Supplementary Information for:
Bayesian Approach to Multilayer Stochastic Blockmodel and Network Changepoint Detection

Contents

1 Stochastic Blockmodels and Spectral Graph Theory

Spectral graph theory is a deterministic correspondence to our Bayesian approach. A large volume of literature in spectral graph theory provides information on the properties of the decomposed traits in our approach (Mieghem, 2011; Peixoto, 2013; X. Zhang, Nadakuditi, & Newman, 2014). Since analytic scrutiny of Bayesian approach is infeasible, we introduce recent findings in spectral graph theory to clarify the modeling scheme in the main text.

Following a formulation from Peixoto (2013), well-known representations of undirected networks, including adjacency matrix, graph Laplacian, normalized Laplacian and modularity matrix, can be expressed as $\mathbf{C} + \mathbf{M} = (\mathbf{C} + \mathbf{M} - \langle\mathbf{M}\rangle) + \langle\mathbf{M}\rangle$, where $\mathbf{C}$ is a random diagonal matrix and $\mathbf{M}$ is a random symmetric matrix. Each element in the mean interaction matrix $\mathbf{C} + \langle\mathbf{M}\rangle$, where off-diagonal entries of $\mathbf{M} - \langle\mathbf{M}\rangle$ have zero mean, is associated with the connection probability among the corresponding pair of nodes in the planted block diagram.

The eigenvalue spectrum of $\mathbf{C} + \mathbf{M}$ consists of two qualitatively different compartments for all matrix types mentioned. To be specific, $\mathbf{M} - \langle\mathbf{M}\rangle$ is associated with a bulk of continuous band of eigenvalues which does not involve block-diagram related information. The other compartment is associated with the eigenvalue spectrum of $\mathbf{C}+\langle\mathbf{M}\rangle$. Since $\mathbf{C}+\langle\mathbf{M}\rangle$ corresponds to the planted block diagram of the realized network, the set of eigenvalues belonging to its compartment approximate to the eigenvalues of a transformed version of the block diagram matrix represented by $\langle\mathbf{M}\rangle$. Accordingly, their corresponding eigenvectors reveal node group related information entailed in the planted block diagram.

The rank of $\mathbf{C}+\langle\mathbf{M}\rangle$ is associated with the number of dimensions one needs to recover for a block diagram assumed. If the block diagram has rank R, the rank of a degree-corrected version of $\mathbf{C} + \langle\mathbf{M}\rangle$ becomes $R - 1$. This relationship can be easily derived from the nullity theorem, $rank(\mathbf{C} + \langle\mathbf{M}\rangle) + nul(\mathbf{C} + \langle\mathbf{M}\rangle) = numcol(\mathbf{C} + \langle\mathbf{M}\rangle)$. For the $N \times N$ matrix $\mathbf{C}+\langle\mathbf{M}\rangle$, there are $N \times N - R + 1$ non-zero basis vectors for its null space. Because $\mathbf{C}+\langle\mathbf{M}\rangle$ has R unique column vectors due to the rank of the planted block diagram, and the degree correction procedure reduces its rank by 1. Same result holds for adjacency matrix $\mathbf{A}$, when excluding its principal eigenvalue which is associated with its eigenvector centrality. Figure S1 and Figure S2 supplement the theoretical results by showing the eigenspectra of homophilic block diagram induced networks with different number of node groups. The graphs demonstrate that the number of eigenvalues separated from the bulk is equal to the rank of the planted block structure.

For an arbitrary block diagram induced network, in which assortative and dissortative interactions are simultaneously present, one needs to choose $R - 1$ eigenvalues and eigenvectors. Such low-rank approximation successfully recovers planted group labels for an arbitrary block diagram induced network as exemplified in Figure S3. Eigenvalue associated with each axis characterizes the rule of node-to-node association on each dimension. Positive values indicate assortative interactions whereas negative values indicate disassortative interactions. For $\mathbf{A} - \mathbf{P}^{modul}$ and $\mathbf{A} - \mathbf{P}^{princ}$, eigenvalues associated with the block structures, which correspond to the spectrum for $\mathbf{C} + \langle\mathbf{M}\rangle$, are the ones with $R - 1$ eigenvalues largest in their absolute values (Peixoto, 2013). Their corresponding eigenvectors are the estimates for node positions in the role equivalence space. The discrepancy between two node trait distributions increases as the shapes of the induced diagrams or their reflections (exactly

opposite diagrams) become dissimilar.

