Enhanced detectability of community structure in multilayer networks through layer aggregation

Dane Taylor[1], Saray Shai[1], Natalie Stanley[1,2] and Peter J. Mucha[1]

[1]Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599, USA
[2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC 27599, USA

Abstract

Many systems are naturally represented by a multilayer network in which edges exist in multiple layers that encode different, but potentially related, types of interactions such as categorical social ties, types of critical infrastructure, or a temporal network at different instances in time. It is important to understand limitations on the detectability of community structure in these networks. In this research [1], we develop random matrix theory to analyze detectability limitations for multilayer (specifically, multiplex) stochastic block models (SBMs) in which L layers are derived from a common SBM [2]. We study the effect of layer aggregation on detectability for several aggregation methods, including summation of the layers' adjacency matrices for which we show the detectability limit vanishes as $\mathcal{O}(L^{-1/2})$ with increasing number of layers, L. Importantly, we find a similar scaling behavior when the summation is thresholded at an optimal value, providing insight into the common—but not well understood—practice of thresholding pairwise-interaction data to obtain sparse network representations.

Multilayer Stochastic Block Models (MLSBMs)

We study multiplex networks [3,4] that are encoded by L adjacency matrices $\{\mathbf{A}^{(l)}\}$. The layers all follow a common SBM that has two communities of size $N/2$ such that the edge probability is p_{in} if two nodes are in the same community and p_{out} if they are in different communities. We denote the nodes' community labels by $\{c_i\}$. To facilitate analysis, it is also convenient to define the *mean edge probability* $\rho = (p_{in}+p_{out})/2$ and the *probability difference* $\Delta = p_{in} - p_{out} > 0$.

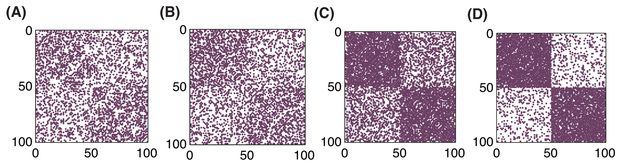

Figure 1. Adjacency matrices with $N = 100$ nodes for single-layer networks drawn from two-community SBMs with $\rho = 0.5$ and increasing values of Δ. Parameter Δ controls the prevalence of community structure and reveals phase transitions in the detectability of community structure.

Detectability Limit Δ^*

Community structure cannot be found if it is too weak. For a single-layer SBM with two-equal-sized communities, and assuming the network is sparse [i.e., $\rho = \mathcal{O}(N^{-1})$], then in the asymptotic $N \to \infty$ limit there exists a phase transition in detectability at the solution Δ^* to [5,6]

$$N\Delta = \sqrt{4N\rho}. \tag{1}$$

That is, the communities are detectable if $\Delta > \Delta^*$ and are undetectable if $\Delta \leq \Delta^*$. For example, the communities are detectable only in panels (B)–(D) of Fig. 1. Importantly, Eq. (1) has been derived via complementary analyses, Bayesian inference [5] and random matrix theory [6], and it identifies a detectability limit Δ^* for all polynomial-time community detection algorithms.

Layer Aggregation

We extend the study of detectability to multilayer networks, and in particular, we analyze the effect of layer aggregation on Δ^*. We study two methods of layer aggregation:

- 1. The *summation network* corresponds to summing the layers' adjacency matrices, $\overline{\mathbf{A}} = \sum_l \mathbf{A}^{(l)}$.
- 2. The *thresholded networks* are encoded by matrices $\{\hat{\mathbf{A}}^{(\tilde{L})}\}$ and are obtained by thresholding the summation at some value $\tilde{L}$. That is, $\hat{A}^{(\tilde{L})}_{ij} = 1$ if $\overline{A}_{ij} \geq \tilde{L}$.

Note: The summation network is often weighted and dense, whereas a thresholded network is unweighted and its sparsity depends on the choice of threshold, $\tilde{L} \in [1, L]$.

Random Matrix Theory for the Modularity Matrix

Inspired by [6], we develop random matrix theory for the modularity matrix. We extend this work to multilayer networks by first studying the modularity matrix $\overline{\mathbf{B}}$ for the summation network, which we define by $\overline{B}_{ij} = \overline{A}_{ij} - \rho L$.

