Community Estimation in General Multilayer Stochastic Block Models

Jing Lei
Department of Statistics & Data Science
Carnegie Mellon University

2022.06.02

Simons Big Data Workshop in Honor of Peter Bickel

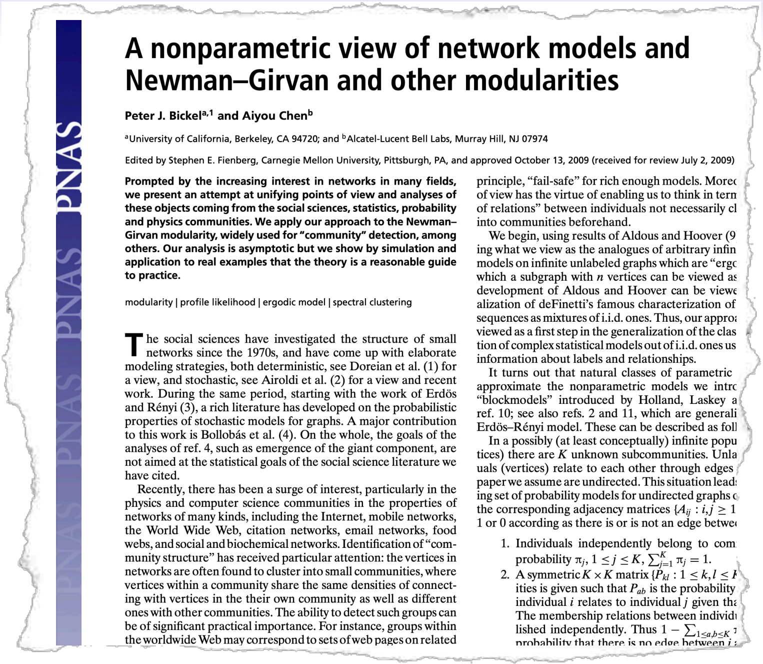
Collaborator

Kevin Z. Lin

- Former student at CMU
- Currently postdoc at U. Penn.

Network Data

- Network data record interactions (edges) between individuals (nodes).
- From WIKIPEDIA: "... a complex network is a graph (network) with non-trivial topological features ..."
- Examples of "non-trivial topological features"
  - heavy-tail degree distribution (a.k.a "scale-free", "power law")
  - large clustering coefficient (transitivity)
  - community structure: the nodes can be grouped into subsets with similar connectivity.
  - . . .



PNAS PNAS PNAS PNAS

A nonparametric view of network models and Newman–Girvan and other modularities

Peter J. Bickel[a,1] and Aiyou Chen[b]

[a]University of California, Berkeley, CA 94720; and [b]Alcatel-Lucent Bell Labs, Murray Hill, NJ 07974

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved October 13, 2009 (received for review July 2, 2009)

Prompted by the increasing interest in networks in many fields, we present an attempt at unifying points of view and analyses of these objects coming from the social sciences, statistics, probability and physics communities. We apply our approach to the Newman–Girvan modularity, widely used for "community" detection, among others. Our analysis is asymptotic but we show by simulation and application to real examples that the theory is a reasonable guide to practice.

modularity | profile likelihood | ergodic model | spectral clustering

The social sciences have investigated the structure of small networks since the 1970s, and have come up with elaborate modeling strategies, both deterministic, see Doreian et al. (1) for a view, and stochastic, see Airoldi et al. (2) for a view and recent work. During the same period, starting with the work of Erdös and Rényi (3), a rich literature has developed on the probabilistic properties of stochastic models for graphs. A major contribution to this work is Bollobás et al. (4). On the whole, the goals of the analyses of ref. 4, such as emergence of the giant component, are not aimed at the statistical goals of the social science literature we have cited.

Recently, there has been a surge of interest, particularly in the physics and computer science communities in the properties of networks of many kinds, including the Internet, mobile networks, the World Wide Web, citation networks, email networks, food webs, and social and biochemical networks. Identification of "community structure" has received particular attention: the vertices in networks are often found to cluster into small communities, where vertices within a community share the same densities of connecting with vertices in the their own community as well as different ones with other communities. The ability to detect such groups can be of significant practical importance. For instance, groups within the worldwide Web may correspond to sets of web pages on related

principle, "fail-safe" for rich enough models. Moreo
of view has the virtue of enabling us to think in term
of relations" between individuals not necessarily cl
into communities beforehand.

