Journal of Machine Learning Research 24 (2023) 1-69 Submitted 7/22; Revised 3/23; Published 6/23

Autoregressive Networks

Binyan Jiang BY.JIANG@POLYU.EDU.HK
Department of Applied Mathematics
Hong Kong Polytechnic University
Hong Kong, China

Jialiang Li STALJ@NUS.EDU.SG
Department of Statistics and Applied Probability
National University of Singapore
Singapore 117546

Qiwei Yao Q.YAO@LSE.AC.UK
Department of Statistics
London School of Economics
London, WC2A 2AE

Editor: Qiaozhu Mei

Abstract

We propose a first-order autoregressive (i.e. AR(1)) model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference methods including a permutation test for diagnostic checking for the fitted network models. The proposed model can be applied to the network processes with various underlying structures but with independent edges. As an illustration, an AR(1) stochastic block model has been investigated in depth, which characterizes the latent communities by the transition probabilities over time. This leads to a new and more effective spectral clustering algorithm for identifying the latent communities. We have derived a finite sample condition under which the perfect recovery of the community structure can be achieved by the newly defined spectral clustering algorithm. Furthermore the inference for a change point is incorporated into the AR(1) stochastic block model to cater for possible structure changes. We have derived the explicit error rates for the maximum likelihood estimator of the change-point. Application with three real data sets illustrates both relevance and usefulness of the proposed AR(1) models and the associate inference methods.

Keywords: AR(1) networks; Change point; Dynamic stochastic block model; Hamming distance; Maximum likelihood estimation; Spectral clustering algorithm; Yule-Walker equation.

1. Introduction

Understanding and being able to model the network changes over time are of immense importance for, e.g., monitoring anomalies in internet traffic networks, predicting demand and setting prices in electricity supply networks, managing natural resources in environmental readings in sensor networks, and understanding how news and opinion propagates in online social networks. In spite of the existence of a large body of literature on dynamic networks, the development of the foundation for dynamic network models is still in its infancy (Kolaczyk, 2017). As for dealing with dynamic changes of networks, early attempts are based on the evolution analysis of network snapshots over time (Aggarwal and Subbian, 2014; Donnat and Holmes, 2018). Although this reflects the fact that most networks change slowly over time, it provides little insight on the dynamics underlying the changes and is almost powerless for future prediction. The popular approaches for modelling dynamic changes include, among others, Markov process models (Snijders, 2005; Ludkin et al., 2018), the exponential random graph models (Hanneke et al., 2010; Krivitsky and Handcock, 2014), and latent process based models (Friel et al., 2016; Durante et al., 2016; Matias and Miele, 2017). The estimation for those models is compute-intensive, relying on various MCMC or EM algorithms.

In this paper we propose a simple first-order autoregressive (i.e. AR(1)) model for dynamic network processes of which the edges changes over time while the nodes are unchanged. Though our setting is a special case of the Markov chain network models (see Yudovina et al. (2015), and also Snijders (2005) and Ludkin et al. (2018)), a simple AR(1) structure makes it possible to measure explicitly the underlying dynamic properties such as autocorrelation coefficients, and the Hamming distance. It facilitates the maximum likelihood estimation (MLE) in a simple and direct manner with uniform error rates. Furthermore diagnostic checking for the fitted network models can be performed in terms of an easy-to-use permutation test, which is impossible under a merely Markovian structure.

Our setting can be applied to any network processes with various underlying structures as long as the edges are independent with each other, which we illustrate through an AR(1) stochastic block model. The latent communities in our setting are characterized by the transition probabilities over time, instead of the (static) connection probabilities – the approach often adopted from static stochastic block models; see Pensky (2019) and the references therein. This new structure also paves the way for a new spectral clustering algorithm which identifies the latent communities more effectively – a phenomenon corroborated by both the asymptotic theory and the simulation results. To cater for possible structure changes of underlying processes, we incorporate a change point detection mechanism in the AR(1) stochastic block modeling. Again the change point is estimated by the maximum likelihood method. The AR(1) continuous time stochastic block model of Ludkin et al. (2018) is based on a sophisticated construction. Its estimation is based on a reversible jump MCMC, though a discrete-time version of their model admits the same Markov Chain representation (3) below.

Theoretical developments for dynamic stochastic block models in the literature were typically based on the assumption that networks observed at different times are independent; see Pensky (2019); Bhattacharjee et al. (2020) and references therein. The autoregressive structure considered in this paper brings the extra complexity due to serial dependence. By establishing the α-mixing property with exponentially decaying coefficients for the AR(1) network processes, we are able to show that the proposed spectral clustering algorithm leads to a consistent recovery of the latent community structure. On the other hand, an extra challenge in detecting a change point in the dynamic stochastic block network process is that the estimation for latent community structures before and after a possible change point is typically not consistent during the search for the change point. To overcome this obstacle, we introduce a truncation technique which breaks the searching interval into two parts such that the error bounds for the estimated change point can be established.

The proposed methods in this paper only apply to the dynamic networks observed on discrete times. Even so the relevant literature is large, across mathematics, computer science, engineer, statistics, biology, genetics and social sciences. We can only list a small selection of more statistics-oriented papers in addition to the aforementioned references. Fu et al. (2009) proposed a state space mixed membership stochastic block model (with a logistic normal prior). Crane et al. (2016) studied the limit properties of Markovian, exchangeable and càdlàg (i.e. every edge remains in each state which it visits for a positive amount of time) dynamic network. Pensky (2019) studied the theoretical properties (such as the minimax lower bounds for the risk) of a dynamic stochastic block model, assuming 'smooth' connectivity probabilities. The literature on change point detection in dynamic networks includes Yang et al. (2011); Wang et al. (2021); Wilson et al. (2019); Zhao et al. (2019); Bhattacharjee et al. (2020); Zhu et al. (2020a). Knight et al. (2016); Zhu et al. (2017, 2019); Chen et al. (2023); Zhu et al. (2020b) adopted autoregressive models for modelling continuous responses observed from the nodes of a network process. Kang et al. (2020) used dynamic network as a tool to model non-stationary vector autoregressive processes. For the development on continuous-time dynamic networks, we refer readers to Snijders (2005), Matias et al. (2018), Ludkin et al. (2018) and Corneli et al. (2018).

The new contributions of this paper include: (i) We propose a new and simple AR(1) model for edge dymanics (see (1) below), which facilitates the easy-to-use inference methods including a permutation test for model diagnostic checking. (ii) The AR(1) setting can be applied to various network processes with specific underlying structures such as dynamic stochastic block models, as illustrated in Section 3 below, and also dynamic dot product model, dynamic graphon model, etc. (iii) The AR(1) structure also makes it possible to develop the theoretical guarantees for the serial dependent network processes. For example, based on a concentration inequality, we have derived a finite sample condition, under which the perfect recovery of the community structure can be achieved by the newly defined spectral clustering algorithm (Theorems 10 and 11 in Section 3.2.2 below). Furthermore, we have shown that the MLE for the change-point in the AR(1) stochastic block process is consistent with explicit error rates (Theorem 14 in Section 3.3 below). Those results are based on some rigorous technical development for the dependent network processes. Note that both Pensky (2019) and Bhattacharjee et al. (2020) assume that networks observed at different times are independent with each other in their asymptotic theories for dynamic stochastic block models. Illustration with the three real network data sets indicates convincingly that the proposed AR(1) model and the associated inference methods are practically relevant and fruitful. In particular, the analysis of the contact network data in a French hospital reveals the different interaction patterns during days and nights. Our proposed AR(1) stochastic block model exhibits clear and meaningful communities in a school social network data. The global trade networks, observed over the recent 6 decades, present some dynamic changes in trade activities across different countries, and our analysis discovers to the discovery of a remarkable change point in 1991, corresponding to the end of the Cold War.

The rest of the paper is organized as follows. A general framework of AR(1) network processes, the probabilistic properties, and the MLE are presented in Section 2. It also contains a new and easy-to-use permutation test for the diagnostic checking for the fitted network models. Section 3 deals with AR(1) stochastic block models. The asymptotic theory is developed for the new spectral clustering algorithm based on the transition probabilities. Further extension of both the inference method and the asymptotic theory to the setting with a change point is established. Simulation results are reported in Section 4, and the illustration with three real dynamic network data sets is presented in Section 5. All technical proofs are relegated to the Appendix.

2. Autoregressive network models

2.1 AR(1) models

Let $\{\mathbf{X}_t, t = 0, 1, 2, \cdots\}$ be a dynamic network process defined on the p fixed nodes, denoted by $\{1, \cdots, p\}$, where $\mathbf{X}_t \equiv (X_{i,j}^t)$ denotes the $p \times p$ adjacency matrix at time t. We also assume that all networks are Erdös-Renyi in the sense that $X_{i,j}^t$, $(i,j) \in \mathcal{J}$, are independent and take values either 1 or 0, where $\mathcal{J} = \{(i,j) : 1 \le i \le j \le p\}$ for undirected networks, $\mathcal{J} = \{(i,j) : 1 \le i < j \le p\}$ for undirected networks without selfloops, $\mathcal{J} = \{(i,j) : 1 \le i, j \le p\}$ for directed networks, and $\mathcal{J} = \{(i,j) : 1 \le i \ne j \le p\}$ for directed networks without selfloops. Note that an edge from node i to j is indicated by $X_{i,j} = 1$, and no edge is denoted by $X_{i,j} = 0$. For undirected networks, $X_{i,j}^t = X_{j,i}^t$.

Definition 1. *An AR(1) network process is defined as*

$$X_{i,j}^t \;=\; X_{i,j}^{t-1}\, I(\varepsilon_{i,j}^t = 0) \;+\; I(\varepsilon_{i,j}^t = 1), \quad t \ge 1, \tag{1}$$

where $I(\cdot)$ denotes the indicator function, the innovations $\varepsilon_{i,j}^t$, $(i,j) \in \mathcal{J}$, are independent, and

$$P(\varepsilon_{i,j}^t = 1) = \alpha_{i,j}^t, \quad P(\varepsilon_{i,j}^t = -1) = \beta_{i,j}^t, \quad P(\varepsilon_{i,j}^t = 0) = 1 - \alpha_{i,j}^t - \beta_{i,j}^t. \tag{2}$$

In the above expression, $\alpha_{i,j}^t, \beta_{i,j}^t$ are non-negative constants, and $\alpha_{i,j}^t + \beta_{i,j}^t \le 1$.

Equation (1) is an analogue of the noisy network model of Chang et al. (2022). The innovation (or noise) $\varepsilon_{i,j}^t$ is ‘added’ via the two indicator functions to ensure that $X_{i,j}^t$ is still binary. Obviously, $\{\mathbf{X}_t, t = 0, 1, 2, \cdots\}$ is a Markov chain, and

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0) = \alpha_{i,j}^t, \quad P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1) = \beta_{i,j}^t, \tag{3}$$

or collectively,

$$P(\mathbf{X}_t|\mathbf{X}_{t-1}, \cdots, \mathbf{X}_0) \;=\; P(\mathbf{X}_t|\mathbf{X}_{t-1}) \;=\; \prod_{(i,j)\in\mathcal{J}} P(X_{i,j}^t | X_{i,j}^{t-1}) \tag{4}$$

$$= \prod_{(i,j)\in\mathcal{J}} (\alpha_{i,j}^t)^{X_{i,j}^t(1-X_{i,j}^{t-1})} (1-\alpha_{i,j}^t)^{(1-X_{i,j}^t)(1-X_{i,j}^{t-1})} (\beta_{i,j}^t)^{(1-X_{i,j}^t)X_{i,j}^{t-1}} (1-\beta_{i,j}^t)^{X_{i,j}^t X_{i,j}^{t-1}}.$$

It is clear that the smaller $\alpha_{i,j}^t$ is, the more likely the no-edge status at time $t-1$ (i.e. $X_{i,j}^{t-1} = 0$) will be retained at time t (i.e. $X_{i,j}^t = 0$); and the smaller $\beta_{i,j}^t$ is, the more likely an edge at time $t-1$ (i.e. $X_{i,j}^{t-1}=1$) will be retained at time t (i.e. $X_{i,j}^{t}=1$). For most slowly changing networks (such as social networks), we expect $\alpha_{i,j}^t$ and $\beta_{i,j}^t$ to be small.

It is natural to model dynamic networks by a Markov chain. See, e.g. Hanneke et al. (2010); Krivitsky and Handcock (2014); Xu (2015); Yudovina et al. (2015); Friel et al. (2016); Crane et al. (2016); Matias and Miele (2017); Rastelli et al. (2018); Ludkin et al. (2018). For example, the Markovian transition probabilities under a discrete version of the stationary independent arcs network model of Snijders (2005, Section 5) can be written equivalently as (3) with $\alpha_{i,j}^t \equiv \alpha$ and $\beta_{i,j}^t \equiv \beta$. In this paper we build the Markovian structure based on the explicit AR(1) model (1), which enables us to study the theoretical properties of the network processes, and to develop simple and efficient inference methods with appropriate theoretical guarantee.

2.2 Stationarity

Note that $\{\mathbf{X}_t\}$ is a homogeneous Markov chain if

$$\alpha_{i,j}^t \equiv \alpha_{i,j} \quad \text{and} \quad \beta_{i,j}^t \equiv \beta_{i,j} \quad \text{for all } t \geq 1 \ \text{ and } \ (i,j) \in \mathcal{J}. \tag{5}$$

Specify the distribution of the initial network $\mathbf{X}_0 = (X_{i,j}^0)$ as follows:

$$P(X_{i,j}^0 = 1) = \pi_{i,j} = 1 - P(X_{i,j}^0 = 0), \tag{6}$$

where $\pi_{i,j} \in (0,1)$, $(i,j) \in \mathcal{J}$, are constants.

Proposition 2. *Let the homogeneity condition (5) hold with $\alpha_{i,j} + \beta_{i,j} \in (0,1]$, and*

$$\pi_{i,j} = \alpha_{i,j}/(\alpha_{i,j} + \beta_{i,j}), \ \ (i,j) \in \mathcal{J}. \tag{7}$$

Then $\{\mathbf{X}_t, t = 0,1,2,\cdots\}$ is a strictly stationary process. Furthermore for any $(i,j),(\ell,m) \in \mathcal{J}$ and $t,s \geq 0$,

$$E(X_{i,j}^t) = \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}, \qquad \text{Var}(X_{i,j}^t) = \frac{\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}, \tag{8}$$

$$\rho_{i,j}(|t-s|) \equiv \text{Corr}(X_{i,j}^t, X_{\ell m}^s) = \begin{cases} (1-\alpha_{i,j}-\beta_{i,j})^{|t-s|} & \text{if } (i,j) = (\ell,m), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The Hamming distance counts the number of different edges in the two networks, and is a measure the closeness of two networks (Donnat and Holmes, 2018).

Definition 3. *For any two matrices $\mathbf{A} = (A_{i,j})$ and $\mathbf{B} = (B_{i,j})$ of the same size, the Hamming distance is defined as $D_H(\mathbf{A},\mathbf{B}) = \sum_{i,j} I(A_{i,j} \neq B_{i,j})$.*

Proposition 4. *Let $\{\mathbf{X}_t, t = 0,1,\cdots\}$ be a stationary network process satisfying the condition of Proposition 2. Let $d_H(|t-s|) = E\{D_H(\mathbf{X}_t,\mathbf{X}_s)\}$ for any $t,s \geq 0$. Then $d_H(0) = 0$, and it holds for any $k \geq 1$ that*

$$d_H(k) = d_H(k-1) + \sum_{(i,j)\in\mathcal{J}} \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}(1-\alpha_{i,j}-\beta_{i,j})^{k-1} \tag{10}$$

$$= \sum_{(i,j)\in\mathcal{J}} \frac{2\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}\{1-(1-\alpha_{i,j}-\beta_{i,j})^k\}. \tag{11}$$

Proposition 4 indicates that the expected Hamming distance $d_H(d) = E\{D_H(\mathbf{X}_t, \mathbf{X}_{t+k})\}$ increases strictly, as k increases, initially from $d_H(1) = \sum \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}$ towards the limit $d_H(\infty) = \sum \frac{2\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}$ which is also the expected Hamming distance of the two independent networks sharing the same marginal distribution of $\mathbf{X}_t$.

Proposition 5 below shows that $\{\mathbf{X}_t,\, t = 0, 1, \cdots\}$ is α-mixing with exponentially decaying coefficients. Note that the conventional mixing results for ARMA processes do not apply here, as they typically require that the innovation distribution is continuous; see, e.g., Section 2.6.1 of Fan and Yao (2003). Let $\mathcal{F}_a^b$ be the σ-algebra generated by $\{X_{i,j}^k, a \le k \le b\}$. The α-mixing coefficient of process $\{X_{i,j}^t, t = 0, 1, \cdots\}$ is defined as

$$\alpha^{i,j}(\tau) = \sup_{k\in\mathbb{N}} \sup_{A\in\mathcal{F}_0^k, B\in\mathcal{F}_{k+\tau}^\infty,} |P(A\cap B) - P(A)P(B)|.$$

Proposition 5. *Let condition (5) hold,* $\alpha_{i,j}, \beta_{i,j} > 0$, *and* $\alpha_{i,j} + \beta_{i,j} \le 1$. *Then* $\alpha^{i,j}(\tau) \le \rho_{i,j}(\tau) = (1 - \alpha_{i,j} - \beta_{i,j})^\tau$ *for any* $\tau \ge 1$.

2.3 Estimation

To simplify the notation, we assume the availability of the observations $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_n$ from a stationary network process which satisfies the condition of Proposition 2. Without imposing any further structure on the model, the parameters $(\alpha_{i,j}, \beta_{i,j})$, for different (i, j), can be estimated separately. Conditionally on $\mathbf{X}_0$, the joint probability function of $\mathbf{X}_1, \cdots, \mathbf{X}_n$ is $\prod_{1\le t\le n} P(\mathbf{X}_t|\mathbf{X}_{t-1})$. It then follows from (4) that the maximum likelihood estimators are

$$\widehat{\alpha}_{i,j} = \frac{\sum_{t=1}^n X_{i,j}^t(1 - X_{i,j}^{t-1})}{\sum_{t=1}^n (1 - X_{i,j}^{t-1})}, \qquad \widehat{\beta}_{i,j} = \frac{\sum_{t=1}^n (1 - X_{i,j}^t) X_{i,j}^{t-1}}{\sum_{t=1}^n X_{i,j}^{t-1}}. \tag{12}$$

For definiteness we shall set $0/0 = 1$. To state the asymptotic properties, we list some regularity conditions first.

C1. There exists a constant $l > 0$ such that $\alpha_{i,j}, \beta_{i,j} \ge l$ and $\alpha_{i,j}+\beta_{i,j} \le 1$ for all $(i, j) \in \mathcal{J}$.

C2. $n, p \to \infty$, and $(\log n)(\log\log n)\sqrt{\frac{\log p}{n}} \to 0$.

Condition C1 defines the parameter space, and condition C2 indicates that the number of nodes is allowed to diverge in a smaller order than $\exp\left\{\frac{n}{(\log n)^2(\log\log n)^2}\right\}$.

Proposition 6. *Let conditions (5), C1 and C2 hold. For any constant* $c > 2$, *there exists a large enough constant* $C > 0$ *such that,*

$$P\left(\max_{(i,j)\in\mathcal{J}} |\widehat{\alpha}_{i,j} - \alpha_{i,j}| \ge l^{-1}C\sqrt{\frac{\log p}{n}}\right) \ge 2p^2 \exp\{-c\log p\},$$
$$P\left(\max_{(i,j)\in\mathcal{J}} |\widehat{\beta}_{i,j} - \beta_{i,j}| \ge l^{-1}C\sqrt{\frac{\log p}{n}}\right) \ge 2p^2 \exp\{-c\log p\}.$$

Consequently, as $n, p \to \infty$, we have:

$$\max_{(i,j)\in\mathcal{J}} |\widehat{\alpha}_{i,j} - \alpha_{i,j}| = O_p\left(\sqrt{\frac{\log p}{n}}\right) \quad \text{and} \quad \max_{(i,j)\in\mathcal{J}} |\widehat{\beta}_{i,j} - \beta_{i,j}| = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Proposition 6 provides a uniform convergence rate for the MLEs in (12). Proposition 7 below implies that any fixed number of estimators $\{\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}\}$ are jointly asymptotically normal; noting that p diverges together with n. To state this joint asymptotic normality, we introduce some notation first: let $J_1 = \{(i_1, j_1), \ldots, (i_{m_1}, j_{m_1})\}$, $J_2 = \{(k_1, \ell_1), \ldots, (k_{m_2}, \ell_{m_2})\}$ be two arbitrary subsets of $\mathcal{J}$ with $m_1, m_2 \geq 1$ fixed. Denote $\mathbf{\Theta}_{J_1,J_2} = (\alpha_{i_1,j_1}, \ldots, \alpha_{i_{m_1},j_{m_1}}, \beta_{k_1,\ell_1}, \ldots, \beta_{k_{m_2},\ell_{m_2}})^\top$, and correspondingly denote the MLEs as $\widehat{\mathbf{\Theta}}_{J_1,J_2} = (\widehat{\alpha}_{i_1,j_1}, \ldots, \widehat{\alpha}_{i_{m_1},j_{m_1}}, \widehat{\beta}_{k_1,\ell_1}, \ldots, \widehat{\beta}_{k_{m_2},\ell_{m_2}})^\top$.

Proposition 7. *Let conditions (5), C1 and C2 hold. Then* $\sqrt{n}(\widehat{\mathbf{\Theta}}_{J_1,J_2} - \mathbf{\Theta}_{J_1,J_2}) \to N(\mathbf{0}, \mathbf{\Sigma}_{J_1,J_2})$, *where* $\mathbf{\Sigma}_{J_1,J_2} = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{m_1+m_2,m_1+m_2})$ *is a diagonal matrix with*

$$\sigma_{rr} = \frac{\alpha_{i_r,j_r}(1-\alpha_{i_r,j_r})(\alpha_{i_r,j_r}+\beta_{i_r,j_r})}{\beta_{i_r,j_r}}, \quad 1 \leq r \leq m_1,$$
$$\sigma_{rr} = \frac{\beta_{k_r,\ell_r}(1-\beta_{k_r,\ell_r})(\alpha_{k_r,\ell_r}+\beta_{k_r,\ell_r})}{\alpha_{k_r,\ell_r}}, \quad m_1+1 \leq r \leq m_1+m_2.$$

2.4 Model diagnostic check

Based on estimators in (12), we define 'residual' $\widehat{\varepsilon}_{i,j}^t$, resulted from fitting model (1) to the data, as the estimated value of $E(\varepsilon_{i,j}^t | X_{i,j}^t, X_{i,j}^{t-1})$, i.e.

$$\begin{aligned}\widehat{\varepsilon}_{i,j}^t =& \frac{\widehat{\alpha}_{i,j}}{1-\widehat{\beta}_{i,j}} I(X_{i,j}^t = 1, X_{i,j}^{t-1} = 1) - \frac{\widehat{\beta}_{i,j}}{1-\widehat{\alpha}_{i,j}} I(X_{i,j}^t = 0, X_{i,j}^{t-1} = 0) \\ &+ I(X_{i,j}^t = 1, X_{i,j}^{t-1} = 0) - I(X_{i,j}^t = 0, X_{i,j}^{t-1} = 1), \quad (i,j) \in \mathcal{J},\ t = 1, \cdots, n.\end{aligned}$$

One way to check the adequacy of the model is to test for the independence of $\widehat{\mathbf{\mathcal{E}}}_t \equiv (\widehat{\varepsilon}_{i,j}^t)$ for $t = 1, \cdots, n$. Since $\widehat{\varepsilon}_{i,j}^t$, $t = 1, \cdots, n$, only take 4 different values for each $(i,j) \in \mathcal{J}$, we adopt the two-way, or three-way contingency table to test the independence of $\widehat{\mathbf{\mathcal{E}}}_t$ and $\widehat{\mathbf{\mathcal{E}}}_{t-1}$, or $\widehat{\mathbf{\mathcal{E}}}_t, \widehat{\mathbf{\mathcal{E}}}_{t-1}$ and $\widehat{\mathbf{\mathcal{E}}}_{t-2}$. For example the test statistic for the two-way contingency table is

$$T = \frac{1}{n\,|\mathcal{J}|} \sum_{(i,j)\in\mathcal{J}} \sum_{k,\ell=1}^{4} \{n_{i,j}(k,\ell) - n_{i,j}(k,\cdot)n_{i,j}(\cdot,\ell)/(n-1)\}^2 / \{n_{i,j}(k,\cdot)n_{i,j}(\cdot,\ell)/(n-1)\}, \tag{13}$$

where $|\mathcal{J}|$ denotes the cardinality of $\mathcal{J}$, and for $1 \leq k, \ell \leq 4$,

$$n_{i,j}(k,\ell) = \sum_{t=2}^{n} I\{\widehat{\varepsilon}_{i,j}^t = u_{i,j}(k),\ \widehat{\varepsilon}_{i,j}^{t-1} = u_{i,j}(\ell)\},$$
$$n_{i,j}(k,\cdot) = \sum_{t=2}^{n} I\{\widehat{\varepsilon}_{i,j}^t = u_{i,j}(k)\}, \quad n_{i,j}(\cdot,\ell) = \sum_{t=2}^{n} I\{\widehat{\varepsilon}_{i,j}^{t-1} = u_{i,j}(\ell)\}.$$

In the above expressions, $u_{i,j}(1) = -1, u_{i,j}(2) = -\frac{\widehat{\beta}_{i,j}}{1-\widehat{\alpha}_{i,j}}, u_{i,j}(3) = \frac{\widehat{\alpha}_{i,j}}{1-\widehat{\beta}_{i,j}}$ and $u_{i,j}(4) = 1$. We calculate the P-values of the test T based on the following permutation algorithm:

1. Permute $\widehat{\mathcal{E}}_1, \cdots, \widehat{\mathcal{E}}_n$ to obtain a new sequence $\mathcal{E}_1^\star, \cdots, \mathcal{E}_n^\star$. Calculate the test statistic $T^\star$ in the same manner as T with $\{\widehat{\mathcal{E}}_t\}$ replaced by $\{\mathcal{E}_t^\star\}$.

