A consistent dot product embedding for stochastic blockmodel graphs

Daniel Sussman, Minh Tang, Donniell Fishkind,
Joshua Vogelstein & Carey Priebe

Department of Applied Mathematics and Statistics
Johns Hopkins University

Interface 2012

Graphs and inference

Network analysis is rapidly becoming a key tool in the analysis of modern datasets in fields ranging from neuroscience to sociology to biochemistry. In each of these fields, there are objects, such as neurons, people, or genes, and there are relationships between objects, such as synapses, friendships, or protein interactions.

The formation of these relationships can depend on attributes of the individual objects as well as properties of the network as a whole. Objects with similar attributes can form communities with similar connective structure, while their unique properties fine tune the shape of these relationships.

Clustering objects based on a graph enables identification of communities and objects of interest as well as illumination of overall network structure. However, finding optimal clusters is difficult and will depend on the particular setting and task.

Graphs and inference (continued)

By using statistical models with inherent community structure as models for graphs [Handcock et al.(2007), Snijders and Nowicki(1997)], the notion of optimal clusters is well-defined. We study the problem of clustering and classification for some random graphs models, namely the stochastic block model [Wang and Wong(1987), Holland et al.(1983)] and a latent space model [Hoff et al.(2002)].

One example of a latent space model is the random dot product graph (RDPG) model [Young and Scheinerman(2007)]. In this model, the probability of an edge between two nodes is given by the dot product of their respective latent vectors. For example, in a social network, the vectors may be interpreted as representing the interest of the individuals.

We present an embedding motivated by the RDPG model which uses a decomposition of a low rank approximation of the adjacency matrix giving an embedding of the nodes as vectors in a low dimensional space. We show that the resulting embedding is useful for clustering and classification purposes.

Stochastic blockmodel graphs

Let K be a positive integer. The stochastic blockmodel (SBM) for graphs is a random graphs model parametrized by a matrix $\mathbf{B} \in [0,1]^{K \times K}$ and a probability vector $\rho \in (0,1)^K$ and we denote by $G_n \sim \mathrm{SBM}(\mathbf{B}, \rho)$ a SBM graph on n vertices.

A $G_n \sim \mathrm{SBM}(\mathbf{B}, \rho)$ can be sampled via the following procedure.

1. Sample $\tau_1, \tau_2, \dots, \tau_n \overset{\text{i.i.d}}{\sim} \mathrm{categorical}(\rho)$, i.e., $\mathbb{P}[\tau_i = k] = \rho_k$.
2. Conditioned on the $\{\tau_v\}$, the edges of G_n are independent Bernoulli random variables with parameters $\mathbf{B}(\tau_u, \tau_v)$, i.e., if we denote by $\mathbf{A}$ the adjacency matrix of G, then

$$\mathbb{P}[\mathbf{A} \mid \{\tau_v\}] = \prod_{u \neq v} (\mathbf{B}(\tau_u, \tau_v))^{\mathbf{A}(u,v)} (1 - \mathbf{B}(\tau_u, \tau_v))^{1 - \mathbf{A}(u,v)}$$

We will denote by $\mathbf{P}$ the $n \times n$ matrix with entries $\mathbf{P}(u,v) = \mathbf{B}(\tau_u, \tau_v)$.

Inferential task for SBM graphs

Given a graph $G \sim \mathrm{SBM}(\mathbf{B}, \rho)$ on n vertices, where $\mathbf{B}$ and ρ are unobserved, we wish to

1. Determine the number of groups K
2. Estimate the group memberships $\tau\colon [n] \mapsto K$ (i.e., the set $\{\tau_v\}_{v \in V}$)

A consistent estimator $\widehat{\tau}$ of τ is one whose proportion of mis-assigned nodes goes to zero (probabilistically) as n, the number of nodes goes to ∞. The issue of consistent partition for SBM has received much attention.

