Popularity Adjusted Block Models are Generalized Random Dot Product Graphs

John Koo
Department of Statistics, Indiana University
and
Minh Tang
Department of Statistics, North Carolina State University
and
Michael Trosset
Department of Statistics, Indiana University

Abstract

We connect two random graph models, the Popularity Adjusted Block Model (PABM) and the Generalized Random Dot Product Graph (GRDPG), by demonstrating that a PABM is a GRDPG in which communities correspond to mutually orthogonal subspaces of latent vectors. This insight allows us to construct new algorithms for community detection and parameter estimation with the PABM, as well as improve an existing algorithm that relies on Sparse Subspace Clustering. Using established asymptotic properties of Adjacency Spectral Embedding for the GRDPG, we derive asymptotic properties of these algorithms. In particular, we demonstrate that the absolute number of community detection errors tends to zero as the number of graph vertices tends to infinity. Simulation experiments illustrate these properties.

1 Introduction

Statistical inference on random graphs requires a suitable probability model. A general probability model for unweighted and undirected graphs is the Bernoulli Graph (also known as the inhomogeneous Erös-Rényi model), which assumes that edges occur as independent Bernoulli trials. A Bernoulli Graph is characterized by an edge probability matrix $P = [P_{ij}]$, where an edge between vertices i and j occurs with success probability P_{ij}. A trivial example of a Bernoulli Graph is the Erdös-Rényi model proposed by Gilbert (1959), in which the vertices of the random graph are fixed and possible edges occur independently with fixed probability $P_{ij} = p$. The requirement that p is fixed is too strong for most applications, and various researchers have weakened that requirement in various ways. The present work relates two lines of generalization.

Network analysis is often concerned with community detection. One form of community detection assumes that each vertex belongs to an unobserved community, with the probability of an edge between vertices i and j depending on the communities to which i and j belong. Formally, one assigns each vertex v_i a community label z_i and assumes a Bernoulli Graph in which P_{ij} is a function of z_i and z_j. Such models, called Block Models, define the goal of community detection as a problem in statistical inference: identify the true community (up to permutation of labels) to which each vertex belongs.

The classical Stochastic Block Model (SBM) of Lorrain and White (1971) specifies that each edge probability P_{ij} depends only on the labels z_i and z_j, i.e., $P_{ij} = \omega_{z,z_j}$. Subsequent researchers have weakened this assumption. The Degree-Corrected Block Model (DCBM) of Karrer and Newman (2011) assigns an additional parameter θ_i to each vertex and sets $P_{ij} = \theta_i\theta_j\omega_{z,z_j}$. The Popularity Adjusted Block Model (PABM) of Sengupta and Chen (2018) generalizes the DCBM, allowing heterogeneity of edge probabilities within and between communities while still maintaining distinct community structure.

Another type of Bernoulli Graph was proposed by Young and Scheinerman (2007). A Random Dot Product Graph (RDPG) specifies that each vertex corresponds to a latent position vector in Euclidean space and that the probability of an edge between two vertices is

the dot product of their latent position vectors. Thus, if the latent positions $x_1, \ldots, x_n \in \mathbb{R}^d$ and $X = [\ x_1 \ \cdots \ x_n\ ]^\top$, then the edge probability matrix is $P = XX^\top$. Clearly, any Bernoulli Graph with positive semidefinite P is an RDPG. The positive definite Euclidean inner product in the RDPG model was replaced by an indefinite inner product in Rubin-Delanchy, Priebe, and Tang (2017), resulting in the *Generalized* RDPG (GRDPG).

In contrast to Block Models, neither RDPGs nor GRDPGs inherently specify distinct communities. However, one can easily impose community structure by assuming that the latent positions lie in distinct clusters. Hence, it is not surprising that Block Models can be studied by reformulating them as RDPGs or GRDPGs. For example, an assortative SBM (an SBM for which P is positive semidefinite) is equivalent to an RDPG for which all vertices in the same community correspond to the same latent position vector (Lyzinski et al., 2014). Likewise, the DCBM is equivalent to an RDPG for which all vertices in the same community correspond to latent position vectors that lie on a straight line (Lyzinski et al., 2014; Rubin-Delanchy, Priebe, and Tang, 2017).

Because the edge probability matrix of a PABM is not necessarily positive semidefinite, a PABM is not necessarily an RDPG. In Section 2.3 we demonstrate that every PABM is in fact a specific type of GRDPG for which the latent position vectors lie in distinct orthogonal subspaces, each subspace corresponding to a community. This identification is our central result. In Section 3, we use the geometry of the GRDPG to derive more efficient algorithms for detecting the communities and estimating the parameters in the PABM. We report the results of simulation studies in Section 4 and apply our methods to three well-known data sets in Section 5. Section 6 concludes.

2 PABMs are GRDPGs

In this section, we show that the PABM is a special case of the GRDPG. More specifically, graph G drawn from the PABM can be represented by a collection of latent vectors in Euclidean space. We further show that the latent configuration that induces the PABM consists of orthogonal subspaces with each subspace corresponding to a community.

2.1 Notation and Scope

Let $G = (V, E)$ be an unweighted, undirected, and hollow graph with vertex set V ($|V| = n$) and edge set E. $A \in \{0,1\}^{n \times n}$ represents the adjacency matrix of G such that $A_{ij} = 1$ if there exists an edge between vertices i and j and 0 otherwise. Because G is symmetric and hollow, $A_{ij} = A_{ji}$ and $A_{ii} = 0$ for each $i, j \in [n]$. We further restrict our analyses to Bernoulli graphs. Let $P \in [0,1]^{n \times n}$ be a symmetric matrix of edge probabilities. Graph G is sampled from P by drawing $A_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(P_{ij})$ for each $1 \leq i < j \leq n$ (setting $A_{ji} = A_{ij}$ and $A_{ii} = 0$). We denote $A \sim \text{BernoulliGraph}(P)$ as a graph with adjacency matrix A sampled from edge probability matrix P in this manner. If each vertex has a hidden label in $[K]$, they are denoted as $z_1, ..., z_n$. Finally, we denote $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n \times d}$ as the collection of n latent vectors $x_1, ..., x_n \in \mathbb{R}^d$.

2.2 Two Probability Models for Graphs

Definition 1 (Popularity Adjusted Block Model)**.** Let $P \in [0,1]^{n \times n}$ be a symmetric edge probability matrix for a graph $G = (V, E)$ with adjacency matrix A such that $A \sim \text{Bernoulli}(P)$. Let each vertex have a community label between 1 and K. Then G is drawn from a Popularity Adjusted Block Model if each vertex has K popularity parameters that describe its affinity toward each of the K communities, i.e., vertex i has popularity parameters $\lambda_{i1}, ..., \lambda_{iK}$, and $P_{ij} = \lambda_{iz_j}\lambda_{jz_i}$ for all $i \leq j$.

Another characterization of the PABM is as follows. Let the rows and columns of P be arranged by community label such that $n_k \times n_\ell$ block $P^{(k\ell)}$ describes the edge probabilities between vertices in communities k and ℓ; note that $P^{(k\ell)} = (P^{(k\ell)})^\top$. Now suppose that each block $P^{(k\ell)}$ can be written as the outer product of two vectors:

$$P^{(k\ell)} = \lambda^{(k\ell)}(\lambda^{(\ell k)})^\top \tag{1}$$

for a set of K^2 popularity vectors $\{\lambda^{(st)}\}_{1 \leq s \leq K, 1 \leq t \leq K}$ where each $\lambda^{(st)}$ is a column vector of dimension n_s. Then a graph G is drawn from a PABM with parameters $\{\lambda^{(st)}\}$ if its adjacency matrix A satisfies $A \sim \text{Bernoulli}(P)$.

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

Definition 2 (Generalized Random Dot Product Graph)**.** Let graph $G = (V, E)$ be drawn as $A \sim \text{BernoulliGraph}(P)$. If there exists a data matrix $X \in \mathbb{R}^{n \times d}$ such that

$$P = X I_{p,q} X^\top \tag{2}$$

for some $d, p, q \in \mathbb{N}$ and $p + q = d$, then G is drawn from the Generalized Random Dot Product Graph with latent positions $x_1, ..., x_n \in \mathbb{R}^d$ and signature (p, q).

We will use the notation $A \sim \text{GRDPG}_{p,q}(X)$ to denote a random adjacency matrix A drawn from latent positions X and signature (p, q). If instead of fixed latent positions, they are drawn $X_1, ..., X_n \overset{\text{iid}}{\sim} F$, we denote the GRDPG as $(A, X) \sim \text{GRDPG}_{p,q}(F, n)$.

Remark. We can use Adjacency Spectral Embedding (ASE) (Sussman et al., 2012) to recover the latent vectors of a GRDPG. This procedure consists of taking the spectral decomposition of A (or P if available) and choosing the p most positive and q most negative eigenvalues and their corresponding eigenvectors to construct the embedding $\hat{Z} = V|D|^{1/2}$.

Definition 3 (Indefinite Orthogonal Group)**.** The indefinite orthogonal group with signature (p, q) is the set $\{Q \in \mathbb{R}^{d \times d} : Q I_{p,q} Q^\top = I_{p,q}\}$, denoted as $\mathbb{O}(p, q)$.

