Global spectral clustering in dynamic networks

Fuchen Liu[a], David Choi[b], Lu Xie[a], and Kathryn Roeder[a,c,1]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213; and [c]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved December 8, 2017 (received for review October 23, 2017)

Community detection is challenging when the network structure is estimated with uncertainty. Dynamic networks present additional challenges but also add information across time periods. We propose a global community detection method, persistent communities by eigenvector smoothing (PisCES), that combines information across a series of networks, longitudinally, to strengthen the inference for each period. Our method is derived from evolutionary spectral clustering and degree correction methods. Data-driven solutions to the problem of tuning parameter selection are provided. In simulations we find that PisCES performs better than competing methods designed for a low signal-to-noise ratio. Recently obtained gene expression data from rhesus monkey brains provide samples from finely partitioned brain regions over a broad time span including pre- and postnatal periods. Of interest is how gene communities develop over space and time; however, once the data are divided into homogeneous spatial and temporal periods, sample sizes are very small, making inference quite challenging. Applying PisCES to medial prefrontal cortex in monkey rhesus brains from near conception to adulthood reveals dense communities that persist, merge, and diverge over time and others that are loosely organized and short lived, illustrating how dynamic community detection can yield interesting insights into processes such as brain development.

community detection | gene expression networks | dynamic networks

Networks or graphs are used to display connections within a complex system. The vertices in a network often reveal clusters with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, could arise from functionality of distinct components of the network, e.g., genes coregulating a cellular process.

Statistical theory (1, 2) has mostly focused on static networks, observed as a single snapshot in time or developmental epoch. In reality, networks are generally dynamic, and it is of substantial interest to visualize and model their evolution. Applications abound, e.g., social networks in Twitter, dynamic diffusion networks in physics, and gene coexpression networks for developing brains. Community detection is vital in all of these areas to illustrate the structure of the relationship of network nodes and how they change over time. While statistical inference in static networks is well established (3–7), how to combine the information in dynamic networks is comparatively less understood. Recent works have sought to extend community detection to dynamic networks (8–15) and to centrality (16) and to extend clustering to dynamic data (17).

Our method, persistent communities by eigenvector smoothing (PisCES), implements degree-corrected spectral clustering, with a smoothing term to promote similarity across time periods, and iterates until a fixed point is achieved. Specifically, this global spectral clustering approach combines the current network with the leading eigenvector of both the previous and future results. The combination is formed as an optimization problem that can be solved globally under moderate levels of smoothing when the number of communities is known. We find that it is important to choose appropriate levels of both smoothing and model order, as well as to balance regularization with "letting the data speak," and we use data-driven methods to do so.

Dynamic networks derived from gene coexpression networks reveal community structure among the genes that develops over spatial or temporal periods, providing a fine-scale view of the inner workings of cellular mechanisms. While it is known that gene expression varies dramatically over developmental periods in the brain, the specific changes in gene communities for a developing brain are not fully understood. Understanding brain disorders like autism spectrum disorder and schizophrenia have been particularly challenging to scientists because of the large number of genes implicated. The clustering of risk genes for neurodevelopmental disorders in specific spatiotemporal periods can help to explain the nature of these disorders.

Recently a rich source of data has become available pertaining to this question. Transcription for numerous samples in rhesus monkeys is assessed over a dense set of pre- and postnatal periods (18). Once the data are divided into fine anatomical regions and developmental periods, however, sample sizes are very small (<20), making it difficult to estimate the gene–gene adjacency matrix from correlated expression of genes. PisCES can significantly improve the power of community detection in this scenario.

We illustrate the power of dynamic community detection methods by investigating the gene communities as they develop over age and cortical layers in the medial prefrontal cortex. The analysis reveals that while many communities are restricted to particular developmental periods, others persist, illustrating the existence of change points as well as periods of persistent community structure. For example, communities enriched for neural projection guidance (NPG) are much more tightly clustered during prenatal development, peaking just before birth. This pattern is consistent with existing knowledge about neurodevelopment. Genes with the annotation NPG have been linked to autism spectrum disorder (9) and our method can provide critical insight into the interactions of these genes.

Significance

Statistical theory has mostly focused on static networks observed as a single snapshot in time. In reality, networks are generally dynamic, and it is of substantial interest to discover the clusters within each network to visualize and model their connectivities. We propose the persistent communities by eigenvector smoothing algorithm for detecting time-varying community structure and apply it to a recent dataset in which gene expression is measured during a broad range of developmental periods in rhesus monkey brains. The analysis suggests the existence of change points as well as periods of persistent community structure; these are not well estimated by standard methods due to the small sample size of any one developmental period or region of the brain.

Author contributions: F.L., D.C., and K.R. designed research; F.L., D.C., and K.R. performed research; F.L., D.C., and K.R. contributed new reagents/analytic tools; F.L., D.C., L.X., and K.R. analyzed data; and F.L., D.C., L.X., and K.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: roeder@andrew.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1718449115/-/DCSupplemental.

