Statistical and Computational Challenges in
Networks and Cybersecurity

Community Detection and Classification in Hierarchical Stochastic Blockmodels

Carey E. Priebe

Department of Applied Mathematics & Statistics
Johns Hopkins University, Baltimore, MD, USA

May 4-8, 2015

Abstract

We propose a robust, scalable, integrated methodology for *community detection* and *community comparison* in graphs. In our procedure, we first embed a graph into an appropriate Euclidean space to obtain a low-dimensional representation, and then cluster the vertices into communities. We next employ nonparametric graph inference techniques to identify structural similarity among these communities. These two steps are then applied recursively on the communities, allowing us to detect more fine-grained structure. We describe a *hierarchical stochastic blockmodel*—namely, a stochastic blockmodel with a natural hierarchical structure—and establish conditions under which our algorithm yields consistent estimates of model parameters and *motifs*, which we define to be stochastically similar groups of subgraphs. Finally, we demonstrate the effectiveness of our algorithm in both simulated and real data. Specifically, we address the problem of locating similar subcommunities in a partially reconstructed *Drosophila* connectome and in the social network Friendster.

`http://www.cis.jhu.edu/~parky/HSBM/`

V.Lyzinski

M.Tang

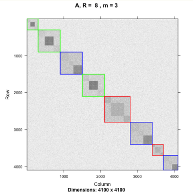
A.Athreya

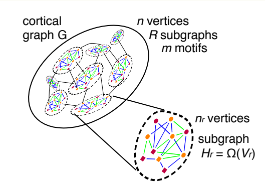
Y.Park

Introduction

Problem

In disciplines as diverse as social network analysis and neuroscience, graphs are believed to be composed of loosely connected smaller graph primitives.





Goal

To develop theoretically-sound robust, scalable, integrated methodology for hierarchical community detection and community classification in graphs.

Background

Definition (Random Dot Product Graph (RDPG))

Let F be a distribution on a set $\mathcal{X} \subset \mathbb{R}^d$ such that $\langle x, x' \rangle \in [0, 1]$ for all $x, x' \in \mathcal{X}$. We say that $(A, X) \sim \mathrm{RDPG}(F)$ is an instance of a random dot product graph (RDPG) if $X = [X_1, \ldots, X_n]^\top$ with $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$, and $A \in \{0, 1\}^{n \times n}$ is a symmetric hollow matrix satisfying

$$\mathbb{P}[A|X] = \prod_{i>j} (X_i^\top X_j)^{A_{ij}} (1 - X_i^\top X_j)^{1 - A_{ij}}.$$

Background

Definition (Stochastic Blockmodel (SBM))

We say that an n vertex graph $(A, X) \sim \mathrm{RDPG}(F)$ is a (positive semidefinite) stochastic blockmodel (SBM) with K blocks if the distribution F is a mixture of K point masses,

$$F = \sum_{i=1}^{K} \pi(i)\delta_{\xi_i},$$

where $\vec{\pi} \in (0,1)^K$ satisfies $\sum_i \pi(i) = 1$, and the distinct latent positions are given by $\xi = [\xi_1, \xi_2, \ldots, \xi_K]^\top \in \mathbb{R}^{K \times d}$. In this case, we write $G \sim SBM(n, \vec{\pi}, \xi\xi^\top)$, and we refer to $\xi\xi^\top \in \mathbb{R}^{K,K}$ as the *block probability matrix* of G.

Background

Definition (Hierarchical Stochastic Blockmodel (HSBM))

We say that $(A, X) \sim \mathrm{RDPG}(F)$ is an instantiation of a D-dimensional hierarchical stochastic blockmodel if F can be written as the mixture

$$F = \sum_{i=1}^{R} \pi(i) F_i,$$

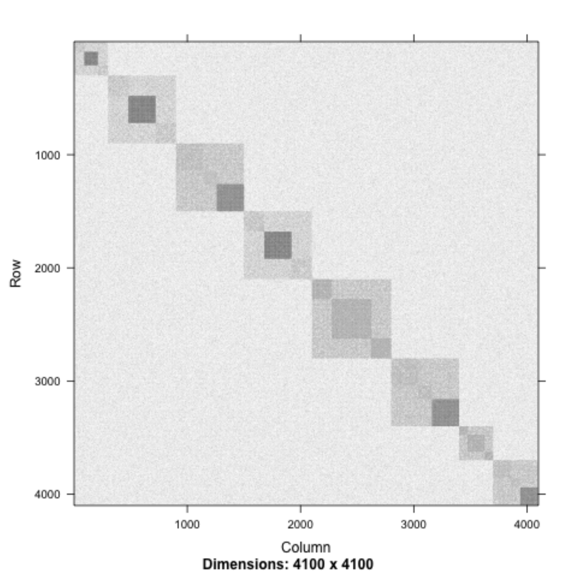
where $\vec{\pi} \in (0,1)^R$ satisfies $\sum_i \pi(i) = 1$, and for each $i \in [R]$, F_i is itself a mixture of point mass distributions

