Electronic Journal of Statistics
Vol. 13 (2019) 678–709
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1533

# Spectral clustering in the dynamic stochastic block model

**Marianna Pensky*** **and Teng Zhang†**

*University of Central Florida,*
*Orlando FL 32816-1353, USA*
*e-mail:* Marianna.Pensky@ucf.edu; Teng.Zhang@ucf.edu

**Abstract:** In the present paper, we have studied a Dynamic Stochastic Block Model (DSBM) under the assumptions that the connection probabilities, as functions of time, are smooth and that at most $s$ nodes can switch their class memberships between two consecutive time points. We estimate the edge probability tensor by a kernel-type procedure and extract the group memberships of the nodes by spectral clustering. The procedure is computationally viable, adaptive to the unknown smoothness of the functional connection probabilities, to the rate $s$ of membership switching, and to the unknown number of clusters. In addition, it is accompanied by non-asymptotic guarantees for the precision of estimation and clustering.

**MSC 2010 subject classifications:** Primary 62F12, 05C80; secondary 62H30.
**Keywords and phrases:** Time-varying network, dynamic stochastic block model, spectral clustering, adaptive estimation.

Received March 2018.

## Contents



*supported by National Science Foundation (NSF), grant DMS-1712977
†supported by National Science Foundation (NSF), grant CNS-1739736

## 1. Introduction

Stochastic networks arise in many areas of research and applications and are used, for example, to study brain connectivity, gene regulatory networks, protein signaling networks, to monitor cyber and homeland security, and to evaluate and predict social relationships within groups or between groups such as countries. Many of those networks evolve in time and therefore require modeling by time-dependent random graphical models. While the literature on statistical modeling of time-independent random graphs is immense (see, e.g., recent surveys [21] and [14]), dynamic stochastic network models are much more recent and less explored. In this paper, we study a dynamic version of the stochastic block model where the probability of a connection between a pair of nodes is determined by a cluster to which the nodes belong, and our main goal is to obtain the time-dependent cluster assignment for each node.

Specifically, we consider a dynamic network defined as an undirected graph with $n$ nodes with connection probabilities changing in time. We observe adjacency matrices $\mathbf{A}_t$ of the graph at time instances $\tau_t$ where $0 < \tau_1 < \cdots < \tau_T = b$. Here, $\mathbf{A}_t(i,j)$ are the Bernoulli random variables with $\mathbf{P}_t(i,j) = \Pr(\mathbf{A}_t(i,j) = 1)$ that are independent for any $1 \le i < j \le n$ and $\mathbf{A}_t(i,j) = \mathbf{A}_t(j,i) = 1$ if a connection between nodes $i$ and $j$ is observed at time $\tau_t$, and $\mathbf{A}_t(i,j) = \mathbf{A}_t(j,i) = 0$ otherwise. For simplicity, we assume that time instances are equi-spaced and the time interval is scaled to one, i.e. $b = 1$ and $\tau_t = t/T$. As we show later (see Remark 1), our method can be modified (in a straightforward manner) to handle the case of non-equal intervals.

Furthermore, we assume that the network can be described by a Dynamic Stochastic Block Model (DSBM): at each time instant $\tau_t$ the nodes are grouped into $K$ classes $G_{t,1}, \cdots, G_{t,K}$, where $K$ is fixed (i.e., independent of $t$) and the probability of a connection $\mathbf{P}_t(i,j)$ is entirely determined by the groups to which the nodes $i$ and $j$ belong at the moment $\tau_t$. In particular, if $i \in G_{t,k}$ and $j \in G_{t,k'}$, then $\mathbf{P}_t(i,j) = \mathbf{B}_t(k,k')$, where $\mathbf{B}_t$ is the *connectivity matrix* at time $\tau_t$ with $\mathbf{B}_t(k,k') = \mathbf{B}_t(k',k)$. In this case, for any $t = 1, \ldots, T$, one has

$$\mathbf{P}_t = \mathbf{\Theta}_t \mathbf{B}_t \mathbf{\Theta}_t^T \tag{1.1}$$

where $\mathbf{\Theta}_t$ is a *clustering* matrix such that $\mathbf{\Theta}_t$ has exactly a single 1 per row, and $\mathbf{\Theta}_t(i,k) = 1$ if and only if node $i$ belongs to the class $G_{t,k}$ and is zero otherwise. One of the main problems in this setting is to cluster the nodes and identify the groups that have common probabilities of connections. If one had an oracle that would give the membership assignments (matrices $\mathbf{\Theta}_t$), then one could obtain accurate estimators of matrices $\mathbf{B}_t$ and $\mathbf{P}_t$ by averaging elements of the adjacency matrices.

The objective of the present paper is to suggest a modification of a popular spectral clustering procedure to the case of the DSBM and study its precision at a time instant $\tau_t$ in a non-asymptotic setting.

The DSBM can be viewed as a natural extension of a Stochastic Block Model (SBM) which was extensively studied in the last decade. Indeed, after After

Olhede and Wolfe [30] showed that SBM provides a kind of a network histogram that can be used for summarizing any network, many authors worked on various problems associated with the SBM such as community detection and clustering (see, e.g., [3], [6], [8], [13], [18], [19], [24], [32], [33], [35], [44], [45] among others) or estimation of the probability matrix (see, e.g., [12] and [20]).

By contrast, there are many fewer results concerning the DSBM model. Although approaches developed for time-independent networks can be applied to a temporal network frame-by-frame, they totally ignore temporal continuity of the network structures and parameters. Nonetheless, by taking advantage of continuity and observations at multiple time instances, one can gain a better insight into a variety of issues and significantly improve precision of the inference (see [16] and [31]).

A survey of the papers published before 2010 can be found in Goldenberg *et al.* [14]. In the last few years, several authors have investigated the SBM in this dynamic setting. The majority of them described changes in the memberships via Markov-type structures that allow modeling of smooth evolution of groups across times. For example, Yang *et al.* [41] assumed that, for each node, its membership forms a Markov chain independent of other nodes; however, the connection probabilities do not change in time. Xu and Hero III [39] allowed both the connection probabilities and the group memberships to change with time via a latent state-space model. Later, Xu [38] and Xu *et al.* [40] further refined the model by introducing a Markov structure on the memberships. In both papers, the logits of connection probabilities are modeled via a dynamical system model. Some authors [17], [41] presented Bayesian variants of similar ideas. We should also mention Fu *et al.* [11] and Xing *et al.* [37] who extended the DSBM to the case of the mixed memberships under the assumption that data follows the multivariate logistic-normal distribution. For example, Xing *et al.* [37] assumed that the data followed the dynamic logistic-normal mixed membership block model and inferred parameter values by using a Laplace variational approximation scheme which generalizes the variational approximation developed in Airoldi *et al.* [1].

None of the papers cited above inferred the number of classes. This shortcoming was corrected by Matias and Miele [28] who propose a Markov chain model for the membership transitions and infer the unknown parameters including the unknown number of classes via variational approximations of the EM algorithm. The approach of [28] was further extended by a very recent paper of Zhang *et al.* [43] who assumed the Poisson model on the number of connections with the time-independent probabilities of edges appearing or disappearing at any time instant. We should also cite an early work of Chi *et al.* [7] that made no assumptions on the mechanism that governs changes in the cluster memberships and deals with a problem by introducing two cost functions: the snapshot cost associated with the error of current clustering, and the temporal cost that measures how the clustering preserves continuity in terms of cluster memberships, where both cost functions are based on the results of the $k$-mean clustering algorithm.

While some of the procedures described in these papers show good computational properties, they come without any guarantees of the accuracy of

estimation and clustering. To the best of our knowledge, the only paper that investigates precision of temporal clustering is [16], where the authors apply the spectral clustering to the matrix of averages under the assumption that the sequences $\mathbf{B}_t(k,k'), t = 1,\dots,T$ form stationary ergodic processes for each $k$ and $k'$, and then prove consistency of the procedure as $T$ and $n$ tend to infinity.

In this paper, we likewise consider a dynamic network that possesses some kind of continuity in a sense that neither connection probabilities $\mathbf{B}_t$ in (1.1), nor class memberships, change drastically from one time instant to another. The setting is motivated by analysis of social networks data where a set of individuals (nodes) can be partitioned into several groups and one can record interactions between the nodes over regular intervals of time (hours, days, weeks). Examples of such data include company e-mails (e.g., ENRON emails), children's interactions during recess, or analysis of Facebook data to name a few. Usually these data are combined over a period of time. We are, however, interested in temporal analysis of such data. For data of this sort it is natural to assume its evolution in time, but temporal changes occurring gradually.

In particular, we assume that, for any pair $k$ and $k'$ of classes, the connection probabilities $\mathbf{B}_t(k,k')$ represent values of some smooth function at time $\tau_t$ and, therefore, can be treated as functional data. In addition, we suppose that at most $s$ nodes can switch their class memberships between two consecutive time points. Both assumptions guarantee some degree of stability of the network in time. Under those assumptions, we extract group memberships of the nodes at every time point by using a spectral clustering procedure and evaluate the error of this procedure. The clustering technique is applied to kernel-type estimators of the edge probability matrices $\mathbf{P}_t$ that we construct in the paper. By using Lepskii's method, we achieve adaptivity of the suggested procedure to the unknown temporal smoothness of the functional connection probabilities $\mathbf{B}_t(k,k')$ and to the rate $s$ of membership switching. Finally, by setting a threshold on the ratio of the eigenvalues of the estimated probability matrix, we find $\hat{K}$, the estimated number of clusters, that coincides with the true number of clusters $K$ with high probability.

Our paper makes several key contributions. We present a computationally viable methodology for estimating an edge probability matrix and clustering of a time-dependent network that follows the DSBM. The procedure is adaptive to the set of unknown parameters, and is accompanied by non-asymptotic guarantees for the precision of estimation and clustering. In order to obtain those results, we develop a variety of new mathematical techniques. In particular, we develop a discrete kernel estimator for an unknown matrix and obtain its adaptive version using Lepskii's method. To the best of our knowledge, neither of these methods have been used in the discrete matrix setting so far. In addition, in Lemma 1, we adapt the methodology used by Lei and Rinaldo [24, Theorem 1.1 in the Supplement], for construction of the upper bound for the spectral norm of a random matrix to derive an upper bound for the spectral norm of a weighted sum of independent random matrices (this methodology was originally introduced in Friedman et al. [10] and Feige and Ofek [9]). Finally, we estimate the number of clusters, and provide guarantees of the accuracy of the estimator.

Our upper bounds for the clustering error are tight. In particular, we show that in the case when the nodes do not switch their memberships in time, under our assumptions, we deliver tighter upper bounds for the error that in the recent paper [5] (see Remark 4).

The rest of the paper is organized as follows. Section 2 introduces notations and the main assumptions of the paper. Section 3 describes the construction of a kernel-type estimator of the probability matrix $\mathbf{P}_t$ at each time point $t$ and evaluates its error. While Sections 3.1 and 3.2 assume that the degree of smoothness $\beta$ of the connection probabilities and the rate $s$ of switching of nodes' memberships are known, Section 3.3 utilizes the Lepskii method for construction of adaptive estimators of the connection probability matrices $\mathbf{P}_t$. Section 4 studies the spectral clustering algorithm and evaluates its error. In particular, Section 4.1 provides an expression for the clustering error at time $t$ in terms of the estimation error of the matrix of the connection probabilities $\mathbf{P}_t$. Furthermore, Section 4.2 presents upper bounds for the clustering errors in terms of the model parameters and discusses when application of the discrete kernel estimator derived in Section 3 improves the clustering accuracy. Section 5 offers an estimator for the number of clusters and provide precision guarantees for the clustering procedure with the estimated number of clusters. Section 6 concludes the paper with a limited simulation study that proves that, over a large variety of model parameters, our technique leads to smaller clustering errors than the two baseline methods which, respectively, cluster the adjacency matrices themselves or their averages. The Appendix (Section 7) describes construction of a discrete kernel and also contains proofs of all statements in the paper.

## 2. Notations and assumptions

For any $a, b \in \mathbb{R}$, denote $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$. For any two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means that there exists a constant $C > 0$ independent of $n$ such that $C^{-1}a_n \le b_n \le Ca_n$ for any $n$. For any set $\Omega$, denote cardinality of $\Omega$ by $|\Omega|$. For any $x$, $\lfloor x \rfloor$ is the largest integer strictly smaller than $x$, $\lceil x \rceil$ is the largest integer no larger than $x$.

For any vector $\mathbf{t} \in \mathbb{R}^p$, denote its $\ell_2$, $\ell_1$, $\ell_0$ and $\ell_\infty$ norms by, respectively, $\|\mathbf{t}\|$, $\|\mathbf{t}\|_1$, $\|\mathbf{t}\|_0$ and $\|\mathbf{t}\|_\infty$. Denote by $\mathbf{1}$ and $\mathbf{0}$ vectors that have, respectively, only unit or zero elements. Denote by $\mathbf{e}_j$ the vector with 1 in the $j$-th position and all other elements equal to zero.

For a matrix $\mathbf{Q}$, its $i$-th row and $j$-th columns are denoted, respectively, by $\mathbf{Q}_{i,*}$ and $\mathbf{Q}_{*,j}$. Similarly, reductions of $\mathbf{Q}$ to a set of rows or columns in a set $G$ are denoted, respectively, by $\mathbf{Q}_{G,*}$ and $\mathbf{Q}_{*,G}$. For any matrix $\mathbf{Q}$, denote its spectral and Frobenius norms by, respectively, $\|\mathbf{Q}\|$ and $\|\mathbf{Q}\|_F$, Denote the largest in absolute value element of $\mathbf{Q}$ by $\|\mathbf{Q}\|_\infty$ and the number of nonzero elements of $\mathbf{Q}$ by $\|\mathbf{Q}\|_0$. Let $\lambda_{\max}(\mathbf{Q})$ and $\lambda_{\min}(\mathbf{Q})$ be the largest and the smallest nonzero eigenvalues of $\mathbf{Q}$.