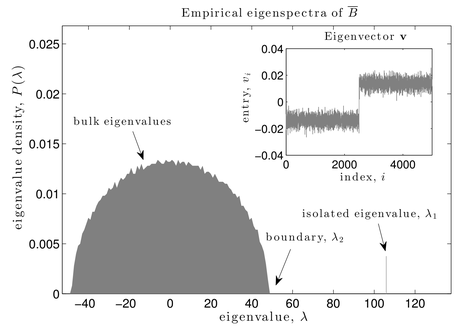

Figure 2. The eigenvalues $\{\lambda_i\}$ of $\overline{\mathbf{B}}$ consists of two parts: bulk eigenvalues following the spectral density $P(\lambda)$ and an isolated eigenvalue λ_1. As shown by the inset, the eigenvector $\mathbf{v}$ corresponding to λ_1 encodes the two-community structure; however, this only occurs if there is a gap between λ_1 and λ_2. The parameters are $N = 5000$, $L = 4$, $\rho = 0.03$, and $\Delta = 0.01$.

We find the $N \to \infty$ limiting spectral density to be

$$P(\lambda) = \frac{\sqrt{\lambda_2^2 - \lambda^2}}{\pi\lambda_2^2/2} \tag{2}$$

for $|\lambda| < \lambda_2$ and $P(\lambda) = 0$ otherwise, where

$$\lambda_2 = \sqrt{4NL[\rho(1-\rho) - \Delta^2/4]} \tag{3}$$

is the upper bound on the support of $P(\lambda)$ and is the limiting value of the second-largest eigenvalue. The largest eigenvalue is an isolated eigenvalue that limits to

$$\lambda_1 = NL\Delta/2 + 2[\rho(1-\rho) - \Delta^2/4]/\Delta. \tag{4}$$

Phase Transition for the Dominant Eigenvector $\mathbf{v}$

The phase transition in detectability corresponds to a phase transition for the dominant eigenvector $\mathbf{v}$ (i.e., $\overline{\mathbf{B}}\mathbf{v} = \lambda_1\mathbf{v}$). Specifically, the nodes' community labels are inferred based on $\mathbf{v}$. When there is a gap between λ_1 and λ_2, the eigenvector entries $\{v_i\}$ are correlated with the community labels $\{c_i\}$—that is, $v_i > 0$ for nodes $\{i\}$ in one community and $v_i < 0$ for nodes $\{i\}$ in the other community. We find that the entries $\{v_i\}$ within a community are Gaussian distributed with mean

$$|\langle v_i \rangle| = \sqrt{\frac{1}{N}}\sqrt{1 - \frac{\lambda_2^2}{(NL\Delta)^2}}, \tag{5}$$

and we use $|\langle v_i \rangle|$ as an order parameter to observe the phase transition in detectability.

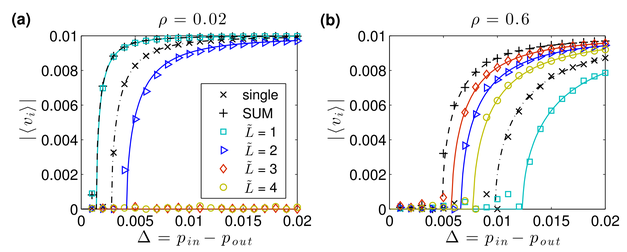

Figure 3. Phase transition for the dominant eigenvector $\mathbf{v}$ of $\overline{\mathbf{B}}$ for a single layer, the summation network, and thresholded networks. We show observed (symbols) and predicted values given by Eq. (5) (curves) for $N = 10^4$ and $L = 4$. The transitions occur at a critical values, Δ^*.

Δ^* for Summation Network

For the summation network, Δ^* corresponds to when $\lambda_1 = \lambda_2$ [see Eqs. (3) and (4)], which gives a new detectability equation

$$NL\Delta = \sqrt{4NL\rho(1-\rho)}. \tag{6}$$

Note that Eq. (6) recovers Eq. (1) in the sparse network limit [i.e., $\rho(1-\rho) \approx \rho$] for a single-layer network, $L = 1$. Importantly, Eq. (6) implies that Δ^* vanishes as $\mathcal{O}(1/\sqrt{NL})$ with increasing N or L.

