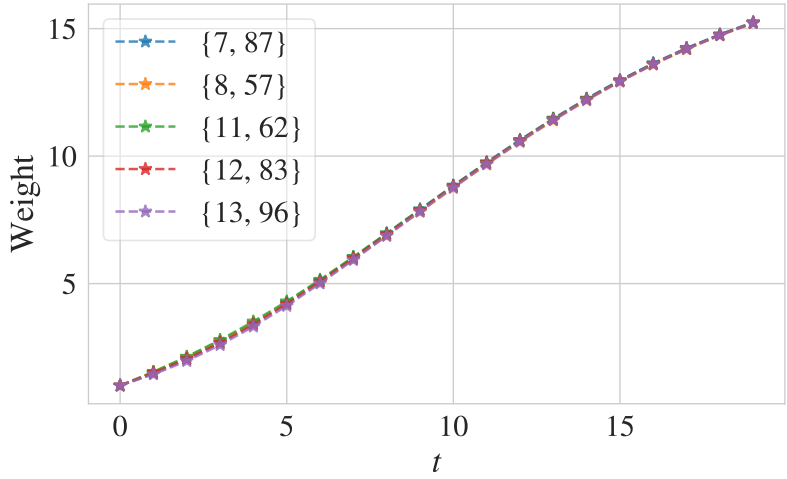
In Ricci flow, we evolve the edge weights via

$$w^{t+1}(\{u,v\}) \leftarrow (1-\kappa^t(\{u,v\}))w^t(\{u,v\})$$

This further emphasizes the communities.

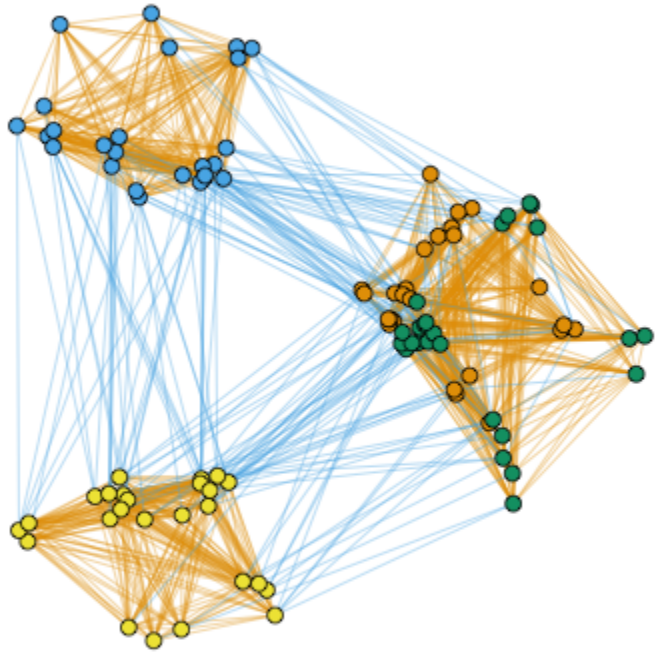






Ricci Flow

In a stochastic blockmodel random graph:



Correctness of Ricci Flow

We see that Ricci Flow *coarsens* graphs by

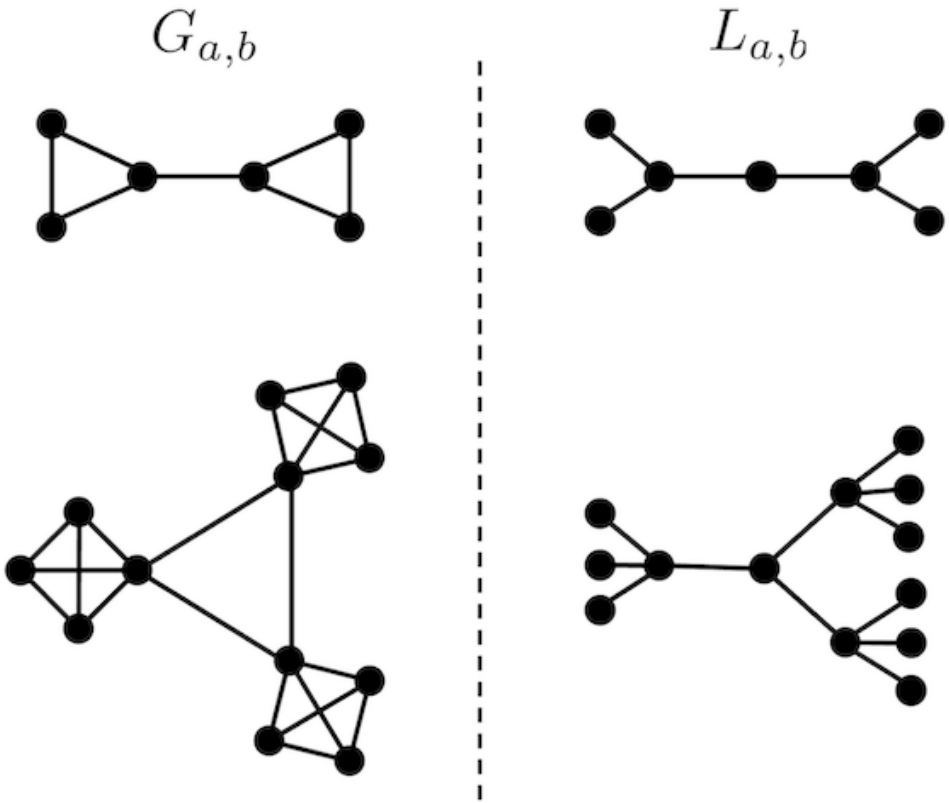
- Pulling vertices within a community closer together, and
- Pushing vertices from distinct communities apart.