Inferential task for SBM graphs (continued)

1. [Snijders and Nowicki(1997)] provided an algorithm to consistently assign nodes to $K = 2$ blocks and [Condon and Karp(2001)] provided a consistent estimator for the planted l-partition model, a special type of SBM with equal-sized blocks.
2. [Bickel and Chen(2009)] showed that maximizing the Newman–Girvan modularity or the likelihood modularity provides consistent estimation of block membership and [Choi et al.(2012)] showed consistency for a MLE with rapidly growing numbers of blocks ($K = O(\sqrt{n})$).
3. [Rohe et al.(2011)] showed consistency for a spectral clustering algorithm (for undirected graphs) using the normalized Laplacian.
4. Maximizing modularities or likelihoods are computationally demanding and thus spectral methods provide a computationally appealing alternative.

SBM and random dot product graphs

An example of a class of SBM graphs is the class of random dot product graphs [Young and Scheinerman(2007)].

Let $\mathcal{U} = [u_1|u_2|\cdots|u_K]^T$ and $\mathcal{V} = [v_1|v_2|\cdots|v_K]^T$; $u_i, v_i \in \mathbb{R}^d$ for all i.

Let $\mathbf{X} = [X_1|X_2|\cdots|X_n]^T$ and $\mathbf{Y} = [Y_1|Y_2|\cdots|Y_n]^T$ be $n \times d$ matrix and the (X_i, Y_i) pairs are independent samples with X_i from $\mathcal{U}$ and Y_i from $\mathcal{V}$ according to a probability vector $\boldsymbol{\rho} > \mathbf{0}$.

If $\langle u_i, v_j \rangle \in [0, 1]$ for all $i, j \in [K]$, then $\mathbf{B} = \mathcal{U}\mathcal{V}^T (\mathbf{P} = \mathbf{X}\mathbf{Y}^T)$ and $\boldsymbol{\rho}$ defines a stochastic blockmodel.

Conversely, if $\mathbf{B}$ and $\boldsymbol{\rho}$ defines a stochastic blockmodel, then there exists $\mathcal{U}$ and $\mathcal{V}$ with $\mathrm{rank}(\mathcal{U}) = \mathrm{rank}(\mathcal{V}) = \mathrm{rank}(\mathbf{B})$ such that $\mathbf{B} = \mathcal{U}\mathcal{V}^T$.

Thus, if $\mathbf{P}$ (or an estimate $\widehat{\mathbf{P}}$) is known, then one can recover the block assignment τ (or an estimate $\widehat{\tau}$) via the SVD decomposition $\mathbf{P} = \mathbf{X}\mathbf{Y}^T (\widehat{\mathbf{P}} = \widehat{\mathbf{X}}\widehat{\mathbf{Y}}^T)$.

An adjacency-spectral partitioning algorithm

In their investigation of vertex nomination, [Marchette et al.(2011)] suggested the following spectral-paritioning algorithm. The algorithm makes no distinction between undirected and directed graphs.

Input: $\mathbf{A} \in \{0,1\}^{n \times n}$
Parameters: $d \in \{1, 2, \dots, n\}$, $K \in \{2, 3, \dots, n\}$
Output: $\widehat{\tau}\colon [n] \mapsto [K]$ the block assignment function.
Step 1: Compute the SVD of $\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{S}}\widehat{\mathbf{V}}^T$.
Step 2: Let $\widehat{\mathbf{S}}_d$ be the matrix corresponding to the d largest singular values. Let $\widehat{\mathbf{U}}_d$ and $\widehat{\mathbf{V}}_d$ be the corresponding columns of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$.
Step 3: Define $\widehat{\mathbf{Z}}$ as $[\widehat{\mathbf{U}}_d \widehat{\mathbf{S}}_d^{1/2} | \widehat{\mathbf{V}}_d \widehat{\mathbf{S}}_d^{1/2}]$.
Step 4: Solve the following clustering problem

$$(\widehat{\psi}, \widehat{\tau}) = \underset{\substack{\psi \in \mathbb{R}^{K \times 2d} \\ \tau\colon [n] \mapsto K}}{\operatorname{argmin}} \sum_{u=1}^{n} \| \widehat{Z}_u - \psi_{\tau(u)} \|^2 \tag{1}$$

where $\widehat{Z}_u$ is the u-th row of $\widehat{\mathbf{Z}}$.