Remark. The latent vectors that produce $X I_{p,q} X^\top = P$ are not unique (Rubin-Delanchy et al., 2017). More specifically, if $P_{ij} = x_i^\top I_{p,q} x_j$, then we also have for any $Q \in \mathbb{O}(p, q)$, $(Qx_i)^\top I_{p,q} (Qx_j) = x_i^\top (Q^\top I_{p,q} Q) x_j = x_i^\top I_{p,q} x_j = P_{ij}$. Unlike in the RDPG case, transforming the latent positions via multiplication by $Q \in \mathbb{O}(p, q)$ does not necessarily maintain interpoint angles or distances.

2.3 The Geometry of PABMs

Now that we defined the PABM and GRDPG, we show the special geometry of the PABM when viewed as a GRDPG.

Theorem 1 (The latent configuration of the PABM)**.** *Let* $A \sim \text{PABM}(\{\lambda^{(k\ell)}\}_K)$ *be an instance of a PABM with* $K \geq 1$ *blocks and latent vectors* $\{\lambda^{(k\ell)} \colon 1 \leq k \leq K, 1 \leq \ell \leq K\}$*. Then there exists a block diagonal matrix* $X \in \mathbb{R}^{n \times K^2}$ *defined by* $\{\lambda^{(k\ell)}\}$ *and a* $K^2 \times K^2$ fixed *orthonormal matrix* U *such that* $A \sim \text{GRDPG}_{K(K+1)/2, K(K-1)/2}(XU)$*.*

Proof. We will prove this theorem in two parts. First, for demonstration purposes, we focus on the case for $K = 2$. Then we generalize this to $K \geq 2$.

For the $K = 2$ case, the proof is straightforward. Let

$$X = \begin{bmatrix} \lambda^{(11)} & \lambda^{(12)} & 0 & 0 \\ 0 & 0 & \lambda^{(21)} & \lambda^{(22)} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Then by straightforward matrix multiplication, we obtain

$$XUI_{3,1}U^\top X^\top = \begin{bmatrix} \lambda^{(11)}(\lambda^{(11)})^\top & \lambda^{(12)}(\lambda^{(21)})^\top \\ \lambda^{(21)}(\lambda^{(12)})^\top & \lambda^{(22)}(\lambda^{(22)})^\top \end{bmatrix},$$

which is a GRDPG with latent vectors described by XU.

It is nevertheless instructive to look at a few intermediate steps. More specifically, the product $UI_{3,1}U^\top$ yields a permutation matrix Π with fixed points at positions 1 and 4 and a cycle of order 2 swapping positions 2 and 3, i.e.,

$$\Pi = UI_{3,1}U^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Furthermore, since U is orthonormal and $I_{3,1}$ is diagonal, $UI_{3,1}U^\top$ is also an eigendecomposition of Π where the fixed points of Π are mapped to the eigenvectors e_1 and e_4 while

the cycles of order two are mapped to the eigenvectors $\frac{1}{\sqrt{2}}(e_2 + e_3)$ and $\frac{1}{\sqrt{2}}(e_2 - e_3)$; here e_i denote the i^{th} basis vector in $\mathbb{R}^4$.

For the general case, we first define the following matrices

$$\Lambda^{(k)} = \left[\lambda^{(k1)} \mid \cdots \mid \lambda^{(kK)}\right] \in \mathbb{R}^{n_k \times K}, \quad X = \text{blockdiag}(\Lambda^{(1)}, \ldots, \Lambda^{(K)}) \in \mathbb{R}^{n \times K^2}, \tag{3}$$

$$L^{(k)} = \text{blockdiag}(\lambda^{(1k)}, \ldots, \lambda^{(Kk)}) \in \mathbb{R}^{n \times K}, \quad Y = \left[L^{(1)} \mid \cdots \mid L^{(K)}\right] \in \mathbb{R}^{n \times K^2}. \tag{4}$$

It is then straighforward to verify that

$$XY^\top = \text{blockdiag}(\Lambda^{(1)}, \ldots, \Lambda^{(K)}) \begin{bmatrix} L_1^\top \\ \vdots \\ L_K^\top \end{bmatrix} = \begin{bmatrix} \Lambda^{(1)}(L^{(1)})^\top \\ \vdots \\ \Lambda^{(K)}(L^{(K)})^\top \end{bmatrix}$$

$$\Lambda^{(k)}(L^{(k)})^\top = \left[\lambda^{(k1)}(\lambda^{(1k)})^\top \mid \cdots \mid \lambda^{(kK)}(\lambda^{(Kk)})^\top\right] = \left[P^{(k1)} \mid P^{(k2)} \mid \cdots \mid P^{(kK)}\right].$$

We therefore have $P = XY^\top$. Similar to the $K = 2$ case, we also have $Y = X\Pi$ for some permutation matrix Π and hence $P = X\Pi X^\top$. 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 $\{-1, +1\}$ and a pair of eigenvectors $\{(e_s + e_t)/\sqrt{2}, (e_s - e_t)/\sqrt{2}\}$.

Let $p = K(K+1)/2$ and $q = K(K-1)/2$. We therefore have

$$\Pi = U I_{K(K+1)/2, K(K-1)/2} U^\top \tag{5}$$

where U is a $K^2 \times K^2$ orthogonal matrix and hence

$$P = XU I_{p,q} (XU)^\top. \tag{6}$$

In summary we can describe the PABM with K communities as a GRDPG with latent positions XU and signature $(p, q) = \left(\frac{1}{2}K(K+1), \frac{1}{2}K(K-1)\right)$. □

Example 1. Let A be a 3 blocks PABM with latent vectors $\{\lambda^{(k\ell)} \colon 1 \leq k \leq 3, 1 \leq \ell \leq 3\}$. Using the same notation as in Theorem 1, we can define

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

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

Then $Y = X\Pi$ and $P = XY^\top$ where Π is a 9×9 permutation matrix of the form

$$\Pi = \Big[e_1 \mid e_4 \mid e_7 \mid e_2 \mid e_5 \mid e_8 \mid e_3 \mid e_6 \mid e_9 \Big].$$

where e_i denotes the i^{th} basis vector in $\mathbb{R}^9$. The matrix Π corresponds to a permutation of $\{1, 2, \ldots, 9\}$ with the following decomposition.

1. Positions 1, 5, 9 are fixed.
2. There are three cycles of length 2, namely $(2, 4)$, $(3, 7)$, and $(6, 8)$.

We can therefore write Π as $\Pi = UI_{6,3}U^\top$ where the first three columns of U consist of e_1, e_5, and e_9 corresponding to the *fixed* points, the next three columns consist of eigenvectors $(e_k + e_\ell)/\sqrt{2}$, and the last three columns consist of eigenvectors $(e_k - e_\ell)/\sqrt{2}$ for $(k, \ell) \in \{(2, 4), (3, 7), (6, 8)\}$.

The matrix P is then the edge probabilities matrix for a Generalized Random Dot Product Graph whose latent positions are the rows of the matrix

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

3 Algorithms

Two inference objectives arise from the PABM:

1. Community membership identification (up to permutation).
2. Parameter estimation (estimating $\lambda^{(k\ell)}$'s).

In our methods, we assume that K, the number of communities, is known beforehand and does not require estimation.

3.1 Previous Work

Sengupta and Chen (2018) used Modularity Maximization (MM) and the Extreme Points (EP) algorithm (Le, Levina, and Vershynin, 2016) for community detection and parameter estimation. They were able to show that as the sample size increases, the proportion of misclassified community labels (up to permutation) goes to 0.

Noroozi, Rimal, and Pensky (2021) used Sparse Subspace Clustering (SSC) (Elhamifar and Vidal, 2009) for community detection in the PABM. The SSC algorithm can be described as follows: Given $X \in \mathbb{R}^{n\times d}$ with vectors $x_i^\top \in \mathbb{R}^d$ as rows of X, the optimization problem $c_i = \arg\min_c ||c||_1$ subject to $x_i = Xc$ and $c^{(i)} = 0$ is solved for each $i = 1, ..., n$. The solutions are collected collected into matrix $C = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}^\top$ to construct an affinity matrix $B = |C| + |C^\top|$. If each x_i lie perfectly on one of K subspaces, B describes an undirected graph consisting of K disjoint subgraphs, i.e., $B_{ij} = 0$ if x_i, x_j lie on different subspaces. If X instead represents points near K subspaces with some noise, a final graph partitioning step may be required (e.g., edge thresholding or spectral clustering).

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

$$c_i = \arg\min_c \frac{1}{2}||x_i - X_{-i}c||_2^2 + \lambda||c||_1 \tag{7}$$

for some sparsity parameter $\lambda > 0$. The c_i vectors are then collected into C and B as before.

Definition 4 (Subspace Detection Property)**.** Let $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ be noisy points sampled from K subspaces, i.e., $x_i = y_i + z_i$ where the y_i belongs to the union of K subspaces and the z_i are noise vectors. Let $\lambda \geq 0$ be given and let C and B be constructed from the solutions of LASSO problems as described in Eq. (7) with this given choice of λ. Then X is

said to satisfy the subspace detection property with sparsity parameter λ if each column of C has nonzero norm and $B_{ij} = 0$ whenever y_i and y_j are from different subspaces.

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

Theorem 1 suggests that SSC is appropriate for community detection for the PABM. More precisely, Theorem 1 says that each community consists of a K-dimensional subspace, and together the subspaces lie in $\mathbb{R}^{K^2}$. The natural approach then is to perform SSC on the ASE of P or A. Noroozi, Rimal, and Pensky (2021) instead applied SSC to P and A, foregoing embedding altogether.

