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

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, and it is important to understand limitations on the detectability of community structure in these networks. Using random matrix theory, we analyze detectability limitations for multilayer (specifically, multiplex) stochastic block models (SBMs) in which L layers are derived from a common SBM. 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.

PACS numbers: 89.75.Hc, 02.70.Hm, 64.60.aq

The analysis of complex networks [1] has far-reaching applications ranging from social systems [2] to the brain [3]. Often, a natural representation is that of a multilayer network (see reviews [4, 5]), whereby network layers encode different classes of interactions, such as categorical social ties [6], types of critical infrastructure [7], or a network at different instances in time [8]. In principle, the multilayer framework offers a more comprehensive representation of a data set or system, as compared to an aggregation of network layers that produces a simplified model but does so at the cost of information loss. For example, neglecting the layered structure can lead to severe and unintended consequences regarding structure [9] and dynamics [10–12], which can fundamentally differ between single-layer and multilayer networks [13, 14].

However, layer aggregation also implements an information processing that can yield beneficial results. Network layers are often correlated with one another and can encode redundant information [15]. In some cases a multilayer representation is an over-modeling, which can negatively impact the computational and memory requirements for storage and analysis. In such situations, it is beneficial to seek a more concise representation in which certain layers are aggregated [16, 17]. Identifying sets of repetitive layers amounts to a clustering problem, and it is closely related to the topic of clustering networks in an ensemble of networks [17, 18]. Much remains to be studied regarding *when* layer aggregation is appropriate and *how* it should be implemented.

We study here the effect of layer aggregation on community structure in multilayer networks in which each layer is drawn from a common stochastic block model (SBM). SBMs are a paradigmatic model [19] for complex structure in networks and are particularly useful for studying limitations on detectability—that is, if the community structure is too weak, it cannot be found upon inspection of the network [20–25]. Recently, the detectability limit has been explored for networks with degree heterogeneity [26] and hierarchical structure [27, 28], for temporal networks [29], and for the detection of communities using multi-resolution methods [30]. Despite growing interest in multilayer SBMs [31–35] (which we note, focus on *multiplex* networks in which nodes are identical in every layer and edges are restricted to connecting nodes in the same layer [4, 5]), the effect of layer aggregation on detectability limitations has yet to be explored outside the infinite layer limit [35].

To this end, we study detectability limitations for multilayer SBMs with layers following from identical SBM parameters and find that the method of aggregation significantly influences detectability. When the aggregate network corresponds to the summation of the adjacency matrices encoding the network layers, aggregation always improves detectability. In particular, the detectability limit vanishes with increasing number of layers, L, and decays as $\mathcal{O}(L^{-1/2})$. Because the summation of L adjacency matrices can often yield a weighted and dense network—which increases the complexity of community detection [36]—we also study binary adjacency matrices obtained by thresholding this summation at some value $\tilde{L}$. We find that the detectability limit is very sensitive to the choice of $\tilde{L}$; however, we also find that there exist thresholds (e.g., mean edge probability for homogeneous communities) that are optimal in that the detectability limit also decays as $\mathcal{O}(L^{-1/2})$. These results provide insight into the use of thresholding pairwise-interaction data so as to produce sparse networks—a practice that is commonplace but for which the effects are not well understood.

We begin by describing the multilayer SBM. We consider N nodes divided into K communities, and we denote by $c_i \in \{1,\dots,K\}$ the community index for each node $i \in \{1,\dots,N\}$. The multiplex network is defined by L layers encoded by a set of adjacency matrices, $\{\mathbf{A}^{(l)}\}$, where $A_{ij}^{(l)} = 1$ if (i,j) is an edge in layer l and $A_{ij}^{(l)} = 0$ otherwise. The probability of edge (i,j) in layer l is given by $\Pi_{c_ic_j} \in [0,1]$, where $\mathbf{\Pi}$ is a $K \times K$ matrix.

The detectability of community structure relates to the ability to recover the nodes' community labels $\{c_i\}$. To connect with previous research [21, 23–25], we focus on the case of $K = 2$ communities of equal size with edge probabilities $\Pi_{11} = \Pi_{22} = p_{in}$ and $\Pi_{12} = \Pi_{21} = p_{out}$. Below, we will simultaneously refer to these respective probabilities as $p_{in,out}$. We assume $p_{in} \geq p_{out}$ to study "assortative" communities in which there is a prevalence of edges between nodes in the same community [37].

It has been shown for the large network $N \to \infty$ limit that there exists a detectability limit characterized [23, 24] by the solution curve (Δ^*, ρ) to

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

where $\Delta = p_{in} - p_{out}$ is the difference in probability and $\rho = (p_{in} + p_{out})/2$ is the mean edge probability. For given ρ, the communities are detectable only when the presence of community structure is sufficiently strong, i.e., $\Delta > \Delta^*$. Equation (1) describes a phase transition that has been obtained via complementary analyses—Bayesian inference [23] and random matrix theory [24]—and represents a critical point that is independent of the community detection method (see [23] and footnote 11 in [24]). We further note that Eq. (1) was derived for sparse networks [i.e., constant ρN so that $\rho = \mathcal{O}(N^{-1})$]. Here, we must consider the full range of densities, $\rho \in [0, 1]$, to allow for aggregated networks that are potentially dense [i.e., $\rho = \mathcal{O}(1)$ as $N \to \infty$].

