Biometrika (?), **?**, ?, *p.* 1
Printed in Great Britain

Spectral clustering for heterophilic stochastic block models with dynamic node memberships

BY K. Z. LIN
Department of Statistics & Data Science, University of Pennsylvania
kevinl1@wharton.upenn.edu

J. LEI
Department of Statistics & Data Science, Carnegie Mellon University
jinglei@andrew.cmu.edu

SUMMARY

We consider a time-ordered sequence of networks stemming from stochastic block models where nodes are gradually changing memberships over time and no network contains sufficient signal strength to recover its community structure. To estimate the time-varying community structure, we develop KD-SoS (kernel debiased sum-of-square), a method performing spectral clustering after a debiased sum-of-squared aggregation of adjacency matrices. Our theory demonstrates via a novel bias-variance decomposition that KD-SoS achieves consistent community detection of each network even when the networks are heterophilic, and do not require smoothness in the time-varying dynamics of between-community connectivities. We also prove the identifiability of aligning community structures across time based on how rapidly nodes change communities, and develop a data-adaptive bandwidth tuning procedure for KD-SoS. We demonstrate the utility and advantages of KD-SoS through simulations and a novel analysis on the time-varying dynamics in gene coordination in the human developing brain system.

Some key words: Gene co-expression network, human brain development, network analysis, non-parametric analysis, single-cell RNA-seq, time-varying model

1. INTRODUCTION

Longitudinal analyses of a network reveals insights on how communities of nodes lose or create new creations over time. Due to the complexity of most networks, statistical methods are a necessity to uncover these broad dynamics. Simply put, suppose we observe a time-ordered sequence of networks among the same n nodes represented as symmetric binary matrices $A^{(0)}, \ldots, A^{(1)} \in \{0,1\}^{n\times n}$, where for time $t \in [0,1]$, the (i,j)-entry of A denotes the presence or absence of interaction between two nodes. Due to the non-Euclidean nature of the data, it is often difficult to assess if the larger-scale community structures changed over time, and if so, which specific nodes were changing communities at what rate. Sarkar & Moore (2006) developed one of the first methods to investigate these time-varying dynamics. However, research on the statistical properties of such estimators is recent by comparison (Han et al., 2015). See Kim et al. (2018); Pensky & Zhang (2019) for a comprehensive overview. Our goal in this paper is to provide new theoretical advancements in this area by providing a method that is computationally-efficient and can handle a wide range of network dynamics.

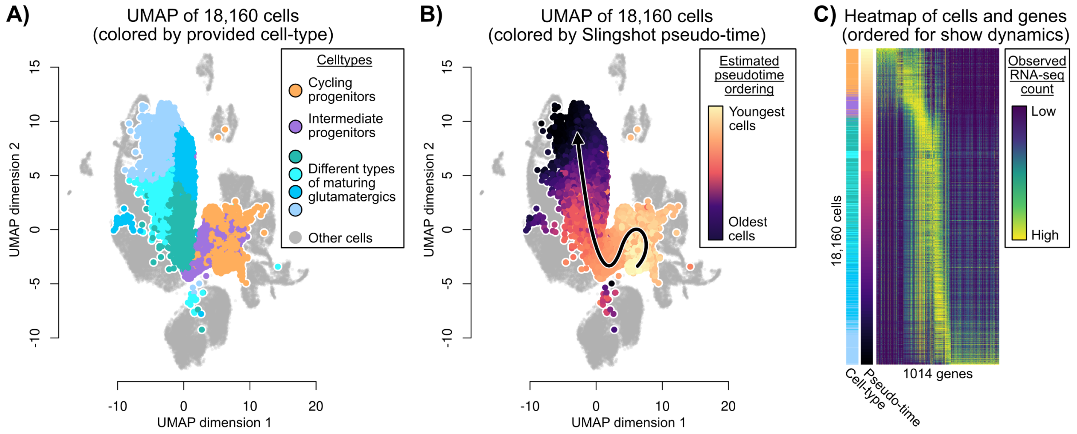

Fig. 1. A) UMAP of the cells among the human developing brain, highlighting the 18,160 cells relevant to our analysis. These denotes cell types such as cycling progenitors (orange) and maturing glutamatergics (shades of teal). B) The 18,160 cells colored based on their estimated pseudotime using Slingshot (Street et al., 2018), colored from youngest (bright yellow) to oldest (dark purple). C) Heatmap ordering the cells based on their estimated pseudotime, and ordering the 1014 relevant genes for this development. The gene expression for each cell is colored based on their expression (high as yellow, low as dark blue).