A modicum of consistency

Even though $\mathbf{A}$ might not be a good estimator for $\mathbf{P}$ with respect to the Frobenius norm, i.e., $\|\mathbf{A} - \mathbf{P}\|_F$ is of order $O(n)$, $\mathbf{A}\mathbf{A}^T$ turns out to be a good estimator of $\mathbf{P}\mathbf{P}^T$. Specifically, we have

Proposition

Let $\mathbf{P}_n \in [0,1]^{n \times n}$ be a sequence of matrices and let $\mathbf{A}_n \in \{0,1\}^{n \times n}$ be a sequence of adjacency matrices corresponding to a sequence of random graphs on n nodes for $n \in \mathbb{N}$. Suppose the probability of an edge from u to v is given by $\mathbf{P}_n(u,v)$ and that the presence of edges are conditionally independent given $\mathbf{P}_n$. Then the following holds almost always:

$$\|\mathbf{A}_n\mathbf{A}_n^T - \mathbf{P}_n\mathbf{P}_n^T\|_F \leq \sqrt{3n^3 \log n}. \tag{2}$$

A modicum of consistency (continued)

By viewing $\mathbf{AA}^T$ as a pertubation of $\mathbf{PP}^T$, the special structure of SBM graphs along with an application of a Davis-Kahan theorem [Davis and Kahan(1970)] gives

Theorem ([Sussman et al.(2012)])

Let $\mathcal{S}_K$ be the set of permutations on $[K]$ and let $\hat{\tau}$ be the estimated block-assignment function. Suppose that the number of blocks K and $\mathrm{rank}(\mathbf{B})$ are known. Then it almost always hold that

$$\min_{\pi \in \mathcal{S}_K} |\{u \in V : \tau(u) \neq \pi(\hat{\tau}(u))\}| \leq C \log n. \tag{3}$$

for some constant C dependent on $\mathbf{B}$ and $\boldsymbol{\rho}$. The proportion of mis-assigned vertices thus goes to zero as n increases.

A modicum of consistency (continued)

The previous theorem requires the knowledge of K and $\mathrm{rank}(\mathbf{B})$. A more careful analysis gives the following stronger result (but with a weaker bound on the error rate).

Theorem ([Fishkind et al.(preprint)])

Let $\mathcal{S}_K$ be the set of permutations on $[K]$ and let $\widehat{\tau}$ be the estimated block-assignment function. Suppose that it is known that $\mathrm{rank}(\mathbf{B}) \leq R$ for some R. Then, for any $\epsilon > \frac{3}{4}$, it almost always hold that

$$\min_{\pi \in \mathcal{S}_K} |\{u \in V \colon \tau(u) \neq \pi(\widehat{\tau}(u))\}| \leq C n^{\epsilon} \tag{4}$$

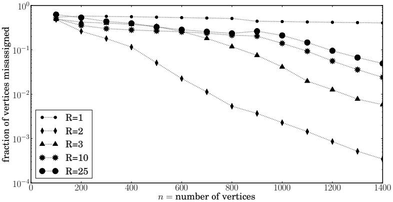
for some constant C dependent on $\mathbf{B}$ and $\boldsymbol{\rho}$. The proportion of mis-assigned vertices thus goes to zero as n increases.