In this Letter, we study the behavior of Δ^* for two methods of aggregating layers. We define the *summation* network corresponding to the weighted adjacency matrix $\overline{\mathbf{A}} = \sum_l \mathbf{A}^{(l)}$ as well as a family of *thresholded* networks with unweighted adjacency matrices $\{\hat{\mathbf{A}}^{(\tilde{L})}\}$ that are obtained by applying a threshold $\tilde{L} \in \{1, \ldots, L\}$ to the entries of $\overline{\mathbf{A}}$. Specifically, we define $\hat{A}_{ij}^{(\tilde{L})} = 1$ if $\overline{A}_{ij} \geq \tilde{L}$ and $\hat{A}_{ij}^{(\tilde{L})} = 0$ otherwise. Of particular interest are the limiting cases $\tilde{L} = L$ and $\tilde{L} = 1$, which respectively correspond to applying logical AND and OR operations to the original multiplex data $\{A_{ij}^{(l)}\}$ for fixed (i, j). We refer to these thresholded networks as the AND and OR networks, respectively.

We study the detectability limit for the layer-aggregated networks using random matrix theory [38, 39]. This approach is particularly suited for detectability analysis since community labels $\{c_i\}$ can be identified using spectral partitioning and phase transitions [24, 27, 28] in detectability correspond to the disappearance of gaps between isolated eigenvalues (whose corresponding eigenvectors reflect community structure) and bulk eigenvalues [which arise due to stochasticity and whose $N \to \infty$ limiting distribution is given by a spectral density $P(\lambda)$]. We develop theory based on the modularity matrix $\overline{B}_{ij} = \overline{A}_{ij} - \rho L$ [40]. Note that we do not use the configuration model as the null model. Instead, since all nodes are identical under the SBM, the appropriate null model is Erdős-Rényi with repeated edges allowed in which the expected number of edges between any pair of nodes is ρL.

We first study Δ^* for the summation network. We analyze the distribution of real eigenvalues $\{\lambda_i\}$ of $\overline{\mathbf{B}}$ (in descending order) using methodology developed in [24, 38]; we extend this work to networks that are multiplex and possibly dense. We outline our results here and provide further details in the Supplemental Material. We begin by describing the statistical properties of entries $\{\overline{A}_{ij}\}$, which are independent random variables following a binomial distribution $P\left(\overline{A}_{ij} = a\right) = f(a; L, \Pi_{c_i c_j})$, where

$$f(a; L, p) = \left(\begin{array}{c} L \\ a \end{array}\right) p^a (1-p)^{L-a} \tag{2}$$

has mean Lp and variance $Lp(1-p)$. Provided that there is sufficiently large variance in the edge probabilities (i.e., $NL\rho(1-\rho) \gg 1$), we find that the limiting $N \to \infty$ distribution of bulk eigenvalues for $\overline{\mathbf{B}}$ is given by a semi-circle distribution,

$$P(\lambda) = \frac{\sqrt{\lambda_2^2 - \lambda^2}}{\pi \lambda_2^2/2} \tag{3}$$

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

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

is the upper bound on the support of this spectral density and is the limiting $N \to \infty$ value of the second-largest eigenvalue. The largest eigenvalue of $\overline{\mathbf{B}}$ in the $N \to \infty$ limit is an isolated eigenvalue

$$\lambda_1 = NL\Delta/2 + 2[\rho(1-\rho) - \Delta^2/4]/\Delta. \tag{5}$$

As we shall show, $\Delta^* \to 0$ as N increases, and therefore the $\Delta^2/4$ terms in Eq. (4) and (29) are negligible near the detectability limit (i.e., $\Delta \approx \Delta^*$). The eigenvector $\mathbf{v}$ corresponding to λ_1 gives the spectral bipartition—the inferred community label of node i is determined by the sign of v_i—and provided that the largest eigenvalue corresponds to this isolated eigenvalue, λ_1, the eigenvector entries $\{v_i\}$ are correlated with the community labels $\{c_i\}$. To obtain the detectability limit, we set $\lambda_1 = \lambda_2$, neglect the $\Delta^2/4$ terms and simplify, yielding a modified detectability equation

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

Note that Eq. (6) recovers Eq. (1) when $L = 1$ and $\rho \to 0$ [i.e., for sparse networks, $\rho(1-\rho) \approx \rho$]. Defining $p_{in}^* = \rho + \Delta^*/2$ and $p_{out}^* = \rho - \Delta^*/2$, we find for fixed ρ and increasing N and/or L that $p_{in,out}^* \to \rho$ and $\Delta^* \to 0$, decaying as $\mathcal{O}(1/\sqrt{NL})$.

We now study Δ^* for the thresholded networks, which correspond to single-layer SBMs in which the community labels $\{c_j\}$ are identical to those of the multilayer SBM, but there are new *effective* block edge probabilities

