Community Detection Methods for Random Dot Product Graphs and Generalized Random Dot Product Graphs

A dissertation proposal submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy
in
Statistical Science

John Koo

Research Committee Members
Dr. Michael Trosset
Dr. Minh Tang
Dr. Julia Fukuyama
Dr. Roni Khardon
Dr. Fangzheng Xie

Date TBA

Department of Statistics
Indiana University
Bloomington, Indiana

Contents

1 Introduction

1.1 Research Goal

Graph and network data have become increasingly more widespread in various fields including sociology, neuroscience, biostatistics, and computer science. This has resulted in certain challenges for researchers, as most traditional statistical and machine learning methods are incompatible with graphs and instead require the data to exist in a subset of Euclidean space. This includes clustering and community detection, which are the typical problems of interest when working with graphs. Common clustering methods often involve calculating some central or representative point around which the data belonging to that cluster lie (e.g., Lloyd's algorithm for K-means clustering [5], Gaussian Mixture Models [4]). Because these methods involve computing summary statistics within each cluster, they cannot be applied directly to graphs as community detection algorithms.

Some proposed methods for unifying graph community detection with traditional clustering techniques involve embedding the graph into Euclidean space, of which the most popular is Spectral Clustering [12]. The Random Dot Product Graph (RDPG) [14] and Generalized Random Dot Product Graph (GRDPG) [9] models take this further by explicitly constructing generative models such that latent positions in Euclidean space are used to induce graphs. A community detection method motivated by this would involve learning the latent positions given an observed graph and then learning the community labels given the latent positions.

The aim of our research is to develop consistent community detection techniques under the RDPG and GRDPG frameworks. First, we explore existing generative graph models with underlying community structures that can be inferred by connecting the generative models to the RDPG or GRDPG. Then we explore other latent structures or mixture distributions in the latent space that induce graphs for which consistent community detection is possible.

1.2 Notation

Let $G = (V, E)$ be an undirected, unweighted graph with no self-loops with n vertices. Denote $A \in \{0,1\}^{n \times n}$ as the adjacency matrix for G such that $A_{ij} = 1$ if there exists an edge between vertices i and j and $A_{ij} = 0$ otherwise. Because G is symmetric and contains no self-loops, $A_{ij} = A_{ji}$ and $A_{ii} = 0$ for $i, j \in [n]$. We further restrict our analyses to independent Bernoulli graphs. Let $P \in [0,1]^{n \times n}$ be a symmetric matrix of edge probabilities. Graph G is sampled by $A_{ij} \overset{\text{indep}}{\sim} P_{ij}$ for each $0 \le i < j \le n$ ($A_{ji} = A_{ij}$ and $A_{ii} = 0$). Denote $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n \times d}$ as the sample $x_1, ..., x_n \in \mathbb{R}^d$, and denote $z_1, ..., z_n \in [K]$ as their corresponding (hidden) labels.

2 Literature Review

2.1 Generative Graph Models and Community Detection

Generative models for symmetric Bernoulli graphs involve defining the edge probability matrix P whose ij^{th} entry is the probability of an edge between vertices i and j for each $i, j \in [n]$. Community detection and recovery starts by restricting the generative model such that for each pair of vertices, the probability of an edge between the vertices is conditioned on the labels of the vertices. One such model is the Stochastic Block Model (SBM) [6]: Given K communities and each vertex belonging to one community, the SBM restricts P to $K(K+1)/2$ unique entries such that $P_{ij} = B_{z_i z_j}$ where $B \in [0,1]^{K \times K}$ with entries B_{kl} denoting the edge probability of each vertex in community k having an edge with each vertex in community l. The homogeneous SBM further restricts P to two unique entries such that $P_{ij} = p$ if $z_i = z_j$ and $P_{ij} = q$ otherwise. Multiple generalizations of the SBM have been introduced since, including the Degree Corrected Block Model (DCBM) and the Popularity Adjusted Block Model (PABM). Like the SBM, these models involve edge probability matrix P that is restricted to rank $d < n$ in part based on community labels.