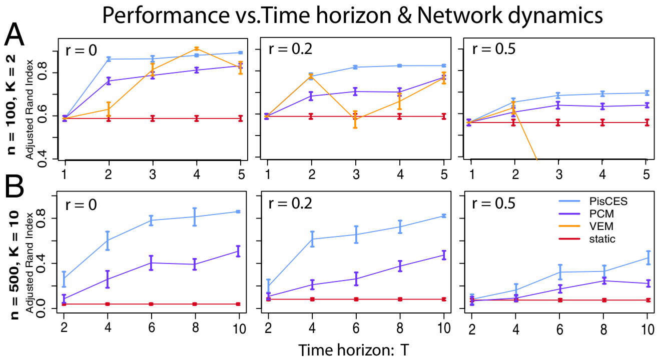

Fig. 1. Performance on synthetic networks as a function of time horizon and class dynamics, as measured by the adjusted Rand index between true and estimated community labels. Networks were generated under a dynamic DCBM (Eqs. **13**–**15**) with three key parameters: p_{in} and p_{out}, which determine the in-cluster and out-of-cluster edge probability/density, and r, which determines the amount of change in cluster memberships between consecutive networks (0 for no change). For A, $K=2, n=100, p_{\text{in}}=(0.2, 0.25)$, $p_{\text{out}}=0.1$, and for B, $K=10$, $n=500$, $p_{\text{in}}=(0.2, 0.35)$, $p_{\text{out}}=0.1$. Shown are 100 simulations per data point.

Methods

Spectral Methods for Static Networks. Spectral clustering is a popular class of methods for finding communities in a static network, and many variations have been discussed in the literature (19–22). A prototypical method is given by ref. 6. Given a symmetric $n \times n$ adjacency matrix A and a fixed number of communities K, the method computes the degree-normalized (or "Laplacianized") adjacency matrix L, which is given by

$$L = D^{-1/2}AD^{-1/2} \quad \text{where} \quad D = \text{diag(degree)}. \qquad [1]$$

The method then returns the clusters found by K-means clustering on the eigenvectors of L corresponding to its K largest eigenvalues in absolute value. Methods for choosing K include refs. 23–26.

Eigenvector Smoothing for Dynamic Networks. Let $A_1, \ldots, A_T$ denote a time series of symmetric adjacency matrices, and for $t=1,\ldots,T$, let L_t denote the Laplacianized version of A_t, as given by Eq. **1**. Let K be fixed, and let $V_t \in \mathbb{R}^{n\times K}$ denote the matrix whose columns are the K leading eigenvectors of L_t. Let $U_t = V_tV_t^T$, the projection matrix onto the column space of V_t.

In static spectral clustering, one would apply K-means clustering to $V_1, \ldots, V_T$ separately. To share signal strength over time, a simplified form of PisCES would solve the following optimization problem, which returns a sequence of matrices $\bar{U}_1, \ldots, \bar{U}_T$ that are smoothed versions of $U_1, \ldots, U_T$,

$$\min_{\bar{U}_1,\ldots,\bar{U}_T} \sum_{t=1}^{T} \|U_t - \bar{U}_t\|_F^2 + \alpha \sum_{t=1}^{T-1} \|\bar{U}_t - \bar{U}_{t+1}\|_F^2 \qquad [2]$$
$$\text{subject to } \bar{U}_t \in \{VV^T : V \in \mathbb{R}^{n\times K}, V^TV = I\} \ \forall\ t,$$

and then apply K-means clustering to the eigenvectors of each smoothed matrix $\bar{U}_1, \ldots, \bar{U}_T$ separately.

The optimization problem Eq. **2** is nonconvex and, to the best of our knowledge, no efficient methods for its global solution currently exist. We propose the following iteration,

$$\bar{U}_1^{\ell+1} = \Pi_K(U_1 + \alpha\bar{U}_2^{\ell}) \qquad [3]$$

$$\bar{U}_t^{\ell+1} = \Pi_K(\alpha\bar{U}_{t-1}^{\ell} + U_t + \alpha\bar{U}_{t+1}^{\ell}),\ t = 2, \ldots, T-1 \qquad [4]$$

$$\bar{U}_T^{\ell+1} = \Pi_K(\alpha\bar{U}_{T-1}^{\ell} + U_T), \qquad [5]$$

where the mapping Π_K extracts the K leading eigenvectors and is given for a matrix M by

$$\Pi_K(M) = \sum_{k=1}^{K} v_kv_k^T,$$

where $v_1, \ldots, v_K$ are the K leading eigenvectors of M. To initialize, we set $\bar{U}_t^0 = U_t$ for $t=1,\ldots,T$.

Convergence result. *Theorem 1* is proved in *SI Appendix*, section S1, and states that for proper choice of α, the iterative algorithm given by Eqs. **3**–**5** converges to the global optimum of Eq. **2**:

Theorem 1. *For* $\alpha < \frac{1}{4\sqrt{2}+2} \approx 0.13$, *the iterations* Eqs. **3**–**5** *converge to the global minimizer of Eq.* **2** *under any feasible initialization.*

Intuition. To build intuition for the behavior of the method, observe that if $\bar{U}_{t-1}^{\ell}$ and $\bar{U}_{t+1}^{\ell}$ are orthogonal to U_t, and if $\alpha < 1/2$, then Eq. **4** implies that $\bar{U}_t^{\ell+1} = U_t$, so that the information at neighboring times is effectively ignored. Along these lines, in simulations where a change point exists in the community memberships, smoothing is suppressed automatically at this time point. This suggests that the method applies a variable amount of smoothing to each time step, which goes to zero as the community memberships at neighboring times become uncorrelated.