Example 1: $\mathrm{rank}(\mathbf{B})$ and K unknown

$$\rho = (0.3, 0.3, 0.4)$$

$$\mathbf{B} = \begin{bmatrix} .205 & .045 & .150 \\ .045 & .205 & .150 \\ .150 & .150 & .180 \end{bmatrix}$$

$$K = 3, \quad \mathrm{rank}(\mathbf{B}) = 2$$

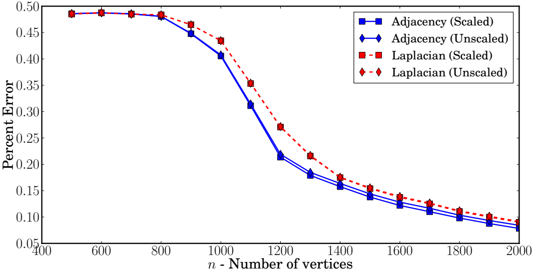


Example 2: Spectral-clustering on simulated data

$$\rho = (0.6, 0.4)$$

$$\mathbf{B} = \begin{bmatrix} .42 & .42 \\ .42 & .5 \end{bmatrix}$$

$$K = 2, \quad \mathrm{rank}(\mathbf{B}) = 2$$



A paired Wilcoxon test shows that the differences in performance between the adjacency-spectral and the normalized Laplacian clustering are statistically signficant for $n \geq 1400$.

Example 3: Spectral clusterings for a Wikipedia dataset

G is an undirected graph on $n = 1382$ nodes. Each node is a Wikipedia page related to "Algebraic geometry" and the edges corresponds to the presence of a hyperlink between the pages. The nodes were then manually assigned labels in $\{1, 2, \ldots, 5\}$.

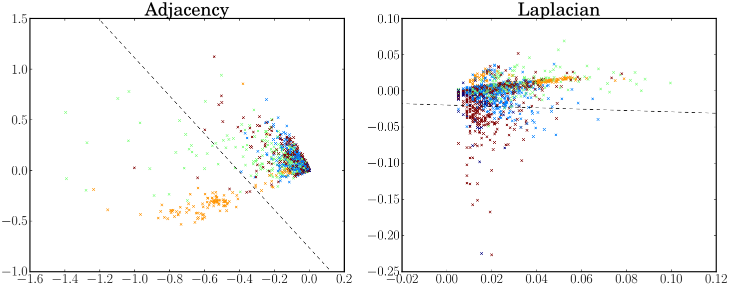


Figure: Scatter plots for the Wikipedia graph. Each point is colored according to the manually assigned labels. The dashed line represents the discriminant boundary determined by K-means with $K = 2$.

A latent space model

Let Ω be a compact subset of $\mathbb{R}^d$ and let F be an absolutely continuous probability measure on Ω. Suppose also that $\langle \omega_1, \omega_2 \rangle \in [0,1]$ for $\omega_1, \omega_2 \in \Omega$. For example, Ω is the unit simplex in $\mathbb{R}^d$ and F is the Dirichlet distribution with parameter $\boldsymbol{\alpha} > 0$.

Let G_n be a graph on n vertices generated as follows.

1. Let $\xi_1, \xi_2, \dots, \xi_n \overset{i.i.d}{\sim} F$.
2. Conditioned on the $\{\xi_i\}$, the edges of G_n are independent Bernoulli random variables with parameters $\langle \xi_u, \xi_v \rangle$ i.e., if we denote by $\mathbf{A}$ the adjacency matrix of G, then

$$\mathbb{P}[\mathbf{A} \mid \{\xi_u\}] = \prod_{u \neq v} (\langle \xi_u, \xi_v \rangle)^{\mathbf{A}(u,v)} (1 - \langle \xi_u, \xi_v \rangle)^{1-\mathbf{A}(u,v)}$$

G_n is then an instance of a latent space model. This latent space model is a random dot product model where the rows of $\mathbf{X} = [\xi_1 | \xi_2 | \cdots | \xi_n]^T$ are sampled from a continuous distribution F.