2.2 (Generalized) Random Dot Product Graphs

The RDPG starts with points in latent space $X \in \mathcal{X} \subset \mathbb{R}^d$ such that $\forall x, y \in \mathcal{X}$, $x^\top y \in [0, 1]$. P is then constructed as $P = XX^\top$, and graph G with adjacency matrix A is drawn from P. We provide a more formal definition of the RDPG and GRDPG below.

Definition 1 ((Generalized) Random Dot Product Graph). *Let $X \in \mathbb{R}^{n \times d}$ be a collection of n points in $\mathcal{X} \subset \mathbb{R}^d$ such that $\forall x, y \in \mathcal{X}$, $x^\top y. \in [0, 1]$. $G = (V, E)$ is a Random Dot Product Graph if its adjacency matrix A is drawn such that $A_{ij} \sim Bernoulli(x_i^\top x_j)$ for $i < j$, with $A_{ji} = A_{ij}$ and $A_{ii} = 0 \ \forall i, j \in [n]$. If on the other hand $A_{ij} \sim Bernoulli(x_i^\top I_{p,q} x_j)$ where $I_{p,q} = blockdiag(I_p, -I_q)$ and $p + q = d$, then A is the adjacency matrix of a Generalized Random Dot Product Graph. These are denoted by $A \sim RDPG(X)$ and $A \sim GRDPG_{p,q}(X)$ respectively.*

In addition, let F be a probability distribution with support $\mathcal{X}$, and $x_1, ..., x_n \overset{iid}{\sim} F$ with $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$. If A is drawn from X as before, then $(A, X) \sim RDPG(F, n)$ or $(A, X) \sim GRDPG_{p,q}(F, n)$.

The structure of the RDPG and GRDPG provides an intuitive method for recovery of the latent positions via spectral embedding.

Definition 2 (Adjacency Spectral Embedding). *Let $A \sim RDPG(X)$ for $X \in \mathcal{X} \subset \mathbb{R}^{n \times d}$. Let $A = V \Lambda V^\top$ be the approximate spectral decomposition of A corresponding to the d largest eigenvalues and their corresponding eigenvectors. Then the rows of $V \Lambda^{1/2}$ are the scaled Adjacency Spectral Embedding (ASE) of A, and the rows of V are the unscaled ASE of A.*

If $A \sim GRDPG_{p,q}(X)$, then let $A = V \Lambda V^\top$ be the approximate spectral decomposition of A corresponding to the p most positive and q most negative eigenvalues of A and their corresponding eigenvectors. Then the rows of $V |\Lambda|^{1/2}$ and V are the scaled and unscaled ASE of A respectively.

Athreya et al. showed that under mild conditions, if $(A_n, X_n) \sim \text{RDPG}(F, n)$ and $\hat{X}_n$ is the scaled ASE of A_n, for some sequence of orthogonal matrices W_n,

$$\max_i \|(\hat{X}_n)_i - W_n (X_n)_i\| \overset{a.s.}{\to} 0 \tag{1}$$

Similarly, Rubin-Delanchy et al. [9] showed that for $(A_n, X_n) \sim \text{GRDPG}_{p,q}(F, n)$,

$$\max_i \|(\hat{X}_n)_i - Q_n (X_n)_i\| \overset{a.s.}{\to} 0 \tag{2}$$

where Q_n is a sequence of matrices in $\mathbb{O}(p, q)$, the indefinite orthogonal group of order p, q.

It is straightforward to show that all Bernoulli graphs with positive semidefinite P are special cases of the RDPG, which is a special case of the GRDPG, and all graphs generated by P are special cases of the GRDPG. This includes the SBM, DCBM, and PABM.