So if we apply Ricci Flow for some amount of time, then cut the longest edges, this should recover the true graph community labels.

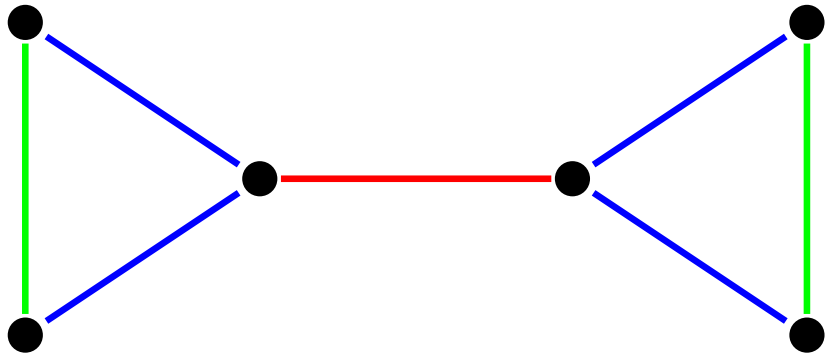
While we have simulation evidence for this, it was not known whether this idea actually works!

Correctness of Ricci Flow

We consider a class of networks comprised of b copies of K_a, $a > b \geq 2$, each with a specified bridge node. These bridge nodes are then connected to form a K_b. We call these graphs $G_{a,b}$, and choose $L_{a,b}$ so that its line graph is $G_{a,b}$.



Three Types of Edges



There are three edge types that need to be analyzed in these graphs:

1. Bridge edges, which connect bridge nodes;
2. Internal edges that connect a bridge node to an internal node; and
3. Internal edges between internal nodes.

Correctness of Ricci Flow

To prove the correctness result, we find the exact ORC values for each type of edge at each iteration, and prove that

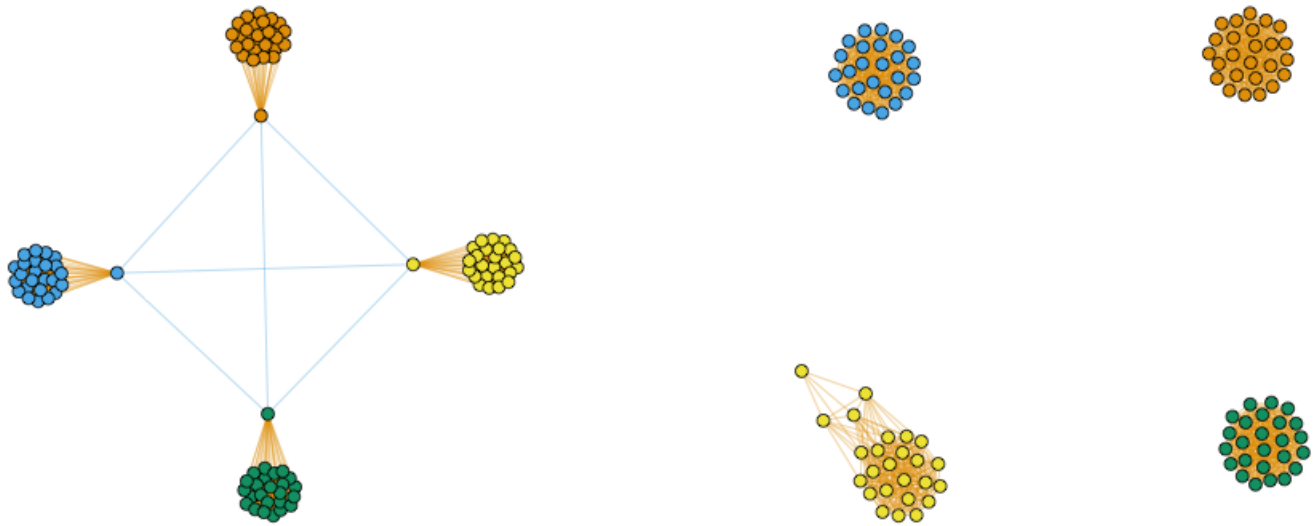
- Type 1 edge weights are strictly larger than Type 2/3 edge weights at every iteration;
- Type 3 edge weights go to 0.

