A statistical interpretation of spectral embedding: The generalised random dot product graph

Patrick Rubin-Delanchy

School of Mathematics, University of Bristol

Bayesian Computation for High-Dimensional Statistical Models, 28th August 2018

Outline

I. The generalised random dot product graph

II. A statistical interpretation of spectral embedding and methodological applications

Acknowledgements

- Carey Priebe, Minh Tang, Joshua Cape (Johns Hopkins University)
- Nick Heard (Imperial College London)
- Melissa Turcotte, Alex Kent (Los Alamos National Laboratory), Josh Neil (Microsoft)
- Heilbronn Institute for Mathematical Research

Context

Data: undirected graph with a binary symmetric adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$.

Model:

Definition (Latent position model)

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}\left\{f(X_i, X_j)\right\},$$

for $i < j$, where:

- $X_1, \ldots, X_n \in \mathcal{X} \subseteq \mathbb{R}^d$ are unobserved latent positions
- $f : \mathcal{X}^2 \to [0,1]$ is a symmetric function (sometimes called a kernel)

Two popular choices:

1. distance-based model (Hoff et al., 2002):

$$f(x, y) = \mathrm{logistic}\left(\alpha - |x - y|\right)$$

2. random dot product graph (Athreya et al., 2018):

$$f(x, y) = x^\top y$$

Context

Data: undirected graph with a binary symmetric adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$.

Model:

Definition (Latent position model)

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}\left\{f(X_i, X_j)\right\},$$

for $i < j$, where:

- $X_1, \ldots, X_n \in \mathcal{X} \subseteq \mathbb{R}^d$ are unobserved latent positions
- $f : \mathcal{X}^2 \to [0,1]$ is a symmetric function (sometimes called a kernel)

Two popular choices:

1. distance-based model (Hoff et al., 2002):

$$f(x, y) = \text{logistic}\left(\alpha - |x - y|\right)$$

2. random dot product graph (Athreya et al., 2018):

$$f(x, y) = x^\top y$$

Problematic data

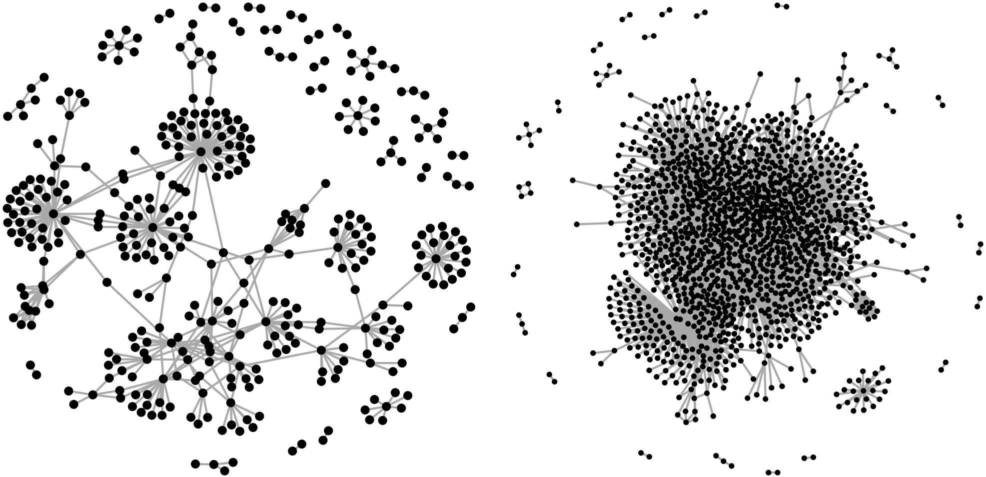


Figure: NetFlow (left: 1 min, right: 5 min), Los Alamos National Laboratory (LANL) network (Kent, 2016)

Stochastic block model

$k \in \mathbb{N}$, $\mathbf{B} \in [0,1]^{k \times k}$.

Definition (Stochastic block model)

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}\{\mathbf{B}_{\ell(i),\ell(j)}\},$$

for $i < j$, where $\ell : \{1, \ldots, n\} \to \{1, \ldots, k\}$ groups the nodes into communities

Mixed membership stochastic block model

$k \in \mathbb{N}$, $\mathbf{B} \in [0,1]^{k \times k}$ and $\alpha \in \mathbb{R}_{+}^{k}$.