Example (Connecting the SBM to the RDPG). *Let $G = (V, E)$ with adjacency matrix A be sampled from the homogeneous SBM with two communities such that within-community edge probability p and between-community edge probability q where $p > q$. Let community 1 have n_1 vertices and community 2 have n_2 vertices such that $n_1 + n_2 = n$. Without loss of generality, organize P and A such that the kl^{th} block represents edges between communities k and l. Then $P = \begin{bmatrix} P^{(11)} & P^{(12)} \\ P^{(21)} & P^{(22)} \end{bmatrix}$ where each block is a constant value, e.g., $P_{ij}^{(11)} = p$. One RDPG representation of this SBM is:*

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

where the first n_1 rows are $\begin{bmatrix} \sqrt{p} & 0 \end{bmatrix}$ and the next n_2 rows are $\begin{bmatrix} \sqrt{r^2/p} & \sqrt{q - r^2/p} \end{bmatrix}$. Then it can be shown that

$$P = XX^\top$$

The ASE of the assortative SBM consists of points in $\mathbb{R}^K$ that lie near one of K centers, depending on the community label, leading to ASE followed by K-means clustering (or similar, e.g., GMM) as a consistent community detection algorithm [7]. A similar result can be shown for the DCBM and PABM but with different structures in the ASE.

Given sufficiently large sample size n, the scaled ASE of affinity matrix A drawn from a RDPG or GRDPG will asymptotically approach the original latent positions X with probability 1, up to a linear transformation (orthogonal transformation for the RDPG, a composition of an orthogonal and scale transformations for the GRDPG). Thus if X consists of points that lie on subspaces of $\mathbb{R}^d$, then both the scaled and unscaled ASE of $A \sim \text{RDPG}(X)$ or $A \sim \text{GRDPG}(X)$ will consist of points that lie near subspaces, with some noise that almost surely goes to 0 as $n \to \infty$, motivating ASE followed by SSC as an asymptotically consistent method for community detection. The Popularity Adjusted Block Model (PABM) [10] is a generative graph model with underlying communities such that each community lies on a subspace. Noroozi et al. [8] showed that SSC is able to recover the subspaces and therefore perform community detection for the PABM given $P = XI_{p,q}X^\top$, the edge probability matrix, rather than A, the adjacency matrix. Combining the results of Rubin-Delanchy et al. and Wang and Xu, it should be possible to recover the communities using A as well.

2.3 Subspace Clustering

Subspace clustering, which assumes that points in $\mathbb{R}^d$ each lie on one of K subspaces of $\mathbb{R}^d$, is an approach that has found a wide range of uses by the Statistics and Machine Learning communities, particularly within the field of Computer Vision [3].

Of particular interest is Sparse Subspace Clustering (SSC), which is performed by solving an optimization problem for each observed point in a sample. Given $X \in \mathbb{R}^{n \times d}$ with vectors $x_i^\top \in \mathbb{R}^d$ as rows of X, the optimization problem $c_i = \min_c \|c\|_1$ subject to $x_i = X_{-i}c$ and $c^{(i)} = 0$ is solved for each $i \in [n]$. The solutions are collected into matrix $C = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}^\top$ to construct affinity matrix $B = |C| + |C^\top|$. If each x_i lie perfectly on one of K subspaces, B is sparse such that $B_{ij} = 0 \; \forall x_i, x_j$ belonging to different subspaces. Then B can describe a graph with at least K disjoint subgraphs, and if the number of subgraphs is exactly K, each subgraph maps onto a subspace.

In practice, SSC is performed by solving the LASSO problems:

$$c_i = \arg\min_c \frac{1}{2}\|x_i - X_{-i}c\|_2^2 + \lambda\|c\|_1 \tag{3}$$

for some sparsity parameter $\lambda > 0$. The c_i vectors are then collected into C and B as described before. If X is noisy in that each x_i does not lie exactly on one of K subspaces but near it, the choice of λ becomes important in guaranteeing the Subspace Detection Property (SDP) [13].