Using results from Soltanolkotabi and Candés (2012), it can be easily shown that the subspace detection property holds for XU, which is an ASE of P. More specifically, if points lie exactly on mutually orthogonal subspaces, then the subspace detection property will hold with probability 1, and this is exactly the case for the PABM (Theorem 1). Much of our work is then built on Rubin-Delanchy et al. (2017), who describe the convergence behavior of the ASE of A to the ASE of P, and Wang and Xu (2016), who show the necessary conditions for the subspace detection property to hold in noisy cases where the points lie near subspaces.

3.2 Algorithms for Community Detection

We previously stated one possible set of latent positions that result in the edge probability matrix of a PABM, $P = (XU)I_{p,q}(XU)^\top$. If we have (or can estimate) XU directly, then both the community detection and parameter identification problem are trivial since U is

Algorithm 1: Orthogonal Spectral Clustering.

Data: Adjacency matrix A, number of communities K

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

1 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.

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 (e.g., using edge thresholding or spectral clustering).

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

orthonormal and fixed for each value of K. However, direct identification or estimation of XU is not possible (Rubin-Delanchy et al., 2017).

If we decompose $P = ZI_{p,q}Z^\top$, then $\exists\ Q \in \mathbb{O}(p,q)$ such that $XU = ZQ$. Even if we start with the exact edge probability matrix, we cannot recover the "original" latent positions XU. Note that unlike in the case of the RDPG, Q is not necessarily an orthogonal matrix. If z_i's are the rows of XU, then $||z_i - z_j||^2 \neq ||Qz_i - Qz_j||^2$, and $\langle z_i, z_j\rangle \neq \langle Qz_i, Qz_j\rangle$. This prevents us from using the orthogonality property of XU directly.

The explicit form of XU represents points in $\mathbb{R}^{K^2}$ such that points within each community lie on K-dimensional orthogonal subspaces. Multiplication by $Q \in \mathbb{O}(p,q)$ removes the orthogonality property but retains the property that each community is represented by a K-dimensional subspace. Therefore, the ASE of P still consists of subspaces that correspond to communities of the PABM, even if they are not orthogonal, suggesting the use of SSC. Before exploring SSC, we will first consider a different approach.

Theorem 2. *Let $P = VDV^\top$ be the spectral decomposition of the edge probability matrix. Let $B = nVV^\top$. Then $B_{ij} = 0$ if vertices i and j are from different communities.*

Proof. We first show that $VV^\top = X(X^\top X)^{-1}X^\top$ where X is defined as in Eq.(3). Indeed,

by Theorem 2, $P = XUI_{p,q}U^\top X^\top$ for $p = K(K+1)/2$ and $q = K(K-1)/2$. The eigendecomposition $P = VDV^\top$ also yields $P = V|D|^{1/2}I_{p,q}|D|^{1/2}V^\top$ where $|\cdot|^{1/2}$ is applied entry-wise. Now let $Y = XU$ and $Z = V|D|^{1/2}$; note that Y and Z both have full column ranks. Since $P = YI_{p,q}Y^\top = ZI_{p,q}Z^\top$, we have

$$Y = ZI_{p,q}Z^\top Y(Y^\top Y)^{-1}I_{p,q}.$$

Let $Q = I_{p,q}Z^\top Y(Y^\top Y)^{-1}I_{p,q}$ and note that $Y = ZQ$. We then have

$$\begin{aligned} Q^\top I_{p,q}Q &= I_{p,q}(Y^\top Y)^{-1}Y^\top ZI_{p,q}I_{p,q}I_{p,q}Z^\top Y(Y^\top Y)^{-1}I_{p,q} \\ &= I_{p,q}(Y^\top Y)^{-1}Y^\top YI_{p,q}Y^\top Y(Y^\top Y)^{-1}I_{p,q} = I_{p,q} \end{aligned}$$

and hence Q is an indefinite orthogonal matrix.

Let $R = UQ|D|^{-1/2}$ and note that $V = XR$. Since R is *invertible*, we have

$$X(X^\top X)^{-1}X^\top = XR(R^\top X^\top XR)^{-1}R^\top X^\top.$$

Furthermore, as V has orthonormal columns, $R^\top X^\top XR = V^\top V = I$. We thus conclude

$$X(X^\top X)^{-1}X^\top = V(V^\top V)^{-1}V^\top = VV^\top$$

as desired.

Recall that X is block diagonal with each block corresponding to one community and hence $X(X^\top X)^{-1}X^\top$ is also a block diagonal matrix with each block corresponding to a community. As $B = nVV^\top = nX(X^\top X)^{-1}X^\top$, we conclude that $B_{ij} = 0$ whenever vertices i and j belong to different communities. □

Theorem 2 provides perfect community detection given P. Letting $|B|$ be the affinity matrix for graph G', G' is partitioned into at least K disjoint subgraphs since each of the K communities have no edges between them. Similar to the subspace detection property, it could be the case that some of the communities are represented by multiple disjoint subgraphs in G', in which case additional reconstruction is required to identify the communities exactly.

Using A instead of P introduces error, which converges to 0 almost surely. We use this fact to motivate Orthogonal Spectral Clustering (Alg. 1). In order to show that OSC is

consistent, we introduce the sparsity factor $\rho_n \in (0, 1]$, which can be characterized as the average expected degree of the vertices divided by the number of vertices. The situation is the same as in Sengupta and Chen, 2018 (PABM) and Rubin-Delanchy et al., 2017 (GRDPG) in which the popularity vectors are scaled as $\rho_n^{1/2}\lambda^{(k\ell)}$. For ease of notation, we restrict our results to the case in which all nonzero popularity parameters are greater than some positive minimum value.

Theorem 3. *Let $\hat{B}$ with entries $\hat{B}_{ij}$ be the affinity matrix from OSC (Alg. 1). Then $\forall$ pairs (i, j) belonging to different communities and sparsity factor satisfying $n\rho_n = \omega\Big((\log n)^{4c}\Big)$,*

$$\max_{i,j} |n\hat{v}_i^\top \hat{v}_j| = O_P\Big(\frac{(\log n)^c}{\sqrt{n\rho_n}}\Big) \tag{8}$$

This provides the result that for i, j in different communities, $\hat{B}_{ij} \xrightarrow{a.s.} 0$.

Theorems 1, 2, and 3 also provide a natural path toward using SSC for community detection. We established in Theorem 1 that an ASE of the edge probability matrix P can be constructed from a latent vector configuration consisting of orthogonal subspaces. Theorem 2 shows how this property can be recovered from the eigenvectors of P. Then Theorem 3 shows that replacing P with A approximates these properties with error that goes to zero as $n \to \infty$. This allows us to obtain the subspace detection property in Theorem 4.

Theorem 4. *Let P describe the edge probability matrix of the PABM with n vertices, and let $A \sim$ Bernoulli(P). Let $\hat{V}$ be the matrix of eigenvectors of A corresponding to the K^2 largest eigenvalues in modulus. Then for all $\epsilon > 0$ there exists a choice of $\lambda > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$, $\sqrt{n}\hat{V}$ obeys the subspace detection property with probability at least $1 - \epsilon$.*

3.3 Algorithms for Parameter Estimation

For any edge probability matrix P for the PABM such that the rows and columns are organized by community, the kl^{th} block is an outer product of two vectors, i.e., $P^{(k\ell)} = \lambda^{(k\ell)}(\lambda^{(\ell k)})^\top$. Therefore, given $P^{(k\ell)}$, $\lambda^{(k\ell)}$ and $\lambda^{(\ell k)}$ are solvable up to multiplicative constant

Algorithm 2: Sparse Subspace Clustering using LASSO.

Data: Adjacency matrix A, number of communities K, hyperparameter λ

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

Find V, the matrix of eigenvectors of A corresponding to the $K(K+1)/2$ most positive and the $K(K-1)/2$ most negative eigenvalues.

Normalize $V \leftarrow \sqrt{n}V$.

for $i = 1, ..., n$ **do**

  Assign $v_i^\top$ as the i^{th} row of V. Assign $V_{-i} = \begin{bmatrix} v_1 & \cdots & v_{i-1} & v_{i+1} & \cdots & v_n \end{bmatrix}^\top$.

  Solve the LASSO problem $c_i = \arg\min_\beta \frac{1}{2}||v_i - V_{-i}\beta||_2^2 + \lambda||\beta||_1$.

  Assign $\tilde{c}_i = \begin{bmatrix} c_i^{(1)} & \cdots & c_i^{(i-1)} & 0 & c_i^{(i)} & \cdots & c_i^{(n-1)} \end{bmatrix}^\top$ such that the superscript is the index of $\tilde{c}_i$.

end

Assign $C = \begin{bmatrix} \tilde{c}_1 & \cdots & \tilde{c}_n \end{bmatrix}$.

Compute the affinity matrix $B = |C| + |C^\top|$.

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 $1, ..., K$.

using singular value decomposition. More specifically, let $P^{(k\ell)} = (\sigma^{(k\ell)})^2 u^{(k\ell)} (v^{(k\ell)})^\top$ be the singular value decomposition of $P^{(k\ell)}$. $u^{(k\ell)} \in \mathbb{R}^{n_k}$ and $v^{(k\ell)} \in \mathbb{R}^{n_\ell}$ are vectors and $\sigma^{(k\ell)}$ is a scalar. Then $\lambda^{(k\ell)} = s_1 u^{(k\ell)}$ and $\lambda^{(\ell k)} = s_2 v^{(k\ell)}$ for unidentifiable $s_1 s_2 = (\sigma^{(k\ell)})^2$. Since each $\lambda^{(k\ell)}$ is not strictly identifiable, we instead estimate each $\tilde{\lambda}^{(k\ell)} = \sigma^{(k\ell)} u^{(k\ell)}$. Given the adjacency matrix A instead of edge probability matrix P, we can simply use plug-in estimators by taking the SVD of each $A^{(k\ell)}$ to obtain $\hat{\lambda}^{(k\ell)} = \hat{\sigma}^{(k\ell)} \hat{u}^{(k\ell)}$ using the largest signular value of A and its corresponding singular vectors.