For $t=1,\ldots,T$ and $i=1,\ldots,n$, let $x_{ti} \in \mathbb{R}^K$ denote the ith row of the matrix V_t. For each time step t, static spectral clustering seeks to find cluster centroids $\mu_k \in \mathbb{R}^K$ for $k=1,\ldots,K$ and a cluster assignment vector $z \in [K]^n$ to optimize the K-means objective function

$$\min_{\{\mu_k\},z} \sum_{i=1}^{n} \|x_{ti} - \mu_{z(i)}\|^2.$$

In *SI Appendix*, section S2, we show that Eq. **2** can be derived as a spectral relaxation of the following smoothed K-means objective, over time-varying centroids and assignment vectors $\{\mu_{tk}\}_{t=1,k=1}^{T,K}$ and $\{z_t\}_{t=1}^{T}$,

$$\min_{\{\mu_{tk}\},\{z_t\}} \sum_{t=1}^{T}\sum_{i=1}^{n} \|x_{ti} - \mu_{t,z_t(i)}\|^2 + \frac{\alpha}{2}\sum_{t=1}^{T-1}\Delta(z_t, z_{t+1}), \qquad [6]$$

where the penalty term $\Delta(z_t, z_{t+1})$ utilizes the "chi-square" metric for comparing partitions (27, 28), which ensures smoothness of the cluster assignments. However, the objective function allows the density of the blocks to change drastically across different developmental periods.

Laplacian smoothing. Eqs. **7**–**9** give a variation in which $L_1, \ldots, L_T$ are used more directly,

$$\bar{U}_1^{\ell+1} = \Pi_K(|L_1| + \alpha\bar{U}_2^{\ell}) \qquad [7]$$

$$\bar{U}_t^{\ell+1} = \Pi_K(\alpha\bar{U}_{t-1}^{\ell} + |L_t| + \alpha\bar{U}_{t+1}^{\ell}) \quad t = 2, \ldots, T-1 \qquad [8]$$

$$\bar{U}_T^{\ell+1} = \Pi_K(\alpha\bar{U}_{T-1}^{\ell} + |L_T|), \qquad [9]$$

where $|L_t|$ denotes the matrix L_t with its eigenvalues replaced by their absolute values.

How should these iterations be interpreted? Analogous to eigenvector smoothing, we show in *SI Appendix*, section S3 that Eqs. **7**–**9** globally solve the optimization problem

$$\min_{\bar{U}_1,\ldots,\bar{U}_T} \sum_{t=1}^{T} \||L_t| - \bar{U}_t\|_F^2 + \alpha\sum_{t=1}^{T-1}\|\bar{U}_t - \bar{U}_{t+1}\|_F^2 \qquad [10]$$
$$\text{subject to } \bar{U}_t \in \{VV^T : V \in \mathbb{R}^{n\times K}, V^TV = I\} \ \forall\ t,$$

for certain values of α and that this problem can be derived as a spectral relaxation to an analogous version of Eq. **6**, in which x_{ti} now denotes the ith row of the square root of $|L_t|$.

Cross-validation. To choose α, we use a cross-validation method for degree-corrected clustering, found in ref. 24 (*SI Appendix*, section S4).

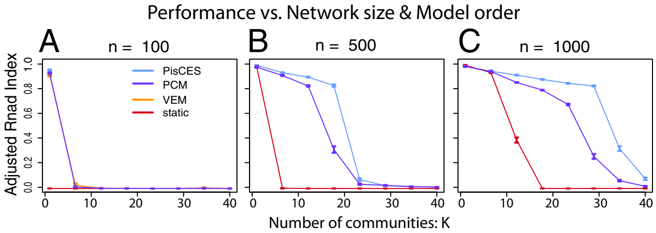

Fig. 2. Performance on synthetic networks as a function of network size and number of communities, as measured by ARI between true and estimated community labels. (*A*–*C*) Networks are generated under the dynamic DCBM with $n \in \{100, 500, 1{,}000\}$, $K \in \{1, 4, 8, 12, \ldots, 40\}$, $r = 0.1$, $p_{\text{in}} = (0.2, 0.5)$, $p_{\text{out}} = 0.1$, $T = 10$. Shown are 100 simulations per data point.

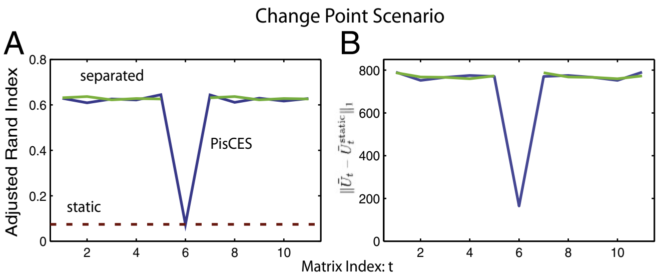

Fig. 3. Performance of PisCES in a scenario with outlier at time $t = 6$ (main text). Simulations were generated from a dynamic DCBM with $n = 500$, $K = 10$, $p_{\mathrm{in}} = (0.3, 0.3)$, $p_{\mathrm{out}} = 0.1$, $r = 1$ for A_6, and $r = 0.1$ outside the change point. (*A*) ARI performance of PisCES applied to $A_1, \ldots, A_{11}$ (blue line), static spectral clustering (static, red dashed line), and PisCES applied separately to $A_1, \ldots, A_5$ and to $A_7, \ldots, A_{11}$ ("separated," green lines). (*B*) $\|\bar{U}_t - \bar{U}_t^{\mathrm{static}}\|_1$, where $\bar{U}_t$ is the output of PisCES (green) or "separate" (blue), and $\bar{U}_t^{\mathrm{static}}$ is the output of static.

PisCES. PisCES extends Laplacian smoothing by allowing the number of classes K to be unknown and possibly varying over time.

