Popularity Adjusted Block Models are Generalized Random Dot Product Graphs

Future Leaders Summit 2022 Lightning Presentation

John Koo, Indiana University

April 2022

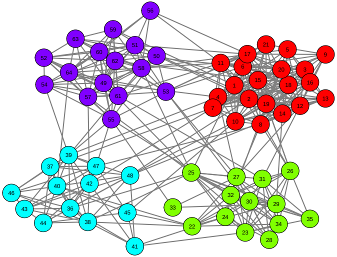
Contributors

John Koo,
PhD Student in
Statistical Science,
Indiana University

Minh Tang,
Assistant Professor of
Statistics,
NC State University

Michael Trosset,
Professor of Statistics,
Indiana University

Community Detection for Networks



How might we cluster the nodes of a network?

Bernoulli Graphs

Let $G = (V, E)$ be an undirected and unweighted graph with $|V| = n$.

G is described by adjacency matrix A such that

$$A_{ij} = \begin{cases} 1 & \exists \text{ edge between } i \text{ and } j \\ 0 & \text{else} \end{cases}$$

$A_{ji} = A_{ij}$ and $A_{ii} = 0 \ \forall i, j \in [n]$.

$A \sim \text{BernoulliGraph}(P)$ iff:

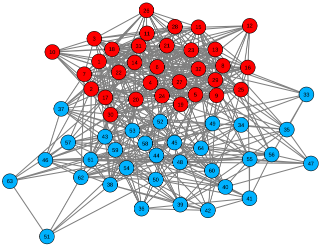
1. $P \in [0, 1]^{n \times n}$ describes edge probabilities between pairs of vertices.
2. $A_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(P_{ij})$ for each $i < j$.

Block Models

Suppose each vertex $v_1, ..., v_n$ has labels $z_1, ..., z_n \in \{1, ..., K\}$, and each P_{ij} depends on labels z_i and z_j.
Then $A \sim \text{BernoulliGraph}(P)$ is a *block model*.

Example 1: Stochastic Block Model with $K = 2$ communities

- $z_1, ..., z_n \in \{1, 2\}$
- $P_{ij} = \begin{cases} p & z_i = z_j = 1 \\ q & z_i = z_j = 2 \\ r & z_i \neq z_j \end{cases}$



Popularity Adjusted Block Model

Def Popularity Adjusted Block Model (Sengupta and Chen, 2017):

Let each vertex $i \in [n]$ have K popularity parameters $\lambda_{i1}, ..., \lambda_{iK} \in [0, 1]$. Then $A \sim$ PABM($\{\lambda_{ik}\}_K$) if each $P_{ij} = \lambda_{i z_j} \lambda_{j z_i}$.

Lemma (Noroozi, Rimal, and Pensky, 2020):

A is sampled from a PABM if P can be described as:

1. Let each $P^{(kl)}$ denote the $n_k \times n_l$ matrix of edge probabilities between communities k and l.
2. Organize popularity parameters as vectors $\lambda^{(kl)} \in \mathbb{R}^{n_k}$ such that $\lambda_i^{(kl)} = \lambda_{k_i l}$ is the popularity parameter of the i^{th} vertex of community k towards community l.
3. Each block can be decomposed as $P^{(kl)} = \lambda^{(kl)} (\lambda^{(lk)})^\top$.

Generalized Random Dot Product Graph

Def Generalized Random Dot Product Graph
(Rubin-Delanchy, Cape, Tang, Priebe, 2020)

Let $I_{p,q} = \text{blockdiag}(I_p, -I_q)$ and suppose that $x_1, \ldots, x_n \in \mathbb{R}^{p+q}$ are such that $x_i^\top I_{p,q} x_j \in [0, 1]$.
Then $A \sim \text{GRDPG}_{p,q}(X)$ iff $A \sim \text{BernoulliGraph}(X I_{p,q} X^\top)$, where $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$

Adjacency Spectral Embedding (Sussman et al., 2012) estimates $x_1, ..., x_n \in \mathbb{R}^{p+q}$ from A:

1. Let $\hat{\Lambda}$ be the diagonal matrix that contains the absolute values of the p most positive and the q most negative eigenvalues.
2. Let $\hat{V}$ be the matrix whose columns are the corresponding eigenvectors.
3. Compute $\hat{X} = \hat{V}\hat{\Lambda}^{1/2}$.

Theorem: $\max_i \|\hat{X}_i - Q_n X_i\| = O_P\left(\frac{(\log n)^c}{n^{1/2}}\right)$

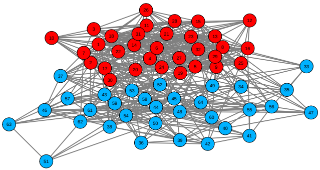
Connecting Block Models to the (G)RDPG Model

All Bernoulli Graphs are RDPG (if P is positive semidefinite) or GRDPG (in general).

