1 Asymptotic behavior of the differences between latent positions and their estimates using omnibus embedding of multiple random graphs

1.1 Definitions

Definition 1.1. *(Random Dot Product Graph [1]) Let F be a distribution on a set $\mathcal{X} \subset \mathbb{R}^d$ satisfying $\langle x, x' \rangle \in [0,1]$ for all $x, x' \in \mathcal{X}$. We say (X, A) ~RDPG(F, n) if the following hold. Let $X_1, X_2, \cdots, X_n \sim F$ be independent random variables and define*

$$X = [X_1, \cdots, X_n]^T \in \mathbb{R}^{n \times d} \text{ and } P = XX^T \in [0,1]^{n \times n}.$$

The X_i are the latent positions for the random graph. The matrix $A \in \{0,1\}^{n \times n}$ is defined to be a symmetric, hollow matrix such that for all $i < j$, conditioned on X_i, X_j,

$$A_{i,j} \sim^{ind} Bern(X_i^T X_j).$$

We say that A is the adjacency matrix of a random dot product graph with latent positions given by the rows of X.

Definition 1.2. *(Joint Random Dot Product Graph [2]) Let F be a d- dimensional inner product distribution on $\mathbb{R}^d$(as in 1.1). We say that random graphs $A^{(1)}, A^{(2)} \cdots, A^{(m)}$ are distributed as a joint random dot product graph (JRDPG) and write $(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, X) \sim JRDPG(F, n, m)$ if $X = [X_1, X_2, \cdots, X_n]^T \in \mathbb{R}^{n \times d}$ has its (transposed) rows distributed i.i.d as $X_i \sim F$, and we have marginal distributions $(A^{(k)}, X)$ ~RDPG(F, n) for each $k = 1, 2, \cdots, m$. That is, the $A^{(k)}$ are conditionally independent given X, with edges independently distributed as $A_{i,j}^{(k)} \sim Bernoulli((XX^T)_{ij})$ for all $1 \leq i < j \leq n$ and all $k \in [m]$.*

Definition 1.3. *(Adjacency Spectral Embedding [3]) Let $A \in \mathbb{R}^{n \times n}$ be the adjacency matrix of an undirected d-dimensional random dot product graph. The d-dim adjacency spectral embedding (ASE) of A is a spectral decomposition of A based on its top d eigenvalues, obtained by $ASE(A, d) = U_A S_A^{1/2}$, where $S_A \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose entries are the top eigenvalues of A(in nondecreasing order) and $U_A \in \mathbb{R}^{n \times d}$ is the matrix whose columns are orthonormal eigenvectors corresponding to the eigenvalues in S_A.*

Definition 1.4. *(Omnibus Embedding [2]) Let $A^{(1)}, A^{(2)}, \cdots, A^{(m)} \in \mathbb{R}^{n \times n}$ be adjacency matrices of a collection of m undirected graphs. We define the mn-by-mn omnibus matrix of $A^{(1)}, A^{(2)}, \cdots, A^{(m)}$ by*

$$M = \begin{bmatrix} A^{(1)} & \frac{A^{(1)}+A^{(2)}}{2} & \frac{A^{(1)}+A^{(3)}}{2} & \cdots & \frac{A^{(1)}+A^{(m)}}{2} \\ \frac{A^{(2)}+A^{(1)}}{2} & A^{(2)} & \frac{A^{(2)}+A^{(3)}}{2} & \cdots & \frac{A^{(2)}+A^{(m)}}{2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{A^{(m)}+A^{(1)}}{2} & \frac{A^{(m)}+A^{(2)}}{2} & \frac{A^{(m)}+A^{(3)}}{2} & \cdots & A^{(m)} \end{bmatrix},$$

and the d-dimensional omnibus embedding of $A^{(1)}, A^{(2)}, \cdots, A^{(m)}$ is the adjacency spectral embedding of M:

$$OMNI(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, d) = ASE(M, d).$$

Remark 1. *The matrix $\widehat{Z} = OMNI(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, d)$ is a natural estimate of the mn latent positions collected in the matrix $Z = [X^T, X^T, \cdots, X^T]^T \in \mathbb{R}^{mn \times d}$.*

1.2 Main Results

In Athreya et al.([1]), they prove a central limit theorem for the scaled differences between the estimated and true latent positions in the RDPG setting.