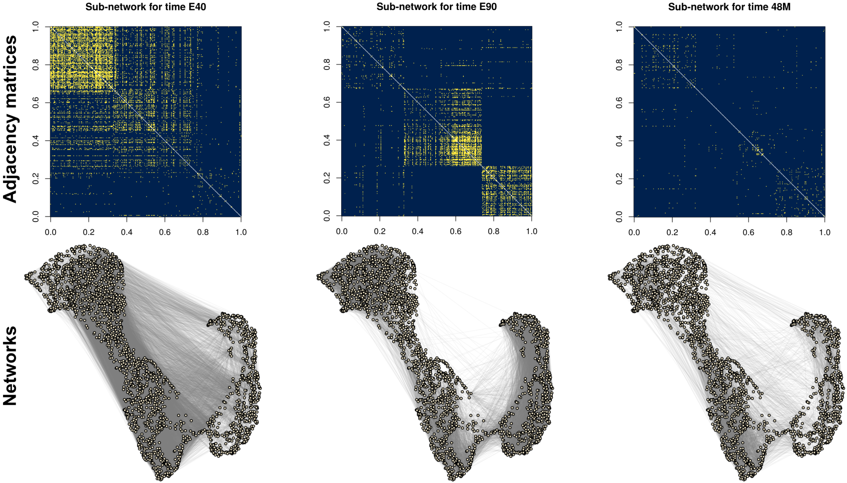
We begin, using results of Aldous and Hoover (9
ing what we view as the analogues of arbitrary infin
models on infinite unlabeled graphs which are "ergo
which a subgraph with n vertices can be viewed as
development of Aldous and Hoover can be viewe
alization of deFinetti's famous characterization of
sequences as mixtures of i.i.d. ones. Thus, our approa
viewed as a first step in the generalization of the clas
tion of complex statistical models out of i.i.d. ones us
information about labels and relationships.

It turns out that natural classes of parametric
approximate the nonparametric models we intro
"blockmodels" introduced by Holland, Laskey a
ref. 10; see also refs. 2 and 11, which are generali
Erdös–Rényi model. These can be described as foll

In a possibly (at least conceptually) infinite popu
tices) there are K unknown subcommunities. Unla
uals (vertices) relate to each other through edges
paper we assume are undirected. This situation lead
ing set of probability models for undirected graphs o
the corresponding adjacency matrices $\{A_{ij} : i, j \geq 1$
1 or 0 according as there is or is not an edge betwe

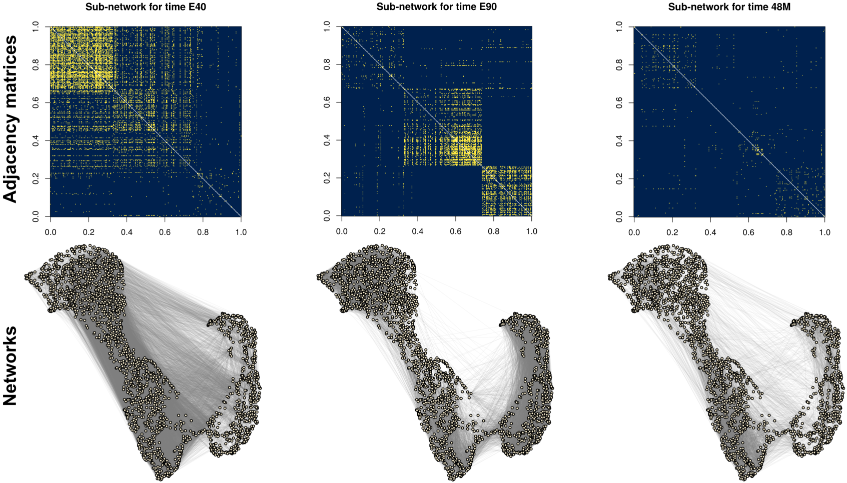
1. Individuals independently belong to com
probability π_j, $1 \leq j \leq K$, $\sum_{j=1}^{K} \pi_j = 1$.
2. A symmetric $K \times K$ matrix $\{P_{kl} : 1 \leq k, l \leq$
ities is given such that P_{ab} is the probability
individual i relates to individual j given tha
The membership relations between individu
lished independently. Thus $1 - \sum_{1 \leq a,b \leq K}$
probability that there is no edge between i