Definition 3 (Subspace Detection Property)**.** *Let $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ be noisy points sampled from K subspaces. Let C and B be constructed from the solutions of LASSO problems as described in (3). If each column of C has nonzero norm and $B_{ij} = 0 \; \forall \; x_i$ and x_j sampled from different subspaces, then X obeys the Subspace Detection Property.*

Remark. In practice, a noisy sample X often does not obey SDP. In such cases, B is treated as an affinity matrix for a graph which is then partitioned into K subgraphs to obtain the clustering. On the other hand, if X does obey the SDP, B describes a graph with at least K disconnected subgraphs. Ideally, when SDP holds, there are exactly K subgraphs which map to each subspace, but it could be the case that some of the subspaces are represented by multiple disconnected subgraphs. SDP is contingent on choosing a sufficiently large sparsity parameter λ.

2.4 Manifold Learning

Trosset et al. [11] showed that the ASE of a RDPG can be used to recover one-dimensional manifolds. Suppose $f : [0, 1] \mapsto \mathcal{X}$ such that f is smooth and $\mathcal{X}$ represents a curve or one-dimensional manifold in $\mathbb{R}^d$. If $t_1, ..., t_n \overset{iid}{\sim} F$ such that F has support $[0, 1]$, the latent positions are $x_i = f(t_i)$ with y_i is its corresponding point in the scaled ASE, and $d_\epsilon(\cdot, \cdot)$ is the shortest path distance of an ϵ-neighborhood graph. Under certain mild conditions, the shortest path distances of the ϵ-neighborhood graph of the ASE approaches the arc lengths along f:

$$d_\epsilon(y_i, y_j) \xrightarrow{p} \int_{t_i}^{t_j} \sqrt{\sum_r^d \left(\frac{df_r}{dt}\right)^2} dt \tag{4}$$

Athreya et al. [2] extended this further by generating a RDPG from a mixture of distributions on a curve. In their example, points were sampled from a mixture of two Beta distributions on the Hardy-Weinberg curve to construct the latent positions of a RDPG, with the goal of recovering the hidden mixture distribution from an observed graph.

3 Proposed Research

3.1 Connecting Existing Generative Graph Models to the RDPG and GRDPG

3.1.1 Subspace Clustering

As discussed in §2.2, if the latent positions of a RDPG or GRDPG model are such that they lie on a small number of subspaces, ASE followed by SSC may be able to identify whether vertices of the graph v_i, v_j belong to the same subspace or to different subspaces, and one such model that is consistent with this construction is the PABM.

3.1.2 Popularity Adjusted Block Model

We will first define the PABM.

Definition 4 (Popularity Adjusted Block Model)**.** *Let $P \in [0,1]^{n \times n}$ be a symmetric edge probability matrix for a set of n vertices, V. Each vertex has a community label $1, ..., K$, and the rows and columns of P are arranged by community label such that $n_k \times n_l$ block $P^{(kl)}$ describes the edge probabilities between vertices in communities k and l ($P^{(lk)} = (P^{(kl)})^\top$). Let graph $G = (V, E)$ be an undirected, unweighted graph such that its corresponding adjacency matrix $A \in \{0,1\}^{n \times n}$ is a realization of Bernoulli(P), i.e., $A_{ij} \stackrel{indep}{\sim}$ Bernoulli(P_{ij}) for $i > j$ ($A_{ij} = A_{ji}$ and $A_{ii} = 0$).*

If each block $P^{(kl)}$ can be written as the outer product of two vectors:

$$P^{(kl)} = \lambda^{(kl)} (\lambda^{(lk)})^\top \tag{5}$$

for a set of K^2 fixed vectors $\{\lambda^{(st)}\}_{s,t=1}^K$ where each $\lambda^{(st)}$ is a column vector of dimension n_s, then graph G and its corresponding adjacency matrix A is a realization of a popularity adjusted block model with parameters $\{\lambda^{(st)}\}_{s,t=1}^K$.