$$\hat{\Pi}_{nm}^{(\tilde{L})} = 1 - F(\tilde{L} - 1; L, \Pi_{nm}), \tag{7}$$

where $F(a; L, p)$ is the cumulative distribution function for the binomial distribution $f(a; L, p)$. The effective probabilities for the AND and OR networks are $\hat{\Pi}_{nm}^{(L)} = (\Pi_{nm})^L$ and $\hat{\Pi}_{nm}^{(1)} = 1 - (1 - \Pi_{nm})^L$, respectively. For the two-community SBM, the effective probabilities are $\hat{p}_{in,out}^{(\tilde{L})} = 1 - F(\tilde{L} - 1; L, p_{in,out})$, $\hat{\Delta}^{(\tilde{L})} = \hat{p}_{in}^{(\tilde{L})} - \hat{p}_{out}^{(\tilde{L})}$, and $\hat{\rho}^{(\tilde{L})} = (\hat{p}_{in}^{(\tilde{L})} + \hat{p}_{out}^{(\tilde{L})})/2$. The modularity matrices for the thresholded networks become $\hat{B}_{ij}^{(\tilde{L})} = \hat{A}_{ij}^{(\tilde{L})} - \hat{\rho}^{(\tilde{L})}$. We identify the detectability limit by substituting $\hat{\Delta}^{(\tilde{L})} \mapsto \Delta$ and $\hat{\rho}^{(\tilde{L})} \mapsto \rho$ into Eq. (6) (with $L = 1$) and numerically finding a solution (Δ^*, ρ) using a root-finding algorithm. Note that the detectability equation holds for the effective probabilities, $N\hat{\Delta}^{(\tilde{L})} = \sqrt{4N\hat{\rho}^{(\tilde{L})}(1 - \hat{\rho}^{(\tilde{L})})}$, and not the single-layer probabilities, $N\Delta \neq \sqrt{4N\rho(1 - \rho)}$.

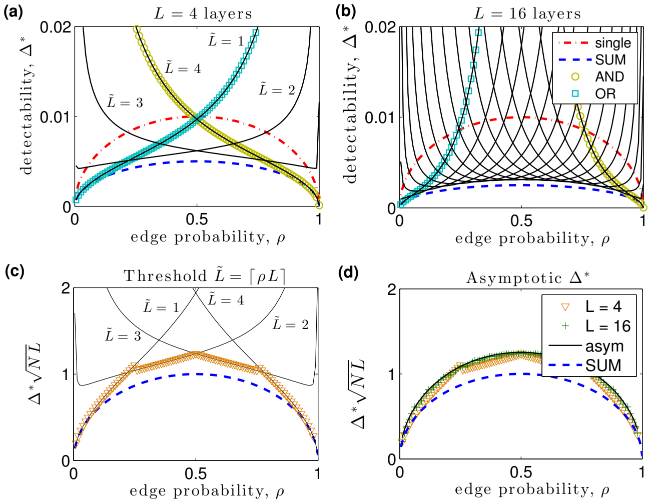

FIG. 1. (Color online) *Layer aggregation enhances the detectability of community structure.* (a)–(b) We plot the detectability limit Δ^* versus mean edge probability ρ for a single network layer (red dot-dashed curves), the aggregate network obtained by summation (blue dashed curves), and aggregate networks obtained by thresholding this summation at $\tilde{L} \in \{1, 2, 3, 4\}$ (solid curves). Gold circles and cyan squares highlight $\tilde{L} = L$ and $\tilde{L} = 1$, which we refer to as AND and OR networks, respectively. Results are shown for $N = 10^4$ nodes with (a) $L = 4$ and (b) $L = 16$ layers. (c) For $L = 4$, we show Δ^* versus ρ for the optimal threshold $\tilde{L} = \lceil \rho L \rceil$ (orange triangles), which lies on the solution curves for $\tilde{L} \in \{1, \ldots, L\}$ (solid curves). (d) We show Δ^* for $\tilde{L} = \lceil \rho L \rceil$ with $L \in \{4, 16\}$. These piecewise-continuous solutions collapse onto the asymptotic solution $\Delta^*_{(asym)}$ (black curve) as L increases. In panels (c)–(d), we additionally plot Δ^* for the summation network (blue dashed curves).

In Figs. 1(a)–(b), we show Δ^* versus the mean edge probability ρ for the different aggregation methods: (i) a single layer (red dot-dashed curves), which is identical in panels (a) and (b); (ii) the summation network (blue dashed curves), for which the curve in (b) corresponds to the curve in panel (a) rescaled by a factor of 1/2; and (iii) thresholded networks (solid curves), which shift left-to-right with increasing $\tilde{L}$. This is evident by comparing Δ^* for the AND ($\tilde{L} = L$, gold circles) and OR ($\tilde{L} = 1$, cyan squares) networks. We find when ρ is large that the AND (OR) network has a relatively small (large) detectability limit; in contrast, when ρ is small the AND (OR) network has a relatively large (small) detectability limit. In other words, aggregating layers using the AND (OR) operation is beneficial for dense (sparse) networks.

It is interesting to ask if there are choices of ρ and $\tilde{L}$ for which the detectability limit vanishes as $\mathcal{O}(L^{-1/2})$ with increasing L—that is, a behavior similar to that of the summation network. To this end, we study the threshold $\tilde{L} = \lceil \rho L \rceil$, which we numerically observe to be the best $\tilde{L}$ for most values of ρ. This choice is also convenient as it only requires knowledge of the mean edge probability, ρ, which is easy to obtain in practice. In Fig. 1(c), we plot Δ^* versus ρ for $L = 4$ and $\tilde{L} = \lceil \rho L \rceil$ (orange triangles), which lies along the solution curves for $\tilde{L} \in \{1, \ldots, L\}$ (solid curves). In Fig. 1(d), we plot Δ^* for threshold $\tilde{L} = \lceil \rho L \rceil$ with $L = 4$ (orange triangles) and $L = 64$ (green crosses). These curves align due to the rescaling of the vertical axis by $\sqrt{NL}$. In fact, we find in the large L limit that these solutions Δ^* collapse onto a single curve $(\Delta^*_{(asym)}, \rho)$ that solves