Theorem 5. *Let each $\tilde{\lambda}^{(k\ell)}$ be the popularity vector derived from its corresponding $P^{(k\ell)}$, and let $\hat{\lambda}^{(k\ell)}$ be its estimate from $A^{(k\ell)}$ (Alg. 3). Then if $n\rho_n = \omega\Big((\log n)^{4c}\Big)$,*

Algorithm 3: PABM parameter estimation.

Data: Adjacency matrix A, community assignments $1, ..., K$

Result: PABM parameter estimates $\{\hat{\lambda}^{(k\ell)}\}_K$.

Arrange the rows and columns of A by community such that each $A^{(k\ell)}$ block consists of estimated edge probabilities between communities k and l.

for $k, l = 1, ..., K,\ k \leq l$ **do**

  Compute $A^{(k\ell)} = U\Sigma V^\top$, the SVD of the kl^{th} block.

  Assign $u^{(k\ell)}$ and $v^{(k\ell)}$ as the first columns of U and V. Assign $(\sigma^{(k\ell)})^2 \leftarrow \Sigma_{11}$.

  Assign $\hat{\lambda}^{(k\ell)} \leftarrow \pm\sigma^{(k\ell)}u^{(k\ell)}$ and $\hat{\lambda}^{(\ell k)} \leftarrow \pm\sigma^{(k\ell)}v^{(k\ell)}$.

end

$$\max_{k,l\in\{1,...,K\}} \|\hat{\lambda}^{(k\ell)} - \tilde{\lambda}^{(k\ell)}\|_\infty = O\left(\frac{(\log n_k)^c}{\sqrt{n_k}}\right) \tag{9}$$

with high probability.

4 Simulation Study

For each simulation, community labels are drawn from a multinomial distribution, the popularity vectors $\{\lambda^{(k\ell)}\}_K$ are drawn from two types of joint distributions depending on whether $k = \ell$, the edge probability matrix P is constructed using the popularity vectors, and finally the adjacency matrix A is drawn from P. OSC is then used for community detection, and this method is compared against SSC (Noroozi, Rimal, and Pensky, 2021; Soltanolkotabi, Elhamifar, and Candès, 2014) and MM (Csardi and Nepusz, 2006; Sengupta and Chen, 2018). True community labels are used with Algorithm 3 to estimate the popularity vectors $\{\lambda^{(k\ell)}\}_K$, and this method is then compared against an MLE-based estimator described in Noroozi, Rimal, and Pensky (2021) and Sengupta and Chen (2018).

Modularity Maximization is NP-hard, so Sengupta and Chen (2018) used the Extreme Points (EP) algorithm (Le, Levina, and Vershynin, 2016), which is $O(n^{K-1})$, as a greedy relaxation of the optimization problem. For these simulations, the Louvain algorithm was

used for modularity maximization, as Sengupta and Chen (2018)'s implementation proved to be prohibitively computationally expensive for $K > 2$. For $K = 2$, it was verified that the Louvain algorithm produces comparable results to EP-MM.

Two implementations of SSC are shown here. The first method, denoted as SSC-A, treats the columns of the adjacency matrix A as points in $\mathbb{R}^n$, as described in Noroozi, Rimal, and Pensky (2021). The second method, denoted as SSC-ASE, first embeds A and then performs SSC on the embedding, as described in Algorithm 2. The sparsity parameter λ was chosen via a preliminary cross-validation experiment. For the final clustering step, a Gaussian Mixture Model was fit on the normalized Laplacian eigenmap of the affinity matrix B.

For comparing methods, we define the community detection error as:

$$L_c(\hat{\sigma}, \sigma; \{v_i\}) = \min_{\pi} \sum_i I(\pi \circ \hat{\sigma}(v_i) = \sigma(v_i))$$

where $\sigma(v_i)$ is the true community label of vertex v_i, $\hat{\sigma}(v_i)$ is the predicted label of v_i, and π is a permutation operator. This is effectively the "misclustering count" of clustering function $\hat{\sigma}$.

We also define two types of parameter estimation error. First, we estimate the popularity vectors directly and compute the RMSE. In this case, the "true" popularity vectors are derived from taking the SVD of each edge probability block $P^{(k\ell)}$ to avoid the unidentifiable multiplicative constants.

$$\text{RMSE}(\{\hat{\lambda}^{(k\ell)}\}, \{\tilde{\lambda}^{(k\ell)}\}) = \sqrt{\frac{1}{n} \sum_{k<l} \min_{s=\pm 1} \|s\hat{\lambda}^{(k\ell)} - \tilde{\lambda}^{(k\ell)}\|_2^2}$$

We can also avoid the unidentifiable multiplicative constant more directly by reconstructing each $\hat{P}^{(k\ell)} = \hat{\lambda}^{(k\ell)}(\hat{\lambda}^{(\ell k)})^\top$, which we use to define another parameter estimation error.

$$\text{RMSE}(\hat{P}, P) = \sum_{k,l} \sqrt{\frac{1}{n_k n_\ell} \|P^{(k\ell)} - \hat{P}^{(k\ell)}\|_F^2}$$

4.1 Balanced Communities

In each simulation, community labels $z_1, ..., z_n$ were drawn from a multinomial distribution with mixture parameters $\{\alpha_1, ..., \alpha_K\}$, then $\{\lambda^{(k\ell)}\}_K$ according to the drawn community labels, P was constructed using the drawn $\{\lambda^{(k\ell)}\}_K$, and A was drawn from P.

For these examples, we set the following parameters:

- Number of vertices $n = 128, 256, 512, 1024, 2048, 4096$
- Number of underlying communities $K = 2, 3, 4$
- Mixture parameters $\alpha_k = 1/K$ for $k = 1, ..., K$, (i.e., each community label has an equal probability of being drawn)
- Community labels $z_k \overset{\text{iid}}{\sim} \text{Multinomial}(\alpha_1, ..., \alpha_K)$
- Within-group popularities $\lambda^{(kk)} \overset{\text{iid}}{\sim} \text{Beta}(2, 1)$
- Between-group popularities $\lambda^{(k\ell)} \overset{\text{iid}}{\sim} \text{Beta}(1, 2)$ for $k \neq \ell$

50 simulations were performed for each (n, K) pair.

Fig. 1 shows the number of mislabeled vertices going to 0 for large n. SSC on both the embedding and on the adjacency matrix produces similar trends for $K > 2$. Weaker performance of SSC for $K = 2$ can be attributed to the final spectral clustering step of the affinity matrix. While the subspace detection property is guaranteed for large n, in our simulations, setting the sparsity parameter to the required value resulted in more than K disconnected subgraphs. We instead chose a smaller sparsity parameter, necessitating a final clustering step. A GMM was fit to the normalized Laplacian eigenmap of B, but visual inspection suggests that the communities are not distributed as a mixture of Gaussians in the eigenmap. A different choice of mixture distribution may result in better performance.

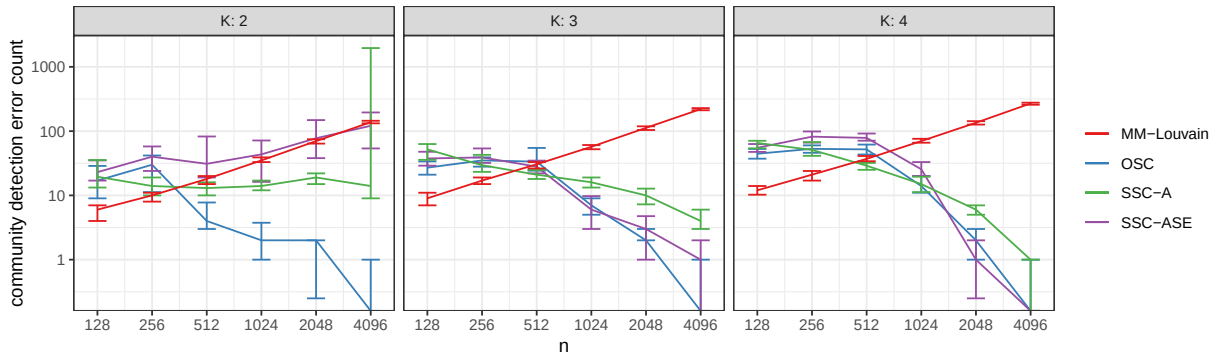

Figure 1: Median and IQR of community detection error. Communities are approximately balanced. Simulations were repeated 50 times for each sample size.

Given ground truth community labels, Algorithm 3 and the MLE-based plug-in estimators perform similarly, with root mean square error decaying at rate approximately $n^{-1/2}$ (Fig. 2). We observe similar behavior when comparing errors in the edge probability blocks (Fig. 3).