$$F_i = \sum_{j=1}^{K} \pi_i(j) \delta_{\xi^{(i)}(j,:)}$$

where $\vec{\pi}_i \in (0,1)^K$ satisfies $\sum_j \pi_i(j) = 1$. The distinct latent positions $\xi = [(\xi^{(1)})^\top | \cdots | (\xi^{(R)})^\top]^\top \in \mathbb{R}^{RK \times D}$ further satisfy $\langle \xi^{(i)}(\ell,:), \xi^{(j)}(h,:) \rangle \leqslant p$ for $1 \leqslant i \neq j \leqslant R$ and $\ell, h \in [K]$. We then write

$$G \sim \mathsf{HSBM}(n, \vec{\pi}, \{\vec{\pi}_i\}_{i=1}^{R}, \xi\xi^\top).$$

Background


Row
1000
2000
3000
4000
1000
2000
3000
4000
Column
Dimensions: 4100 x 4100

Background

Denoting by $J_{K,d}$ the $K \times d$ matrix of all ones, we write

$$\xi = \begin{bmatrix} \xi^{(1)} \\ \xi^{(2)} \\ \vdots \\ \xi^{(R)} \end{bmatrix} = \begin{bmatrix} \chi_1 & \delta J_{K,d} & \cdots & \delta J_{K,d} \\ \delta J_{K,d} & \chi_2 & \cdots & \delta J_{K,d} \\ \vdots & \vdots & \ddots & \vdots \\ \delta J_{K,d} & \delta J_{K,d} & \cdots & \chi_R \end{bmatrix} \in \mathbb{R}^{RK \times D}, \qquad (1)$$

where for each $i \in [R]$, $\chi_i \in \mathbb{R}^{K \times d}$, and

$$\delta := \frac{\sqrt{d + pd(R-2)} - \sqrt{d}}{d(R-2)}$$

is chosen to make the off block-diagonal elements of the corresponding edge probability matrix $\xi\xi^T$ bounded above by an absolute constant p. In this setting, for each $i \in [R]$ the latent positions

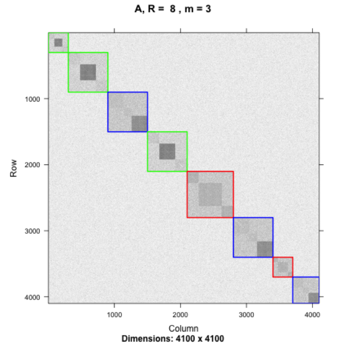
$$\xi^{(i)} := \left[\delta J_{K,d(i-1)} \quad \chi_i \quad \delta J_{K,d(R-i)}\right] \in \mathbb{R}^{K \times D}, \qquad (2)$$

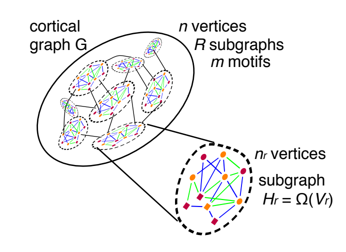
are those associated with H_i, the i-th induced SBM subgraph of G.

Background

Definition (Motifs)

Let $(A, X) \sim RDPG(F)$ and $(B, Y) \sim RDPG(G)$. We say that A and B are of the same *motif* if there exists a unitary transformation U such that $F = G \circ U$.





Main algorithm

The following algorithm is our methodology for identifying and estimating the structural properties of repeated motifs.

Algorithm 1 Detecting hierarchical structure for graphs

```
Input: Adjacency matrix $A \in \{0,1\}^{n \times n}$ for a latent position random graph.
Output: Subgraphs and characterization of their dissimilarity
while Cluster size exceeds threshold do
    Step 1: Compute the adjacency spectral embedding $\widehat{X}$;
    Step 2: Project the rows of $\widehat{X}$ onto the sphere yielding $\widehat{Y}$; i.e., for each $i \in [n]$,
        $\widehat{Y}_i := \widehat{X}_i / \|\widehat{X}_i\|_2$;
    Step 3: Cluster $\widehat{Y}$ to obtain subgraphs $\widehat{H}_1, \cdots, \widehat{H}_R$;
    Step 4: For each $i \in [R]$, use ASE to re-embed $\widehat{H}_i$, obtaining $\widehat{X}_{\widehat{H}_i}$;
    Step 5: Compute $\widehat{S} := [T_{\hat{n}_r, \hat{n}_s}(\widehat{X}_{\widehat{H}_r}, \widehat{X}_{\widehat{H}_s})]$ producing a pairwise dissimilarity matrix on induced subgraphs;
    Step 6: Cluster induced subgraphs into motifs according to $\widehat{S}$;
    Step 7: Recurse on each motif;
end while
```