Thus, there is a choice of threshold that separates all of the bridge edges from all of the internal edges.

Since $G_{a,b}$ is just $L(L_{a,b})$, this also proves correctness of the mixed-membership version of the algorithm for a class of graphs with a single overlapping node.

Correctness of Ricci Flow

Before and after 10 steps of Ricci Flow, and cutting the edges with the top 5% of weights:



Correctness of Ricci Flow

- Before this theorem, there was only simulation evidence for correctness
- A curvature gap makes it much more likely that Ricci Flow will succeed!
- This motivates learning the distribution of ORCs in large, random networks.

Challenges of getting distributional results for ORC

- ORC is calculated by solving an optimal transport problem
- Having no closed form of ORC prevents us from getting exact distributional results for finite n
- Local cycle counts for approximating ORC can be dependent in complicated ways
- So distributional results require tight approximation accuracy while preserving distributional tractability.

We divide the parameter space into three signal strength regimes and derive the distributional results using different approximations of ORC within each of them.

Erdös-Rényi graph and two block SBM

- Erdös-Rényi graph:
  Let $G \sim ER(n, p_n)$ be a Erdos Renyi graph. We suppose $p_n = c\left(\frac{\log n}{n}\right)^{\lambda}, \lambda \in [0,1], c \in \mathbb{R}^+$.

- Two block Stochastic Block Model:
  Let $G \sim SSBM(n, p_n, q_n)$ be a two block model with equal size $n/2$, with block-connection probability matrix $\begin{bmatrix} p_n & q_n \\ q_n & p_n \end{bmatrix}$. We suppose
  $p_n = a\left(\frac{\log n}{n}\right)^{\lambda}, q_n = b\left(\frac{\log n}{n}\right)^{\lambda}, \lambda \in [0,1], a, b \in \mathbb{R}^+$.

Signal strength

We consider the following three signal strength regimes:

- The strong signal regime: $\lambda \in [0, 1/2]$. Triangles and 4-cycles are the most important structures for ORC in this case.
- The moderate signal regime: $\lambda \in (1/2, 2/3]$. With high probability, there are no triangles supported on (u, v), so 4- and 5-cycles are the most important structures determining ORC.
- The weak signal regime: $\lambda \in (2/3, 1]$. Now there are very few 4-cycles, so the ORC is determined by the balance of 5- and 6-cycles supported on the edge.

The case with $\lambda = 1$ is the information-theoretic limit for exact community recovery.

Signal strength

From prior work by Bhattacharya and Mukherjee (2015), we have the following convergences in probability for ER graphs:

- If $p_n = p$ is constant: $\kappa_n(u, v) \to p$
- If $p_n \to 0, np_n^2 \to \infty$: $\kappa_n(u, v) \to 0$.
- If $np_n^2 \to 0$, $n^2 p_n^3 \to \infty$: $\kappa_n(u, v) \to -1$.
- If $n^2 p_n^3 \to 0$, $np_n \to \infty$: $\kappa_n(u, v) \to -2$.

Our cases $\lambda = 0, \lambda \in (0, 1/2], \lambda \in (1/2, 2/3], \lambda \in (2/3, 1]$ correspond to these results.