Denote by $\mathcal{M}_{n,K}$ the collection of clustering matrices $\mathbf{\Theta} \in \{0,1\}^{n\times K}$. Denote by $n_t(k) = |G_{t,k}|$ the number of elements in class $G_{t,k}$ and let $n_{t,\max} =$

$\max_k n_t(k)$ and $n_{t,\min} = \min_k n_t(k)$, $k = 1, \ldots, K$. We assume that there exists $\alpha_n$ independent of $T$ and an absolute constant $C_\alpha$ independent of $n$ and $T$ such that

$$C_\alpha^{-1}\alpha_n \le \|\mathbf{B}_t\|_\infty \le C_\alpha\,\alpha_n, \quad 1 \le C_\alpha < \infty. \tag{2.1}$$

If the network is sparse, then $\alpha_n$ is small for large $n$ and $\|\mathbf{P}_t\|_\infty \le C_\alpha\,\alpha_n$, otherwise, one can just set $\alpha_n = 1$. Denote

$$\mathbf{H}_t = \alpha_n^{-1}\,\mathbf{B}_t, \quad \mathbf{B}_t = \alpha_n\,\mathbf{H}_t.$$

We shall carry out time-dependent clustering of the nodes in the situation where neither the connection probabilities nor the cluster memberships change drastically from one time point to another. In addition, to make successful clustering possible, the values of probabilities of connection should be sufficiently different from each other, which is guaranteed by the smallest eigenvalues of matrices $\mathbf{H}_t$ being separated from zero.

In order to quantify those notions, we consider a Hölder class $\Sigma(\beta, L)$ of functions $f$ on $[0,1]$ such that $f$ are $l_\beta$ times differentiable and

$$|f^{(l_\beta)}(x) - f^{(l_\beta)}(x')| \le L|x - x'|^{\beta - l_\beta} \quad \text{for any} \quad x, x' \in [0,1], \tag{2.2}$$

where $l_\beta = \lfloor \beta \rfloor$ is the largest integer strictly smaller than $\beta$. We suppose that the following assumptions hold.

**(A1).** For any $1 \le k \le k' \le K$, there exist a function $f(\cdot; k, k')$ such that $\mathbf{H}_t(k, k') = f(t/T; k, k')$ and $f(\cdot; k, k') \in \Sigma(\beta, L)$.

**(A2).** At most $s$ nodes can change their memberships between any consecutive time instances.

**(A3).** There exists an absolute constant $C_\lambda$, $1 \le C_\lambda < \infty$, independent of $n$ and $T$ such that

$$C_\lambda^{-1} \le \lambda_{\min}(\mathbf{H}_t) \le \lambda_{\max}(\mathbf{H}_t) \le C_\lambda.$$

Clustering of the nodes can be recovered only up to column permutations. However, in order condition **A1** can hold, we shall assume that the node's labels are fixed and do not depend on $t$. We denote the set of $K \times K$ permutation matrices by $\mathcal{E}_K$ and, following [24], consider two measures of clustering precision at time $\tau_t$. The first is the overall relative clustering error at time $\tau_t$

$$R_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) = n^{-1} \min_{\mathbf{J} \in \mathcal{E}_K} \|\widehat{\mathbf{\Theta}}_t\mathbf{J} - \mathbf{\Theta}_t\|_0 \tag{2.3}$$

that measures the overall proportion of mis-clustered nodes. The second measure is the highest relative clustering error over the communities at time $\tau_t$

$$\tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) = \min_{\mathbf{J} \in \mathcal{E}_K} \max_{1 \le k \le K} n_{t,k}^{-1}\, \|(\widehat{\mathbf{\Theta}}_t\mathbf{J} - \mathbf{\Theta}_t)_{G_{t,k},*}\|_0. \tag{2.4}$$

In addition, we study two global measures of clustering accuracy such as the overall highest relative error over the communities and the overall highest relative error

$$\tilde{R}_{\max} = \max_{1 \le t \le T} \tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t), \quad R_{\max} = \max_{1 \le t \le T} R_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \tag{2.5}$$

In order to take advantage of the temporal continuity of the network, we construct estimators of the matrices $\mathbf{P}_t$ at every point $t \in \{1, \cdots, T\}$. It follows from [24] that the clustering errors depend on the error of estimation of matrices $\mathbf{P}_t$ in operational norm.

## 3. Estimation of the edge probability matrices

### 3.1. Construction of the estimators

In order to estimate $\mathbf{P}_t$, we choose an integer $r \geq 0$, the width of the window, and consider three pairs of sets of integers

$$\begin{aligned} \mathcal{F}_{r,1} &= \{-r, \cdots, r\}, \quad \mathcal{D}_{r,1} = \{1+r, \cdots, T-r\}; \\ \mathcal{F}_{r,2} &= \{0, \cdots, r\}, \quad \mathcal{D}_{r,2} = \{1, \cdots, r\}; \\ \mathcal{F}_{r,3} &= \{-r, \cdots, 0\}, \quad \mathcal{D}_{r,3} = \{T-r+1, \cdots, T\}. \end{aligned}$$

If $r = 0$, then $\mathcal{D}_{0,2}$ and $\mathcal{D}_{0,3}$ are just empty sets. If $t \in \mathcal{D}_{r,j}$, we construct an estimator of $\mathbf{P}_t$ on the basis of $\mathbf{A}_{t+i}$ where $i \in \mathcal{F}_{r,j}$, $j = 1, 2, 3$. For this purpose, we introduce discrete kernel functions $W_{r,l}^{(j)}(i)$ of an integer argument $i$ that satisfy the following assumption:

**(A4).** Functions $W_{r,l}^{(j)}$, $j = 1, 2, 3$, are such that $|W_{r,l}^{(j)}(i)| \leq W_{\max}$, where $W_{\max}$ is independent of $r$, $j$ and $i$, and for $j = 1, 2, 3$, one has

$$\frac{1}{|\mathcal{F}_{r,j}|} \sum_{i \in \mathcal{F}_{r,j}} i^k \, W_{r,l}^{(j)}(i) = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } k = 1, \ldots, l. \end{cases} \tag{3.1}$$

Here $|\mathcal{F}_{r,j}|$ is the cardinality of the set $\mathcal{F}_{r,j}$.

One can easily see that function $W_{r,l}^{(j)}$ are discrete versions of order $l$ continuous kernels, where $W_{r,l}^{(1)}$ corresponds to a regular kernel designed for the internal points of the interval while $W_{r,l}^{(j)}$, $j = 2, 3$, mimic the boundary kernels (the left boundary kernel for $j = 2$ and the right boundary kernel for $j = 3$). Section 7.1 provides an algorithm for the explicit construction of $W_{r,l}^{(j)}$ for any values of $r, l$ and $j$. We ought to point out that the dependence of $W_{r,l}^{(j)}$ on $r$ is a weak one, especially as $r$ grows. We estimate the edge probability matrix $\mathbf{P}_t$ by

$$\widehat{\mathbf{P}}_{t,r} = \sum_{j=1}^{3} \mathbb{I}(t \in \mathcal{D}_{r,j}) \left\{ \frac{1}{|\mathcal{F}_{r,j}|} \sum_{i \in \mathcal{F}_{r,j}} W_{r,l}^{(j)}(i) \mathbf{A}_{t+i} \right\}. \tag{3.2}$$

Note that since the sets $\mathcal{D}_{r,j}$ are disjoint for different values of $j$, the estimator of $\mathbf{P}_t$ always involve just one expression in figure brackets in formula (3.2).

**Remark 1.** Note that our method can be modified to handle the case of non-equal intervals. Indeed, in this case, one needs to modify condition (3.1) by introducing weights that account for the differences in the values of $\Delta_{\tau,i} = \tau_{i+1} -$

$\tau_i$, $i = 0, \dots, T-1$, the same manner as it is done in non-equispaced regression estimation. However, while this modification adds very little conceptually, it would make the paper very technical and hard to follow.

### *3.2. Estimation error*

In order to figure out how to choose the value of $r$, we evaluate the error $\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_t\|$. Denote

$$\mathbf{P}_{t,r} = \sum_{j=1}^{3} \mathbb{I}(t \in \mathcal{D}_{r,j}) \left\{ \frac{1}{|\mathcal{F}_{r,j}|} \sum_{i \in \mathcal{F}_{r,j}} W_{r,l}^{(j)}(i) \mathbf{P}_{t+i} \right\}$$

and observe that

$$\Delta_t(r) = \|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_t\| \le \|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\| + \|\mathbf{P}_{t,r} - \mathbf{P}_t\| \equiv \Delta_{1,t}(r) + \Delta_{2,t}(r), \quad (3.3)$$

where $\Delta_{1,t}(r) = \|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\|$ and $\Delta_{2,t}(r) = \|\mathbf{P}_{t,r} - \mathbf{P}_t\|$ represent, respectively, the variance and the bias portions of the error. The following statements provide upper bounds for those quantities.

**Lemma 1.** *Let (2.1) be valid and Assumption* **A4** *hold with $l \ge l_\beta = \lfloor \beta \rfloor$. If $\alpha_n \ge C_\alpha^{-1} c_0 \log n / n$, then, for any $\tau > 0$ there exists a set $\Omega_{t,\tau}$ and a constant $C_{0,\tau} = C(\tau, c_0, C_\alpha, W_{\max})$ such that $\Pr(\Omega_{t,\tau}) \ge 1 - 4\, n^{-\tau}$ and, for any $\omega \in \Omega_{t,\tau}$, one has*

$$\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\| \le C_{0,\tau} \sqrt{n\, \alpha_n / (r \vee 1)}. \quad (3.4)$$

*The exact expression for $C_{0,\tau}$ is given by formula (7.16) in the Appendix.*

The proof of Lemma 1 generalizes the methodology used by Lei and Rinaldo [24, Theorem 1.1 of the Supplement], from derivation of an upper bound for the spectral norm of $\mathbf{P}_t - \mathbf{A}_t$ (i.e., $\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\|$ with $r = 0$) to the upper bound of the spectral norm of $\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\|$, the weighted sum of independent matrices $\mathbf{P}_i - \mathbf{A}_i$ with $i \in \mathcal{F}_{r,j}$. While one can use a simpler method applying [24, Theorem 1.1 of the Supplement] to each $\mathbf{P}_i - \mathbf{A}_i$ and then combining these upper bounds together by a matrix concentration inequality such as the matrix Bernstein inequality [34], the result of this more straightforward technique is looser by a logarithmic factor $\log n$.

**Lemma 2.** *Let $l \ge l_\beta = \lfloor \beta \rfloor$. Then, under Assumptions* **A1**–**A4** *and (2.1), one has*

$$\|\mathbf{P}_{t,r} - \mathbf{P}_t\| \le \frac{L}{(l_\beta)!} W_{\max}\, \alpha_n n \left(\frac{r}{T}\right)^{\beta} + 2\sqrt{2}\, W_{\max}\, C_\lambda\, \alpha_n \sqrt{n_{\max}\, r s}. \quad (3.5)$$

Lemmas 1 and 2 together with inequality (3.3) provide an upper bound for $\Delta_t(r)$. Since $\Delta_{1,t}(r)$ is decreasing and $\Delta_{2,t}(r)$ is increasing in $r$, there exist a value $r^*$ that ensures the best bias-variance balance. Denote

$$r^* = \operatorname*{argmin}_r \left( \|\mathbf{P}_{t,r} - \mathbf{P}_t\| + C_{0,\tau} \sqrt{n\, \alpha_n / (r \vee 1)} \right), \quad (3.6)$$

$$\delta_1 = \sqrt{n\,\alpha_n}, \quad \delta_2 = \left(\frac{\alpha_n^{\beta+1} n^{\beta+1}}{T^\beta}\right)^{\frac{1}{2\beta+1}} + (\alpha_n^3\, n\, n_{\max}\, s)^{\frac{1}{4}}, \tag{3.7}$$

Then, the following lemma yields an upper bound for $\Delta_t(r)$.

**Lemma 3.** *Let (2.1) be valid and Assumptions* **A1**–**A4** *hold with* $l \geq l_\beta = \lfloor \beta \rfloor$ *and* $\alpha_n \geq C_\alpha^{-1}\, c_0\, \log n/n$. *Then, the optimal value of* $r$ *is*

$$r^* \leq \min\left(\left\lceil C_T\, \left(n^{-1}\alpha_n T^{2\beta}\right)^{1/(2\beta+1)}\right\rceil,\ \left\lceil C_s\, \sqrt{(\alpha_n n_{\max}\, s)^{-1}\, n}\right\rceil\right) \tag{3.8}$$

*where* $\lceil x \rceil$ *is the largest integer no greater than* $x$ *and* $C_T$ *and* $C_s$ *are positive constants independent of* $n$, $T$, $n_{\max}$, $s$ *and* $\alpha_n$. *Also, with probability at least* $1 - 4\, n^{-\tau}$ *one has*

$$\Delta_t(r^*) \leq C_\Delta\ \min(\delta_1, \delta_2), \tag{3.9}$$

*where constant* $C_\Delta$ *depends on* $\tau, c_0, W_{\max}, \beta, L, C_\alpha$ *and* $C_\lambda$.