Our Bayesian rank reduction algorithm naturally recovers latent traits that correspond to the $R-1$ largest eigenvalues in their absolute values. The MCMC procedure entailed in the algorithm proceeds to minimize $\|W - \hat{W}\|_F$ in which W is the matrix to be approximated and $\|W - \hat{W}\|_F$ is Frobenius norm of the error function calculated from the actual and the estimated. Due to Eckart-Young theorem, which states that for $\hat{W}$ with rank R, $\|W - \hat{W}\|_F$ is minimized when $\hat{W} = \sum_{r=1}^{R} \lambda_r \mathbf{u}^r \mathbf{u}^{rT}$ where λ_r and $\mathbf{u}^r$ are the rth eigenvalue and eigenvector of W when sorted in a decreasing order by its absolute value (Eckart & Young, 1936), such procedure outputs probabilistic estimates of $\{\lambda_r | 1 \leq r \leq R\}$ that are largest in their absolute values.
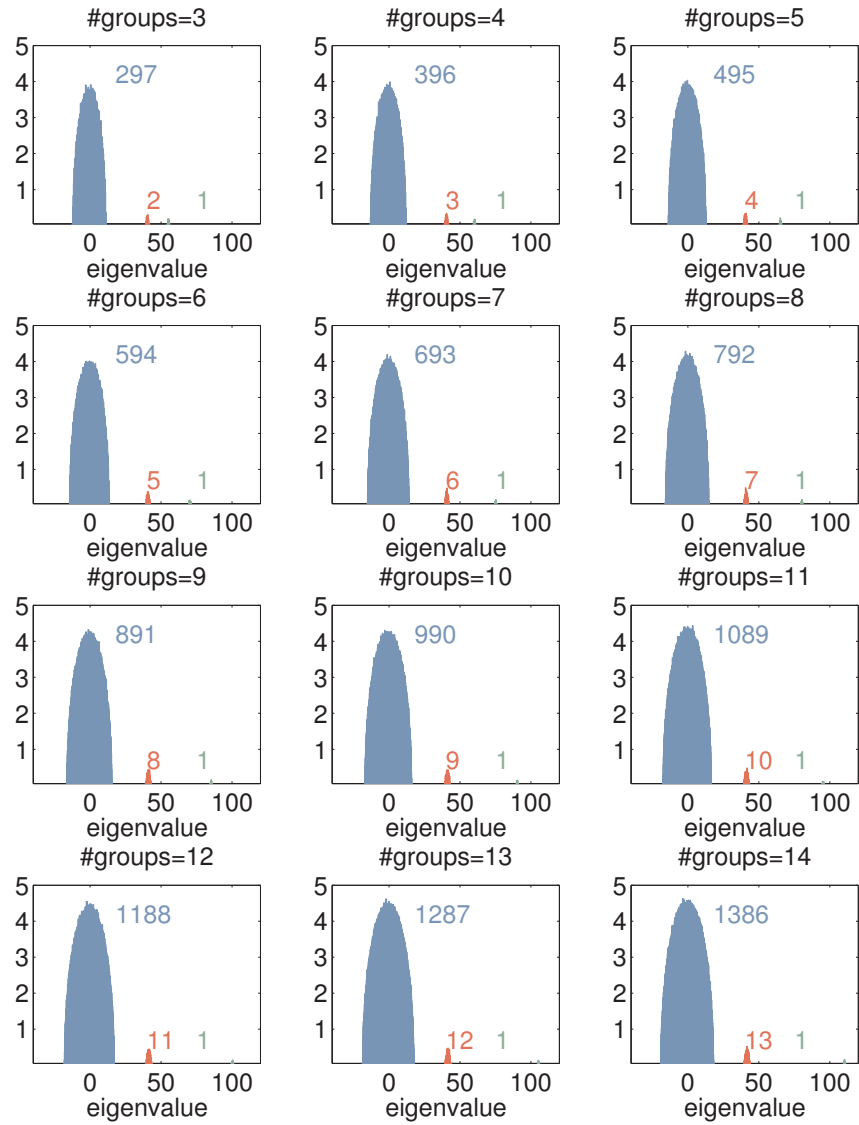
Figure S1: Eigenspectra of adjacency matrices for assortative networks with various number of groups. All networks entail assortative block diagrams where within-group link density is 0.45 and inter-group link desnity is 0.05. Each group consists of 100 nodes. Each number corresponds to the average number of eigenvalues belonging to the subspectrum in the same color contained in 100 stochastic realizations.