$$NL\Delta = \sqrt{2\pi NL\rho(1 - \rho)}, \tag{8}$$

which we plot by the black line in Fig. 1(d). To obtain Eq. (8), we use the central limit theorem [41] to approximate $\hat{p}_{in,out}^{(\lceil \rho L \rceil)} \approx \hat{p}_{in,out}^{(asym)} = 1 - G\left(L\rho; Lp_{in,out}, Lp_{in,out}\left(1 - p_{in,out}\right)\right)$, where $G\left(p; \mu, \sigma^2\right) = \frac{1}{2} + \frac{1}{2}\text{erf}\left((p - \mu)/\sigma\sqrt{2}\right)$ is the value of the cumulative distribution function of the normal distribution with mean μ and variance σ^2 evaluated at p. In particular, we approximate $\hat{\Delta}^{(\lceil \rho L \rceil)} \approx \hat{\Delta}^{(asym)} = \text{erf}\left(\Delta\sqrt{L}/\sqrt{8\rho(1 - \rho)}\right)$ and $\hat{\rho}^{(\lceil \rho L \rceil)} \approx \hat{\rho}^{(asym)} = 1/2$. Equation (8) is recovered after substituting $\hat{\Delta}^{(asym)} \mapsto \Delta$ and $\hat{\rho}^{(asym)} \mapsto \rho$ into Eq. (6) with $L = 1$ and using the first-order expansion $\text{erf}^{-1}(N^{-1/2}) \approx \sqrt{\pi/4N}$. Importantly, Eq. (8) implies that Δ^* decays as $\mathcal{O}(1/\sqrt{NL})$ for thresholded networks with $\tilde{L} = \lceil \rho L \rceil$.

In Fig. 2, we illustrate the limiting $L \to \infty$ behavior for thresholded networks with $\tilde{L} = \lceil \rho L \rceil$. In panels (a)–(b), we plot $\hat{p}_{in}^{(\lceil \rho L \rceil)}$ (blue triangles) and $\hat{p}_{out}^{(\lceil \rho L \rceil)}$ (red circles) versus ρ for $\Delta = 0.1$ with (a) $L = 4$ and (b) $L = 64$. We also plot the effective probabilities $\hat{p}_{in}^{(\tilde{L})}$ (solid curves) and $\hat{p}_{out}^{(\tilde{L})}$ (dashed curves) for the AND (gold curves) and OR (cyan curves) networks. In panel (b), we additionally plot the limiting effective probabilities $\hat{p}_{in}^{(asym)}$ (blue solid curve) and $\hat{p}_{out}^{(asym)}$ (red dashed curve). Comparing panel (b) to (a), one can observe that as L increases, the piecewise-continuous solutions $\hat{p}_{in,out}^{(\lceil \rho L \rceil)}$ separate and align with the respective asymptotic solutions $\hat{p}_{in,out}^{(asym)}$.

In Figs. 2(c)–(f), we illustrate adjacency matrices $\hat{\mathbf{A}}^{(\lceil \rho L \rceil)}$ of thresholded networks with $\rho = 0.3$ and $\Delta = 0.1$ for various L. We note that the community structure is undetectable for $L = 1$ since $\Delta^* = 0.1095$, whereas it is detectable (and visually apparent) for $L = 128$. Comparing (c)–(f) illus-

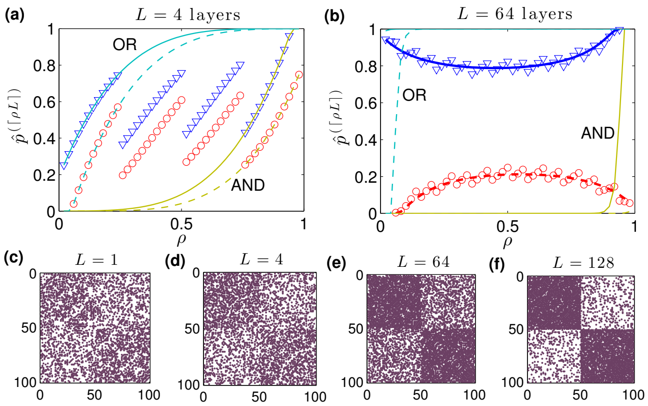

FIG. 2. (Color online) *Effective edge probabilities for layer aggregation at an optimal threshold.* (a)–(b) The summation and thresholding at $\tilde{L} = \lceil \rho L \rceil$ of L adjacency matrices yields a new SBM with effective edge probabilities $\hat{p}_{in}^{(\lceil \rho L \rceil)}$ (blue triangles) and $\hat{p}_{out}^{(\lceil \rho L \rceil)}$ (red circles). Results are for $\Delta = 0.1$ (i.e., $p_{in,out} = \rho \pm 0.05$) with (a) $L = 4$ and (b) $L = 64$ layers. We also show effective probabilities for the AND (gold curves) and OR (cyan curves) networks. (Solid and dashed curves give $\hat{p}_{in}^{(\tilde{L})}$ and $\hat{p}_{out}^{(\tilde{L})}$, respectively.) Note for the larger L value in (b) that $\hat{p}_{in}^{(\lceil \rho L \rceil)}$ and $\hat{p}_{out}^{(\lceil \rho L \rceil)}$ have separated and aligned with the asymptotic probabilities $\hat{p}_{in}^{(asym)}$ (blue solid curve) and $\hat{p}_{out}^{(asym)}$ (red dashed curve), respectively. (c)–(f) Adjacency matrices of thresholded networks with $\rho = 0.3$, $\Delta = 0.1$, $\tilde{L} = \lceil \rho L \rceil$ and various L.