Recall that $r^*$ is the value that ensures the best bias-variance balance in the right-hand side of (3.6). Since we have only an upper bound for the bias, we obtain an upper bound for the optimal value in (3.6), and hence, we have an inequality in (3.8). Furthermore, we have two possible scenarios in (3.6): $r^* = 0$ and $r^* \geq 1$. We obtain the values of $\Delta_t(r^*)$ by plugging the upper bound for $r^*$ into the expression (3.4) for the variance. If $r^* = 0$, then the right-hand side of (3.6) reduces to $C_{0,\tau}\, \sqrt{n\,\alpha_n} \propto \delta_1$. If $r^* \geq 1$, then we choose the value of $r$ that minimizes the sum of the upper bound for the bias (3.5) and the variance $C_{0,\tau}\, \sqrt{n\,\alpha_n\, r^{-1}}$. Since the bias in (3.5) consists of two terms, we obtain respective two terms in the expression (3.6) for $r^*$. The estimation error in this case is $\Delta_t(r^*) \leq C_\Delta\, \delta_2$.

Note that $\delta_1 < \delta_2$ corresponds to the case where $r^* = 0$ and this situation occurs only if $T$ is rather small or $s$ is large. Otherwise, $r^* \geq 1$ and one can take an advantage of the smoothness of the connection probabilities and the relative stability of group memberships.

### 3.3. Adaptive estimation

Observe that the value of $r^*$ depends on the values of $s$, $n_{\max}$, $\alpha_n$ and $\beta$ that are unknown, therefore, in practice, the value $r^*$ in (3.8) is unavailable. In order to construct an adaptive estimator we use the Lepskii method [25], [26]. For any $t$, set $\widehat{r} \equiv \widehat{r}_t$ where

$$\widehat{r}_t = \max\left\{0 \leq r \leq T/2 :\ \|\widehat{\mathbf{P}}_{t,r} - \widehat{\mathbf{P}}_{t,\rho}\| \leq 4\, C_{0,\tau}\, \sqrt{\frac{n\,\alpha_n}{\rho \vee 1}},\ \ \forall \rho < r\right\} \tag{3.10}$$

Observe that evaluation of $\widehat{r}$ does not require the knowledge of $s$, $n_{\max}$ or $\beta$. If the network is not sparse, one can set $\alpha_n = 1$. Otherwise, one needs to know $\alpha_n$ for choosing an optimal value of $r$. If $\alpha_n$ is known, the following lemma ensures that the replacement of $r^*$ by $\widehat{r}$ changes the upper bound by a constant factor only.

**Lemma 4.** *Let (2.1) be valid and Assumptions* **A1**–**A4** *hold with* $l \geq l_\beta = \lfloor \beta \rfloor$ *and* $\alpha_n \geq C_\alpha^{-1} c_0 \log n/n$. *Then, for any* $\tau > 0$, *with probability at least* $1-4n^{-\tau}$, *one has*

$$\|\widehat{\mathbf{P}}_{t,\widehat{r}} - \mathbf{P}_t\| \leq 10 \min_r \left\{ \|\mathbf{P}_{t,r} - \mathbf{P}_t\| + C_{0,\tau} \sqrt{n\,\alpha_n/(r \vee 1)} \right\} \leq 10\,\Delta_t(r^*). \quad (3.11)$$

Lemma 4 implies that the error of the adaptive procedure lies within a constant multiple of the optimal error $\Delta_t(r^*)$. The idea of the proof is based on the fact that if $\widehat{r} < r^*$, then $\|\widehat{\mathbf{P}}_{t,\widehat{r}} - \mathbf{P}_{t,r^*}\| \leq C_{0,\tau} \sqrt{n\,\alpha_n/(r^* \vee 1)}$, which guarantees (3.11). On the other hand, due to Lemma 1, the probability of the opposite inequality is very low. See the proof of Lemma 4 for details.

## 4. Spectral clustering and its error

### *4.1. Spectral clustering algorithm*

Spectral clustering is a common method for community recoveries (see, e.g., [29], [18], [19], [24], [27], [32] and [33] among others). The accuracy of spectral clustering depends on how well one can relate the eigenvectors of $\mathbf{P}_t = \mathbf{\Theta}_t \mathbf{B}_t \mathbf{\Theta}_t^T$ to the eigenvectors of its estimator $\widehat{\mathbf{P}}_t$. For this reason, our first goal will be to obtain an estimator $\widehat{\mathbf{P}}_t$ of $\mathbf{P}_t$. Subsequently we shall apply the spectral clustering based on the approximate $k$-means algorithm suggested by Lei and Rinaldo [24]. Although one can read a description of the algorithm in their paper, for completeness we review it here.

Given a matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$, let $\mathbf{U} \in \mathbb{R}^{n \times K}$ be the matrix that consists of the first $K$ eigenvectors of $\mathbf{P}$. Then [24] suggested to investigate the $(1+\epsilon)-$approximate solution to the $k$-means problem applied to the $n$ row vectors of $\mathbf{U}$, specifically, finding $\hat{\mathbf{\Theta}} \in \mathcal{M}_{n,K}$ and $\hat{\mathbf{X}} \in \mathbb{R}^{K \times K}$ that satisfy

$$\|\hat{\mathbf{\Theta}}\hat{\mathbf{X}} - \mathbf{U}\|_F^2 \leq (1+\epsilon) \min_{\substack{\mathbf{\Theta} \in \mathcal{M}_{n,K} \\ \mathbf{X} \in \mathbb{R}^{K \times K}}} \|\mathbf{\Theta}\mathbf{X} - \mathbf{U}\|_F^2. \quad (4.1)$$

Then the cluster assignments are given by the estimated $\hat{\mathbf{\Theta}}_t$. There exist efficient algorithms for solving (4.1), see, e.g., [22]. The procedure is summarized as Algorithm 1.

**Algorithm 1** Spectral clustering in the dynamic stochastic block model

**Input:** Adjacency matrices $\mathbf{A}_t$, $t = 1, \ldots, T$; number of communities $K$; approximation parameter $\epsilon$.
**Output:** Estimators of the membership matrices $\widehat{\mathbf{\Theta}}_t$ for any $t = 1, \ldots, T$.
**Steps:**
**1:** Estimate $\mathbf{P}_t$ by $\widehat{\mathbf{P}}_{t,r}$ defined in (3.2).
**2:** Let $\widehat{\mathbf{U}}_t \in \mathbb{R}^{n \times K}$ be a matrix representing the first $K$ eigenvectors of $\widehat{\mathbf{P}}_{t,r}$.
**3:** Apply the $(1+\epsilon)$-approximate $k$-means algorithm to the row vectors of $\widehat{\mathbf{U}}_t$
**4:** Obtain the solution $\widehat{\mathbf{\Theta}}_t$.

The difference between Algorithm 1 and the one suggested in [24] is that spectral clustering is applied not to the adjacency matrix directly but to the estimators of the probability matrices $\widehat{\mathbf{P}}_{t,r}$.

The clustering errors $R_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t)$ in (2.3) and $\tilde{R}_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t)$ in (2.4) are determined by the precision of estimation of $\mathbf{P}_t$ by $\widehat{\mathbf{P}}_t$. In particular, the following statement (which is a straightforward modification of arguments in [24]) holds.

**Lemma 5.** *Let clustering be carried out according to the Algorithm 1 on the basis of an estimator $\widehat{\mathbf{P}}_{t,r}$ of $\mathbf{P}_t$. Let $\boldsymbol{\Theta}_t \in \mathcal{M}_{n,K}$. Then,*

$$\tilde{R}_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \le \frac{64(2+\epsilon)K}{\lambda^2_{\min}(\mathbf{P}_t)} \, \|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_t\|^2 \tag{4.2}$$

*and*

$$R_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \le \frac{64(2+\epsilon)K}{\lambda^2_{\min}(\mathbf{P}_t)} \, \frac{n_{t,\max}}{n} \, \|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_t\|^2. \tag{4.3}$$

*Here, $\lambda_{\min}(\mathbf{P}_t)$ is the smallest nonzero eigenvalue of $\mathbf{P}_t$.*

**Remark 2. The value of $\boldsymbol{\alpha_n}$.** Observe that the only unknown parameter in our algorithm is the value of $\alpha_n$. In general, $\alpha_n$ is difficult to estimate. One of the possible approaches may be to construct an initial estimator of $\alpha_n$ as the maximum of the infinity norm of the moving average estimator

$$\tilde{\alpha}_n = \frac{1}{2r+1} \, \max_t \left\| \sum_{i=-r}^{r} \mathbf{A}_{t+i} \right\|$$

and then validate it by finding $\widehat{\alpha}_n = \max_t \, \|\widehat{\mathbf{B}}_t\|_\infty$ after the clustering procedure is completed and changing the value of $r$ if necessary.

### *4.2. The clustering error*

Lemmas 4 and 5 allow to obtain upper bounds for the clustering errors.

**Theorem 1.** *Let clustering be carried out according to the Algorithm 1. Let $\mathbf{P}_t = \boldsymbol{\Theta}_t \mathbf{B}_t \boldsymbol{\Theta}_t^T$ where $\mathbf{B}_t = \alpha_n \mathbf{H}_t$. If (2.1) and Assumptions* **A1**–**A4** *hold with $\alpha_n \ge C_\alpha^{-1} \, c_0 \, \log n / n$, then for any $\tau > 0$, with probability at least $1 - 4\, n^{-\tau}$, one has*

$$\tilde{R}_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \;\le\; C_R(2+\epsilon) \, \frac{K \, \min(\delta_1^2, \delta_2^2)}{\alpha_n^2 \, n_{\min}^2} \tag{4.4}$$

$$R_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \;\le\; \tilde{R}_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \, \frac{n_{\max}}{n}, \tag{4.5}$$

*where $\delta_1$ and $\delta_2$ are defined in (3.7) and $C_R = C_R(\tau, c_0, W_{\max}, \beta, L, C_\alpha, C_\lambda)$.*

*In addition, if $T$ grows at most polynomialy with $n$, so that $T \le n^{\tau_1}$ for some $\tau_1 < \infty$, then for any $\tau > 0$, with probability at least $1 - 4\, n^{-(\tau - \tau_1)}$, one has*

$$\tilde{R}_{\max} \le C_R \, (2+\epsilon) \, \frac{K \, \min(\delta_1^2, \delta_2^2)}{\alpha_n^2 \, n_{\min}^2}, \quad R_{\max} \le \tilde{R}_{\max} \, \frac{n_{\max}}{n}. \tag{4.6}$$

Theorem 1 provides upper bounds for the local clustering errors at time $\tau_t$ as well as for the maximum clustering errors on the whole time interval. In order to assess when employing the kernel estimator is beneficial, recall that the latter happens when $r^* \geq 1$ in (3.8) and $\delta_2 < \delta_1$ in (4.4)–(4.6). In particular, it follows from (4.4)–(4.6) that the ratio $\Delta R$ of the clustering errors when $\mathbf{P}_t$ is estimated by the kernel estimator (3.2) with $r = r^* > 1$ and $\mathbf{P}_t$ is estimated by $\mathbf{A}_t$ (which corresponds to the direct application of Lei and Rinaldo's procedure [24]) is

$$\Delta R \asymp \min\left(1; \frac{\delta_2^2}{\delta_1^2}\right) = \min\left\{1; \left(\frac{n\,\alpha_n}{T^{2\beta}}\right)^{\frac{1}{2\beta+1}} + \sqrt{\frac{n_{\max}\,\alpha_n s}{n}}\right\}. \tag{4.7}$$

Hence, (4.7) yields that application of the kernel estimator is advantageous if $\Delta R < 1$, which is equivalent to

$$T \geq (n/\alpha_n)^{\frac{1}{2\beta}} \quad \text{and} \quad s \leq (n_{\max}\alpha_n)^{-1}\, n. \tag{4.8}$$

Therefore, as long as (4.8) holds, the clustering errors obtained by our algorithm will be smaller than those obtained by separately clustering snapshots of the network at each individual time point as in [24]. This is true for the clustering errors at every time instance $\tau_t$ as well as overall. Specifically, one can formulate the following corollary.

**Corollary 1.** *Under the assumptions of Theorem 1, for any $\tau > 0$, with probability at least $1 - 4\,n^{-\tau}$, one has*

$$\tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \leq C_R(2+\epsilon)\,\frac{Kn}{\alpha_n\, n_{\min}^2}\,\Delta R, \tag{4.9}$$

*where $C_R = C_R(\tau, c_0, W_{\max}, \beta, L, C_\alpha, C_\lambda)$ and $\Delta R$ is defined in (4.7). Moreover, (4.5) holds provided the right hand side of (4.9) is bounded by one. In addition, if $T \leq n^{\tau_1}$ for some $\tau_1 < \infty$, then, with probability at least $1 - 4\,n^{-(\tau-\tau_1)}$, one has*

$$\tilde{R}_{\max} \leq \tilde{C}_R(2+\epsilon)\,\frac{Kn}{\alpha_n\, n_{\min}^2}\,\Delta R.$$

If the community sizes are balanced, i.e. there exist positive constants $C_1$ and $C_2$ such that

$$C_1\,\frac{n}{K} \leq n_{\min} \leq n_{\max} \leq C_2\,\frac{n}{K}, \tag{4.10}$$

one can obtain more transparent upper bounds for the clustering errors.