Background

Definition (Adjacency Spectral Embedding (ASE))

Given an adjacency matrix $A \in \{0,1\}^{n \times n}$ of a d-dimensional RDPG(F), the *adjacency spectral embedding* of A into $\mathbb{R}^d$ is given by $\widehat{X} = U_A S_A^{1/2}$ where

$$|A| = [U_A|\tilde{U}_A][S_A \oplus \tilde{S}_A][U_A|\tilde{U}_A]$$

is the spectral decomposition of $|A| = (A^\top A)^{1/2}$, S_A is the diagonal matrix with the (ordered) d largest eigenvalues of $|A|$ on its diagonal, and $U_A \in \mathbb{R}^{n \times d}$ is the matrix whose columns are the corresponding orthonormal eigenvectors of $|A|$.

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*, vol. 107, no. 499, pp. 119-1128, 2012.

Background

Theorem

Let $(A, X) \sim RDPG(F)$ *and* $(B, Y) \sim RDPG(G)$ *be* d*-dimensional random dot product graphs. Consider the hypothesis test*

$$H_0 \colon F = G \circ U \quad \textit{against} \quad H_A \colon F \neq G \circ U$$

Denote by $\widehat{X} = \{\widehat{X}_1, \ldots, \widehat{X}_n\}$ *and* $\widehat{Y} = \{\widehat{Y}_1, \ldots, \widehat{Y}_m\}$ *the adjacency spectral embedding of* A *and* B*, respectively. Define the test statistic* $T_{n,m} = T_{n,m}(\widehat{X}, \widehat{Y})$ *as follows:*

$$\begin{aligned} T_{n,m}(\widehat{X}, \widehat{Y}) &= \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(\widehat{X}_i, \widehat{X}_j) \\ &\quad - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \kappa(\widehat{X}_i, \widehat{Y}_k) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(\widehat{Y}_k, \widehat{Y}_l) \end{aligned} \tag{3}$$

where κ *is a radial basis kernel, e.g.,* $\kappa = \exp(-\| \cdot - \cdot \|^2 / \sigma^2)$*.*

Background

Theorem (cont.)

Suppose that $m, n \to \infty$ and $m/(m+n) \to \rho \in (0,1)$. Then under the null hypothesis of $F = G \circ U$,

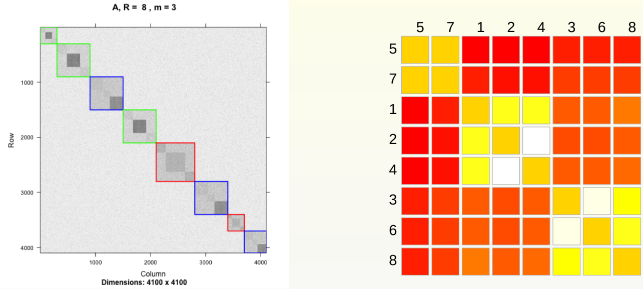
$$(m+n)(T_{n,m}(\widehat{X}, \widehat{Y}) - T_{n,m}(X, YW)) \xrightarrow{\text{a.s.}} 0 \tag{4}$$

where W is any orthogonal matrix such that $F = G \circ W$. In addition, under the alternative hypothesis of $F \neq G \circ U$, there exists an orthogonal matrix $W \in \mathbb{R}^{d \times d}$, depending on F and G but independent of m and n, such that

$$\frac{(m+n)}{\log^2(m+n)}(T_{n,m}(\widehat{X}, \widehat{Y}) - T_{n,m}(X, YW)) \xrightarrow{\text{a.s.}} 0. \tag{5}$$

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe,
"A nonparametric two-sample hypothesis for random dot product graphs," 2014,
arXiv preprint. `http://arxiv.org/abs/1403.7249`.

Simulation



(a)

(b)

(a) Depiction of the adjacency matrix of a two-level HSBM graph with 3 distinct motifs. The subgraphs corresponding to these motifs are outlined in blue (H_3, H_6, H_8), green (H_1, H_2, H_4), and red (H_5, H_7).

