Community Recovery in random geometric graphs

Hariprasad M
Chanakya University

March 3, 2026

Outline

- Introduction and background: Random geometric graphs and community recovery.
- Adjacency limiting spectra of SGBM (soft geometric block model).
- Adjacency eigenvectors and community recovery.
- Community recovery in 1D-GBM using Spectral seriation.

Communities in Graphs

Communities are groups of nodes with dense internal connections and sparse external links.

- Capture underlying structure of networks
- Reveal functional or social groupings
- Help simplify and interpret complex graphs

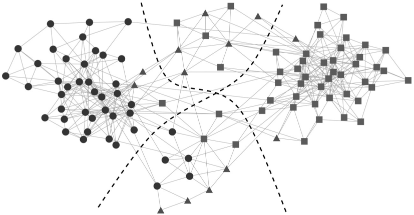

Figure 1: Books are nodes; co-purchase is denoted by edges, four communities: circles are liberal, squares are conservative, and triangles are centrist or unaligned.

Stochastic Block Model (SBM)

Nodes are assigned to one of the k communities, and edges are placed independently with probabilities depending only on the community pair.

Figure 2: SBM with two communities

Spectral Clustering via Fiedler Vector Sign

For a 2-way partition, use the **Fiedler eigenvector** v_2, the eigenvector of the graph Laplacian L corresponding to the **second-smallest eigenvalue**.

Key Idea: The sign of each entry in v_2 reveals the natural bipartition of the graph.

$$\text{Cluster 1: } \{i : v_2(i) \geq 0\}, \qquad \text{Cluster 2: } \{i : v_2(i) < 0\}$$

Why this works:

- v_2 minimizes a relaxed version of the *RatioCut* / *NCut* objective.
- Sign structure corresponds to the sparsest balanced cut indicated by Laplacian geometry.
- No need for k-means when $k = 2$; the sign split is optimal in the relaxation.

Result: A simple and robust 2-clustering: *just threshold at zero.*

Random Geometric Graphs

- Nodes placed randomly in a metric space.
- An edge exists if the distance is less than r.

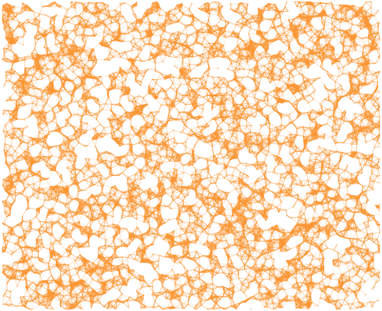

Figure 3: A Random geometric graph

.

- Modeling connectivity between sensors distributed in a region.
- Multi-agent systems and autonomous swarms: visibility and communication graphs for robots.

Random geometric graphs in one dimension and interval graphs

Consider n intervals on the real line, $I_j = [a_j, b_j] \subset \mathbb{R}$. These intervals corresponds to nodes $V = \{I_j\}$. When two intervals intersect $I_j \cap I_k \neq \Phi$, we have an edge between those nodes i.e. $(v_i, v_k) \in E$. Graphs generated in this way are known as interval graphs. When all the intervals considered are of the same size, the graph is known as a unit interval graph.

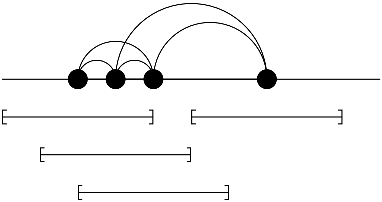

Figure 4: Geometric graph and its interval correspondence

Geometric Block Model (GBM) — Definition

The GBM extends random geometric graphs by incorporating community structure.

- Nodes placed in a metric space; edges drawn based on spatial distance
- Within-community radius: r_{in}; cross-community radius: r_{out} with $r_{out} < r_{in}$
- Captures interplay of geometry and communities in spatial networks

[1] Galhotra, Sainyam, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. *"The Geometric Block Model."* In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.

[2] Galhotra, Sainyam, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. *"Community recovery in the geometric block model."* Journal of Machine Learning Research 24, no. 338 (2023): 1-53.

GBM as interval graph

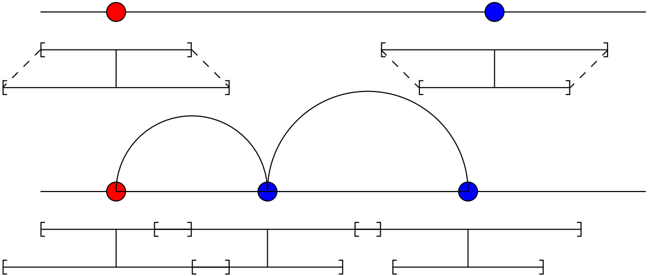

Figure 5: Nodes of different communities correspond to different interval sets. intersection of these interval sets is used in representing geometric graph with two communities as an intersection graph. Equivalently these nodes can be identified with two types of trapezoids, so these graphs are special cases of trapezoidal graphs.