Definition (Mixed membership stochastic block model)

Let $\pi_1, \ldots, \pi_n \overset{i.i.d.}{\sim}$ Dirichlet(α), representing the nodes' community affinities. Next, let

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}\left(\mathbf{B}_{z_{i \to j}, z_{j \to i}}\right),$$

where

$$z_{i \to j} \overset{ind}{\sim} \text{multinomial}(\pi_i) \quad \text{and} \quad z_{j \to i} \overset{ind}{\sim} \text{multinomial}(\pi_j)$$

With this notion of mixed membership in place, consider how we might want to interpret latent space. One obvious idea would be for $x = (y + z)/2$ to represent a 50/50 mixture of connectivity behaviours at y and z.

Property (Reproducing mixtures of connectivity behaviour)

A symmetric function $f : \mathcal{X}^2 \to [0, 1]$ reproduces mixtures of connectivity behaviour over a convex set $\mathcal{X}$ if, whenever $x = \sum \alpha_r u_r$, where $u_r \in \mathcal{X}$, $0 \leq \alpha_r \leq 1$ and $\sum \alpha_r = 1$, we have

$$f(x, y) = \sum_r \alpha_r f(u_r, y),$$

for any $y \in \mathcal{X}$.

Example:

if $X_1 = \frac{1}{2}X_2 + \frac{1}{2}X_3$, then

$$\mathbf{A}_{14} \overset{ind}{\sim} \mathrm{B}\{f(X_1, X_4)\},$$

or equivalently

$$\mathbf{A}_{14} \overset{ind}{\sim} \frac{1}{2}\mathrm{B}\{f(X_2, X_4)\} + \frac{1}{2}\mathrm{B}\{f(X_3, X_4)\}.$$

A canonical representation

Theorem

If $\mathcal{X}$ is a subset of $\mathbb{R}^l$, then f reproduces mixtures of connectivity behaviour over $\mathcal{X}$ if and only if there exists an affine transformation $\mathcal{A}$ such that $f(x, y) = \mathcal{A}(x)^\top \mathbf{I}_{p,q} \mathcal{A}(y)$, where $\mathbf{I}_{p,q} = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$, with $p \geq 1$ ones followed by $q \geq 0$ minus ones on its diagonal.

$p \geq 1$, $q \geq 0$, $p + q = d$, $\mathcal{X}$ a subset of $\mathbb{R}^d$ such that $x^\top \mathbf{I}_{p,q} y \in [0, 1]$ for all $x, y \in \mathcal{X}$.

Definition (Generalised random dot product graph)

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}(X_i^\top \mathbf{I}_{p,q} X_j),$$

for $i < j$, where $X_1, \ldots, X_n \in \mathcal{X} \subseteq \mathbb{R}^d$ are unobserved latent positions.

A canonical representation

Theorem

If $\mathcal{X}$ is a subset of $\mathbb{R}^l$, then f reproduces mixtures of connectivity behaviour over $\mathcal{X}$ if and only if there exists an affine transformation $\mathcal{A}$ such that $f(x, y) = \mathcal{A}(x)^\top \mathbf{I}_{p,q} \mathcal{A}(y)$, where $\mathbf{I}_{p,q} = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$, with $p \geq 1$ ones followed by $q \geq 0$ minus ones on its diagonal.

$p \geq 1$, $q \geq 0$, $p + q = d$, $\mathcal{X}$ a subset of $\mathbb{R}^d$ such that $x^\top \mathbf{I}_{p,q} y \in [0, 1]$ for all $x, y \in \mathcal{X}$.

Definition (Generalised random dot product graph)

$$\mathbf{A}_{ij} \overset{ind}{\sim} \mathrm{B}(X_i^\top \mathbf{I}_{p,q} X_j),$$

for $i < j$, where $X_1, \ldots, X_n \in \mathcal{X} \subseteq \mathbb{R}^d$ are unobserved latent positions.