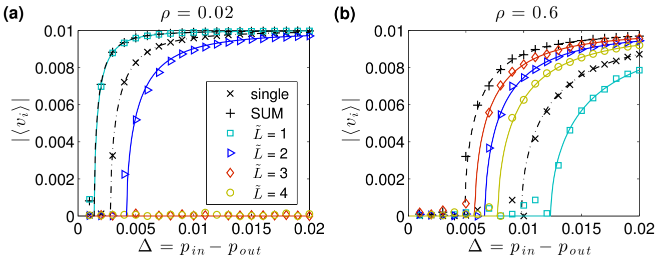

FIG. 3. (Color online) *Phase transition at* Δ^* *for the dominant eigenvector* $\mathbf{v}$ *of the modularity matrix.* We show observed (symbols) and predicted values given by Eq. (9) (curves) for the mean eigenvector entry $|\langle v_i \rangle|$ within a community for $N = 10^4$ and $L = 4$.

trates the $L \to \infty$ limiting behavior of $\hat{\mathbf{A}}^{(\lceil \rho L \rceil)}$. Specifically, application of Hoeffding's inequality [42] (and using that $p_{in,out} - \rho = \pm \Delta/2$) yields $p_{in}^{(\lceil \rho L \rceil)} \geq 1 - e^{-L\Delta^2/2}$ and $p_{out}^{(\lceil \rho L \rceil)} \leq e^{-L\Delta^2/2}$, which implies that $\hat{p}_{in}^{(\lceil \rho L \rceil)} \to 1$ and $\hat{p}_{out}^{(\lceil \rho L \rceil)} \to 0$ with increasing L so that $\hat{A}_{ij}^{(\lceil \rho L \rceil)} \to \delta_{c_i c_j}$, where δ_{nm} is the Kronecker delta function.

We conclude by studying the dominant eigenvector $\mathbf{v}$ of the appropriate modularity matrix, which undergoes a phase transition at Δ^*: $\{v_i\}$ and the community labels $\{c_i\}$ are uncorrelated for $\Delta < \Delta^*$, whereas they are correlated for $\Delta > \Delta^*$. Using methodology developed in [38], we find that the entries $\{v_i\}$ within a community are Gaussian distributed with mean

$$|\langle v_i \rangle| = \sqrt{\frac{1}{N}} \sqrt{1 - \frac{\lambda_2^2}{(NL\Delta)^2}}, \tag{9}$$

which we use as an order parameter to observe the phase transition. In Fig. 3, we depict observed (symbols) and predicted values given by Eq. (9) (curves) of $|\langle v_i \rangle|$ for a single layer (×-symbols), the summation network (+-symbols) and thresholded networks (open symbols). We focus on a range of Δ that contains Δ^* for most aggregation methods. Note for the thresholded networks that there is no simple ordering to Δ^*, which can be deduced by examining Fig. 1(a) for $\rho \in \{0.02, 0.6\}$. Finally, we note that finite-size effects amplify disagreement between observed and predicted values near the phase transitions.

In this Letter, we studied the limitations on community detection for multilayer networks with layers drawn from a common SBM. As an illustrative model, we analyzed the effect of layer aggregation on the detectability limit Δ^* for two equal-sized communities. When layers are aggregated by summation, we analytically showed detectability is always enhanced and Δ^* vanishes as $\mathcal{O}(L^{-1/2})$. When layers are aggregated by thresholding this summation, Δ^* depends sensitively on the choice of threshold, $\tilde{L}$. For $\tilde{L} = \lceil \rho L \rceil$, we analytically found Δ^* to also vanish as $\mathcal{O}(L^{-1/2})$. We note that our analysis also describes layer aggregation by taking the mean, $L^{-1} \sum_l \mathbf{A}^{(l)}$, since the multiplication of a matrix by a constant (e.g., L^{-1}) simply scales all eigenvalues by that constant. Thus, our results are in excellent agreement with previous work [35] that proved spectral clustering via the mean adjacency matrix to be a consistent estimator for the community labels.

Finally, it is commonplace to threshold pairwise-interaction data to construct network representations that are sparse and unweighted and can be studied at a lower computational cost. Our research provides insight into this common—yet not well understood—practice. It is important to extend our work to more-complicated settings. We believe fruitful directions should include allowing the SBMs of layers to be correlated [25] (that is, rather than identical) as well as allowing layers to be organized into "strata" [17], so that layers within a stratum follow a similar SBM but the SBMs can greatly differ between strata. We are currently extending our analysis to hierarchical SBMs using methodology developed in [27].

The authors were supported by the NIH (R01HD075712, T32GM067553, T32CA201159), the James S. McDonnell Foundation (#220020315) and the UNC Lineberger Comprehensive Cancer Center with funding provided by the University Cancer Research Fund via the State of North Carolina. The content does not necessarily represent the views of the funding agencies. We thank the reviewers for their helpful comments.

* dane.r.taylor@gmail.com

[1] M. E. J. Newman, SIAM Rev. **45**(2), 167–256 (2003).
[2] J. Moody, Amer. Soc. Rev. **69**(2), 213–238 (2004).
[3] D. S. Bassett *et al.*, Proc. Natl. Acad. of Sci. **108**, 7641–8646 (2011).

[4] S. Boccaletti *et al.*, Phys. Reports, **544**(1), 1–122 (2014).

[5] M. Kivelä *et al.*, J. of Complex Networks **2**(3), 203–271 (2014).