Adjacency of a gemetric block model on a circle

- Sparse: $r = O(1/n)$.
- Logarithmic: $r = O(\log n/n)$.
- Dense: $r = c$ (constant).
- GBM in log regime: $r_{in} = a \log n/n$, $r_{out} = b \log n/n$.

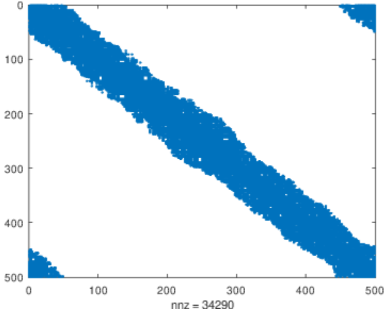

Figure 6: A GBM on a circle with 1000 nodes

Task : Given the adjacency matrix, recover the communities.

Real world GBMs

Dataset	Total no. of nodes	T_1	T_2	T_3	Accuracy		Running Time (sec)	
					Motif-Counting	Spectral clustering	Motif-Counting	Spectral clustering
Political Blogs	1222	20	2	1	**0.788**	0.53	1.62	**0.29**
DBLP	12138	10	1	2	**0.675**	0.63	**3.93**	18.077
LiveJournal	2366	20	1	1	**0.7768**	0.64	**0.49**	1.54

Table 3: Performance on real world networks

Figure 7: Performance of various algorithms on real world networks.

The Core Question

Can spectral clustering recover communities in geometric random graphs?

In SBM:

- Community eigenvalues are top eigenvalues

In geometric models:

- Geometry injects additional spectral structure
- Leading eigenvectors reflect spatial modes

Key phenomenon: Community signal lives at a specific frequency.

Traditional Spectral clustering fails on GBM

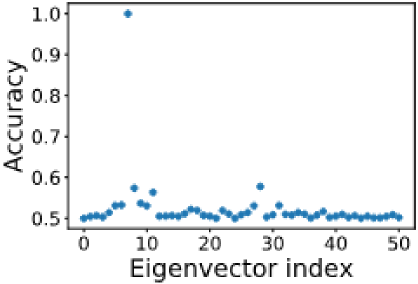

Figure 8: Accuracy obtained on a 1 dimensional GBM ($n = 2000, rin = 0.08, rout = 0.02$) using the different eigenvectors of the adjacency matrix. The eigenvector of index k corresponds to the eigenvector associated with the k th largest eigenvalue of A

[3] Avrachenkov, Konstantin, Andrei Bobu, and Maximilien Dreveton. "Higher-order spectral clustering for geometric graphs." Journal of Fourier Analysis and Applications 27, no. 2 (2021): 22.

Spectral clustering on GBM

Algorithm 1: Higher-Order Spectral Clustering (HOSC).

Input: Adjacency matrix A, average intra- and inter-cluster edge densities $\mu_{\text{in}}, \mu_{\text{out}}$.
Output: Node labelling $\tilde{\sigma} \in \{1, 2\}^n$.

Global step:
Let $\tilde{\lambda}$ be the eigenvalue of A closest to $\lambda_* = \frac{(\mu_{\text{in}} - \mu_{\text{out}})}{2} n$, and $\tilde{v}$ be the associated eigenvector.
for $i = 1, \ldots, n$ **do**
  If $\tilde{v}_i > 0$, let $\tilde{\sigma}_i = 1$; otherwise, let $\tilde{\sigma}_i = 2$.

Figure 9: Spectral clustering Algorithm on GBM

GBM with Multiple communities

- Nodes of multiple communities are placed uniformly randomly in a metric space.
- Connectivity depends on the distance between the nodes and the communities $r_{in}(i,j), r_{out}(i,j)$.
- Task: Given the adjacency matrix, recover the communities.

Allem, Luiz Emilio, Konstantin Avrachenkov, Carlos Hoppen, Hariprasad Manjunath, and Lucas Siviero Sibemberg. "Multi-Community Spectral Clustering for Geometric Graphs." arXiv preprint arXiv:2508.00893 (2025).