**Corollary 2.** *If the assumptions of Theorem 1 and condition (4.10) hold, then, for any $\tau > 0$, with probability at least $1 - 4\,n^{-\tau}$, one has*

$$\tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \leq \tilde{C}_R(2+\epsilon)\,\frac{K^3}{n\alpha_n}\,\Delta R. \tag{4.11}$$

*where $\tilde{C}_R = \tilde{C}_R(\tau, c_0, W_{\max}, \beta, L, C_\alpha, C_\lambda)$ and $\Delta R$ is defined in (4.7). Moreover, (4.5) holds provided the right hand side of (4.11) is bounded by one. In addition,*

*if* $T \le n^{\tau_1}$ *for some* $\tau_1 < \infty$*, then, with probability at least* $1 - 4\,n^{-(\tau-\tau_1)}$*, one has*

$$\tilde{R}_{\max} \le \tilde{C}_R(2+\epsilon)\,\frac{K^3}{n\alpha_n}\,\Delta R.$$

**Remark 3. Dense network.** Inequalities (4.4), (4.5), (4.11) and (4.6) imply that precision of clustering is better when $\alpha_n$ is larger. Indeed, if the network is dense, then $\alpha_n = 1$, the estimator $\widehat{\mathbf{P}}_{t,\widehat{r}}$ is fully adaptive and with probability at least $1 - 4\,n^{-(\tau-\tau_1)}$,

$$\tilde{R}_{\max} \le C_R\,(2+\epsilon)\,\frac{K\,n}{n_{\min}^2}\,\min\left(1; \left(\frac{n}{T^{2\beta}}\right)^{\frac{1}{2\beta+1}} + \sqrt{\frac{n_{\max}\,s}{n}}\right).$$

**Remark 4. Constant memberships and comparison with [5].** If group memberships of the nodes remain unchanged over time, then $s = 0$ and one can cluster the average $\overline{\mathbf{P}}$ of edge probability matrices on the basis of its observed counterpart $\widehat{\mathbf{P}}$ where

$$\overline{\mathbf{P}} = T^{-1}\,\sum_{t=1}^{T}\mathbf{P}_t, \quad \widehat{\mathbf{P}} = T^{-1}\,\sum_{t=1}^{T}\mathbf{A}_t$$

rather than the individual matrices $\mathbf{P}_t$. In this case, $2r+1 = T$, $W_{\max} = 1$ and the bias portion of the error disappears, hence,

$$\|\widehat{\mathbf{P}} - \overline{\mathbf{P}}\| \le C_{0,\tau}\sqrt{n\alpha_n/T}. \tag{4.12}$$

Observe that $\lambda_{\min}(\widehat{\mathbf{P}}) \ge C_\lambda^{-1}\,\alpha_n\,n_{\min}$. Therefore, for any $\tau > 0$, with probability at least $1 - 4\,n^{-\tau}$, one has

$$\begin{aligned} \tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \quad &\equiv \quad \tilde{R}(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}) \le 64(2+\epsilon)\,C_{0\tau}^2\,C_\lambda^2\,\frac{K\,n}{T\,\alpha_n n_{\min}^2}, \\ R(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}) \quad &\le \quad \tilde{R}(\widehat{\mathbf{\Theta}}, \mathbf{\Theta})\,\frac{n_{\max}}{n}. \end{aligned} \tag{4.13}$$

Recently, the case of constant memberships was considered by Bhattacharyya and Chatterjee [5] in the context of multiple networks. Below, we compare our clustering error with the clustering error in [5] under Assumption **(A3)**. Under this assumption, Bhattacharyya and Chatterjee [5] define $f_k$, $k = 1, \cdots, K$, the proportion of misclassified nodes in community $k$, $k = 1, \cdots, K$, and provide an upper bound for their sum. In the case of the balanced model satisfying (4.10), if $K^5 \le CTn\alpha_n$ and $Tn\alpha_n \to \infty$, Theorem 3.1 in [5] yields the following upper bound (with a high probability)

$$\sum_{k=1}^{K} f_k \le CK^3\,(Tn\alpha_n)^{-1/2}. \tag{4.14}$$

In comparison, inequality (4.13) implies that

$$\sum_{k=1}^{K} f_k \leq K\,\tilde{R}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \leq CKK^3(Tn\alpha_n)^{-1} \leq CK^3(Tn\alpha_n)^{-4/5}$$

which is a tighter upper bound than (4.14). However, the upper bounds are obtained under different assumptions.

**Remark 5. Constant matrix of connection probabilities.** Consider the situation where nodes of the network can switch memberships in time ($s > 0$) but the matrix of the connection probabilities is constant: $\mathbf{B}_t \equiv \mathbf{B}$. In this case, Assumption **A1** is valid with $\beta = \infty$ and one can choose $C_\alpha = 1$ in (2.1). Then, for $r < T$ one has $\|\mathbf{P}_{t,r} - \mathbf{P}_t\| \leq 2\sqrt{2}\, W_{\max}\, C_\lambda\, \alpha_n \sqrt{n_{\max}\, rs}$. Hence, $\delta_2 = (\alpha_n^3\, n\, n_{\max}\, s)^{\frac{1}{4}}$ and for any $\tau > 0$, with probability at least $1 - 4\,n^{-\tau}$, the clustering error at time $\tau_t$ appears as

$$\tilde{R}_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) \leq C_R(2+\epsilon)\,\frac{Kn}{\alpha_n\, n_{\min}^2}\,\min\left(1; \sqrt{\frac{n_{\max}\,\alpha_n s}{n}}\right)$$

## 5. Estimating the number of clusters

Estimating the number of clusters is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem with a known number of clusters. A common method for finding the number of clusters is the so called "elbow" method that looks at the fraction of the variance explained as a function of the number of clusters. The method is based on the idea that one should choose the smallest number of clusters such that adding another cluster does not significantly improve fitting of the data by a model. There are many ways to define the "elbow". For example, one of the methods is based on evaluation of the clustering error in terms of an objective function [42], while another one monitors the eigenvalues of the non-backtracking matrix or the Bethe Hessian matrix [23]. In the present paper, we employ a very simply strategy of finding the number of clusters by checking the eigen-gaps of matrices $\widehat{\mathbf{P}}_{t,\hat{r}}$, the technique that has been discussed in [36]. The intuition behind the technique is that, for the error-free case where the $(i,j)$-th entry of the adjacency matrix is 1 when the $i$-th node and the $j$-th node are in the same cluster, and 0 otherwise, the rank of the adjacency matrix should be $K$. Combining this observation with Lemma 4, one derives that the eigenvalues of $\widehat{\mathbf{P}}_{t,\hat{r}}$ can be used to estimate the true number of clusters $K$ as long as there exists a sufficient eigen-gap between the $K$-th and the $(K+1)$-th eigenvalues.

Indeed, denote the sorted eigenvalues of any symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ by $\lambda_1(\mathbf{X}) \geq \lambda_2(\mathbf{X}) \geq \ldots \geq \lambda_n(\mathbf{X})$. Then, due to the matrix perturbation inequality $|\lambda_i(\mathbf{X}) - \lambda_i(\mathbf{Y})| \leq \|\mathbf{X} - \mathbf{Y}\|$ (see, e.g., Corollary III.2.6 of [4]), obtain

$$\lambda_{K+1}(\widehat{\mathbf{P}}_{t,\hat{r}}) \leq \|\widehat{\mathbf{P}}_{t,\hat{r}} - \mathbf{P}_t\|, \quad \lambda_j(\widehat{\mathbf{P}}_{t,\hat{r}}) \geq \lambda_j(\mathbf{P}_t) - \|\widehat{\mathbf{P}}_{t,\hat{r}} - \mathbf{P}_t\|, \quad j = 1, \cdots, K.$$

Denote $\lambda_{j,t} = \lambda_j(\mathbf{P}_t)$, $\widehat{\lambda}_{j,t} = \lambda_j(\widehat{\mathbf{P}}_{t,\widehat{r}})$ and

$$\epsilon_t = \Delta_t(\widehat{r}_t)/\lambda_{K,t} \quad \text{with} \quad \Delta_t(\widehat{r}_t) = \|\widehat{\mathbf{P}}_{t,\widehat{r}} - \mathbf{P}_t\|.$$

Then, one has

$$\frac{\widehat{\lambda}_{j+1,t}}{\widehat{\lambda}_{j,t}} \geq \frac{1-\epsilon_t}{\lambda_{j,t}/\lambda_{j+1,t} + \epsilon_t}, \ j = 1, \cdots, K-1, \qquad \frac{\widehat{\lambda}_{K+1,t}}{\widehat{\lambda}_{K,t}} \leq \frac{\epsilon_t}{1-\epsilon_t}. \tag{5.1}$$

Hence, if $\epsilon_t$ is small enough, then there exists a threshold $\varpi$ such that

$$\frac{\widehat{\lambda}_{j+1,t}}{\widehat{\lambda}_{j,t}} > \varpi, \ j = 1, \cdots, K-1, \quad \frac{\widehat{\lambda}_{K+1,t}}{\widehat{\lambda}_{K,t}} \leq \varpi \tag{5.2}$$

while $\widehat{\lambda}_{j+1,t}/\widehat{\lambda}_{j,t}$ can exhibit chaotic behavior for $j \geq K+1$. Therefore, one can estimate $K$ by

$$\widehat{K} = \min\left\{k : \sum_{t=1}^{T} \widehat{\lambda}_{k+1,t} < \varpi \sum_{t=1}^{T} \widehat{\lambda}_{k,t}\right\} \tag{5.3}$$

where $\varpi$ is a tuning parameter. Note that the expression for $\widehat{K}$ is somewhat similar to the one suggested by Le and Levina [23] with the difference that we use the eigenvalues of the adjacency matrix in the situation of a time-dependent network while they use the eigenvalues of the non-backtracking matrix in the stationary case.

The following statement shows that if eigenvalues of $\mathbf{P}_t$ grow at most exponentially and $\lambda_{K,t} = \lambda_{\min}(\mathbf{P}_t)$ is large enough, then $\widehat{K}$ is an accurate estimator of $K$ with high probability.

**Proposition 1.** *Let Assumptions* **A1**–**A3** *hold and* $\alpha_n \geq C_\alpha^{-1} c_0 \log n/n$. *Let for some* $w > 0$

$$\lambda_j(\mathbf{P}_t) \leq (1+w)\,\lambda_{j+1}(\mathbf{P}_t), \quad j = 1, \cdots, K-1, \tag{5.4}$$

*where* $K$ *is the true number of clusters. If* $T \leq n^{\tau_1}$ *for some* $\tau_1 < \infty$,

$$\lambda_{\min}(\mathbf{P}_t) \geq (40 + 10w)\,\Delta_t(r^*) \tag{5.5}$$

*where* $\Delta_t(r^*)$ *is defined in (3.9), then for any* $\tau > 0$, *with probability at least* $1 - 4T\,n^{-(\tau-\tau_1)}$, *inequalities (5.2) hold with* $\varpi = (3+w)^{-1}$ *and* $\widehat{K} = K$.

Observe that condition (5.5) on the lowest nonzero eigenvalue of $\mathbf{P}_t$ is essentially a necessary condition required for accurate clustering. Indeed, $\Delta_t(r^*) \leq C_\Delta \ \min(\delta_1, \delta_2)$ by (3.9) and, by Assumption **A3**, $\lambda_{\min}(\mathbf{P}_t) \geq C_\lambda^{-1}\,\alpha_n n_{\min}$, so that (5.5) is guaranteed by

$$\aleph = \frac{\min(\delta_1^2, \delta_2^2)}{\alpha_n^2 n_{\min}^2} \leq \tilde{C} = [C_\Delta\, C_\lambda\,(40+10w)]^{-2}. \tag{5.6}$$

On the other hand, the clustering error in (4.4) is bounded above by $\tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \leq C_R(2+\epsilon)\, K\, \aleph$ where $\aleph$ is defined in (5.6). Therefore, a small value $\tilde{R}_t(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t) \leq \delta$ of the clustering error implies that $\aleph \leq (C_R(2+\epsilon))^{-1}\delta/K$ which ensures (5.6) provided that $K$ is large enough.

Note also that assumption (5.4) is not restrictive. Indeed, since $\lambda_K(\mathbf{P}_t) = \lambda_{\min}(\mathbf{P}_t) \geq C_\lambda^{-1}\, \alpha_n n_{\min}$ and $\lambda_1(\mathbf{P}_t) = \lambda_{\max}(\mathbf{P}_t) \leq C_\lambda\, \alpha_n n_{\max}$, obtain that

$$\frac{\lambda_1(\mathbf{P}_t)}{\lambda_K(\mathbf{P}_t)} \leq C_\lambda^2\, \frac{n_{\max}}{n_{\min}},$$

so condition (5.4) always holds, for example, in the case of a balanced model satisfying (4.10).

Combination of Theorem 1 and Proposition 1 immediately yield the following corollary.

**Corollary 3.** *Let clustering be carried out according to the Algorithm 1 with $\widehat{K}$ clusters. Let assumptions of Theorem 1 and conditions (5.4) and (5.5) be valid. Then, for any $\tau > 0$, with probability at least $1 - 4\, n^{-(\tau-\tau_1)}$, inequalities (4.4), (4.5) and (4.6) hold.*

## 6. Simulations

In this section, we evaluate the accuracy of Algorithm 1 via a limited simulation study and compare it with two other "baseline" methods. Specifically, the first method applies spectral clustering separately to each observed adjacency matrix $\mathbf{A}_t$ (instead of $\widehat{\mathbf{P}}_{t,r}$ in Algorithm 1), thus, essentially following Lei and Rinaldo [24]. The second method applies spectral clustering to the sum of all observed adjacency matrices $\sum_{t=1}^{n} \mathbf{A}_t$ as it is done in [5]. The precision is measured in terms of overall relative clustering error defined in (2.3).

In the simulations, we set $\alpha_n = 1$ and use $n = 100$, $T = 1000$, $K = 3$. We set all the diagonal entries of matrices $\mathbf{B}_t$ to be $f_1(t/T)$ and all the off-diagonal entries of $\mathbf{B}_t$ to be $f_2(t/T)$, with the choices of $f_1$ and $f_2$ given later. For $t = 1$, we randomly assign each of the $n$ nodes to one of the $K$ clusters with equal probabilities, and for each $t > 1$, we randomly choose $s$ nodes from the previous time step and assign their memberships at random.

We generate the kernel in (3.2) following Section 7.1 with the parameters $(l, m, m_0) = (4, 2, 1)$. Furthermore, we replace the $(1+\epsilon)$-approximate $k$-means algorithm in the step 3 of Algorithm 1 with the standard iterative refinement $K$-means algorithm, which is used in the built-in `kmeans` function in MATLAB. Since it performs well empirically, we do not expect that application of the $(1+\epsilon)$-approximate $k$-means algorithm will lead to notably different results.