Special cases

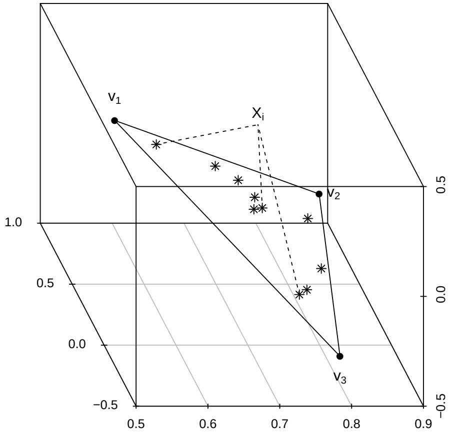
1. The stochastic block model (see below)
2. The mixed membership stochastic block model: eigendecomposition $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. Let $\mathrm{v}_1, \ldots, \mathrm{v}_k \in \mathbb{R}^k$ denote rows of $\mathbf{U}|\mathbf{\Sigma}|^{1/2}$
3. The random dot product graph: $p = d, q = 0$, yielding standard inner product. What the GRDPG adds is the possibility of modelling disassortative connectivity behaviour, e.g. where 'opposites attract'

Special cases

1. The stochastic block model (see below)
2. The mixed membership stochastic block model: eigendecomposition $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$. Let $\mathrm{v}_1, \ldots, \mathrm{v}_k \in \mathbb{R}^k$ denote rows of $\mathbf{U}|\boldsymbol{\Sigma}|^{1/2}$
3. The random dot product graph: $p = d, q = 0$, yielding standard inner product. What the GRDPG adds is the possibility of modelling disassortative connectivity behaviour, e.g. where 'opposites attract'

Special cases

1. The stochastic block model (see below)
2. The mixed membership stochastic block model: eigendecomposition $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. Let $\mathrm{v}_1, \ldots, \mathrm{v}_k \in \mathbb{R}^k$ denote rows of $\mathbf{U}|\mathbf{\Sigma}|^{1/2}$



Convex combination parameters

$$X_i = \sum_{l=1}^{k} \pi_{il} \mathrm{v}_l$$

give the community membership probability π_i

3. The random dot product graph: $p = d, q = 0$, yielding standard inner product. What the GRDPG adds is the possibility of modelling disassortative connectivity behaviour, e.g. where 'opposites attract'

Special cases

1. The stochastic block model (see below)
2. The mixed membership stochastic block model: eigendecomposition $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. Let $\mathrm{v}_1, \ldots, \mathrm{v}_k \in \mathbb{R}^k$ denote rows of $\mathbf{U}|\mathbf{\Sigma}|^{1/2}$
3. The random dot product graph: $p = d, q = 0$, yielding standard inner product. What the GRDPG adds is the possibility of modelling disassortative connectivity behaviour, e.g. where 'opposites attract'

Conclusion (Part I)

1. The generalised random dot product graph provides a canonical latent position model that encompasses earlier (very popular) models as fairly straightforward special cases.
2. I think developing Bayesian inference tools for this model would be interesting and useful.
3. We will now show that several popular spectral embedding/clustering techniques can be understood and improved through the generalised random dot product graph.

Spectral embedding

1. $\mathbf{M}$ some 'regularised' version of the adjacency matrix $\mathbf{A}$, e.g.
   - $\mathbf{M} = \mathbf{A}$
   - $\mathbf{M} = \mathbf{D} - \mathbf{A}$, $\mathbf{D}$ degree matrix (Laplacian)
   - $\mathbf{M} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ or $\mathbf{M} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ (normalised Laplacian)
2. Consider the eigendecomposition[1] $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{U}^T$, where
   - $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an orthonormal matrix
   - $\mathbf{S} = \text{diag}\{\lambda_1, \dots, \lambda_n\}$ contains the eigenvalues of $\mathbf{M}$ in some order (e.g. decreasing)

   The first d columns of $\mathbf{U}|\mathbf{S}|^{1/2}$ provide a *spectral embedding* of $\mathbf{A}$, i.e. a mapping of the nodes $1, \dots, n$ to points $\hat{X}_1, \dots, \hat{X}_n \in \mathbb{R}^d$
3. Use some clustering algorithm, e.g. k-means (typically $k = d$), on $\hat{X}_1, \dots, \hat{X}_n$, to identify communities