Multilayer Network Data: An Example



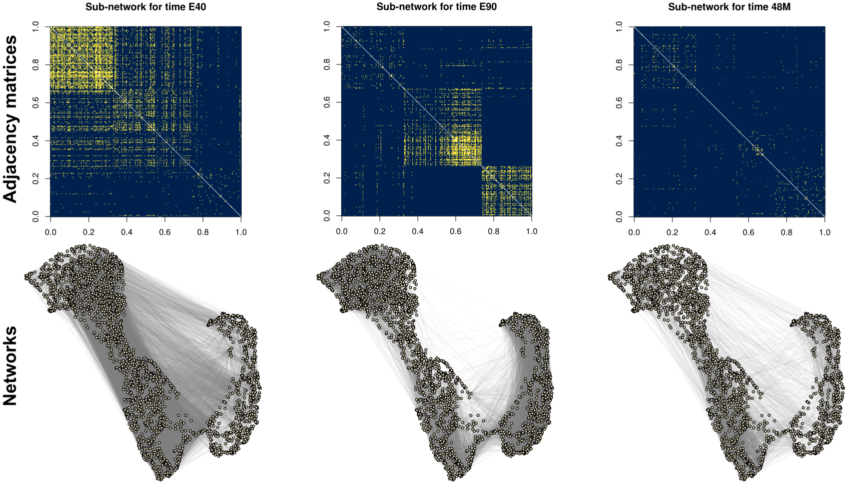
Temporal gene co-expression networks in the medial prefrontal cortex of rhesus monkeys (10 layers, ∼8k nodes, Bakken et al ’16, Liu et al ’18).

Multilayer Network Data: An Example



Visual inspection suggests four groups. But no single network can distinguish all four groups.

Multilayer Network Data: An Example



Must aggregate the layers in an informative way to reveal the complete group structure.

Multilayer Stochastic Block Models

- n nodes; K groups; L layers
- Let B_1,..., B_L be $K \times K$ symmetric matrices with all entries in $[0,1]$.
- Let $g \in \{1,\ldots,K\}^n$ be a membership vector, same for all layers.
- For each $1 \leq l \leq L$, $1 \leq i < j \leq n$

  $$A_{l,ij} \sim \text{Bernoulli}(\rho B_{l,g_i g_j})$$

  independent of everything else, where $\rho \in (0,1)$ is a global sparsity parameter.
- Goal: Estimate g from multilayer adjacency matrices $(A_l : 1 \leq l \leq L)$.

Singe layer SBMs

- When $L = 1$, this reduces to the well-known stochastic block model (Holland et al 1983).
- Community recovery is well understood for single layer SBMs (Bickel & Chen ’09, L. & Rinaldo ’14, Abbe & Sandon ’15, and many more...)
- When $L = 1$, K is constant, and group sizes are balanced, consistent estimation of g is possible if and only if

$$n\rho \to \infty\,.$$

- How will multiple layers facilitate community estimation?

Multilayer SBM with layer-wise positivity

- If each B_l is positive definite with minimum eigenvalue bounded away from 0, then consistent community recovery is possible if

  $$n\rho L \to \infty.$$

  For example, consider the summed adjacency matrix $A = \sum_{l=1}^{L} A_l$ and use variants of spectral clustering (Paul & Chen 2017, Bhattacharyya & Chatterjee, 2018).
- Dynamic SBM and smoothing: Pensky and Zhang ’19, Pensky ’19.
- Layer aggregation and denoising: Levin, Lodhia and Levina ’19.
- Global and local clustering: Chen, Liu and Ma ’20.

This work: aggregating general multilayer SBMs

- When the layer-wise positivity assumption is dropped, summing up the layers may lose signal.
- Example: All entries of B_l's are iid $\mathrm{U}(0,1)$, subject to symmetry.
- How to aggregate general multilayer SBM's?
- What is the threshold for consistent estimation in general multilayer SBM's?

First approach: least squares

- Least squares estimate:
  $(\hat{g}_{\text{ls}}, \hat{B}_{\text{ls}}) = \arg\min_{g,B} \sum_{l=1}^{L} \sum_{1 \le i < j \le n} (B_{l,g_i g_j} - A_{l,ij})^2$

Theorem (L., Chen, & Lynch 2020)

If $K = O(1)$, $L = O(n)$, community sizes are balanced, and $L^{-1} \sum_{l=1}^{L} B_l^2 \succeq cI$ for some $c > 0$, then, with probability tending to 1, the least squares estimate satisfies

$$n^{-1} D_{\text{Ham}}(\hat{g}_{\text{ls}}, g) \le \text{Const.} \times \frac{(\log n)^{3/2}}{n\rho L^{1/2}},$$

where $D_{\text{Ham}}(\hat{g}, g) = \min_{\sigma:[K]\to[K]} \sum_{i=1}^{n} \mathbb{1}(\hat{g}_i \neq \sigma(g_i))$ is the Hamming distance.