Furthermore, we study the performances of our technique and the two baseline methods under the following five models

1. $f_1(x) = 0.5 + 0.1\sin(2\pi x + 0.1\pi)$, $f_2(x) = 0.3 + 0.1\sin(2\pi x)$, $n = 100$;
2. $f_1(x) = 0.5 + 0.1\sin(2\pi x + 0.1\pi)$, $f_2(x) = 0.3 + 0.1\sin(2\pi x)$, $n = 400$;
3. $f_1(x) = 0.45 + 0.1\sin(2\pi x + 0.1\pi)$, $f_2(x) = 0.3 + 0.1\sin(2\pi x)$, $n = 100$;

TABLE 1
*The overall relative clustering errors of Algorithm 1, Baseline 1 (point-per-point clustering) and Baseline 2 (clustering of the sum) algorithms for various values of s and r.*

| s | 0 | 1 | 2 | 4 | 8 | 16 | 32 | 50 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| Model 1 | | | | | | | | | |
| Baseline 1 | 21.52 | 21.97 | 23.75 | 20.90 | 23.14 | 24.15 | 23.43 | 23.87 | 23.13 |
| Baseline 2 | 0.00 | 13.31 | 19.83 | 31.88 | 41.67 | 53.09 | 56.35 | 58.83 | 60.26 |
| Algorithm 1 | 0.00 | 0.32 | 0.71 | 1.46 | 2.71 | 5.79 | 11.57 | 20.17 | 24.41 |
| Model 2 | | | | | | | | | |
| Baseline 1 | 0.33 | 0.33 | 0.34 | 0.30 | 0.27 | 0.35 | 0.33 | 0.30 | 0.30 |
| Baseline 2 | 0.00 | 4.96 | 7.48 | 14.27 | 22.30 | 33.06 | 43.02 | 49.37 | 57.31 |
| Algorithm 1 | 0.00 | 0.01 | 0.03 | 0.03 | 0.09 | 0.17 | 0.28 | 0.44 | 0.78 |
| Model 3 | | | | | | | | | |
| Baseline 1 | 41.85 | 41.84 | 41.48 | 43.38 | 43.24 | 41.48 | 40.89 | 42.01 | 41.41 |
| Baseline 2 | 0.00 | 12.48 | 23.74 | 40.16 | 45.59 | 54.36 | 58.59 | 59.79 | 59.92 |
| Algorithm 1 | 0.08 | 0.87 | 1.39 | 2.83 | 5.62 | 12.17 | 26.54 | 36.75 | 42.06 |
| Model 4 | | | | | | | | | |
| Baseline 1 | 17.19 | 17.25 | 18.42 | 17.10 | 17.69 | 18.32 | 17.55 | 18.23 | 17.26 |
| Baseline 2 | 0.00 | 13.42 | 26.57 | 39.68 | 45.15 | 51.98 | 56.21 | 57.67 | 58.50 |
| Algorithm 1 | 0.03 | 0.42 | 1.03 | 1.79 | 4.50 | 8.72 | 14.36 | 18.72 | 18.82 |
| Model 5 | | | | | | | | | |
| Baseline 1 | 26.60 | 24.13 | 28.30 | 24.18 | 23.37 | 21.35 | 23.25 | 24.16 | 22.40 |
| Baseline 2 | 0.00 | 11.51 | 24.16 | 36.62 | 42.29 | 52.31 | 57.84 | 58.30 | 59.58 |
| Algorithm 1 | 0.00 | 0.29 | 0.87 | 1.41 | 2.48 | 5.17 | 11.90 | 19.88 | 23.66 |

4. $f_1(x) = 0.6 - 0.1\sin(2\pi x + 0.1\pi)$, $f_2(x) = 0.3 + 0.1\sin(2\pi x)$, $n = 100$;
5. $f_1(x) = 0.5 + 0.1\sin(20\pi x + \pi)$, $f_2(x) = 0.3 + 0.1\sin(20\pi x)$, $n = 100$;

and various choices of $s$ and $r$. The Model 1 can be viewed as a basic model; the Model 2 represents a more difficult case with a larger "signal to noise ratio" (since $f_1$ is smaller than in Model 1); the Model 3 is similar to Model 1 with larger variations of connection probabilities (i.e., less smoothness) in time.

Table 1 presents the overall relative clustering errors for the three methods, three models and various choices of the number of membership switches $s$. In our study, $r$ are chosen appropriately according to each model, from the choices of $r = 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20, 25, 30$. Empirically, our choices of $r$ decreases as $s$ increases: it is intuitive that when $s = 0$, the membership does not change over time and the largest $r$ would perform best; and when $s = 100$, any two consecutive membership vectors do not have any correlation and the optimal $r$ should be small.

Table 1 shows that Algorithm 1 delivers better precision than the baseline methods for the most cases. Indeed, Baseline 1 method is more accurate than Algorithm 1 only when $s$ is very large and close to $T$, i.e., when the membership vector changes almost completely between two consecutive time points. On the other hand, Baseline 2 method is more accurate than Algorithm 1 only when $s$ is very small and close to 0, i.e., when the membership vector is almost fixed across all time points. Moreover, even when one of the baseline algorithm outperforms our method, the differences in the clustering errors are very minor.

Table 1 also shows that some settings are more difficult than the others.

For example, all methods have larger errors under Models 3 and 5 than under Model 1 since the differences between $f_1$ and $f_2$ are smaller in Model 3, and $f_1$ and $f_2$ are less "smooth" in Model 5. On the other hand, all methods have smaller errors under Model 2 than under Model 1 since a larger $n$ with fixed $K$ provides more information about the underlying matrix $\mathbf{B}_t$, which in turn leads to a better clustering precision. In Model 4, the difference $f_1 - f_2$ has a larger mean and a larger variance than Model 1, which leads to the clustering precision similar to Model 1. Nevertheless, the advantage of our algorithm over the two baseline methods is consistent across all models.

The computational cost of the Algorithm 1 can be summarized as follows: Step 1 requires $O(n^2 r)$ and Step 2 requires $O(n^2 K)$ operations. The computational cost of Step 3 depends on the empirical implementation of the $K$-means method. In particular, the standard iterative refinement procedure used in the simulations above requires $O(n^2 K)$ operations per iteration, while the approximate algorithm in [22] has a computational cost of $O(n)$ for any fixed $K$.

## 7. Appendix

### *7.1. Construction of kernels of integer arguments*

First consider construction of the kernel $W^{(1)}_{r,l}$ designed for internal points. Since it has symmetric domain, we construct a symmetric version of the kernel

$$W^{(1)}_{r,l}(i) = \sum_{j=0}^{m} a_j (r - |i|)^{j+m_0} r^{-(j+m_0)}, \quad i = -r, \dots, r, \tag{7.1}$$

where $m_0$ is a nonnegative integer and coefficients $a_j, j = 0, 1 \dots, m$, are to be determined. Note that if $m_0 \geq 1$, then $W^{(1)}_{r,l}(\pm r) = 0$. We need to find $a_j, j = 0, \dots, m$, such that

$$\sum_{i=-r}^{r} W^{(1)}_{r,l}(i) = 2r + 1; \quad \sum_{i=-r}^{r} i^k W^{(1)}_{r,l}(i) = 0, \quad k = 1, \dots, l. \tag{7.2}$$

Note that due to the symmetry of the kernel, the second equation in (7.2) holds for any odd value of $k$, hence, we need to consider only even values $k = 2k_0$. The latter also means that we can consider $l = 2l_0$ and $l = 2l_0 + 1$ simultaneously: any order $2l_0$ kernel is also automatically an order $2l_0 + 1$ kernel. Plugging (7.1) into (7.2) and simplifying the expressions, we rewrite (7.2) as

$$\sum_{j=0}^{m} a_j \left[ 1 + 2r^{-(j+m_0)} \sum_{i=0}^{r-1} i^{j+m_0} \right] = 2r + 1,$$

$$\sum_{j=0}^{m} a_j r^{-(j+m_0+2k_0)} \sum_{i=0}^{r-1} i^{j+m_0} (r - i)^{2k_0} = 0, \quad k_0 = 1, \dots, l_0.$$

Denote

$$\mathcal{P}_{h,k}(r) = r^{-(h+k+1)} \sum_{i=0}^{r-1} i^h (r-i)^k, \quad h = 0, 1, \ldots, \ k = 0, 1, 2 \tag{7.3}$$

Observe that $\mathcal{P}_{h,k}(r)$ are polynomials in $1/r$ of degree $h+k+1$ and that expressions for $\mathcal{P}_{h,k}(r)$ can be found exactly for every $h$ and $k$ using, e.g., [15], formula 0.121.

Then, vector $\mathbf{a} = (a_0, a_1, \cdots, a_m)^T$ can be found as a solution of the following system of linear equations

$$\mathbf{K}\mathbf{a} = (2r+1)\mathbf{e}_1, \tag{7.4}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ is the canonical vector in $\mathbb{R}^{m+1}$ and matrix $\mathbf{K} \in \mathbb{R}^{(l_0+1)\times(m+1)}$ has elements

$$\mathbf{K}_{0,j} = 1 + 2r\mathcal{P}_{j+m_0,0}, \quad \mathbf{K}_{k_0,j} = \mathcal{P}_{j+m_0,2k_0}, \quad j = 0, \ldots, m+1, \ k_0 = 1, \ldots, l_0.$$

Since the rows of matrix $\mathbf{K}$ are linearly independent, the system of equations (7.4) has a solution whenever $m \geq l_0$. If $m = l_0$, then this solution is unique. If $m > l_0$, then (7.4) has multiple solutions and one can find vector $\mathbf{a}$ such that, for example, $\|W_{r,l}^{(1)}\|_\infty = \max_i |W_{r,l}^{(1)}(i)|$ takes the minimal value. The latter can be accomplished by solving the following linear programming problem

$$w \Rightarrow \min \quad \text{s.t.} \quad \mathbf{K}\mathbf{a} = (2r+1)\mathbf{e}_1, \quad \mathbf{q}^{(i)}\mathbf{a} \leq w; \ -\mathbf{q}^{(i)}\mathbf{a} \leq w, \ i = 0, \ldots, r. \tag{7.5}$$

where $\mathbf{q}^{(i)}$ are vectors with components $\mathbf{q}_j^{(i)} = r^{-(j+m_0)}(r-i)^{j+m_0}$, $j = 0, \ldots, m$.

Construction of the boundary kernels $W_{r,l}^{(j)}$, $j = 2, 3$, are very similar to $W_{r,l}^{(1)}$. For the sake of brevity, we describe only construction of $W_{r,l}^{(2)}$. Write $W_{r,l}^{(2)}$ in a form

$$W_{r,l}^{(2)}(i) = \sum_{j=0}^{m} a_j (r-i)^{j+m_0} r^{-(j+m_0)}, \quad i = 0, \ldots, r,$$

and choose the coefficients, so that the kernel satisfies condition (3.1). The latter leads to the system of $(l+1)$ linear equations of the form

$$\sum_{j=0}^{m} a_j \, (1 + r\mathcal{P}_{j+m_0,0}(r)) = 1 + r, \quad \sum_{j=0}^{m} a_j \mathcal{P}_{j+m_0,k}(r) = 0, \ k = 1, \ldots, l, \ j = 0, \ldots, m,$$

which can be written in a matrix form as $\mathbf{K}\mathbf{a} = (r+1)\mathbf{e}_1$ where $\mathbf{e}$ is as before and and matrix $\mathbf{K} \in \mathbb{R}^{(l+1)\times(m+1)}$ has elements

$$\mathbf{K}_{0,j} = 1 + r\mathcal{P}_{j+m_0,0}, \quad \mathbf{K}_{k_0,j} = \mathcal{P}_{j+m_0,k}, \quad j = 0, \ldots, m+1, \ k = 1, \ldots, l.$$

Similarly to the case of construction of $W_{r,l}^{(1)}$, one can obtain an unique solution of the system of equations by choosing $m = l$ or find it as a solution of a linear programming problem similarly to (7.5).

### *7.2. Proofs of the statements in the paper*

**Proof of Lemma 5.** The proof of Lemma 5 can be split into two steps. In the first step, it bound the difference of the eigenvectors of $\mathbf{P}_t$ and $\widehat{\mathbf{P}}_t$ by $\|\widehat{\mathbf{P}}_t - \mathbf{P}_t\|$. In the second step, it applies the analysis from [24] to bound the classification error by the difference of eigenvectors.

If $\mathbf{P}_t = \mathbf{U}\mathbf{D}\mathbf{U}^T$ and $\widehat{\mathbf{P}}_t = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^T$, then, by Lemma 5.1 of [24], obtain that there exists and orthogonal matrix $\mathbf{O}$ such that

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\| \leq \frac{2\sqrt{2K}}{\lambda_{\min}(\mathbf{P}_t)} \|\widehat{\mathbf{P}}_t - \mathbf{P}_t\|.$$

Let $S_{t,k}$ is a subset of nodes in class $G_{t,k}$ that are misclassified. Then, Lemma 5.3 and Theorem 3.1 of Lei and Rinaldo (2015) imply that

$$\sum_{k=1}^{K} \frac{|S_{t,k}|}{n_t(k)} \leq 8(2+\epsilon)\,\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|^2 \leq \frac{64K(2+\epsilon)}{[\lambda_{\min}(\mathbf{P}_t)]^2}\,\|\widehat{\mathbf{P}}_t - \mathbf{P}_t\|^2. \tag{7.6}$$

In order to derive (4.3), observe that

$$R_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{\Theta}_t) = \sum_{k=1}^{K} |S_{t,k}| \leq \frac{n_{t,\max}}{n} \sum_{k=1}^{K} \frac{|S_{t,k}|}{n_t(k)}.$$

**Proof of Lemma 1.** Since the case $r = 0$ follows directly from Theorem 1.1 in the Supplementary material of [24], we can assume that $r \geq 1$. Also, in order to simplify the proof, we do not consider kernels $W_{r,l}^{(j)}$ for each $j = 1, 2, 3$, separately, but instead remove the index $j$ since the proofs are practically identical for all three values of $j$.