This is accomplished by replacing the operator Π_K in Eqs. **7**–**9** with a new operator Π, which requires a model order selection method κ to choose the number of eigenvectors from the data:

$$\Pi(M) = \sum_{k=1}^{\kappa(M)} v_k v_k^T.$$

Here κ: $\mathbb{R}^{n\times n} \mapsto \mathbb{N}$ is a function that determines the number of eigenvectors to be returned, and $v_1, \ldots, v_{\kappa(M)}$ are the eigenvectors of M corresponding to its $\kappa(M)$ largest eigenvalues in absolute value.

To choose the model order $\kappa(M)$, in principle one could use or adapt existing methods for eigenvector selection, such as refs. 23–26. Alternatively, in *Model Order Selection* κ we describe a new method that can be adapted to the specific assumptions of the data-generating process.

The iterates for PisCES (that is, Eqs. **7**–**9** with Π_K replaced by Π) are heuristic in that no convergence theorems are known. However, simulations suggest that they can help when K cannot be accurately estimated from any single network A_t due to noise.

To estimate the clusters, K means is applied to the eigenvectors of $\bar{U}_1, \ldots, \bar{U}_T$.

Model Order Selection κ**.** Given a Laplacianized matrix $L \in \mathbb{R}^{n\times n}$ with eigenvalues $|\lambda_1| \geq \cdots \geq |\lambda_n|$, $\kappa(L)$ is given by

$$\kappa(L) = \min\{K : |\lambda_i| - |\lambda_{i+1}| < \delta, \text{ for all } i > K\}, \qquad \textbf{[11]}$$

where δ is the threshold for the "noise" eigenvalues of a null model for the data-generating process.

In generic network settings, a suitable null model could be to simulate Laplacianized Erdos–Renyi adjacency matrices $L^{(ER)}$ with size and density matching L and eigenvalues $|\lambda_1^{(ER)}| \geq \cdots |\lambda_n^{(ER)}|$ and to return the 0.95 quantile of the largest eigengap excluding $\lambda_1^{(ER)}$:

$$\delta = \mathrm{quantile}_{0.95}\left[\max\{|\lambda_i^{(ER)}| - |\lambda_{i+1}^{(ER)}|, i \geq 2\}\right]. \qquad \textbf{[12]}$$

This approach may be appropriate when the observation noise is assumed to be independent across dyads (such as a stochastic block model)—e.g., when the dyads are the observations.

A null model assuming dyadically independent observation noise may not be appropriate when networks $A_1, \ldots, A_T$ are transformations of empirical correlation matrices, as in *Results*. Instead, a more appropriate choice may be to generate random samples that are matched in number to the observations that are used to form the correlation matrices underlying $A_1, \ldots, A_T$. Further details on such null models can be found in *SI Appendix*, section S5 and Figs. S1–S3.

Simulations

Simulations suggest that PisCES works well in practice. Here we show three examples of simulation performance; more results can be found in *SI Appendix*, section S6 and Figs. S4–S7.

Figs. 1 and 2 show simulations where $A_1, \ldots, A_T$ are symmetric adjacency matrices each generated by a dynamic degree-corrected block model (DCBM) (8), where

$$[A_t]_{ij} \sim \text{ Bernoulli}\left(\psi_{ti}\psi_{tj} B^{(t)}_{z_{ti},z_{tj}}\right) \qquad i, j \in [n], j > i, \qquad \textbf{[13]}$$

with $[A_t]_{ij} = [A_t]_{ji}$. Here $z_t \in [K]^n$ and $\psi_t \in \mathbb{R}^n$ are vectors of class labels and degree parameters, and $B^{(t)} \in [0,1]^{K\times K}$ is a connectivity matrix. z_t evolves over time by

$$z_{(t+1)i} = \begin{cases} z_{ti} & \text{with prob. } 1 - r \\ \text{Multinomial}(\frac{1}{K}, \ldots, \frac{1}{K}) & \text{otherwise,} \end{cases} \qquad \textbf{[14]}$$

where r denotes the probability that a node changes clusters, and $\psi^{(t)}$ and $B^{(t)}$ are randomized at each stage by

$$\psi_t = 1/2 + \pi_t/n \qquad \textbf{[15]}$$

$$B^{(t)}_{lk} = \begin{cases} \text{Unif}(p_{\text{in}}^{(1)}, p_{\text{in}}^{(2)}) & l = k \\ p_{\text{out}} & l \neq k, \end{cases} \qquad \textbf{[16]}$$

where π_t is a random permutation of $1{:}n$, and $p_{\text{in}} = (p_{\text{in}}^{(1)}, p_{\text{in}}^{(2)})$ and p_{out} are in-cluster and between-cluster density parameters.

For comparison, we evaluate a variational–expectation-maximization ("VEM") likelihood method for the dynamic DCBM (8) and a spectral method ["PCM" (preserving cluster membership)] (17); as a baseline we contrast these with static spectral clustering ("static"). (Due to its computational complexity, results for VEM are shown for $n = 100$ only, as its runtimes were impractical for $n \geq 500$.)