[6] D. Krackhardt, Social Networks **9**(2), 109–134 (1987).

[7] Y. Y. Haimes and P. Jiang, J. of Infrast. Sys. **7**(1) 1–12 (2001).

[8] P Holme and J. Saramäki, Phys. Reports **519**(3), 97–125 (2012).

[9] P. J. Mucha and *et al.*, Science **328**(5980), 876–878 (2010).

[10] R. J. Sánchez-García, E. Cozzo and Y Moreno, Phys. Rev. E **89**(5), 052815 (2014).

[11] A. Sole-Ribalta *et al.*, Phys. Rev. E **88**(3), 032807 (2013).

[12] C. D. Brummitt, R. M. D'Souza and E. A. Leicht, Proc. Natl. Acad. Sci. **109**(12), E680–E689 (2012).

[13] A. Bashan, Y. Berezin, S. V. Buldyrev and S. Havlin, Nat. Phys. **9**(10), 667–672 (2013).

[14] F. Radicchi and A. Arenas, Nat. Phys. **9**(11), 717–720 (2013).

[15] G. Menichetti, D. Remondini and G. Bianconi, Phys. Rev. E **90**(6) 062817 (2014).

[16] M. De Domenico, V. Nicosia, A. Arenas and V. Latora, *et al.*, Nat. Comms. **6**, 6864 (2015).

[17] N. Stanley, S. Shai, D. Taylor, P. J. Mucha, Preprint available online at http://arxiv.org/abs/1507.01826 (2015).

[18] J.-P. Onnela *et al.* Phys. Rev. E **86**, 036104 (2012).

[19] A Lancichinetti, S. Fortunato and F. Radicchi, Phys. Rev. E **78**(4), 046110 (2008).

[20] A. Lancichinetti and S. Fortunato. Phys. Rev. E **84**(6), 066122 (2011).

[21] J. Reichardt and M. Leone, Phys. Rev. Lett. **101**, 078701 (2008).

[22] D. Hu, P. Ronhovde and Z. Nussinov, Philo. Mag. **92**(4), 406–445 (2012).

[23] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Phys. Rev. Lett. **107**(6), 065701 (2011).

[24] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**(18), 188701 (2012).

[25] E. Abbe, A. S. Bandeira and G. Hall, IEEE Trans. on Info. Theory, **62**(1), 471–487 (2016).

[26] F. Radicchi, Phys. Rev. E **88**(1), 010801 (2013).

[27] T. P. Peixoto, Phys. Rev. Lett. **111**(9), 098701 (2013).

[28] S. Sarkar, J. A. Henderson, and P. A. Robinson. PloS one **8**(1), e54383 (2013).

[29] A. Ghasemian *et al.*, Preprint available online at http://arXiv.org/abs/1506.06179 (2015).

[30] T. Kawamoto and Y. Kabashima, EPL (Europhysics Letters), 112(4), 40007 (2015).

[31] T. Valles-Catala, F. A. Massucci, R. Guimera, and M. Sales-Pardo, Preprint available online at http://arxiv.org/abs/1411.1098 (2014).

[32] S. Paul and Y. Chen, Preprint available online at http://arxiv.org/abs/1506.02699 (2015).

[33] P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen, Preprint available online at http://arxiv.org/abs/1501.06444 (2015).

[34] T. P. Peixoto, Phys. Rev. E 92, 042807 (2015).

[35] Q. Han, K. Xu, and E. Airoldi, Proc. of the 32nd Int. Conf. on Machine Learn., 1511–1520 (2015).

[36] C. Aicher, A. Z. Jacobs and A. Clauset, J. of Complex Networks **3**(2), 221–248 (2015).

[37] M. P. Rombach *et al.*, SIAM J. on A. Math. **74**(1) 167–190 (2014).

[38] F. Benaych-Georges and R. R. Nadakuditi, Adv. in Math. **227**, 494 (2011).

[39] R. R Nadakuditi and M. E. J Newman, Phys. Rev. E **87**(1), 012803 (2013).

[40] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**(2), 026113 (2004).

[41] O. Kallenberg, *Foundations of modern probability,* (Springer Science & Business Media, 2006).

[42] W. Hoeffding, J. of the Amer. Stat. Assoc. **58**(301), 13–30 (1963).

Supplemental Material: Enhanced detectability of community structure in multilayer networks through layer aggregation

Dane Taylor,[1,*] Saray Shai,[1] Natalie Stanley[1,2] 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

Eigenspectra of Modularity Matrix $\overline{B}$

Here, we provide further details about the limiting $N \to \infty$ distribution of eigenvalues for modularity matrix $\overline{\mathbf{B}} = \overline{\mathbf{A}} - \rho L \mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is a vector of ones, $\overline{\mathbf{A}} = \sum_l \mathbf{A}^{(l)}$ is the summation of the layers' adjacency matrices, and each $\mathbf{A}^{(l)}$ is drawn from a single stochastic block model with two equal-sized communities. Our analysis is based on methodology developed in [1,2], which we extend to layer-aggregated multiplex networks including those that are potentially dense. As shown in Fig. 4, the spectrum consists of two parts—an isolated eigenvalue λ_1 (whose corresponding eigenvector $\mathbf{v}$ encodes the spectral bi-partition) and bulk eigenvalues which have an $N \to \infty$ limiting distribution $P(\lambda)$. In the analysis to follow, we will assume that the community structure is detectable. We begin by defining random matrix