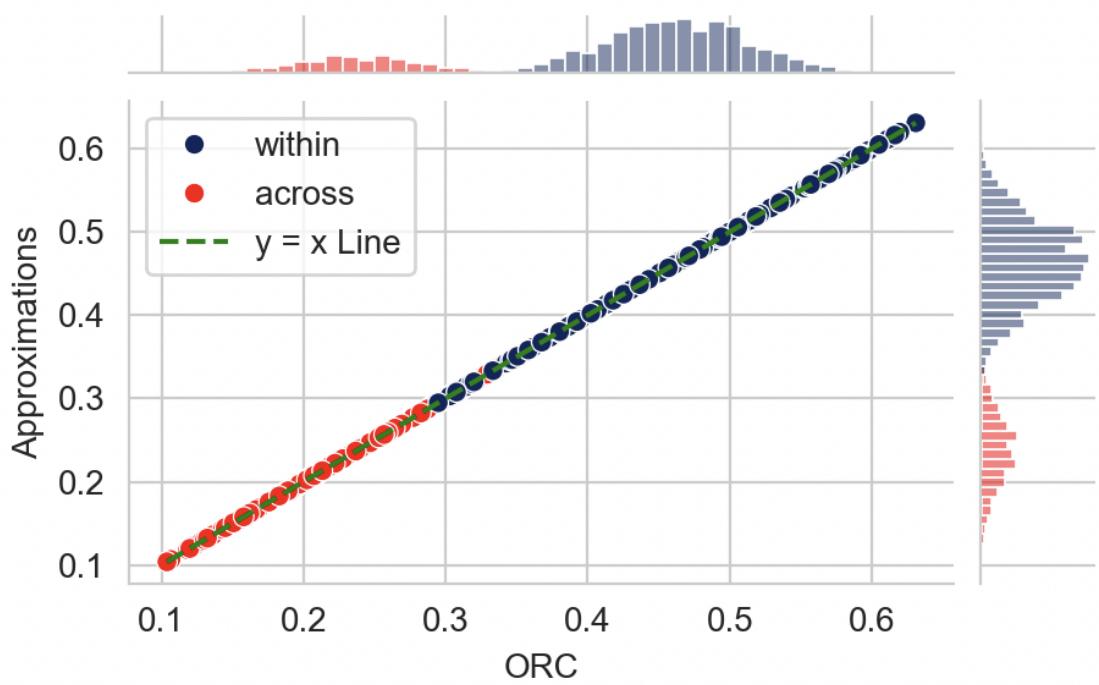
Approximation under strong signal

For $\lambda \in [0, 1/2]$, we approximate ORC via

$$\mathrm{ORC} \approx \frac{\triangle_{uv}}{d_u \vee d_v}.$$

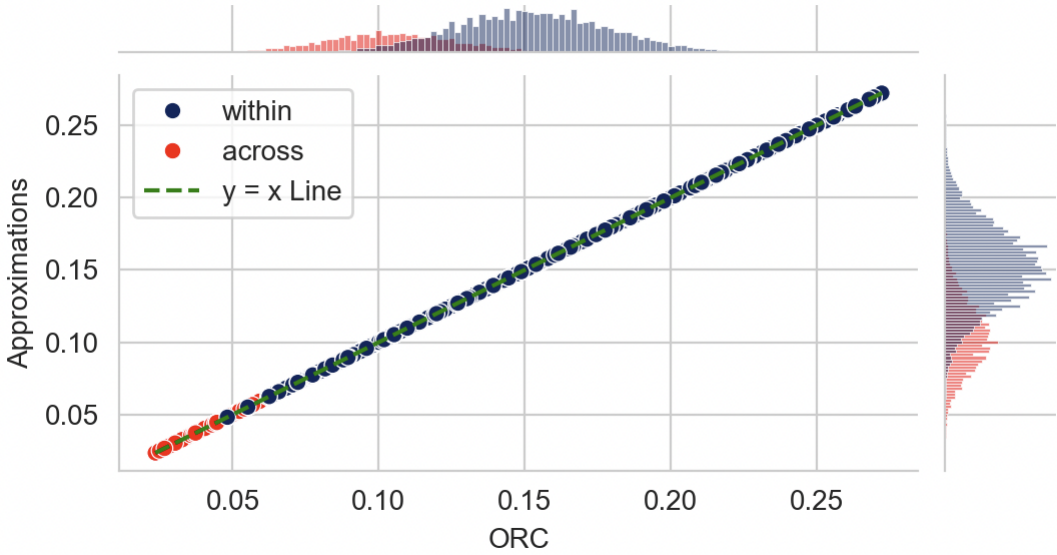
- $\triangle_{uv}$ is the number of common neighbors of u, v
- d_u, d_v are the degrees of u and v
- This is a combinatorial formula, whereas ORC is the solution to an optimization problem, so it's much easier to work with.

Curvature Gap under strong signal



- $n_1 = n_2 = 100,$
- $p = 3.6\sqrt{\frac{\log 200}{200}} = 0.5859,$
- $q = 1\sqrt{\frac{\log 200}{200}} = 0.1628.$

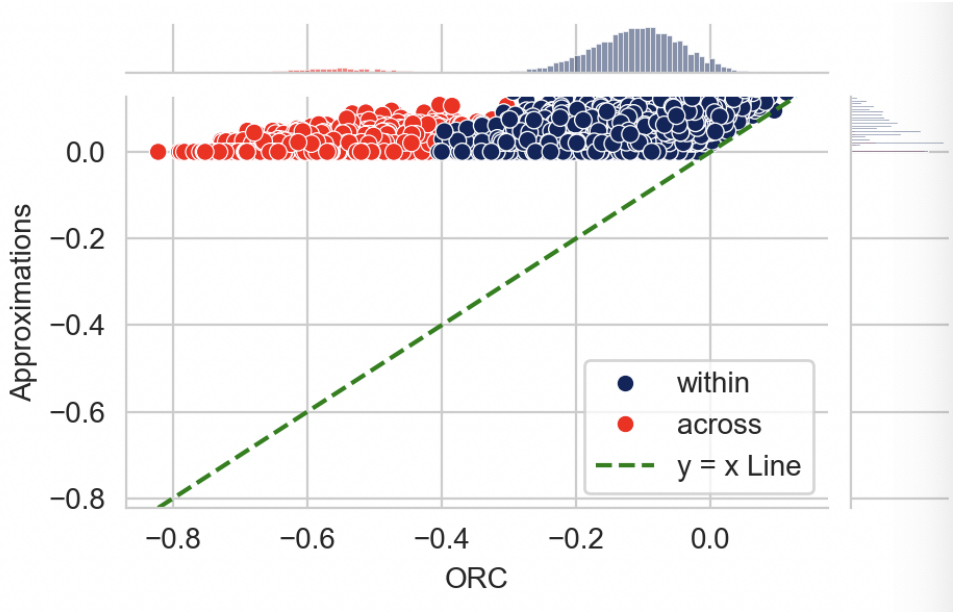
Curvature Gap under strong signal



- $n_1 = n_2 = 600,$
- $p = 2.5\sqrt{\frac{\log 1200}{1200}} = 0.1922,$
- $q = 1\sqrt{\frac{\log 1200}{1200}} = 0.0769.$