2. Repeat 1 above M times, obtaining permutation test statistics $T_j^\star$, $j = 1, \cdots, M$, where $M > 0$ is a large integer. The P-value of the test (for rejecting the stationary AR(1) model) is then

$$\frac{1}{M} \sum_{j=1}^{M} I(T < T_j^\star).$$

3. Autoregressive stochastic block models

The general setting in Section 2 may apply to various network processes with some specific underlying structures as long as the edges are independent with each other. In this section we illustrate the idea with a new dynamic stochastic block (DSB) model.

3.1 Models

The DSB networks are undirected (i.e. $X_{i,j}^t \equiv X_{j,i}^t$) with no self-loops (i.e. $X_{i,i}^t \equiv 0$). Most available DSB models assume that the networks observed at different times are independent (Pensky, 2019; Bhattacharjee et al., 2020) or conditionally independent (Xu and Hero, 2014; Durante et al., 2016; Matias and Miele, 2017) as connection probabilities and node memberships evolve over time. We take a radically different approach as we impose autoregressive structure (1) in the network process. Furthermore, instead of assuming that the members in the same communities share the same (unconditional) connection probabilities, we entertain the idea that the transition probabilities (3) for the members in the same communities are the same. This reflects more directly the dynamic behavior of the process, and implies the former assumption on the unconditional connection probabilities under the stationarity. See (7). Furthermore, it is only natural under our AR(1) DSB model to identify the latent community structure using the information on both $\alpha_{i,j}$ and $\beta_{i,j}$, instead of that on $\pi_{i,j} = \alpha_{i,j}/(\alpha_{i,j} + \beta_{i,j})$ only. Both our theory (Theorem 10 and also Remark 3 below) and the numerical experiments (Section 4.2 below) confirm that the newly proposed spectral clustering algorithm based on $\alpha_{i,j}$ and $\beta_{i,j}$ provides more accurate estimation than that based on $\pi_{i,j}$ only.

Let ν_t be the membership function at time t, i.e. for any $1 \le i \le p$, $\nu_t(i)$ takes an integer value between 1 and q ($\le p$); indicating that node i belongs to the $\nu_t(i)$-th community at time t, where q is a fixed integer. This effectively assumes that the p nodes are divided into the q communities. We assume that q is fixed though some communities may contain no nodes at some times.

Definition 8. *An AR(1) stochastic block network process* $\{\mathbf{X}_t = (X_{i,j}^t),\ t = 0, 1, 2, \cdots\}$ *is defined by (1), where for* $1 \le i < j \le p$,

$$P(\varepsilon_{i,j}^t = 1) = \alpha_{i,j}^t = \theta_{\nu_t(i),\nu_t(j)}^t, \quad P(\varepsilon_{i,j}^t = -1) = \beta_{i,j}^t = \eta_{\nu_t(i),\nu_t(j)}^t, \tag{14}$$
$$P(\varepsilon_{i,j}^t = 0) = 1 - \alpha_{i,j}^t - \beta_{i,j}^t = 1 - \theta_{\nu_t(i),\nu_t(j)}^t - \eta_{\nu_t(i),\nu_t(j)}^t.$$

In the above expressions, $\theta_{k,\ell}^t, \eta_{k,\ell}^t$ *are non-negative constants, and* $\theta_{k,\ell}^t + \eta_{k,\ell}^t \le 1$ *for all* $1 \le k \le \ell \le q$.

The evolution of membership process ν_t and/or the connection probabilities was often assumed to be driven by some latent/Markov processes. The statistical inference for those models is carried out using computational methods such as MCMC, EM or extended Kalman filters. See, for example, Yang et al. (2011); Xu and Hero (2014); Xu (2015); Durante et al. (2016); Matias and Miele (2017); Rastelli et al. (2018). Bhattacharjee et al. (2020) adopted a change point approach: assuming both the membership and the connection probabilities remain constants either before or after a change point. See also Ludkin et al. (2018); Wilson et al. (2019). This reflects the fact that many networks (e.g. social networks) hardly change, and a sudden change is typically triggered by some external events.

We adopt a change point approach in this paper. Section 3.2 considers the estimation for both the community membership and transition probabilities when there are no change points in the process. This will serve as a building block for the inference with a change point in Section 3.3. Note that detecting change points in dynamic networks is a surging research area. In addition to the aforementioned references, more recent developments include Wang et al. (2021); Zhu et al. (2020a). Also note that the method of Zhao et al. (2019) can be applied to detect multiple change points for any dynamic networks.

3.2 Estimation without change points

We first consider a simple scenario of no change points in the observed period, i.e.

$$\nu_t(\cdot) \equiv \nu(\cdot) \quad \text{and} \quad (\theta_{k,\ell}^t, \eta_{k,\ell}^t) \equiv (\theta_{k,\ell}, \eta_{k,\ell}), \quad t = 1, \cdots, n, \ \ 1 \le k \le \ell \le q. \tag{15}$$

Then fitting the DSB model consists of two steps: (i) estimating $\nu(\cdot)$ to cluster the p nodes into q communities, and (ii) estimating transition probabilities $\theta_{k,\ell}$ and $\eta_{k,\ell}$ for $1 \le k \le \ell \le q$. To simplify the presentation, q is assumed to be known, which is the assumption taken by most papers on change point detection for DSB networks. In practice, one can determine q by, for example, the jittering method of Chang et al. (2020), or a Bayesian information criterion; see an example in Section 5.2 below. Below we first introduce a new Laplacian eigendecomposition which provides the underpinning for the proposed spectral clustering algorithm for estimating network memberships. With the estimated membership, the MLEs for $\theta_{k,\ell}$ and $\eta_{k,\ell}$ can be derived in the similar manner as in (12).

3.2.1 Laplacian eigendecomposition

The stochastic block model with p nodes and q communities can be parameterized by a pair of matrices $(\mathbf{Z}, \mathbf{\Omega})$, where $\mathbf{Z} = (z_{i,j}) \in \{0,1\}^{p \times q}$ is the membership matrix such that it has exactly one 1 in each row and at least one 1 in each column, and $\mathbf{\Omega} = (\omega_{k,\ell})_{q\times q} \in [0,1]^{q\times q}$

is a symmetric and full rank connectivity matrix, with $\omega_{k,\ell} = \frac{\theta_{k,\ell}}{\theta_{k,\ell}+\eta_{k,\ell}}$. Then $z_{i,j} = 1$ if and only if the i-th node belongs to the j-th community. On the other hand, $\omega_{k,\ell}$ is the connection probability between the nodes in community k and the nodes in community ℓ, and $s_k \equiv \sum_{i=1}^p z_{i,k}$ is the size of community $k \in \{1, \dots, q\}$. Clearly, matrix $\mathbf{Z}$ and function $\nu(\cdot)$ are the two equivalent representations for the community membership of the network nodes.

Under model (15), the edge connection probability matrix is given as $E(\mathbf{X}_t) = \mathbf{W} - \text{diag}(\mathbf{W})$, where $\mathbf{W} = \mathbf{Z}\mathbf{\Omega}\mathbf{Z}^\top = (\omega_{\nu(i),\nu(j)})$. Then the standard spectral clustering algorithm for estimating the community memberships are based on Laplacian eigendecomposition of $\mathbf{W}$ or its normalized version (Rohe et al., 2011), for which connection probability $\omega_{\nu(i),\nu(j)}$ is taken as a similarity measure between nodes i and j. To take the advantage of the dynamic structure of dynamic networks, we take the transition probabilities $\theta_{\nu(i),\nu(j)}$ and $(1 - \eta_{\nu(i),\nu(j)})$, instead of conNection probability $\omega_{\nu(i),\nu(j)}$, as the similarity measures (see Remark 1(i) below). To this end, define

$$\mathbf{\Omega}_1 = (\theta_{k,\ell})_{q\times q}, \qquad \mathbf{\Omega}_2 = (1 - \eta_{k,\ell})_{q\times q},$$
$$\mathbf{W}_1 = \mathbf{Z}\mathbf{\Omega}_1\mathbf{Z}^\top = (\alpha_{i,j})_{p\times p}, \qquad \mathbf{W}_2 = \mathbf{Z}\mathbf{\Omega}_2\mathbf{Z}^\top = (1 - \beta_{i,j})_{p\times p},$$

where $\alpha_{i,j} = \theta_{\nu(i),\nu(j)}$, $\beta_{i,j} = \eta_{\nu(i),\nu(j)}$. Let $\mathbf{D}_1$ and $\mathbf{D}_2$ be two $p \times p$ diagonal matrices with, respectively, $d_{i,1}, d_{i,2}$ as their (i, i)-th elements, where

$$d_{i,1} = \sum_{j=1}^p \alpha_{i,j}, \qquad d_{i,2} = \sum_{j=1}^p (1 - \beta_{i,j}).$$

The normalized Laplacian matrices based on $\mathbf{W}_1$ and $\mathbf{W}_2$ are then defined as:

$$\mathbf{L}_1 = \mathbf{D}_1^{-1/2}\mathbf{W}_1\mathbf{D}_1^{-1/2}, \qquad \mathbf{L}_2 = \mathbf{D}_2^{-1/2}\mathbf{W}_2\mathbf{D}_2^{-1/2}, \qquad \mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2. \tag{16}$$

The following lemma shows that the block structure in the membership matrix $\mathbf{Z}$ can be recovered by the leading eigenvectors of $\mathbf{L}$.

Proposition 9. *Let rank($\mathbf{L}$) $= q$, and $\mathbf{\Gamma}_q\mathbf{\Lambda}\mathbf{\Gamma}_q^\top$ be the eigen-decomposition of $\mathbf{L}$, where $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_q\}$ is the diagonal matrix consisting of the nonzero eigenvalues of $\mathbf{L}$ arranged in the order $|\lambda_1| \geq \cdots \geq |\lambda_q| > 0$. There exists a matrix $\mathbf{U} \in \mathcal{R}^{q\times q}$ such that $\mathbf{\Gamma}_q = \mathbf{Z}\mathbf{U}$. Furthermore, for any $1 \leq i, j \leq p$, $\mathbf{z}_i\mathbf{U} = \mathbf{z}_j\mathbf{U}$ if and only if $\mathbf{z}_i = \mathbf{z}_j$, where $\mathbf{z}_i$ denotes the i-th row of $\mathbf{Z}$.*

Remark 1. (i) Both $\theta_{\nu(i),\nu(j)}$ and $(1 - \eta_{\nu(i),\nu(j)})$ can be regarded as a similarity measure between nodes i and j. This is due to the fact that the communities in a network are often formed in the way that the members within the same community are more likely to be connected with each other, and the members belong to different communities are unlikely or less likely to be connected. Hence when nodes i and j belong to the same community, $\alpha_{i,j}$ tends to be large and $\beta_{i,j}$ tends to be small (see (3)). The converse is true when the two nodes belong to two different communities.

(ii) As $\mathbf{Z}$ is $p \times q$ with rank q, it is reasonable to assume rank($\mathbf{L}$) $= q$ in Proposition 9. The q columns of $\mathbf{\Gamma}_q$ are the orthonormal eigenvectors of $\mathbf{L}$ corresponding to the q non-zero eigenvalues. Proposition 9 implies that there are only q distinct rows in the $p \times q$ matrix $\mathbf{\Gamma}_q$, and two nodes belong to a same community if and only if the corresponding rows in $\mathbf{\Gamma}_q$ are the same. Intuitively the discriminant power of $\mathbf{\Gamma}_q$ can be understood as follows. For any unit vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\top$, it holds that

$$\boldsymbol{\gamma}^\top \mathbf{L} \boldsymbol{\gamma} = 2 - \sum_{1 \le i < j \le p} \alpha_{i,j} \Big(\frac{\gamma_i}{\sqrt{d_{i,1}}} - \frac{\gamma_j}{\sqrt{d_{j,1}}} \Big)^2 - \sum_{1 \le i < j \le p} (1 - \beta_{i,j}) \Big(\frac{\gamma_i}{\sqrt{d_{i,2}}} - \frac{\gamma_j}{\sqrt{d_{j,2}}} \Big)^2. \quad (17)$$

For $\boldsymbol{\gamma}$ being an eigenvector corresponding to the largest (positive) eigenvalue of $\mathbf{L}$, the sum of the 2nd and the 3rd terms on the RHS (17) is minimized. Thus $|\gamma_i - \gamma_j|$ is minimized when $\alpha_{i,j}$ and/or $(1 - \beta_{i,j})$ are large; noting that $d_{i,k} = d_{j,k}$ for $k = 1, 2$ when nodes i and j belong to the same community. The eigenvectors corresponding to negative eigenvalues are capable to identify the so-called heterophilic communities, see pp.1892-3 of Rohe et al. (2011).

3.2.2 Estimating membership $\nu(\cdot)$

It follows from Proposition 2, (14) and (15) that

$$P(X_{ij}^t = 1) = \theta_{\nu(i),\nu(j)} / (\theta_{\nu(i),\nu(j)} + \eta_{\nu(i),\nu(j)}) \equiv \omega_{\nu(i),\nu(j)}, \quad 1 \le i < j \le p,$$

provided that X_{ij}^0 is initiated with the same marginal distribution. A simple approach adopted in literature is to apply a community detection method for static stochastic block models using the averaged data $\bar{\mathbf{X}} = \sum_{1 \le t \le n} \mathbf{X}_t / n$ to detect the latent communities characterized by the connection probabilities $\{\omega_{k,\ell}, 1 \le k \le \ell \le q\}$. We take a different approach based on estimators $\{(\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}), 1 \le i < j \le p\}$ defined in (12) to identify the clusters determined by the transition probabilities $\{(\theta_{k,\ell}, \eta_{k,\ell}), 1 \le k \le \ell \le q\}$ instead. More precisely, we propose a new spectral clustering algorithm to estimate $\mathbf{\Gamma}_q$ based on Proposition 9 above.

Let $\widehat{\mathbf{W}}_1, \widehat{\mathbf{W}}_2$ be two $p \times p$ matrices with, respectively, $\widehat{\alpha}_{i,j}, (1 - \widehat{\beta}_{i,j})$ as their (i,j)-th elements for $i \ne j$, and 0 on the main diagonals. Let $\widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2$ be two $p \times p$ diagonal matrices with, respectively, $\widehat{d}_{i,1}, \widehat{d}_{i,2}$ as their (i,i)-th elements, where

$$\widehat{d}_{i,1} = \sum_{j=1}^p \widehat{\alpha}_{i,j}, \qquad \widehat{d}_{i,2} = \sum_{j=1}^p (1 - \widehat{\beta}_{i,j}).$$

Define two (normalized) Laplacian matrices

$$\widehat{\mathbf{L}}_1 = \widehat{\mathbf{D}}_1^{-1/2} \widehat{\mathbf{W}}_1 \widehat{\mathbf{D}}_1^{-1/2}, \qquad \widehat{\mathbf{L}}_2 = \widehat{\mathbf{D}}_2^{-1/2} \widehat{\mathbf{W}}_2 \widehat{\mathbf{D}}_2^{-1/2}. \quad (18)$$

Perform the eigen-decomposition for the sum of $\mathbf{L}_1$ and $\mathbf{L}_2$:

$$\widehat{\mathbf{L}} \equiv \widehat{\mathbf{L}}_1 + \widehat{\mathbf{L}}_2 = \widehat{\mathbf{\Gamma}} \, \mathrm{diag}(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_p) \widehat{\mathbf{\Gamma}}^\top, \quad (19)$$

where the eigenvalues are arranged in the order $\widehat{\lambda}_1^2 \ge \ldots \ge \widehat{\lambda}_p^2$, and the columns of the $p \times p$ orthogonal matrix $\widehat{\mathbf{\Gamma}}$ are the corresponding eigenvectors. We call $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ the q leading eigenvalues of $\widehat{\mathbf{L}}$. Denote by $\widehat{\mathbf{\Gamma}}_q$ the $p \times q$ matrix consisting of the first q columns of $\widehat{\mathbf{\Gamma}}$, which are called the leading eigenvectors of $\widehat{\mathbf{L}}$. The spectral clustering applies the k-means clustering algorithm to the p rows of $\widehat{\mathbf{\Gamma}}_q$ to obtain the community assignments for the p nodes $\widehat{\nu}(i) \in \{1, \cdots, q\}$ for $i = 1, \cdots, p$.

Remark 2. Proposition 9 implies that the true memberships can be recovered by the q distinct rows of $\mathbf{\Gamma}_q$. Note that

$$\widehat{\mathbf{L}} = \widehat{\mathbf{L}}_1 + \widehat{\mathbf{L}}_2 \approx \mathbf{L}_1 - \mathrm{diag}(\mathbf{L}_1) + \mathbf{L}_2 - \mathrm{diag}(\mathbf{L}_2) = \mathbf{L} - \mathrm{diag}(\mathbf{L}).$$

We shall see that the effect of the term $\mathrm{diag}(\mathbf{L})$ on the eigenvectors $\mathbf{\Gamma}_q$ is negligible when p is large (see for example (A.6) in the proof of Lemma 19 in Appendix A), and hence the rows of $\widehat{\mathbf{\Gamma}}_q$ should be slightly perturbed versions of the q distinct rows in $\mathbf{\Gamma}_q$.

The following theorem justified the validity of using $\widehat{\mathbf{L}}$ for spectral clustering. Note that $\|\cdot\|_2$ and $\|\cdot\|_F$ denote, respectively, the L_2 and the Frobenius norm of matrices.

Theorem 10. *Let conditions (2.5), C1 and C2 hold, and* $\lambda_q^{-2}\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right) \to 0$*, as* $n, p \to \infty$*. Then it holds that*

$$\max_{i=1,\ldots,p} |\lambda_i^2 - \widehat{\lambda}_i^2| \le \|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_2 \le \|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_F = O_p\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right). \tag{20}$$

Moreover, for any constant $B > 0$*, there exists a constant* $C > 0$ *such that the inequality*

$$\|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q \mathbf{O}_q\|_F \le 4\lambda_q^{-2} C \left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right) \tag{21}$$

holds with probability greater than $1 - 16p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right]$*, where* $\mathbf{O}_q$ *is a* $q \times q$ *orthogonal matrix.*

It follows from (20) that the leading eigenvalues of $\mathbf{L}$ can be consistently recovered by the leading eigenvalues of $\widehat{\mathbf{L}}$. By (21), the leading eigenvectors of $\mathbf{L}$ can also be consistently estimated, subject to a rotation (due to the possible multiplicity of some leading eigenvalues $\mathbf{L}$). Proposition 9 indicates that there are only q distinct rows in $\mathbf{\Gamma}_q$, and, therefore, also q distinct rows in $\mathbf{\Gamma}_q \mathbf{O}_q$, corresponding to the q latent communities for the p nodes. This paves the way for the k-means algorithm stated below. Put

$$\mathcal{M}_{p,q} = \{\mathbf{M} \in \mathcal{R}^{p \times q} : \mathbf{M} \text{ has } q \text{ distinct rows}\}.$$

The k-means clustering algorithm: Let

$$(\widehat{\mathbf{c}}_1, \cdots, \widehat{\mathbf{c}}_p)^\top = \arg\min_{\mathbf{M} \in \mathcal{M}_{p,q}} \|\widehat{\mathbf{\Gamma}}_q - \mathbf{M}\|_F^2.$$

There are only q distinct vectors among $\widehat{\mathbf{c}}_1, \cdots, \widehat{\mathbf{c}}_p$, forming the q communities. Theorem 11 below shows that they are identical to the latent communities of the p nodes under (21) and (22). The latter holds if $\sqrt{s_{\max}}\lambda_q^{-2} C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right) \to 0$, where $s_{\max} = \max\{s_1, \ldots, s_q\}$ is the size of the largest community.

Theorem 11. *Let* (21) *hold and*

$$\sqrt{\frac{1}{s_{\max}}} > 4\sqrt{2}\lambda_q^{-2} C \left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p} \right). \tag{22}$$

Then $\widehat{\mathbf{c}}_i = \widehat{\mathbf{c}}_j$ *if and only if* $\nu(i) = \nu(j)$, $1 \le i, j \le p$.

Remark 3. By Lemma A.1 of Rohe et al. (2011), the error bound for the standard spectral clustering algorithm (with $n = 1$) is $O_p\left(\frac{\log p}{\sqrt{p}} + \frac{1}{p}\right)$, where the term $\frac{1}{p}$ reflects the bias caused by the inconsistent estimation of diagonal terms (see equation (A.5) and subsequent derivations in Rohe et al. (2011)). This bias comes directly from the removal of the diagonal elements of $\mathbf{L}$, as pointed out in Remark 2 above. Although the algorithm was designed for static networks, it has often been applied to dynamic networks using $\frac{1}{n}\sum_t \mathbf{X}_t$ in the place of a single observed network; see, e.g. Bhattacharjee et al. (2020). With some simple modification to the proof of Lemma A.1 of Rohe et al. (2011), it can be shown that the error bound is then reduced to

$$O_p\left(\frac{\log(pn)}{\sqrt{np}} + \frac{1}{p} \right), \tag{23}$$

provided that the observed networks are i.i.d. The error would only increase when the observations are not independent. On the other hand, our proposed spectral clustering algorithm for (dependent) dynamic networks entails the error rate specified in (20) and (21) which is smaller than (23) as long as n is sufficiently large (i.e. $(p/n)^{\frac{1}{2}} / \log(np) \to 0$). Note that we need n to be large enough in relation to p in order to capture the dynamic dependence of the networks. From the proof of Theorem 11 we can see that left hand side of (22) is a lower bound for the "signal" strength (c.f. equation (A.17)), and the right hand side of (22) is an upper bound for the overall estimation error.If we relax the lower bound for the "signal" strength in (22) to $\lambda_q^{-2} \frac{C}{s_{n,p}} \left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p} \right)$ for some $s_{n,p} = o(\sqrt{p})$, from (21) we have, there are at most $O(s_{n,p}^2)$ rows in $\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q \mathbf{O}_q$ such that the estimation error is larger than $\lambda_q^{-2} \frac{C}{s_{n,p}} \left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p} \right)$. Consequently, under this relaxed lower bound, our algorithm will still be consistent with a negligible misclassification rate $r_{n,p} := O_p(p^{-1} s_{n,p}^2)$.

3.2.3 Estimation for $\theta_{k,\ell}$ and $\eta_{k,\ell}$

For any $1 \le k \le \ell \le q$, we define

$$S_{k,l} = \begin{cases} \{(i,j) : 1 \le i \ne j \le p, \nu(i) = k, \nu(j) = \ell\} & \text{if } k \ne l, \\ \{(i,j) : 1 \le i < j \le p, \nu(i) = k = \nu(j) = \ell\} & \text{if } k = l, \end{cases} \tag{24}$$

Clearly the cardinality of $S_{k,\ell}$ is $n_{k,\ell} = s_k s_\ell$ when $k \ne \ell$ and $n_{k,\ell} = s_k(s_k - 1)/2$ when $k = \ell$.

Based on the procedure presented in Section 3.2.2, we obtain an estimated membership function $\widehat{\nu}(\cdot)$. Consequently, the MLEs for $(\theta_{k,\ell}, \eta_{k,\ell})$, $1 \le k \le \ell \le q$, admit the form

$$\widehat{\theta}_{k,\ell} = \sum_{(i,j)\in\widehat{S}_{k,\ell}} \sum_{t=1}^{n} X_{i,j}^t (1 - X_{i,j}^{t-1}) \Big/ \sum_{(i,j)\in\widehat{S}_{k,\ell}} \sum_{t=1}^{n} (1 - X_{i,j}^{t-1}), \tag{25}$$

$$\widehat{\eta}_{k,\ell} = \sum_{(i,j)\in\widehat{S}_{k,\ell}} \sum_{t=1}^{n} (1 - X_{i,j}^t) X_{i,j}^{t-1} \Big/ \sum_{(i,j)\in\widehat{S}_{k,\ell}} \sum_{t=1}^{n} X_{i,j}^{t-1}, \tag{26}$$

where

$$\widehat{S}_{k,\ell} = \begin{cases} \{(i,j) : 1 \le i \ne j \le p,\ \widehat{\nu}(i) = k,\ \widehat{\nu}(j) = \ell\} & \text{if } k \ne \ell, \\ \{(i,j) : 1 \le i < j \le p,\ \widehat{\nu}(i) = \widehat{\nu}(j) = k\} & \text{if } k = \ell. \end{cases}$$

See (12) and also (14).

Theorem 11 implies that the memberships of the nodes can be consistently recovered. Consequently, the consistency and the asymptotic normality of the MLEs $\widehat{\theta}_{k,\ell}$ and $\widehat{\eta}_{k,\ell}$ can be established in the same manner as for Propositions 6 and 7. We state the results below.

Let $\mathcal{K}_1 = \{(i_1, j_1), \ldots, (i_{m_1}, j_{m_1})\}$ and $\mathcal{K}_2 = \{(k_1, \ell_1), \ldots, (k_{m_2}, \ell_{m_2})\}$ be two arbitrary subsets of $\{(k,\ell) : 1 \le k \le \ell \le q\}$ with $m_1, m_2 \ge 1$ fixed. Let

$$\mathbf{\Psi}_{\mathcal{K}_1,\mathcal{K}_2} = (\theta_{i_1,j_1}, \ldots, \theta_{i_{m_1},j_{m_1}}, \eta_{k_1,\ell_1}, \ldots, \eta_{k_{m_2},\ell_{m_2}})',$$

and let $\widehat{\mathbf{\Psi}}_{\mathcal{K}_1,\mathcal{K}_2}$ denote its MLE. Put $\mathbf{N}_{\mathcal{K}_1,\mathcal{K}_2} = \text{diag}(n_{i_1,j_1}, \ldots, n_{i_{m_1},j_{m_1}}, n_{k_1,\ell_1}, \ldots, n_{k_{m_2},\ell_{m_2}})$ where $n_{k,\ell}$ is the cardinality of $S_{k,\ell}$ defined as in (24).

Theorem 12. *Let conditions (2.5), C1 and C2 hold, and* $\frac{\sqrt{s_{\max}}}{\lambda_q^2}\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right) \to 0.$ *Then it holds that*

$$\max_{1\le k,\ell\le q} |\widehat{\theta}_{k,\ell} - \theta_{k,\ell}| = O_p\left(\sqrt{\frac{\log q}{n s_{\min}^2}}\right) \quad \text{and} \quad \max_{1\le k,\ell\le q} |\widehat{\eta}_{k,\ell} - \eta_{k,\ell}| = O_p\left(\sqrt{\frac{\log q}{n s_{\min}^2}}\right),$$

where $s_{\min} = \min\{s_1, \ldots, s_q\}$.