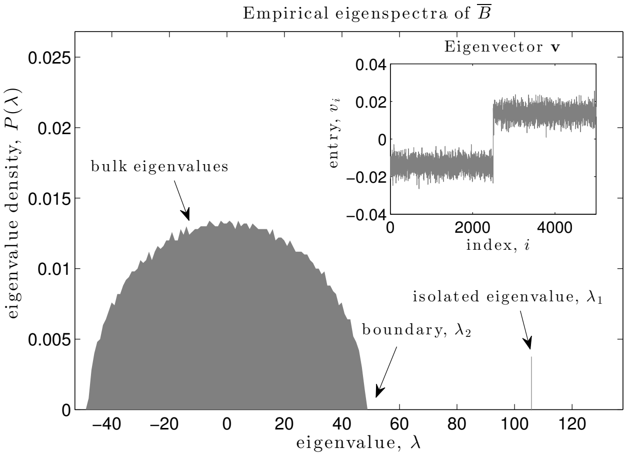

FIG. 4. (Color online) *Empirical eigenspectra of the modularity matrix* $\overline{\mathbf{B}}$. We plot the distribution of eigenvalues of $\overline{\mathbf{B}} = \overline{\mathbf{A}} - \rho L \mathbf{1}\mathbf{1}^T$, which consists of two parts: bulk eigenvalues that solve Eq. (12) and an isolated eigenvalue that solves Eq. (13). The subplot depicts the eigenvector $\mathbf{v}$ corresponding to the largest eigenvalue λ_1, which encodes community structure and gives the spectral bi-partition. Results are shown for $N = 5000$ nodes, $L = 4$ layers, mean edge probability $\rho = 0.03$, and probability difference $\Delta = 0.01$ (see main text).

$$\mathbf{X} = \overline{\mathbf{B}} - \langle \overline{\mathbf{B}} \rangle, \tag{10}$$

where $\langle \overline{B}_{ij} \rangle$ indicates the mean value of $\overline{B}_{ij}$ across the random matrix ensemble. The decomposition of $\overline{\mathbf{B}}$ facilitates the analysis of spectra through the following relation,

$$\begin{aligned} 0 &= \det\left(z\mathbf{I} - \overline{\mathbf{B}}\right) \\ &= \det\left(z\mathbf{I} - (\mathbf{X} + \langle \overline{\mathbf{B}} \rangle)\right) \\ &= \det\left(z\mathbf{I} - \mathbf{X}\right) \det\left(\mathbf{I} - (z\mathbf{I} - \mathbf{X})^{-1} \langle \overline{\mathbf{B}} \rangle\right), \end{aligned} \tag{11}$$

which assumes the invertibility of $(z\mathbf{I} - \mathbf{X})$. Equation (11) highlights that the spectra of $\overline{\mathbf{B}}$ can be studied in two parts: a distribution $P(z)$ of bulk eigenvalues that solve the first term,

$$0 = \det\left(z\mathbf{I} - \mathbf{X}\right), \tag{12}$$

and an isolated eigenvalue that solves the second term,

$$0 = \det\left(\mathbf{I} - (z\mathbf{I} - \mathbf{X})^{-1} \langle \overline{\mathbf{B}} \rangle\right). \tag{13}$$

Before describing the solutions to Eq. (12) and Eq. (13), we comment on the matrices $\mathbf{X}$ and $\langle \overline{\mathbf{B}} \rangle$. Recall that each entry $\overline{A}_{ij}$ follows a binomial distribution [see Eq. (2) in the main text], so that their mean and variance is

$$\begin{aligned} \langle A_{ij} \rangle &= \begin{cases} Lp_{in}, & c_i = c_j \\ Lp_{out}, & c_i \neq c_j. \end{cases} \\ \langle A_{ij}^2 \rangle - \langle A_{ij} \rangle^2 &= \begin{cases} Lp_{in}(1-p_{in}), & c_i = c_j \\ Lp_{out}(1-p_{out}), & c_i \neq c_j, \end{cases} \end{aligned} \tag{14}$$

where $c_i, c_j \in \{1, 2\}$ indicate the community labels of nodes i and j. It follows that $\{X_{ij}\}$ have mean and variance

$$\begin{aligned} \langle X_{ij} \rangle &= 0 \\ \langle X_{ij}^2 \rangle &= \begin{cases} Lp_{in}(1-p_{in}), & c_i = c_j \\ Lp_{out}(1-p_{out}), & c_i \neq c_j. \end{cases} \end{aligned} \tag{15}$$

We next consider $\langle \overline{\mathbf{B}} \rangle$. Using that $\overline{B}_{ij} = \overline{A}_{ij} - \rho L$ and $\rho = (p_{in} + p_{out})/2$ (i.e., $p_{in,out} - \rho = \pm\Delta/2$), we find

$$\langle \overline{B}_{ij} \rangle = \begin{cases} L\Delta/2, & c_i = c_j \\ -L\Delta/2, & c_i \neq c_j. \end{cases} \tag{16}$$

Importantly, $\langle \overline{\mathbf{B}} \rangle$ is a rank-one matrix [2]

$$\langle \overline{B} \rangle = \theta_1 \mathbf{u}\mathbf{u}^T, \tag{17}$$

where $\mathbf{u} = N^{-1/2}[1, \ldots, 1, -1, \ldots, -1]^T$ and

$$\theta_1 = \frac{NL\Delta}{2}. \tag{18}$$

We point out that without loss of generality, we have assumed that nodes $\{1,\ldots,N/2\}$ are in community 1 (i.e., $u_i > 1$ for these nodes) and nodes $\{1+N/2,\ldots,N\}$ are in community 2 (i.e., $u_i < 1$ for these nodes).