Approximation fails under moderate signal

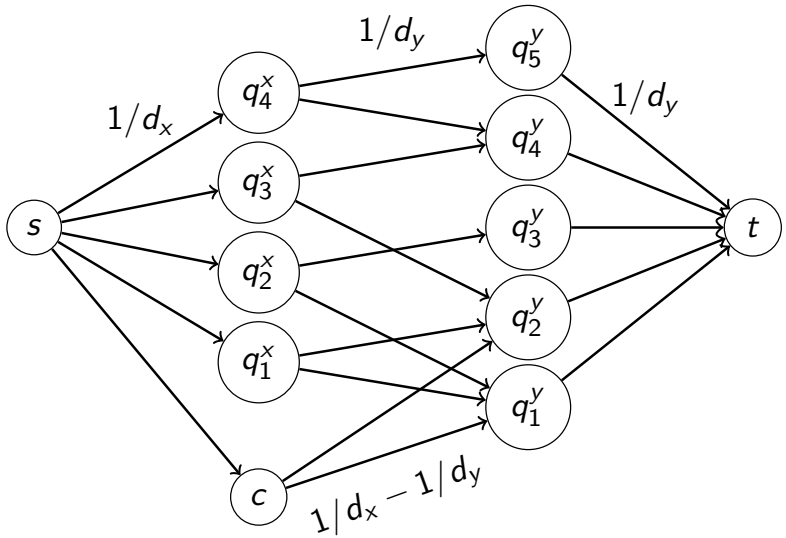
This approximation completely fails in the moderate signal strength regime!



- $n_1 = n_2 = 500$,
- $p = 2\left(\frac{\log 1000}{1000}\right)^{2/3} = 0.0725$,
- $q = 0.25\left(\frac{\log 1000}{1000}\right)^{2/3} = 0.0091$.

Approximation improved by maximum flow

In the strong signal strength setting, triangles are the dominant structure determining ORC. In the moderate signal strength, however, we need to quantify the proportion of mass that can be transported on 4-cycles. We accomplish this by solving a max-flow problem:



c denotes a common vertex, q_i^x / q_j^y are neighbors of only one vertex. We introduce s and t to define the max-flow problem.

Approximation under moderate signal

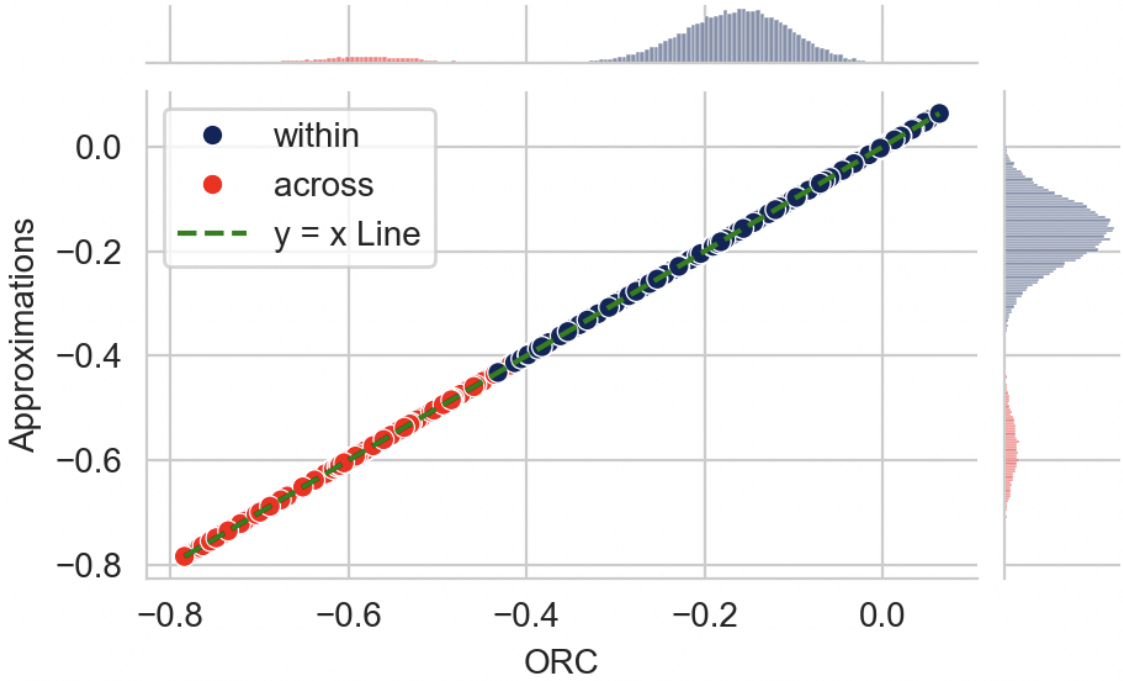
This lets us approximate ORC via

$$\kappa(u,v) \approx 1 - f_{uv} - 2\left(1 - f_{uv} - \frac{\triangle_{uv}}{d_u \vee d_v}\right)$$