(b) Heatmap depicting the dissimilarity matrix $\widehat{S}$ produced by Algorithm 1 for the 2-level HSBM.

Spherical R-means Clustering

Theorem

Suppose G is a hierarchical stochastic blockmodel whose latent position structure is of the form in Eq. (1). Suppose that R is fixed and the $\{H_r\}$ correspond to M different motifs, i.e., the set $\{\chi_1, \chi_2, \dots, \chi_R\}$ has $M \leqslant R$ distinct elements. Assume that the constants defined above satisfy (with $\pi_{\min} := \min_i \pi(i)$)

$$\delta_D > 0; \tag{6}$$

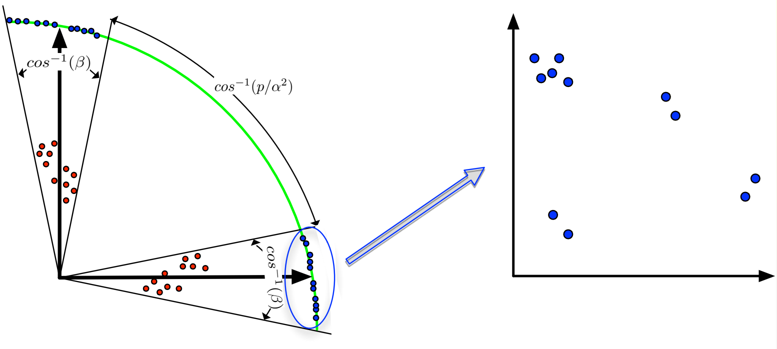
$$\sqrt{1 - \frac{p}{\alpha^2}} > \left(2 + \frac{1}{\sqrt{\pi_{\min}}}\right)\sqrt{1-\beta}. \tag{7}$$

Let c be arbitrary. There exists a constant $n_0 = n_0(c)$ such that if $n > n_0$, then for any η satisfying $n^{-c} < \eta < 1/2$, the procedure in Algorithm 1 yields consistent estimates $\widehat{H}_1, \cdots, \widehat{H}_R$ for $H_1, \cdots, H_R$ and $\widehat{S}$ for S with probability greater than $1 - \eta$.

Corollary

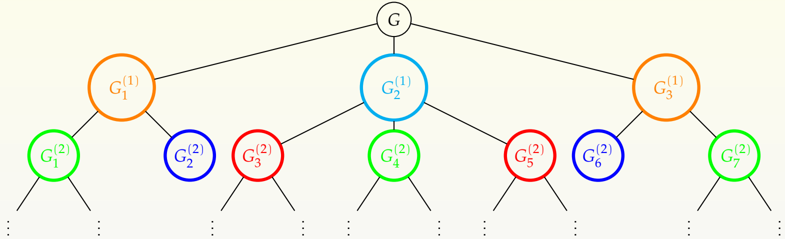
Clustering the matrix of p-values associated with $\widehat{S}$ yields a consistent clustering of $\{\widehat{H}_i\}_{i=1}^{R}$ into motifs.

Spherical R-means Clustering



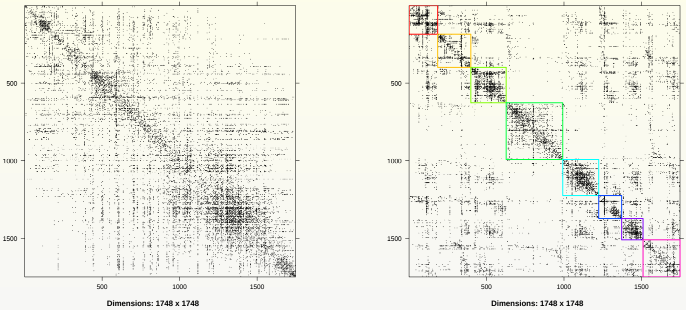
Notional illustration of projection onto the sphere after embedding and its effect on K-means clustering. The embedded points are colored red prior to projection and colored blue after projection. Eq. (7) specified that the angles between the blocks after projection (bounded from below by $\cos^{-1}(p/\alpha^2)$) should be sufficiently large when compared to the angles within the blocks (bounded from above by $\cos^{-1}(\beta)$).

Hierarchical Graph Structure



Notional depiction of a general hierarchical graph structure. The colored nodes in the first and second level of the tree (below the root node) correspond to induced subgraphs and associated motifs.