In this work, we focus on understanding the dynamics of gene coordination over human brain development, but our methods are applicable more broadly to investigate any time-ordered sequence of networks. Consider the single-cell RNA-seq (scRNA-seq) dataset initially published in Trevino et al. (2021), where the authors delineated a specific set of 18,160 cells representing how cycling progenitors (orange) that develop into numerous types of maturing glutamatergics (shades of teal). These cells were annotated by the authors, and the authors also discovered a set 1014 genes that were associated with these cells' development. This data can be visualized through a UMAP (McInnes et al., 2018), a non-linear dimension-reduction method (Figure 1A). Using typical tools in the single-cell analysis toolbox such as Slingshot (Street et al., 2018), we can order the cells in this lineage from the youngest to oldest cells (Figure 1B) and visualize how the gene expression evolves across this lineage (Figure 1C). However, while this simple analysis shows apparent dynamics of the mean gene expression across pseudo-time, the evolution of the gene coordination patterns is unknown. Do the genes that are tightly-coordinated at the beginning of development remain tightly-coordinated at the end of development, and are there tightly-coordinated genes that are not highly expressed?

As reviewed in Kim et al. (2018), there are a plethora of statistical models for time-varying networks. In this work, we focus on time-varying stochastic block models (SBMs). SBMs are a class of prototypical networks that reveal insightful theory while being flexible enough to model many networks in practice. Broadly speaking, a SBM represents each node as part of K (unobserved) communities, and the presence of an edge between two nodes is determined solely by the nodes' community label. Previous work have proven that there is a fundamental limit on how sparse the SBM can be before recovering the communities is impossible (Abbe, 2017). However, when there is a collection of SBMs, this fundamental sparsity can become even smaller. This has led to many different lines of work. For example, one line of work studies the fixed community structure, where T SBMs are observed with all the same community structure (Bhattacharyya & Chatterjee, 2020; Paul & Chen, 2020; Arroyo et al., 2021; Lei & Lin, 2022). Another line of work is when T time-ordered SBMs are observed but there is a changepoint – all the networks before or all the networks after the changepoint share the same community

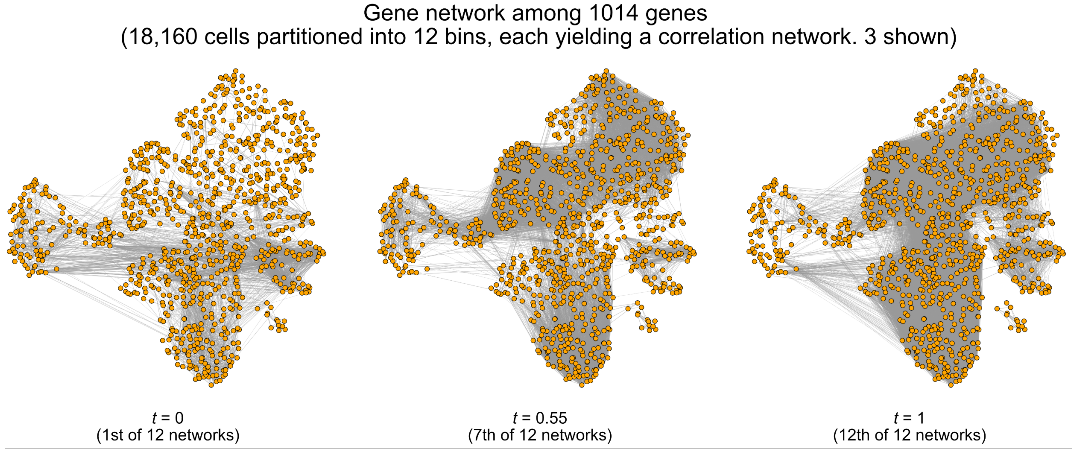
Gene network among 1014 genes
(18,160 cells partitioned into 12 bins, each yielding a correlation network. 3 shown)
t = 0
(1st of 12 networks)
t = 0.55
(7th of 12 networks)
t = 1
(12th of 12 networks)