Theorem 1. *Let $(X, A) \sim RDPG(F)$ be a d-dimensional random dot product graph, and let $\widehat{X} = ASE(A, d)$ be our estimate for X. Let $\Phi(z, \Sigma)$ denote the cumulative distribution function for the multivariate normal, with mean zero and covariance matrix Σ, evaluated at z. Then, there exists a sequence of orthogonal matrices W_n converging to the identity almost surely such that for each component i and any $z \in \mathbb{R}^d$,*

$$\mathbb{P}\left\{\sqrt{n}(W_n\widehat{X}_i - X_i) \leq z\right\} \to \int \Phi(z_i, \Sigma(x_i))dF(x_i) \tag{1}$$

where $\Sigma(x) = \Delta^{-1}\mathbb{E}[X_jX_j^T(x^TX_j - (x^TX_j)^2)]\Delta^{-1}$ and $\Delta = \mathbb{E}[X_1X_1^T]$ is the second matrix. That is, the sequence of random variables $\sqrt{n}(W_n\widehat{X}_i - X_i)$ converges in distribution to a mixture of multivariate normals. We denote the mixture by $\mathcal{N}(0, \Sigma(X_i))$.

In Levin et al.([2]), the above result is extended to multiple random graphs using the Omnibus matrix,

Theorem 2. *Let $(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, X) \sim JRDPG(F, n, m)$ for some d-dimensional inner product distribution F and let M denote the omnibus matrix as in 1.4. Let Z and its estimate $\widehat{Z}$ as in Remark 1. Let $h = n(s-1)+i$ for $i \in [n]$, $s \in [m]$, so that $\widehat{Z}_h$ denotes the estimated latent position of the i-th vertex in the s-th graph $A^{(s)}$. That is, $\widehat{Z}_h$ is the column vector formed by transposing the h-th row of the matrix $\widehat{Z} = U_MS_M^{1/2} = OMNI(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, d)$. Let $\Phi(x, \Sigma)$ denote the cdf of a (multivariate) Gaussian with mean zero and covariance matrix Σ, evaluated at $x \in \mathbb{R}^d$. There exists a sequence of orthogonal d-by-d matrices $(\tilde{W}_n)_{n=1}^\infty$ such that for all $x \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\sqrt{n}(\widehat{Z}\tilde{W}_n - Z)_h \leq x\right\} \to \int_{supp\ F} \Phi(x, \Sigma(y))dF(y),$$

where $\Sigma(y) = (m+3)\Delta^{-1}\widetilde{\Sigma}(y)\Delta^{-1}/4m$, $\Delta = \mathbb{E}[X_1X_1^T] \in \mathbb{R}^{d\times d}$ and $\widetilde{\Sigma}(y) = \mathbb{E}[(y^TX_1 - (y^TX_1)^2)X_1X_1^T]$.

1.3 Extension-General case with unknown coefficients

In this section, the off-diagonal entries of the omnibus matrix M are of the form $\frac{c_iA_i+c_jA_j}{c_i+c_j}$, where $c_i, c_j \in (0,1)$.

$$M = \begin{bmatrix} A^{(1)} & \frac{c_1A^{(1)}+c_2A^{(2)}}{c_1+c_2} & \frac{c_1A^{(1)}+c_3A^{(3)}}{c_1+c_3} & \cdots & \frac{c_1A^{(1)}+c_mA^{(m)}}{c_1+c_m} \\ \frac{c_2A^{(2)}+c_1A^{(1)}}{c_2+c_1} & A^{(2)} & \frac{c_2A^{(2)}+c_3A^{(3)}}{c_2+c_3} & \cdots & \frac{c_2A^{(2)}+c_mA^{(m)}}{c_2+c_m} \\ \frac{c_3A^{(3)}+c_1A^{(1)}}{c_3+c_1} & \frac{c_3A^{(3)}+c_2A^{(2)}}{c_3+c_2} & A^{(3)} & \cdots & \frac{c_3A^{(3)}+c_mA^{(m)}}{c_3+c_m} \\ \vdots & \vdots & \ddots & \vdots & \\ \frac{c_mA^{(m)}+c_1A^{(1)}}{c_m+c_1} & \frac{c_mA^{(m)}+c_2A^{(2)}}{c_m+c_2} & \frac{c_mA^{(m)}+c_3A^{(3)}}{c_m+c_3} & \cdots & A^{(m)} \end{bmatrix}$$

