Manifold Clustering in the Setting of Generalized Random Dot Product Graphs

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

[1] Indiana University
[2] North Carolina State University

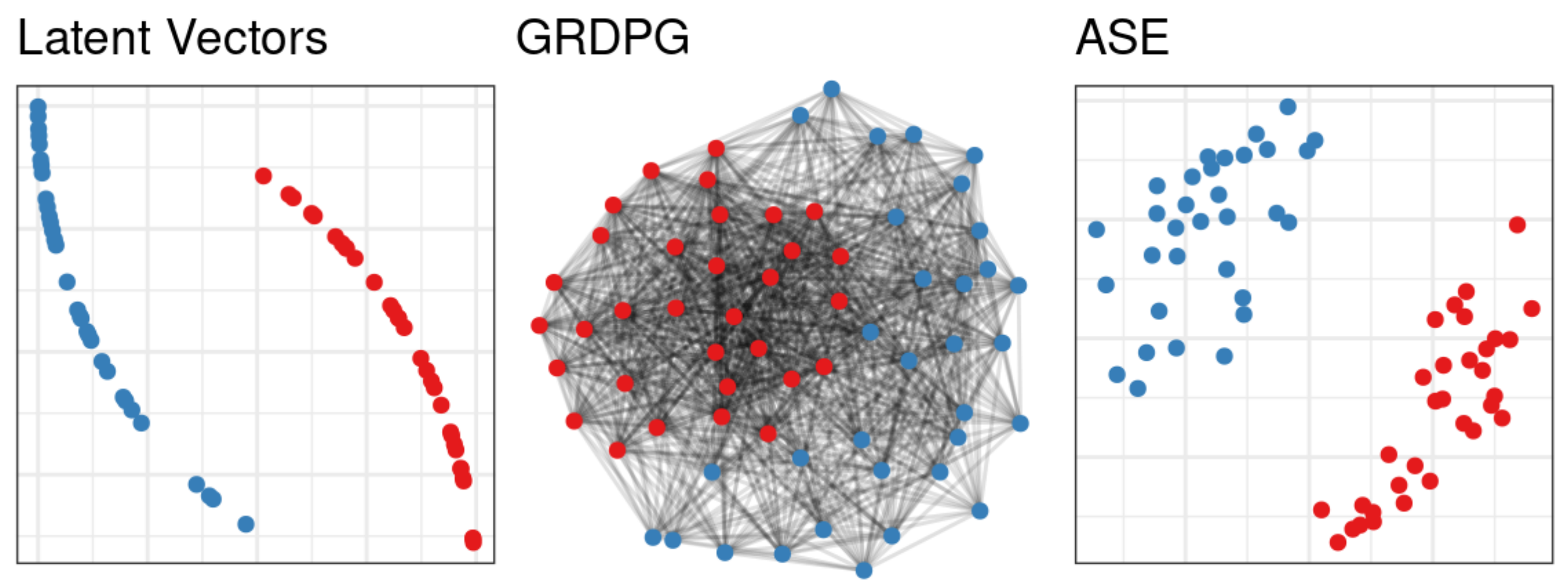
Generalized Random Dot Product Graphs

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

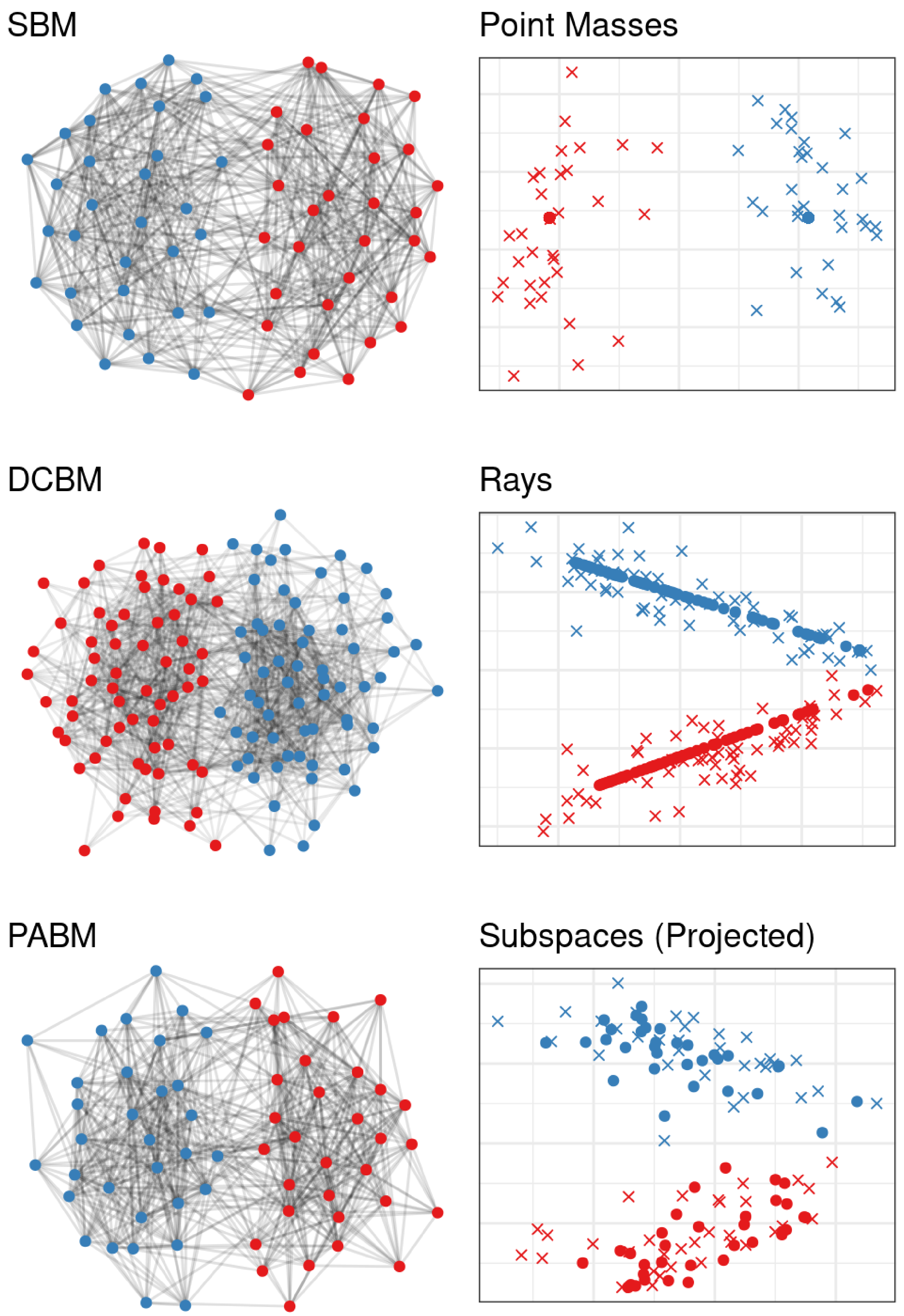
Approximate A by spectral decomposition $A \approx V_{p,q} \Lambda_{p,q} V_{p,q}^\top$. The subscript p, q denotes the p most positive and q most negative eigenvalues and corresponding eigenvectors. Each $\hat{x}_i$, the i^{th} row of $\hat{X} = V_{p,q} |\Lambda_{p,q}|^{1/2}$, estimates the relative 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 $Q_n \in \mathbb{O}(p, q)$.