We will use the notation $A \sim \text{PABM}(\{\lambda^{(kl)}\}_K)$ to denote a random adjacency matrix A drawn from a PABM with parameters $\lambda^{(kl)}$ consisting of K underlying communities.

It is trivial to show that the PABM, as well as all graphs such that the adjacency matrix is drawn such that $A_{ij} \sim Bernoulli(P_{ij})$, is a special case of the GRDPG. It can also be shown that the latent positions of the PABM under the GRDPG framework consists of K K-dimensional subspaces in $\mathbb{R}^{K^2}$. While there is no unique latent configuration X such that $XX^\top = P$, the edge probability P for the PABM, they all have this subspace structure, and one in particular consists of *orthogonal* subspaces.

Theorem 1 (Connecting the PABM to the GRDPG for $K = 2$)**.** *Let*

$$X = \begin{bmatrix} \lambda^{(11)} & \lambda^{(12)} & 0 & 0 \\ 0 & 0 & \lambda^{(21)} & \lambda^{(22)} \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

where each $\lambda^{(kl)}$ *is a vector as in Definition 1. Then* $A \sim GRDPG_{3,1}(XU)$ *and* $A \sim PABM(\{(\lambda^{(kl)}\}_2)$ *are equivalent.*

Theorem 2 (Generalization to $K > 2$)**.** *There exists a block diagonal matrix* $X \in \mathbb{R}^{n \times K^2}$ *defined by PABM parameters* $\{\lambda^{(kl)}\}_K$ *and orthonormal matrix* $U \in \mathbb{R}^{K^2 \times K^2}$ *that is fixed for each* K *such that* $A \sim GRDPG_{K(K+1)/2, K(K-1)/2}(XU)$ *and* $A \sim PABM(\{(\lambda^{(kl)}\})_K)$ *are equivalent.*

Proof. Define the following matrices from $\{\lambda^{(kl)}\}_K$:

$$\Lambda^{(k)} = \begin{bmatrix} \lambda^{(k,1)} & \cdots & \lambda^{(k,K)} \end{bmatrix} \in \mathbb{R}^{n_k \times K}$$

$$X = \text{blockdiag}(\Lambda^{(1)}, ..., \Lambda^{(K)}) \in \mathbb{R}^{n \times K^2} \tag{6}$$

$$L^{(k)} = \text{blockdiag}(\lambda^{(1k)}, ..., \lambda^{(Kk)}) \in \mathbb{R}^{n \times K}$$

$$Y = \begin{bmatrix} L^{(1)} & \cdots & L^{(K)} \end{bmatrix} \in \mathbb{R}^{n \times K^2}$$

Then $P = XY^\top$.

Similar to the $K = 2$ case, we have $Y = X\Pi$ for a permutation matrix Π, resulting in $P = X\Pi X^\top$. The permutation described by Π has K fixed points, which correspond to K eigenvalues equal to 1 with corresponding eigenvectors e_k where $k = r(K+1)+1$ for $r = 0, ..., K-1$. It also has $\binom{K}{2} = K(K-1)/2$ cycles of order 2. Each cycle corresponds to a pair of eigenvalues $+1$ and -1 and a pair of eigenvectors $(e_s + e_t)/\sqrt{2}$ and $(e_s - e_t)/\sqrt{2}$.

Then Π has $K(K+1)/2$ eigenvalues equal to 1 and $K(K-1)/2$ eigenvalues equal to -1. Π has the decomposed form

$$\Pi = U I_{K(K+1)/2, K(K-1)/2} U^\top \tag{7}$$

The edge probability matrix then can be written as:

$$P = XUI_{p,q}(XU)^\top \tag{8}$$

$$p = K(K+1)/2 \tag{9}$$

$$q = K(K-1)/2 \tag{10}$$

and we can describe the PABM with K communities as a GRDPG with latent positions XU with signature $\big(K(K+1)/2, K(K-1)/2\big)$. □