Theorem 13. *Let the condition of Theorem 12 hold. Then*

$$\sqrt{n}\mathbf{N}_{\mathcal{K}_1,\mathcal{K}_2}^{\frac{1}{2}}(\widehat{\mathbf{\Psi}}_{\mathcal{K}_1,\mathcal{K}_2} - \mathbf{\Psi}_{\mathcal{K}_1,\mathcal{K}_2}) \to N(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_{\mathcal{K}_1,\mathcal{K}_2}),$$

where $\widetilde{\mathbf{\Sigma}}_{\mathcal{K}_1,\mathcal{K}_2} = \text{diag}(\widetilde{\sigma}_{11}, \ldots, \widetilde{\sigma}_{m_1+m_2,m_1+m_2})$ *with*

$$\widetilde{\sigma}_{rr} = \frac{\theta_{i_r,j_r}(1 - \theta_{i_r,j_r})(\theta_{i_r,j_r} + \eta_{i_r,j_r})}{\eta_{i_r,j_r}}, \quad 1 \le r \le m_1,$$
$$\widetilde{\sigma}_{rr} = \frac{\eta_{k_r,\ell_r}(1 - \eta_{k_r,\ell_r})(\theta_{k_r,\ell_r} + \eta_{k_r,\ell_r})}{\theta_{k_r,\ell_r}}, \quad m_1 + 1 \le r \le m_1 + m_2.$$

Finally to prepare for the inference in Section 3.3 below, we introduce some notations. First we denote $\widehat{\nu}$ by $\widehat{\nu}^{1,n}$, to reflect the fact that the community clustering was carried out using the data $\mathbf{X}_1, \cdots, \mathbf{X}_n$ (conditionally on $\mathbf{X}_0$). See Section 3.2.2 above. Further we denote the maximum log likelihood by

$$\widehat{l}(1, n;\ \widehat{\nu}^{1,n}) = l(\{\widehat{\theta}_{k,\ell}, \widehat{\eta}_{k,\ell}\};\ \widehat{\nu}^{1,n}) \tag{27}$$

to highlight the fact that both the node clustering and the estimation for transition probabilities are based on the data $\mathbf{X}_1, \cdots, \mathbf{X}_n$.

3.3 Inference with a change point

Now we assume that there is a change point τ_0 at which both the membership of nodes and the transition probabilities $\{\theta_{k,\ell}, \eta_{k,\ell}\}$ change. It is necessary to assume $n_0 \le \tau_0 \le n - n_0$, where n_0 is an integer and $n_0/n \equiv c_0 > 0$ is a small constant, as we need enough information before and after the change in order to detect τ_0. We assume that within the time period $[0, \tau_0]$, the network follows a stationary model (15) with parameters $\{(\theta_{1,k,\ell}, \eta_{1,k,\ell}) : 1 \le k, l \le q\}$ and a membership map $\nu^{1,\tau_0}(\cdot)$. Within the time period $[\tau_0 + 1, n]$ the network follows a stationary model (15) with parameters $\{(\theta_{2,k,\ell}, \eta_{2,k,\ell}) : 1 \le k, l \le q\}$ and a membership map $\nu^{\tau_0+1,n}(\cdot)$. Though we assume that the number of communities is unchanged after the change point, our results can be easily extended to the case that the number of communities also changes.

We estimate the change point τ_0 by the maximum likelihood method:

$$\widehat{\tau} = \arg \max_{n_0 \le \tau \le n - n_0} \{\widehat{l}(1, \tau;\ \widehat{\nu}^{1,\tau}) + \widehat{l}(\tau + 1, n;\ \widehat{\nu}^{\tau+1,n})\}, \tag{28}$$

where $\widehat{l}(\cdot)$ is given in (27).

To measure the difference between the two sets of transition probabilities before and after the change, we put

$$\Delta_F^2 = \frac{1}{p^2}\big(\|\mathbf{W}_{1,1} - \mathbf{W}_{2,1}\|_F^2 + \|\mathbf{W}_{1,2} - \mathbf{W}_{2,2}\|_F^2\big),$$

where the four $p \times p$ matrices are defined as

$$\mathbf{W}_{1,1} = (\theta_{1,\nu^{1,\tau_0}(i),\nu^{1,\tau_0}(j)}), \quad \mathbf{W}_{1,2} = (1 - \eta_{1,\nu^{1,\tau_0}(i),\nu^{1,\tau_0}(j)}),$$

$$\mathbf{W}_{2,1} = (\theta_{2,\nu^{\tau_0+1,n}(i),\nu^{\tau_0}(j)+1,n}), \quad \mathbf{W}_{2,2} = (1 - \eta_{1,\nu^{\tau_0+1,n}(i),\nu^{\tau_0+1,n}(j)}).$$

Note that Δ_F can be viewed as the signal strength for detecting the change point τ_0. Let $s_{\max},\ s_{\min}$ denote, respectively, the largest, and the smallest community size among all the communities before and after the change. Similar to (16), we denote the normalized Laplacian matrices corresponding to $\mathbf{W}_{i,j}$ as $\mathbf{L}_{i,j}$ for $i, j = 1, 2$. Let $|\lambda_{i,1}| \ge |\lambda_{i,2}| \ge \ldots \ge |\lambda_{i,q}|$ be the absolute nonzero eigenvalues of $\mathbf{L}_{i,1} + \mathbf{L}_{i,2}$ for $i = 1, 2$, and we denote $\lambda_{\min} = \min\{|\lambda_{1,q}|, |\lambda_{2,q}|\}$. Now some regularity conditions are in order.

C3. For some constant $l > 0$, $\theta_{i,k,\ell}, \eta_{i,k,\ell} > l$, and $\theta_{i,k,\ell} + \eta_{i,k,\ell} \le 1$ for all $i = 1, 2$ and $1 \le k \le \ell \le q$.

C4. $\log(np)/\sqrt{p} \to 0$, and $\sqrt{s_{\max}}\lambda_{\min}^{-2}\big(\sqrt{\log(pn)/np}+\frac{1}{n}+\frac{1}{p}+\frac{\log(np)/n+\sqrt{\log(np)/(np^2)}}{\Delta_F^2}\big) \to 0$.

C5. $\frac{\Delta_F^2}{\log(np)/n+\sqrt{\log(np)/(np^2)}} \to \infty$.

Condition C3 is similar to C1. The condition $\log(np)/\sqrt{p} \to 0$ in C4 controls the misclassification rate of the k-means algorithm. Recall that there is a bias term $O(p^{-1})$ in spectral clustering caused by the removal of the diagonal of the Laplacian matrix (see Remark 2 above). Intuitively, as p increases, the effect of this bias term on the misclassification rate of the k-means algorithm becomes negligible. On the other hand, note that the length of the time interval for searching for the change point in (28) is of order $O(n)$; the $\log(n)$ term here in some sense reflects the effect of the difficulty in detecting the true change point when the searching interval is extended as n increases. The second condition in C4 is similar to (22), which ensures that the true communities can be recovered. Condition C5 requires that the average signal strength $\Delta_F^2 = p^{-2}\big[\|\mathbf{W}_{1,1} - \mathbf{W}_{2,1}\|_F^2 + \|\mathbf{W}_{1,2} - \mathbf{W}_{2,2}\|_F^2\big]$ is of higher order than $\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}$ for change point detection.

Yudovina et al. (2015) deals with the MLE for a change-point in a network Markov chain but without a latent community structure. Hence it does not have the complication to estimate the community memberships in addition. Allowing the membership change in our setting leads to an extra challenge: in the process of searching for the location of the change-point, the estimation for the latent communities before or after a specified location may not be consistent. To overcome this obstacle, we introduce a truncation which breaks the searching interval into two parts such that the error in the estimated change-point can be bounded. Bhattacharjee et al. (2020) also allows the membership change. But it assumes that the networks observed at different times are independent with each other.

Theorem 14. *Let conditions C2-C5 hold. Then the following assertions hold.*

(i) When $\nu^{1,\tau_0} \equiv \nu^{\tau_0+1,n}$,

$$\frac{|\tau_0 - \widehat{\tau}|}{n} = O_p\left(\frac{\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}}{\Delta_F^2} \times \min\left\{1, \frac{\min\left\{1, (n^{-1}p^2\log(np))^{\frac{1}{4}}\right\}}{\Delta_F s_{\min}}\right\}\right).$$

(ii) When $\nu^{1,\tau_0} \neq \nu^{\tau_0+1,n}$,

$$\frac{|\tau_0 - \widehat{\tau}|}{n} = O_p\left(\frac{\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}}{\Delta_F^2} \times \min\left\{1, \frac{\min\left\{1, (n^{-1}p^2\log(np))^{\frac{1}{4}}\right\}}{\Delta_F s_{\min}} + \frac{1}{\Delta_F^2}\right\}\right).$$

Notice that for $\tau < \tau_0$, the observations in the time interval $[\tau + 1, n]$ are a mixture of the two different network processes if $\nu^{1,\tau_0} \neq \nu^{\tau_0+1,n}$. In the worst case scenario then, all q communities can be changed after the change point τ_0. This causes the extra estimation error term $\frac{1}{\Delta_F^2}$ in Theorem 14(ii).

4. Simulations

4.1 Parameter estimation

We generate data according to model (1) in which the parameters α_{ij} and β_{ij} are drawn independently from $U[0.1, 0.5]$, $1 \leq i, j \leq p$. The initial value $\mathbf{X}_0$ was simulated according to (6) with $\pi_{ij} = 0.5$. We calculate the estimates according to (12). For each setting (with different p and n), we replicate the experiment 500 times. Furthermore we also calculate the 95% confidence intervals for α_{ij} and β_{ij} based on the asymptotically normal distributions specified in Proposition 7, and report the relative frequencies of the intervals covering the true values of the parameters. There are a few cases with denominators being exactly zero when evaluating the asymptotic variance. In such cases we follow a traditional approach by adding a small number $n^{-1} \times 10^{-4}$ to the denominator. This small value is negligible when the denominators are non-zero. The results are summarized in Table 1.

Table 1: The mean squared errors (MSE) of the estimated parameters in AR(1) network model (1) and the relative frequencies (coverage rates) of the event that the asymptotic 95% confidence intervals cover the true values in a simulation with 500 replications.

		$\widehat{\alpha}_{i,j}$		$\widehat{\beta}_{i,j}$	
n	p	MSE	Coverage (%)	MSE	Coverage (%)
5	100	.130	39.2	.131	39.3
5	200	.131	39.3	.131	39.4
20	100	.038	86.1	.037	86.0
20	200	.037	86.1	.037	86.0
50	100	.012	92.3	.012	92.2
50	200	.011	92..2	.012	92.2
100	100	.005	93.7	.005	93.8
100	200	.005	93.8	.005	93.9
200	100	.002	94.5	.002	94.5
200	200	.002	94.6	.002	94.5

The MSE decreases as n increases, showing steadily improvement in performance. The coverage rates of the asymptotic confidence intervals are very close to the nominal level when $n \geq 50$. The results hardly change between $p = 100$ and 200.

4.2 Community Detection

We now consider model (14) with $q = 2$ or 3 clusters, in which $\theta_{i,i} = \eta_{i,i} = 0.4$ for $i = 1, \cdots, q$, and $\theta_{i,j}$ and $\eta_{i,j}$, for $1 \leq i, j \leq q$, are drawn independently from $U[0.05, 0.25]$. For each setting, we replicate the experiment 500 times.

We identify the q latent communities using the newly proposed spectral clustering algorithm based on matrix $\widehat{\mathbf{L}} = \widehat{\mathbf{L}}_1 + \widehat{\mathbf{L}}_2$ defined in (19). For the comparison purpose, we also implement the standard spectral clustering method for static networks (cf. Rohe et al.

(2011)) but using the average

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t \tag{29}$$

in place of the single observed adjacency matrix. This idea has been frequently used in spectral clustering for dynamic networks; see, for example, Wilson et al. (2019); Zhao et al. (2019); Bhattacharjee et al. (2020). We report the normalized mutual information (NMI) and the adjusted Rand index (ARI): Both metrics take values between 0 and 1, and both measure the closeness between the true communities and the estimated communities in the sense that the larger the values of NMI and ARI are, the closer the two sets of communities are; see Vinh et al. (2010). The results are summarized in Table 2. The newly proposed algorithm based on $\widehat{\mathbf{L}}$ always outperforms the algorithm based on $\bar{\mathbf{X}}$, even when n is as small as 5. The differences between the two methods are substantial in terms of the scores of both NMI and ARI. For example when $q = 2, p = 100$ and $n = 5$, NMI and ARI are, respectively, 0.621 and 0.666 for the new method, and they are merely 0.148 and 0.158 for the standard method based on $\bar{\mathbf{X}}$. This is due to the fact that the standard method uses only the information on $\pi_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}$, and fails to take the advantage of the AR(1) structure in which the information on both $\alpha_{i,j}$ and $\beta_{i,j}$ is available.

After the communities were identified, we estimate $\theta_{i,j}$ and $\eta_{i,j}$ by (25) and (26), respectively. The mean squared errors (MSE) are evaluated for all the parameters. The results are summarized in Table 3. For the comparison purpose, we also report the estimates based on the identified communities by the $\bar{\mathbf{X}}$-based clustering. The MSE values of the estimates based on the communities identified by the new clustering method are always smaller than those of based on $\bar{\mathbf{X}}$. Noticeably now the estimates with small n such as $n = 5$ are already reasonably accurate, as the information from all the nodes within the same community is pulled together.

5. Illustration with real data

We illustrate the proposed methodology through three real data examples in this section. More real data analysis can be found in Appendix B.

5.1 RFID sensors data

Contacts between patients, patients and health care workers (HCW) and among HCW represent one of the important routes of transmission of hospital-acquired infections. Vanhems et al. (2013) collected records of contacts among patients and various types of HCW in the geriatric unit of a hospital in Lyon, France, between 1pm on Monday 6 December and 2pm on Friday 10 December 2010. Each of the $p = 75$ individuals in this study consented to wear Radio-Frequency IDentification (RFID) sensors on small identification badges during this period, which made it possible to record when any two of them were in face-to-face contact with each other (i.e. within 1-1.5 meters) in every 20-second interval during the period. This data set is now available in R packages `igraphdata` and `sand`.

Following Vanhems et al. (2013), we combine together the recorded information in each 24 hours to form 5 daily networks ($n = 5$), i.e. an edge between two individuals is equal to 1 if they made at least one contact during the 24 hours, and 0 otherwise. Those 5 networks

Table 2: Normalized mutual information (NMI) and adjusted Rand index (ARI) of the true communities and the estimated communities in the simulation with 500 replications. The communities are estimated by the spectral clustering algorithm (SCA) based on either matrix $\widehat{\mathbf{L}}$ in (19) or matrix $\bar{\mathbf{X}}$ in (29).

			SCA based on $\widehat{\mathbf{L}}$		SCA based on $\bar{\mathbf{X}}$	
q	p	n	NMI	ARI	NMI	ARI
2	100	5	.621	.666	.148	.158
		20	.733	.755	.395	.402
		50	.932	.938	.572	.584
		100	.994	.995	.692	.696
2	200	5	.808	.839	.375	.406
		20	.850	.857	.569	.589
		50	.949	.953	.712	.722
		100	.994	.995	.790	.796
3	100	5	.542	.536	.078	.057
		20	.686	.678	.351	.325
		50	.931	.929	.581	.562
		100	.988	.987	.696	.670
3	200	5	.729	.731	.195	.175
		20	.779	.763	.550	.542
		50	.954	.952	.726	.711
		100	.994	.994	.822	.802

are plotted in Figure 1. We fit the data with stationary AR(1) model (1) and (5). Some summary statistics of the estimated parameters, according to the 4 different roles of the individuals, are presented in Table 4, together with the direct relatively frequency estimates $\widetilde{\pi}_{i,j} = \bar{X}_{i,j} = \sum_{t=1}^{5} X_{i,j}^t/5$. We apply the permutation test (13) (with 500 permutations) to the residuals resulted from the fitted AR(1) model. The P-value is 0.45, indicating no significant evidence against the stationarity assumption.

Since the original data were recorded for each 20 seconds, they can also be combined into half-day series with $n = 10$. Figure 2 presents the 10 half-day networks. We repeat the above exercise for this new sequence. Now the P-value of the permutation test is 0.008, indicating the stationary AR(1) model should be rejected for this sequence of 10 networks. This is intuitively understandable, as people behave differently at the different times during a day (such as daytime or night). Those within-day nonstationary behaviour shows up in the data accumulation over every 12 hours, and it disappears in the accumulation over 24 hour periods. Also overall the adjacent two networks in Figure 2 look more different from each other than the adjacent pairs in Figure 1.

There is no evidence of the existence of any communities among the 75 individuals in this data set. Our analysis confirms this too. For example the results of the spectral clustering algorithm based on, respectively, $\widehat{\mathbf{L}}$ and $\bar{\mathbf{X}}$ do not corroborate with each other at all as the NMI is smaller than 0.1.

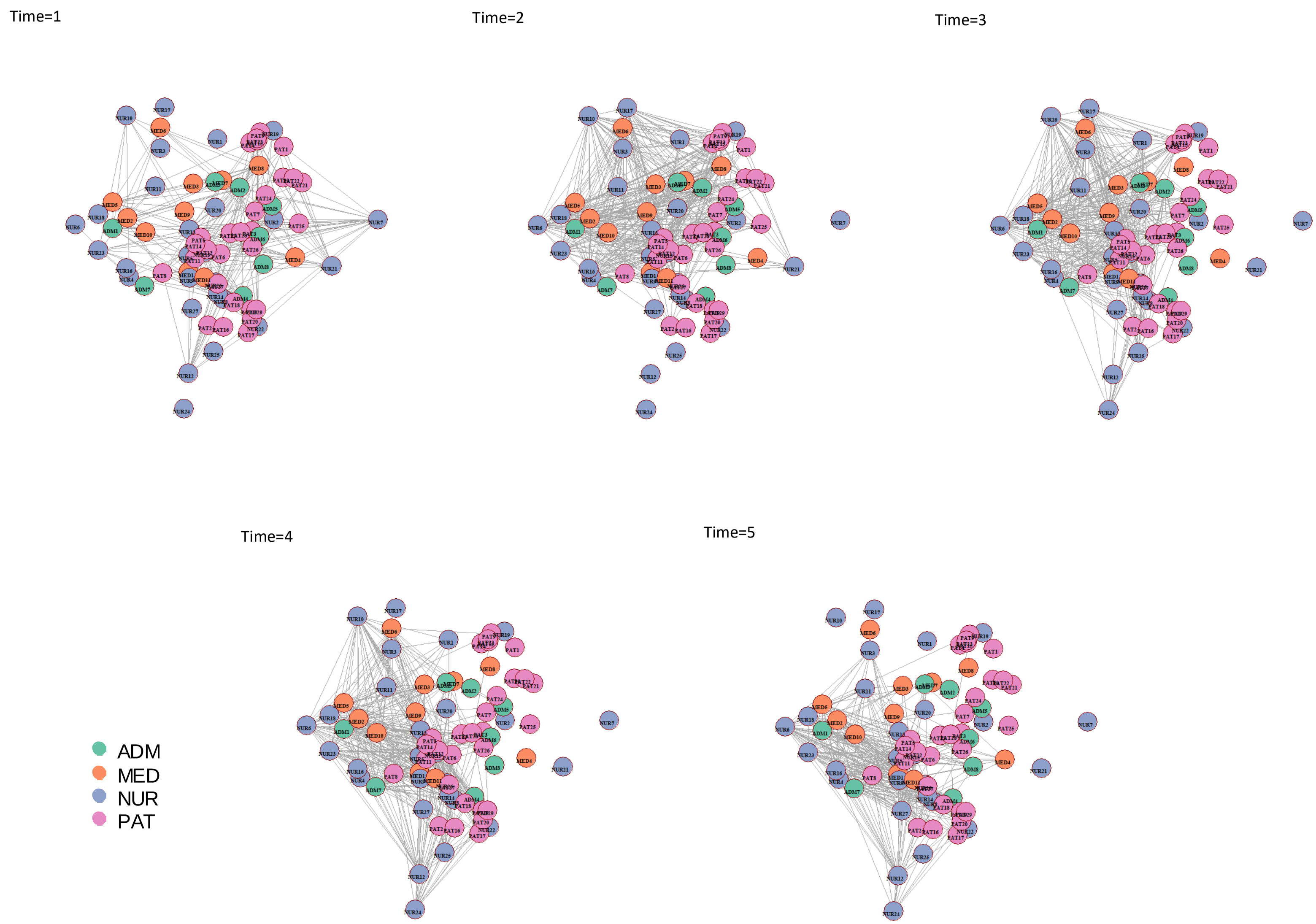

Figure 1: The RFID sensors data: the 5 networks obtained by combining together the information within each of the five 24-hour periods. The four different identities of the individuals are marked in four different colours.

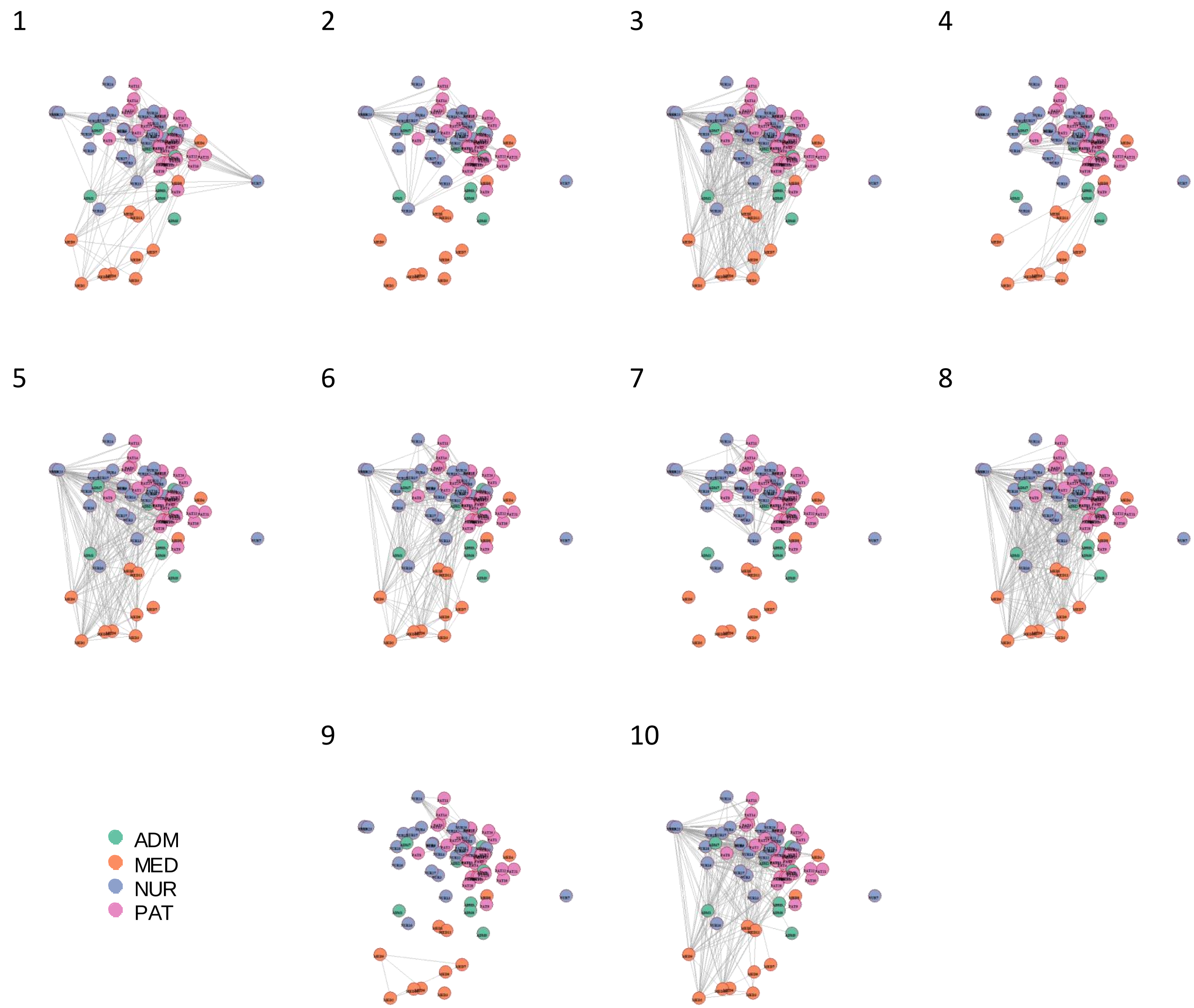

Figure 2: The RFID sensors data: the 10 networks obtained by combining together the information within each of the ten 12-hour periods. The four different identities of the individuals are marked in four different colours.

Table 3: The mean squared errors (MSE) of the estimated parameters in AR(1) network models with q communities. The communities are estimated by the spectral clustering algorithm (SCA) based on either matrix $\widehat{\mathbf{L}}$ in (19) or matrix $\bar{\mathbf{X}}$ in (29).

			SCA based on $\widehat{\mathbf{L}}$		SCA based on $\bar{\mathbf{X}}$	
q	p	n	$\widehat{\theta}_{i,j}$	$\widehat{\eta}_{i,j}$	$\widehat{\theta}_{i,j}$	$\widehat{\eta}_{i,j}$
2	100	5	.0149	.0170	.0298	.0312
		20	.0120	.0141	.0229	.0233
		50	.0075	.0083	.0178	.0177
		100	.0058	.0061	.0147	.0148
2	200	5	.0099	.0116	.0223	.0248
		20	.0093	.0111	.0219	.0248
		50	.0068	.0073	.0140	.0145
		100	.0061	.0062	.0117	.0118
3	100	5	.0194	.0211	.0318	.0325
		20	.0156	.0181	.0251	.0255
		50	.0093	.0104	.0193	.0193
		100	.0081	.0085	.0163	.0162
3	200	5	.0143	.0162	.0287	.0301
		20	.0134	.0156	.0200	.0205
		50	.0090	.0093	.0156	.0153
		100	.0079	.0083	.0130	.0131

5.2 French high school contact data

Now we consider a contact network data collected in a high school in Marseilles, France (Mastrandrea et al., 2015). The data are the recorded face-to-face contacts among the students from 9 classes during $n = 5$ days in December 2013, measured by the SocioPatterns infrastructure. Those are students in the so-called *classes preparatoires* – a part of the French post-secondary education system. We label the 3 classes majored in mathematics and physics as MP1, MP2 and MP3, the 3 classes majored in biology as BIO1, BIO2 and BIO3, the 2 classes majored in physics and chemistry as PC1 and PC2, and the class majored in engineering as EGI. The data are available at `www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/`. We have removed the individuals with missing values, and include the remaining $p = 327$ students in our clustering analysis based on the AR(1) stochastic block network model (see Definition 8).