We remark that while the proof of Lemma 1 is a generalization of and based on the proof of Lemma 2.1 in the Supplementary material of [24], some steps still require nontrivial derivation. For example, Lemma 9 is derived specifically for our setting.

Lemma 2.1 in the Supplementary material of [24] (with $\delta = 1/2$) implies that

$$\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\| \leq 4 \sup_{\mathbf{x},\mathbf{y}\in\mathcal{T}} |\mathbf{x}^T(\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r})\mathbf{y}|, \tag{7.7}$$

where

$$\mathcal{T} = \{\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n, \|\mathbf{x}\| = 1, \quad 2\sqrt{n}x_i \text{ are all integers.}\} \tag{7.8}$$

Hence, we bound above the right-hand side of (7.7) by dividing the coordinates of $\mathbf{x}$ and $\mathbf{y}$ into "light pairs" and "heavy pairs" as follows

$$\mathcal{L}(\mathbf{x},\mathbf{y}) = \left\{(i,j) : |x_i y_j| \leq \sqrt{\frac{C_\alpha \alpha_n r}{n}}\right\}, \quad \bar{\mathcal{L}}(\mathbf{x},\mathbf{y}) = \left\{(i,j) : |x_i y_j| > \sqrt{\frac{C_\alpha \alpha_n r}{n}}\right\}.$$

Note that here the definition is different from the proof in [24] by a factor of $r$, since we consider weighted sum of random matrices (instead of a single random matrix). Partitioning $|\mathbf{x}^T(\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r})\mathbf{y}|$ into the portions containing the "light pairs" and the "heavy pairs", obtain

$$\begin{aligned}\left|\mathbf{x}^T(\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r})\mathbf{y}\right| \;&\le\; \Big|\sum_{(i,j)\in\mathcal{L}(\mathbf{x},\mathbf{y})} x_i[\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}](i,j)y_j\Big| \\ &+\; \Big|\sum_{(i,j)\in\bar{\mathcal{L}}(\mathbf{x},\mathbf{y})} x_i[\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}](i,j)y_i\Big|. \qquad (7.9)\end{aligned}$$

In order to obtain upper bounds for the right-hand side of (7.9), we need three supplementary statements, Lemmas 6, 7 and 8, that generalize, respectively, Lemmas 3.1, 4.1 and 4.2 of Lei and Rinaldo [24] to our setting. The proofs of Lemmas 6, 7, 8 and 9 are deferred till Section 7.3.

**Lemma 6.** *Under assumptions of Lemma 1, with probability at least* $1 - 2n^{-\tau}$, *one has*

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{T}}\Big|\sum_{(i,j)\in\mathcal{L}(\mathbf{x},\mathbf{y})} x_i y_j[\widehat{\mathbf{P}}_{t,r}(i,j) - \mathbf{P}_{t,r}(i,j)]\Big| \le C_{\tau,1}W_{\max}\sqrt{\frac{C_\alpha\, n\,\alpha_n}{r}}$$

*provided*

$$C_{\tau,1} \ge \max\left\{2\sqrt{(\tau+\log 14)},\ 8(\tau+\log 14)/3\right\}. \qquad (7.10)$$

**Lemma 7.** *Let* $d_t(i)$ *be the degree of the* $i$*-th node in the network with connection probabilities given by the matrix* $\mathbf{P}_t$. *Denote*

$$d_{t,r}(i) = \sum_{k\in\mathcal{F}_r} W_{r,l}(k)d_{t+k}(i).$$

*Then, under assumptions of Lemma 1, one has*

$$\Pr\left\{\max_{1\le i\le n} d_{t,r}(i) \le 3C_\alpha\left(W_{\max}C_{\tau,2}+1\right)n\alpha_n r\right\} \ge 1 - n^{-\tau} \qquad (7.11)$$

*provided*

$$C_{\tau,2} \ge \max\left(\sqrt{\frac{2(\tau+1)}{c_0}}, \frac{\tau+1}{3c_0}\right). \qquad (7.12)$$

**Lemma 8.** *Let* $I, J \subseteq \{1,\cdots,n\}$ *with* $|I| \le |J|$. *Denote*

$$\bar{\mu}(I,J) = C_\alpha\alpha_n|I||J|\,|\mathcal{F}_r|, \quad e_{t,r}(I,J) = \sum_{k\in\mathcal{F}_r} W_{r,l}(k)e_{t+k}(I,J),$$

*where* $e_l(I,J)$ *represents the number of distinct edges between* $I$ *and* $J$ *in the network at time* $l$. *Assume that the event in* (7.11) *holds. Then with probability at least* $1 - n^{-\tau}$, *at least one of the following is true:*

1. $e_{t,r}(I,J) \le e\, C_{\tau,3}\, \bar{\mu}(I,J)$,
2. $e_{t,r}(I,J)\, \log\left(\frac{e_{t,r}(I,J)}{\bar{\mu}(I,J)}\right) \le C_{\tau,4}\, |J|\, \log\left(\frac{n}{|J|}\right)$.

*Here*

$$C_{\tau,3} = \max\{3(W_{\max}C_{\tau,2}+1), e^{3W_{\max}}+1\}, \quad C_{\tau,4} = 8\, W_{\max}(\tau+6). \tag{7.13}$$

**Lemma 9.** *Let* $\{X_i\}_{i=1}^n$ *be independent random variables such that* $\Pr(X_i = 1-p_i) = p_i$, $\Pr(X_i = -p_i) = 1-p_i$ *for some* $p_i > 0$. *Let* $X = \sum_{i=1}^n w_i X_i$, $p = \frac{1}{n}\sum_{i=1}^n p_i$, $p_{\max} = \max_{1\le i\le n} p_i$, $w_{\max} = \max_{1\le i\le n} w_i$, *then for* $k > \max(e^{3w_{\max}}, 2)$,

$$\Pr(X \ge k p_{\max} n) < e^{-(k+1)p_{\max} n \ln(k+1)/2w_{\max}}.$$

The first term in the right-hand side of (7.9) corresponding to the "light pairs" is bounded by Lemma 6. In order to bound the "heavy pairs" in the second term, observe that

$$\left|\sum_{(i,j)\in\bar{\mathcal{L}}(\mathbf{x},\mathbf{y})} x_i y_j\, \mathbf{P}_{t,r}(i,j)\right| \le \frac{1}{|\mathcal{F}_r|} \sum_{(i,j)\in\bar{\mathcal{L}}(\mathbf{x},\mathbf{y})} \sum_{k\in\mathcal{F}_r} \frac{x_i^2 y_j^2}{|x_i y_j|}\, |W_{r,l}(k)|\, \mathbf{P}_{t+k}(i,j) \tag{7.14}$$

$$\le \frac{1}{|\mathcal{F}_r|}\sqrt{\frac{n}{C_\alpha \alpha_n r}} \sum_{k\in\mathcal{F}_r} |W_{r,l}(k)|\, C_\alpha \alpha_n \sum_{(i,j)\in\bar{\mathcal{L}}(\mathbf{x},\mathbf{y})} x_i^2 y_j^2 \le W_{\max}\sqrt{\frac{C_\alpha\, n\, \alpha_n}{r}}.$$

Applying Lemmas 7 and 8, and the same argument as in Section 4 in the Supplementary material of [24] with (with $C_\alpha\, n\, \alpha_n\, r$ replacing $d$), we derive

$$\Pr\left\{\frac{1}{|\mathcal{F}_r|}\left|\sum_{\substack{(i,j)\in\bar{\mathcal{L}}(\mathbf{x},\mathbf{y})\\ k\in\mathcal{F}_r}} x_i\, y_j\, W_{r,l}(k)\, \mathbf{A}_{t+k}(i,j)\right| \le \tilde{C}_\tau\, \sqrt{\frac{C_\alpha n\, \alpha_n}{r}}\right\} \ge 1 - 2n^{-\tau}, \tag{7.15}$$

where $\tilde{C}_\tau = 8\,\{16\delta^{-2} + e\, C_{\tau,3}\delta^{-2} + 24\,(W_{\max}C_{\tau,2}+1) + 40\, C_{\tau,4} + 8\}$ with $\delta = 1/2$. Combining (7.14) and (7.15), obtain that the second term in the right-hand side of (7.9) is bounded above by $(W_{\max} + \tilde{C}_\tau)\,\sqrt{C_\alpha\, n\, \alpha_n / r}$, with probability at least $1 - 2n^{-\tau}$.

Combining (7.7), (7.9), Lemma 6, (7.14) and (7.15), we derive

$$\Pr\left\{\|\widehat{\mathbf{P}}_{t,r} - \mathbf{P}_{t,r}\| \le 4(W_{\max} + W_{\max}C_{\tau,1} + \tilde{C}_\tau)\sqrt{\frac{C_\alpha\, n\, \alpha_n}{r}}\right\} \ge 1 - 4n^{-\tau},$$

and obtain the expression for $C_{0,\tau}$ in (3.4):

$$C_{0,\tau} = 4\sqrt{C_\alpha}\,\{W_{\max}(1+C_{\tau,1}) + 32(24\, W_{\max}C_{\tau,2} + 4e\, C_{\tau,3} + 40\, C_{\tau,4} + 96)\} \tag{7.16}$$

where $C_{\tau,1}, C_{\tau,2}, C_{\tau,3}$ and $C_{\tau,4}$ are defined in (7.10), (7.12) and (7.13), respectively.

**Proof of Lemma 2.** First, let us prove that under Assumption **A2**, one has for any $k$ such that $1 \le t+k \le T$ one has

$$\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k}\| \le \sqrt{2|k|s}. \tag{7.17}$$

Let, without loss of generality, $k > 0$. Note that matrix $\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k}$ at most $ks$ nonzero rows in which one entry is 1 and another is -1. If we permute the rows of matrix $\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k}$ so that those nonzero rows are the first ones, we obtain that $(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k})(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k})^T$ is the block-diagonal matrix with the only nonzero block matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{ks \times ks}$ in the top left corner that has elements with absolute values equal to 0,1 or 2. Then

$$\lambda_{\max}\left[(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k})(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t+k})^T\right] = \lambda_{\max}(\boldsymbol{\Lambda}) \le \max_{1\le i\le ks} \sum_{j=1}^{ks} \boldsymbol{\Lambda}_{i,j} \le 2|k|s$$

which implies (7.17). Note that the upper bound is tight (to see this, consider the case where s elements move from class $i$ to class $j$ at each of $k$ time points.

Next, note that $\Delta_{2,t}$ in (3.3) can be decomposed as

$$\|\mathbf{P}_{t,r} - \mathbf{P}_t\| \le \|\mathbf{P}_{t,r} - \tilde{\mathbf{P}}_{t,r}\| + \|\tilde{\mathbf{P}}_{t,r} - \mathbf{P}_t\| \equiv \Delta_{2,1,t} + \Delta_{2,2,t}, \tag{7.18}$$

where

$$\begin{aligned} \mathbf{P}_{t,r} &= r^{-1} \sum_{i=-r}^{r} W_{r,l}(i) \mathbf{P}_{t+i} \\ \tilde{\mathbf{P}}_{t,r} &= r^{-1} \sum_{i=-r}^{r} W_{r,l}(i) \boldsymbol{\Theta}_t \mathbf{B}_{t+i} \boldsymbol{\Theta}_t^T \end{aligned}$$

Let us show that

$$\Delta_{2,1,t} = \|\mathbf{P}_{t,r} - \tilde{\mathbf{P}}_{t,r}\| \le 2\sqrt{2}\, W_{\max}\, C_\lambda\, \alpha_n \sqrt{n_{\max}\, rs} \tag{7.19}$$

For this purpose observe that $\|\boldsymbol{\Theta}_t\| \le \sqrt{n_{\max}}$ and that

$$\begin{aligned} \Delta_{2,1,t} &= \|r^{-1} \sum_{i=-r}^{r} W(i) \left[\boldsymbol{\Theta}_{t+i} \mathbf{B}_{t+i} \boldsymbol{\Theta}_{t+i}^T - \boldsymbol{\Theta}_t \mathbf{B}_{t+i} \boldsymbol{\Theta}_t^T\right] \| \\ &\le W_{\max} \max_{|i|\le r} \|\boldsymbol{\Theta}_{t+i} \mathbf{B}_{t+i} \boldsymbol{\Theta}_{t+i}^T - \boldsymbol{\Theta}_t \mathbf{B}_{t+i} \boldsymbol{\Theta}_t^T\| \end{aligned}$$

where $\|\boldsymbol{\Theta}_{t+i} \mathbf{B}_{t+i} \boldsymbol{\Theta}_{t+i}^T - \boldsymbol{\Theta}_t \mathbf{B}_{t+i} \boldsymbol{\Theta}_t^T\| \le [\|\boldsymbol{\Theta}_{t+i}\| + \|\boldsymbol{\Theta}_t\|]\, \|\mathbf{B}_{t+i}\|\, \|\boldsymbol{\Theta}_{t+i} - \boldsymbol{\Theta}_t\|$. The last two inequalities together with (7.17) imply (7.19).