As a consequence, $\hat{g}_{\text{ls}}$ is consistent if $n\rho L^{1/2}/(\log n)^{3/2} \to \infty$.

From least squares to spectral clustering

- Let $G = [G_1, \ldots, G_K]$ be the normalized membership matrix.
- Example: $n = 5$, $K = 2$,

$$g = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix} \Leftrightarrow G = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$$

- Similar to the total variance decomposition in ANOVA, the least squares problem is approximately

$$\max_G \sum_{l=1}^{L} \|G^T A_l G\|_F^2 \approx \max_G \operatorname{tr}\left[G^T \left(\sum_{l=1}^{L} A_l^2\right) G\right].$$

Spectral clustering on sum of squares

- Now the least squares is approximately $\max_G \mathrm{tr}\left[G^T\left(\sum_{l=1}^L A_l^2\right)G\right]$, over all normalized membership matrices.
- Relaxing the requirement of normalized membership matrix: Let $U = [U_1, ..., U_K]$ be the top K eigenvectors of $\sum_{l=1}^L A_l^2$.
- Let $\hat{g}$ be the output of a clustering algorithm applied to the rows of U.
- How does it work?

Spectral clustering

- In general, let $A = S + N$ where S is a signal matrix whose leading eigenvectors contain useful information, and N is a noise matrix.
- In order for the leading eigenvectors of A to carry similar information as those of S, the eigengap of S needs to dominate $\|N\|$.

An error decomposition in squared SBM

- Let $P_l = \mathbb{E}A_l \approx \rho G B_l G^T$ (except diagonal), $E_l = A_l - P_l$, then

$$\begin{aligned} A_l^2 =& (P_l + E_l)^2 \\ =& P_l^2 + P_l E_l + E_l P_l + E_l^2 \,. \end{aligned}$$

- The firs term P_l^2 is the signal term.
- The second and third terms $P_l E_l$, $E_l P_l$ are mean zero noise terms, so that $\| \sum_{l=1}^{L} P_l E_l \| \lesssim \sqrt{L} \| P_1 E_1 \|$ (matrix Bernstein).
- The last term E_l^2 has positive expectation, so that $\| \sum_{l=1}^{L} E_l^2 \| \asymp L \| E_1^2 \|$. This is the bias term.

The bias term

- $(E_l^2)_{ij} = \sum_{k \neq i,j} E_{l,ik} E_{l,jk} = \sum_{k \neq i,j} (A_{l,ik} - P_{l,ik})(A_{l,jk} - P_{l,jk})$
- $\mathbb{E}(E_l^2)_{ij} = 0$ if and only if $i \neq j$.
- When $i = j$,

$$\begin{aligned} (E_l^2)_{ii} &= \sum_{k \neq i} E_{l,ik}^2 = \sum_{k \neq i} (A_{l,ik} - P_{l,ik})^2 \\ &= \sum_{k \neq i} (A_{l,ik}^2 - 2A_{l,ik} P_{l,ik} + P_{l,ik}^2) \end{aligned}$$

- The leading term is $\sum_{k \neq i} A_{l,ik} \equiv d_{l,i}$, the degree of node i in layer l, which can be computed from the data and hence can be removed.

The bias term

- $(E_l^2)_{ij} = \sum_{k \neq i,j} E_{l,ik} E_{l,jk} = \sum_{k \neq i,j} (A_{l,ik} - P_{l,ik})(A_{l,jk} - P_{l,jk})$
- $\mathbb{E}(E_l^2)_{ij} = 0$ if and only if $i \neq j$.
- When $i = j$,

$$
\begin{aligned}
(E_l^2)_{ii} &= \sum_{k \neq i} E_{l,ik}^2 = \sum_{k \neq i} (A_{l,ik} - P_{l,ik})^2 \\
&= \sum_{k \neq i} ({\color{red}A_{l,ik}} - 2A_{l,ik} P_{l,ik} + P_{l,ik}^2)
\end{aligned}
$$

- The leading term is $\sum_{k \neq i} A_{l,ik} \equiv d_{l,i}$, the degree of node i in layer l, which can be computed from the data and hence can be removed.