We start the analysis with $q = 2$. The detected 2 clusters by the spectral clustering algorithm (SCA) based on either $\widehat{\mathbf{L}}$ in (19) or $\bar{\mathbf{X}}$ are reported in Table 5. The two methods lead to almost identical results: 3 classes majored in biology are in one cluster and the other 6 classes are in the other cluster. The number of 'misplaced' students is 2 and 1, respectively, by the SCA based on $\widehat{\mathbf{L}}$ and $\bar{\mathbf{X}}$. Figure 3 shows that the identified two clusters are clearly separated from each other across all the 5 days. The permutation test (13) on the residuals indicates that the stationary AR(1) stochastic block network model seems to

Table 4: Mean estimated coefficients (standard errors) for the four types of individuals in RFID data. Status codes: administrative staff (ADM), medical doctor (MED), paramedical staff, such as nurses or nurses' aides (NUR), and patients (PAT).

	$\widehat{\alpha}_{ij}$			
Status	ADM	NUR	MED	PAT
ADM	.1249 (.2212)	.1739 (.2521)	.1666 (.2641)	.1113 (.2021)
NUR		.2347 (.2927)	.2398 (.3022)	.1922 (.2513)
MED			.3594 (.3883)	.1264 (.2175)
PAT				.0089 (.0552)
	$\widehat{\beta}_{ij}$			
Status	ADM	NUR	MED	PAT
ADM	.1666 (.3660)	.2326 (.3883)	.2925 (.4235)	.2061 (.3798)
NUR		.3714 (.4470)	.3001 (.4167)	.3656 (.4498)
MED			.4187 (.3973)	.2311 (.4066)
PAT				.0198 (.1331)
	$\widehat{\pi}_{ij} = \widehat{\alpha}_{ij}/(\widehat{\alpha}_{ij} + \widehat{\beta}_{ij})$			
Status	ADM	NUR	MED	PAT
ADM	.2265 (.3900)	.2478 (.3672)	.1893 (.3119)	.1239 (.2490)
NUR		.2488 (.3244)	.2729 (.3491)	.2088 (.3016)
MED			.3310 (.3674)	.1398 (.2660)
PAT				.0124 (.0928)
	$\widetilde{\pi}_{i,j} = \bar{X}_{ij}$			
Status	ADM	NUR	MED	PAT
ADM	.1250 (.3312)	.1583 (.3652)	.1704 (.3764)	.0887 (.2845)
NUR		.1854 (.3887)	.1730 (.3784)	.1542 (.3612)
MED			.3901 (.4881)	.0927 (.2902)
PAT				.0090 (.0946)

be appropriate for this data set, as the P-value is 0.676. We repeat the analysis for $q = 3$, leading to equally plausible results: 3 biology classes are in one cluster, 3 mathematics and physics classes are in another cluster, and the 3 remaining classes form the 3rd cluster. See also Figure 4 for the graphical illustration with the 3 clusters. We remark that to avoid the plot to be overcrowded, only the random selected 25% of nodes in each cluster were used in Figures 3 and 4.

Table 5: French high school contact network data: the detected clusters by spectral clustering algorithm (SCA) based on either $\widehat{\mathbf{L}}$ in (17) or $\bar{\mathbf{X}}$. The number of clusters is set at $q = 2$.

	SCA based on $\widehat{\mathbf{L}}$		SCA based on $\bar{X}$	
Class	Cluster 1	Cluster 2	Cluster 1	Cluster 2
BIO1	0	37	1	36
BIO2	1	32	0	33
BIO3	1	39	0	40
MP1	33	0	33	0
MP2	29	0	29	0
MP3	38	0	38	0
PC1	44	0	44	0
PC2	39	0	39	0
EGI	34	0	34	0

To choose the number of clusters q objectively, we define the Bayesian information criteria (BIC) as follows:

$$\text{BIC}(q) = -2 \max \log(\text{likelihood}) + \log\{n(p/q)^2\}q(q+1).$$

For each fixed q, we effectively build $q(q+1)/2$ models independently and each model has 2 parameters $\theta_{k,\ell}$ and $\eta_{k,\ell}$, $1 \le k \le \ell \le q$. The number of the available observations for each model is approximately $n(p/q)^2$, assuming that the numbers of nodes in all the q clusters are about the same, which is then p/q. Thus the penalty term in the BIC above is $\sum_{1\le k\le \ell\le q} 2\log\{n(p/q)^2\} = \log\{n(p/q)^2\}q(q+1)$.

Table 6 lists the values of BIC(q) for different q. The minimum is obtained at $q = 9$, exactly the number of original classes in the school. Performing the SCA based on $\widehat{\mathbf{L}}$ with $q = 9$, we obtain almost perfect classification: all the 9 original classes are identified as the 9 clusters with only in total 4 students being placed outside their own classes. The estimated $\theta_{i,j}$ and $\eta_{i,j}$, together with their standard errors calculated based on the asymptotic normality presented in Theorem 13, are reported in Table 7. As $\widehat{\theta}_{i,j}$ for $i \ne j$ are very small (i.e. ≤ 0.027), the students from different classes who have not contacted with each other are unlikely to contact next day. See (14) and (3). On the other hand, as $\widehat{\eta}_{i,j}$ for $i \ne j$ are large (i.e. ≥ 0.761), the students from different classes who have contacted with each other are likely to lose the contacts next day. Note that $\widehat{\theta}_{i,i}$ are greater than $\widehat{\theta}_{i,j}$ for $i \ne j$ substantially, and $\widehat{\eta}_{i,i}$ are smaller than $\widehat{\eta}_{i,j}$ for $i \ne j$ substantially. This implies that the students in the same class are more likely to contact with each other than those across the different classes.

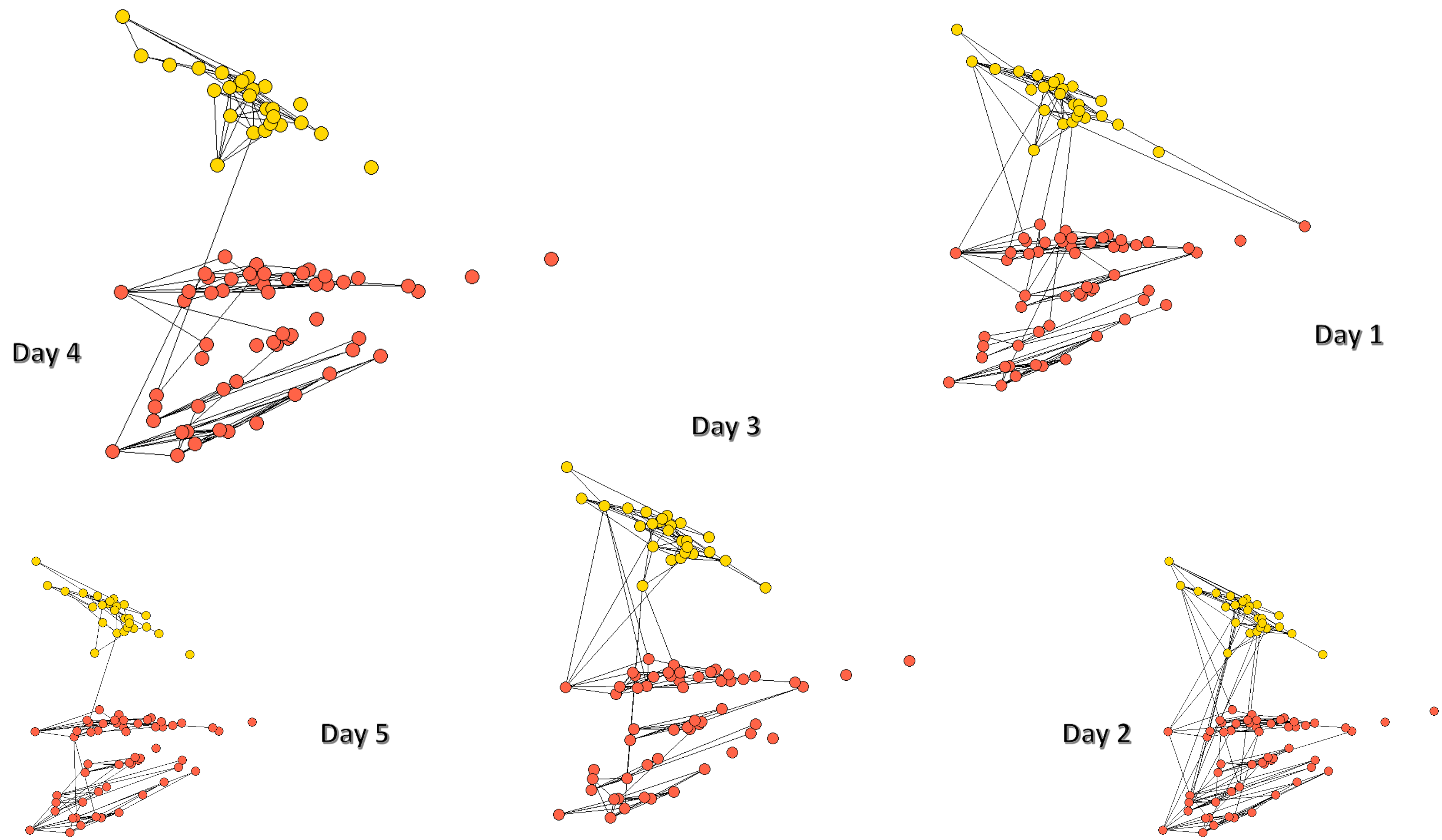

Figure 3: French high school contact networks over 5 days: the nodes marked in two colours represent the $q = 2$ clusters determined by SCA based on $\widehat{\mathbf{L}}$ in (17).

Table 6: Fitting AR(1) stochastic block models to the French high school data: BIC values for different cluster numbers q.

q	2	3	5	7	8	9	10	11
BIC(q)	43624	40586	37726	36112	35224	34943	35002	35120

To apply the variational EM algorithm of Matias and Miele (2017) to analyze this data set, we use the `R` package `dynsbm`. The algorithm is designed to identify time-varying dynamic stochastic block structure in the sense that both the membership of nodes and the transition probabilities may vary with time. Furthermore it also identifies the nodes not belonging to any clusters. The number of the clusters selected by the so-called integrated classification likelihood criterion is also 9. The identified 9 clusters are always dominated by the 9 original classes in the school, though they vary from day to day. The number of the identified students not belonging to any of the 9 clusters was 15, 17, 24, 32 and 28, respectively, in those 5 days. Furthermore the number of the students who were not put in their own classes was 14, 9, 12, 10 and 12, respectively. The more detailed results are reported in Appendix B. Those findings are less clear-cut than those obtained from our method above. This is hardly surprising as Matias and Miele (2017) adopts a general setting without imposing stationarity.

5.3 Global trade data

Our last example concerns the annual international trades among $p = 197$ countries between 1950 and 2014 (i.e. $n = 65$). We define an edge between two countries to be 1 as long as there exist trades between the two countries in that year (regardless the direction), and 0 otherwise. We take this simplistic approach to illustrate our AR(1) stochastic block model with a change point. The data used are a subset of the openly available trade data for 205 countries in 1870 − 2014 (Barbieri et al., 2009; Barbieri and Keshk, 2016). We leave out several countries, e.g. Russia and Yugoslavia, which did not exist for the whole period concerned.

Setting $q = 2$, we fit the data with an AR(1) stochastic block model with two clusters. The P-value of the permutation test for the residuals resulted from the fitted model is 0, indicating overwhelmingly that the stationarity does not hold for the whole period. Applying the maximum likelihood estimator (28), the estimated change point is at year 1991. Before this change point, the identified Cluster I contains 26 countries, including the most developed industrial countries such as USA, Canada, UK and most European countries. Cluster II contains 171 countries, including all African and Latin American countries, and most Asian countries. After 1991, 41 countries switched from Cluster II to Cluster I, including Argentina, Brazil, Bulgaria, China, Chile, Columbia, Costa Rica, Cyprus, Hungary, Israel, Japan, New Zealand, Poland, Saudi Arabia, Singapore, South Korea, Taiwan, and United Arab Emirates. There was no single switch from Cluster I to II. Note that 1990 may be viewed as the beginning of the globalization. With the collapse of the Soviet Union in 1989, the fall of Berlin Wall and the end of the Cold War in 1991, the world became more interconnected. The communist bloc countries in East Europe, which had

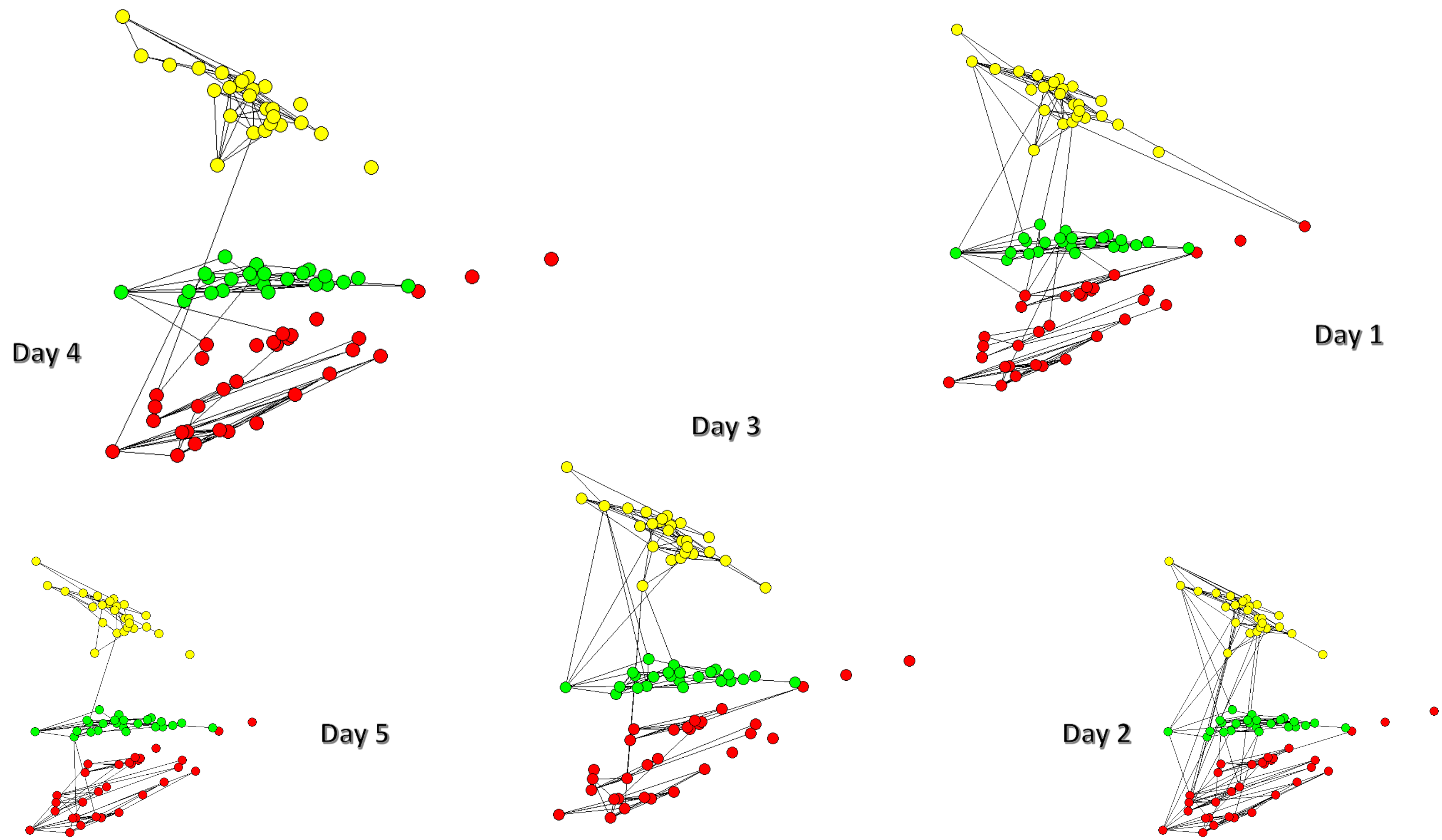

Figure 4: French high school contact networks over 5 days: the nodes marked in three colours represent the $q = 3$ clusters determined by SCA based on $\widehat{\mathbf{L}}$ in (17).

Table 7: Fitting AR(1) stochastic block models with $q = 9$ clusters to the French high school data: the estimation parameters and their standard errors (in parentheses).

	Cluster	1	2	3	4	5	6	7	8	9
	1	.246	.001	.004	.006	.001	.009	.003	.024	.003
		(.008)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.002)	(.001)
	2		.136	.024	.0018	.001	.007	.001	.001	.027
			(.009)	(.002)	(.001)	(.001)	(.001)	(.000)	(.001)	(.002)
	3			.252	.001	.002	.007	.001	.001	.022
				(.011)	(.001)	(.001)	(.001)	(.001)	(.001)	(.002)
	4				.234	.020	.001	.024	.002	.001
					(.010)	(.002)	(.001)	(.002)	(.001)	(.001)
$\widehat{\theta}_{i,j}$	5					.196	.001	.020	.002	.004
						(.008)	(.001)	(.002)	(.000)	(.001)
	6						.181	.001	.010	.007
							(.008)	(.001)	(.001)	(.001)
	7							.252	.003	.006
								(.009)	(.001)	(.001)
	8								.202	.001
									(.006)	(.001)
	9									.219
										(.008)
	1	.563	.999	.959	.976	.999	.867	.870	.792	.909
		(.015)	(.001)	(.036)	(.098)	(.001)	(.054)	(.001)	(.000)	(.051)
	2		.472	.761	.888	.999	.866	.999	.999	.866
			(.024)	(.036)	(.097)	(.001)	(.054)	(.001)	(.000)	(.026)
	3			.453	.999	.928	.864	.999	.999	.772
				(.016)	(.000)	(.066)	(.048)	(.000)	(.000)	(.031)
	4				.509	.868	.999	.784	.956	.999
					(.017)	(.028)	(.000)	(.029)	(.041)	(.000)
$\widehat{\eta}_{i,j}$	5					.544	.999	.929	.842	.935
						(.017)	(.001)	(.021)	(.078)	(.041)
	6						.589	.999	.793	.923
							(.019)	(.001)	(.040)	(.036)
	7							.480	.999	.814
								(.014)	(.000)	(.051)
	8								.504	.999
									(.127)	(.000)
	9									.471
										(.014)

been isolated from the capitalist West, began to integrate into the global market economy. Trade and investment increased, while barriers to migration and to cultural exchange were lowered.

Figure 5 presents the average adjacency matrix of the 197 countries before and after the change point, where the cold blue color indicates small value and the warm red color indicates large value. Before 1991, there are only 26 countries in Cluster 1. The intensive red in the small lower left corner indicates the intensive trades among those 26 countries. After 1991, the densely connected lower left corner is enlarged as now there are 67 countries in Cluster 1. Note some members of Cluster 2 also trade with the members of Cluster 1, though not all intensively.

The estimated parameters for the fitted AR(1) stochastic block model with $q = 2$ clusters are reported in Table 8. Since estimated values for $\widehat{\theta}_{1,2}, \widehat{\eta}_{1,2}$ before and after the change point are always small, the trading status between the countries across the two clusters are unlikely to change. Nevertheless $\widehat{\theta}_{1,2}$ is 0.154 after 1991, and 0.053 before 1991; indicating greater possibility for new trades to happen after 1991.

Table 8: Fitting AR(1) stochastic block model with a change point and $q = 2$ to the Global trade data: the estimated AR coefficients before and after 1991.

	$t \leq 1991$		$t > 1991$	
Coefficients	Estimates	SE	Estimates	SE
$\theta_{1,1}$	.062	.0092	.046	.0005
$\theta_{1,2}$	.053	.0008	.154	.0013
$\theta_{2,2}$	.023	.0002	.230	.0109
$\eta_{1,1}$	.003	.0005	.144	.0016
$\eta_{1,2}$	.037	.0008	.047	.0007
$\eta_{2,2}$	.148	.0012	.006	.0003

6. Miscellaneous remarks

We proposed in this paper a simple AR(1) setting to represent the dynamic dependence in network data explicitly. It also facilitates easy inference such as the maximum likelihood estimation and model diagnostic checking. A new class of dynamic stochastic block models illustrates the usefulness of the setting in handling more complex underlying structures including structure breaks due to change points.

Model (1) can be easily extended to higher-order autoregressive forms, e.g. linear AR(2) model

$$X_{i,j}^t = X_{i,j}^{t-1} I(\varepsilon_{i,j}^t = 1) + X_{i,j}^{t-2} I(\varepsilon_{i,j}^t = 2) + I(\varepsilon_{i,j}^t = 0),$$

or nonlinear AR(2)

$$X_{i,j}^t = X_{i,j}^{t-1} I(\varepsilon_{i,j}^t = 1) + X_{i,j}^{t-1} X_{i,j}^{t-2} I(\varepsilon_{i,j}^t = 2) + I(\varepsilon_{i,j}^t = 0),$$

where $\{\varepsilon_{i,j}^t,\, t \geq 0\}$ a sequence of i.i.d. innovations, and each $\varepsilon_{i,j}^t$ takes four possible values $-1, 0, 1, 2$. The nonlinear AR(2) above may be particularly appealing for some slowly

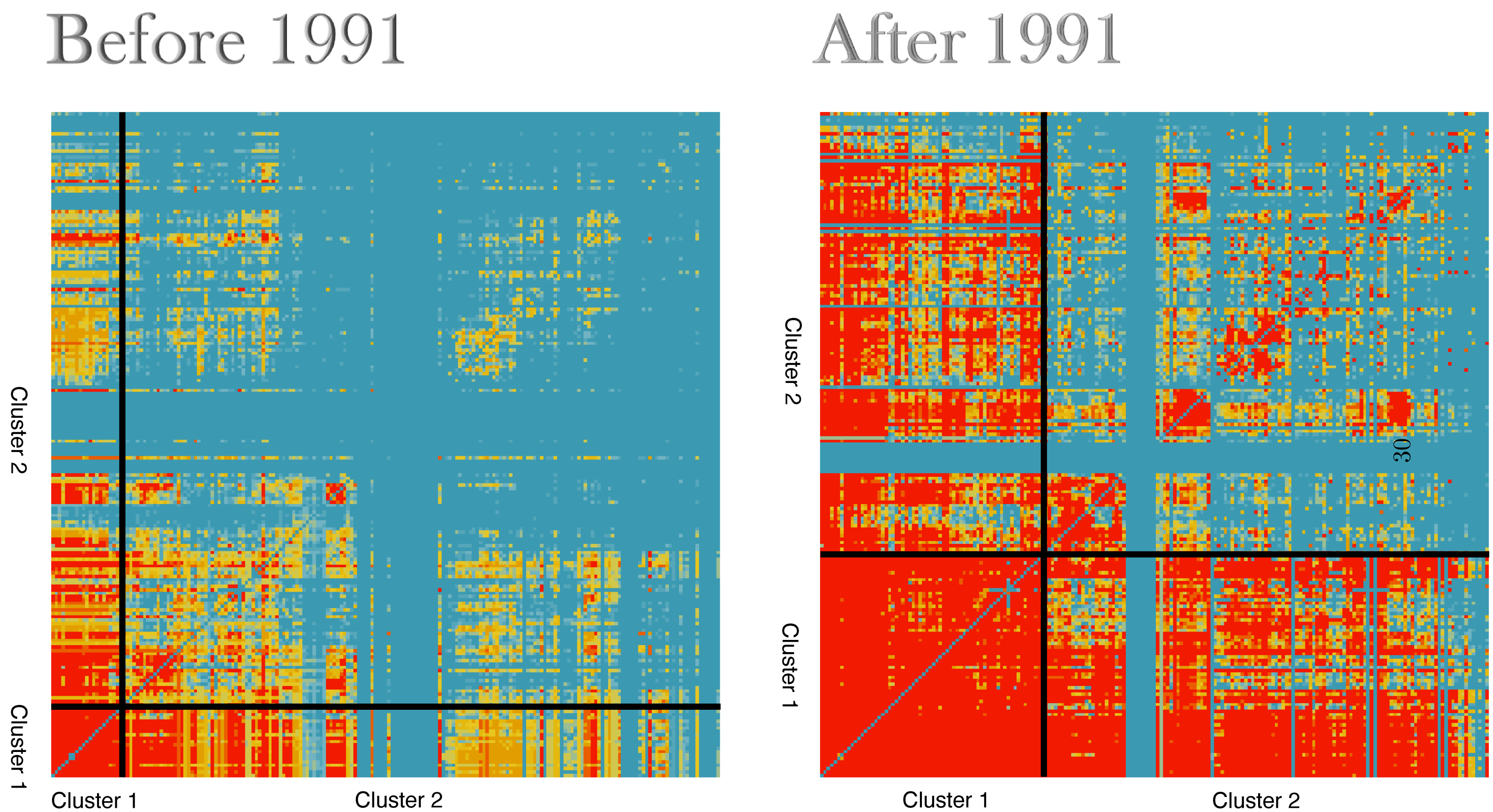

Figure 5: Average adjacency matrix for the trades among the 197 countries before and after 1991. The values from 0 to 1 are colour-coded from blue, light blue, light red to red.

changing networks. Furthermore we can model $X_{i,j}^t$ based on binary DARMA models of MacDonald and Zucchini (1997), Section 1.4.

Furthermore, a more fertile exploration is perhaps to tailor the model to incorporate various stylized features of network data, such as edge sparsity, node heterogeneity, transitivity and homophily. For example, to model the transitivity (i.e. the individuals with common friends are likely to become friends), we may let in (2) and (3)

$$\alpha_{i,j}^t = \frac{e^{aU_{i,j}^{(t-1)}}}{1 + e^{aU_{i,j}^{(t-1)}} + e^{bU_{i,j}^{(t-1)}}}, \qquad \beta_{i,j}^t = \frac{e^{bV_{i,j}^{(t-1)}}}{1 + e^{aU_{i,j}^{(t-1)}} + e^{bU_{i,j}^{(t-1)}}},$$

where a, b are unknown parameters, $U_{i,j}^t = \sum_k X_{i,k}^t X_{j,k}^t$, and $V_{i,j}^t = 0.5 \sum_k (X_{i,k}^t + X_{j,k}^t) - U_{i,j}^t$. The developments in those directions will be reported in some follow-up papers.

