Popularity Adjusted Block Models are Generalized Random Dot Product Graphs

John Koo[1], Minh Tang[2], Michael W. Trosset[1]

[1] Department of Statistics, Indiana University
[2] Department of Statistics, North Carolina State University

Block Models

A *block model* is a random graph model in which each vertex v_i has a community label $z_i \in [K]$ and the edge probability between each pair of vertices is determined in part by the pair of labels.

A is an adjacency matrix of a block model if and only if each $A_{ij} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(P_{ij})$ and $P_{ij} = f(z_i, z_j, \cdot)$.

Stochastic Block Model

Each edge probability is fixed for each pair of communities, i.e., $P_{ij} = \theta_{z_i, z_j}$.

Degree Corrected Block Model

Each edge probability is determined by the community edge probability as well as each vertex's degree factor, i.e., $P_{ij} = \theta_{z_i, z_j} \omega_i \omega_j$.

Popularity Adjusted Block Model (Sengupta and Chen 2018)

Each vertex has a popularity parameter for each community that determines its affinity toward that community, i.e., $P_{ij} = \lambda_{i, z_j} \lambda_{j, z_i}$.

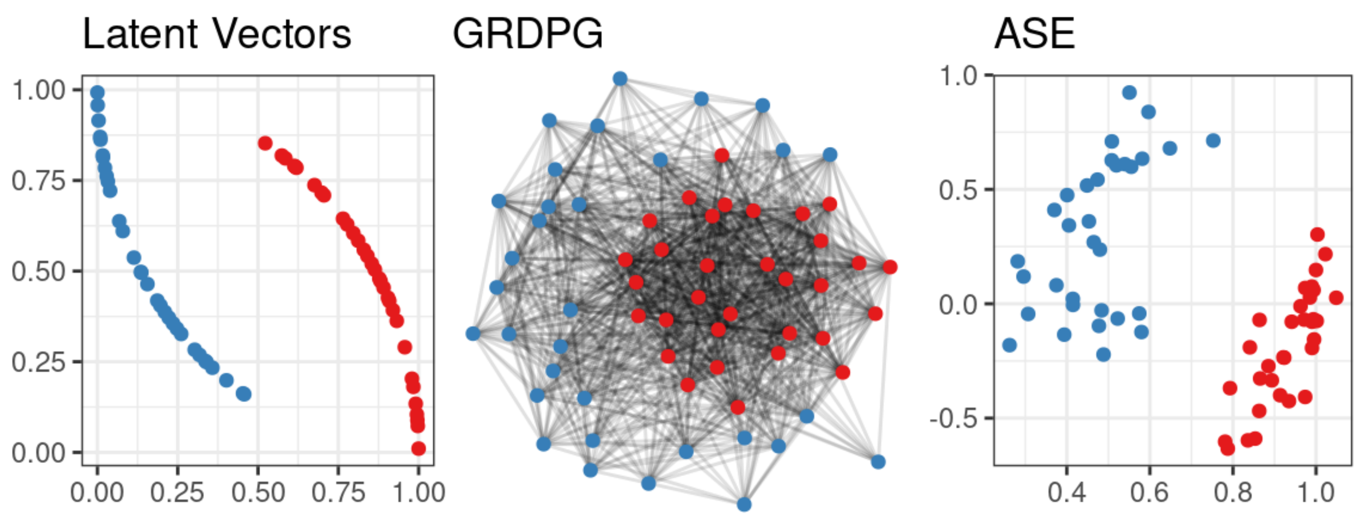
Generalized Random Dot Product Graphs

The *generalized random dot product graph* is a random graph model in which each vertex v_i has a corresponding hidden vector $x_i \in \mathbb{R}^{p+q}$ and each edge probability is the indefinite inner product of the corresponding pair of hidden vectors, i.e., $P_{ij} = x_i^\top I_{p,q} x_j$, $I_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$

Adjacency Spectral Embedding

Approximate A by spectral decomposition $A \approx V_{p,q} \Lambda_{p,q} V_{p,q}^\top$. The subscript p, q denotes the p most positive and q most negative eigenvalues and corresponding eigenvectors. Each $\hat{x}_i$, the i^{th} row of $\hat{X} = V_{p,q} |\Lambda_{p,q}|^{1/2}$, estimates the relative position of its corresponding latent vector x_i, up to an indefinite orthogonal transformation.