Bias-adjusted squared spectral clustering

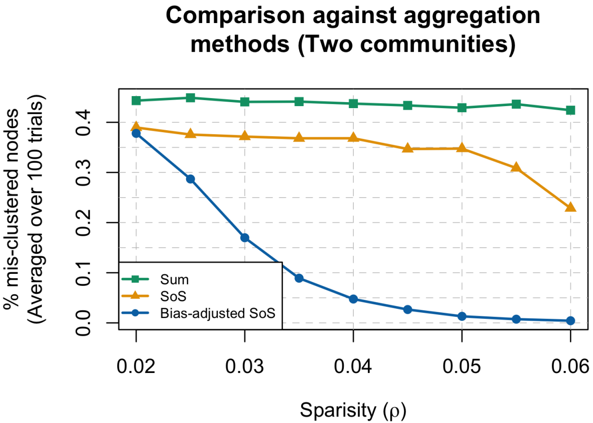
- Let $D_l = \mathrm{diag}(d_{l,1}, \ldots, d_{l,n})$
- Let $\hat{g}_{\mathrm{sc}}$ be obtained by applying spectral clustering on

$$\sum_{l=1}^{L} (A_l^2 - D_l)$$

- In practice, one can just calculate $\sum_l A_l^2$ and then zero out the diagonal.

A simulation

$L = 30$, $n = 200$, $K = 2$, B_l equals $B^{(1)}$ or $B^{(2)}$ with equal probability, where $B^{(1)} = \begin{bmatrix} 3/4 & \sqrt{3}/8 \\ \sqrt{3}/8 & 1/2 \end{bmatrix}$, $B^{(2)} = \begin{bmatrix} 7/8 & 3\sqrt{3}/8 \\ 3\sqrt{3}/8 & 1/8 \end{bmatrix}$



Analysis of bias removal.

$$\sum_l (A_l^2 - D_l) = \sum_l P_l^2 + \sum_l (P_l E_l + E_l P_l) + \sum_l (E_l^2 - D_l)$$

1. $\sum_l P_l^2$ has K large eigenvalues, with gap $\gtrsim n^2 \rho^2 L$.
2. $\| \sum_l P_l E_l \| \lesssim n^{3/2} \rho^{3/2} L^{1/2} \log^{1/2}(L+n)$ (gen. Bernstein)
3. $\| \sum_l (E_l^2 - D_l) \| \lesssim n \rho L^{1/2} \log^{1/2}(L+n)$ (gen. Hanson–Wright)

Theorem (L. & Lin '20)

Assume $K = O(1)$, balanced community sizes, and $L^{-1} \sum_l B_l^2 \succ cI$. The bias-adjusted squared spectral clustering is consistent if $n\rho L^{1/2} / \log^{1/2}(L+n) \to \infty$.

Not just a log factor improvement, but computationally much simpler.

The linear term: concentration of matrix sums

Theorem (generalized Bernstein's inequality)

Let $(X_l : 1 \leq l \leq L)$ be a sequence of $n \times r$ matrices with independent mean zero entries satisfying $\mathbb{E}|X_{l,ij}^k| \leq (v/2)R^{k-2}k!$ for some constants (v, R), and $(H_l : 1 \leq l \leq L)$ be a sequence of non-random $r \times m$ matrices, then for all $t > 0$

$$\mathbb{P}\left(\left\|\sum_l X_l H_l\right\| \geq t\right)$$
$$\leq 2(m+n)\exp\left\{-\frac{t^2/2}{v\left[(n\|\sum_l H_l^T H_l\|) \vee (\sum_l \|H_l\|_F^2)\right] + tR\max_l \|H_l\|_{2,\infty}}\right\},$$

where $\|\cdot\|_{2,\infty}$ denotes the maximum ℓ_2 norm of rows.

Controlling the quadratic term

Theorem

Let $E_l = A_l - \mathbb{E}A_l$. If $n\rho = O(1)$, $n\rho L^{1/2} \gtrsim \log^{1/2}(L+n)$, then with high probability

$$\left\| \sum_l (E_l^2 - D_l) \right\| \lesssim n\rho L^{1/2} \log^{1/2}(L+n)\,.$$

Main ingredient of the proof: two rounds of de-coupling (de la Penã and Montgomery-Smith ’95).