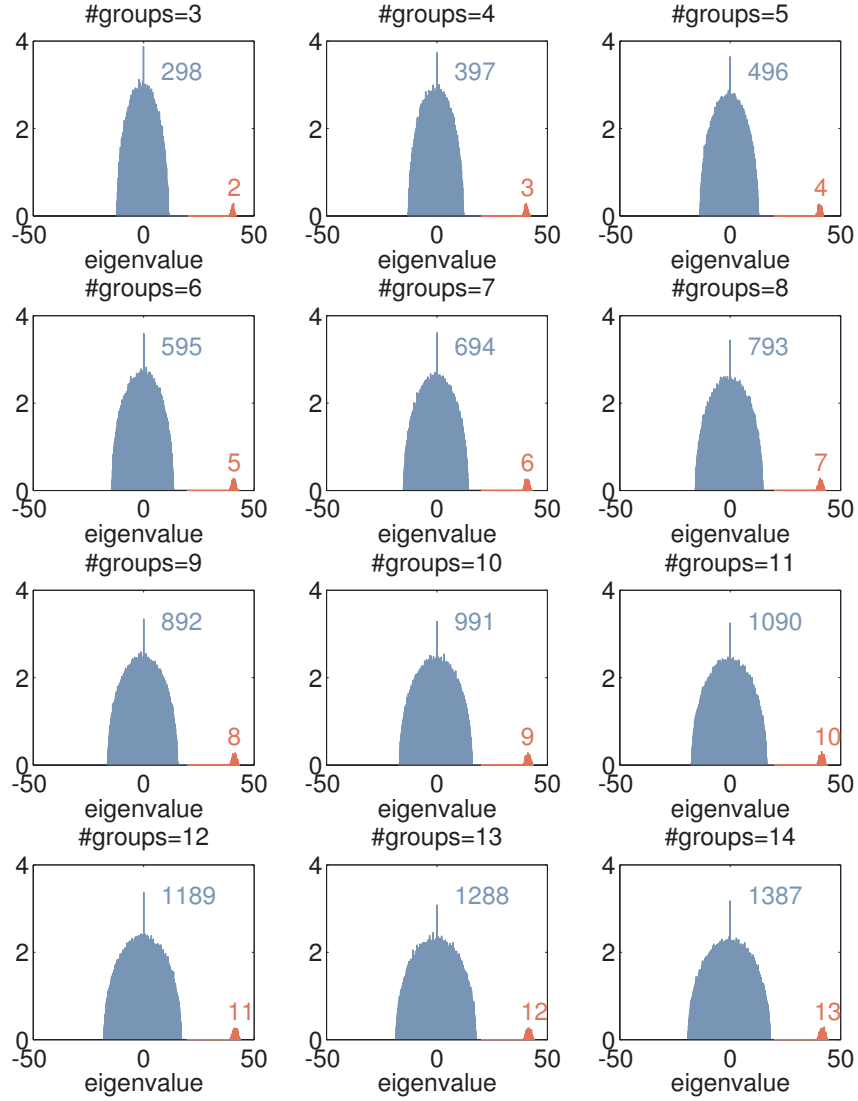
Figure S2: Eigenspectra of modularity matrices, $\mathbf{A} - \mathbf{P}^{modul}$, for assortative networks with various number of groups. All networks entail assortative block diagrams where within-group link density is 0.45 and inter-group link desnity is 0.05. Each group consists of 100 nodes. Each number corresponds to the average number of eigenvalues belonging to the subspectrum in the same color contained in 100 stochastic realizations.
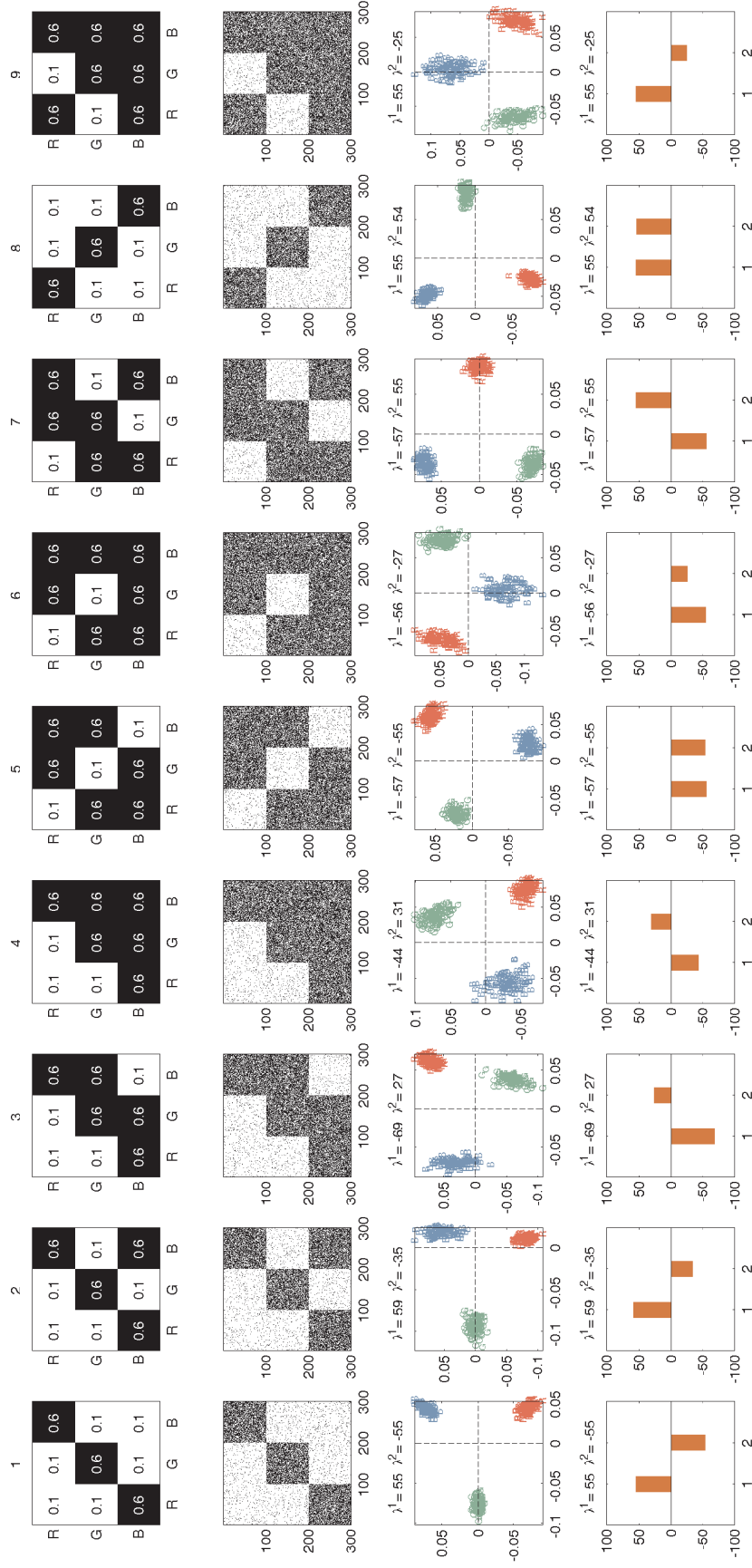
Figure S3: (First Row) Block diagrams embedded in the middle panel. All 9 rank 3 symmetric block diagrams with 3 groups are presented. Number in each block element indicates the connection probability of the given group pair. (Second) Realized networks from the Stochastic Blockmodels. (Third) 2-dimensional node traits, corresponding to eigenvectors associated with the first and second largest absolute eigenvalues, identified from the networks depicted in the above. (Fourth) Recovered λ corresponding to each dimension. Node color corresponds to its group label in the planted block diagram.