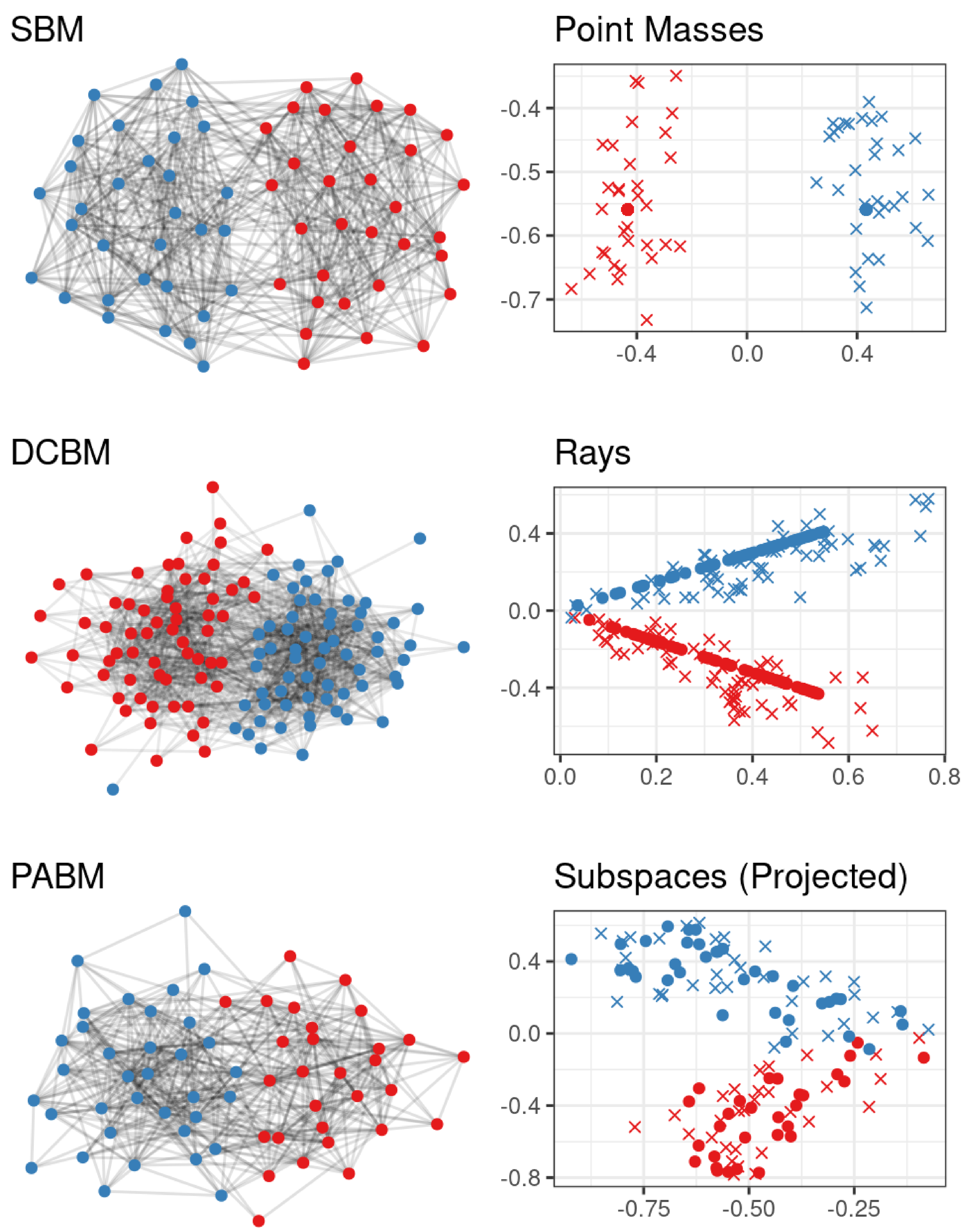
Theorem (Rubin-Delanchy et al. 2022): $\max_i \|\hat{x}_i - Q_n x_i\| = O_P\left(\frac{\log^c n}{n^{1/2}}\right)$ for some $Q_n \in \mathbb{O}(p, q)$.