We now return our attention to solving Eq. (11) for the eigenvalues of $\overline{\mathbf{B}}$. We first solve Eq. (12) to study the bulk eigenvalues. The limiting $N \to \infty$ spectral density $P(z)$ of $\mathbf{X}$ can be solved via its average resolvent $\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle$ and the Stieltjes transform [1]

$$P(z) = \frac{-1}{N\pi}\text{Im Tr}\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle, \tag{19}$$

where $z \in \mathbb{C}$ approaches the real line $\mathbb{R}$ from above. Our analysis of Eq. (19) directly follows the methodology presented in [2], albeit for an aggregated multiplex network and allowing for potentially dense networks. In particular, the average resolvent can be expanded as

$$\text{Tr}\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle = \frac{1}{z}\sum_{k=0}^{\infty}\frac{\text{Tr}\langle \mathbf{X}^k\rangle}{z^k}, \tag{20}$$

where

$$\text{Tr}\langle \mathbf{X}^k\rangle = \sum_{i_1,\ldots,i_k}\langle X_{i_1 i_2}X_{i_2 i_3}\ldots X_{i_k i_1}\rangle, \tag{21}$$

and the sequence $\{i_1, i_2, \ldots, i_k, i_1\}$ defines an Euler tour at node i_1. Because $\langle X_{ij}\rangle = 0$, any term in Eq. (21) that contains a variable just once will be mean zero across the ensemble. Moreover, terms containing a variable more than twice become negligible when the nodes' degrees are large. As shown in [2], the only terms remaining are those that contain each variable exactly twice and for which k is even, implying that

$$\text{Tr}\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle \approxeq \frac{1}{z}\sum_{k=0}^{\infty}\frac{\text{Tr}\langle \mathbf{X}^{2k}\rangle}{z^{2k}}, \tag{22}$$

where

$$\begin{aligned}\text{Tr}\langle \mathbf{X}^{2k}\rangle &\approxeq \sum_{i_1,\ldots,i_k}\langle X^2_{i_1 i_2}X^2_{i_2 i_3}\cdots X^2_{i_k i_1}\rangle \\ &= N\left(NL\tilde{p}\right)^k C_k, \end{aligned}\tag{23}$$

C_k is the Catalan number, and

$$\begin{aligned}\tilde{p} &= [p_{in}(1-p_{in}) + p_{out}(1-p_{out})]/2 \\ &= \rho(1-\rho) - \Delta^2/4 \end{aligned}\tag{24}$$

is the average variance across the matrix entries $\{X_{ij}\}$. We note for sparse networks that $\tilde{p} \approx \rho$, which was the case considered by [2]. After substituting Eq. (23) into Eq. (22), we obtain $\text{Tr}\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle = t(z)$, where

$$t(z) = \frac{z - \sqrt{z^2-\lambda_2^2}}{\lambda_2^2/2}, \tag{25}$$

and

$$\lambda_2 = \sqrt{4NL\tilde{p}}, \tag{26}$$

which recovers Eq. (4) in the main text. Moreover, we substitute Eq. (25) into Eq. (19) to obtain Eq. (3) in the main text.

We now study the isolated eigenvalue λ_1 by solving the ensemble average of Eq. (13),

$$0 = \det\left(I - \langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle\langle\overline{\mathbf{B}}\rangle\right), \tag{27}$$

Because $\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle_{ij} = 0$ for $i \neq j$, we have that $\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle = t(z)\mathbf{I}$. It follows that $t(z)\theta_1$ is the largest eigenvalue of $\langle (z\mathbf{I}-\mathbf{X})^{-1}\rangle\langle\overline{\mathbf{B}}\rangle$. Because Eq. (27) requires this matrix to have an eigenvalue equal to one, we find that λ_1 solves

$$1 = t(\lambda_1)\theta_1. \tag{28}$$

Using that $t(z)$ has the inverse $t^{-1}(z) = z^{-1} + z\lambda_2^2/4$, we solve Eq. (28) for λ_1 to obtain

$$\begin{aligned}\lambda_1 &= t^{-1}(\theta_1^{-1}) \\ &= \theta_1 + \frac{\lambda_2^2}{4\theta_1}. \end{aligned}\tag{29}$$

After substituting the definition of θ_1 given by Eq. (18), we recover Eq. (5) in the main text. Setting $\lambda_1 = \lambda_2$ gives the solution $\lambda_2 = 2\theta_1$, which recovers Eq. (6) in the main text. As shown in [1], the corresponding eigenvector $\mathbf{v}$ is correlated with $\mathbf{u}$, which can be measured by the inner product

$$|\mathbf{u}^T\mathbf{v}|^{\mathbf{2}} = 1 - \frac{\lambda_2^2}{4\theta_1^2}. \tag{30}$$

We note that in the large N limit,

$$\frac{|\mathbf{u}^T\mathbf{v}|}{N^{1/2}} \approx \left|\frac{1}{N/2}\sum_{i=1}^{N/2} v_i\right|, \tag{31}$$

where the right hand side is the mean entry within a community. Therefore, we divide Eq. (30) by N and take the square root to obtain Eq. (9) in the main text.

[1] F. Benaych-Georges and R. R. Nadakuditi, Adv. in Math. **227**, 494 (2011).

[2] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**(18), 188701 (2012).