Concentration of matrix quadratic forms

Theorem (generalized Hanson–Wright inequality)

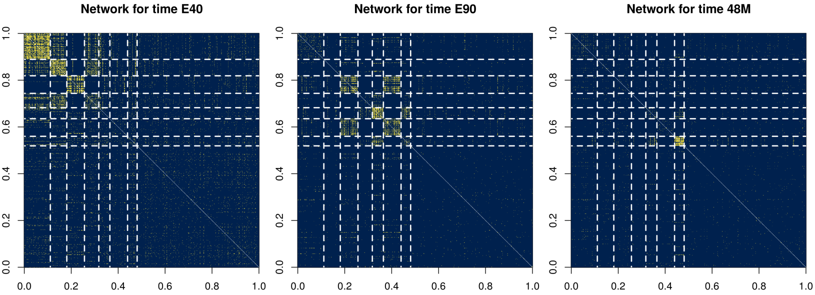
Let $(X_l : 1 \le l \le L)$ be a sequence of $n \times r$ matrices with independent zero mean sub-Gaussian entries: $\mathbb{E} e^{X_{l,ij}^2/v} \le 2$ for some constant v, and $(Q_l : 1 \le l \le L)$ be a sequence of non-random $r \times r$ matrices, then there is a constant C such that with high probability

$$\left\| \sum_l \left(X_l Q_l X_l^T - \mathbb{E} X_l Q_l X_l^T \right) \right\| \le C v n \log(L+n) \left(\sum_l \|Q_l\|^2 \right)^{1/2} .$$

Monkey brain gene co-expression network

- $n = 7836$ genes
- $L = 10$: from 40 days in the embryo to 48 months after birth.
- Co-expression network from the medial prefrontal cortex (related to developmental brain disorders such as ASD).
- $K = 8$ by scree plot.

Adjacency matrices grouped by $K = 8$ clusters



Network for time E40

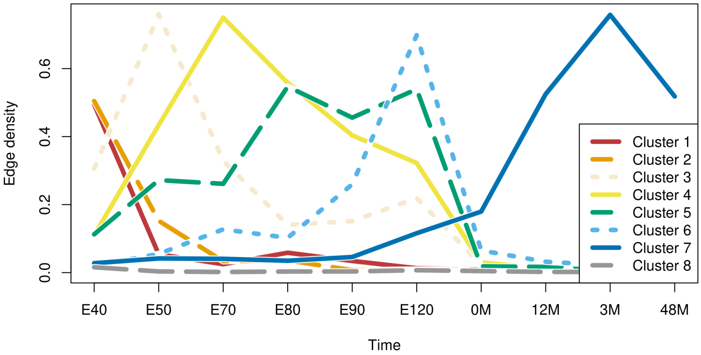
Network for time E90

Network for time 48M

Meanings of the communities

Group	Description	GO ID	p-value
1	RNA splicing	0008380	2.91×10^{-15}
2	Mitotic cell cycle process	1903047	1.26×10^{-25}
3	Chemical synaptic transmission	0007268	1.51×10^{-11}
4	Tissue development	0009888	8.00×10^{-5}
5	Neurotransmitter transport	0006836	2.22×10^{-5}
6	Regulation of transporter activity	0032409	5.68×10^{-6}
7	Transmembrane transporter activity	0022857	4.33×10^{-4}
8	None		

Within-cluster connectivity over time



Next steps

- Optimality of the threshold $n\rho L^{1/2}$?
    - Conjecture 1: Without computational constraints, the threshold is $n\rho L$. Consider $\max_{w:\|w\|=1} \|\sum_l w_l A_l\|$, and spectral clustering on $\sum_l w_l A_l$.
    - Conjecture 2: $n\rho L^{1/2}$ is the right threshold for polynomial time algorithms.
- Time-varying SBMs with general connectivity structure
    - What if both g and B changes smoothly from layer to layer?
    - Optimal smoothing determined by triplet (n, L, ρ)?
- Estimating K: cross-validation or spectral methods?

References

1. L., Chen, & Lynch (2020) “Consistent community detection in multi-layer network data”, *Biometrika*, **107**(1), 61-73.
2. L. & Lin (2022+) “Bias-adjusted spectral clustering in multi-layer stochastic block models”, *JASA*, to appear. `arXiv:2003.08222`

Thank you!