2   Clustering Techniques

Previous research discusses proper clustering schemes for latent node traits (Shi & Malik, 2000; Murphy, 2012; Richardson, Mucha, & Porter, 2009; Riolo & Newman, 2014; Z. Zhang, Jordan, et al., 2008). The most popular approach to classify the latent trait distribution is employing k-means clustering (Shi & Malik, 2000; Murphy, 2012). k-means clustering detects k subgroups of nodes, in which mean distance of node traits to the nearest k centroid positions becomes minimum. In our case, this procedure detect k groups with similar connection profiles in the structural equivalence space.

Subsequent studies introduced clustering techniques on a multidimensional space through simplex vector projection (Richardson et al., 2009; Riolo & Newman, 2014; Z. Zhang et al., 2008). A set of transformed simplex vectors is obtained as follow. For R dimensional node traits, $R + 1$ simplex vectors $\{\mathbf{w}_r | 1 \leq r \leq R + 1\}$ and $\{\mathbf{w}_s | 1 \leq s \leq R + 1\}$ are obtained satisfying $\mathbf{w}_r^T \mathbf{w}_s = c_{rs} \ \forall r, s$ where c_{rs} is a constant value assigned to $\{r, s\}$ group pair. For the group of vectors $\{\mathbf{w}_r | 1 \leq r \leq R+1\}$, each vectors are adjusted by group size distribution, rotated and stretched. After calculating the distances between these transformed simplex vectors and each node's trait vector, one selects the nearest transformed simplex vector, and label each node accordingly. For details of the iterative algorithms, see Richardson et al. (2009), Riolo and Newman (2014) and Z. Zhang et al. (2008).

3 Mutilayer Rank Reduction Algorithm

Let $\mathbf{B}_t \equiv \mathbf{A}_t - \mathbf{P}_t$ and $\mathcal{B} \equiv [\mathbf{B}_t | t \in T]$. Assuming a generalized latent space with R dimensions, the multilayer stochastic blockmodel using a layer specific null model $\mathbf{P}_t$ can be written as follows:

$$
\begin{aligned}
\mathbf{B}_t &= \mathbf{A}_t - \mathbf{P}_t \\
\mathbf{B}_t &= \mathbf{U}\mathbf{\Lambda}_t\mathbf{U}^T + \mathbf{E}_t \\
\mathbf{E}_t &\sim \text{ matrix normal}_{N \times N}(\mathbf{0}, \sigma^2\mathbf{I}_N, \mathbf{I}_N).
\end{aligned}
$$

where $\mathbf{U}$ indicates an $N \times R$ vector array and $\mathbf{\Lambda}_t = diag(\lambda_t^1, \cdots, \lambda_t^R)$ is $R \times R$ diagonal matrix at t.