On the other hand, dynamic networks with weighted edges may be treated as matrix time series for which effective modelling procedures have been developed based on various tensor decompositions (Wang et al., 2019; Chang et al., 2023).

Acknowledgments

The authors would like to thank the action editor and the three anonymous referees for their valueable comments. Jiang was supported by the Hong Kong RGC Gerneral Research Fund 15302722 and the NSFC Grant 12001459. Li was supported in part by the MOE Tier 1 Academic Research Funds A-8000016-00-00 in Singapore. Yao was supported by the EPSRC grants EP/V007556/1 and EP/X002195/1.

References

Charu Aggarwal and Karthik Subbian. Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR)*, 47(1):1–36, 2014.

Katherine Barbieri and Omar M. G. Keshk. *Correlates of War Project Trade Data Set Codebook, Version 4.0.* Online: http://correlatesofwar.org., 2016.

Katherine Barbieri, Omar M. G. Keshk, and Brian Pollins. Trading data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science*, 26(5):471–491, 2009.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Monika Bhattacharjee, Moulinath Banerjee, and George Michailidis. Change point estimation in a dynamic stochastic block model. *Journal of Machine Learning Research*, 21 (107):1–59, 2020.

Richard C Bradley. *Introduction to strong mixing conditions.* Kendrick press, 2007.

Jinyuan Chang, Eric D Kolaczyk, and Qiwei Yao. Discussion of "network cross-validation by edge sampling". *Biometrika*, 107(2):277–280, 2020.

Jinyuan Chang, Eric D Kolaczyk, and Qiwei Yao. Estimation of subgraph densities in noisy networks. *Journal of the American Statistical Association*, 117(537):361–374, 2022.

Jinyuan Chang, Jin He, Lin Yang, and Qiwei Yao. Modelling matrix time series via a tensor cp-decomposition. *Journal of the Royal Statistical Society,* **B**, in press, 2023.

Elynn Y. Chen, Jianqing Fan, and Xuening Zhu. Community network auto-regression for high-dimensional time series. *Journal of Econometrics*, in press, 2023.

M. Corneli, P. Latouche, and F. Rossi. Multiple change points detection and clustering in dynamic networks. *Statistics and Computing*, 28(5):989–1007, 2018.

Harry Crane et al. Dynamic random networks and their graph limits. *The Annals of Applied Probability*, 26(2):691–721, 2016.

Claire Donnat and Susan Holmes. Tracking network dynamics: A survey of distances and similarity metrics. *The Annals of Applied Statistics*, 12(2):971–1012, 2018.

Daniele Durante, David B Dunson, et al. Locally adaptive dynamic networks. *The Annals of Applied Statistics*, 10(4):2203–2232, 2016.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Jianqing Fan and Qiwei Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York, 2003.

N Friel, R Rastelli, J Wyse, and AE Raftery. Interlocking directorates in irish companies using a latent space model for bipartite networks. *Proceedings of the national academy of sciences*, 113(24):6629–6634, 2016.

Wenjie Fu, Le Song, and Eric P Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 329–336, 2009.

Steve Hanneke, Wenjie Fu, and Eric P Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

Xinyu Kang, Apratim Ganguly, and Eric D Kolaczyk. Dynamic networks with multi-scale temporal structure. *Sankhya* **A**, 84:218–260, 2020.

M.I. Knight, M.A. Nunes, and G.P. Nason. Modelling, detrending and decorrelation of network time series. *arXiv preprint arXiv:1603.03221*, 2016.

Eric D Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis*. Cambridge University Press, 2017.

Pavel N Krivitsky and Mark S Handcock. A separable model for dynamic networks. *Journal of the Royal Statistical Society,* **B**, 76(1):29, 2014.

Zhengyan Lin and Zhidong Bai. *Probability inequalities*. Springer Science & Business Media, 2011.

Matthew Ludkin, Idris Eckley, and Peter Neal. Dynamic stochastic block models: parameter estimation and detection of changes in community structure. *Statistics and Computing*, 28(6):1201–1213, 2018.

Iain L. MacDonald and Walter Zucchini. *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London, 1997.

R Mastrandrea, J Fournet, and A Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, 10(9):e0136497, 2015.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society,* **B**, 79(4): 1119–1141, 2017.

Catherine Matias, Tabea Rebafka, and Fanny Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(5):989–1007, 2018.

Florence Merlevède, Magda Peligrad, Emmanuel Rio, et al. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

Marianna Pensky. Dynamic network models and graphon estimation. *Annals of Statistics*, 47(4):2378–2403, 2019.

R. Rastelli, P. Latouche, and N. Friel. Choosing the number of groups in a latent stochastic block model for dynamic networks. *Network Science*, 6(4):469–493, 2018.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

T. A. B. Snijders. Models for longitudinal network data. In P.J. Carrington, J. Scott, and S. S. Wasserman, editors, *Models and Methods in Social Network Analysis*, chapter 11. Cambridge University Press, New York, 2005.

P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Regis, B. a. Kim, and N. Voirin B. Comte. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS ONE*, 8:e73970, 2013.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *Annals of statistics*, 49(1):203–232, 2021.

Dong Wang, Xialu Liu, and Rong Chen. Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248, 2019.

James D Wilson, Nathaniel T Stevens, and William H Woodall. Modeling and detecting change in temporal networks via the degree corrected stochastic block model. *Quality and Reliability Engineering International*, 35(5):1363–1378, 2019.

Kevin Xu. Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087. PMLR, 2015.

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

Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networks?a bayesian approach. *Machine learning*, 82(2):157–189, 2011.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Elena Yudovina, Moulinath Banerjee, and George Michailidis. Changepoint inference for erdös-rényi random graphs. In *Stochastic Models, Statistics and Their Applications*, pages 197–205. Springer, 2015.

Zifeng Zhao, Li Chen, and Lizhen Lin. Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*, 2019.

Tingting Zhu, Ping Li, Lanlan Yu, Kaiqi Chen, and Yan Chen. Change point detection in dynamic networks based on community identification. *IEEE Transactions on Network Science and Engineering*, 2020a.

Xuening Zhu, Rui Pan, Guodong Li, Yuewen Liu, and Hansheng Wang. Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123, 2017.

Xuening Zhu, Weining Wang, Hansheng Wang, and Wolfgang Karl Härdle. Network quantile autoregression. *Journal of Econometrics*, 212(1):345–358, 2019.

Xuening Zhu, Danyang Huang, Rui Pan, and Hansheng Wang. Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics*, 215(2):591–606, 2020b.

"Autoregressive Networks" by B. Jiang, J. Li and Q. Yao

Appendix: Technical proofs and further real data analysis

A.1 Proof of Proposition 2

Note all $X_{i,j}^t$ take binary values 0 or 1. Hence

$$P(X_{i,j}^1 = 1) = P(X_{i,j}^0 = 1)P(X_{i,j}^1 = 1|X_{i,j}^0 = 1) + P(X_{i,j}^0 = 0)P(X_{i,j}^1 = 1|X_{i,j}^0 = 0)$$
$$= \pi_{i,j}(1-\beta_{i,j}) + (1-\pi_{i,j})\alpha_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}(1-\beta_{i,j}) + \frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\alpha_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}} = \pi_{i,j}.$$

Thus $\mathcal{L}(X_{i,j}^1) = \mathcal{L}(X_{i,j}^0)$. Since all $\mathbf{X}^t$ are Erdós-Renyi, $\mathcal{L}(\mathbf{X}^1) = \mathcal{L}(\mathbf{X}^0)$. Condition (5) ensures that $\{\mathbf{X}_t\}$ is a homogeneous Markov chain. Hence $\mathcal{L}(\mathbf{X}^t) = \mathcal{L}(\mathbf{X}^0)$ for any $t \geq 1$. This implies the required stationarity.

As $E(X_{i,j}^t) = P(X_{i,j}^t = 1)$, and $\text{Var}(X_{i,j}^t) = E(X_{i,j}^t) - \{E(X_{i,j}^t)\}^2$, (8) follows from the stationarity, (6) and (7).

Note that (1) implies a Yule-Walker equation

$$\gamma_{i,j}(k) = (1-\alpha_{i,j}-\beta_{i,j})\gamma_{i,j}(k-1), \quad k = 1, 2, \cdots, \tag{A.1}$$

where $\gamma_{i,j}(k) = \text{Cov}(X_{i,j}^{t+k}, X_{i,j}^t)$.

Since the networks are all Erdös-Renyi, (9) follows from the Yule-Walker equation (A.1) immediately, noting $\rho_{i,j}(k) = \gamma_{i,j}(k)/\gamma_{i,j}(0)$ and $\rho_{i,j}(0) = 1$. To prove (A.1), it follows from (1) that for any $k \geq 1$,

$$E(X_{i,j}^{t+k}X_{i,j}^t) = E(X_{i,j}^{t+k-1}X_{i,j}^t)P(\varepsilon_{i,j}^{t+k} = 0) + P(\varepsilon_{i,j}^{t+k} = 1)EX_{i,j}^t$$
$$= (1-\alpha_{i,j}-\beta_{i,j})E(X_{i,j}^{t+k-1}X_{i,j}^t) + \alpha_{i,j}^2/(\alpha_{i,j}+\beta_{i,j}).$$

Thus

$$\gamma_{i,j}(k) = E(X_{i,j}^{t+k}X_{i,j}^t) - (EX_{i,j}^t)^2 = E(X_{i,j}^{t+k}X_{i,j}^t) - \frac{\alpha_{i,j}^2}{(\alpha_{i,j}+\beta_{i,j})^2}$$
$$= (1-\alpha_{i,j}-\beta_{i,j})E(X_{i,j}^{t+k-1}X_{i,j}^t) + \frac{\alpha_{i,j}^2}{\alpha_{i,j}+\beta_{i,j}}(1 - \frac{1}{\alpha_{i,j}+\beta_{i,j}})$$
$$= (1-\alpha_{i,j}-\beta_{i,j})\{E(X_{i,j}^{t+k-1}X_{i,j}^t) - \frac{\alpha_{i,j}^2}{(\alpha_{i,j}+\beta_{i,j})^2}\} = (1-\alpha_{i,j}-\beta_{i,j})\gamma_{i,j}(k-1).$$

This completes the proof.

A.2 Proof of Proposition 4

We only prove (11), as (10) follows from (11) immediately. To prove (11), we only need to show

$$d_{i,j}(k) \equiv P(X_{i,j}^t \neq X_{i,j}^{t+k}) = \frac{2\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}\{1-(1-\alpha_{i,j}-\beta_{i,j})^k\}, \quad k = 1, 2, \cdots. \tag{A.2}$$

We Proceed by induction. It is easy to check that (A.2) holds for $k = 1$. Assuming it also holds for $k \geq 1$, then

$$\begin{aligned}
& d_{i,j}(k+1) = P(X_{i,j}^t = 0, X_{i,j}^{t+k+1} = 1) + P(X_{i,j}^t = 1, X_{i,j}^{t+k+1} = 0) \\
= & P(X_{i,j}^t = 0, X_{i,j}^{t+k} = 1, X_{i,j}^{t+k+1} = 1) + P(X_{i,j}^t = 0, X_{i,j}^{t+k} = 0, X_{i,j}^{t+k+1} = 1) \\
& + P(X_{i,j}^t = 1, X_{i,j}^{t+k} = 0, X_{i,j}^{t+k+1} = 0) + P(X_{i,j}^t = 1, X_{i,j}^{t+k} = 1, X_{i,j}^{t+k+1} = 0) \\
= & P(X_{i,j}^t = 0, X_{i,j}^{t+k} = 1)(1 - \beta_{i,j}) + \{P(X_{i,j}^t = 0) - P(X_{i,j}^t = 0, X_{i,j}^{t+k} = 1)\}\alpha_{i,j} \\
& + P(X_{i,j}^t = 1, X_{i,j}^{t+k} = 0)(1 - \alpha_{i,j}) + \{P(X_{i,j}^t = 1) - P(X_{i,j}^t = 1, X_{i,j}^{t+k} = 0)\}\beta_{i,j} \\
= & \{P(X_{i,j}^t = 0, X_{i,j}^{t+k} = 1) + P(X_{i,j}^t = 1, X_{i,j}^{t+k} = 0)\}(1 - \alpha_{i,j} - \beta_{i,j}) + \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} \\
= & d_{i,j}(k)(1 - \alpha_{i,j} - \beta_{i,j}) + \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} = \frac{2\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j} + \beta_{i,j})^2}\{1 - (1 - \alpha_{i,j} - \beta_{i,j})^{k+1}\}.
\end{aligned}$$

Hence (A.2) also holds for $k + 1$. This completes the proof.

A.3 Proof of Proposition 5

Proof Note that for any nonempty elements $A \in \mathcal{F}_0^k, B \in \mathcal{F}_{k+\tau}^\infty$, there exist $A_0 \in \mathcal{F}_0^{k-1}$ and $B_0 \in \mathcal{F}_{k+\tau+1}^\infty$ such that $A = A_0 \times \{0\}, A_0 \times \{1\}$, or $A_0 \times \{0, 1\}$, and $B = B_0 \times \{0\}, B_0 \times \{1\}$, or $B_0 \times \{0, 1\}$. We first consider the case where $B = B_0 \times \{x_k\}$ and $A = A_0 \times \{x_{k+\tau}\}$ where $x_k, x_{k+\tau} = 0$ or 1. Note that

$$\begin{aligned}
& P(A_0, X_{i,j}^k = x_k, B_0, X_{i,j}^{k+\tau} = x_{k+\tau}) \\
= \quad & P(B_0 | X_{i,j}^{k+\tau} = x_{k+\tau}) P(X_{i,j}^{k+\tau} = x_{k+\tau}, A_0, X_{i,j}^k = x_k) \\
= \quad & P(B_0, X_{i,j}^{k+\tau} = x_{k+\tau}) P(A_0, X_{i,j}^k = x_k) \cdot \frac{P(X_{i,j}^{k+\tau} = x_{k+\tau} | X_{i,j}^k = x_k)}{P(X_{i,j}^{k+\tau} = x_{k+\tau})} \\
= \quad & P(B_0, X_{i,j}^{k+\tau} = x_{k+\tau}) P(A_0, X_{i,j}^k = x_k) \cdot \frac{P(X_{i,j}^{k+\tau} = x_{k+\tau}, X_{i,j}^k = x_k)}{P(X_{i,j}^{k+\tau} = x_{k+\tau}) P(X_{i,j}^k = x_k)}
\end{aligned}$$

On the other hand, note that

$$P(X_{i,j}^{k+\tau} = 1, X_{i,j}^k = 1) - P(X_{i,j}^{k+\tau} = 1) P(X_{i,j}^k = 1) = \rho_{i,j}(\tau);$$

$$\begin{aligned}
& P(X_{i,j}^{k+\tau} = 1, X_{i,j}^k = 0) - P(X_{i,j}^{k+\tau} = 1) P(X_{i,j}^k = 0) \\
= \quad & P(X_{i,j}^{k+\tau} = 1) - P(X_{i,j}^{k+\tau} = 1, X_{i,j}^k = 1) - P(X_{i,j}^{k+\tau} = 1)[1 - P(X_{i,j}^k = 1)] \\
= \quad & -\rho_{i,j}(\tau);
\end{aligned}$$

$$\begin{aligned}
& P(X_{i,j}^{k+\tau} = 0, X_{i,j}^k = 1) - P(X_{i,j}^{k+\tau} = 0) P(X_{i,j}^k = 1) \\
= \quad & P(X_{i,j}^k = 1) - P(X_{i,j}^{k+\tau} = 1, X_{i,j}^k = 1) - [1 - P(X_{i,j}^{k+\tau} = 1)] P(X_{i,j}^k = 1) \\
= \quad & -\rho_{i,j}(\tau);
\end{aligned}$$

$$\begin{aligned}
& P(X_{i,j}^{k+\tau}=0, X_{i,j}^{k}=0) - P(X_{i,j}^{k+\tau}=0)P(X_{i,j}^{k}=0) \\
= \;& P(X_{i,j}^{k+\tau}=0) - P(X_{i,j}^{k+\tau}=0, X_{i,j}^{k}=1) - P(X_{i,j}^{k+\tau}=0)[1-P(X_{i,j}^{k}=1)] \\
= \;& \rho_{i,j}(\tau).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
& |P(A_0, X_{i,j}^{k}=x_k, B_0, X_{i,j}^{k+\tau}=x_{k+\tau}) - P(A_0, X_{i,j}^{k}=x_k)P(B_0, X_{i,j}^{k+\tau}=x_{k+\tau})| \\
= \;& \left| P(A_0, X_{i,j}^{k}=x_k)P(B_0, X_{i,j}^{k+\tau}=x_{k+\tau}) \left[\frac{P(X_{i,j}^{k+\tau}=x_{k+\tau}, X_{i,j}^{k}=x_k)}{P(X_{i,j}^{k+\tau}=x_{k+\tau})P(X_{i,j}^{k}=x_k)} - 1 \right] \right| \\
\leq \;& \rho_{i,j}(\tau).
\end{aligned}$$

In the case where $A = A_0 \times \{0,1\}$ and/or $B = B_0 \times \{0,1\}$, since A and B are nonempty, there exist integers $0 < k_1 < k$ and/or $k_2 > k+1$, and correspondingly $A_1 \in \mathcal{F}_0^{k_1-1} \times \{x_{k_1}\}$ and/or $B \in \mathcal{F}_{k_2+\tau+1}^{\infty} \times \{x_{k_2+\tau}\}$ with $x_{k_1}, x_{k_2+\tau} = 0$ or 1, such that $P(A\cap B) - P(A)P(B) = P(A_1 \cap B_1) - P(A_1)P(B_1)$. Following similar arguments above we have $P(A\cap B) - P(A)P(B) \leq \rho_{i,j}(\tau + k_2 - k_1) < \rho_{ij}(\tau)$. We thus proved that $\alpha^{i,j}(\tau) \leq \rho_{i,j}(\tau)$. The conclusion of Proposition 5 follows from Proposition 2. ■

A.4 Proof of Proposition 6

We introduce some technical lemmas first.

Lemma 15. *For any $(i,j) \in \mathcal{J}$, denote $Y_{i,j}^t := X_{i,j}^t(1 - X_{i,j}^{t-1})$, and let $\mathbf{Y}_t = (Y_{i,j}^t)_{1\leq i,j\leq p}$ be the $p\times p$ matrix at time t. Under the assumptions of Proposition 2, we have $\{\mathbf{Y}_t, t = 1, 2\ldots\}$ is stationary such that for any $(i,j), (l,m) \in \mathcal{J}$, and $t, s \geq 1$, $t \neq s$,*

$$EY_{i,j}^t = \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}, \quad \mathrm{Var}(Y_{i,j}^t) = \frac{\alpha_{i,j}\beta_{i,j}(\alpha_{i,j}+\beta_{i,j}-\alpha_{i,j}\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2},$$

$$\rho_{Y_{i,j}}(|t-s|) \equiv \mathrm{Corr}(Y_{i,j}^t, Y_{lm}^s) = \begin{cases} -\frac{\alpha_{i,j}\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^{|t-s|-1}}{\alpha_{i,j}+\beta_{i,j}-\alpha_{i,j}\beta_{i,j}} & \text{if } \ (i,j) = (l,m), \\ 0 & \text{otherwise.} \end{cases}$$

Proof Note that $Y_{i,j}^t = X_{i,j}^t(1 - X_{i,j}^{t-1}) = (1 - X_{i,j}^{t-1})I(\varepsilon_{i,j}^t = 1)$. We thus have:

$E(Y_{i,j}^t) = P(X_{i,j}^{t-1} = 0)\alpha_{i,j} = (1 - EX_{i,j}^{t-1})\alpha_{i,j} = \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}$.

$\mathrm{Var}(Y_{i,j}^t) = E(Y_{i,j}^t)[1 - E(Y_{i,j}^t)] = \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\left(1 - \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\right) = \frac{\alpha_{i,j}\beta_{i,j}(\alpha_{i,j}+\beta_{i,j}-\alpha_{i,j}\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2}$.

For $k = 1$ we have $E(Y_{i,j}^t Y_{i,j}^{t+1}) = E[(1 - X_{i,j}^{t-1})X_{i,j}^t(1 - X_{i,j}^t)X_{i,j}^{t+1}] = 0$. For any $k \geq 2$, using the fact that $E(X_{ij}^t X_{ij}^{t+k}) = \frac{\alpha_{ij}}{(\alpha_{ij}+\beta_{ij})^2}\{\beta_{ij}(1 - \alpha_{ij} - \beta_{ij})^k + \alpha_{ij}\}$, we have

$$\begin{aligned} E(Y_{i,j}^t Y_{i,j}^{t+k}) &= E[X_{i,j}^t(1 - X_{i,j}^{t-1})(1 - X_{i,j}^{t+k-1})I(\varepsilon_{i,j}^{t+k} = 1)] \\ &= \alpha_{i,j}E[X_{i,j}^t(1 - X_{i,j}^{t-1})(1 - X_{i,j}^{t+k-1})] \\ &= \alpha_{i,j}P(X_{i,j}^{t+k-1} = 0|X_{i,j}^t = 1)P(X_{i,j}^t = 1|X_{i,j}^{t-1} = 0)P(X_{i,j}^{t-1} = 0) \\ &= \frac{\alpha_{i,j}^2\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}[1 - P(X_{i,j}^{t+k-1} = 1|X_{i,j}^t = 1)] \\ &= \frac{\alpha_{i,j}^2\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}\left[1 - \frac{E(X_{i,j}^{t+k-1}X_{i,j}^t)}{EX_{i,j}^t}\right] \\ &= \frac{\alpha_{i,j}^2\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}\left[1 - \frac{\beta_{i,j}(1 - \alpha_{i,j} - \beta_{i,j})^{k-1} + \alpha_{i,j}}{\alpha_{i,j} + \beta_{i,j}}\right] \\ &= \frac{\alpha_{i,j}^2\beta_{i,j}^2[1 - (1 - \alpha_{i,j} - \beta_{i,j})^{k-1}]}{(\alpha_{i,j} + \beta_{i,j})^2}. \end{aligned}$$

Therefore we have for any $k \geq 1$,

$$\begin{aligned} \text{Cov}(Y_{i,j}^t, Y_{i,j}^{t+k}) &= E(Y_{i,j}^t Y_{i,j}^{t+k}) - EY_{i,j}^t EY_{i,j}^{t+k} \\ &= \frac{\alpha_{i,j}^2\beta_{i,j}^2[1 - (1 - \alpha_{i,j} - \beta_{i,j})^{k-1}]}{(\alpha_{i,j} + \beta_{i,j})^2} - \frac{\alpha_{i,j}^2\beta_{i,j}^2}{(\alpha_{i,j} + \beta_{i,j})^2} \\ &= -\frac{\alpha_{i,j}^2\beta_{i,j}^2(1 - \alpha_{i,j} - \beta_{i,j})^{k-1}}{(\alpha_{i,j} + \beta_{i,j})^2}. \end{aligned}$$

Consequently, for any $|t - s| = 1, 2, \ldots$, the ACF of the process $\{Y_{i,j}^t, t = 1, 2\ldots\}$ is given as:

$$\begin{aligned} \rho_{Y_{i,j}}(|t - s|) &= -\frac{\alpha_{i,j}^2\beta_{i,j}^2(1 - \alpha_{i,j} - \beta_{i,j})^{|t-s|-1}}{(\alpha_{i,j} + \beta_{i,j})^2} \cdot \frac{(\alpha_{i,j} + \beta_{i,j})^2}{\alpha_{i,j}\beta_{i,j}(\alpha_{i,j} + \beta_{i,j} - \alpha_{i,j}\beta_{i,j})} \\ &= -\frac{\alpha_{i,j}\beta_{i,j}(1 - \alpha_{i,j} - \beta_{i,j})^{|t-s|-1}}{\alpha_{i,j} + \beta_{i,j} - \alpha_{i,j}\beta_{i,j}}. \end{aligned}$$

■

Since the mixing property is hereditary, $Y_{i,j}^t$ is also α-mixing. From Proposition 5 and Theorem 1 of Merlevède et al. (2009), we obtain the following concentration inequalities:

Lemma 16. *Let conditions (2.5) and C1 hold. There exist positive constants* C_1 *and* C_2 *such that for all* $n \geq 4$ *and* $\varepsilon < \frac{1}{(\log n)(\log\log n)}$,

$$P\left(\left|n^{-1}\sum_{t=1}^n X_{i,j}^t - EX_{i,j}^t\right| > \varepsilon\right) \leq \exp\{-C_1 n\varepsilon^2\},$$
$$P\left(\left|n^{-1}\sum_{t=1}^n Y_{i,j}^t - EY_{i,j}^t\right| > \varepsilon\right) \leq \exp\{-C_2 n\varepsilon^2\}.$$

Now we are ready to prove Proposition 6.

Proof of Proposition 6:

Let $\varepsilon = C_0\sqrt{\frac{\log p}{n}}$ with $C_0^2 C_1 > 2$ and $C_0^2 C_2 > 2$. Note that under condition (C2) we have $\varepsilon = o\big(\frac{1}{(\log n)(\log\log n)}\big)$. Consequently by Lemma 16, Proposition 2 and Lemma 15, we have

$$P\left(\left|n^{-1}\sum_{t=1}^{n} X_{i,j}^t - \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\right| > C_0\sqrt{\frac{\log p}{n}}\right) \le \exp\{-C_0^2 C_1 \log p\},$$

$$P\left(\left|n^{-1}\sum_{t=1}^{n} Y_{i,j}^t - \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\right| > C_0\sqrt{\frac{\log p}{n}}\right) \le \exp\{-C_0^2 C_2 \log p\}.$$

Consequently, with probability greater than $1-\exp\{-C_0^2C_1\log p\}-\exp\{-C_0^2C_2\log p\}$,

$$\frac{\frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} - C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \frac{1}{n} + C_0\sqrt{\frac{\log p}{n}}} \le \widehat{\alpha}_{i,j} \le \frac{\frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} - \frac{1}{n} - C_0\sqrt{\frac{\log p}{n}}}.$$

Note that when n and $\frac{n}{\log p}$ are large enough such that, $\frac{1}{n} \le C_0\sqrt{\frac{\log p}{n}} \le l/4$, we have

$$\alpha_{i,j} - \frac{\frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} - C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \frac{1}{n} + C_0\sqrt{\frac{\log p}{n}}} \le \frac{2C_0\alpha_{i,j}\sqrt{\frac{\log p}{n}} + C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}} \le 3l^{-1}C_0\sqrt{\frac{\log p}{n}},$$

and

$$\frac{\frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} - \frac{1}{n} - C_0\sqrt{\frac{\log p}{n}}} - \alpha_{i,j} \le \frac{2C_0\alpha_{i,j}\sqrt{\frac{\log p}{n}} + C_0\sqrt{\frac{\log p}{n}}}{\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} - \frac{l}{2}} \le 6l^{-1}C_0\sqrt{\frac{\log p}{n}},$$

Therefore we conclude that when when n and $\frac{n}{\log p}$ are large enough,

$$P\left(|\widehat{\alpha}_{i,j} - \alpha_{i,j}| \ge 6l^{-1}C_0\sqrt{\frac{\log p}{n}}\right) \le \exp\{-C_0^2C_1\log p\} + \exp\{-C_0^2C_2\log p\}. \tag{A.3}$$

For any $c > 2$, the concentration inequalities in Proposition 6 can then be concluded by setting $C_0 = \max\{\sqrt{c/C_1}, \sqrt{c/C_2}\}$. Further, as $n, p \to \infty$, we immediately have $\max_{(i,j)\in\mathcal{J}} |\widehat{\alpha}_{i,j} - \alpha_{i,j}| = O_p\left(\sqrt{\frac{\log p}{n}}\right)$. Convergence of $\widehat{\beta}_{i,j}$ can be proved similarly.