Fig. 1 shows improving performance for PisCES as the time horizon T increases (which allows greater sharing of information) and decreasing performance as the nodal classes evolve more rapidly over time. Fig. 2 shows increasing performance for all methods as the network size n increases and decreasing

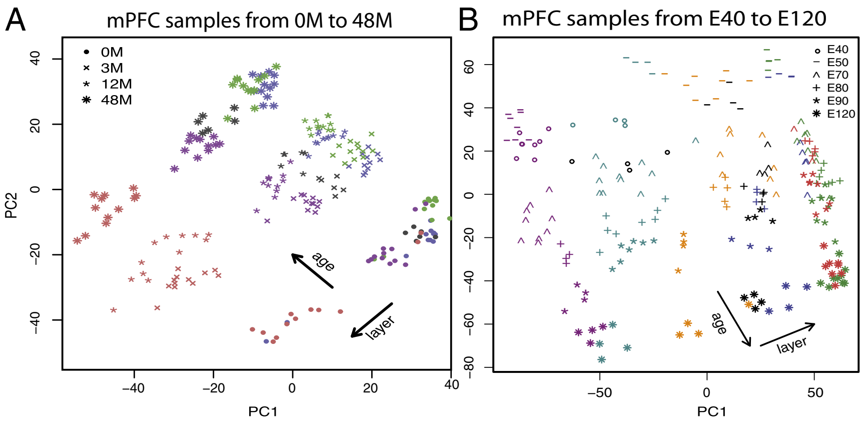

Fig. 4. (*A* and *B*) Top two principal components for mPFC samples in (*A*) postnatal ages 0 M to 48 M and (*B*) prenatal ages E40–E120. Age and layer of each sample are depicted by marker shape and color, respectively. Shown are postnatal layers L2 (blue), L3 (green), L4 (black), L5 (purple), and L6 (red) and prenatal layers VZ (purple), SZ/IFZ (cyan), IZ/OFZ (orange), SP (blue), L5/L6/CPi (green), L2/L3/CPo (red), and MZ (black). Samples obtained from a given age and layer are relatively homogeneous in their rates of transcription.

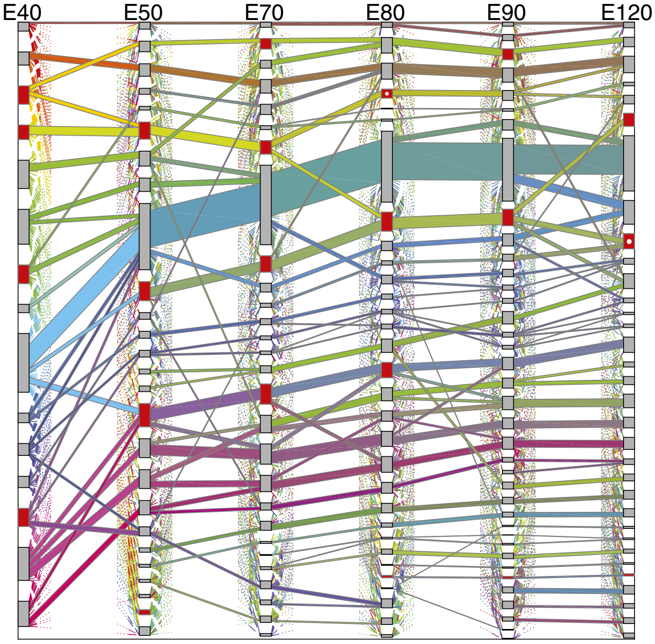

Fig. 5. Sankey plot for prenatal cases. Gray and red boxes denote communities, with height indicating community size. Colored "flows" denote groups of genes moving between communities, with height indicating flow size. To reduce clutter, only large flows (>100 genes or >15% of its source and destination community) are shown; small flows are partially drawn using dotted lines. Each flow's color is determined by its gene membership and equals the mixture of the colors of its input flows. NPG-enriched communities are denoted by red boxes, with ASD-enriched communities further marked by a white circle (E80, fourth from top, and E120, ninth from top).

performance as the number of nodal classes K increases. In all cases, PisCES performs comparably to or better than the other methods.

In *Methods*, we mention that PisCES suppresses the amount of smoothing at times where an outlier or a change point exists in the data. To demonstrate this, Fig. 3 shows the results of a simulation with $T = 11$ where the class labels evolve as a dynamic DCBM except at times $t = 6$ and $t = 7$, at which point they are randomized with no dependence on past time steps. Fig. 3 shows that the estimated $\bar{U}_t$ at $t = 6$ closely resembles the output of static spectral clustering, while at the other time steps the output closely resembles that of PisCES applied to $A_1, \ldots, A_5$ and to $A_7, \ldots, A_{11}$.

Results

Background. The transcriptional patterns of the developing primate brain are of keen interest to neuroscientists and others interested in neurological and psychiatric disorders. Bakken et al. (18) provide a high-resolution transcriptional atlas of rhesus monkeys (*Macaca mulatta*) built from recorded samples of gene expression, including expression of 9,173 genes that can be mapped directly to humans. The samples span six prenatal ages from 40 embryonic days to 120 embryonic days (E40–E120) and four postnatal ages from 0 mo to 48 mo after birth (0 M to 48 M; *SI Appendix*, section S7). These ages represent key stages of development in the prenatal phase and key milestones postnatally (newborn, juvenile, teen, and mature). The samples can also be divided into different regions of the brain and further into different layers (i.e., subregions) within each region. For example, in postnatal ages, the medial prefrontal cortex region (mPFC) can be divided into contiguous layers L6–L2, with a more complex layer parcellation in prenatal ages (*SI Appendix*, Table S2).

The changes in gene expression over spatiotemporal periods have been fairly well studied, but changes in coexpression are less well understood. Bakken et al. (18) used weighted gene coexpression network analysis (WGCNA) (29) to identify gene communities in the cortex of rhesus monkeys at each age; roughly speaking, this corresponds to dividing the samples by age, constructing gene coexpression networks from each subset of samples, and then performing a clustering analysis on each network. Our goal is to take this investigation farther using PisCES, but in this present work we give exploratory results only to demonstrate proof of concept.