Fig. 2. Three of 12 networks, for $t = 0$ (i.e., gene network among the youngest cells), $t = 0.55$ and $t = 1$ (i.e., gene network among the oldest cells). These are constructed based on thresholding the correlation matrix among the 1014 genes. The visual position of each gene is fixed for each network, but the edges among the gene varies.

structure (Liu et al., 2018; Wang et al., 2021). One last noteworthy line of work is when no temporal structure is imposed across the T networks, but instead each network slightly deviates from a common community structure at random (Chen et al., 2020).

Despite the abundance of aforementioned SBM models equipped with rigorous theory, they are not quite applicable for our intended analysis of the human developing brain. To provide the reader with a scope of the analysis, we plot the correlation network among the 1014 genes for three different time points in Figure 2. These networks were constructed from 12 non-overlapping partitioning of cells across the estimated time, and we observe potentially gradual changes in community structure across time. See Appendix for more details on the preprocessing. Hence, we turn towards time-varying SBM models where the community structure changes slowly over time. To date, Pensky & Zhang (2019) and Keriven & Vaiter (2022) are among the only works that study this setting. This difficulty is induced by the simple observation that changes in community structure are discrete, which prevent typical non-parameteric techniques from being easily applied. However, as we will discuss later, we take a different theoretical approach to analyze this problem and prove consistent estimation of each network's community under broader assumptions. We briefly note that beyond time-varying SBMs, there are works that studying time-varying latent-position graphs (Gallagher et al., 2021; Athreya et al., 2022). Latent-position graphs are more general than SBMs, as they do not impose a community structure. However, these are not the focus of our work, and also not immediately applicable for understanding the gene coordination dynamics in the developing brain.

The main contribution of this paper is a novel and computationally-efficient method equipped with theoretical guarantees regarding community estimation. Specifically, our method is inspired by Lei & Lin (2022), where a debiased sum-of-squared estimator was proven to consistently estimate of communities for fixed-community multi-layer networks, allowing for both homophilic and heterophilic networks. We adapt this to the time-varying setting by introducing a kernel smoother, and prove through a novel bias-variance decomposition that it can consistently estimate the time-varying communities, holding all other assumptions the same. In particular, while the nodes are gradually changing communities, we impose almost no conditions on how the communities are related within a network or how the community-relations change across networks.

We also formalize the information-theoretic relation between the number of networks and the rate nodes change communities as an identifiability condition.

Our second contribution is a tuning procedure for an appropriate kernel bandwidth that also does not impose restrictions on how the community-relations change across networks. Leave-one-out tuning procedures designed in other matrix applications (Yang & Peng, 2020) where network t is predicted using other temporally-surrounding networks are inappropriate since these procedures require community-relations to change smoothly over time. This also precludes Lepskii-based procedures (Pensky & Zhang, 2019). In contrast, our procedure is designed based on the cosine distance between eigenspaces – for network $t \in \mathcal{T}$, the cosine distance is computed between the eigenspaces of kernel-weighted networks for time less than t and of kernel-weighted networks for time greater than t respectively. The bandwidth that minimizes this distance, averaged over all $t \in \mathcal{T}$, is deemed the most appropriate bandwidth. We show through simulation studies and a thorough investigation of the scRNA-seq data that this procedure selects a desirable bandwidth.

2. DYNAMIC STOCHASTIC BLOCK MODEL

We define the statistical model and assumptions for the dynamic stochastic block model (SBM) throughout this section, which models the evolution of the same set of nodes through time. Let n denote the number of nodes, and $m^{(0)} \in \{1, \ldots, K\}^n$ denote the initial membership vector, where K is a fixed number of communities. That is, $m_i^{(0)} = k$ for $k \in \{1, \ldots, K\}$ if node i starts in community k. We posit that each of the n nodes changes communities according to a $\mathrm{Poisson}(\gamma)$ process with $\gamma > 0$, independent of all other nodes. This means node $i \in \{1, \ldots, n\}$ changes communities at random times $0 < x_{i,1} < x_{i,2} < \ldots < 1$ where the expected difference between consecutive times is $1/\gamma$, and the node changes to one of the $K-1$ other communities with arbitrary probability. This process generates membership vectors $m^{(t)}$ for $t \in [0, 1]$.