Example $(K = 3)$**.** *Using the same notation as in Theorem 2:*

$$X = \begin{bmatrix} \lambda^{(11)} & \lambda^{(12)} & \lambda^{(13)} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda^{(21)} & \lambda^{(22)} & \lambda^{(23)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda^{(31)} & \lambda^{(32)} & \lambda^{(33)} \end{bmatrix}$$

$$Y = \begin{bmatrix} \lambda^{(11)} & 0 & 0 & \lambda^{(12)} & 0 & 0 & \lambda^{(13)} & 0 & 0 \\ 0 & \lambda^{(21)} & 0 & 0 & \lambda^{(22)} & 0 & 0 & \lambda^{(23)} & 0 \\ 0 & 0 & \lambda^{(31)} & 0 & 0 & \lambda^{(32)} & 0 & 0 & \lambda^{(33)} \end{bmatrix}$$

Then $P = XY^\top$ and $Y = X\Pi$ where Π is a permutation matrix consisting of 3 fixed points and 3 cycles of order 2:

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

** Positions 1, 5, 9 are fixed.*

** The cycles of order 2 are $(2, 4)$, $(3, 7)$, and $(6, 8)$.*

Therefore, we can decompose $\Pi = UI_{6,3}U^\top$ where the first three columns of U consist of e_1, e_5, and e_9 corresponding to the fixed positions 1, 5, and 9, the next three columns consist of eigenvectors $(e_k + e_l)/\sqrt{2}$, and the last three columns consist of eigenvectors $(e_k - e_l)/\sqrt{2}$, where pairs (k, l) correspond to the cycles of order 2 described above.

The latent positions are the rows of

$$XU = \begin{bmatrix} \lambda^{(11)} & 0 & 0 & \lambda^{(12)}/\sqrt{2} & \lambda^{(13)}/\sqrt{2} & 0 & \lambda^{(12)}/\sqrt{2} & \lambda^{(13)}/\sqrt{2} & 0 \\ 0 & \lambda^{(22)} & 0 & \lambda^{(21)}/\sqrt{2} & 0 & \lambda^{(23)}/\sqrt{2} & -\lambda^{(21)}/\sqrt{2} & 0 & \lambda^{(23)}/\sqrt{2} \\ 0 & 0 & \lambda^{(33)} & 0 & \lambda^{(31)}/\sqrt{2} & \lambda^{(32)}/\sqrt{2} & 0 & -\lambda^{(31)}/\sqrt{2} & -\lambda^{(32)}/\sqrt{2} \end{bmatrix}$$

This leads to the following theorem.

Theorem 3. *Let $P = V\Lambda V^\top$ be the spectral decomposition of the edge probability matrix of a PABM. Define $B = nVV^\top$. Then $B_{ij} = 0$ $\forall i, j$ in different communities.*

If $\hat{V}$ is the *unscaled* ASE of A, Theorem 3 and results from Rubin-Delanchy et al. together imply $n\hat{V}\hat{V}^\top \stackrel{a.s.}{\to} 0$, leading to the following result:

Theorem 4. *Let $\hat{B}_n$ with entries $\hat{B}_n^{(ij)}$ be the affinity matrix from OSC (Alg. 1). Then $\forall$ pairs (i, j) belonging to different communities and sparsity factor satisfying $n\rho_n = \omega\{(\log n)^{4c}\}$,*

Algorithm 1: Orthogonal Spectral Clustering.

Data: Adjacency matrix A, number of communities K

Result: Community assignments $1, ..., K$

Compute the eigenvectors of A that correspond to the $K(K+1)/2$ most positive eigenvalues and $K(K-1)/2$ most negative eigenvalues. Construct V using these eigenvectors as its columns.

Compute $B = |nVV^\top|$, applying $|\cdot|$ entry-wise.

Construct graph G using B as its similarity matrix.

Partition G into K disconnected subgraphs (e.g., using edge thresholding or spectral clustering).