Due to its importance in understanding developmental brain disorders, we focus on the mPFC and use only the 423 samples corresponding to that region in our analysis (*SI Appendix*, Fig. S8). For these samples, variability in the first two components of principal components analysis is explained roughly equally by their layer or by their age, suggesting that they may be grouped by either variable (Fig. 4). To illustrate that PisCES can be used over time or location, we choose to divide the 209 prenatal samples by age category (E40–E120) and divide the 214 postnatal samples by their layer within the mPFC region (L6–L2).

For each group of samples, coexpression networks (gene networks in which edges encode coexpression level between gene pairs) are constructed by the procedure of refs. 18 and 29. Specifically, adjacency matrix A_t for group $t \in$ (E40–E120, L6–L2) is generated by soft thresholding the empirical correlation matrix

$$[A_t]_{ij} = |\operatorname{corr}(g_i, g_j)|^6, \quad i \neq j, \qquad \textbf{[17]}$$

where g_i and g_j are the recorded expression levels for genes i and j in the samples belonging to group t.

PisCES Results. PisCES—using Eq. **S17** for δ—is run separately on the pre- and postnatal samples. The detected communities vary in composition, size, and density (Dataset S1); see *SI Appendix*, Table S3 for descriptive summaries. For pre- and postnatal analyses we compare the performance of algorithms, using a measure of log-likelihood (*SI Appendix*, Fig. S9). PisCES generally performs the best across ages and layers, suggesting that the dynamic progression is informative.

To assess the similarity of communities across periods we compute the adjusted Rand index (ARI) (27) between subsequent periods. We find that, for the most part, communities in adjacent ages or layers are more similar in composition than those in noncontiguous periods (*SI Appendix*, Table S4); for example, we find that communities in E120 are most similar to those in E90 (ARI = 0.23) and least similar to those in E40 (ARI = 0.05).

Many communities persist in pre- and postnatal samples as illustrated by the Sankey plots (Fig. 5 and *SI Appendix*, Fig. S10). To better understand the Sankey display, we focus on the transition from E40 to E50 and display thresholded weighted adjacency matrices along with the mapping between communities (Fig 6*A*). The ninth community (blue) appears to dominate Fig. 5; however, closer inspection of Fig. 6*A* suggests that this cluster may be dominated by genes that are connected with many genes outside of the cluster, indicating that this cluster is not likely to be interesting biologically.

To facilitate visual comparison of the adjacency matrices, only those nodes belonging to the major flows between E40 and E50 have been included in Fig. 6*B*. Cluster boundaries are delineated in red (featured) or gray. The density of some clusters decreases markedly from E40 to E50 (clusters 2 → 3, 3 → 6, 9 → 10, 9 → 16), while others increase (clusters 5 → 7, 6 → 8, 6 → 9). Genes in dense clusters 7, 8, 10, 11, and 13 go to cluster 9 and lose their tight connectivity as they enter this catch-all cluster, suggesting that some genes act together, but only for brief periods.