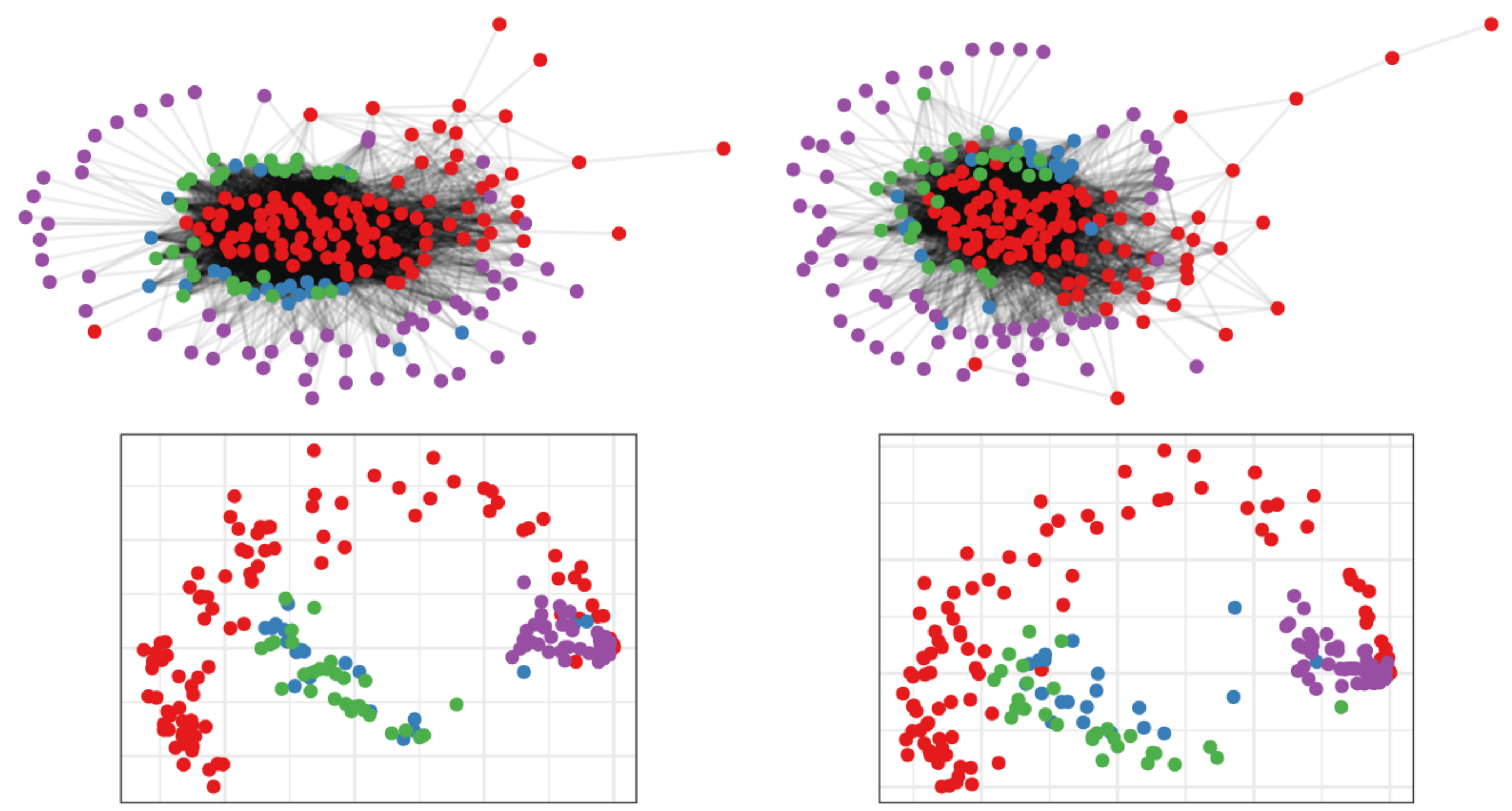


Block Models as GRDPGs

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



GRDPGs with Nonlinear Latent Structure



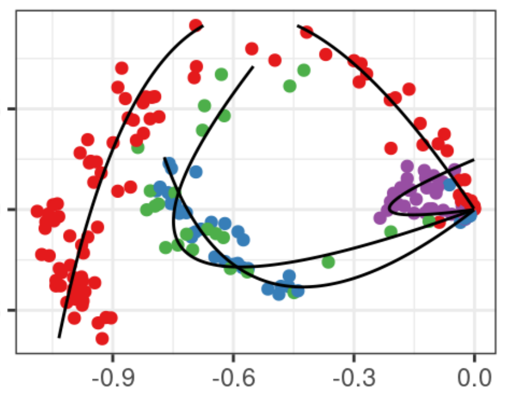
Manifold Block Model

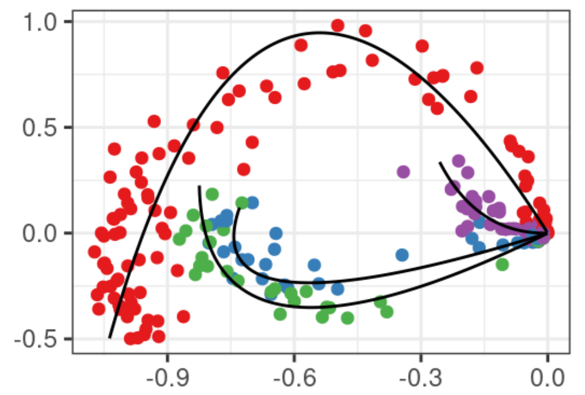
Let $p, q \geq 0$, $d = p + q \geq 1$, $1 \leq r < d$, $K \geq 2$, and $n > K$ be integers. Define manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_K \in \mathcal{X}$ for $\mathcal{X} = \{x, y \in \mathbb{R}^d : x^\top I_{p,q} y \in [0, 1]\}$ each by continuous function $g_k : [0, 1]^r \to \mathcal{X}$. Define probability distribution F with support $[0, 1]^r$. Then the following mixture model is a *manifold block model*:

1. Draw labels $z_1, \ldots, z_n \overset{\text{iid}}{\sim} \text{Cat}(\alpha_1, \ldots, \alpha_K)$.
2. Draw latent vectors by first taking $t_1, \ldots, t_n \overset{\text{iid}}{\sim} F$ and then computing each $x_i = g_{z_i}(t_i)$.
3. Compile the latent vectors into data matrix $X = [x_1 \mid \cdots \mid x_n]^\top$ and define the adjacency matrix as $A \sim \text{GRDPG}_{p,q}(X)$.

K-Curves Clustering

```
1. Compute $X$, the ASE of $A$ using the $p$ most positive and $q$ most negative eigenvalues and their corresponding eigenvectors.
2. Initialize community labels $z_1, \ldots, z_n$.
3. While change in $\sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$ is less than $\epsilon$:
   i. For $k = 1, \ldots, K$:
      a. Define $X_k$ as the rows of $X$ for which $z_i = k$.
      b. Fit curve $g_k$ and positions $t_{k_i}$ to $X_k$ by minimizing $\sum_{k_i} \|x_{k_i} - g_k(t_{k_i})\|^2$.
   ii. For $k = 1, \ldots, K$:
      a. Assign $z_i \leftarrow \arg\min_\ell \|x_i - g_\ell(t_i)\|^2$
         .
```





Conclusion

Block models can be expressed as GRDPGs in which the communities are linear structures in the latent space. We propose the manifold block model to characterize nonlinear latent structures and the K-curves clustering algorithm to estimate these structures for community detection.