Δ^* for Thresholded Networks

Thresholded networks are SBMs with the same community labels as the original SBM, but they have *effective* edge probabilities $\hat{p}_{in}^{(\tilde{L})}$ and $\hat{p}_{out}^{(\tilde{L})}$ as well as effective values for $\hat{\rho}^{(\tilde{L})} = (\hat{p}_{in}^{(\tilde{L})} + \hat{p}_{out}^{(\tilde{L})})/2$ and $\hat{\Delta}^{(\tilde{L})} = \hat{p}_{in}^{(\tilde{L})} - \hat{p}_{out}^{(\tilde{L})}$. (See [1] for details.) For a given $\tilde{L}$, the detectability limit can be found by substituting $\hat{\rho}^{(\tilde{L})} \mapsto \rho$ and $\hat{\Delta}^{(\tilde{L})} \mapsto \Delta$ into Eq. (6) with $L = 1$ and using a root-finding algorithm to numerically obtain a solution Δ^*.

Optimal Threshold $\tilde{L} = \lceil \rho L \rceil$

The threshold $\lceil \rho L \rceil$ is optimal in that it is typically the best choice of $\tilde{L}$ for any edge density ρ. More importantly, for threshold $\lceil \rho L \rceil$ in the asymptotic $L \to \infty$ limit, we obtain another detectability equation for Δ^*,

$$NL\Delta = \sqrt{2\pi NL\rho(1-\rho)}. \tag{7}$$

This result implies that Δ^* also vanishes as $\mathcal{O}(1/\sqrt{NL})$ for threshold $\tilde{L} = \lceil \rho L \rceil$.

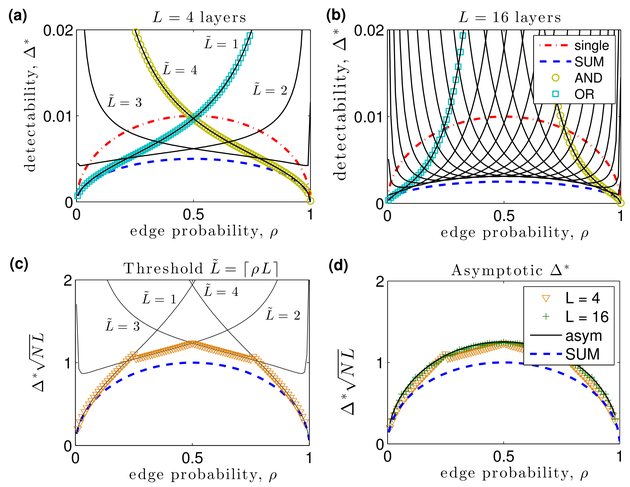

Figure 4. (a)–(b) Detectability limit Δ^* versus ρ for a single-layer network [Eq. (1)], the summation network [Eq. (6)], and thresholded networks with $\tilde{L} \in \{1, \ldots, L\}$ [Eq. (6) with effective probabilities]. The gold circles and cyan squares highlight $\tilde{L} = 1$ and $\tilde{L} = L$, which are equivalent to aggregating layers using logical OR and AND operations, respectively. (c) Δ^* for the optimal threshold $\tilde{L} = \lceil \rho L \rceil$ (orange triangles) is nearly as small as Δ^* for the summation network. (d) Considering the threshold $\lceil \rho L \rceil$, as L increases Δ^* limits to an asymptotic solution that solves Eq. (7) (black curve).

References

[1] D. Taylor, S. Shai, N. Stanley and P. J. Mucha. *Physical Review Letters*, in press (2016).
[2] N. Stanley, S. Shai, D. Taylor, P. J. Mucha, *IEEE Transaction on Network Science and Engineering*, in press (2016).
[3] P. J. Mucha, T. Richardsons, K. Macon, M. A. Porter, J.-P. Onnela, *Science* 328(5980), 876–878 (2010).
[4] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter, *J. of Complex Net.* 2, 203–271 (2014).
[5] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, *Physical Review Letters* 107(6), 065701 (2011).
[6] R. R. Nadakuditi and M. E. J. Newman, *Physical Review Letters* 108(18), 188701 (2012).