[1] fast methods available for large sparse matrices, e.g. irlba (Lewis, 2009)

Los Alamos National Laboratory data — NetFlow

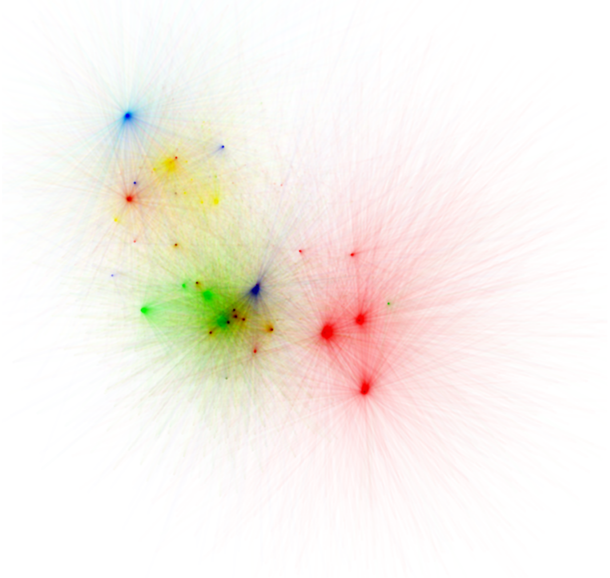


Figure: NetFlow (full graph), LANL network

Los Alamos National Laboratory data — NetFlow

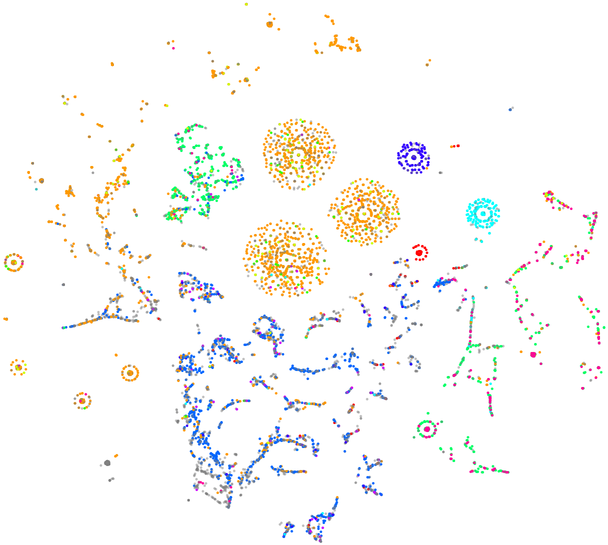


Figure: NetFlow (full graph), spectral embedding ($d = 10$) followed by t-SNE

Spectral clustering and the stochastic block model

Theorem by Rohe et al. (2011):

- Let $\mathbf{M} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ (normalised Laplacian), eigendecomposition $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{U}^T$ with the eigenvalues ordered in decreasing **magnitude***, k-dimensional embedding $\hat{X}_1, \ldots, \hat{X}_n$
- k-means clustering $\Rightarrow$ consistent estimate for the SBM

*and not by decreasing value, as commonly recommended e.g. "A tutorial on spectral clustering" (Von Luxburg, 2007)

A closer look at spectral embedding

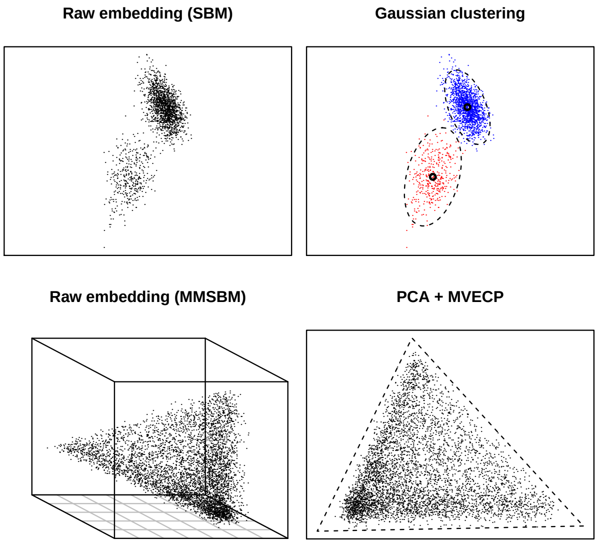


Figure: Adjacency spectral embedding for the SBM and MMSBM

Potential applications

From the pictures we could (optimistically) conjecture that the points were in fact estimates of the latent positions of a generalised random dot product graph. Could we then use such a connection to obtain:

1. A mathematical characterisation of the observed ellipsoids (SBM)?
2. A guarantee on the extreme values of $\hat{X}_i - X_i$ (MMSBM)?