Latent position model and a classification problem

Let $F_{\xi,Y}$ be a probability measure on $\Omega \times \{0,1\}$ such that F_ξ is absolutely continuous on Ω. $F_{\xi,Y}$ then induces a classification problem.

We consider a sequence of (random) mappings $T_n \colon \Omega \mapsto \mathbb{R}^d$ for $n \geq 2$ defined as follows.

Sample $n-1$ points $\{\xi_i\}_{i=1}^{n-1} \overset{i.i.d}{\sim} F_\xi$. Let $\xi_n = \xi \sim F_\xi$ be given.

1. Let $\mathbf{X}_n = [\xi_1 | \xi_2 | \cdots | \xi_n]^T$.
2. Sample a graph G_n on n vertices from the latent space model with parameter $\mathbf{X}_n$.
3. Let $\widehat{\mathbf{X}}$ be the adjacency-spectral embedding of G_n and let $\{\widehat{\xi}_i\}_{i=1}^n$ denote the rows of $\widehat{\mathbf{X}}$.
4. $T_n(\xi) = T_n(\xi_n) = \widehat{\xi} := \widehat{\xi}_n$.

Example 4: $\widehat{\mathbf{X}}$ and LDA

Two-class problem for multivariate normal data

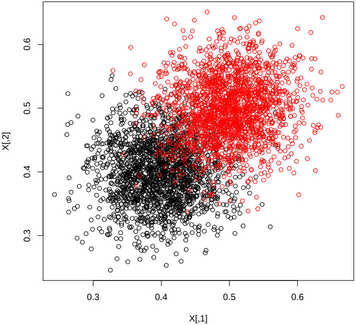
$$n = 1000, \quad \mu_1 = (0.4, 0.4), \quad \mu_2 = (0.5, 0.5), \quad \Sigma = \begin{bmatrix} 1/400 & 0 \\ 0 & 1/400 \end{bmatrix}$$

$$\{\xi_i\}_{i=1}^{n} \sim \mathcal{N}(\mu_1, \Sigma), \quad Y_i \equiv 0 \quad \text{for } 1 \leq i \leq n$$

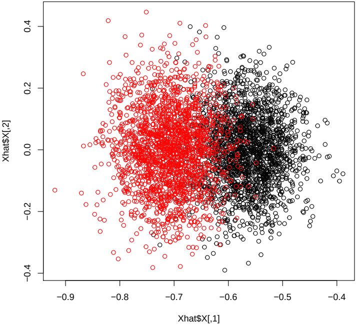
$$\{\xi_i\}_{i=n+1}^{2n} \sim \mathcal{N}(\mu_2, \Sigma), \quad Y_i \equiv 1 \quad \text{for } n+1 \leq i \leq 2n$$

Let $\mathbf{X} = \{\xi_i\}_{i=1}^{2n}$ be the sampled points. Let $\mathbf{A}$ be the adjacency matrix of an instance of the latent space model with parameter $\mathbf{X}$ and let $\widehat{\mathbf{X}} = \{\widehat{\xi_i}\}_{i=1}^{2n}$ be the resulting embedding of $\mathbf{A}$ via the adjacency-spectral partitioning algorithm. We train a LDA classifier on $\mathbf{X}$ using 1000 training data points, 500 from each class, and test the resulting classifier on the remaining 1000 data points, again with 500 from each class. We also train a LDA classifier on $\widehat{\mathbf{X}}$ with the same settings.

Example 4: $\widehat{\mathbf{X}}$ and LDA (continued)



(a) Scatter plot for $\mathbf{X}$



(b) Scatter plot for $\widehat{\mathbf{X}}$

Figure: Scatter plots for $\mathbf{X}$ and $\widehat{\mathbf{X}}$ for the current example. The error rate for a LDA classifier trained and test on $\mathbf{X}$ is 0.077. The error rate for a LDA classifier trained and test on $\widehat{\mathbf{X}}$ is 0.085. The Bayes-risk is $\Phi(-\sqrt{2}) \approx 0.0786$.