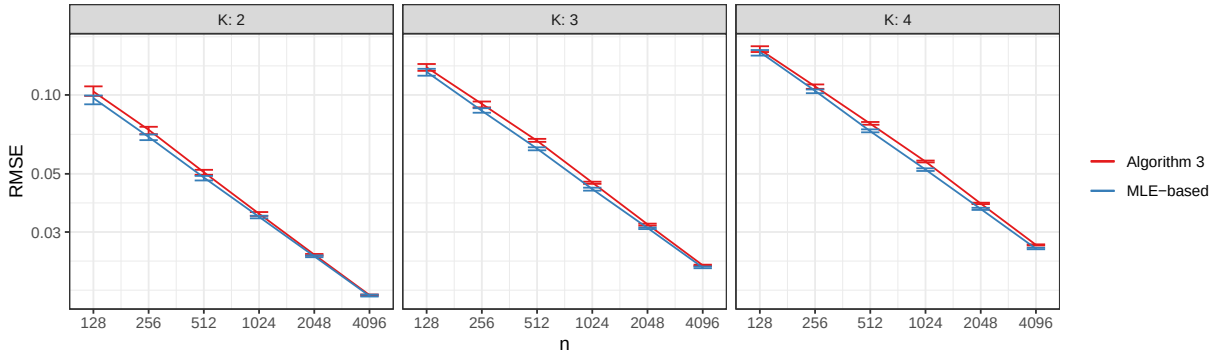

Figure 2: Median and IQR RMSE for popularity vectors from Algorithm 3 (red) compared against an MLE-based method (blue). Simulations were repeated 50 times for each sample size. Communities were drawn to be approximately balanced.

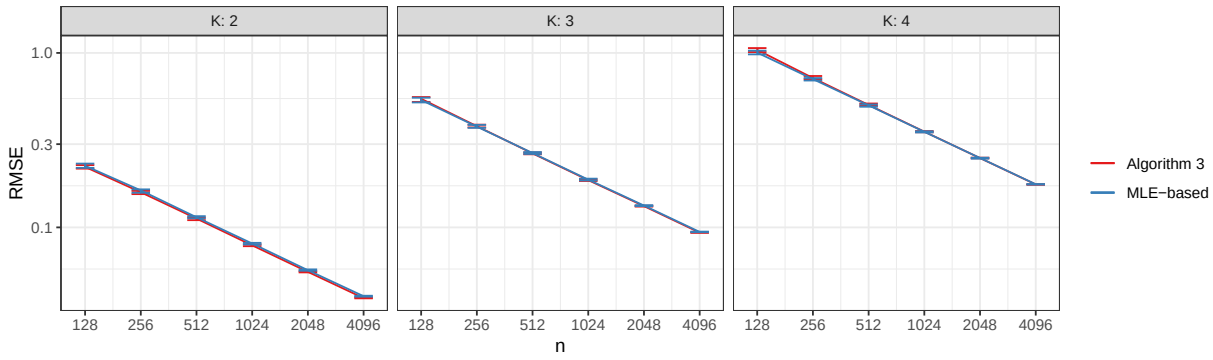

Figure 3: Median and IQR RMSE for edge probability blocks reconstructed from the outputs of Algorithm 3 (red) compared against outputs of an MLE-based method (blue). Simulations were repeated 50 times for each sample size. Communities were drawn to be approximately balanced.

4.2 Imbalanced Communities

Simulations performed in this section are the same as those in the previous section with the exception of the mixture parameters $\{\alpha_1, ..., \alpha_K\}$ used to draw community labels from the multinomial distribution. For these examples, we set the following parameters:

- Number of vertices $n = 128, 256, 512, 1024, 2048, 4096$
- Number of underlying communities $K = 2, 3, 4$
- Mixture parameters $\alpha_k = \frac{k^{-1}}{\sum_{\ell=1}^{K} \ell^{-1}}$ for $k = 1, ..., K$
- Community labels $z_k \overset{\text{iid}}{\sim} \text{Multinomial}(\alpha_1, ..., \alpha_K)$
- Within-group popularities $\lambda^{(kk)} \overset{\text{iid}}{\sim} \text{Beta}(2, 1)$
- Between-group popularities $\lambda^{(k\ell)} \overset{\text{iid}}{\sim} \text{Beta}(1, 2)$ for $k \neq \ell$

50 simulations were performed for each (n, K) pair.

We again see community detection error trending to 0 for OSC, as well as for SSC when $K > 2$ (Fig. 4). Alg. 3 continues to see $n^{-1/2}$ decay in parameter estimation error; however, the MLE-based estimators for the popularity vectors decay at a slower rate (Fig. 5). The reduced rate of decay is removed if we reconstruct the edge probability blocks by taking the

outer products of the popularity vectors (Fig. 6).

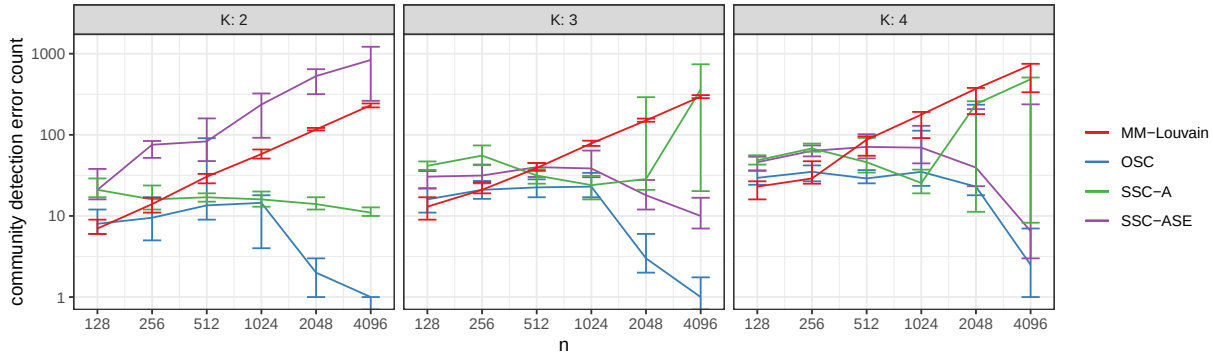

Figure 4: Median and IQR of community detection error. Communities are imbalanced. Simulations were repeated 50 times for each sample size.

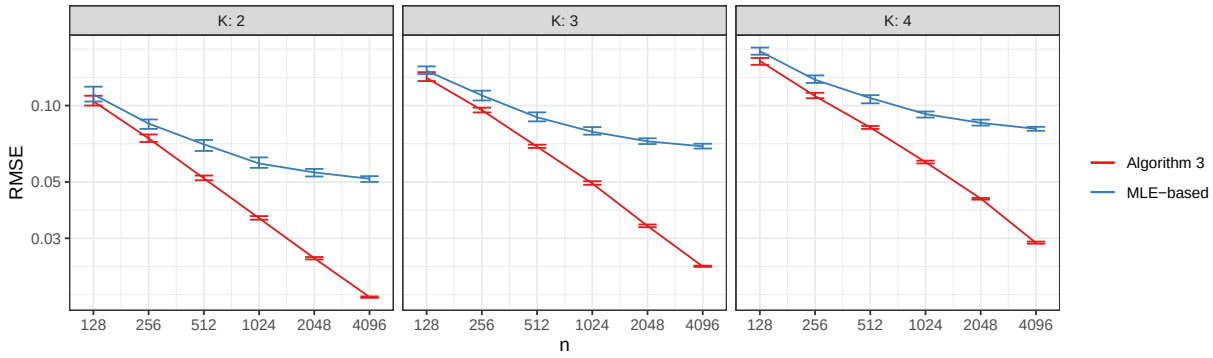

Figure 5: Median and IQR RMSE from Algorithm 3 (red) compared against an MLE-based method (blue). Simulations were repeated 50 times for each sample size. Communities were drawn to be imbalanced.

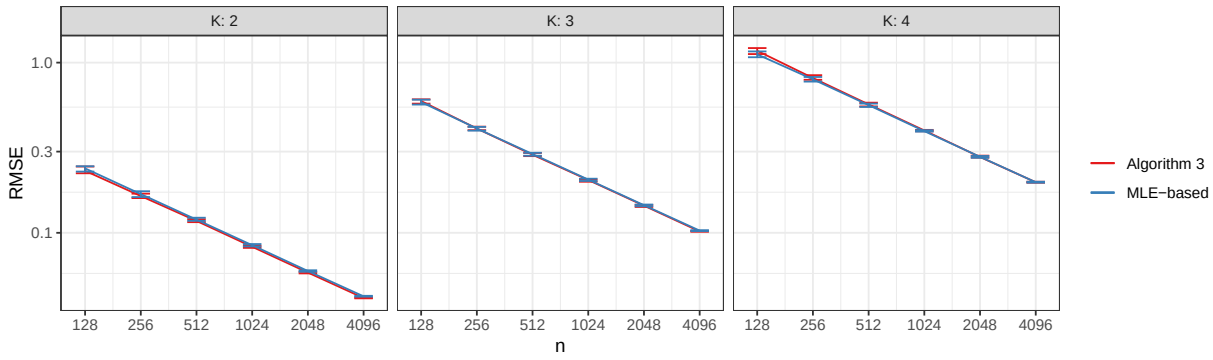

Figure 6: Median and IQR RMSE of edge probability blocks derived from the outputs of Algorithm 3 (red) compared against an MLE-based method (blue). Simulations were repeated 50 times for each sample size. Communities were drawn to be imbalanced.

5 Applications

In the first example, we applied OSC to the Leeds Butterfly dataset (Wang et al., 2018) consisting of visual similarity measurements among 832 butterflies across 10 species. The graph was modified to match the example from Noroozi, Rimal, and Pensky (2021): Only the $K = 4$ most frequent species were considered, and the similarities were discretized to $\{0, 1\}$ via thresholding. Fig. 7 shows a sorted adjacency matrix sorted by the resultant clustering.