Although each node can potentially change communities at multiple times throughout $t \in [0, 1]$, we assume that only T graphs at fixed time points are observed, for

$$\mathcal{T} = \Big\{0,\ \frac{1}{T-1},\ \frac{2}{T-1},\ \ldots,\ 1\Big\}.$$

The generative model for a specific graph $A^{(t)} \in \{0, 1\}^{n \times n}$ for a time $t \in \mathcal{T}$ is as follows. Let $B^{(t)} \in [0, 1]^{K \times K}$ be a symmetric matrix that denotes the connectivity matrix among the K communities for a fixed positive integer K, and let $m^{(t)}$ be the random membership vector based on the above $\mathrm{Poisson}(\gamma)$ process. Each membership vector $m^{(t)}$ can be encoded as one-hot membership matrix $M^{(t)} \in \{0, 1\}^{n \times K}$ where $M_{ik}^{(t)} = 1$ if and only if node i is in community k, and 0 otherwise. Then, the probability matrix $Q^{(t)} \in [0, 1]^{n \times n}$ is defined as

$$Q^{(t)} = \rho_n \cdot M^{(t)} B^{(t)} \big(M^{(t)}\big)^\top, \tag{1}$$

for a network density parameter $\rho_n \in (0, 1)$, and $P^{(t)} = Q^{(t)} - \mathrm{diag}(Q^{(t)})$. The observed graph $A^{(t)}$ for time $t \in \mathcal{T}$ is then sampled according to

$$A_{ij}^{(t)} = \begin{cases} \text{Bernoulli}\left(P_{ij}^{(t)}\right), & \text{if } i > j, \\ 0 & \text{if } i = j, \\ A_{ji}^{(t)} & \text{otherwise.} \end{cases} \tag{2}$$

This implies the following relation:

$$\mathbb{E}\big[A^{(t)}\big] = P^{(t)} = Q^{(t)} - \mathrm{diag}\,\big(Q^{(t)}\big).$$

We define additional to facilitate our theoretical development. Let $n_1^{(t)}, \ldots, n_K^{(t)}$ denote the number of nodes in each community at time t, and $n_{\min}^{(t)} = \min\{n_1^{(t)}, \ldots, n_K^{(t)}\}$. Define $L(M^{(s)}, M^{(t)})$ as the relative Hamming distance between the two membership matrices $M^{(s)}$ and $M^{(t)}$,

$$L(M^{(s)}, M^{(t)}) = \min_{R \in \mathbb{Q}_K} \frac{1}{n} \big\| M^{(s)} R - M^{(t)} \big\|_0, \tag{3}$$

where $\mathbb{Q}_K$ is the set of $K \times K$ permutation matrices. For the rest of the paper, we reserve the letters $i, j \in \{1, \ldots, n\}$ for indices of nodes, and $s, t \in [0, 1]$ for indices of time. We reserve K to denote the fixed number of clusters, and the letter $k, \ell \in \{1, \ldots, K\}$ for indices of cluster.

Our theoretic goal is to show the interplay between the number of nodes n, the number of observed networks T, community-switching rate γ, the network-sparsity parameter ρ_n, and smallest eigenvalue across all T connectivity matrices squared,

$$\lambda_{\min}^2 = \min_{t \in \{1, \ldots, T\}} \lambda_{\min}\big[\big(B^{(t)}\big)^2\big]$$

needed to consistently estimate the T membership matrices across time. Existing theory of single-layer SBMs have already shown that for single network, if $\rho_n \gtrsim \log(n)$, then spectral clustering can asymptotically recover the community structure, while no method can achieve exact recovery if $\rho_n \lesssim \log(n)$ (Bickel & Chen, 2009; Lei & Rinaldo, 2015; Abbe, 2017). Hence, we are interested primarily in the latter setting, where we show that the time-varying community structures can nonetheless be estimated in this challenging setting by aggregating information across networks. Previous methods and theoretical analyses for this setting require strict assumptions on connectivity matrices $\{B^{(t)}\}$ (Pensky & Zhang, 2019; Keriven & Vaiter, 2022) – these matrices are required to smoothly-vary across time and strictly positive eigenvalues, i.e., cannot display patterns of heterophily where edges between communities are more frequent than edges within communities. We seek to develop a method that does not require these assumptions.