Written in an array form,

$$
\begin{aligned}
\mathcal{B} &= \langle \mathbf{U}, \mathbf{V}, \mathbf{U} \rangle + \mathbf{E} \\
\mathbf{E} &\sim \text{ array normal}_{N \times N \times T}(\mathbf{0}, \sigma^2\mathbf{I}_N, \mathbf{I}_N, \mathbf{I}_T)
\end{aligned}
$$

and, equivalently in a scalar form,

$$
\begin{aligned}
b_{i,j,t} &= \langle \mathbf{u}_i, \mathbf{v}_t, \mathbf{u}_j \rangle + \epsilon_{i,j,t} \\
\epsilon_{i,j,t} &\sim \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

where $\mathbf{u}_i = (u_i^1, \cdots, u_i^R)$ denotes the ith row of $\mathbf{U}$ and $\mathbf{v}_t = (\lambda_t^1, \cdots, \lambda_t^R)$ the tth row of $\mathbf{V}$.

For parameter estimation, we adopt (Hoff, 2011)'s hierarchical Bayesian scheme to avoid overfitting. Prior distributions for hierarchical parameters of $\mathbf{U}$ are as follows:

$$
\begin{aligned}
\{\mathbf{u}_1, \ldots, \mathbf{u}_N\} &\sim \text{ multivariate normal}(\boldsymbol{\mu}_u, \Psi_u) \\
\boldsymbol{\mu}_u | \Psi_u &\sim \text{ multivariate normal}(\boldsymbol{\mu}_{0,u}, \Psi_u/\kappa_0) \\
\Psi_u &\sim \text{ inverse Wishart}(W_{0,u}, v_{0,u})
\end{aligned}
$$

Equivalently, prior distributions for hierarchical parameters of $\mathbf{V}$ are

$$
\begin{aligned}
\{\mathbf{v}_1, \ldots, \mathbf{v}_T\} &\sim \text{ multivariate normal}(\boldsymbol{\mu}_v, \Psi_v) \\
\boldsymbol{\mu}_v | \Psi_v &\sim \text{ multivariate normal}(\boldsymbol{\mu}_{0,v}, \Psi_v/\kappa_0) \\
\Psi_v &\sim \text{ inverse Wishart}(W_{0,v}, v_{0,v}).
\end{aligned}
$$

The prior distribution for σ^2 is an inverse gamma distribution:

$$
\sigma^2 \sim \text{ inverse gamma}(c_0, d_0).
$$

The resulting posterior distribution is

$$
\begin{aligned}
p(\mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_u, \boldsymbol{\mu}_v, \Psi_v, \sigma^2 | \mathcal{B}) \propto\; & p(\mathcal{B} | \mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_u, \boldsymbol{\mu}_v, \Psi_v, \sigma^2) \\
& p(\mathbf{U} | \boldsymbol{\mu}_u, \Psi_u) p(\boldsymbol{\mu}_u | \Psi_u) p(\Psi_u) \\
& p(\mathbf{V} | \boldsymbol{\mu}_v, \Psi_v) p(\boldsymbol{\mu}_v | \Psi_v) p(\Psi_v) p(\sigma^2),
\end{aligned}
$$

the samples of which are equivalent to samples from $p(\mathbf{U}, \mathbf{V}, \sigma^2 | \mathcal{B})$ after integrating out augmented parameters:

$$p(\mathbf{U}, \mathbf{V}, \sigma^2 | \mathcal{B}) = \int p(\mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_u, \boldsymbol{\mu}_v, \Psi_v, \sigma^2 | \mathcal{B}) d\boldsymbol{\mu}_u d\Psi_u d\boldsymbol{\mu}_v d\Psi_v.$$

The MCMC sampling sequence of the above model can be summarized multiple steps of the Gibbs sampling:

Step 1 For each layer, generate $\mathbf{B}_t = \mathbf{A}_t - \mathbf{P}_t$ by choosing a null model ($\mathbf{P}_t$).

Step 2 Initialize $(\mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_v, \boldsymbol{\mu}_v, \Psi_v, \sigma^2)$. For the starting values of $\mathbf{U}$ and $\mathbf{V}$, find the first R largest *absolute* eigenvalues and the corresponding eigenvectors of the mean matrix of $\mathcal{B}$ over T.

Step 3 Sample $\mathbf{U}$.

    1. Sample Ψ_u from inverse Wishart$((\mathbf{U}^T\mathbf{U} + W_{0,u}^{-1})^{-1}, N + R + 1)$.

    2. Sample $\boldsymbol{\mu}_u | \Psi_u$ from multivariate normal$(\mathbf{U}^T\mathbf{1}/(N+1), \Psi_u/(N+1))$.

    3. Sample $\mathbf{U}$ from matrix normal$_{N \times R}(\tilde{\mathbf{M}}_u, \mathbf{I}_N, \tilde{\Psi}_u)$ where

$$
\begin{aligned}
\tilde{\Psi}_u &= (\mathbf{Q}_u/\sigma^2 + \Psi_u^{-1})^{-1} \\
\tilde{\mathbf{M}}_u &= (\mathbf{L}_u/\sigma^2 + \mathbf{1}\boldsymbol{\mu}_u^T \Psi_u^{-1})\tilde{\Psi}_u \\
\mathbf{Q}_u &= (\mathbf{U}^T\mathbf{U}) \circ (\mathbf{V}^T\mathbf{V}) \\
\mathbf{L}_u &= \sum_{j,t} b_{\cdot,j,t} \otimes (\mathbf{U}_{j,\cdot} \circ \mathbf{V}_{t,\cdot})
\end{aligned}
$$

where $\circ$ indicates element wise matrix multiplication and $\otimes$ indicates the Kronecker product.