$L^*_{T_n(\xi)} \xrightarrow{p} L^*_\xi$

Theorem

Let $(\xi, Y) \sim F_{\xi,Y}$ *where* $F_{\xi,Y}$ *is a distribution function on* $\Omega \times \{0,1\}$*. Suppose that* $\langle \omega_1, \omega_2 \rangle \in [0,1]$ *for almost every* $\omega_1, \omega_2 \in \Omega$*. Let* $\eta(x) = \mathbb{P}[Y = 1 | \xi = x]$ *and suppose that* η *is also continuous. Let* L^*_ξ *be the Bayes-error for* ξ *and let* $L^*_{T_n(\xi)}$ *be the Bayes-error for the transformation* $T_n \colon \Omega \mapsto \mathbb{R}^d$*. We then have, as* $n \to \infty$

$$L^*_{T_n(\xi)} \xrightarrow{p} L^*_\xi \tag{5}$$

References I

P.J. Bickel and A. Chen.

A nonparametric view of network models and newman-girvan and other modularities.

Proceedings of the National Academy of Sciences of the United States of America, 106(50):21068–73, 2009.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi.

Stochastic blockmodels with growing number of classes.

Biometrika, 2012.

A. Condon and R.M. Karp.

Algorithms for graph partitioning on the planted partition model.

Random Structures and Algorithms, 18(2):116–140, 2001.

References II

C. Davis and W. Kahan.

The rotation of eigenvectors by a pertubation. III.

Siam Journal on Numerical Analysis, 7:1–46, 1970.

D. E. Fishkind, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe.

Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown.

arxiv.org/abs/1205.0309.

References III

M.S. Handcock, A.E. Raftery, and J.M. Tantrum.

Model-based clustering for social networks.

Journal of the Royal Statistical Society: Series A (Statistics in Society), 170:301–354, 2007.

P. Hoff, A. Rafferty, and M. Handcock.

Latent space approaches to social network analysis.

Journal of the American Statistical Association, 97:1090–1098, 2002.

P. W. Holland, K. Laskey, and S. Leinhardt.

Stochastic blockmodels: First steps.

Social Networks, 5:109–317, 1983.

References IV

D. Marchette, C. Priebe, and G. Coppersmith.

Vertex nomination via attributed random dot product graphs.

In *Proceedings of the 57th ISI World Statistics Congress*, 2011.

K. Rohe, S. Chatterjee, and B. Yu.

Spectral clustering and the high-dimensional stochastical block model.

Annals of Statistics, 39:1878–1915, 2011.

T. Snijders and K. Nowicki.

Estimation and prediction for stocchastic block models for graphs with latent block structure.

Journal of Classification, 14:75–100, 1997.

References V

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, 2012.

Y. J. Wang and G. Y. Wong.

Stochastic blockmodels for directed graphs.

Journal of the American Statistical Association, 82:8–19, 1987.

S. Young and E. Scheinerman.

Random dot product models for social networks.

In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149, 2007.

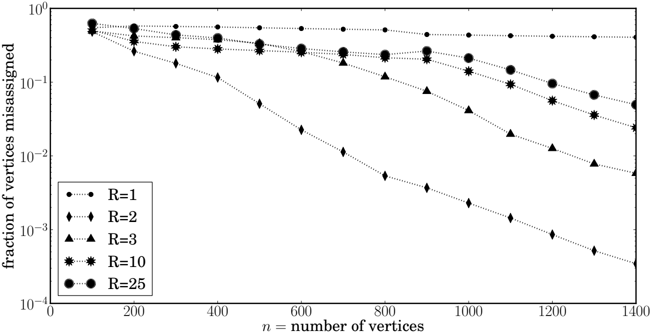


Figure: Adjacency spectral partitioning when $\mathrm{rank}(\mathbf{B})$ and K are unknown.