Comparing against the ground truth species labels, OSC achieves an adjusted Rand index of 92%, while SSC on the ASE achieves an adjusted Rand index of 96%. In comparison, SSC on the adjacency matrix yields an adjusted Rand index of 73% (Noroozi, Rimal, and Pensky, 2021).

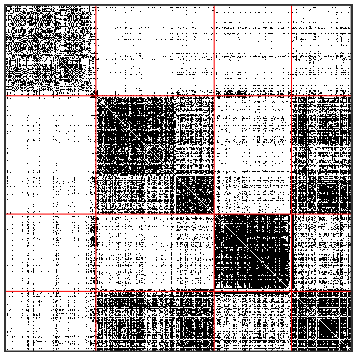

Figure 7: Adjacency matrix of the Leeds Butterfly dataset after sorting by the clustering outputted by OSC.

In the second example, we applied OSC to the British MPs Twitter network (Greene and Cunningham, 2013), the Political Blogs network (Adamic and Glance, 2005), and the DBLP network (Gao et al., 2009; Ji et al., 2010). For this data analysis, we subsetted the data as described in Sengupta and Chen (2018) for their analysis of the same networks. Our methods underperformed compared to modularity maximization, although performance is comparable. In addition, OSC's runtime is much lower than that of modularity maximization.

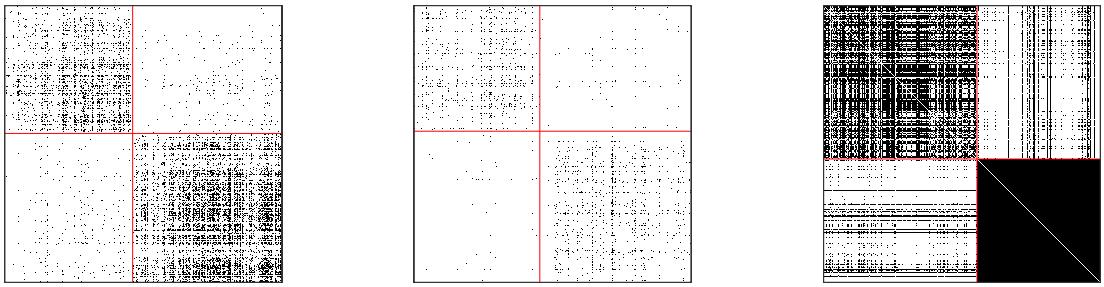

Figure 8: Adjacency matrices of (from left to right) the British MPs, Political Blogs, and DBLP networks after sorting by the clustering outputted by OSC.

In the third example, we analyzed the Karantaka villages data studied by Banerjee et al. (2013). We chose the `visitgo` networks from villages 12, 31, and 46 at the household level. The label of interest is the religious affiliation. The networks were truncated to religions "1"

Table 1: Community detection error rates on the British MPs Twitter, Political Blogs, and DBLP networks using modularity maximization, sparse subspace clustering, and OSC.

Network	MM	SSC-ASE	OSC
British MPs	0.003	0.018	0.009
Political Blogs	0.050	0.196	0.062
DBLP	0.028	0.087	0.059

Table 2: Community detection error rates for identifying household religion.

Network	MM	SSC-ASE	OSC
Village 12	0.270	0.291	0.227
Village 31	0.125	0.066	0.110
Village 46	0.052	0.463	0.078

and "2", and vertices of degree 0 were removed. The results are displayed in Fig. 9 and Table 2.

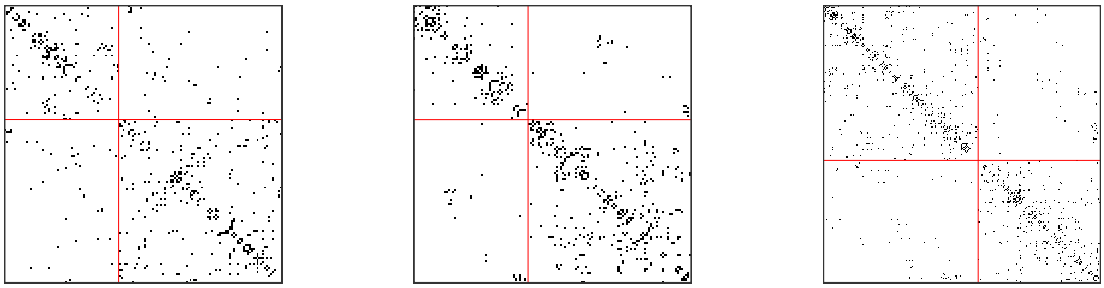

Figure 9: Adjacency matrix of the Karnataka villages data, arranged by the clustering produced by OSC (left). The villages studied here are, from left to right, 12, 31, and 46.

6 Discussion

Our central result states that the Popularity Adjusted Block Model is a special case of the Generalized Random Dot Product Graph. In particular, the PABM with K communities is a GRDPG for which the communities are represented by mutually orthogonal K-dimensional subspaces of the latent space. This result extends previous results that connected the Stochastic Block Model and the Degree Corrected Block Model to Random Dot Product Graphs. Replacing RDPGs with GRDPGs is a critical step in this line of research, as a PABM is not necessarily a RDPG.

Because all Bernoulli Graphs are GRDPGs, it should be possible to invent and study new families of Bernoulli Graphs by characterizing them as special cases of GRDPGs and expliting the latent structures that define them. The present work illustrates the powr of this approach. We recover the latent structure of the PABM by Adjacency Spectral Embedding, then exploit that structure to improve statistical inference. Exploiting the fact that PABM communities correspond to orthogonal subspaces, we propose Orthogonal Spectral Clustering for community detection and demonstrate that the number of misclassified vertices approaches zero with high probability as the size of the graph increases. This is a stronger result than previously proposed algorithms (Sengupta and Chen, 2018), which only guarantee that the error rate (and not count) approaches zero asymptotically. Parameter estimation can be performed in a similar fashion using the ASE, for which we also prove that the per-parameter error approaches zero asymptotically.

A secondary benefit of the GRDPG approach is that the latent structure may be used to improve existing algorithms. For example, one algorithm for PABM community detection (Noroozi, Rimal, and Pensky, 2021) relies on Sparse Subspace Clustering. The latent structure of the PABM provides a natural justification for SSC for the PABM and leads to an improvement over the previous implementation. The improved algorithm applies SSC to the ASE, and we prove that the ASE of the PABM obeys the Subspace Detection Property with high probability if the graph is large.

Finally, one might well inquire what one gains and what one sacrifices by assuming that

a Bernoulli Graph is a PABM. The GRDPG model offers a plausible way to pursue this inquiry. Absent a known latent structure that can be exploited by specialized methods, the GRDPG-ASE approach transforms the problem of network community detection to the much-studied problem of clustering vectors in Euclidean space. Communities of vertices are defined as clusters of latent vectors. After ASE, a standard clustering algorithm, e.g., single linkage, is used ot infer the communities. In future research, we intend to use such general algorithms as baselines and measure the efficiency of the PABM algorithms (and other specialized algorithms) by studying how much they improve on general algorithms when the specified latent structure obtains.

References

Adamic, L. A. and Glance, N. (2005). "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". In: *Proceedings of the 3rd International Workshop on Link Discovery*. LinkKDD '05. Chicago, Illinois: Association for Computing Machinery, 36–43. ISBN: 1595932151. DOI: 10.1145/1134271.1134277.

Banerjee, A. et al. (2013). *The Diffusion of Microfinance*. Version V9. DOI: 10.7910/DVN/U3BIHX.

Cape, J., Tang, M., and Priebe, C. E. (2019). "Signal-plus-noise matrix models: eigenvector deviations and fluctuations". *Biometrika* 106, pp. 243–250.

Csardi, G. and Nepusz, T. (2006). "The igraph software package for complex network research". *InterJournal* Complex Systems, p. 1695.

Elhamifar, E. and Vidal, R. (2009). "Sparse subspace clustering". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2790–2797. DOI: 10.1109/CVPR.2009.5206547.

Gao, J. et al. (2009). "Graph-based Consensus Maximization among Multiple Supervised and Unsupervised Models". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. Curran Associates, Inc., pp. 585–593.

Gilbert, E. N. (1959). "Random Graphs". *The Annals of Mathematical Statistics* 30.4, pp. 1141 –1144. DOI: 10.1214/aoms/1177706098. URL: https://doi.org/10.1214/aoms/1177706098.

Greene, D. and Cunningham, P. (2013). "Producing a Unified Graph Representation from Multiple Social Network Views". *CoRR* abs/1301.5809. arXiv: 1301.5809.

Ji, M. et al. (2010). "Graph Regularized Transductive Classification on Heterogeneous Information Networks". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by J. L. Balcázar et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 570–586. ISBN: 978-3-642-15880-3.

Karrer, B. and Newman, M. E. J. (2011). "Stochastic blockmodels and community structure in networks". *Physical Review E* 83.1. ISSN: 1550-2376. DOI: 10.1103/physreve.83.016107.

Le, C. M., Levina, E., and Vershynin, R. (Feb. 2016). "Optimization via low-rank approximation for community detection in networks". *Ann. Statist.* 44.1, pp. 373–400. DOI: 10.1214/15-AOS1360.

Lorrain, F. and White, H. C. (1971). "Structural equivalence of individuals in social networks". *The Journal of Mathematical Sociology* 1.1, pp. 49–80. DOI: 10.1080/0022250X.1971.9989788.