Map each partition to the community labels $1, ..., K$.

$$\max_{i,j} |n(\hat{v}_n^{(i)})^\top \hat{v}_n^{(j)}| = O_P\Big(\frac{(\log n)^c}{\sqrt{n\rho_n}}\Big) \tag{11}$$

This provides the result that $\forall i, j$ in different communities, $\hat{B}_n^{(ij)} \stackrel{a.s.}{\to} 0$.

Since every ASE of the PABM consists of subspaces and as $n \to \infty$ each point of the ASE approaches its subspace almost surely, SSC should also work for PABM community detection. Combining this with the results by Wang and Xu, which state that if the points lie sufficiently close to their respective subspaces and the cosine of the angles between subspaces is sufficiently small, SDP will hold. We show that the unscaled ASE exhibits exactly these conditions for sufficiently large n.

Theorem 5. *Let P_n describe the edge probability matrix of the PABM with n vertices, and let $A_n \sim Bernoulli(P_n)$. Let $\hat{V}_n$ be the matrix of eigenvectors of A_n corresponding to the $K(K+1)/2$ most positive and $K(K-1)/2$ most negative eigenvalues. Then $\exists \lambda > 0$ and $N \in \mathbb{N}$ such that when $n > N$, $\sqrt{n}\hat{V}_n$ obeys the subspace detection property with probability 1.*

3.2 Manifold Clustering

We would like to extend subspace clustering to manifold clustering. The problem setup is as follows: Suppose that in the latent space $\mathcal{X} \subset \mathbb{R}^d$, sample X of n points lie on a union of K disjoint manifolds with each manifold corresponding to a community. If $A \sim \text{RDPG}(X)$, we wish to recover the community labels (up to permutation) from A.

Similarly, suppose that probability distribution F is described as follows:

1. Define functions $f_1, ..., f_K$ such that $f_k : [0, 1] \mapsto \mathcal{X}$ and $f_k(t) \neq f_l(t)$ $\forall k, l \in [K]$ and $t \in [0, 1]$.
2. Sample labels $z_1, ..., z_n \stackrel{iid}{\sim} Categorical(\pi_1, ..., \pi_K)$.
3. Sample $t_1, ..., t_n \stackrel{iid}{\sim} D$ where D has support $[0, 1]$.
4. Set latent positions $x_i = f_{z_i}(t_i)$ and $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$.

Then if $(A, X) \sim \text{RDPG}(F, n)$ and we observe A, we wish to recover hidden labels $z_1, ..., z_n$.

3.2.1 Affine Subspaces

We will motivate an approach by the following example.

Example. *Let $U_1, ..., U_n \stackrel{iid}{\sim} Uniform(0, 1)$ with order statistics $U_{(1)}, ..., U_{(n)}$. Then $\forall a \in (0, 1)$ and $\delta \in (0, 1)$, $\exists N = N(\delta, a) < \infty$ such that $\forall n \geq N$,*

$$P(\max_i U_{(i+1)} - U_{(i)} \leq a) \geq 1 - \delta/2 \tag{12}$$

Where $N(\delta, a)$ is monotone increasing w.r.t. δ and a. To prove this, we start with the fact that $U_{(i+1)} - U_{(i)} \sim Beta(1, n)$. Then

$$P(U_{(i+1)} - U_{(i)} \leq a) = 1 - (1-a)^n \tag{13}$$

and

$$P(\max_i U_{(i+1)} - U_{(i)} \leq a) \geq (P(U_{(i+1)} - U_{(i)} \leq a))^n = (1 - (1-a)^n)^n \tag{14}$$

This expression is monotone increasing $\forall n \geq N_1$ for some $N_1 < \infty$. Setting $(1-(1-a)^{N_2})^{N_2} \geq 1-\delta/2$, we can solve for a finite N_2. Then $N = \max(N_1, N_2)$.