$$= -1 + f_{uv} + 2\frac{\triangle_{uv}}{d_u \vee d_v}$$

- The first line separates the mass moving distance 1 (the max flow value) from the mass moving distance 2
- The second line shows that $\kappa(u,v) \geq -1$, and is converging to -1 as $n \to \infty$.

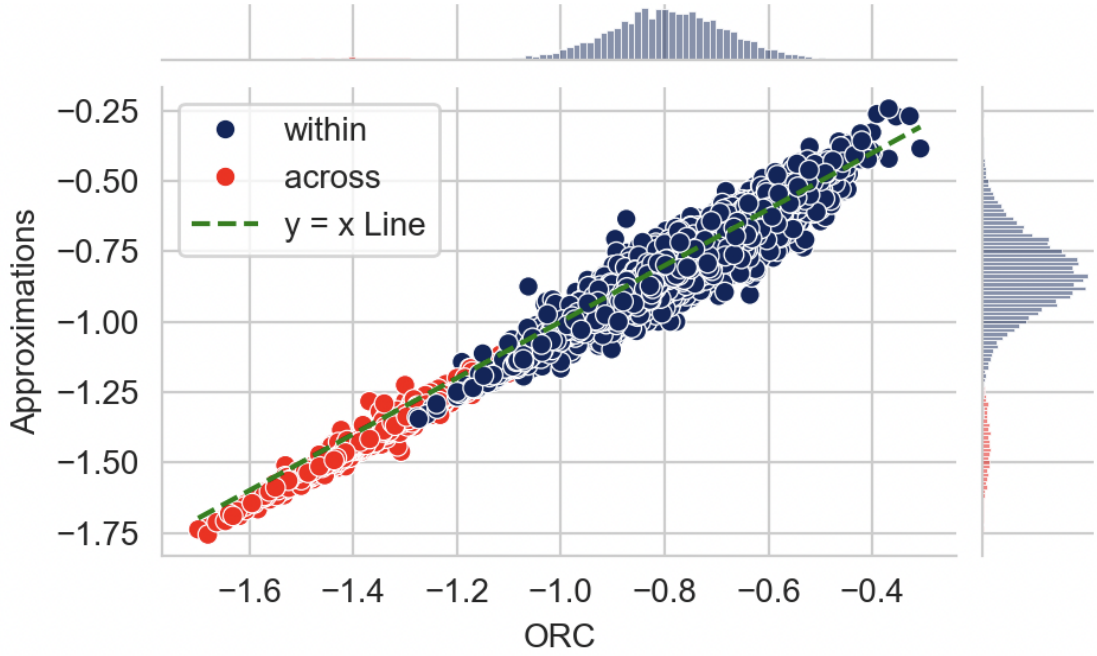
Curvature Gap under moderate signal



- $n_1 = n_2 = 1000$,
- $p = 2\left(\frac{\log 2000}{2000}\right)^{2/3} = 0.0487$,
- $q = 0.25\left(\frac{\log 2000}{2000}\right)^{2/3} = 0.0061$.

Curvature Gap under low signal

Approximation becomes much harder in the low signal regime!



(After further refinement beyond 4-cycles)

- $n_1 = n_2 = 1000$,
- $p = 4\,\frac{\log 2000}{2000} = 0.0152$
- $q = 0.25\,\frac{\log 2000}{2000} = 0.0009$

Bounds in ER graph

Theorem (Jost and Liu (2014))

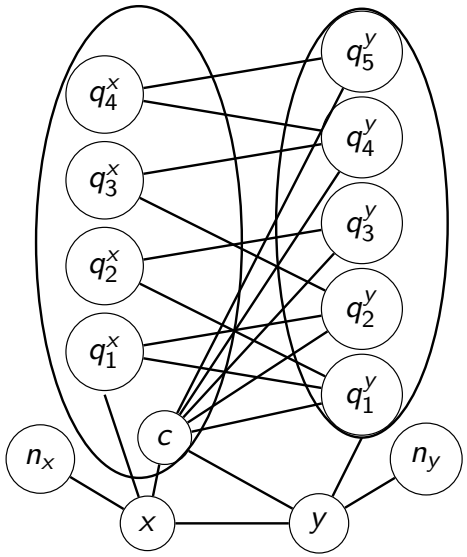
$$\frac{\triangle_{uv}}{d_u \vee d_v} \geq \kappa(u,v) \geq \frac{\triangle_{uv}}{d_u \vee d_v} - E(u,v)$$

$$E(u,v) = \left(1 - \frac{1}{d_u} - \frac{1}{d_v} - \frac{\triangle_{uv}}{d_u \vee d_v}\right)_+ + \left(1 - \frac{1}{d_u} - \frac{1}{d_v} - \frac{\triangle_{uv}}{d_u \wedge d_v}\right)_+$$

where $\triangle_{uv}$ denotes the shared neighbors of u and v.
$\triangle_{uv} = \{x | x \in N(u), x \in N(v)\}$