A.5 Proof of Proposition 7

Note that the log-likelihood function for $(\alpha_{i,j}, \beta_{i,j})$ is:

$$l(\alpha_{i,j},\beta_{i,j}) = \log(\alpha_{i,j})\sum_{t=1}^{n} X_{i,j}^t(1-X_{i,j}^{t-1}) + \log(1-\alpha_{i,j})\sum_{t=1}^{n}(1-X_{i,j}^t)(1-X_{i,j}^{t-1})$$
$$+ \log(\beta_{i,j})\sum_{t=1}^{n}(1-X_{i,j}^t)X_{i,j}^{t-1} + \log(1-\beta_{i,j})\sum_{t=1}^{n} X_{i,j}^t X_{i,j}^{t-1}.$$

Our first observation is that, owing to the independent edge formation assumption, all the $(\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}), (i,j) \in \mathcal{J}$ pairs are independent. For each pair $(\alpha_{i,j}, \beta_{i,j})$, the score equations of the log-likelihood function are:

$$
\begin{aligned}
\frac{\partial l(\alpha_{i,j}, \beta_{i,j})}{\partial \alpha_{i,j}} &= \frac{1}{\alpha_{i,j}} \sum_{t=1}^{n} X_{i,j}^{t}(1 - X_{i,j}^{t-1}) - \frac{1}{1-\alpha_{i,j}} \sum_{t=1}^{n} (1 - X_{i,j}^{t})(1 - X_{i,j}^{t-1}), \\
&= \left(\frac{1}{\alpha_{i,j}} + \frac{1}{1-\alpha_{i,j}} \right) \sum_{t=1}^{n} Y_{i,j}^{t} - \frac{1}{1-\alpha_{i,j}} \sum_{t=1}^{n} (1 - X_{i,j}^{t}) + O(1), \\
\frac{\partial l(\alpha_{i,j}, \beta_{i,j})}{\partial \beta_{i,j}} &= \frac{1}{\beta_{i,j}} \sum_{t=1}^{n} (1 - X_{i,j}^{t}) X_{i,j}^{t-1} - \frac{1}{1-\beta_{i,j}} \sum_{t=1}^{n} X_{i,j}^{t} X_{i,j}^{t-1} \\
&= \frac{1}{\beta_{i,j}} \sum_{t=1}^{n} X_{i,j}^{t-1} + \left(\frac{1}{\beta_{i,j}} + \frac{1}{1-\beta_{i,j}} \right) \sum_{t=1}^{n} (Y_{i,j}^{t} - X_{i,j}^{t}) \\
&= \left(\frac{1}{\beta_{i,j}} + \frac{1}{1-\beta_{i,j}} \right) \sum_{t=1}^{n} Y_{i,j}^{t} - \frac{1}{1-\beta_{i,j}} \sum_{t=1}^{n} X_{i,j}^{t} + O(1).
\end{aligned}
$$

Clearly, for any $0 < \alpha_{i,j}, \beta_{i,j}, \alpha_{i,j} + \beta_{i,j} \leq 1$, $\left(\frac{1}{\alpha_{i,j}} + \frac{1}{1-\alpha_{i,j}}, \frac{1}{1-\alpha_{i,j}} \right)$ and $\left(\frac{1}{\beta_{i,j}} + \frac{1}{1-\beta_{i,j}}, \frac{-1}{1-\beta_{i,j}} \right)$ are linearly independent. On the other hand, from Proposition 5, Lemma 16 and classical central limit theorems for weakly dependent sequences (Bradley, 2007; Durrett, 2019), we have $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Y_{i,j}^{t}$ and $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_{i,j}^{t}$ and any of their nontrivial linear combinations are asymptotically normally distributed. Consequently, any nontrivial linear combination of $\frac{1}{\sqrt{n}} \frac{\partial l(\alpha_{i,j}, \beta_{i,j})}{\partial \alpha_{i,j}}$, $(i,j) \in J_1$ and $\frac{1}{\sqrt{n}} \frac{\partial l(\alpha_{i,j}, \beta_{i,j})}{\partial \beta_{i,j}}, (i,j) \in J_2$ converges to a normal distribution. By standard arguments for consistency of MLEs, we conclude that $(\sqrt{n}(\widehat{\alpha}_{i,j} - \alpha_{i,j}), \sqrt{n}(\widehat{\beta}_{i,j} - \beta_{i,j}))'$ converges to the normal distribution with mean $\mathbf{0}$ and covariance matrix $I(\alpha_{i,j}, \beta_{i,j})^{-1}$, where $I(\alpha_{i,j}, \beta_{i,j})$ is the Fisher information matrix given as:

$$
I(\alpha_{i,j}, \beta_{i,j}) = \frac{1}{n} E \begin{bmatrix} \frac{\sum_{t=1}^{n} X_{i,j}^{t}(1-X_{i,j}^{t-1})}{\alpha_{i,j}^{2}} + \frac{\sum_{t=1}^{n}(1-X_{i,j}^{t})(1-X_{i,j}^{t-1})}{(1-\alpha_{i,j})^{2}} & 0 \\ 0 & \frac{\sum_{t=1}^{n}(1-X_{i,j}^{t})X_{i,j}^{t-1}}{\beta_{i,j}^{2}} + \frac{\sum_{t=1}^{n} X_{i,j}^{t} X_{i,j}^{t-1}}{(1-\beta_{i,j})^{2}} \end{bmatrix}.
$$

Note that

$$
\begin{aligned}
&\frac{1}{n} E \sum_{t=1}^{n} X_{i,j}^{t}(1 - X_{i,j}^{t-1}) = \frac{1}{n} E \sum_{t=1}^{n} (1 - X_{i,j}^{t}) X_{i,j}^{t-1} = \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}, \\
&\frac{1}{n} E \sum_{t=1}^{n} (1 - X_{i,j}^{t})(1 - X_{i,j}^{t-1}) = \frac{\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} - \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} = \frac{(1 - \alpha_{i,j})\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}, \\
&\frac{1}{n} E \sum_{t=1}^{n} X_{i,j}^{t} X_{i,j}^{t-1} = \frac{\alpha_{i,j}(1 - \beta_{i,j})}{\alpha_{i,j} + \beta_{i,j}}.
\end{aligned}
$$

We thus have

$$\begin{aligned} I(\alpha_{i,j},\beta_{i,j}) &= \begin{bmatrix} \frac{\beta_{i,j}}{\alpha_{i,j}(\alpha_{i,j}+\beta_{i,j})} + \frac{\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})(1-\alpha_{i,j})} & 0 \\ 0 & \frac{\alpha_{i,j}}{\beta_{i,j}(\alpha_{i,j}+\beta_{i,j})} + \frac{\alpha_{i,j}}{(1-\beta_{i,j})(\alpha_{i,j}+\beta_{i,j})} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta_{i,j}}{\alpha_{i,j}(\alpha_{i,j}+\beta_{i,j})(1-\alpha_{i,j})} & 0 \\ 0 & \frac{\alpha_{i,j}}{\beta_{i,j}(\alpha_{i,j}+\beta_{i,j})(1-\beta_{i,j})} \end{bmatrix}. \end{aligned}$$

Consequently, we have

$$\begin{bmatrix} \sqrt{n}(\widehat{\alpha}_{i,j}-\alpha_{i,j}) \\ \sqrt{n}(\widehat{\beta}_{i,j}-\beta_{i,j}) \end{bmatrix} \to N\left(\mathbf{0}, \begin{bmatrix} \frac{\alpha_{i,j}(\alpha_{i,j}+\beta_{i,j})(1-\alpha_{i,j})}{\beta_{i,j}} & 0 \\ 0 & \frac{\beta_{i,j}(\alpha_{i,j}+\beta_{i,j})(1-\beta_{i,j})}{\alpha_{i,j}} \end{bmatrix}\right).$$

This together with the independence among the $(\widehat{\alpha}_{i,j},\widehat{\beta}_{i,j}),(i,j)\in\mathcal{J}$ pairs proves the proposition.

A.6 Proof of Proposition 9

Denote $\mathbf{N}=\text{diag}\{\sqrt{s_1},\dots,\sqrt{s_q}\}$. Note that

$$\begin{aligned} \mathbf{L} &= \mathbf{D}_1^{-1/2}\mathbf{Z}\mathbf{\Omega}_1\mathbf{Z}^\top\mathbf{D}_1^{-1/2} + \mathbf{D}_2^{-1/2}\mathbf{Z}\mathbf{\Omega}_2\mathbf{Z}^\top\mathbf{D}_2^{-1/2} \\ &= \mathbf{Z}\widetilde{\mathbf{D}}_1^{-1/2}\mathbf{\Omega}_1\widetilde{\mathbf{D}}_1^{-1/2}\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{D}}_2^{-1/2}\mathbf{\Omega}_2\widetilde{\mathbf{D}}_2^{-1/2}\mathbf{Z}^\top \\ &= \mathbf{Z}(\widetilde{\mathbf{\Omega}}_1+\widetilde{\mathbf{\Omega}}_2)\mathbf{Z}^\top \\ &= (\mathbf{Z}\mathbf{N}^{-1})\mathbf{N}\widetilde{\mathbf{\Omega}}\mathbf{N}(\mathbf{Z}\mathbf{N}^{-1})^\top. \end{aligned}$$

Note that the columns of $\mathbf{Z}\mathbf{N}^{-1}$ are orthonormal, we thus have $rank(\mathbf{L})=q$. Let $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top=\mathbf{N}\widetilde{\mathbf{\Omega}}\mathbf{N}$ be the eigen-decomposition of $\mathbf{N}\widetilde{\mathbf{\Omega}}\mathbf{N}$, we immediately have $\mathbf{L}=(\mathbf{Z}\mathbf{N}^{-1})\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top(\mathbf{Z}\mathbf{N}^{-1})^\top$. Again, since the columns of $\mathbf{Z}\mathbf{N}^{-1}$ are orthonormal, we conclude that $\mathbf{\Gamma}_q=\mathbf{Z}\mathbf{N}^{-1}\mathbf{Q}$, and $\mathbf{U}=\mathbf{N}^{-1}\mathbf{Q}$. On the other hand, note that $\mathbf{U}$ is invertible, we conclude that $\mathbf{z}_{i,\cdot}\mathbf{U}=\mathbf{z}_{j,\cdot}\mathbf{U}$ and $\mathbf{z}_{i,\cdot}=\mathbf{z}_{j,\cdot}$ are equivalent.

A.7 Proof of Theorem 10

The key step is to establish an upper bound for the Frobenius norm $\|\widehat{\mathbf{L}}\widehat{\mathbf{L}}-\mathbf{L}\mathbf{L}\|_F$, and the theorem can be proved by Weyl's inequality and the Davis-Kahan theorem. We first introducing some technical lemmas.

Lemma 17. *Under the assumptions of Proposition 2, we have, there exists a constant $C_l>0$ such that*

$$\begin{aligned} &Cov\left(\sum_{t=1}^n Y_{i,j}^t, \sum_{t=1}^n (1-X_{i,j}^{t-1})\right) = -Cov\left(\sum_{t=1}^n Y_{i,j}^t, \sum_{t=1}^n X_{i,j}^{t-1}\right) \\ = &\quad \frac{n\alpha_{i,j}\beta_{i,j}[2\alpha_{i,j}(1-\beta_{i,j})+\alpha_{i,j}+\beta_{i,j}-2\beta_{i,j}^2]}{(\alpha_{i,j}+\beta_{i,j})^3} + C_{i,j}, \end{aligned}$$

with $|C_{i,j}|\le C_l$ for any $C_{i,j},(i,j)\in\mathcal{J}$.

Proof In the following we shall be using the fact that for any $0 \le x < 1$, $\sum_{h=1}^{n-1} x^{h-1} = \frac{1-x^n}{1-x} = \frac{1}{1-x} + o(1)$, and $\sum_{h=1}^{n-1} hx^{h-1} = \frac{1-x^n-n(1-x)x^{n-1}}{(1-x)^2} = O(1)$. In particular, when $x = 1-\alpha_{i,j}-\beta_{i,j}$, under condition C1, we have $2l \le 1-x < 1$, the $O(1)$ term in will become bounded uniformly for any $(i,j) \in \mathcal{J}$. In what follows, with some abuse of notation, we shall use $O_l(1)$ to denote a generic constant term with magnitude bounded by a large enough constant C_l that depends on l only.

$$
\begin{aligned}
& Cov\left(\sum_{t=1}^{n} Y_{i,j}^t, \sum_{t=1}^{n}(1-X_{i,j}^{t-1})\right) = -Cov\left(\sum_{t=1}^{n} Y_{i,j}^t, \sum_{t=1}^{n} X_{i,j}^{t-1}\right) \\
= & -\sum_{t=1}^{n}\sum_{s=1}^{n}\left[E(1-X_{i,j}^{t-1})X_{i,j}^{t}X_{i,j}^{s-1} - \frac{\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\cdot\frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}\right] \\
= & -\sum_{t=1}^{n}\sum_{s=1}^{n}\left\{\frac{\alpha_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}\left[\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^{|t-s+1|}+\alpha_{i,j}\right] - \frac{\alpha_{i,j}^2\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}\right\} \\
& +\sum_{t=1}^{n}\sum_{s=1}^{n} E(X_{i,j}^{t-1}X_{i,j}^{t}X_{i,j}^{s-1}) \\
= & -\sum_{t=1}^{n}\sum_{s=1}^{n}\frac{\alpha_{i,j}\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^{|t-s+1|}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{n^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} \\
& +(2n-1)E(X_{i,j}^{t-1}X_{i,j}^{t}) + \sum_{s<t} E(X_{i,j}^{t-1}X_{i,j}^{t}X_{i,j}^{s-1}) + \sum_{s>t+1} E(X_{i,j}^{t-1}X_{i,j}^{t}X_{i,j}^{s-1}). \quad \text{(A.4)}
\end{aligned}
$$

For the first three terms on the right hand side of (A.4), we have

$$
\begin{aligned}
& -\sum_{t=1}^{n}\sum_{s=1}^{n}\frac{\alpha_{i,j}\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^{|t-s+1|}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{n^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} + (2n-1)E(X_{i,j}^{t-1}X_{i,j}^{t}) \\
= & -\frac{\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2}\left[n + \frac{2n(1-\alpha_{i,j}-\beta_{i,j})}{\alpha_{i,j}+\beta_{i,j}}\right] - \frac{n^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} \\
& +\frac{2n\alpha_{i,j}\left[\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})+\alpha_{i,j}\right]}{(\alpha_{i,j}+\beta_{i,j})^2} + O_l(1) \\
= & \frac{3n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{2n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} - \frac{2n\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \frac{2n\alpha_{i,j}^2}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{n^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} + O_l(1).
\end{aligned}
$$

For the last two terms on the right hand side of (A.4), we have

$$\begin{aligned}
& \sum_{s<t} E(X_{i,j}^{t-1} X_{i,j}^{t} X_{i,j}^{s-1}) + \sum_{s>t+1} E(X_{i,j}^{t-1} X_{i,j}^{t} X_{i,j}^{s-1}) \\
= \; & \sum_{s<t} P(X_{i,j}^{t} = 1 | X_{i,j}^{t-1} = 1) P(X_{i,j}^{t-1} = 1, X_{i,j}^{s-1} = 1) \\
& + \sum_{s>t+1} P(X_{i,j}^{s-1} = 1 | X_{i,j}^{t} = 1) P(X_{i,j}^{t} = 1, X_{i,j}^{t-1} = 1) \\
= \; & (1-\beta_{i,j}) \sum_{s<t} E(X_{i,j}^{t-1} X_{i,j}^{s-1}) + (1-\beta_{i,j}) \sum_{s>t+1} E(X_{i,j}^{s-1} X_{i,j}^{t}) \\
= \; & \frac{(1-\beta_{i,j})\alpha_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} \sum_{h=1}^{n-1} (n-h)[\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^h + \alpha_{i,j}] \\
& + \frac{(1-\beta_{i,j})\alpha_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} \sum_{h=2}^{n-1} (n-h)[\beta_{i,j}(1-\alpha_{i,j}-\beta_{i,j})^{h-1} + \alpha_{i,j}] \\
= \; & \frac{(n-1)^2 \alpha_{i,j}^2 (1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} + \frac{2n(1-\beta_{i,j})\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} + O_l(1).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
& Cov\left(\sum_{t=1}^{n} Y_{i,j}^{t}, \sum_{t=1}^{n} (1 - X_{i,j}^{t-1})\right) = -Cov\left(\sum_{t=1}^{n} Y_{i,j}^{t}, \sum_{t=1}^{n} X_{i,j}^{t-1}\right) \\
= \; & \frac{3n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{2n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} - \frac{2n\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \frac{2n\alpha_{i,j}^2}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{n^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} \\
& + \frac{(n-1)^2\alpha_{i,j}^2(1-\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^2} + \frac{2n(1-\beta_{i,j})\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} + O_l(1) \\
= \; & \frac{3n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{2n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} - \frac{2n\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \frac{2n\alpha_{i,j}^2\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} \\
& + \frac{2n(1-\beta_{i,j})\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^3} + O_l(1) \\
= \; & \frac{3n\alpha_{i,j}\beta_{i,j}}{(\alpha_{i,j}+\beta_{i,j})^2} - \frac{2n\alpha_{i,j}\beta_{i,j}^2(1+\alpha_{i,j}+\beta_{i,j})}{(\alpha_{i,j}+\beta_{i,j})^3} + O_l(1).
\end{aligned}$$

This proves the lemma. ■

Lemma 18. (Bias of $\widehat{\alpha}_{i,j}$ and $\widehat{\beta}_{i,j}$) *Ket conditions C1, C2 and the assumptions of Proposition 2 hold. We have*

$$E\widehat{\alpha}_{i,j} - \alpha_{i,j} = -\frac{\alpha_{i,j}[2\alpha_{i,j}(1-\beta_{i,j}) + \alpha_{i,j} + \beta_{i,j} - 2\beta_{i,j}^2]}{n(\alpha_{i,j}+\beta_{i,j})\beta_{i,j}} + \frac{R_{i,j}^{(1)}}{n} + O(n^{-2}),$$

$$E\widehat{\beta}_{i,j} - \beta_{i,j} = \frac{\beta_{i,j}[2\alpha_{i,j}(1-\beta_{i,j}) + \alpha_{i,j} + \beta_{i,j} - 2\beta_{i,j}^2]}{n(\alpha_{i,j}+\beta_{i,j})\alpha_{i,j}} + \frac{R_{i,j}^{(2)}}{n} + O(n^{-2}),$$

where $R_{i,j}^{(1)}$ and $R_{i,j}^{(2)}$ are constants such that when n is large enough we have $0 \le R_{i,j}^{(1)}, R_{i,j}^{(2)} \le R_l$ for some constant R_l and all $(i,j) \in \mathcal{J}$.

Proof From Lemma 16 we have, under Condition C2, the event $\{|n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}| \le (1-\pi_{i,j})/2, 1 \le i,j \le p\}$ holds with probability larger than $1 - O(n^{-2})$. Denote $\mathcal{I} := I(|n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}| \le (1-\pi_{i,j})/2, 1 \le i,j \le p)$. By expanding $\frac{1}{1-n^{-1}\sum_{t=1}^n X_{i,j}^{t-1}}$ around $\frac{1}{1-\pi_{i,j}}$, we have

$$
\begin{aligned}
E\widehat{\alpha}_{i,j}\mathcal{I} &= E\frac{n^{-1}\sum_{t=1}^n X_{i,j}^t(1-X_{i,j}^{t-1})}{n^{-1}\sum_{t=1}^n(1-X_{i,j}^{t-1})}\mathcal{I} \\
= & \frac{1}{n}E\sum_{t=1}^n X_{i,j}^t(1-X_{i,j}^{t-1})\left[\frac{1}{1-\pi_{i,j}} + \frac{(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j})}{(1-\pi_{i,j})^2} + \sum_{k=2}^{\infty}\frac{(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j})^k}{(1-\pi_{i,j})^{k+1}}\right]\mathcal{I}.
\end{aligned}
$$

Write $R_{i,j}^{(1)} := E\sum_{t=1}^n X_{i,j}^t(1-X_{i,j}^{t-1})\Big(\sum_{k=2}^{\infty}\frac{(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1}-\pi_{i,j})^k}{(1-\pi_{i,j})^{k+1}}\Big)\mathcal{I}$. By Taylor series with Lagrange remainder we have there exist random scalars $r_{i,j}^t \in [n^{-1}\sum_{t=1}^n X_{i,j}^{t-1}, \pi_{i,j}]$ such that

$$
R_{i,j}^{(1)} = E\sum_{t=1}^n X_{i,j}^t(1-X_{i,j}^{t-1})\left(\frac{(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j})^2}{(1-r_{i,j}^t)^3}\right)\mathcal{I} > 0.
$$

On the other hand, note that

$$
\begin{aligned}
\sum_{k=2}^{\infty}\frac{|n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}|^k}{(1-\pi_{i,j})^{k+1}}\mathcal{I} &\le \left(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}\right)^2\sum_{k=0}^{\infty}\frac{1}{(1-\pi_{i,j})^3 2^k} \\
&= \left(n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}\right)^2\frac{2}{(1-\pi_{i,j})^3}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
R_{i,j}^{(1)} &\le E\sum_{t=1}^n\Big(\sum_{k=2}^{\infty}\frac{|n^{-1}\sum_{t=1}^n X_{i,j}^{t-1} - \pi_{i,j}|^k}{(1-\pi_{i,j})^{k+1}}\Big)\mathcal{I} \\
&\le Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^n X_{ij}^{t-1}\right)\frac{2}{(1-\pi_{i,j})^3} \\
&= \frac{2}{(1-\pi_{i,j})^3}Var(X_{ij}^t)\left[1 + \frac{2}{n}\sum_{h=1}^{n-1}(n-h)\rho_{ij}(h)\right] \\
&= \frac{2}{(1-\pi_{i,j})^3}\cdot\frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij})^2}\left[1 + \frac{2}{n}\sum_{h=1}^{n-1}(n-h)(1-\alpha_{ij}-\beta_{ij})^h\right] \\
&= \frac{2}{(1-\pi_{i,j})^3}\cdot\frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij})^2}\left[1 + \frac{2(1-\alpha_{ij}-\beta_{ij})}{\alpha_{ij}+\beta_{ij}} + O(n^{-1})\right] \\
&= \frac{2}{(1-\pi_{i,j})^4\pi_{i,j}}\cdot\frac{2-\alpha_{ij}-\beta_{ij}}{\alpha_{ij}+\beta_{ij}} + O(n^{-1}).
\end{aligned}
$$

Again, since $0 < l \le \alpha_{i,j}, \beta_{i,j}, \alpha_{i,j}+\beta_{i,j} \le 1$ holds for all $(i,j) \in \mathcal{J}$, we conclude that there exists a constant R_l such that $R^{(1)}_{i,j} \le R_l$. Together with Lemma 17, we have

$$\begin{aligned}
E\widehat{\alpha}_{i,j} &= E\widehat{\alpha}_{i,j}\mathcal{I} + E\widehat{\alpha}_{i,j}(1-\mathcal{I}) \\
&= E\frac{1}{n}\sum_{t=1}^{n} X^t_{i,j}(1-X^{t-1}_{i,j})\left[\frac{1}{1-\pi_{i,j}} + \frac{(n^{-1}\sum_{t=1}^n X^{t-1}_{i,j} - \pi_{i,j})}{(1-\pi_{i,j})^2}\right]\mathcal{I} + \frac{R^{(1)}_{i,j}}{n} + E\widehat{\alpha}_{i,j}(1-\mathcal{I}) \\
&= \alpha_{i,j} + \frac{Cov(\sum_{t=1}^n Y^t_{i,j}, \sum_{t=1}^n X^t_{i,j})}{n^2(1-\pi_{i,j})^2} + \frac{R^{(1)}_{i,j}}{n} + O(n^{-2}) \\
&= \alpha_{i,j} - \frac{\alpha_{i,j}[2\alpha_{i,j}(1-\beta_{i,j}) + \alpha_{i,j} + \beta_{i,j} - 2\beta^2_{i,j}]}{n(\alpha_{i,j}+\beta_{i,j})\beta_{i,j}} + \frac{R^{(1)}_{i,j}}{n} + O(n^{-2}).
\end{aligned}$$

Similarly, write $R^{(2)}_{i,j} := E\sum_{t=1}^n X^t_{i,j}(1-X^{t-1}_{i,j})\Big(\sum_{k=2}^{\infty}\frac{(n^{-1}\sum_{t=1}^n X^{t-1}_{i,j} - \pi_{i,j})^k}{(-1)^k\pi^{k+1}_{i,j}}\Big)\mathcal{I}'$ where $\mathcal{I}' := I\{|n^{-1}\sum_{t=1}^n(1-X^t_{i,j})X^{t-1}_{i,j}| \le \pi_{i,j}/2\}$. We have,

$$\begin{aligned}
E\widehat{\beta}_{i,j} &= E\frac{n^{-1}\sum_{t=1}^n(1-X^t_{i,j})X^{t-1}_{i,j}}{n^{-1}\sum_{t=1}^n X^{t-1}_{i,j}} \\
&= E\frac{1}{n}\sum_{t=1}^{n}(1-X^t_{i,j})X^{t-1}_{i,j}\Big[\frac{1}{\pi_{i,j}} - \frac{(n^{-1}\sum_{t=1}^n X^{t-1}_{i,j} - \pi_{i,j})}{\pi^2_{i,j}} \\
&\quad + \sum_{k=2}^{\infty}\frac{(n^{-1}\sum_{t=1}^n X^{t-1}_{i,j} - \pi_{i,j})^k}{(-1)^k\pi^{k+1}_{i,j}}\Big]\mathcal{I}' + E\widehat{\beta}_{i,j}(1-\mathcal{I}') \\
&= \beta_{i,j} - \frac{Cov(\sum_{t=1}^n Y^t_{i,j}, \sum_{t=1}^n X^t_{i,j} - X^n_{i,j} + X^0_{i,j})}{n^2\pi^2_{i,j}} + \frac{R^{(2)}_{i,j}}{n} + O(n^{-2}) \\
&= \beta_{i,j} + \frac{\beta_{i,j}[2\alpha_{i,j}(1-\beta_{i,j}) + \alpha_{i,j} + \beta_{i,j} - 2\beta^2_{i,j}]}{n(\alpha_{i,j}+\beta_{i,j})\alpha_{i,j}} + \frac{R^{(2)}_{i,j}}{n} + O(n^{-2}).
\end{aligned}$$

Here in the second last step we have used the fact that $En^{-1}(X^0_{i,j} - X^n_{i,j})(n^{-1}\sum_{t=1}^n X^{t-1}_{i,j} - \pi_{i,j}) = O(n^{-2})$, and in the last step we have used the fact that

$$\begin{aligned}
&n^{-2}E\sum_{t=1}^{n} X^t_{i,j}(1-X^{t-1}_{i,j})(X^n_{i,j} - X^0_{i,j}) \\
=\ & n^{-2}E\left[\sum_{t=1}^{n} X^{t-1}_{i,j}X^t_{i,j}X^0_{i,j} - \sum_{t=1}^{n} X^{t-1}_{i,j}X^t_{i,j}X^n_{i,j}\right] + n^{-2}[E(X^n_{i,j})^2 - E(X^n_{i,j}X^0_{i,j})] \\
=\ & n^{-2}\Big[\sum_{t=1}^{n} P(X^t_{i,j}=1|X^{t-1}_{i,j}=1)P(X^{t-1}_{i,j}=1|X^0_{i,j}=1)P(X^0_{i,j}=1) \\
& -\sum_{t=1}^{n} P(X^n_{i,j}=1|X^t_{i,j}=1)P(X^t_{i,j}=1|X^{t-1}_{i,j}=1)P(X^{t-1}_{i,j}=1)\Big] + O(n^{-2}) \\
=\ & O(n^{-2})
\end{aligned}$$

On one hand, similar to $R_{i,j}^{(1)}$, we can show that when n is large enough, there exists a R_l such that $R_{i,j}^{(2)} \leq R_l$ for any $(i,j) \in \mathcal{J}$.