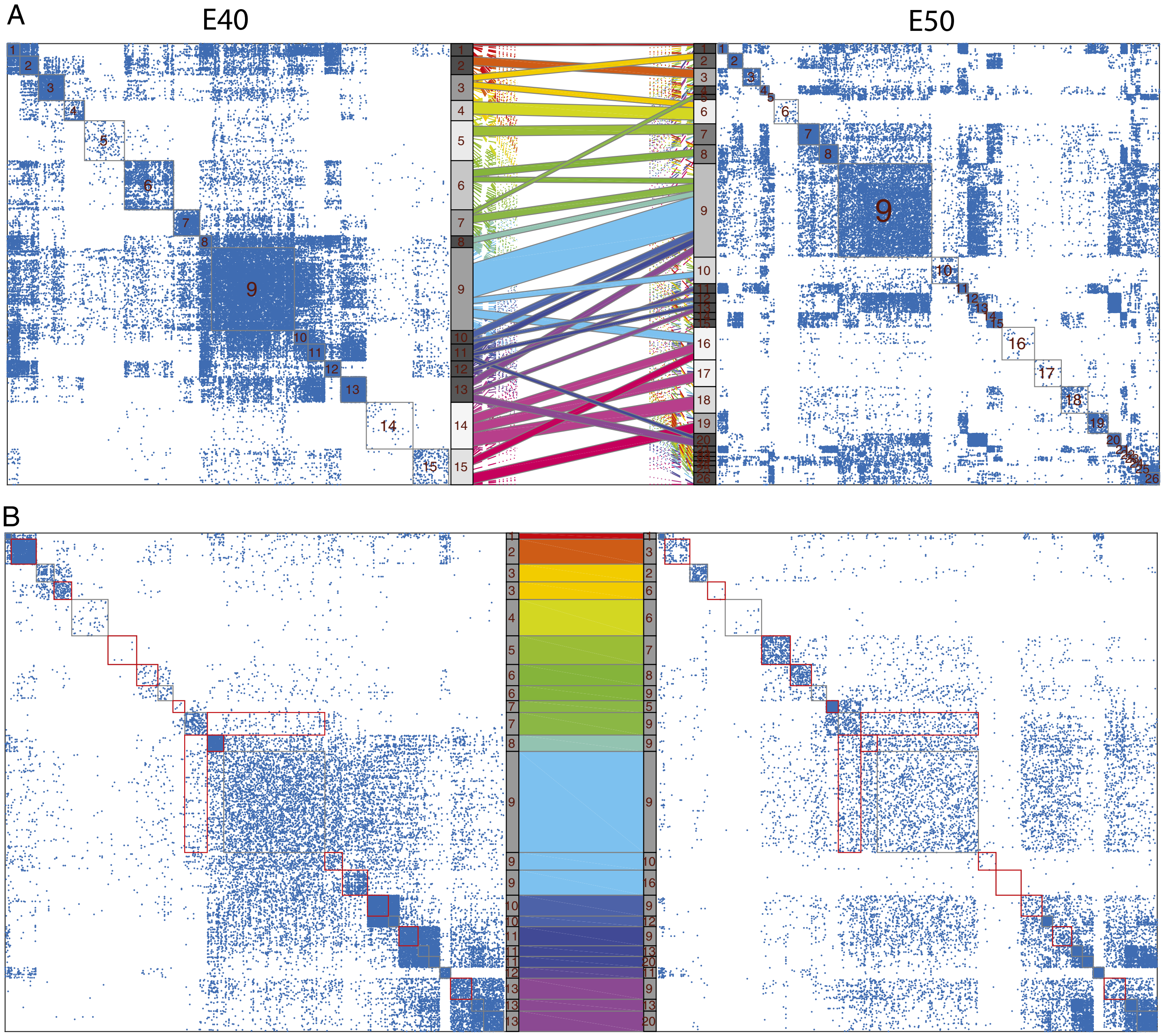

Fig. 6. (*A*) Sankey plot for stages E40 and E50 only, along with the thresholded Laplacian matrices L_{E40} and L_{E50} corresponding to those stages. (*B*) Submatrices of L_{E40} and L_{E50} corresponding to those genes in large flows between E40 and E50 (excluding E40 communities 14 and 15). To facilitate comparison, the genes in L_{E50} have been reordered to match their ordering in L_{E40}. Red boxes highlight large difference between the submatrices. In both *A* and *B*, the flow colors match their coloring in Fig. 5.

For each community, we examine the composition of genes to interpret function using Enrichr (amp.pharm.mssm.edu/Enrichr/) for annotation. Communities are highlighted in red for which neural projection guidance (NPG) genes show significant enrichment (Fig. 5 and *SI Appendix*, Fig. S10). Given the hypothesis that strongly correlated genes share regulation and/or function (30), we examined the communities to determine whether any are enriched with the Simons Foundation Autism Research Initiative (SFARI) autism spectrum disorder (ASD) risk genes (classes 1, 2, 3, and S) (https://gene.sfari.org/autdb/GS_Home.do). Across prenatal communities we observe the strongest clustering of ASD genes in two communities, one at E80 and another at E120 (white dot in red). Both of these communities are enriched for NPG and synaptic transmission (ST) genes and fall in NPG- and ST-enriched paths (*SI Appendix*, Fig. S11). NPG and ST genes have been strongly implicated with ASD (31, 32).

We further investigate the 263 NPG genes that persist in nominally significant enriched communities across at least three consecutive periods (NPG^{+}) to identify the properties of the most highly persistent genes. These 64 NPG^{+} genes show a notable pattern of expression across ages and layers in contrast to the other NPG genes (NPG^{-}). The former set of genes expresses more highly in mPFC at all ages and layers, but especially during the period just before and at birth (*SI Appendix*, Fig. S12) A thresholded-correlation network (Fig. 7) shows how the NPG genes coexpress within three epochs of development: E70–E90, E120–0M, and 3M–12M. Clearly the NPG^{+} genes are more tightly clustered at every epoch and there is a notable pattern to the development of clustered genes over time (*SI Appendix*, Fig. S13). All of these results suggest that a subset of NPG genes plays a strong coordinated role in mid-to-late fetal development in the mPFC, up to and including the time of birth.

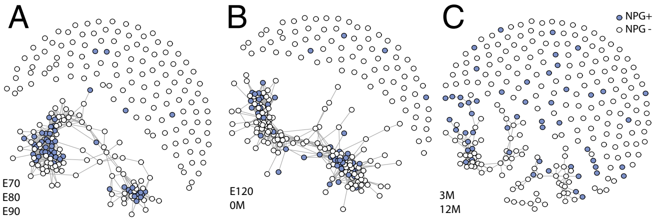

Fig. 7. (*A*–*C*) The correlation networks for all NPG genes in different time bands. NPG genes with pairwise correlations at least 0.7 in the specific time band are connected with edges; nodes for NPG$^+$ genes are filled. Fruchterman–Reingold layout (implemented by igraph's layout_nicely function: igraph.org/r/doc/layout_nicely.html) is applied to illustrate network structures.

ST genes also show strong clustering and persistence in the prenatal analysis; however, unlike NPG genes, the tight clustering of ST genes is not apparent in the postnatal analysis of layers. And yet ST genes become most strongly expressed at birth and continue to be highly expressed throughout life in the mPFC (*SI Appendix*, Fig. S14*A*). This seeming contradiction illustrates a limitation of correlation to accurately capture the relationships between genes under certain conditions. Examining the correlation pattern between two ST genes over two periods of time reveals the problem (*SI Appendix*, Fig. S14*B*). At E120 the genes have considerable variability in expression and show a strong correlation, but at 0M both genes are expressing at their maximum level. When this happens, there is insufficient variability in expression to detect correlation between genes, and hence the correlation is near zero. This suggests that other measures of coexpression will be needed as we continue to investigate gene communities.