Example 1 (cont'd): Assortative SBM ($pq > r^2$) with $K = 2$

$$P_{ij} = \begin{cases} p & z_i = z_j = 1 \\ q & z_i = z_j = 2 \\ r & z_i \neq z_j \end{cases}$$

$$P = \begin{bmatrix} P^{(11)} & P^{(12)} \\ P^{(21)} & P^{(22)} \end{bmatrix} = XX^\top$$

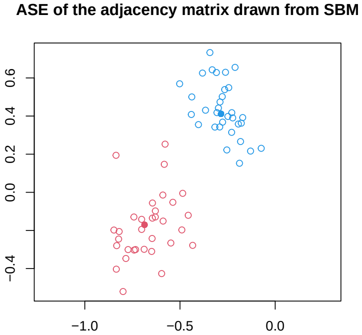


$$X = \begin{bmatrix} \sqrt{p} & 0 \\ \vdots & \vdots \\ \sqrt{p} & 0 \\ \sqrt{r^2/p} & \sqrt{q - r^2/p} \\ \vdots & \vdots \\ \sqrt{r^2/p} & \sqrt{q - r^2/p} \end{bmatrix}$$

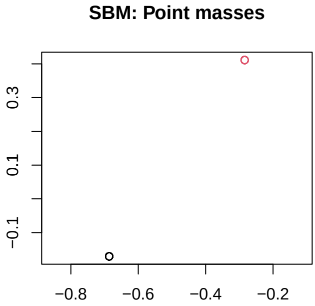
Connecting Block Models to the (G)RDPG Model

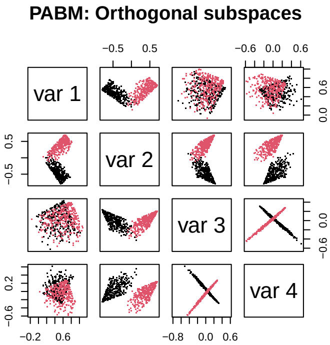
Example 1 (cont'd): If we want to perform community detection,

1. Note that A is a RDPG because $P = XX^\top$.
2. Compute the ASE $A \approx \hat{X}\hat{X}^\top$ with $\hat{X} = \hat{V}\hat{\Lambda}^{1/2}$.
3. Apply clustering algorithm (e.g., K-means) to $\hat{X}$,
   noting that as $n \to \infty$, the ASE approaches point masses.



Connecting Block Models to the (G)RDPG Model





Connecting the PABM to the GRDPG

Theorem (KTT): $A \sim \text{PABM}(\{\lambda^{(kl)}\}_K)$ is equivalent to $A \sim \text{GRDPG}_{p,q}(XU)$ with

- $p = K(K+1)/2$, $q = K(K-1)/2$
- $U \in \mathbb{O}(K^2)$
- $X \in \mathbb{R}^{n \times K^2}$ is block diagonal and composed of $\{\lambda^{(kl)}\}_K$ with each block corresponding to a community.

$$X = \begin{bmatrix} \Lambda^{(1)} & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \Lambda^{(K)} \end{bmatrix} \in \mathbb{R}^{n \times K^2}$$

$$\Lambda^{(k)} = \begin{bmatrix} \lambda^{(k1)} & \cdots & \lambda^{(kK)} \end{bmatrix} \in \mathbb{R}^{n_k \times K}$$

$$A \sim \text{PABM}(\{\lambda_{ik}\}_K) \iff A \sim \text{GRDPG}_{p,q}(XU)$$

Orthogonal Spectral Clustering

Theorem (KTT): If $P = V \Lambda V^\top$ and $B = nVV^\top$, then $B_{ij} = 0$ if $z_i \neq z_j$.

Algorithm: Orthogonal Spectral Clustering:

1. Let V be the eigenvectors of A corresponding to the $K(K+1)/2$ most positive and $K(K-1)/2$ most negative eigenvalues.
2. Compute $B = |nVV^\top|$ applying $|\cdot|$ entry-wise.
3. Construct graph G using B as its similarity matrix.
4. Partition G into K disconnected subgraphs.

Theorem (KTT): Let $\hat{B}_n$ with entries $\hat{B}_n^{(ij)}$ be the affinity matrix from OSC. Then $\forall$ pairs (i, j) belonging to different communities and sparsity factor satisfying $n\rho_n = \omega((\log n)^{4c})$,

$$\max_{i,j} \hat{B}_n^{(ij)} = O_P\Big(\frac{(\log n)^c}{\sqrt{n\rho_n}}\Big)$$

Simulation Results

Simulation setup:

1. $z_1, ..., z_n \overset{\text{iid}}{\sim} \text{Categorical}(1/K, ..., 1/K)$
2. $\lambda_{ik} \overset{\text{iid}}{\sim} \text{Beta}(a_{ik}, b_{ik})$
   $$a_{ik} = \begin{cases} 2 & z_i = k \\ 1 & z_i \neq k \end{cases} \quad b_{ik} = \begin{cases} 1 & z_i = k \\ 2 & z_i \neq k \end{cases}$$
3. $P_{ij} = \lambda_{iz_j} \lambda_{jz_i}$
4. $A \sim \text{BernoulliGraph}(P)$