Connecting Block Models to the GRDPG

It has been previously shown that the SBM and DCBM are GRDPGs in which the communities lie on point masses and line segments respectively.

Theorem (KTT): The PABM is GRDPG in which the communities lie on mutually orthogonal K-dimensional subspaces.



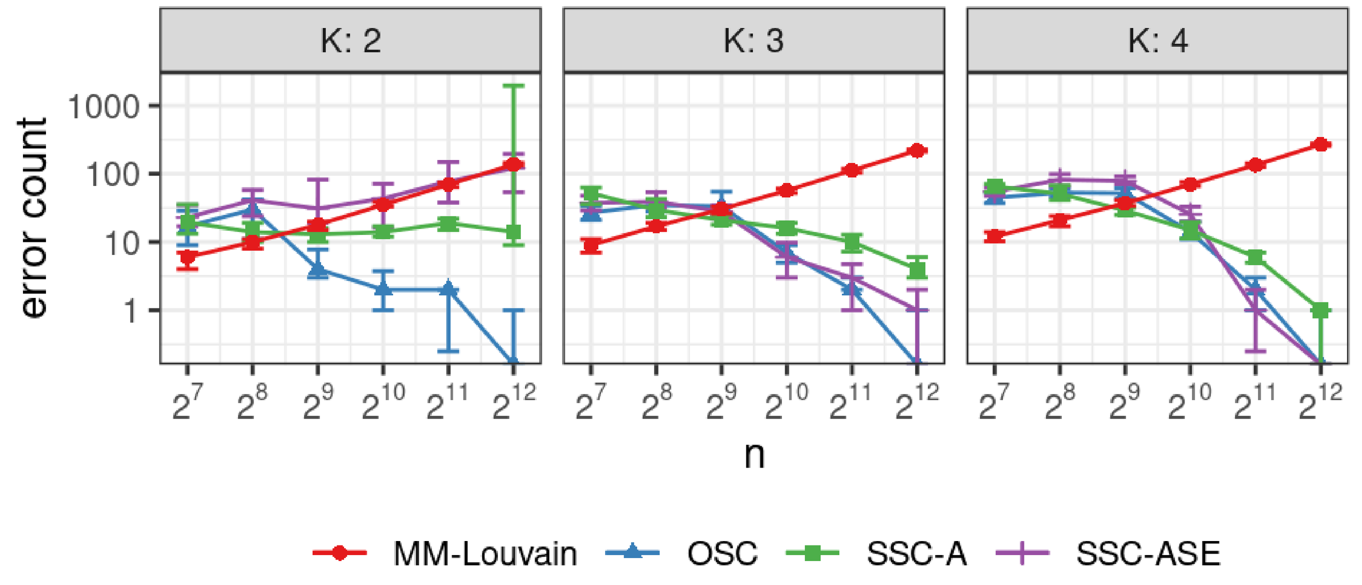
Orthogonal Spectral Clustering

1. Set $p = \frac{K(K+1)}{2}$, $q = \frac{K(K-1)}{2}$ and decompose $A \approx V_{p,q} \Lambda_{p,q} V_{p,q}^\top$.
2. Construct graph G' from $B = |n V_{p,q} V_{p,q}^\top|$.
3. Partition G' into K disconnected subgraphs.

Theorem (KTT): $\max_{i,j} B_{ij} = O_P\left(\frac{\log^c n}{\sqrt{n \rho_n}}\right)$ for each (i, j) in different communities.

Corollary: OSC results in zero community detection error with probability 1 as $n \to \infty$.

Simulation Study



Conclusion

Exploiting the geometry of PABMs (and other block models) results in intuitive and consistent community detection algorithms.