However, note that positions are identifiable only up to transformations in the group $\mathbb{O}(p, q) = \{\mathbf{M} \in \mathbb{R}^{d \times d} \ : \ \mathbf{M}\mathbf{I}_{p,q}\mathbf{M}^\top = \mathbf{I}_{p,q}\}$, which creates initially perturbing complications.

Identifiability

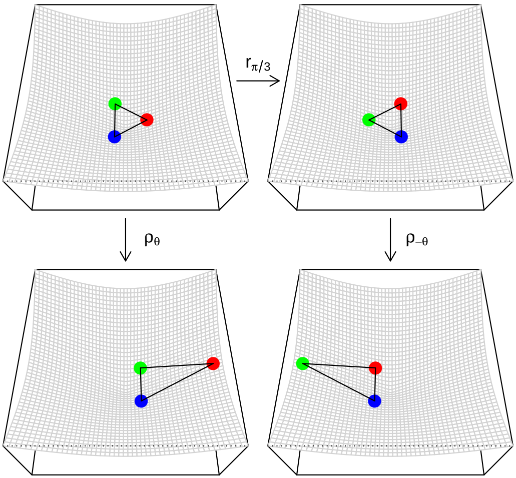


Figure: Identifiability of GRDPG latent positions when $p = 1, q = 2$.

Two theorems

Theorem (GRDPG CLT)

Fix $m > 0$. Conditional on $X_1 = x_1, \ldots, X_m = x_m$, there is a random sequence of indefinite orthogonal matrices $\mathbf{Q}_n \in \mathbb{O}(p,q)$ such that, as $n \to \infty$, the vectors $n^{1/2}(\mathbf{Q}_n\hat{X}_1 - X_1), \ldots, n^{1/2}(\mathbf{Q}_n\hat{X}_m - X_m)$ are independent zero-mean Gaussian vectors with covariances $\Psi(x_1), \ldots, \Psi(x_m)$, respectively, where $\Psi(x)$ is a fixed function of x.

Theorem (GRDPG $2 \to \infty$)

There is a random sequence of indefinite orthogonal matrices $\mathbf{Q}_n \in O(p,q)$ such that

$$\max_i \|\mathbf{Q}_n\hat{X}_i - X_i\| = O_{\mathbb{P}}\left(\frac{d^{1/2}(\log n)^c}{n^{1/2}}\right),$$

where $c > 0$ is fixed constant related to sparsity.