Example: Fly Structural Data

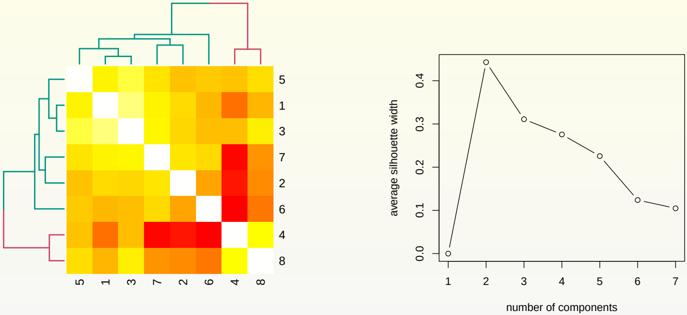


(a) (b)

Visualization of our method applied to the *Drosophila* connectome. We show (a) the adjacency matrix, (b) the clustering derived via ASE, projection to the sphere and k-means clustering.

S. Takemura, el al. "A visual motion detection circuit suggested by drosophila connectomics," *Nature*, vol. 500, no. 7461, pp. 175-181, 2013.

Example: Fly Structural Data



(c)

(d)

Visualization of our method applied to the *Drosophila* connectome. (c) $\widehat{S}$ calculated from these clusters. (d) Average silhouette width of clustering $\widehat{S}$ into k motifs. By this measure, clustering the subgraphs based on this $\widehat{S}$ suggests two repeated motifs: $\{1, 2, 3, 5, 6, 7\}$ and $\{4, 8\}$.

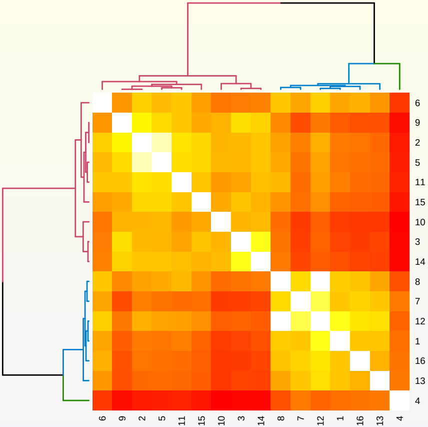
Example: Friendster Network Data

http://snap.stanford.edu/data/com-Friendster.html

Dataset Statstics	
Nodes	65,608,366
Edges	1,806,067,135
Average clustering coefficient	0.1623
Number of triangles	4,173,724,142
Fraction of closed triangles	0.005859
Diameter (longest shortest path)	32
90-percentile effective diameter	5.8

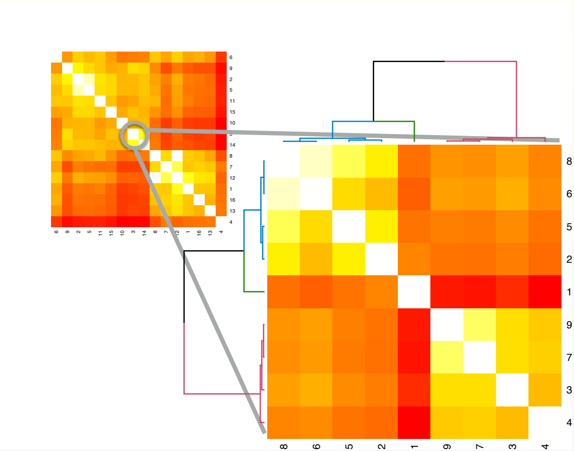
D. Zheng, D. Mhembere, R. Burns, J.T. Vogelstein, C.E. Priebe, and A.S. Szalay, *"Flashgraph: Processing billion-node graphs on an array of commodity SSDs,"* in 13th USENIX Conference on File and Storage Technologies (FAST 15), Santa Clara, CA, Feb. 2015, pp. 45-58. https://www.usenix.org/conference/fast15/technical-sessions/presentation/zheng

Example: Friendster Network Data



Heat map depiction of the **level one** Friendster estimated dissimilarity matrix $\widehat{S} \in \mathbb{R}^{16\times16}$. In addition, we cluster $\widehat{S}$ using hierarchical clustering and display the associated hierarchical clustering dendrogram.

Example: Friendster Network Data



Heat map depiction of the **level two** Friendster estimated dissimilarity matrix $\widehat{S} \in \mathbb{R}^{9\times 9}$ of $\widehat{H}_3$. In addition, we cluster $\widehat{S}$ using hierarchical clustering and display the associated hierarchical clustering dendrogram.

Yogi Berra:

"In theory there is no difference between theory and practice. In practice, there is."

Leopold Kronecker to Hermann von Helmholtz (1888):

“The wealth of your practical experience
with sane and interesting problems
will give to mathematics
a new direction and a new impetus.”

Kronecker

Helmholtz