■

Lemma 18 implies that the bias of the MLEs is of order $O(n^{-1})$. The bound R_l here also implies that the $O(n^{-1})$ order of the bias holds uniformly for all $(i,j) \in \mathcal{J}$.

Lemma 19. *Let conditions (2.5), C1 and C2 hold. For any constant $B > 0$, there exists a large enough constant $C > 0$ such that*

$$P\left\{\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1\|_F \geq C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \leq 8p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right],$$
$$P\left\{\|\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_2 - \mathbf{L}_2\mathbf{L}_2\|_F \geq C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \leq 8p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right],$$
$$P\left\{\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_2 - \mathbf{L}_1\mathbf{L}_2\|_F \geq C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \leq 8p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right],$$
$$P\left\{\|\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_1 - \mathbf{L}_2\mathbf{L}_1\|_F \geq C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \leq 8p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right]. \tag{A.5}$$

Proof We only prove the first inequality in (A.5) here as the other three inequalities can be proved similarly. Denote

$$\widetilde{\mathbf{L}}_1 := \mathbf{L}_1 - \mathrm{diag}(\mathbf{L}_1) = \mathbf{D}_1^{-1/2}\left[\mathbf{W}_1 - \mathrm{diag}(\mathbf{W})_1\right]\mathbf{D}_1^{-1/2},$$

and for any $1 \leq i,j \leq p$ we denote the (i,j)th element of $\widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1$ as $\delta_{i,j}$. Correspondingly, for any $\ell = 1,\ldots,p$, we define $\widetilde{d}_{\ell,1} := d_{\ell,1} - \alpha_{\ell,\ell}$. We first evaluate the error introduced by removing the $\mathrm{diag}(\mathbf{L}_1)$ term. With some abuse of notation, let $\widetilde{\alpha}_{i,j} = \alpha_{i,j}$ for $1 \leq i \neq j \leq p$ and $\widetilde{\alpha}_{i,i} = 0$ for $i = 1,\ldots,p$. We have $\mathbf{W} - \mathrm{diag}(\mathbf{W}) = (\widetilde{\alpha}_{i,j})_{1\leq i,j\leq p}$. Therefore,

$$|\delta_{i,j}| = \left|\sum_{k=1}^{p}\frac{\widetilde{\alpha}_{i,k}\widetilde{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}} - \sum_{k=1}^{p}\frac{\alpha_{i,k}\alpha_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right| \leq \frac{\alpha_{i,i}\alpha_{i,j}}{d_{i,1}\sqrt{d_{i,1}d_{j,1}}} + \frac{\alpha_{i,j}\alpha_{j,j}}{d_{j,1}\sqrt{d_{i,1}d_{j,1}}} \leq \frac{2}{(p-1)^2l^2}.$$

Consequently, we have

$$\begin{aligned}
\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1\|_F^2 &= \|(\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1) + (\widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1)\|_F^2 \\
&\leq 2\left[\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1\|_F^2 + \|\widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1\|_F^2\right] \\
&= 2\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1\|_F^2 + 2\sum_{1\leq i,j\leq p}\delta_{i,j}^2 \\
&\leq 2\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1\|_F^2 + \frac{8p^2}{(p-1)^4l^4}.
\end{aligned} \tag{A.6}$$

Next, we derive the asymptotic bound for $\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1\|_F^2$.

For any $1 \le i, j \le p$, we denote the (i,j)th element of $\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1$ as $\Delta_{i,j}$. By definition we have,

$$\Delta_{i,j} = \sum_{\substack{1\le k\le p\\ k\ne i,j}} \left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{\widehat{d}_{k,1}\sqrt{\widehat{d}_{i,1}\widehat{d}_{j,1}}} - \frac{\alpha_{i,k}\alpha_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}} \right),$$

where $\widehat{d}_{\ell,1} = \sum_{k=1}^p \widehat{\alpha}_{\ell,k}$ and $d_{\ell,1} = \sum_{k=1}^p \alpha_{\ell,k}$ for $l = 1,\ldots,p$. Note that $\widehat{\alpha}_{i,1},\ldots,\widehat{\alpha}_{i,p}$ are independent. Denote $\sigma_{i,k}^2 := Var(\widehat{\alpha}_{i,k})$, and $\tau_i^2 := \sum_{k=1}^p \sigma_{i,k}^2$. Similar to the proofs of Lemma 17 we can show that, when n is large enough, their exists a constant $C_\sigma > (2l)^{-1}$ and $c_\sigma := l(1-l)$ such that $c_\sigma n^{-1} \le \sigma_{i,k}^2 \le C_\sigma n^{-1}$ for any $(i,j) \in \mathcal{J}$. Consequently, $\tau_i^2 \simeq O(n^{-1}p)$. On the other hand, from Lemma 18 we know that there exists a large enough constant $C_\alpha > 0$ such that $|E\widehat{\alpha}_{i,j} - \alpha_{i,j}| \le \frac{C_\alpha}{n}$ for all $(i,j) \in \mathcal{J}$, and consequently, $\frac{|E\widehat{d}_{\ell,1} - d_{\ell,1}|}{p} \le \frac{|E\widehat{d}_{\ell,1} - \widetilde{d}_{\ell,1}|}{p} + \frac{1}{p} < \frac{C_\alpha}{n} + \frac{1}{p}$ for any $l = 1,\ldots,p$. We next break our proofs into three steps:

Step 1. Concentration of $p^{-1}\widehat{d}_{\ell,1}$.

Note that $|\widehat{\alpha}_{\ell,j}| \le 1$. By Bernstein's inequality (Bennett, 1962; Lin and Bai, 2011) we have, for any constant $C_d > 0$:

$$\begin{aligned}
& P\left(\frac{|\widehat{d}_{\ell,1} - d_{\ell,1}|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}} + \frac{C_\alpha}{n} + \frac{1}{p} \right) \\
\le\ & P\left(\frac{|\widehat{d}_{\ell,1} - E(\widehat{d}_{\ell,1})|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}} + \frac{C_\alpha}{n} + \frac{1}{p} - \frac{|E(\widehat{d}_{\ell,1}) - d_{\ell,1}|}{p} \right) \\
\le\ & P\left(\frac{|\widehat{d}_{\ell,1} - E(\widehat{d}_{\ell,1})|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}} \right) \\
\le\ & 2\exp\left\{ -\frac{\sqrt{p}C_d^2 n^{-1}\log(pn)}{2(\sqrt{p}C_\sigma/n + aC_d\sqrt{\log(pn)/n})} \right\} \\
=\ & 2\exp\left\{ -\frac{\sqrt{p}C_d^2 n^{-1}\log(pn)}{2(\sqrt{p}C_\sigma/n + C_d e(6l^{-1} + C_\alpha)\sqrt{\log n/(C_3 n)}\sqrt{\log(pn)/n})} \right\}. \qquad (A.7)
\end{aligned}$$

When $\sqrt{p}C_\sigma/n > C_d e(6l^{-1} + C_\alpha)\sqrt{\log n/(C_3 n)}\sqrt{\log(pn)/n})$, for any constant $B > 0$, by choosing $C_d > 2\sqrt{(B+1)C_\sigma}$, (A.7) reduces to

$$\begin{aligned}
& P\left(\frac{|\widehat{d}_{\ell,1} - d_{\ell,1}|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}} + \frac{C_\alpha}{n} + \frac{1}{p} \right) \\
\le\ & 2\exp\left\{ -\frac{\sqrt{p}C_d^2 n^{-1}\log(pn)}{4\sqrt{p}C_\sigma/n} \right\} < 2(pn)^{-(B+1)}. \qquad (A.8)
\end{aligned}$$

When $\sqrt{p}C_\sigma/n \le C_d e(6l^{-1}+C_\alpha)\sqrt{\log n/(C_3 n)}\sqrt{\log(pn)/n}$, by choosing $C_d = 4Be(6l^{-1}+C_\alpha)/\sqrt{C_3}$, (A.7) reduces to

$$\begin{aligned} & P\left(\frac{|\widehat{d}_{\ell,1}-d_{\ell,1}|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}}+\frac{C_\alpha}{n}+\frac{1}{p}\right) \\ \le\ & 2\exp\left\{-\frac{\sqrt{p}C_d^2 n^{-1}\log(pn)}{4C_d e(6l^{-1}+C_\alpha)\sqrt{\log n/(C_3 n)}\sqrt{\log(pn)/n}}\right\} \\ \le\ & 2\exp\{-B\sqrt{p}\}\,. \end{aligned} \tag{A.9}$$

From (A.7), (A.8) and (A.9) we conclude that for any $B>0$, by choosing C_d to be large enough, we have,

$$\begin{aligned} & P\left(\max_{l=1,\ldots,p}\frac{|\widehat{d}_{\ell,1}-d_{\ell,1}|}{p} \ge C_d\sqrt{\frac{\log(pn)}{np}}+\frac{C_\alpha}{n}+\frac{1}{p}\right) \\ \le\ & 2p\left[(pn)^{-(1+B)}+\exp\{-B\sqrt{p}\}\right]. \end{aligned} \tag{A.10}$$

Step 2. Concentration of $\Delta_{i,j}$.

Using the fact that $\widehat{\alpha}_{k,k}=0$ for $k=1,\ldots,p$, we have,

$$\Delta_{i,j} = \sum_{k=1}^{p}\left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{\widehat{d}_{k,1}\sqrt{\widehat{d}_{i,1}\widehat{d}_{j,1}}}-\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right)+\sum_{\substack{1\le k\le p\\ k\ne i,j}}\left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}-\frac{\alpha_{i,k}\alpha_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right).$$

We next bound the two terms on the right hand side of the above inequality. For the first term, denote $e_k := (\widehat{d}_{k,1}-d_{k,1})/p$. From (A.10) we have there exists a large enough constant C_B such that

$$P\left\{\max_{k=1,\ldots,p}|e_k| \le C_B\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right)\right\} \ge 1-2p\left[(pn)^{-(1+B)}+\exp\{-B\sqrt{p}\}\right].$$

Denote the event $\left\{\max_{k=1,\ldots,p}|e_k| \le C_B\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right)\right\}$ as $\mathcal{E}_B$. Under $\mathcal{E}_B$, we have, when n and p are large enough, $\sqrt{p^{-1}d_{k,1}+e_k} = \sqrt{p^{-1}d_{k,1}}+e_k/(2\sqrt{p^{-1}d_{k,1}})+O(e_k^2)$, and hence there exists a large enough constant $C_{l,B}>0$ such that for any $1\le i,j\le p$,

$$\begin{aligned} & \left|\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{\widehat{d}_{k,1}\sqrt{\widehat{d}_{i,1}\widehat{d}_{j,1}}}-\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right| \\ \le\ & \frac{\left|p^{-1}d_{k,1}\sqrt{p^{-1}d_{i,1}p^{-1}d_{j,1}}-(p^{-1}d_{k,1}+e_k)\sqrt{(p^{-1}d_{i,1}+e_i)(p^{-1}d_{j,1}+e_j)}\right|}{p^2(p^{-1}d_{k,1}+e_k)\sqrt{(p^{-1}d_{i,1}+e_i)(p^{-1}d_{j,1}+e_j)}p^{-1}d_{k,1}\sqrt{p^{-1}d_{i,1}p^{-1}d_{j,1}}} \\ =\ & O(p^{-2}(|e_i|+|e_j|+|e_k|)) \\ \le\ & \frac{C_{l,B}}{p^2}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right). \end{aligned}$$

Consequently, we have, under $\mathcal{E}_B$,

$$\left|\sum_{k=1}^{p}\left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{\widehat{d}_{k,1}\sqrt{\widehat{d}_{i,1}\widehat{d}_{j,1}}}-\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right)\right| \leq \frac{C_{l,B}}{p}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right). \tag{A.11}$$

For the second term, note that for any $1 \leq i, j \leq p$ and $k \neq i, j$,

$$\begin{aligned} & |E\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}-\alpha_{i,k}\alpha_{k,j}| \\ = \; & |E(\widehat{\alpha}_{i,k}-\alpha_{i,k})(\widehat{\alpha}_{k,j}-\alpha_{k,j})+E(\widehat{\alpha}_{i,k}-\alpha_{i,k})\alpha_{k,j}+E\alpha_{i,k}(\widehat{\alpha}_{k,j}-\alpha_{k,j})| \\ \leq \; & |E(\widehat{\alpha}_{i,k}-\alpha_{i,k})(\widehat{\alpha}_{k,j}-\alpha_{k,j})|+\frac{2C_\alpha}{n}. \end{aligned} \tag{A.12}$$

When $i \neq j$, by Lemma 18 and the fact that $\widehat{\alpha}_{i,k}$ and $\widehat{\alpha}_{k,j}$ are independent (since $k \neq i, j$), we have $|E\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}-\alpha_{i,k}\alpha_{k,j}| \leq C_{l,1}n^{-1}$ for some large enough constant $C_{l,1} > 0$. Using the same arguments for obtaining (A.10), we have, there exists a large enough constant $D_{l,B} > 0$ such that when n and p are large enough,

$$\begin{aligned} & P\left(\max_{1\leq i\neq j\leq p}\left|\sum_{\substack{1\leq k\leq p\\ k\neq i,j}}\left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}-\frac{\alpha_{i,k}\alpha_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right)\right| \geq \frac{D_{l,B}}{p}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right)\right) \\ \leq \; & 2p\left[(pn)^{-(1+B)}+\exp\{-B\sqrt{p}\}\right]. \end{aligned} \tag{A.13}$$

Denote the event $\left\{\max_{1\leq i\neq j\leq p}\left|\sum_{\substack{1\leq k\leq p\\ k\neq i,j}}\left(\frac{\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}-\frac{\alpha_{i,k}\alpha_{k,j}}{d_{k,1}\sqrt{d_{i,1}d_{j,1}}}\right)\right| \leq \frac{D_{l,B}}{p}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right)\right\}$ as $\mathcal{A}_B$. From (A.11) and (A.13) we conclude that, when n and p are large enough,

$$\begin{aligned} & P\left\{\max_{1\leq i\neq j\leq p}|\Delta_{i,j}| > \frac{C_{l,B}+D_{l,B}}{p}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{1}{n}+\frac{1}{p}\right)\right\} \\ \leq \; & P(\mathcal{E}_B^c)+P(\mathcal{A}_B^c) \\ < \; & 4p\left[(pn)^{-(1+B)}+\exp\{-B\sqrt{p}\}\right]. \end{aligned} \tag{A.14}$$

When $i = j$, by applying Lemma 18 and (A.3) to (A.12), we have, there exists a large enough constant $C_{l,2} > 0$, such that

$$|E\widehat{\alpha}_{i,k}\widehat{\alpha}_{k,i}-\alpha_{i,k}\alpha_{k,i}| \leq C_{l,2}\left(\frac{\log(pn)}{n}+\frac{1}{n}+\frac{1}{p}\right).$$

Consequently, similar to (A.14), we have, there exists a large enough constant $C_{l,3} > 0$, such that

$$P\left\{\max_{1\leq i\leq p}|\Delta_{i,i}| > \frac{C_{l,3}}{p}\left(\sqrt{\frac{\log(pn)}{np}}+\frac{\log(pn)}{n}+\frac{1}{p}\right)\right\} < 4p\left[(pn)^{-(1+B)}+\exp\{-B\sqrt{p}\}\right]. \tag{A.15}$$

Step 3. Proof of the first inequality in (A.5).

Note that $\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1\|_F = \sqrt{\sum_{1\le i,j\le p}\Delta_{i,j}^2} \le p\max_{1\le i\ne j\le p}|\Delta_{i,j}| + \sqrt{p}\max_{1\le i\le p}|\Delta_{i,i}|$. From (A.6), (A.14), (A.15) and the fact that $\frac{1}{\sqrt{p}}\left(\sqrt{\frac{\log(pn)}{np}} + \frac{\log(pn)}{n} + \frac{1}{p}\right) = o\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)$ we immediately have that there exists a large enough constant $C > 0$ such that when n and p are large enough,

$$P\left\{\|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1\|_F \ge C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \le 8p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right].$$

This proves the first inequality in (A.5).

■

Lemma 20. *Let conditions (2.5), C1 and C2 hold. For any constant $B > 0$, there exists a large enough constant $C > 0$ such that*

$$P\left\{\|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_F \ge 4C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\} \le 16p\left[(pn)^{-(1+B)} + \exp\{-B\sqrt{p}\}\right]. \tag{A.16}$$

Proof

Note that from the triangle inequality we have

$$\begin{aligned}
& \|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_F \\
= & \|(\widehat{\mathbf{L}}_1 + \widehat{\mathbf{L}}_2)(\widehat{\mathbf{L}}_1 + \widehat{\mathbf{L}}_2) - (\mathbf{L}_1 + \mathbf{L}_2)(\mathbf{L}_1 + \mathbf{L}_2)\|_F \\
= & \|(\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1) + (\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_2 - \mathbf{L}_1\mathbf{L}_2) + (\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_1 - \mathbf{L}_2\mathbf{L}_1) + (\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_2 - \mathbf{L}_2\mathbf{L}_2)\|_F \\
\le & \|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_1 - \mathbf{L}_1\mathbf{L}_1\|_F + \|\widehat{\mathbf{L}}_1\widehat{\mathbf{L}}_2 - \mathbf{L}_1\mathbf{L}_2\|_F + \|\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_1 - \mathbf{L}_2\mathbf{L}_1\|_F + \|\widehat{\mathbf{L}}_2\widehat{\mathbf{L}}_2 - \mathbf{L}_2\mathbf{L}_2\|_F.
\end{aligned}$$

Together with Lemma 19 we immediately conclude that (A.16) hold.

■

Proof of Theorem 10

From Weyl's inequality and Lemma 20, we have,

$$\max_{i=1,\dots,p}|\lambda_i^2 - \widehat{\lambda}_i^2| \le \|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_2 \le \|\widehat{\mathbf{L}}\widehat{\mathbf{L}} - \mathbf{L}\mathbf{L}\|_F = O_p\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right).$$

(21) is a direct result of the Davis-Kahan theorem (Rohe et al., 2011; Yu et al., 2015) theorem and Lemma 20.

A.8 Proof of Theorem 11

Recall that $\mathbf{\Gamma}_q = \mathbf{Z}\mathbf{U}$ where $\mathbf{U}$ is defined as in the proof of Proposition 9. For any $1 \le i \ne j \le n$ such that $\mathbf{z}_i \ne \mathbf{z}_j$, we need to show that $\|\mathbf{z}_i\mathbf{U}\mathbf{O}_q - \mathbf{z}_j\mathbf{U}_q\mathbf{O}_q\|_2 = \|\mathbf{z}_i\mathbf{U} - \mathbf{z}_j\mathbf{U}\|_2$ is large enough, so that the perturbed version (i.e. the rows of $\widehat{\mathbf{\Gamma}}_q$) is not changing the clustering structure.

Denote the ith row of $\mathbf{\Gamma}_q\mathbf{O}_q$ and $\widehat{\mathbf{\Gamma}}_q$ as $\boldsymbol{\gamma}_i$ and $\widehat{\boldsymbol{\gamma}}_i$, respectively, for $i = 1, \ldots, p$. Notice that from the proof of Proposition 9, we have $\mathbf{U}\mathbf{U}^\top = \mathbf{N}^{-1}\mathbf{Q}\mathbf{Q}^\top\mathbf{N}^{-1} = \mathbf{N}^{-2} = \mathrm{diag}\{s_1^{-1}, \ldots, s_q^{-1}\}$. Consequently, for any $\mathbf{z}_i \neq \mathbf{z}_j$, we have:

$$\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 = \|\mathbf{z}_i\mathbf{U}\mathbf{O}_q - \mathbf{z}_j\mathbf{U}_q\mathbf{O}_q\|_2 = \|\mathbf{z}_i\mathbf{U} - \mathbf{z}_j\mathbf{U}\|_2 \geq \sqrt{\frac{2}{s_{\max}}}. \tag{A.17}$$

We first show that $\mathbf{z}_i \neq \mathbf{z}_j$ implies $\widehat{\mathbf{c}}_i \neq \widehat{\mathbf{c}}_j$. Notice that $\mathbf{\Gamma}_q\mathbf{O}_q \in \mathcal{M}_{p,q}$. Denote $\widehat{\mathbf{C}} = (\widehat{\mathbf{c}}_1, \cdots, \widehat{\mathbf{c}}_p)^\top$. By the definition of $\widehat{\mathbf{C}}$ we have

$$\|\mathbf{\Gamma}_q\mathbf{O}_q - \widehat{\mathbf{C}}\|_F^2 \leq \|\widehat{\mathbf{\Gamma}}_q - \widehat{\mathbf{C}}\|_F^2 + \|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q\mathbf{O}_q\|_F^2 \leq 2\|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q\mathbf{O}_q\|_F^2. \tag{A.18}$$

Suppose there exist $i, j \in \{1, \ldots, p\}$ such that $\mathbf{z}_i \neq \mathbf{z}_j$ but $\widehat{\mathbf{c}}_i = \widehat{\mathbf{c}}_j$. We have

$$\|\mathbf{\Gamma}_q\mathbf{O}_q - \widehat{\mathbf{C}}\|_F^2 \geq \|\mathbf{z}_i\mathbf{U}\mathbf{O}_q - \widehat{\mathbf{c}}_i\|_2^2 + \|\mathbf{z}_j\mathbf{U}\mathbf{O}_q - \widehat{\mathbf{c}}_j\|_2^2 \geq \|\mathbf{z}_i\mathbf{U}\mathbf{O}_q - \mathbf{z}_j\mathbf{U}\mathbf{O}_q\|_2^2. \tag{A.19}$$

Combining (A.17), (21), (A.18) and (A.19), we have:

$$\sqrt{\frac{2}{s_{\max}}} \leq \|\mathbf{\Gamma}_q\mathbf{O}_q - \widehat{\mathbf{C}}\|_F \leq \sqrt{2}\|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q\mathbf{O}_q\|_F \leq 4\sqrt{2}\lambda_q^{-2}C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right).$$

We have reach a contradictory with (22). Therefore we conclude that $\widehat{\mathbf{c}}_i \neq \widehat{\mathbf{c}}_j$.