Spectral clustering on GBM

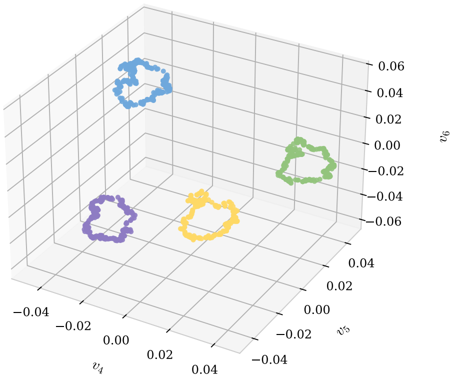

Figure 10: Nodes embedded in eigenvector space

Model: Soft Geometric Block Model (SGBM)

Domain:

$$T^d = \mathbb{R}^d / \mathbb{Z}^d$$

Vertices:

$$\sigma : [n] \to [k], \quad |C_q| = n/k$$

Positions:

$$X_i \overset{i.i.d.}{\sim} \mathrm{Unif}(T^d)$$

Edges:

$$\mathbb{P}(A_{ij} = 1) = F(X_i - X_j, \sigma_i, \sigma_j)$$

Homogeneous case:

$$F = \begin{cases} F_{\mathrm{in}} & \sigma_i = \sigma_j \\ F_{\mathrm{out}} & \sigma_i \neq \sigma_j \end{cases}$$

Edge Densities

Expected intracommunity edge density:

$$\mu_{\text{in}} = \int_{T^d} F_{\text{in}}(x)\, dx$$

Expected intercommunity edge density:

$$\mu_{\text{out}} = \int_{T^d} F_{\text{out}}(x)\, dx$$

As expectations:

- If i, j are in the same community,

$$\mathbb{E}[A_{ij}] = \mu_{\text{in}}.$$

- If i, j are in different communities,

$$\mathbb{E}[A_{ij}] = \mu_{\text{out}}.$$

Expected degree of a vertex:

$$\frac{n}{k}\mu_{\text{in}} + \frac{(k-1)n}{k}\mu_{\text{out}}.$$

Our Approach

- **Step 1: Spectral localization via moment method.**
  - Study empirical spectral measure of A/n.
  - Show convergence to a discrete limiting measure using Fourier expansion.
  - Identify an isolated point $\lambda^* = \frac{n(\mu_{in} - \mu_{out})}{k}$ corresponding to community structure.
- **Step 2: Eigenspace perturbation analysis.**
  - The informative eigenvalue has multiplicity $k-1$.
  - Compare invariant subspaces of A and B_σ.
  - Apply a Davis–Kahan type bound to control $\min_{Q \in O(k-1)} \|VQ - U\|_F$.

Empirical Spectral Measure

Let $\lambda_1, \dots, \lambda_n$ be eigenvalues of A.
Define the empirical spectral measure:

$$\mu_n = \sum_{j=1}^{n} \delta_{\lambda_j/n}.$$

Goal:

$$\mu_n \Rightarrow \mu \quad \text{almost surely.}$$

Strategy:

1. Show $\mathbb{E}\mu_n(P_m) \to \mu(P_m)$ for all m (for $P_m = t^m$).
2. Show $\mu_n(P_m) - \mathbb{E}\mu_n(P_m) \to 0$.

Moment Method (Expectation)

For $P_m(t) = t^m$:

$$\mu_n(P_m) = \frac{1}{n^m} \operatorname{tr}(A^m).$$

Interpretation:

$$\operatorname{tr}(A^m) = \text{number of closed walks of length } m.$$

After conditioning on positions:

$$\mathbb{E}[\operatorname{tr}(A^m)] = \sum_{\alpha} \int \prod F(x_{i_\ell} - x_{i_{\ell+1}}).$$

Convolution Structure

Key identity:

$$\int_{(T^d)^m} \prod F(x_{i_\ell} - x_{i_{\ell+1}}) = (F * \cdots * F)(0).$$

Thus moments reduce to convolution powers.

Fourier diagonalizes convolution:

$$\widehat{F * G}(z) = \hat{F}(z)\hat{G}(z).$$

Moments depend on

$$\hat{F}_{\text{in}}, \hat{F}_{\text{out}}.$$

Limiting Spectral Measure

$$\mu = \sum_{z \in \mathbb{Z}^d} \delta_{\frac{\hat{F}_{\text{in}}(z)+(k-1)\hat{F}_{\text{out}}(z)}{k}} + (k-1)\delta_{\frac{\hat{F}_{\text{in}}(z)-\hat{F}_{\text{out}}(z)}{k}}.$$

Two components:

- Geometric spectrum
- Community spectrum

Community signal corresponds to $z = 0$.

Concentration Around the Mean

Changing one edge changes $\mathrm{tr}(A^m)$ by $O(n^{m-1})$.