Discussion

Community detection, which involves identification of the number of clusters in a network and the membership of each node, is a challenging problem, especially in applications like gene coexpression when the information about the network is uncertain. This paper aims to improve community detection within networks by incorporating available information about the evolution of a network over time. PisCES works by smoothing the signal contained in a series of adjacency matrices, ordered by time or developmental unit, to permit analysis by spectral clustering methods designed for static networks.

Applying PisCES to the medial prefrontal cortex in rhesus monkey brains from near conception to adulthood reveals communities that persist over numerous developmental periods, communities that merge and diverge over time, and others that are loosely organized and ephemeral. PisCES provides a powerful tool to facilitate the discovery of such fine-scale dynamic structures in coexpression data.

Weighted adjacency matrices derived from gene-coexpression data over a number of time frames or developmental periods are ideal for PisCES. These estimates are usually derived from correlation matrices and are often based on a limited number of samples when the spatial–temporal partitions are extremely fine. But, in this situation, dynamic smoothing across partitions can increase the reliability of the resulting communities.

Estimates of community structure provided by PisCES for the rhesus monkeys have highlighted features that comport with known brain development, such as the coordinated expression of NPG and ST genes. This provides a proof of concept for the analysis paradigm. We posit that in-depth study of gene communities over spatial and temporal partitions of the brain will elucidate key developmental periods and communities associated with neurological disorders.

ACKNOWLEDGMENTS. We thank Sivaraman Balakrishnan, Bernie Devlin, Maria Jalbrzikowski, and Lingxue Zhu for insightful comments. We also thank the reviewers whose comments led us to discover *Theorems 1* and *S3.1*. This work was supported by National Institute of Mental Health Grants R37MH057881 and R01MH109900 and the Simons Foundation SFARI 124827.

1. Kolaczyk ED (2009) *Statistical Analysis of Network Data: Methods and Models* (Springer, New York), 1st Ed.
2. Newman M (2010) *Networks: An Introduction* (Oxford Univ Press, New York).
3. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic block models: First steps. *Soc Networks* 5:109–137.
4. Wang YJ, Wong GY (1987) Stochastic blockmodels for directed graphs. *J Am Stat Assoc* 82:8–19.
5. Karrer B, Newman MEJ (2011) Stochastic blockmodels and community structure in networks. *Phys Rev E* 83:016107.
6. Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann Stat* 39:1878–1915.
7. Lei J, Rinaldo A (2015) Consistency of spectral clustering in stochastic block models. *Ann Stat* 43:215–237.
8. Matias C, Miele V (2016) Statistical clustering of temporal networks through a dynamic stochastic block model. *J R Stat Soc Ser B Stat Methodol* 79:1119–1141.
9. Ghasemian A, Zhang P, Clauset A, Moore C, Peel L (2016) Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Phys Rev X* 6:031005.
10. Cribben I, Yu Y (2017) Estimating whole-brain dynamics by using spectral clustering. *J R Stat Soc Ser C Appl Stat* 66:607–627.
11. Nguyen NP, Dinh TN, Shen Y, Thai MT (2014) Dynamic social community detection and its applications. *PLoS One* 9:e91431.
12. Xu KS, Hero AO (2014) Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J Selected Top Signal Process* 8:552–562.
13. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP (2010) Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328:876–878.
14. Bazzi M, et al. (2016) Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Model Simul* 14:1–41.
15. Bassett DS, et al. (2013) Robust detection of dynamic community structure in networks. *Chaos* 23:013142.
16. Taylor D, Myers SA, Clauset A, Porter MA, Mucha PJ (2017) Eigenvector-based centrality measures for temporal networks. *Multiscale Model Simul* 15:537–574.
17. Chi Y, Song X, Zhou D, Hino K, Tseng BL (2007) Evolutionary spectral clustering by incorporating temporal smoothness. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Berkhin P, Caruana R, Wu X (Association for Computing Machinery, New York), pp 153–162.
18. Bakken TE, et al. (2016) A comprehensive transcriptional map of primate brain development. *Nature* 535:367–375.
19. Amini AA, et al. (2013) Pseudo-likelihood methods for community detection in large sparse networks. *Ann Stat* 41:2097–2122.
20. Newman ME (2013) Spectral methods for community detection and graph partitioning. *Phys Rev E* 88:042822.
21. Sussman DL, Tang M, Fishkind DE, Priebe CE (2012) A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J Am Stat Assoc* 107:1119–1128.
22. Sarkar P, et al. (2015) Role of normalization in spectral clustering for stochastic blockmodels. *Ann Stat* 43:962–990.
23. Chen K, Lei J (2017) Network cross-validation for determining the number of communities in network data. *J Am Stat Assoc*, 10.1080/01621459.2016.1246365.
24. Li T, Levina E, Zhu J (2017) Network cross-validation by edge sampling. arXiv:1612.04717v4.
25. Wang YR, et al. (2017) Likelihood-based model selection for stochastic block models. *Ann Stat* 45:500–528.
26. Le CM, Levina E (2015) Estimating the number of communities in networks by spectral methods. arXiv:1507.00827.
27. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218.
28. Meilă M (2012) Local equivalences of distances between clusterings—a geometric perspective. *Mach Learn* 86:369–389.
29. Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.
30. Parikshak N, et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155:1008–1021.
31. Willsey AJ, et al. (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155:997–1007.
32. De Rubeis S, et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515:209–215.