Lyzinski, V. et al. (2014). "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding". *Electron. J. Statist.* 8.2, pp. 2905–2922. DOI: 10.1214/14-EJS978.

Noroozi, M., Rimal, R., and Pensky, M. (2021). "Estimation and clustering in popularity adjusted block model". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. DOI: https://doi.org/10.1111/rssb.12410.

Rubin-Delanchy, P., Priebe, C. E., and Tang, M. (2017). *Consistency of adjacency spectral embedding for the mixed membership stochastic blockmodel*. arXiv: 1705.04518. arXiv: 1705.04518 [stat.ME].

Rubin-Delanchy, P. et al. (2017). *A statistical interpretation of spectral embedding: the generalised random dot product graph*. arXiv: 1709.05506. arXiv: 1709.05506 [stat.ML].

Sengupta, S. and Chen, Y. (Mar. 2018). "A block model for node popularity in networks with community structure". English (US). *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.2, pp. 365–386. ISSN: 1369-7412. DOI: 10.1111/rssb.12245.

Soltanolkotabi, M. and Candés, E. J. (Aug. 2012). "A geometric analysis of subspace clustering with outliers". *Ann. Statist.* 40.4, pp. 2195–2238. DOI: 10.1214/12-AOS1034.

Soltanolkotabi, M., Elhamifar, E., and Candès, E. J. (Apr. 2014). "Robust subspace clustering". *Ann. Statist.* 42.2, pp. 669–699. DOI: 10.1214/13-AOS1199.

Sussman, D. L. et al. (2012). "A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs". *Journal of the American Statistical Association* 107.499, pp. 1119–1128. DOI: 10.1080/01621459.2012.699795. eprint: https://doi.org/10.1080/01621459.2012.699795. URL: https://doi.org/10.1080/01621459.2012.699795.

Wang, B. et al. (2018). "Network enhancement as a general method to denoise weighted biological networks". *Nature Communications* 9.1. ISSN: 2041-1723. DOI: 10.1038/s41467-018-05469-x.

Wang, Y.-X. and Xu, H. (2016). "Noisy Sparse Subspace Clustering". *Journal of Machine Learning Research* 17.12, pp. 1–41.

Young, S. J. and Scheinerman, E. R. (2007). "Random Dot Product Graph Models for Social Networks". In: *Algorithms and Models for the Web-Graph*. Ed. by A. Bonato and F. R. K. Chung. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 138–149. ISBN: 978-3-540-77004-6.

A Proofs of Theorem 3 and Theorem 4

Let V_n and $\hat{V}_n$ be the $n \times K^2$ matrices whose columns are the eigenvectors of P and A corresponding to the K^2 largest eigenvalues (in modulus), respectively. We first state an important technical lemma for bounding the maximum ℓ_2 norm difference between the rows of $\hat{V}_n$ and V_n. See Cape, Tang, and Priebe (2019) and Rubin-Delanchy et al. (2017, Lemma 5) for a proof.

Lemma 1. *Let $A \sim \text{PABM}(\{\lambda^{(k\ell\ell)}\}_K)$ be a K-blocks PABM graph on n vertices and let V_n*

and $\hat{V}_n$ be the $n \times K^2$ matrices whose columns are the eigenvectors of P and A corresponding to the K^2 largest eigenvalues in modulus, respectively. Let $(v_n^{(i)})^\top$ and $(\hat{v}_n^{(i)})^\top$ denote the ith row of V_n and $\hat{V}_n$, respectively. Then there exists a constant $c > 1$ and an orthogonal matrix W such that, with high probability,

$$\max_i \|W\hat{v}_n^{(i)} - v_n^{(i)}\| = O_P\Big(\frac{\log^c n}{n\sqrt{\rho_n}}\Big).$$

In particular we can take $c = 1 + \epsilon$ for any $\epsilon > 0$.

Proof of Theorem 3. Recall the notations in Lemma 1 and note that, under our assumption that the latent vectors $\lambda^{(k\ell)|}$ are all homogeneous, we have $\max_i \|v_n^{(i)}\| = O(n^{-1/2})$. Next recall Theorem 2; in particular $B_{ij} = n(v_n^{(i)})^\top v_n^{(j)}$. We therefore have

$$\begin{aligned}
\max_{ij} |\hat{B}_{ij} - B_{ij}| &= \max_{ij} n|(\hat{v}_n^{(i)})^\top \hat{v}_n^{(j)} - (v_n^{(i)})^\top v_n^{(j)}| \\
&\le n \max_{ij} |(\hat{v}_n^{(i)})^\top WW^\top \hat{v}_n^{(j)} - (v_n^{(i)})^\top v_n^{(j)}| \\
&\le n \max_{i,j} \Big(\|W^\top \hat{v}_n^{(i)} - v_n^{(i)}\| \times \|\hat{v}_n^{(j)}\| + \|W^\top \hat{v}_n^{(j)} - v_n^{(j)}\| \times \|v_n^{(i)}\| \Big) \\
&\le n \Big(\max_{ij} \|W\hat{v}_n^{(i)} - v_n^{(i)}\|^2 + \|W\hat{v}_n^{(i)} - v_n^{(i)}\| \times \|v_n^{(j)}\| + \|W\hat{v}_n^{(j)} - v_n^{(j)}\| \times \|v_n^{(i)}\| \Big) \\
&\le n \max_i \|W\hat{v}_n^{(i)} - v_n^{(i)}\|^2 + 2n \max_i \|W\hat{v}_n^{(i)} - v_n^{(i)}\| \times \max_j \|v_n^{(j)}\| \\
&= O\Big(\frac{\log^c n}{n^{1/2}\rho_n^{1/2}} \Big)
\end{aligned}$$

with high probability. Theorem 3 follows from the above bound together with the conclusion in Theorem 2 that $B_{ij} = 0$ whenever vertices i and j belongs to different communities. □

We now provide a proof of Theorem 4. Our proof is based on verifying the sufficient conditions given in Theorem 6 of Wang and Xu, 2016 under which sparse subspace clustering based on solving the optimization problem in Eq. (7) yields an affinity matrix $B = |C| + |C^\top|$ satisfying the subspace detection property of Definition 4. We first recall a few definitions used in Wang and Xu, 2016; for ease of exposition, these definitions are stated using the notations of the current paper.

Definition 5 (Inradius (Soltanolkotabi and Candés, 2012; Wang and Xu, 2016))**.** The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest

Euclidean ball inscribed in $\mathcal{P}$. Let X be a $n \times d$ matrix with rows $x_1, x_2, \ldots, x_n$. We then define, with a slight abuse of notation, $r(X)$ as the inradius of the convex hull formed by $\{\pm x_1, \pm x_2, \ldots, \pm x_n\}$.

Definition 6 (Subspace incoherence)**.** Let $\hat{V}$ be the eigenvectors of A_n corresponding to the K^2 largest eigenvalues in modulus. Let $\hat{V}^{(k)}$ denote the matrix formed by keeping only the rows of $\hat{V}$ corresponding to the k^{th} community and let $\hat{V}^{(-k)}$ denote the matrix formed by omitting the rows of $\hat{V}$ corresponding to the k^{th} community. Let $(\hat{v}_i^{(k)})^\top$ denote the ith row of $\hat{V}^{(k)}$ and $\hat{V}_{-i}^{(k)}$ be $\hat{V}^{(k)}$ with the i^{th} row omitted. Let V, $V^{(k)}$, $V^{(-k)}$, and $v_i^{(k)}$ be defined similarly using the eigenvectors V of P_n. Finally let $\mathcal{S}^{(k)}$ be the vector space spanned by the rows of $V^{(k)}$.

Now define $\nu_i^{(k)}$ for $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n_k$ as the solution of the following optimization problem

$$\nu_i^{(k)} = \max_{\eta} (\hat{v}_i^{(k)})^\top \eta - \frac{1}{2\lambda} \eta^\top \eta, \quad \text{subject to } \|V_{-i}^{(k)} \eta\|_\infty \leq 1.$$

Given $\nu_i^{(k)}$, let $\mathbb{P}_{\mathcal{S}^{(k)}}(\nu_i^{(k)})$ be the vector in $\mathbb{R}^{K^2}$ corresponding to the orthogonal projection of $\nu_i^{(k)}$ onto the vector space $\mathcal{S}^{(k)}$ and define the projected dual direction $w_i^{(k)}$ as

$$w_i^{(k)} = \frac{\mathbb{P}_{\mathcal{S}^{(k)}}(\nu_i^{(k)})}{\|\mathbb{P}_{\mathcal{S}^{(k)}}(\nu_i^{(k)})\|}.$$

Now let $W^{(k)} = \begin{bmatrix} w_1^{(k)} & \cdots & w_{n_k}^{(k)} \end{bmatrix}^\top$ and define the subspace incoherence for $\hat{V}^{(k)}$ by

$$\mu^{(k)} = \mu(\hat{V}^{(k)}) = \max_{v \in V^{(-k)}} \|W^{(k)} v\|_\infty.$$

With the above definitions in place, we are now ready to state our proof of Theorem 4.

