Popularity Adjusted Block Models are Generalized Random Dot Product Graphs

Future Leaders Summit 2022 Lightning Presentation

John Koo, Indiana University

April 2022

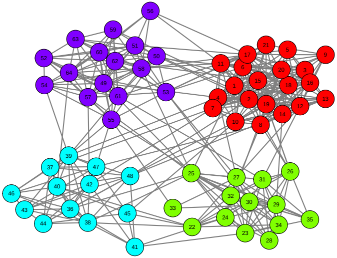
Contributors

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 can we cluster the nodes of a network?

Statistical inference (parametric approach):

1. Define a generative model for graph: $G \mid z_1, ..., z_n, \vec{\theta} \sim P(\vec{z}, \vec{\theta})$.
2. Develop a method for obtaining estimators: $f(G) = (\hat{\vec{z}}, \hat{\vec{\theta}})$.
3. Describe asymptotic properties of estimators: $(\hat{\vec{z}}, \hat{\vec{\theta}}) \to (\vec{z}, \vec{\theta})$.

Bernoulli Graphs

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

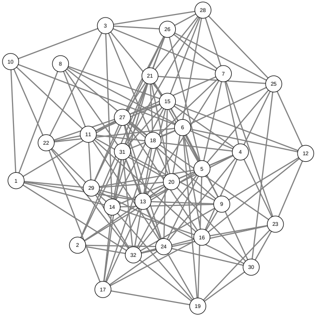
G is described by adjacency matrix A such that $A_{ij} = \begin{cases} 1 & \exists \text{ edge between } i \text{ and } j \\ 0 & \text{else} \end{cases}$

$A_{ji} = A_{ij}$ 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 $i < j$.

Example 1: If every entry $P_{ij} = \theta$, then $A \sim \text{BernoulliGraph}(P)$ is an Erdos-Renyi graph. For this model, $A_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$.

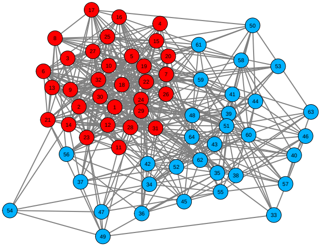


Block Models

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 2: Stochastic Block Model with two 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 $A \sim \text{PABM}(\{\lambda_{ik}\}_K)$ if each $P_{ij} = \lambda_{iz_j} \lambda_{jz_i}$.

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)

$A \sim \text{GRDPG}_{p,q}(X)$ iff

- Latent vectors $x_1, ..., x_n \in \mathbb{R}^{p+q}$ such that $x_i^\top I_{p,q} x_j \in [0,1]$ and $I_{p,q} = \text{blockdiag}(I_p, -I_q)$
- $A \sim \text{BernoulliGraph}(X I_{p,q} X^\top)$ where $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$
  - $P(\text{edge between } v_i, v_j) = x_i^\top I_{p,q} x_j$

(Generalized) Random Dot Product Graph Model

Recovery/Estimation

Want to estimate X from A, or alternatively, interpoint distances, inner products, or angles.

Adjacency Spectral Embedding

To embed in $\mathbb{R}^{p+q}$,

1. Compute $A \approx \hat{V}\hat{\Lambda}\hat{V}^\top$ where $\hat{\Lambda} \in \mathbb{R}^{(p+q)\times(p+q)}$ and $\hat{V} \in \mathbb{R}^{n\times(p+q)}$ by using the p most positive and q most negative eigenvalues.
2. Let $\hat{X} = \hat{V}|\hat{\Lambda}|^{1/2}$.

$\max_i \|\hat{X}_i - Q_n X_i\| = O_P\left(\frac{(\log n)^c}{n^{1/2}}\right)$ (Rubin-Delanchy et al., 2020)

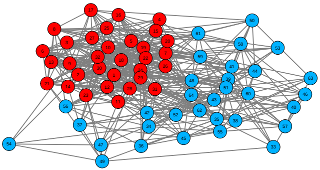
Connecting Block Models to the (G)RDPG Model

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

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

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

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

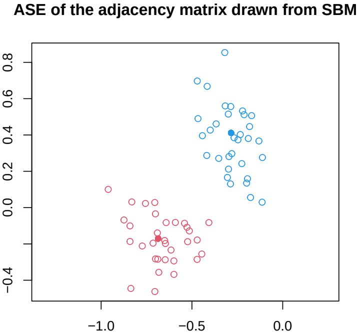


$$X = \begin{bmatrix} \sqrt{p} & 0 \\ \vdots & \vdots \\ \sqrt{p} & 0 \\ \sqrt{r^2/p} & \sqrt{q - r^2/p} \\ \vdots & \vdots \\ \sqrt{r^2/p} & \sqrt{q - r^2/p} \end{bmatrix}$$

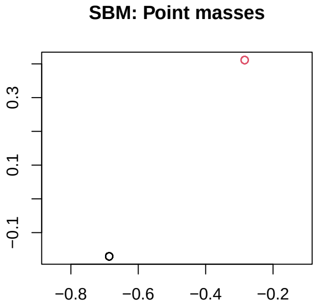
Connecting Block Models to the (G)RDPG Model

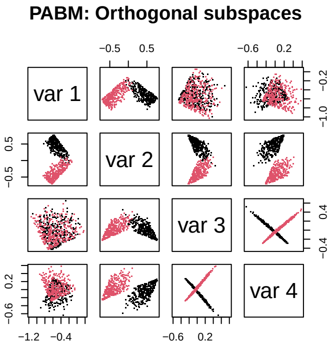
Example 2 (cont'd): If we want to perform community detection,

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.

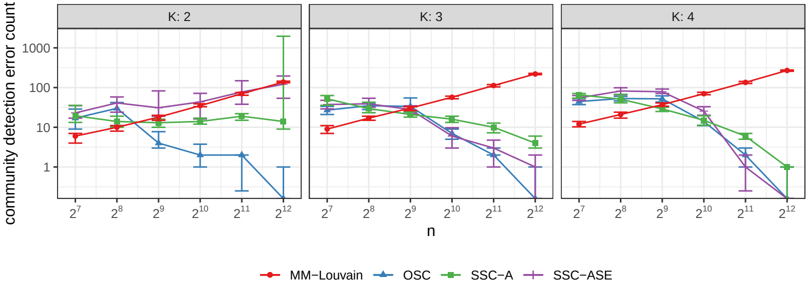
Theorem (KTT): Let $\hat{B}_n$ with entries $\hat{B}_n^{(ij)}$ be the affinity matrix from OSC. Then $\forall$ pairs (i,j) belonging to different communities and sparsity factor satisfying $n\rho_n = \omega((\log n)^{4c})$,

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

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. $A \sim \text{BernoulliGraph}(P)$



Conclusion

1. The PABM is a recently developed flexible block model that can be used to describe graphs with community structure.
2. The GRDPG, which can describe all block models, motivates a spectral approach to statistical inference on graphs.
3. Under the GRDPG framework, the PABM with K communities can be induced by a latent configuration in $\mathbb{R}^{K^2}$ consisting of K K-dimensional subspaces that are orthogonal to each other.
4. The latent configuration of the PABM under the GRDPG framework leads to an intuitive method for community detection with nice theoretical asymptotic properties.