If we extend this example such that n_1 points are sampled uniformly from the segment $f_1(t) = (t, 0)$ and n_2 points are sampled uniformly from the segment $f_2(t) = (t, a)$ for $t \in [0, 1]$, then a sample of size $N(\delta, a)$ is sufficient to satisfy:

$$\begin{aligned} P(\max_i X_{(i+1)} - X_{(i)} \leq \min_{i,j} \|X_i - Y_j\|) \geq 1 - \delta \\ P(\max_j Y_{(j+1)} - Y_{(j)} \leq \min_{i,j} \|X_i - Y_j\|) \geq 1 - \delta \end{aligned} \tag{15}$$

for X_i in the first segment and Y_j in the second segment and $X_{(i)}$, $Y_{(j)}$ are order statistics in the first coordinate. If each segment corresponds to a community, this leads to the following two results:

1. Single linkage clustering with $K = 2$ will perform perfect community detection with probability at least $1 - \delta$.

2. An ϵ-neighborhood graph with $\epsilon \in (0, a)$ will consists of at least 2 disjoint subgraphs such that no subgraph consists of members of two different communities (analogous to the SDP), with probability at least $1 - \delta$.

We can then further extend this to the case where points are drawn from unit segments with noise.

3.2.2 Mainfold Learning

If instead of sampling uniformly from line segments of unit length, we sample uniformly from a 1 dimensional manifolds of unit length, the above property still holds. Let $U_1, ..., U_n \overset{iid}{\sim} Uniform(0, 1)$ and $f : [0, 1] \mapsto \mathbb{R}^d$ be a smooth function such that $\int_u^v \sqrt{1 + (f'(t))^2} dt = \|u - v\|$. Then $U_{(i+1)} - U_{(i)} \geq \|f(U_{(i+1)}) - f(U_{(i)})\|$, so $P(U_{(i+1)} - U_{(i)} \leq \alpha) \leq P(\|f(U_{(i+1)}) - f(U_{(i)})\| \leq a)$. If the shortest distance between the two manifolds defined by f_1 and f_2 with the same restriction is a, then the same N as before is sufficient, although perhaps a more lenient lower bound can be derived based on the shape of $f_i(\cdot)$.

4 Summary

5 Estimated Timeline of Completion

Literature review: August 2021

Complete proofs of main theorems: January 2022

Simulations and real data analyses: March 2022

Dissertation completion: April 2022

References

[1] Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018. URL http://jmlr.org/papers/v18/17-448.html.

[2] Avanti Athreya, Minh Tang, Youngser Park, and Carey E. Priebe. On estimation and inference in latent structure random graphs, 2020.

[3] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797, 2009. doi: 10.1109/CVPR.2009.5206547.

[4] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131. URL https://doi.org/10.1198/016214502760047131.

[5] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[6] François Lorrain and Harrison C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80, 1971. doi: 10.1080/0022250X.1971.9989788. URL https://doi.org/10.1080/0022250X.1971.9989788.

[7] Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Statist.*, 8(2):2905–2922, 2014. doi: 10.1214/14-EJS978. URL https://doi.org/10.1214/14-EJS978.

[8] Majid Noroozi, Ramchandra Rimal, and Marianna Pensky. Estimation and clustering in popularity adjusted block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a). doi: https://doi.org/10.1111/rssb.12410. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12410.

[9] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph, 2017.

[10] Srijan Sengupta and Yuguo Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(2):365–386, March 2018. ISSN 1369-7412. doi: 10.1111/rssb.12245.

[11] Michael W. Trosset, Mingyue Gao, Minh Tang, and Carey E. Priebe. Learning 1-dimensional submanifolds for subsequent inference on random dot product graphs, 2020.

[12] Ulrike von Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007. URL http://arxiv.org/abs/0711.0189.

[13] Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. *Journal of Machine Learning Research*, 17(12):1–41, 2016. URL http://jmlr.org/papers/v17/13-354.html.

[14] Stephen J. Young and Edward R. Scheinerman. Random dot product graph models for social networks. In Anthony Bonato and Fan R. K. Chung, editors, *Algorithms and Models for the Web-Graph*, pages 138–149, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-77004-6.