Proof of Theorem 4. For a given $k = 1, 2 \ldots, K$, let $r^{(k)} = \min_i r(V_{-i}^{(k)})$ be inradius of the convex hull formed by the rows of $V_{-i}^{(k)}$ and let $r_* = \min_k r^{(k)}$. Then Theorem 6 in Wang and Xu (2016) states that there exists a $\lambda > 0$ such that $\sqrt{n}\hat{V}$ satisfies the subspace detection property in Definition 4 whenever the following two conditions are satisfied

$$\mu^{(k)} < r^{(k)} \quad \text{for all } k = 1, 2, \ldots, K, \tag{10}$$

$$\max_i \|W\hat{v}_i - v_i\| \leq \min_k \frac{r_*(r^{(k)} - \mu^{(k)})}{2 + 7r^{(k)}}. \tag{11}$$

We now verify that for sufficiently large n, Eq. (10) and Eq. (11) holds with high probability.

Verifying Eq. (10). If n is sufficiently large then there are enough vertices in each community k so that $\text{span}(V_{-i}^{(k)}) = \mathcal{S}^{(k)}$ for all i and hence $r^{(k)} = \min_i r(V_{-i}^{(k)}) > 0$ for all $k = 1, 2, \ldots, K$.

Next, by Theorem 2 we have that the subspaces $\{\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(K)}\}$ are mutually orthogonal, i.e., $v^\top w = 0$ for all $v \in \mathcal{S}^{(k)}$ and $w \in \mathcal{S}^{(\ell)}$ with $k \neq \ell$. Now let $z \in \mathbb{R}^{K^2}$ be arbitrary and let $\tilde{z} = \mathbb{P}_{\mathcal{S}^{(k)}} z$ be the projection of z onto $\mathcal{S}^{(k)}$. We then have $v^\top \tilde{z} = 0$ for all $v \in V^{(-k)}$. Since z is arbitrary, this implies $\|W^{(k)} v\|_\infty = 0$ for all $v \in V^{(-k)}$ and hence $\mu^{(k)} = 0$ for all $k = 1, 2, \ldots, K$. Therefore $\mu^{(k)} < r^{(k)}$ for all $k = 1 = 2, \ldots, K$ as desired.

Verifying Eq. (11). Let $\delta = \max_i \sqrt{n}\|W\hat{v}_i - v_i\|$. Then from Lemma 1, we have $\delta \xrightarrow{a.s.} 0$ and hence

$$\delta < \min_k \frac{r_*(r^{(k)} - \mu^{(k)})}{2 + 7r^{(k)}}$$

asymptotically almost surely.

In summary $\sqrt{n}\hat{V}$ satisfies the subspace detection property with probability converging to 1 as $n \to \infty$. □

Remark. Theorem 6 of Wang and Xu (2016) assumes that each row v_i of V has unit norm, i.e., $\|v_i\| = 1$ for all i. This assumption has the effect of scaling the $r^{(k)}$ so that $r^{(k)} \leq 1$ for all $k = 1, 2, \ldots, K$. We emphasize that this assumption has no effect on the proof of Theorem 4. Indeed, since $\mu^{(k)} = 0$ for all k, as long as the rows of $V^{(k)}$ spans the subspace $\mathcal{S}^{(k)}$, then $ar^{(k)} > \mu^{(k)}$ for any scalar $a > 0$.

Proof of Theorem 5. Let P be organized by community such that $P^{(k\ell)}$ denote the $n_k \times n_\ell$ matrix obtained by keeping only the rows of P corresponding to vertices in community k

and the columns of P corresponding to vertices in community ℓ. We define $A^{(k\ell)}$ analogously. Recall that $P^{(k\ell)} = \lambda^{(k\ell)}(\lambda^{(\ell k)})^\top$ for all k, ℓ. We now consider estimation of $P^{(k\ell)}$ for the cases when $k = \ell$ versus when $k \neq \ell$.

Case $k = l$. Let $P^{(kk)} = \sigma_{kk}^2 u^{(kk)}(u^{(kk)})^\top$ be the singular value decomposition of $P^{(kk)}$. We can then define $\tilde{\lambda}^{(kk)} = \sigma_{kk} u^{(kk)}$. Now let $\hat{U}^{(kk)}\hat{\Sigma}^{(kk)}(\hat{U}^{(kk)})^\top$ be the singular value decomposition of $A^{(kk)}$, and let $\hat{\sigma}_{kk}^2 \hat{u}^{(kk)}(\hat{u}^{(kk)})^\top$ be the best rank-one approximation of $A^{(kk)}$. Define $\hat{\lambda}^{(kk)} = \hat{\sigma}_{kk}\hat{u}^{(kk)}$. Then $\hat{\lambda}^{(kk)}$ is the adjacency spectral embedding approximation of $\lambda^{(kk)}$ and by Theorem 5 of Rubin-Delanchy et al. (2017), we have

$$\|\hat{\lambda}^{(kk)} - \lambda^{(kk)}\|_\infty = O\Big(\frac{(\log n_k)^c}{\sqrt{n_k}}\Big)$$

with high probability. Here $\|\cdot\|_\infty$ denote the ℓ_∞ norm of a vector.

Case $k \neq l$. Let $P^{(k\ell)} = \sigma_{k\ell}^2 u^{(k\ell)}(v^{(k\ell)})^\top$ and $P^{(\ell k)} = \sigma_{kl}^2 u^{(\ell k)}(v^{(\ell k)})^\top$ be the singular value decompositions and note that $\sigma_{k\ell} = \sigma_{\ell k}$, $u^{(k\ell)} = v^{(\ell k)}$, and $v^{(k\ell)} = u^{(\ell k)}$. Now define $\lambda^{(k\ell)} = \sigma_{k\ell} u^{(k\ell)}$ and $\lambda^{(\ell k)} = \sigma_{k\ell} v^{(k\ell)}$.

Next consider the Hermitian dilation

$$M^{(k\ell)} = 2\begin{bmatrix} 0 & P^{(k\ell)} \\ P^{(\ell k)} & 0 \end{bmatrix}$$

which is a symmetric $(n_k + n_\ell) \times (n_k + n_\ell)$ matrix. The eigendecomposition of $M^{(k\ell)}$ is then

$$M^{(k\ell)} = \begin{bmatrix} u^{(k\ell)} & -u^{(k\ell)} \\ v^{(k\ell)} & v^{(k\ell)} \end{bmatrix} \times \begin{bmatrix} \sigma_{kl}^2 & 0 \\ 0 & -\sigma_{kl}^2 \end{bmatrix} \times \begin{bmatrix} u^{(k\ell)} & -u^{(k\ell)} \\ v^{(k\ell)} & v^{(k\ell)} \end{bmatrix}^\top$$

Thus treating $M^{(k\ell)}$ as the edge probability matrix of a GRDPG, we have latent positions in $\mathbb{R}^2$ given by the $(n_k + n_\ell) \times 2$ matrix

$$\Lambda^{(k\ell)} = \begin{bmatrix} \sigma_{k\ell} u^{(k\ell)} & \sigma_{k\ell} u^{(k\ell)} \\ \sigma_{k\ell} v^{(k\ell)} & -\sigma_{k\ell} v^{(k\ell)} \end{bmatrix} = \begin{bmatrix} \lambda^{(k\ell)} & \lambda^{(k\ell)} \\ \lambda^{(\ell k)} & -\lambda^{(\ell k)} \end{bmatrix}.$$

Now consider

$$\hat{M}^{(k\ell)} = \begin{bmatrix} 0 & A^{(k\ell)} \\ A^{(\ell k)} & 0 \end{bmatrix}$$

We can then view $\hat{M}^{(k\ell)}$ as an adjacency matrix drawn from the edge probabilities matrix $M^{(k\ell)}$. Now suppose that the adjacency spectral embedding of $\hat{M}^{(k\ell)}$ is represented as the $(n_k + n_\ell) \times 2$ matrix

$$\hat{\Lambda}^{(k\ell)} = \begin{bmatrix} \hat{\lambda}^{(k\ell)} & \hat{\lambda}^{(k\ell)} \\ \hat{\lambda}^{(\ell k)} & -\hat{\lambda}^{(\ell k)} \end{bmatrix}$$

where each $\hat{\lambda}^{(k\ell)}$ is defined as in Algorithm 3. Then by Theorem 5 of Rubin-Delanchy et al. (2017), there exists an indefinite orthogonal transformation W^* such that, with high probability,

$$\max_i |W^* \hat{\Lambda}_i^{(k\ell)} - \Lambda_i^{(k\ell)}\| = O\Big(\frac{(\log(n_k + n_\ell))^c}{\sqrt{n_k + n_\ell}}\Big)$$

with high probability. Here $\Lambda_i^{(k\ell)}$ and $\hat{\Lambda}_i^{(k\ell)}$ denote the ith rows of $\Lambda^{(k\ell)}$ and $\hat{\Lambda}^{(k\ell)}$, respectively. Furthermore, by looking at the proof of Theorem 5 in (Rubin-Delanchy et al., 2017), we see that W^* is also blocks diagonal with 2 blocks where the positive eigenvalues of $M^{(k\ell)}$ forming a block and the negative eigenvalues of $M^{(k\ell)}$ forming the remaining block. Since $M^{(k\ell)}$ has one positive eigenvalue and one negative eigenvalue, we see that W^* is necessarily of the form $W^* = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ Using this form for W^*, we obtain

$$\max\{\|\hat{\lambda}^{(k\ell)} - \lambda^{(k\ell)}\|_\infty, \|\hat{\lambda}^{(\ell k)} - \lambda^{(\ell k)}\|_\infty\} = O\Big(\frac{(\log(n_k + n_\ell))^c}{\sqrt{n_k + n_\ell}}\Big)$$

with high probability. Combining this bound with the bound for $\|\hat{\lambda}^{(kk)} - \lambda^{(kk)}\|_\infty$ given above yields Eq. (9) in Theorem 5.