3. Debiasing and kernel smoothing

3·1. *Estimator*

Our estimator, the kernelized debiased sum-of-squared (KD-SoS), is motivated by Lei & Lin (2022), where we adopt using de-biased sum of squared adjacency matrices to handle heterophilic networks. We describe our method using the box kernel for simplicity, but the method and theory can be extended to any kernels that are bounded, continuous, symmetric, non-negative and integrate to 1. Provided a bandwidth $r \in [0, 1]$ and number of clusters K, our estimator applies the following procedure for any $t \in \mathcal{T}$. First compute the de-biased sum of squared adjacency matrices, where the summation is over all networks within a bandwidth r,

$$Z^{(t;r)} = \sum_{s \in \mathcal{S}(t;r)} (A^{(s)})^2 - D^{(s)}, \tag{4}$$

where $\mathcal{S}(t;r) = \mathcal{T} \cap [t-r, t+r]$ and $D^{(t)} \in \mathbb{R}^{n\times n}$ is the (random) diagonal matrix encoding the degrees of the n nodes, i.e.,

$$\big[D^{(t)}\big]_{ii} = \sum_{j=1}^{n} A_{ij}^{(t)}, \quad \text{for all } i \in \{1,\ldots,n\}.$$

Second, compute eigen-decomposition of $\widehat{Z}^{(t;r)}$,

$$\widehat{Z}^{(t;r)} = \widehat{U}^{(t;r)}\widehat{\Lambda}^{(t;r)}\big(\widehat{U}^{(t;r)}\big)^\top,$$

where the diagonal entries of $\widehat{\Lambda}^{(t;r)}$ are in descending order, and lastly apply K-means clustering row-wise on the first K columns of $\widehat{U}^{(t;r)}$. This yields the estimated clustering $\widetilde{m}^{(t)} \in \{1,\ldots,K\}^n$. This debiased sum-of-squared estimator is proven in Lei & Lin (2022) to consistently estimate communities under the fixed-community setting, where the squaring of adjacency matrices enable the population connectivity matrices $\{B^{(t)}\}$ to be semidefinite, and the debiasing corrects for the additive noise incurred by this squaring.

After estimating the communities for all T networks, we align the clusters across time. Specifically, initialize $\widehat{m}^{(0)} = \widetilde{m}^{(0)}$, $\delta = 1/(T-1)$ and for $t = 0$, let $C^{(t,t+\delta)}$ be the confusion matrix between $\widehat{m}^{(t)}$ and $\widetilde{m}^{(t+\delta)}$ where

$$C_{k\ell}^{(t,t+\delta)} = \Big|\Big\{i \in \{1,\ldots,n\} : \widehat{m}_i^{(t)} = k \text{ and } \widetilde{m}_i^{(t+\delta)} = \ell\Big\}\Big|. \tag{5}$$

We then solve the following assignment problem,

$$\widehat{R}^{(t+\delta)} = \max_{R\in\mathbb{Q}_K} \big\|\operatorname{diag}(C^{(t,t+\delta)}R)\big\|_1 \tag{6}$$

This can be formulated as an *Hungarian assignment problem*, which can be solved via linear programming. Then, we let the estimated clustering for time t to be $\widehat{m}^{(t)}$ where $\widehat{m}^{(t)} = k$ if and only if $\widetilde{m}^{(t)} = \ell$ and $\widehat{R}_{\ell k}^{(t+\delta)} = 1$. We repeat this procedure for $t \in \mathcal{T}\backslash\{0\}$, and return the final estimated clusterings $\widehat{m}^{(t)}$ for $t \in \mathcal{T}$.