Distributional result for ORC in ER

Our max flow argument lets us bound the error term $E(u, v)$ in the strong signal case, showing that the upper bound is a tight approximation.



Distributional result for the upper bound in ER

Remark

$(|N_u \setminus \triangle_{uv}|, |\triangle_{uv}|, |N_v \setminus \triangle_{uv}|, n-2-|N_u \cup N_v|)$ follows a Multinomial distribution. For large n, the first three entries approximately follow a joint Normal distribution.

Theorem (ER, Strong signal)

When $p_n \equiv p$, as $n \to \infty$,

$$\sqrt{n}\left(\frac{\triangle_{uv}}{d_u \vee d_v} - p\right) \xrightarrow{\text{dist}} \mathcal{D}(\rho_{12}, \rho_{13}, \rho_{23}),$$

where $\rho_{12}, \rho_{13}, \rho_{23}$ are the correlations between pairs of $|N_u \setminus \triangle_{uv}|, |\triangle_{uv}|, |N_v \setminus \triangle_{uv}|$.

When $p_n \to 0$, $np_n^2 \to \infty$, then this quantity converges in distribution to $\mathcal{N}(0,1)$.

Distributional result for the upper bound in ER

The characteristic function of $\mathcal{D}(\rho_1, \rho_2, \rho_3)$ is

$$\varphi_Z(t) = \exp\left(-\frac{1-p}{2}t^2\right)\left[1 + i\mathrm{Erfi}\left(-\frac{\sqrt{p(1-p)}}{2}t\right)\right],$$

and its density is

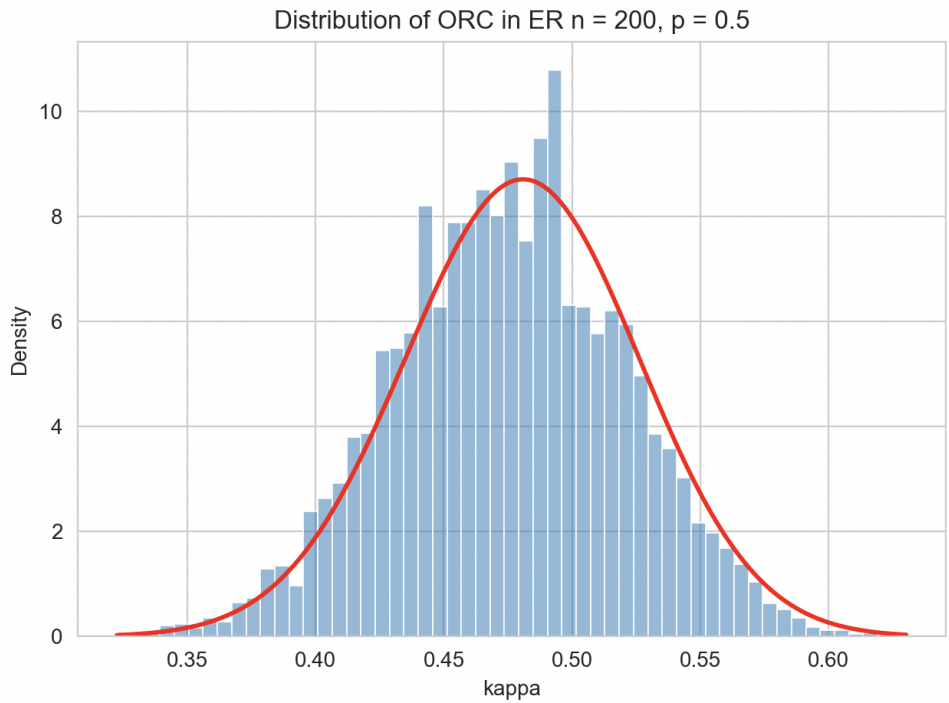
$$f_Z(t) = \frac{1}{\sqrt{2\pi(1-p)}}\exp\left(\frac{-t^2}{2(1-p)}\right) \times \left[1 + \mathrm{Erf}\left(-\frac{\sqrt{p}t}{\sqrt{2(1-p)(2-p)}}\right)\right].$$

The mean and variance of this distribution are given by

$$\mathbb{E}[Z] = -\sqrt{\frac{p(1-p)}{\pi}}, \quad \mathrm{Var}(Z) = \frac{(\pi - p)(1-p)}{\pi}.$$