After normalization:

$$\mu_n(P_m) = \frac{1}{n^m}\,\mathrm{tr}(A^m)$$

has Lipschitz constant $O(1/n)$.

Talagrand inequality:

$$\mathbb{P}(|\mu_n(P_m) - \mathbb{E}\mu_n(P_m)| > t) \leq e^{-cnt^2}.$$

Almost Sure Convergence

Exponential tails imply:

$$\sum_n \mathbb{P}(|\mu_n(P_m) - \mathbb{E}\mu_n(P_m)| > \epsilon) < \infty.$$

By Borel–Cantelli:

$$\mu_n(P_m) - \mathbb{E}\mu_n(P_m) \to 0 \quad a.s.$$

Since $\mathbb{E}\mu_n(P_m) \to \mu(P_m)$,

$$\mu_n(P_m) \to \mu(P_m) \quad a.s.$$

Population Matrix

Define the deterministic matrix

$$(B_\sigma)_{ij} = \begin{cases} \mu_{\text{in}} & \sigma_i = \sigma_j \\ \mu_{\text{out}} & \sigma_i \neq \sigma_j. \end{cases}$$

This is the matrix of expected connection probabilities:

$$\mathbb{E}[A \mid \sigma] = B_\sigma.$$

Example: $k = 4$ Equal Communities

Suppose $k = 4$, each community size $n/4$.
Then B_σ has block form:

$$B_\sigma = \begin{pmatrix} \mu_{\text{in}} J & \mu_{\text{out}} J & \mu_{\text{out}} J & \mu_{\text{out}} J \\ \mu_{\text{out}} J & \mu_{\text{in}} J & \mu_{\text{out}} J & \mu_{\text{out}} J \\ \mu_{\text{out}} J & \mu_{\text{out}} J & \mu_{\text{in}} J & \mu_{\text{out}} J \\ \mu_{\text{out}} J & \mu_{\text{out}} J & \mu_{\text{out}} J & \mu_{\text{in}} J \end{pmatrix}$$

where J is the all-ones matrix of size $n/4$.

Eigenvalues of B_σ

Spectrum:

$$\lambda_1 = \frac{n(\mu_{\text{in}} + (k-1)\mu_{\text{out}})}{k}$$

$$\lambda_2 = \cdots = \lambda_k = \frac{n(\mu_{\text{in}} - \mu_{\text{out}})}{k} = \lambda^*$$

All remaining eigenvalues are 0.

Multiplicity of λ^* is $k-1$.

Structure of Population Eigenvectors

Top eigenvector:

$$\mathbf{1} = (1, \dots, 1).$$

Community eigenvectors:
Constant on each community, with zero global sum.

For $k = 4$, a basis can be chosen as:

$$u^{(1)} = (1, -1, 0, 0), \quad u^{(2)} = (1, 0, -1, 0), \quad u^{(3)} = (1, 0, 0, -1).$$

These span a $(k-1)$-dimensional invariant subspace.

Control of Block Averages

To compare eigenspaces, we analyze compressed quantities.

For communities p, q, define

$$Y_{pq}(i) = \frac{1}{|C_q|} \sum_{j \in C_q} A_{ij}.$$

Then

$$\mathbb{E}[Y_{pq}(i)] = \begin{cases} \mu_{\text{in}} & p = q, \\ \mu_{\text{out}} & p \neq q. \end{cases}$$

Each $Y_{pq}(i)$ is an average of independent Bernoulli variables. By Chernoff bounds,

$$\mathbb{P}\left(|Y_{pq}(i) - \mathbb{E}Y_{pq}(i)| > \delta\right) \leq 2\exp(-cn\delta^2).$$

Subspace Control via Davis–Kahan

From Theorem 4.1 (Davis–Kahan type result):

$$\min_{Q \in O(k-1)} \|VQ - U\|_F \;\le\; \frac{C}{\epsilon n} \|(A - B_\sigma)U\|_F.$$

Thus the problem reduces to controlling

$$\|(A - B_\sigma)U\|_F.$$

For $i \in C_p$,

$$(AU)_i = \sum_{q=1}^{k} \left(\sum_{j \in C_q} A_{ij} \right) u_q,$$

which depends on block sums. Applying a union bound over all p, q,

$$\|(A - B_\sigma)U\|_F = O(\sqrt{n \log n}) \quad \text{w.h.p}$$

Plugging into the Davis–Kahan bound yields

$$\min_{Q \in O(k-1)} \|VQ - U\|_F \leq \frac{\sqrt{12k^5 \log n}}{\epsilon\sqrt{n}} \quad \text{w.h.p.}$$

Consequence for Clustering

Rows of eigenvector matrix U:

- Identical within each community.
- Distinct across communities.