3·2. Bias-variance tradeoff for spectral clustering

We first describe the bias-variance decomposition foundational to our work. Let $\Delta^{(t)} \in \mathbb{R}^{K\times K}$ denote the diagonal matrix where

$$\operatorname{diag}\big(\Delta^{(t)}\big) = \big\{n_1^{(t)},\ldots,n_K^{(t)}\big\}.$$

We then define $\widetilde{M}^{(t)} = M^{(t)}\big(\Delta^{(t)}\big)^{-1/2}$, which is an orthonormal matrix, as well as its projection matrix $\Pi^{(t)} = \widetilde{M}^{(t)}(\widetilde{M}^{(t)})^\top$. Additionally, define the noise matrix $X^{(t)} = P^{(t)} - A^{(t)}$. Observe the following bias-variance decomposition.

LEMMA 1. *Given the model in Section 2, the following deterministic equality holds,*

$$\sum_{s\in\mathcal{S}(t;r)} (A^{(s)})^2 - D^{(s)} = \underbrace{\Big[\sum_{s\in\mathcal{S}(t;r)} (Q^{(s)})^2 - \Pi^{(t)}(Q^{(s)})^2\Pi^{(t)}\Big]}_{I} \tag{7}$$

$$+ \underbrace{\Big[\sum_{s\in\mathcal{S}(t;r)} \big[\mathrm{diag}(Q^{(t)})\big]^2 - Q^{(t)}\,\mathrm{diag}(Q^{(t)}) - \mathrm{diag}(Q^{(t)})Q^{(t)}\Big]}_{II}$$

$$+ \underbrace{\Big[\sum_{s\in\mathcal{S}(t;r)} X^{(s)}P^{(s)} + P^{(s)}X^{(s)}\Big]}_{III} + \underbrace{\Big[\sum_{s\in\mathcal{S}(t;r)} (X^{(s)})^2 - D^{(s)}\Big]}_{IV}$$

$$+ \underbrace{\Big[\sum_{s\in\mathcal{S}(t;r)} \Pi^{(t)}(Q^{(s)})^2\Pi^{(t)}\Big]}_{V}.$$

We deem this decompose as the *bias-variance decomposition* for dynamic SBMs, since term I represents the deterministic bias dictated by nodes changing communities, term II represents the deterministic diagonal bias, term III represents a random error term centered around 0, term IV represents the random variance term, and term V represents the determistic desired matrix with the community information. We note that this decomposition differs from those used in Pensky & Zhang (2019) and Keriven & Vaiter (2022), which instead yield a decomposition that requires smoothness assumptions in $\{B^{(t)}\}$ in order to derive community-consistency.

3·3. Consistency of time-varying communities

Assumption 1 *(Asymptotic regime).* Assume a sequence where n and T are increasing and $n = o(\log(T))$. Additionally, ρ_n and γ can vary with n and T, but there exists a constant c_0 such that $n\rho_n \leq c_0$. Furthermore, assume K is fixed.

We codify the membership dynamics described in Section 2 with the following assumption.

Assumption 2 *(Independent Poisson community changing rate).* Assume for a given community-switching rate $\gamma \geq 0$, each node changes memberships at random times between $t \in [0, 1]$ according to a $\mathrm{Poisson}(\gamma)$ process, independent of all other nodes.

Assumption 3 *(Minimum eigenvalue of aggregated connectivity matrix).* Assume that the sequence $\{B^{(t)}\}$ from $t \in [0, 1]$ is an integrable process across each $(i, j) \in \{1, \ldots, K\}^2$ coordinate. Additionally, assume that for any $t_1, t_2 \in [0, 1]$ for $t_1 \leq t_2$ that

$$\lambda_{\min}\Big(\int_{s=t_1}^{t_2} (B^{(s)})^2 ds\Big) \geq c_1 \cdot (t_2 - t_1)$$

for some constant $c_1 > 0$ independent of n, T, γ, ρ_n and t_1, t_2.

Assumption 4 *(Roughly-equal community sizes).* Assume that across all $t \in [0, 1]$ and all clusters $k \in \{1, \ldots, K\}$, there exists a constant c independent of n, T, λ, ρ_n satisfying $1 \leq c_2$ such that with probability 1 almost surely,

$$n_k^{(t)} \in \Big[\frac{1}{c_2 K} \cdot n,\ \frac{c_2}{K} \cdot n\Big], \quad \text{for all } t \in [0, 1].$$