Our focus in this section shifts on the differences between the estimated latent positions with each other, instead of the differences between the estimated and the true latent positions. For example, we may are interested in the asympotic behavior of the difference between the 1st row of M with the $n+1$st, $2n+1$st,$\cdots,(m-1)n+1$st rows of M. The extension follows naturally, since from 2 we have, for example, the true latent position given by the 1st row of X is close to the estimated latent position given by the 1st row of M, the estimated latent position given by the $n+1$st row of M and so on, hence the corresponding rows of M has to be close to each other as well. The advantage using the rows of M is that you avoid to use Procrustean transformation as you do in Theorems 1 and 2. The simulations produced from the code below, suggest that the statement is true.

Theorem 3. *(Conjecture) Let $(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, X) \sim$JRDPG$(F,n,m)$ for some d-dim inner product distribution F and let M denote the omnibus matrix above. Let $h_1 = n(s_1-1)+i, h_2 = n(s_2-1)+i$ for $i \in [n], s_1, s_2 \in [m]$, so that $\widehat{M}_{h_1}, \widehat{M}_{h_2}$ denote the estimated latent positions of the i-th vertex in the s_1, s_2-th graphs $A^{(s_1)}, A^{(s_2)}$ respectively. That is, $\widehat{M}_{h_1}, \widehat{M}_{h_2}$ are the column vectors formed by transposing the h_1, h_2-th rows of the matrix $ASE(M,d) = U_MS_M^{1/2} = OMNI(A^{(1)}, A^{(2)}, \cdots, A^{(m)}, d)$. Let $\Phi(x,\Sigma)$ denote the cdf of a (multivariate) Gaussian with mean zero and covariance matrix Σ, evaluated at $x \in \mathbb{R}^d$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\sqrt{n}(\widehat{M}_{h_1} - \widehat{M}_{h_2}) \leq x\right\} \to \int_{supp\ F} \Phi(x, \Sigma(y))dF(y),$$

where $\Sigma(y) = \frac{2}{m^2}\Delta^{-1}\widetilde{\Sigma}(y)\Delta^{-1}$, $\Delta = \mathbb{E}[X_1X_1^T] \in \mathbb{R}^{d\times d}$ and

$$\widetilde{\Sigma}(y) = \frac{2}{m^2}\,\mathbb{E}[X_jX_j^T(y^TX_j - (y^TX_j)^2)]\Big[1 + \sum_{k=3}^{m} \frac{c_1}{c_1+c_k} + \frac{c_2}{c_2+c_k} + \frac{c_1^2}{(c_1+c_k)^2} + \frac{c_2^2}{(c_2+c_k)^2}$$
$$-\frac{c_1c_2(c_1c_2+c_1c_k+c_2c_k+c_k^2)}{(c_1+c_k)^2(c_2+c_k)^2} + \sum_{k<l,k\geq 3} \frac{c_1^2c_2(c_2+c_l+c_k)+c_1c_2^2(c_1+c_l+c_k)+c_kc_l(c_1^2+c_2^2)}{(c_1+c_k)(c_2+c_k)(c_1+c_l)(c_2+c_l)}\Big]. \qquad (2)$$

References

[1] *A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs.* Sankhya A, *pages 1–18, 2013.*

[2] Keith Levin, Avanti Athreya, Minh Tang, Vince Lyzinski, Youngser Park, and Carey E. Priebe. A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference, 2017.

[3] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. Journal of the American Statistical Association*, 107(499):1119–1128, 2012.*