Step 4 Sample $\mathbf{V}$.

    1. Sample Ψ_v from inverse Wishart$((\mathbf{V}^T\mathbf{V} + W_{0,v}^{-1})^{-1}, T + R + 1)$.

    2. Sample $\boldsymbol{\mu}_v | \Psi_v$ from multivariate normal$(\mathbf{V}^T\mathbf{1}/(T+1), \Psi_v/(T+1))$.

    3. Sample $\mathbf{V}$ from matrix normal$_{T \times R}(\tilde{\mathbf{M}}_v, \mathbf{I}_T, \tilde{\Psi}_v)$ where

$$
\begin{aligned}
\tilde{\Psi}_v &= (\mathbf{Q}_v/\sigma^2 + \Psi_v^{-1})^{-1} \\
\tilde{\mathbf{M}}_v &= (\mathbf{L}_v/\sigma^2 + \mathbf{1}\boldsymbol{\mu}_v^T \Psi_v^{-1})\tilde{\Psi}_v \\
\mathbf{Q}_v &= (\mathbf{U}^T\mathbf{U}) \circ (\mathbf{U}^T\mathbf{U}) \\
\mathbf{L}_v &= \sum_{i,j} b_{i,j,\cdot} \otimes (\mathbf{U}_{i,\cdot} \circ \mathbf{U}_{j,\cdot}).
\end{aligned}
$$

4   Network Changepoint Detection Algorithm

Here we introduce the network changepoint detection algorithm in detail. Now we consider $\mathcal{B}$ as realizations of the stochastic process $\{\mathbf{B}_t\}_{t=1}^T$. The probability distribution of $\{\mathbf{B}_t\}_{t=1}^T$ depends on the realizations of hidden states $\{S_t\}_{t=0}^T$.

Utilizing the conditional independence of data given hidden states, the joint density of data (likelihood) and hidden states ($\mathbf{S}$) is

$$p(\mathcal{B}, \mathbf{S}|\mathbf{\Theta}, \mathbf{G}) = p(S_0|\mathbf{\Theta}, \mathbf{G}) \prod_{t=1}^T \underbrace{p(\mathbf{B}_t|\mathbf{B}^{t-1}, \mathbf{S}^t, \mathbf{\Theta}, \mathbf{G})}_{\text{one-step ahead predictive density}} \overbrace{p(S_t|\mathbf{S}^{t-1}, \mathbf{B}^{t-1}, \mathbf{\Theta}, \mathbf{G})}^{\text{distribution of hidden states at } t \text{ given the history}}$$

where $\mathbf{\Theta} = \{\mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_u, \boldsymbol{\mu}_v, \Psi_v, \sigma^2\}$, $\mathbf{S}^t$ indicates the history of hidden states from t to T, $\mathbf{G}$ is a transition matrix, and $\mathbf{B}^{t-1}$ indicate the history of observed data from $t-1$ to T.

In order to sample hidden states from the above joint distribution, we have to specify the transition matrix $\mathbf{G}$. Among many variants of the transition matrix that have been developed in the literature of hidden Markov model (Frühwirth-Schnatter, 2006; Cappe, Moulines, & Ryden, 2005), we choose Chib (1998)'s non-ergodic Markov chain for multiple changepoint detection.

Let $\gamma_{ij} = \Pr(S_t = j|S_{t-1} = i)$ denote the probability of going to state j from state i at time t when the state at $t-1$ is i. Then, the $M \times M$ transition matrix for a non-ergodic Markov chain takes the following form:

$$\mathbf{G} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 & \dots & 0 \\ 0 & \gamma_{22} & \gamma_{23} & \dots & 0 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ 0 & 0 & 0 & \gamma_{M-1,M-1} & \gamma_{M-1,M} \\ 0 & 0 & & 0 & 1 \end{pmatrix}.$$

We employ the Beta prior distribution for transition probabilities:

$$\gamma_{ii} \sim \mathcal{B}eta(a_0, b_0).$$

The joint posterior distribution can be decomposed into the product of two terms:

$$\begin{aligned} p(\mathbf{\Theta}, \mathbf{G}|\mathcal{B}) &= \int p(\mathbf{\Theta}, \mathbf{G}, \mathbf{S}|\mathcal{B}) d\mathbf{S} \\ &\quad \int \underbrace{p(\mathbf{\Theta}|\mathcal{B}, \mathbf{G}, \mathbf{S})}_{\text{Part 1}} \underbrace{p(\mathbf{G}, \mathbf{S}|\mathcal{B})}_{\text{Part 2}} d\mathbf{S}. \end{aligned}$$

Draws of the Gibbs sampling from two conditional posterior distributions, $p(\mathbf{\Theta}|\mathcal{B}, \mathbf{G}, \mathbf{S})$ and $p(\mathbf{G}, \mathbf{S}|\mathcal{B}, \mathbf{\Theta})$, provide proper samples from $p(\mathbf{\Theta}, \mathbf{G}|\mathcal{B})$:

Part 1: $p(\mathbf{\Theta}|\mathcal{B}, \mathbf{G}, \mathbf{S})$

Step 1 For each layer, generate $\mathbf{B}_t = \mathbf{A}_t - \mathbf{P}_t$ by choosing a null model ($\mathbf{P}_t$).