Implications of the theorems

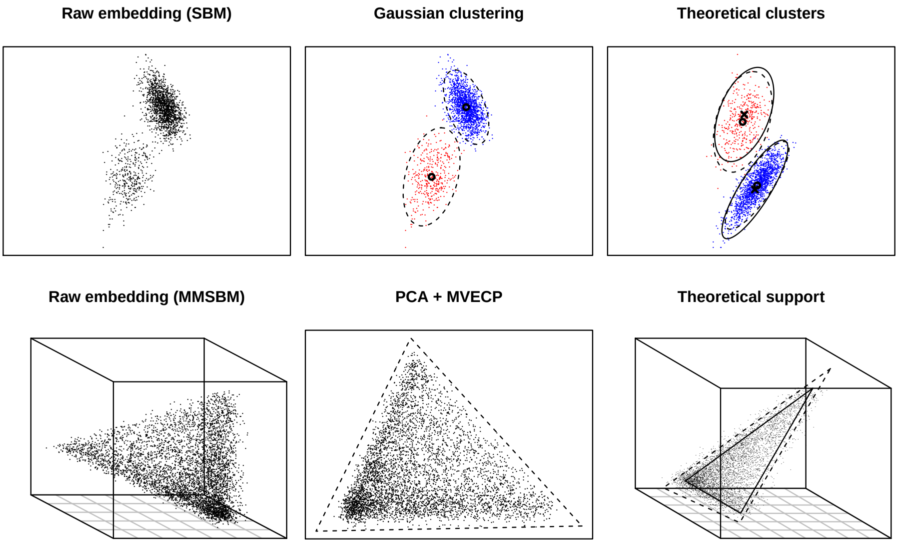


Figure: Adjacency spectral embedding for the SBM and MMSBM

Los Alamos National Laboratory data — red team

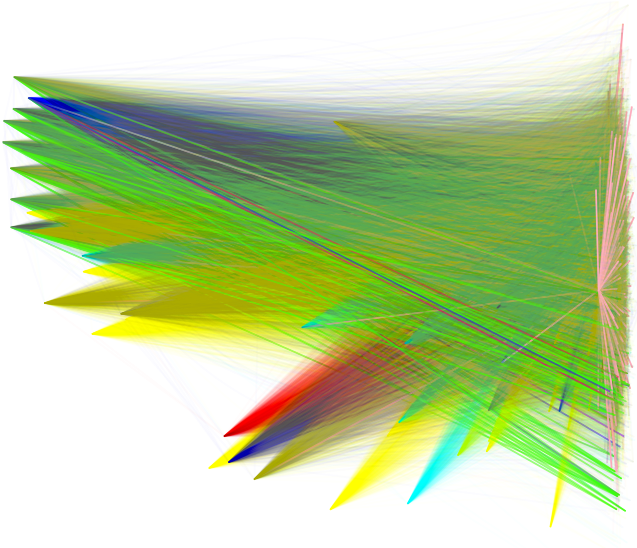


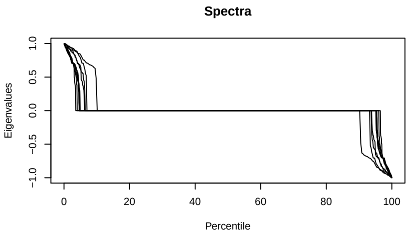
Figure: Authentication data (full graph), LANL network

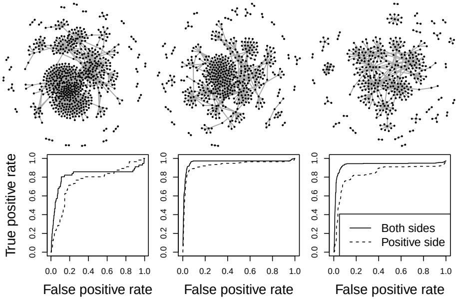
Red team experiment

- The authentication data have 336,806,387 time-indexed events, distributed over
  - 16,230 source computers (4 red team),
  - 15,417 destination computers,
  - 419,744 (source,destination) directed edges.
- Assume destination computers receive new connections according to exponential model.
- Combine p-values by source.

Source computer ID	Anomaly ranking
C17693	2
C18025	384
C19932	550
C22409	1079

Link prediction





Three one-minute observations of the NetFlow graph

Predicting which *new* edges will occur in the next minute

Two approaches:

- Eigenvectors from largest positive eigenvalues only, i.e. standard spectral embedding (Von Luxburg, 2007)
- Eigenvectors from largest magnitude eigenvalues

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Athreya, A., Fishkind, D. E., Levin, K., Lyzinski, V., Park, Y., Qin, Y., Sussman, D. L., Tang, M., Vogelstein, J. T., and Priebe, C. E. (2018). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18:1–92.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining *p*-values. *Biometrika*, 105(1):239–246.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Kent, A. D. (2016). Cybersecurity data sources for dynamic network research. In *Dynamic Networks and Cybersecurity*. World Scientific.

Lewis, B. (2009). irlba: Fast partial SVD by implicitly-restarted lanczos bidiagonalization.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915.

Rubin-Delanchy, P., Heard, N. A., and Lawson, D. J. (2018a). Meta-analysis of mid-p-values: Some new results based on the convex order. *Journal of the American Statistical Association*, (just-accepted):1–25.

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

Rubin-Delanchy, P., Priebe, C. E., and Tang, M. (2018b). A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.