Now, let us prove that

$$\Delta_{2,2,t} = \|\tilde{\mathbf{P}}_{t,r} - \mathbf{P}_t\| \le \frac{L}{(l_\beta)!}\, W_{\max}\, \alpha_n n \left(\frac{r}{T}\right)^\beta. \tag{7.20}$$

Let $j = 1, 2$ or $3$ be determined by the value of $t$. Denote

$$\mathbf{Q}_{r,t} = |\mathcal{F}_{r,j}|^{-1} \sum_{i \in \mathcal{F}_{r,j}} W_{r,l}^{(j)}(i) \left(\mathbf{H}_{t+i} - \mathbf{H}_t\right).$$

Observe that by Assumptions **A1** and **A4**, for any $k, k' = 1, \ldots, K$, using Taylor's expansion at $i = 0$, one derives

$$\begin{aligned}
\mathbf{Q}_{r,t}(k,k') =& |\mathcal{F}_{r,j}|^{-1} \sum_{i \in \mathcal{F}_{r,j}} W_{r,l}^{(j)}(i) \left[ f\left(\frac{t+i}{T}; k, k'\right) - f\left(\frac{t}{T}; k, k'\right)\right] \\
=& \sum_{h=1}^{l_\beta} \frac{1}{h!} f^{(h)}\left(\frac{t}{T}; k, k'\right) \left[\frac{1}{r} \sum_{i \in \mathcal{F}_{r,j}} \left(\frac{i}{T}\right)^h W_{r,l}^{(j)}\right] \\
&+ \frac{1}{|\mathcal{F}_{r,j}|(l_\beta)!} \sum_{i \in \mathcal{F}_{r,j}} W_{r,l}^{(j)} \left(\frac{i}{T}\right)^l \left[ f^{(l_\beta)}\left(\frac{t}{T} + \xi; k, k'\right) - f^{(l_\beta)}\left(\frac{t}{T}; k, k'\right)\right],
\end{aligned}$$

where $|\xi| \le r/T$. Due to Assumption **A1**, the first sum is equal to zero and

$$|\mathbf{Q}_{r,t}(k,k')| \le \frac{1}{|\mathcal{F}_{r,j}|\,(l_\beta)!} \sum_{i \in \mathcal{F}_{r,j}} \left(\frac{i}{T}\right)^{l_\beta} |W_{r,l}^{(j)}(i)|\, L|\xi|^{\beta - l} \le \frac{L W_{\max}}{(l_\beta)!} \left(\frac{r}{T}\right)^\beta. \tag{7.21}$$

Recall that $\mathbf{B}_{t+i} - \mathbf{B}_t = \alpha_n(\mathbf{H}_{t+i} - \mathbf{H}_t)$ and that the spectral norm of a matrix is dominated by the $l_1$ norm. Therefore,

$$\begin{aligned}
\Delta_{2,2,t} &\le \alpha_n \max_{1 \le j' \le n} \sum_{j=1}^n |(\mathbf{\Theta}_t \mathbf{Q}_{r,t} \mathbf{\Theta}_t^T)(j,j')| \\
&\le \alpha_n \max_{k,k'} |\mathbf{Q}_{r,t}(k,k')| \max_{1 \le j' \le n} \sum_{k=1}^K \sum_{k'=1}^K \left[\sum_{j \in G_{t,k}} \mathbf{\Theta}_t(j,k)\right] \mathbf{\Theta}_t(j',k') \\
&= \alpha_n\, n \max_{k,k'} |\mathbf{Q}_{r,t}(k,k')|
\end{aligned}$$

Combination of the last inequality with (7.21) yields (7.20) while (7.18), (7.19) and (7.20) together complete the proof of the lemma.

**Proof of Lemma 3.** If $r^* = 0$, then results of the Lemma follow directly from [24]. Consider $r \ge 1$. Then,

$$\Delta_t(r) \le C \left[ n \left(\frac{r}{T}\right)^\beta + \alpha_n \sqrt{2 r n_{\max} s} + \sqrt{n \alpha_n / r}\right]$$

where $C$ depends on $\tau, c_0, W_{\max}, l, L$ and $\lambda_{\max}$. Denote

$$F_1(r) = n(r/T)^\beta, \quad F_2(r) = \alpha_n \sqrt{2 r n_{\max} s}, \quad F_3(r) = \sqrt{n \alpha_n / r}.$$

It is easy to see that $F_1(r)$ and $F_2(r)$ are growing in $r$ while $F_3(r)$ is declining. Therefore, the minimum is reached at the point $r$ where $F_1(r) + F_2(r) \asymp$

$\max(F_1(r), F_2(r)) = F_3(r)$. Observe that $F_1(r) = F_3(r)$ if $r = r_1$ where $r_1 = \left(n^{-1}\alpha_n T^{2\beta}\right)^{1/(2\beta+1)}$ and $F_2(r) = F_3(r)$ if $r = r_2$ where $r_2 = \sqrt{(\alpha_n n_{\max} s)^{-1}\, n}$. Moreover, $\max(F_1(r), F_2(r))$ occurs at $r^* = \min(r_1, r_2)$ and we need $r^*$ to be an integer. Then, $\min_r(F_1(r) + F_2(r) + F_3(r)) \asymp F_3(r^*)$ and plugging $r^*$ into $F_3(r)$, we obtain (3.7).

**Proof of Lemma 4.** Note that (3.10) implies that for any $r_0 \geq \widehat{r}$ one has

$$\|\widehat{\mathbf{P}}_{t,\widehat{r}} - \widehat{\mathbf{P}}_{t,r_0}\| \leq 4\, C_{0,\tau} \sqrt{n\, \alpha_n/(r_0 \vee 1)}. \tag{7.22}$$

On the other hand, for any $r_0 > \widehat{r}$, there exists $\tilde{r} < r_0$ such that

$$\|\widehat{\mathbf{P}}_{t,\tilde{r}} - \widehat{\mathbf{P}}_{t,r_0}\| > 4\, C_{0,\tau} \sqrt{n\, \alpha_n/(\tilde{r} \vee 1)}. \tag{7.23}$$

Denote, for convenience, $\delta_1(r) = \|\mathbf{P}_{t,r} - \mathbf{P}_t\|$ and $\delta_2(r) = 4\, C_{0,\tau} \sqrt{n\, \alpha_n/(r \vee 1)}$. Note that $\delta_1(r)$ growing in $r$ since using $\mathbf{P}_{t,r}$ as approximations of $\mathbf{P}_t$ are less accurate for larger $r$ due to changes in the underlying probability values and switching of group memberships of nodes in time. Since $\delta_2(r)$ is decreasing in $r$, there exists $r_0$ such that

$$\delta_1(r_0) < \delta_2(r_0), \quad \delta_1(r_0 + 1) \geq \delta_2(r_0 + 1)$$

Then,

$$\begin{aligned} \delta_1(r^*) + \delta_2(r^*) &= \min_r[\delta_1(r) + \delta_2(r)] \geq \min_r \max[\delta_1(r), \delta_2(r)] \\ &= \max[\delta_1(r_0 + 1), \delta_2(r_0)] \geq \delta_2(r_0) > [\delta_1(r_0) + \delta_2(r_0)]/2, \end{aligned}$$

so that

$$\delta_1(r_0) + \delta_2(r_0) < 2[\delta_1(r^*) + \delta_2(r^*)] \leq 2[\delta_1(r_0) + \delta_2(r_0)]. \tag{7.24}$$

Let $\Omega_\tau$ be the set defined in Lemma 1 and let $\omega \in \Omega_\tau$. Now consider two cases: $\widehat{r} \geq r_0$ and $\widehat{r} < r_0$.

If $\widehat{r} \geq r_0$, then by (7.22) one has

$$\begin{aligned} \|\widehat{\mathbf{P}}_{t,\widehat{r}} - \mathbf{P}_t\| &\leq \|\widehat{\mathbf{P}}_{t,\widehat{r}} - \widehat{\mathbf{P}}_{t,r_0}\| + \|\widehat{\mathbf{P}}_{t,r_0} - \mathbf{P}_t\| \\ &\leq 4\, C_{0,\tau} \sqrt{n\, \alpha_n/(r_0 \vee 1)} + \delta_1(r_0) + \delta_2(r_0) \\ &= 5\delta_2(r_0) + \delta_1(r_0) \leq 5[\delta_1(r_0) + \delta_2(r_0)], \end{aligned}$$

so it follows from (7.24) that

$$\|\widehat{\mathbf{P}}_{t,\widehat{r}} - \mathbf{P}_t\| < 10 \min_r[\delta_1(r) + \delta_2(r)]. \tag{7.25}$$

On the other hand, if $\widehat{r} < r_0$, then there exist $\tilde{r} < r_0$ such that (7.23) holds. Therefore, due to $\delta_1(\tilde{r}) < \delta_1(r_0) < \delta_2(r_0) < \delta_2(\tilde{r})$, obtain

$$\begin{aligned} \|\widehat{\mathbf{P}}_{t,r_0} - \widehat{\mathbf{P}}_{t,\tilde{r}}\| &\leq \|\widehat{\mathbf{P}}_{t,r_0} - \mathbf{P}_t\| + \|\widehat{\mathbf{P}}_{t,\tilde{r}} - \mathbf{P}_t\| \\ &\leq \delta_1(r_0) + \delta_2(r_0) + \delta_1(\tilde{r}) + \delta_2(\tilde{r}) < 4\, \delta_2(\tilde{r}) \end{aligned}$$

which contradicts (7.23). Hence, $\widehat{r} \geq r_0$ for $\omega \in \Omega_\tau$ and validity of Lemma 4 follows from (7.25).

**Proof of Theorem 1.** Recall that $\mathbf{P}_t = \alpha_n \mathbf{\Theta}_t \mathbf{H}_t \mathbf{\Theta}_t^T$ with $\lambda_{\min}(\mathbf{H}_t) \geq C_\lambda^{-1}$. Since $\mathbf{\Theta}_t^T \mathbf{\Theta}_t = \mathbf{\Lambda}_t^2$, the diagonal matrix with $n_{t,1}, \cdots, n_{t,K}$ on the diagonal, $\mathbf{\Theta}_t = \tilde{\mathbf{U}}_t \mathbf{\Lambda}_t$ where $\tilde{\mathbf{U}}_t = \mathbf{\Theta}_t \mathbf{\Lambda}_t^{-1}$ is an orthogonal matrix. Therefore, in (4.2) and (4.3)

$$\lambda_{\min}(\mathbf{P}_t) \geq C_\lambda^{-1} \, \alpha_n \, n_{\min}.$$

Combining the last inequality, (4.2) and (4.3) with Lemma 3, immediately obtain (4.4) and (4.5).

**Proof of Proposition 1.** Let $\Omega_{t,\tau}$ be a set with $\Pr(\Omega_{t,\tau}) \geq 1 - 4n^{-\tau}$ where (3.11) hold. Then, due to (3.11) and (5.5), $\epsilon_t \leq 10 \, \Delta_t(r^*)/\lambda_{K,t} \leq (4+w)^{-1}$. Therefore, (5.1) and (5.4) yield that

$$\frac{\widehat{\lambda}_{j+1,t}}{\widehat{\lambda}_{j,t}} \geq \frac{1-(4+w)^{-1}}{1+w+(4+w)^{-1}} > \frac{1}{3+w}, \; j = 1, \cdots, K-1, \quad \frac{\widehat{\lambda}_{K+1,t}}{\widehat{\lambda}_{K,t}} \leq \frac{1}{3+w}$$

which completes the proof.

### *7.3. Proofs of supplementary lemmas*

For Lemmas 6, 7 and 8, their proofs are mainly based on the proofs of (7.26), (7.27) and (7.29), which in turn are based on the measure concentration. In particular, (7.26) and (7.27) are based on Bernstein's inequality and (7.29) are based on the proof of Lemma 9, which estimates the summation of poisson distributions.

**Proof of Lemma 6.** First, it is suffient to prove that for any $C > 0$,

$$\Pr\Big\{ \sup_{\mathbf{x},\mathbf{y} \in \mathcal{T}} \Big| \sum_{(i,j) \in \mathcal{L}(\mathbf{x},\mathbf{y})} x_i y_j [\widehat{\mathbf{P}}_{t,r}(i,j) - \mathbf{P}_{t,r}(i,j)] \Big| \geq C \sqrt{\frac{n \, C_\alpha \alpha_n}{r}} \Big\} \tag{7.26}$$
$$\leq 2 \, e^{-n \left( \frac{C^2}{4 \, W_{\max}^2 + 4C \, W_{\max}/3} - \log 14 \right)}.$$

If (7.26) holds, then $C = C_{\tau,1} W_{\max}$ with (7.10) guarantees that the right-hand side in (7.26) is bounded by $2n^{-\tau}$, and Lemma 6 is proved.

To prove (7.26), denote $u_{ij} = x_i y_j \, \mathbb{I}(|x_i y_j| \leq \sqrt{C_\alpha \alpha_n r/n}) + x_j y_i I(|x_j y_i| \leq \sqrt{C_\alpha \alpha_n r/n})$. Consider

$$\begin{aligned} S &= \sum_{(i,j) \in \mathcal{L}(\mathbf{x},\mathbf{y})} x_i y_j \, (\widehat{\mathbf{P}}_{t,r}(i,j) - \mathbf{P}_{t,r}(i,j)) \\ &= \frac{1}{|\mathcal{F}_r|} \sum_{1 \leq i < j \leq n} \sum_{k \in \mathcal{F}_r} u_{ij} \, W_{r,l}(k) (\mathbf{A}_{t+k}(i,j) - \mathbf{P}_{t+k}(i,j)). \end{aligned}$$

Note that the right-hand side is the sum of $n(n-1)|\mathcal{F}_r|$ independent variables

$$\xi_{i,j,k} = u_{ij} W_{r,l}(k) (\mathbf{A}_{t+k}(i,j) - \mathbf{P}_{t+k}(i,j))/|\mathcal{F}_r|$$

with zero means and absolute values bounded by $|\xi_{i,j,k}| \le 2W_{\max}\sqrt{C_\alpha\alpha_n/n\,|\mathcal{F}_r|}$, due to $|u_{ij}| \le 2\sqrt{C_\alpha\alpha_n r/n}$ and $|\mathbf{A}_{t+k}(i,j) - \mathbf{P}_{t+k}(i,j)| \le 1$. Applying Bernstein's inequality and using the fact that $\sum_{i<j} u_{ij}^2 \le 2$ (as proved in the end of Section 3 in the Supplementary material of [24]), obtain

$$\begin{aligned}\Pr\left\{\left|\sum_{1\le i<j\le n}\sum_{k\in\mathcal{F}_r}\xi_{i,j,k}\right| \ge C\sqrt{\frac{n\,C_\alpha\alpha_n}{r}}\right\} &\le 2\exp\left(-\frac{\frac{C^2 nC_\alpha\alpha_n}{2r}}{\frac{2\,C_\alpha W_{\max}^2\alpha_n}{|\mathcal{F}_r|} + \frac{2W_{\max}}{3}\sqrt{\frac{C_\alpha\alpha_n}{n|\mathcal{F}_r|}}\,C\sqrt{\frac{nC_\alpha\alpha_n}{r}}}\right)\\ &\le 2\exp\left(-\frac{C^2 n}{4\,W_{\max}^2 + \frac{4C}{3}W_{\max}}\right),\end{aligned}$$

since $|\mathcal{F}_r| \ge 1+r$. Using the fact that cardinality $|\mathcal{T}| \le \exp(n\log 14)$ (see Section 3 in the Supplementary material of [24] with $\delta = 1/2$), (7.26) is proved.