Step 2 Initialize $(\mathbf{U}, \mathbf{V}, \boldsymbol{\mu}_u, \Psi_v, \boldsymbol{\mu}_v, \Psi_v, \sigma^2, \mathbf{S}, \mathbf{G})$ in the same way with the above algorithm. Set the total number of changepoints M.

Step 3 Sample $\mathbf{U}_m$ which indicates $\mathbf{U}$ corresponding to state m.

1. Sample $\Psi_{u,m}$ from inverse Wishart$((\mathbf{U}_m^T \mathbf{U}_m + W_{0,u}^{-1})^{-1}, N + R + 1)$.

2. Sample $\boldsymbol{\mu}_{u,m} | \Psi_{u,m}$ from multivariate normal$(\mathbf{U}_m^T \mathbf{1}/(N+1), \Psi_u/(N+1))$.

3. Sample $\mathbf{U}_m$ from matrix normal$_{N \times R}(\tilde{\mathbf{M}}_{u,m}, \mathbf{I}_N, \tilde{\Psi}_{u,m})$ where

$$
\begin{aligned}
\tilde{\Psi}_{u,m} &= (\mathbf{Q}_{u,m}/\sigma_m^2 + \Psi_{u,m}^{-1})^{-1} \\
\tilde{\mathbf{M}}_{u,m} &= (\mathbf{L}_{u,m}/\sigma_m^2 + \mathbf{1}\boldsymbol{\mu}_{u,m}^T \Psi_{u,m}^{-1})\tilde{\Psi}_{u,m} \\
\mathbf{Q}_{u,m} &= (\mathbf{U}_m^T \mathbf{U}_m) \circ (\mathbf{V}_m^T \mathbf{V}_m) \\
\mathbf{L}_{u,m} &= \sum_{j,t:\, t \in S_t = m} b_{\cdot,j,t} \otimes (\mathbf{U}_{m,j,\cdot} \circ \mathbf{V}_{m,t,\cdot})
\end{aligned}
$$

Step 4 Sample $\mathbf{V}_m$ which indicates $\mathbf{V}$ corresponding to state m.

1. Sample $\Psi_{v,m}$ from inverse Wishart$((\mathbf{V}_m^T \mathbf{V}_m + W_{0,v}^{-1})^{-1}, T_m + R + 1)$.

2. Sample $\boldsymbol{\mu}_{v,m} | \Psi_{v,m}$ from multivariate normal$(\mathbf{V}_m^T \mathbf{1}/(T_m+1), \Psi_{v,m}/(T_m+1))$.

3. Sample $\mathbf{V}_m$ from matrix normal$_{T_m \times R}(\tilde{\mathbf{M}}_{v,m}, \mathbf{I}_{T_m}, \tilde{\Psi}_{v,m})$ where

$$
\begin{aligned}
\tilde{\Psi}_{v,m} &= (\mathbf{Q}_{v,m}/\sigma_m^2 + \Psi_{v,m}^{-1})^{-1} \\
\tilde{\mathbf{M}}_{v,m} &= (\mathbf{L}_{v,m}/\sigma_m^2 + \mathbf{1}\boldsymbol{\mu}_{v,m}^{T_m} \Psi_{v,m}^{-1})\tilde{\Psi}_{v,m} \\
\mathbf{Q}_{v,m} &= (\mathbf{U}_m^T \mathbf{U}_m) \circ (\mathbf{U}_m^T \mathbf{U}_m) \\
\mathbf{L}_{v,m} &= \sum_{i,j} b_{i,j,\cdot} \otimes (\mathbf{U}_{m,i,\cdot} \circ \mathbf{U}_{m,j,\cdot})
\end{aligned}
$$

Part 2: $p(\mathbf{G}, \mathbf{S} | \mathcal{B}, \boldsymbol{\Theta})$

Step 5 Sample $\mathbf{S}$

The joint conditional distribution of the latent states $p(S_0, \ldots, S_T | \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B})$ can be written as the product of T numbers of independent conditional distributions:

$$p(S_0, \ldots, S_T | \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B}) = p(S_T | \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B}) \ldots p(S_t | \mathbf{S}^{t+1}, \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B}) \ldots p(S_0 | \mathbf{S}^1, \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B}).$$

Using Bayes' Theorem,

$$p(S_t | \mathbf{S}^{t+1}, \boldsymbol{\Theta}, \mathbf{G}, \mathcal{B}) \quad \propto \quad p(S_t | \boldsymbol{\Theta}, \mathbf{G}, \mathbf{B}_{1:t}) p(S_{t+1} | S_t, \mathbf{G}).$$

The second part on the right hand side is a transition probability at t, which can be obtained from a sampled transition matrix ($\mathbf{G}$). The first part on the right hand side is simulated via a forward-filtering-backward-sampling algorithm as shown in Chib (1998).