Rows of V are small perturbations of those of U.

Therefore:

- Within-community variation is small.
- Between-community separation remains constant.

This yields weak consistency.

A Natural Question

- is there a spectral method in log regime ?
- Can we recover the latent ordering directly?

Task

Recover the community membership of the node with just adjacency matrix.

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

In log regime: Galhotra and Mazumdar's Motif Counting

- Motifs: small subgraph patterns (e.g., triangles, 3-paths) used to infer community structure.
- Demonstrated recovery in the log regime: $r_{in} = a \log n/n$, $r_{out} = b \log n/n$.
- Theoretical guarantee: recovery possible when $a > 2b$.
- Applied to real-world networks: political blogs, DBLP, LiveJournal.

Spectral seriation

- Take Fiedler vector that is second eigenvector of the matrix $L = D - A$.
- Sort the entries of the Fiedler vector.
- Known: Sorted order minimizes the bandwidth.
- Conjecture: Sorted order corresponds to the actual order of vertices on the line.

Spectral Seriation and Bandwidth

- Seriation minimizes bandwidth: $\sum A(i,j)|i-j|$.
- Relaxed to L2 form: $\sum A(i,j)(x_i - x_j)^2$.
- Fiedler vector x minimizes this.

First-zero-r-out Algorithm

	1	2	3	4	5	6
1	0	1	1	0	0	1
2	1	0	1	0	1	0
3	1	1	0	1	0	1
4	0	0	1	0	1	0
5	0	1	0	1	0	1
6	1	0	1	0	1	0

1. Start at any node i.
2. Look right/left in adjacency matrix row i.
3. After first zero, ones indicate same community.
4. Traverse all such nodes recursively.
5. Remaining nodes are in the other community.

Results

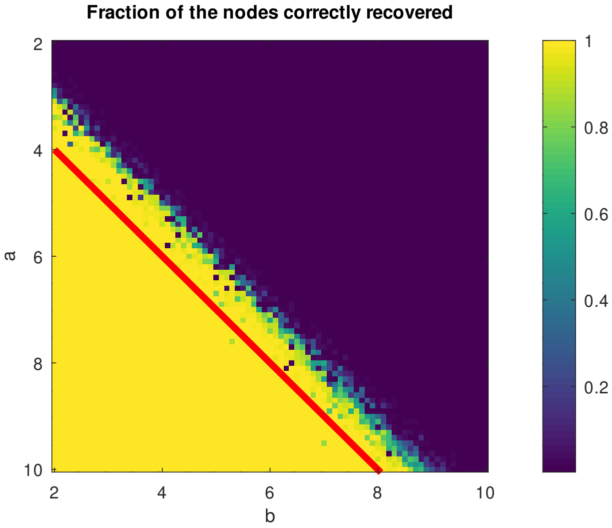

Figure 11: The fraction of recovered nodes from the spectral seriation followed by `first-zero-rout` algorithm for 1000 nodes on [0,1]. a and b are varied from 2 to 10 and red line indicates $a - b = 2$.

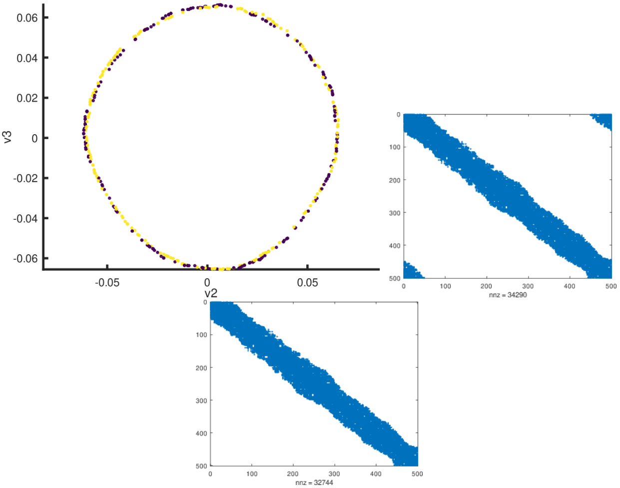

Figure 12: The first two eigenvectors of the Laplacian corresponding to the Geometric block model of 500 nodes on a circle of radius 1 and $r_{in} = 0.624$ and $r_{out} = 0.234$. Corresponding adjacency matrix and the masked adjacency matrix used for recovering community. The algorithm `first-zero-rout` recovers the community label of all the nodes for this numerical example.

Thank you!