**Proof of Lemma 7.** First, it is sufficient to prove that for any $c_1 > 1$, one has

$$\Pr\left\{\max_{1\le i\le n} d_{t,r}(i) \le c_1\,C_\alpha\,n\,\alpha_n\,r\right\} \ge 1 - n^{1-\frac{3(c_1-1)^2c_0\,r}{6\,W_{\max}^2+2W_{\max}(c_1-1)}}. \tag{7.27}$$

If (7.27), then $c_1 = 3(W_{\max}C_{\tau,2}+1)$, the inequality (7.12) and $\max(r,2) \le |\mathcal{F}_r| \le 3r$ for $r \ge 1$ guarantees that the right hand side of (7.27) is bounded below by $1 - n^{-\tau}$, and Lemma 7 is proved.

For a fixed node $i$, using Bernstein's inequality and $C_\alpha\alpha_n n \ge c_0\log n$, obtain

$$\begin{aligned}&\Pr(d_{t,r}(i) > c_1 nC_\alpha\alpha_n|\mathcal{F}_r|)\\ \le&\Pr\Big(\sum_{j=1}^n\sum_{k\in\mathcal{F}_r} W_{r,l}(k)[\mathbf{A}_{t+k}(i,j) - \mathbf{P}_{t+k}(i,j)] \le (c_1-1)C_\alpha n\alpha_n|\mathcal{F}_r|\Big)\\ \le&\exp\Big(-\frac{\frac{1}{2}(c_1-1)^2C_\alpha^2\,n^2\alpha_n^2|\mathcal{F}_r|^2}{C_\alpha|\mathcal{F}_r|W_{\max}^2 n\alpha_n + \frac{1}{3}W_{\max}(c_1-1)C_\alpha n\alpha_n|\mathcal{F}_r|}\Big) \le n^{-\frac{3\,(c_1-1)^2c_0|\mathcal{F}_r|}{6W_{\max}^2+2W_{\max}(c_1-1)}}.\end{aligned}$$

Taking the union bound over $i = 1,\cdots,n$, (7.27) is proved.

**Proof of Lemma 8.** First, if we divide the weights $W_{r,l}(k)$ into two groups: $\mathcal{K}_1 = \{k \in \mathcal{F}_r : W_{r,l}(k) > 0\}$ and $\mathcal{K}_2 = \{k \in \mathcal{F}_r : W_{r,l}(k) \le 0\}$. Define

$$Y_{ijk} = I(\mathbf{A}_k(i,j) = 1)\cdot\mathbb{I}(k \in \mathcal{K}_1).$$

Then each $Y_{ijk}$ is a Bernoulli random variable with expectation $\mathbf{P}_k(i,j)\cdot I(k \in \mathcal{K}_1)$, and by definition

$$e_{t,r}(I,J) = \sum_{k\in\mathcal{F}_r} W_{r,l}(k)e_{t+k}(I,J) \le \sum_{k\in\mathcal{K}_1} W_{r,l}(k)e_{t+k}(I,J) \tag{7.28}$$

$$= \sum_{k\in\mathcal{F}_r}\sum_{i\in I}\sum_{j\in J} W_{r,l}(k)Y_{ijk}.$$

Applying Lemma 9 with $X_i$ replaced by $Y_{ijk} - \mathbb{E}Y_{ijk}$, $w_{\max}$ replaced by $W_{\max}$, $p_{\max}$ replaced by $\alpha_n$, $k$ replaced by $t$ and $n = |\mathcal{F}_r||I||J|$, obtain for $t \geq \max(e^{3W_{\max}}, 2)$:

$$\Pr\Big\{ \sum_{k\in\mathcal{F}_r, i\in I, j\in J} W_{r,l}(k)[Y_{ijk} - \mathbb{E}Y_{ijk}] > t\ \bar{\mu}(I,J) \Big\} \leq \exp\left[ -\frac{(t+1)\ln(t+1)C_\alpha\alpha_n|\mathcal{F}_r||I||J|}{2W_{\max}} \right]. \tag{7.29}$$

Second, since $\mathbb{E}Y_{ijk} < C_\alpha\alpha_n$, application of (7.28) and (7.29) with $t \geq \max(e^{3W_{\max}}, 2) + W_{\max}$ yields

$$\begin{aligned}
&\Pr\{e_{t,r}(I,J) > t\ \bar{\mu}(I,J)\} \leq \Pr\left\{ \sum_{k\in\mathcal{F}_r}\sum_{i\in I}\sum_{j\in J} W_{r,l}(k)Y_{ijk} > t\ \bar{\mu}(I,J) \right\} \\
&\leq \Pr\left\{ \sum_{k\in\mathcal{F}_r}\sum_{i\in I}\sum_{j\in J} W_{r,l}(k)[Y_{ijk} - \mathbb{E}Y_{ijk}] > (t - W_{\max})\bar{\mu}(I,J) \right\} \qquad (7.30) \\
&\leq \exp\Big[ -\frac{(t+1-W_{\max})\ln(t+1-W_{\max})C_\alpha\alpha_n|\mathcal{F}_r||I||J|}{2W_{\max}} \Big] \\
&= \exp\Big[ -\frac{(t+1-W_{\max})\ln(t+1-W_{\max})\bar{\mu}(I,J)}{2W_{\max}} \Big].
\end{aligned}$$

When $s > \max(a,2)$, $(s+a)\ln(s+a) \leq (s+a)[\ln s + \ln\max(a,2)] \leq 4s\ln s$. Setting $a = W_{\max} - 1$ and $s = t + 1 - W_{\max}$, we have that when $t + 1 - W_{\max} > \max(W_{\max} - 1, 2)$,

$$\frac{(t+1-W_{\max})\ln(t+1-W_{\max})\bar{\mu}(I,J)}{2W_{\max}} \geq \frac{t\ln t\bar{\mu}(I,J)}{8W_{\max}}. \tag{7.31}$$

If $t > C_{\tau,5}$ with $C_{\tau,5} = \max(e^{3W_{\max}}, W_{\max} - 1, 2) + W_{\max}$, then (7.31) holds and (7.30) implies

$$\Pr\{e_{t,r}(I,J) > t\bar{\mu}(I,J)\} \leq \exp\Big[ -\frac{t\ \ln t\ \bar{\mu}(I,J)}{8W_{\max}} \Big].$$

The rest of the proof repeats the proof of Lemma 4.2 in [24] (start from the fourth paragraph in Section 4.2, note that the constant 8 is replaced by $C_{\tau,5}$, $\frac{1}{2}$ in the exponent is replaced by $\frac{1}{8W_{\max}}$, $c$, $c_1$, $c_2$ and $c_3$ are replaced, respectively, by $\tau$, $3(W_{\max}C_{\tau,2} + 1)$, $C_{\tau,3}$ and $C_{\tau,4}$). In particular, by following their calculation, we can show that if $C_{\tau,3} = \max\{3(W_{\max}C_{\tau,2} + 1), C_{\tau,5}\}$ and $C_{\tau,4}$ is chosen so that $C_{\tau,4}/(8W_{\max}) - 6 = \tau$, that is $C_{\tau,4} = 8W_{\max}(\tau + 6)$, then Lemma 8 is valid. To complete the proof, note that $\max(W_{\max} - 1, 2) \leq 3(W_{\max}C_{\tau,2} + 1)$.

**Proof of Lemma 9.** By definition, $E(e^{\lambda X_i}) = p_i e^{w_i(1-p_i)\lambda} + (1-p_i)e^{-w_i p_i \lambda}$. Following the same proof as Lemma A.1.8 in [2], we have

$$E(e^{\lambda X}) \le e^{-\sum_{i=1}^n w_i p_i \lambda}[pe^{w_{\max}\lambda} + (1-p)]^n. \tag{7.32}$$

Let $a$ be any positive real number and $\lambda = \frac{1}{w_{\max}} \ln[1 + a/pn]$, then using $pe^{w_{\max}\lambda} + (1-p) = 1 + a/n$ and $(1+a/n)^n \le e^a$, the right hand side of (7.32) is bounded by

$$e^{a - \frac{1}{w_{\max}} \sum_{i=1}^n w_i p_i \ln[1+a/pn]}.$$

The Chernoff bound then implies that

$$\Pr(X \ge a) < e^{-a\lambda} E(e^{\lambda X}) \le e^{a - (\frac{1}{w_{\max}} \sum_{i=1}^n w_i p_i + a) \ln[1+a/pn]}.$$

Let $a = kp_{\max}n$, Lemma 9 is proved as follows:

$$\begin{aligned}
&\Pr(X \ge kp_{\max}n) < e^{kp_{\max}n - (\frac{1}{w_{\max}} \sum_{i=1}^n w_i p_i + kp_{\max}n) \ln[1+k]} \\
&< e^{\frac{1}{w_{\max}} kp_{\max}n(1-\ln(k+1))} \\
&< e^{-\frac{1}{2w_{\max}}(k+1)p_{\max}n \ln(k+1)},
\end{aligned}$$

where the last inequality holds when $k > \max(e^{3w_{\max}}, 2)$.

## References


[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[2] N. Alon and J. H. Spencer. *The Probabilistic Method*. John Wiley & Sons, Inc., 2008. MR2437651

[3] E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Ann. Statist.*, 42(3):940–969, 06 2014. MR3210992

[4] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer, New York, 1997. MR1477662

[5] S. Bhattacharyya and S. Chatterjee. Spectral Clustering for Multiple Sparse Networks: I. *ArXiv e-prints*, May 2018.

[6] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[7] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. pages 153–162. ACM Press, August 2007.

[8] D. Choi, P. J. Wolfe, et al. Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42(1):29–63, 2014. MR3161460

[9] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005. MR2155709

[10] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, STOC '89, pages 587–598, New York, NY, USA, 1989. ACM.

[11] W. Fu, L. Song, and E. P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 329–336, New York, NY, USA, 2009. ACM.

[12] C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652, 12 2015. MR3405606

[13] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015. MR3687603

[14] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[15] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Academic press, 2014. MR0669666

[16] Q. Han, K. S. Xu, and E. M. Airoldi. Consistent estimation of dynamic and multi-layer block models. In *ICML*, pages 1511–1520, 2015.

[17] T. Herlau, M. Mørup, and M. N. Schmidt. Modeling temporal evolution and multiscale structure in networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–960–III–968. JMLR.org, 2013.

[18] J. Jin et al. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015. MR3285600

[19] A. Joseph and B. Yu. Impact of regularization on spectral clustering. *Ann. Statist.*, 44(4):1765–1791, 08 2016. MR3519940

[20] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*, 2015. MR3611494

[21] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition, 2009. MR2724362

[22] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time (1 + epsiv;)-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462, Oct 2004. MR2606080

[23] C. M. Le and E. Levina. Estimating the number of communities in networks by spectral methods. *ArXiv e-prints*, July 2015.

[24] J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015. MR3285605

[25] O. V. Lepski. Asymptotic mimimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36:654–659, 1991. MR1147167

[26] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates

with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 06 1997. MR1447734

[27] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007. MR2409803

[28] C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. MR3689311

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, Cambridge, MA, USA, 2001. MIT Press.

[30] S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

[31] M. Pensky. Dynamic network models and graphon estimation. *arXiv preprint arXiv:1607.00673*, 2016.

[32] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 08 2011. MR2893856

[33] P. Sarkar and P. J. Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.*, 43(3):962–990, 06 2015. MR3346694

[34] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[35] N. Verzelen and E. Arias-Castro. Community detection in sparse random networks. *Ann. Appl. Probab.*, 25(6):3465–3510, 12 2015. MR3404642

[36] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. MR2409803

[37] E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.*, 4(2):535–566, 06 2010. MR2758639

[38] K. Xu. Stochastic block transition models for dynamic networks. In *AISTATS*, 2015.

[39] K. S. Xu and A. O. Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

[40] K. S. Xu, M. Kliger, and A. O. Hero III. Adaptive evolutionary clustering. *Data Mining and Knowledge Discovery*, 28(2):304–336, 2014. MR3147571

[41] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning*, 82(2):157–189, 2011. MR3108191

[42] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100(3):217–240, Dec 2012. MR2979307

[43] X. Zhang, C. Moore, and M. E. J. Newman. Random graph models for dynamic networks. *CoRR*, abs/1607.07570, 2016.

[44] Y. Zhao, E. Levina, and J. Zhu. Community extraction for social networks.

*Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

[45] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292, 08 2012. MR3059083