Step 6 Sample **G**

$$\gamma_{kk} \sim \mathcal{B}eta(a_0 + j_{k,k} - 1, b_0 + j_{k,k+1})$$

where γ_{kk} is the probability of staying when the state is k, and $j_{k,k}$ is the number of jumps from state k to k, and $j_{k,k+1}$ is the number of jumps from state k to $k+1$.

5 Scalability of the Changepoint Detection Algorithm

Figure S4 illustrates the scaling behavior of the run time of our changepoint detection algorithm. The run time scales linearly with the squared value of the number of nodes and the number of layers. As indicated in the main text, the R-implemented version of the algorithm, used in the study, is reasonably fast, producing 100 samples in 30 seconds on a laptop for the simulated examples analyzed. It is still affordable for larger networks with 1,000 nodes, producing 100 samples within an hour. We expect that C/C++ implementation of the algorithm will yield a substantial improvement in its computational speed.
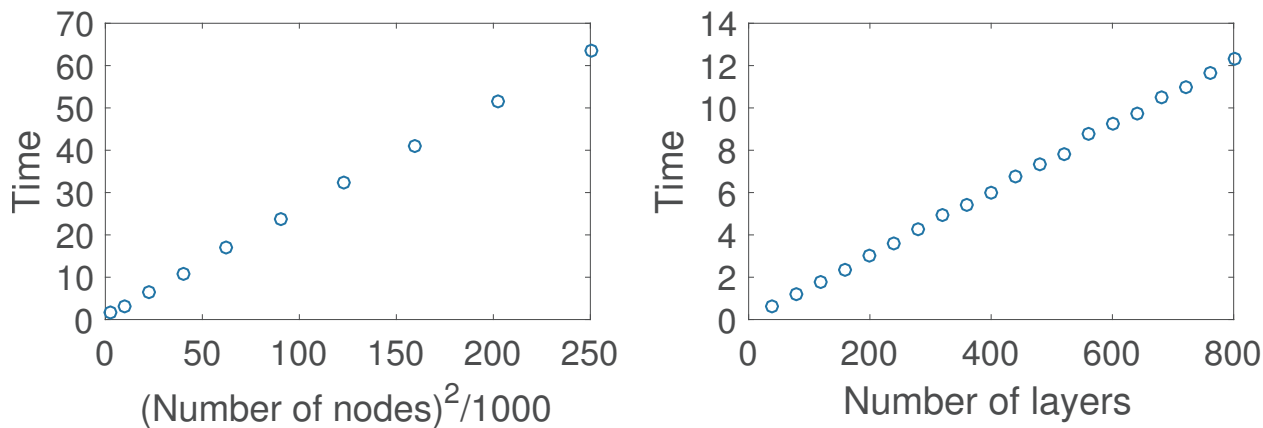

Figure S4: The scaling behavior of the run time for the changepoint detection algorithm by the network size and the number of layers. The unit of time on the y-axis does not convey substantive information. Each point is the average value for 10 runs of the core-periphery transition time series. All types of the transitions show almost identical scaling patterns.

References

Cappe, O., Moulines, E., & Ryden, T. (2005). *Inference in hidden markov models.* Springer-Verlag.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of econometrics*, *86*(2), 221–241.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models.* Heidelberg: Springer Verlag.

Hoff, P. D. (2011). Hierarchical multilinear models for multiway data. *Computational Statistics & Data Analysis*, *55*, 530 – 543.

Mieghem, V. (2011). *Graph spectra for complex networks.* Cambridge University Press.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.* MIT press.

Peixoto, T. P. (2013). Eigenvalue spectra of modular networks. *Physical review letters*, *111*(9), 098701.

Richardson, T., Mucha, P. J., & Porter, M. A. (2009). Spectral tripartitioning of networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, *80*(3), 036111.

Riolo, M. A., & Newman, M. E. (2014). First-principles multiway spectral partitioning of graphs. *Journal of Complex Networks*, *2*(2), 121–140.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(8), 888–905.

Zhang, X., Nadakuditi, R. R., & Newman, M. E. (2014). Spectra of random graphs with community structure and arbitrary degrees. *Phys Rev E Stat Nonlin Soft Matter Phys*, *89*(4), 042816.

Zhang, Z., Jordan, M. I., et al. (2008). Multiway spectral clustering: A margin-based perspective. *Statistical Science*, *23*(3), 383–403.