Experimental results

Histogram of ORCs for an ER(1/2) with large n, and superimposed density of $\mathcal{D}$:



Curvature gap, strong signal

We find the limiting distributions, means, and variances for the ORCs of edges within and between communities:

- $\mu_{\text{within}} = \frac{p_n^2+q_n^2}{p_n+q_n} + o\left(\frac{1}{\sqrt{n}}\right)$,
- $\sigma^2_{\text{within}} = \frac{(p_n^2+q_n^2)\left[(p_n+q_n)\pi-(p_n^2+q_n^2)\right]}{n\pi(p_n+q_n)^4}\left[p_n(1-p_n)+q_n(1-q_n)\right]+o\left(\frac{1}{n}\right)$.
- $\mu_{\text{between}} = \frac{2p_nq_n}{p_n+q_n}$,
- $\sigma^2_{\text{between}} = \frac{2p_nq_n[(p_n+q_n)\pi-2p_nq_n]}{n\pi(p_n+q_n)^4}\left[p_n(1-p_n)+q_n(1-q_n)\right]+o\left(\frac{1}{n}\right)$.

Curvature gap, strong signal

When the difference in means is much larger than the standard deviations, we have a curvature gap:

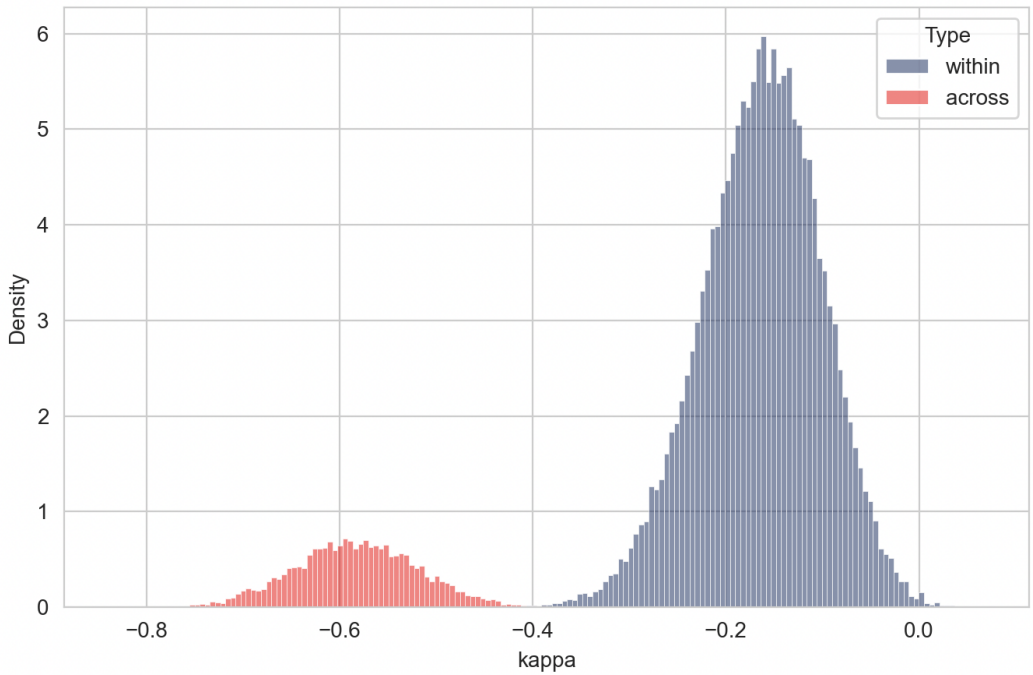
$$\frac{p_n^2 + q_n^2}{p_n + q_n} - \log^{1/2-\epsilon}(n)\sigma_{\text{within}} > \frac{2p_n q_n}{p_n + q_n} + \log^{1/2-\epsilon}(n)\sigma_{\text{between}}$$

$$\Leftrightarrow \frac{(a-b)^2}{\sqrt{a^2+b^2}+\sqrt{2ab}} > \left(\frac{\log(n)}{n}\right)^{1/2-\lambda} \frac{1}{\log^{\epsilon}(n)}$$

So for large n, there is always a curvature gap, and thus we can identify clusters with curvature!

Curvature gap persists for weaker signal

We clearly see the gap in the moderate (shown) and weak signal strengths, too:



Summary

- We have been able to compute the limiting distributions of the ORC for within-block and between-block edges in an SBM under strong signal strength
- We have tight approximations that should lead to similar results in the moderate signal strength case
- We have experimentally demonstrated curvature gaps in SBMs under all three signal strength regimes
- With additional work, we believe it can be shown that graph curvatures identify clusters down to the information-theoretic limit!

Thank you!

Ricci flow work: Y. Tian*, ZL*, M. Weber, "Curvature-based clustering on graphs." *JMLR* 26 (2025)
or arXiv: 2307.10155