Next we show that if $\mathbf{z}_i = \mathbf{z}_j$ we must have $\widehat{\mathbf{c}}_i = \widehat{\mathbf{c}}_j$. Assume that there exist $1 \leq i \neq j \leq p$ such that $\mathbf{z}_i = \mathbf{z}_j$ and $\widehat{\mathbf{c}}_i \neq \widehat{\mathbf{c}}_j$. Notice that from the previous conclusion (i.e., that different z_i implies different $\widehat{\mathbf{c}}_i$), since there are q distinct rows in $\mathbf{Z}$, there are correspondingly q different rows in $\widehat{\mathbf{C}}$. Consequently for any $\mathbf{z}_i = \mathbf{z}_j$, if $\widehat{\mathbf{c}}_i \neq \widehat{\mathbf{c}}_j$ there must exist a $k \neq i, j$ such that $\mathbf{z}_i = \mathbf{z}_j \neq \mathbf{z}_k$ and $\widehat{\mathbf{c}}_j = \widehat{\mathbf{c}}_k$. Let $\widehat{\mathbf{C}}^*$ be $\widehat{\mathbf{C}}$ with the jth row replaced by $\widehat{\mathbf{c}}_i$. We have

$$\begin{aligned}
& \|\widehat{\mathbf{\Gamma}}_q - \widehat{\mathbf{C}}^*\|_F^2 - \|\widehat{\mathbf{\Gamma}}_q - \widehat{\mathbf{C}}\|_F^2 \\
= \;& \|\widehat{\boldsymbol{\gamma}}_j - \widehat{\mathbf{c}}_i\|_2^2 - \|\widehat{\boldsymbol{\gamma}}_j - \widehat{\mathbf{c}}_k\|_2^2 \\
= \;& \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j + \boldsymbol{\gamma}_i - \widehat{\mathbf{c}}_i\|_2^2 - \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j + \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_k + \boldsymbol{\gamma}_k - \widehat{\mathbf{c}}_k\|_2^2 \\
\leq \;& \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j + \boldsymbol{\gamma}_i - \widehat{\mathbf{c}}_i\|_2^2 + \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j + \boldsymbol{\gamma}_k - \widehat{\mathbf{c}}_k\|_2^2 - \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_k\|_2^2 \\
\leq \;& 2\|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}_q\mathbf{O}_q\|_F^2 + \|\mathbf{\Gamma}_q\mathbf{O}_q - \widehat{\mathbf{C}}\|_F^2 - \frac{2}{s_{\max}} \\
\leq \;& 4\left\{4\lambda_q^{-2}C\left(\sqrt{\frac{\log(pn)}{np}} + \frac{1}{n} + \frac{1}{p}\right)\right\}^2 - \frac{2}{s_{\max}} \\
< \;& 0.
\end{aligned}$$

Again, we reach a contradiction and so we conclude that if $\mathbf{z}_i = \mathbf{z}_j$ we must have $\widehat{\mathbf{c}}_i = \widehat{\mathbf{c}}_j$.

A.9 Proof of Theorem 13

Note that from Theorem 11, we have the memberships can be recovered with probability tending to 1, i,e, $P(\widehat{\nu} \neq \nu) \to 0$. On the other hand, given $\widehat{\nu} = \nu$, we have, the log likelihood function of $(\theta_{k,\ell}, \eta_{k,\ell})$, $1 \leq k \leq \ell \leq q$, is

$$\begin{aligned} l(\{\theta_{k,\ell}, \eta_{k,\ell}\}; \nu) &= \sum_{(i,j)\in S_{k,l}} \sum_{t=1}^{n} \Big\{ X_{i,j}^t (1 - X_{i,j}^{t-1}) \log \theta_{k,\ell} + (1 - X_{i,j}^t)(1 - X_{i,j}^{t-1}) \log(1 - \theta_{k,\ell}) \\ &\quad + (1 - X_{i,j}^t) X_{i,j}^{t-1} \log \eta_{k,\ell} + X_{i,j}^t X_{i,j}^{t-1} \log(1 - \eta_{k,\ell}) \Big\}. \end{aligned}$$

Using the same arguments as in the proof of Proposition 7, we can conclude that when $\widehat{\nu} = \nu$, $\sqrt{n}\mathbf{N}_{J_1,J_2}^{\frac{1}{2}}(\widehat{\boldsymbol{\Psi}}_{\mathcal{K}_1,\mathcal{K}_2} - \boldsymbol{\Psi}_{\mathcal{K}_1,\mathcal{K}_2}) \to N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{\mathcal{K}_1,\mathcal{K}_2})$. Let $\mathbf{Y} \sim N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{\mathcal{K}_1,\mathcal{K}_2})$. For any $\mathcal{Y} \subset \mathcal{R}^{m_1+m_2}$, let $\boldsymbol{\Phi}(\mathcal{Y}) := P(\mathbf{Y} \in \mathcal{Y})$, we have:

$$\begin{aligned} &|P(\sqrt{n}\mathbf{N}_{\mathcal{K}_1,\mathcal{K}_2}^{\frac{1}{2}}(\widehat{\boldsymbol{\Psi}}_{\mathcal{K}_1,\mathcal{K}_2} - \boldsymbol{\Psi}_{\mathcal{K}_1,\mathcal{K}_2}) \in \mathcal{Y}) - \boldsymbol{\Phi}(\mathcal{Y})| \\ \leq\ & P(\widehat{\nu} \neq \nu) + |P(\sqrt{n}\mathbf{N}_{\mathcal{K}_1,\mathcal{K}_2}^{\frac{1}{2}}(\widehat{\boldsymbol{\Psi}}_{\mathcal{K}_1,\mathcal{K}_2} - \boldsymbol{\Psi}_{\mathcal{K}_1,\mathcal{K}_2}) \in \mathcal{Y} | \widehat{\nu} = \nu) - \boldsymbol{\Phi}(\mathcal{Y})| \\ =\ & o(1). \end{aligned}$$

This proves the theorem.

A.10 Proof of Theorem 14

Without loss of generality, we consider the case where $\tau \in [n_0, \tau_0]$, as the convergence rate for $\tau \in [\tau_0, n-n_0]$ can be similarly derived. The idea is to break the time interval $[n_0, \tau_0]$ into two consecutive parts: $[n_0, \tau_{n,p}]$ and $[\tau_{n,p}, \tau_0]$, where $\tau_{n,p} = \Big\lfloor \tau_0 - \kappa n \Delta_F^{-2} \Big[\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\Big]\Big\rfloor$ for some large enough $\kappa > 0$. Here $\lfloor \cdot \rfloor$ denotes the least integer function. We shall show that when $\tau \in [n - n_0, \tau_{n,p}]$, in which $\widehat{\nu}^{\tau+1,n}$ might be inconsistent in estimating $\nu^{\tau_0+1,n}$, we have $\sup_{\tau\in[n_0,\tau_{n,p}]}[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)] < 0$ in probability. Hence $\arg\max_{\tau\in[n_0,\tau_0]} \mathbb{M}_n(\tau) = \arg\max_{\tau\in[\tau_{n,p},\tau_0]} \mathbb{M}_n(\tau)$ holds in probability. On the other hand, when $\tau \in [\tau_{n,p}, \tau_0]$, we shall see that the membership maps can be consistently recovered, and hence the convergence rate can be obtained using classical probabilistic arguments. For simplicity, we consider the case where $\nu^{1,\tau_0} = \nu^{\tau_0+1,n} = \nu$ first, and modification of the proofs for the case where $\nu^{1,\tau_0} \neq \nu^{\tau_0+1,n}$ will be provided subsequently.

A.10.1 Change point estimation with $\nu^{1,\tau_0} = \nu^{\tau_0+1,n} = \nu$.

We first consider the case where the membership structures remain unchanged, while the connectivity matrices before/after the change point are different. Specifically, we assume that $\nu^{1,\tau_0} = \nu^{\tau_0+1,n} = \nu$ for some ν, and $(\theta_{1,k,\ell}, \eta_{1,k,\ell}) \neq (\theta_{2,k,\ell}, \eta_{2,k,\ell})$ for some $1 \leq k \leq l \leq q$. For brevity, we shall be using the notations $S_{k,l}$, s_k, $s_{\min}$ and $n_{k,\ell}$ defined as in Section 3, and introduce some new notations as follows:

Define

$$\theta_{2,k,\ell}^{\tau} = \frac{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}} + \frac{n-\tau_0}{n-\tau}\frac{\theta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}{\frac{\tau_0-\tau}{n-\tau}\frac{\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}} + \frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}, \quad \eta_{2,k,\ell}^{\tau} = \frac{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}} + \frac{n-\tau_0}{n-\tau}\frac{\theta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}} + \frac{n-\tau_0}{n-\tau}\frac{\theta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}.$$

Clearly when $\tau = \tau_0$ we have $\theta^{\tau_0}_{2,k,\ell} = \theta_{2,k,\ell}$ and $\eta^{\tau_0}_{2,k,\ell} = \eta_{2,k,\ell}$.

Correspondingly, we denote the MLEs as

$$\widehat{\theta}^{\tau}_{1,k,\ell} = \sum_{(i,j)\in\widehat{S}^{\tau}_{1,k,\ell}} \sum_{t=1}^{\tau} X^t_{i,j}(1 - X^{t-1}_{i,j}) \Big/ \sum_{(i,j)\in\widehat{S}^{\tau}_{1,k,\ell}} \sum_{t=1}^{\tau}(1 - X^{t-1}_{i,j}),$$

$$\widehat{\eta}^{\tau}_{1,k,\ell} = \sum_{(i,j)\in\widehat{S}^{\tau}_{1,k,\ell}} \sum_{t=1}^{\tau}(1 - X^t_{i,j})X^{t-1}_{i,j} \Big/ \sum_{(i,j)\in\widehat{S}^{\tau}_{1,k,\ell}} \sum_{t=1}^{\tau} X^{t-1}_{i,j},$$

$$\widehat{\theta}^{\tau}_{2,k,\ell} = \sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{n} X^t_{i,j}(1 - X^{t-1}_{i,j}) \Big/ \sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{n}(1 - X^{t-1}_{i,j}),$$

$$\widehat{\eta}^{\tau}_{2,k,\ell} = \sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{n}(1 - X^t_{i,j})X^{t-1}_{i,j} \Big/ \sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{n} X^{t-1}_{i,j},$$

where $\widehat{S}^{\tau}_{1,k,\ell}$ and $\widehat{S}^{\tau}_{2,k,\ell}$ are defined in a similar way to $\widehat{S}_{k,\ell}$ (cf. Section 3.2.3), based on the estimated memberships $\widehat{\nu}^{1,\tau}$ and $\widehat{\nu}^{\tau+1,n}$, respectively.

Denote

$$\begin{aligned}
\mathbb{M}_n(\tau) &:= l(\{\widehat{\theta}^{\tau}_{1,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell}\};\ \widehat{\nu}^{1,\tau}) + l(\{\widehat{\theta}^{\tau}_{2,k,\ell}, \widehat{\eta}^{\tau}_{2,k,\ell}\};\ \widehat{\nu}^{\tau+1,n}),\\
\mathbb{M}(\tau) &:= El(\{\theta_{1,k,\ell}, \eta_{1,k,\ell}\}; \nu^{1,\tau}) + El(\{\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}\}; \nu^{\tau+1,n}).
\end{aligned}$$

We first evaluate several terms in (i)-(v), and all these results will be combined to obtain the error bound in (vi). In particular, (vi) states that as a direct result of (v), we can focus on the small neighborhood of $[\tau_{n,p}, \tau_0]$ when searching for the estimator $\widehat{\tau}$. Further, the inequality (A.38) transforms the error bound for $\tau_0 - \widehat{\tau}$ into the error bounds of the terms that we derived in (i)-(iv).

(i) Evaluating $\mathbb{M}(\tau) - \mathbb{M}(\tau_0)$.

Note that $\tau_0 = \arg\max_{n_0 \le \tau \le n-n_0} \mathbb{M}(\tau)$, and for any $\tau \in [n_0, \tau_0]$,

$$\begin{aligned}
\mathbb{M}(\tau) - \mathbb{M}(\tau_0) &= El(\{\theta_{1,k,\ell}, \eta_{1,k,\ell}\}; \nu^{1,\tau}) + El(\{\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}\}; \nu^{\tau+1,n})\\
&\quad - El(\{\theta_{1,k,\ell}, \eta_{1,k,\ell}\}; \nu^{1,\tau_0}) - El(\{\theta_{2,k,\ell}, \eta_{2,k,\ell}\}; \nu^{\tau_0+1,n})\\
&= El(\{\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}\}; \nu^{\tau+1,\tau_0}) - El(\{\theta_{1,k,\ell}, \eta_{1,k,\ell}\}; \nu^{\tau+1,\tau_0})\\
&\quad + El(\{\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}\}; \nu^{\tau_0+1,n}) - El(\{\theta_{2,k,\ell}, \eta_{2,k,\ell}\}; \nu^{\tau_0+1,n}).
\end{aligned}$$

Recall that

$$\begin{aligned}
l(\{\theta_{k,\ell}, \eta_{k,\ell}\}; \nu) = \sum_{1\le k\le \ell\le q} \sum_{(i,j)\in S_{k,l}} \sum_{t=1}^{n} \Big\{ & X^t_{i,j}(1 - X^{t-1}_{i,j}) \log\theta_{k,\ell}\\
& + (1 - X^t_{i,j})(1 - X^{t-1}_{i,j}) \log(1 - \theta_{k,\ell}) + (1 - X^t_{i,j})X^{t-1}_{i,j} \log\eta_{k,\ell} + X^t_{i,j}X^{t-1}_{i,j} \log(1 - \eta_{k,\ell}) \Big\}.
\end{aligned}$$

By Taylor expansion and the fact that the partial derivative of the expected likelihood evaluated at the true values equals zero we have, there exist $\theta^*_{k,\ell} \in [\theta_{1,k,\ell}, \theta^{\tau}_{2,k,\ell}], \eta^*_{k,\ell} \in$

$[\eta_{1,k,\ell},\eta^{\tau}_{2,k,\ell}], 1\le k\le \ell\le q$, such that

$$\begin{aligned}
& El(\{\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell}\};\nu^{\tau+1,\tau_0}) - El(\{\theta_{1,k,\ell},\eta_{1,k,\ell}\};\nu^{\tau+1,\tau_0})\\
= & -\frac{1}{2}\sum_{1\le k\le\ell\le q} n_{k,\ell}(\tau_0-\tau)\Bigg\{\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\Big(\frac{\theta^{\tau}_{2,k,\ell}-\theta_{1,k,\ell}}{\theta^*_{k,\ell}}\Big)^2 + \frac{(1-\theta_{1,k,\ell})\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\Big(\frac{\theta^{\tau}_{2,k,\ell}-\theta_{1,k,\ell}}{1-\theta^*_{k,\ell}}\Big)^2\\
& +\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\Big(\frac{\eta^{\tau}_{2,k,\ell}-\eta_{1,k,\ell}}{\eta^*_{k,\ell}}\Big)^2 + \frac{(1-\eta_{1,k,\ell})\theta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\Big(\frac{\eta^{\tau}_{2,k,\ell}-\eta_{1,k,\ell}}{1-\eta^*_{k,\ell}}\Big)^2\Bigg\}\\
\le & -C_1(\tau_0-\tau)\sum_{1\le k\le\ell\le q} n_{k,\ell}[(\theta_{1,k,\ell}-\theta_{2,k,\ell})^2+(\eta_{1,k,\ell}-\eta_{2,k,\ell})^2]\\
\le & -C_1(\tau_0-\tau)\big[\|\mathbf{W}_{1,1}-\mathbf{W}_{2,1}\|_F^2+\|\mathbf{W}_{1,2}-\mathbf{W}_{2,2}\|_F^2\big],
\end{aligned}$$

for some constant $C_1>0$. Here in the first step we have used the fact that for any $(i,j)\in S_{k,\ell}$ and $t\le\tau_0$, $EX^t_{i,j}(1-X^{t-1}_{i,j}) = EX^{t-1}_{i,j}(1-X^t_{i,j}) = \frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}$, $E(1-X^t_{i,j})(1-X^{t-1}_{i,j}) = \frac{(1-\theta_{1,k,\ell})\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}$, and $EX^t_{i,j}X^{t-1}_{i,j} = \frac{(1-\eta_{1,k,\ell})\theta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}$. Similarly, there exist $\theta^{\dagger}_{k,\ell}\in[\theta_{2,k,\ell},\theta^{\tau}_{2,k,\ell}],\eta^{\dagger}_{k,\ell}\in[\eta_{2,k,\ell},\eta^{\tau}_{2,k,\ell}], 1\le k\le\ell\le q$, such that

$$\begin{aligned}
& El(\{\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell}\};\nu^{\tau_0+1,n}) - El(\{\theta_{2,k,\ell},\eta_{2,k,\ell}\};\nu^{\tau_0+1,n})\\
= & -\frac{1}{2}\sum_{1\le k\le\ell\le q} n_{k,\ell}(n-\tau_0)\Bigg\{\frac{\theta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}\Big(\frac{\theta^{\tau}_{2,k,\ell}-\theta_{2,k,\ell}}{\theta^{\dagger}_{k,\ell}}\Big)^2 + \frac{(1-\theta_{2,k,\ell})\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}\Big(\frac{\theta^{\tau}_{2,k,\ell}-\theta_{2,k,\ell}}{1-\theta^{\dagger}_{k,\ell}}\Big)^2\\
& +\frac{\theta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}\Big(\frac{\eta^{\tau}_{2,k,\ell}-\eta_{2,k,\ell}}{\eta^{\dagger}_{k,\ell}}\Big)^2 + \frac{(1-\eta_{2,k,\ell})\theta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}\Big(\frac{\eta^{\tau}_{2,k,\ell}-\eta_{2,k,\ell}}{1-\eta^{\dagger}_{k,\ell}}\Big)^2\Bigg\}\\
\le & -C_2'(n-\tau_0)\sum_{1\le k\le\ell\le q}\frac{n_{k,\ell}(\tau_0-\tau)^2}{(n-\tau)^2}[(\theta_{1,k,\ell}-\theta_{2,k,\ell})^2+(\eta_{1,k,\ell}-\eta_{2,k,\ell})^2]\\
\le & -\frac{C_2(\tau_0-\tau)^2}{n-\tau}\big[\|\mathbf{W}_{1,1}-\mathbf{W}_{2,1}\|_F^2+\|\mathbf{W}_{1,2}-\mathbf{W}_{2,2}\|_F^2\big],
\end{aligned}$$

for some constants $C_2',C_2>0$. Consequently, we conclude that there exists a constant $C_3>0$ such that for any $n_0\le\tau\le\tau_0$, we have

$$\mathbb{M}(\tau)-\mathbb{M}(\tau_0)\le -C_3(\tau_0-\tau)\big[\|\mathbf{W}_{1,1}-\mathbf{W}_{2,1}\|_F^2+\|\mathbf{W}_{1,2}-\mathbf{W}_{2,2}\|_F^2\big]. \tag{A.20}$$

(ii) Evaluating $\sup_{\tau\in[\tau_{n,p},\tau_0]}P(\widehat{\nu}(\tau)\ne\nu)$.
Let $\widehat{\nu}(\tau)$ be either $\widehat{\nu}^{1,\tau}$ or $\widehat{\nu}^{\tau+1,n}$. Note that the membership maps of the networks before/after τ remain to be ν. From Theorems 10 and 11, we have, under conditions C2-C4, for any constant $B>0$, there exists a large enough constant C_B such that

$$\sup_{\tau\in[\tau_{n,p},\tau_0]}P(\widehat{\nu}(\tau)\ne\nu)\le C_B(\tau_0-\tau_{n,p})p[(pn)^{-(B+1)}+\exp\{-B\sqrt{p}\}].$$

Note that by choosing B to be large enough, we have $p(\tau_0-\tau_{n,p})(pn)^{-(B+1)} = o\left(\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{n^2 s^2_{\min}}}\right)$.
On the other hand, the assumption that $\frac{\log(np)}{\sqrt{p}}\to 0$ in condition C4 implies $pn\sqrt{\frac{(\tau_0-\tau_{n,p})s^2_{\min}}{\log(np)}} =$

$o(\exp\{B\sqrt{p}\})$ for some large enough constant B. Consequently, we have $(\tau_0-\tau_{n,p})p\exp\{-B\sqrt{p}\} = o\left(\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{n^2 s_{\min}^2}}\right)$, and hence we conclude that $\sup_{\tau\in[\tau_{n,p},\tau_0]} P(\widehat{\nu}(\tau)\neq\nu) = o\left(\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{n^2 s_{\min}^2}}\right)$.

(iii) Evaluating $\sup_{\tau\in[\tau_{n,p},\tau_0]}[\mathbb{M}_n(\tau)-\mathbb{M}(\tau)]$ when $\widehat{\nu}(\tau)=\nu$.

From (ii) we have with probability greater than $1-o\left(\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{n^2 s_{\min}^2}}\right)$, $\widehat{\nu}(\tau)=\nu$ for all $\tau\in[\tau_{n,p},\tau_0]$. For simplicity, in this part we assume that $\widehat{S}^{\tau}_{1,k,\ell}=\widehat{S}^{\tau}_{2,k,\ell}=S_{k,l}$ (or equivalently $\widehat{\nu}^{1,\tau}=\widehat{\nu}^{\tau+1,n}=\nu$) holds for all $1\le k\le\ell\le q$ and $\tau_{n,p}\le\tau\le\tau_0$ without indicating that this holds in probability.

Denote

$$g_{1,i,j}(\theta,\eta;\tau)=\sum_{t=1}^{\tau}\Big\{X^t_{i,j}(1-X^{t-1}_{i,j})\log\theta$$
$$+(1-X^t_{i,j})(1-X^{t-1}_{i,j})\log(1-\theta)+(1-X^t_{i,j})X^{t-1}_{i,j}\log\eta+X^t_{i,j}X^{t-1}_{i,j}\log(1-\eta)\Big\},$$

and

$$g_{2,i,j}(\theta,\eta;\tau)=\sum_{t=\tau+1}^{n}\Big\{X^t_{i,j}(1-X^{t-1}_{i,j})\log\theta$$
$$+(1-X^t_{i,j})(1-X^{t-1}_{i,j})\log(1-\theta)+(1-X^t_{i,j})X^{t-1}_{i,j}\log\eta+X^t_{i,j}X^{t-1}_{i,j}\log(1-\eta)\Big\}.$$

When $\widehat{\nu}=\nu$, we have,

$$\begin{aligned}
&\mathbb{M}_n(\tau)-\mathbb{M}(\tau)\\
=\;&\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell},\widehat{\eta}^{\tau}_{1,k,\ell};\tau)+\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell},\widehat{\eta}^{\tau}_{2,k,\ell};\tau)\\
&-E\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau)-E\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau)\\
=\;&\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell},\widehat{\eta}^{\tau}_{1,k,\ell};\tau)+\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell},\widehat{\eta}^{\tau}_{2,k,\ell};\tau)\\
&-\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau)-\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau)\\
&+\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau)+\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau)\\
&-E\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau)-E\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau)
\end{aligned}\tag{A.21}$$

Note that $\{\widehat{\theta}^{\tau}_{1,k,\ell},\widehat{\eta}^{\tau}_{1,k,\ell}\}$ is the maximizer of $\sum_{1\le k\le\ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{k,\ell},\eta_{k,\ell};\tau)$. Applying Taylor's expansion we have, there exist random scalars $\theta^-_{k,\ell}\in[\widehat{\theta}^{\tau}_{1,k,\ell},\theta_{1,k,\ell}],\eta^-_{k,\ell}\in$

$[\widehat{\eta}^{\tau}_{1,k,\ell}, \eta_{1,k,\ell}]$ such that

$$\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell};\tau) - \sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell};\tau)$$
$$\le \frac{1}{2}\sum_{1\le k\le \ell\le q} n_{k,\ell}\tau\Big\{\Big(\frac{\theta_{1,k,\ell}-\widehat{\theta}^{\tau}_{1,k,\ell}}{\theta^{-}_{k,\ell}}\Big)^2 + \Big(\frac{\theta_{1,k,\ell}-\widehat{\theta}^{\tau}_{1,k,\ell}}{1-\theta^{-}_{k,\ell}}\Big)^2 + \Big(\frac{\eta_{1,k,\ell}-\widehat{\eta}^{\tau}_{1,k,\ell}}{\eta^{-}_{k,\ell}}\Big)^2 + \Big(\frac{\eta_{1,k,\ell}-\widehat{\eta}^{\tau}_{1,k,\ell}}{1-\eta^{-}_{k,\ell}}\Big)^2\Big\}.$$

On the other hand, when $\widehat{\nu}=\nu$, similar to Proposition 6 and Theorem 12, we can show that for any $B>0$, there exists a large enough constant C^{-} such that $\max_{1\le k\le \ell\le q,\tau\in[\tau_{n,p},\tau_0]}|\widehat{\theta}^{\tau}_{1,k,\ell}-\theta_{1,k,\ell}|\le C^{-}\sqrt{\frac{\log(np)}{ns^2_{\min}}}$, and $\max_{1\le k\le \ell\le q,\tau\in[\tau_{n,p},\tau_0]}|\widehat{\eta}^{\tau}_{1,k,\ell}-\eta_{1,k,\ell}| = C^{-}\sqrt{\frac{\log(np)}{ns^2_{\min}}}$ hold with probability greater than $1-O((np)^{-B})$. Consequently, we have, when $\widehat{\nu}=\nu$, there exits a large enough constant $C_4>0$ such that

$$\begin{aligned}&\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell};\tau) - \sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell};\tau)\\ &\le C_4\tau\sum_{1\le k\le \ell\le q} n_{k,\ell}\frac{\log(np)}{ns^2_{\min}}\\ &\le \frac{C_4\tau p^2\log(np)}{ns^2_{\min}}. \end{aligned}\tag{A.22}$$

Similarly, we have there exists a large enough constant $C_5>0$ such that with probability greater than $1-O((np)^{-B})$,

$$\begin{aligned}&\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell}, \widehat{\eta}^{\tau}_{2,k,\ell};\tau) - \sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell};\tau)\\ &\le \frac{C_5(n-\tau)p^2\log(np)}{ns^2_{\min}}. \end{aligned}\tag{A.23}$$

On the other hand, similar to Lemma 16, there exists a constant $C_6>0$ such that with probability greater than $1-O((np)^{-B})$,

$$\begin{aligned}&\left|\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell};\tau) - E\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell};\tau)\right|\\ &\le C_6\tau p^2\sqrt{\frac{\log(np)}{\tau p^2}}, \end{aligned}\tag{A.24}$$

and

$$\begin{aligned}&\left|\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell};\tau) - E\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell};\tau)\right|\\ &\le C_6(n-\tau)p^2\sqrt{\frac{\log(np)}{(n-\tau)p^2}}. \end{aligned}\tag{A.25}$$

Combining (A.21), (A.22), (A.23), (A.24) and (A.25) we conclude that when $\widehat{\nu} = \nu$, there exists a large enough constant $C_0 > 0$ such that with probability greater than $1-O((np)^{-B})$,

$$\sup_{\tau\in[\tau_{n,p},\tau_0]} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau)| \leq C_0 np^2 \left\{ \frac{\log(np)}{ns_{\min}^2} + \sqrt{\frac{\log(np)}{np^2}} \right\} = O\left(np^2 \sqrt{\frac{\log(np)}{ns_{\min}^2}} \right). \tag{A.26}$$

(iv) Evaluating $E\sup_{\tau\in[\tau_{n,p},\tau_0]} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau) - \mathbb{M}_n(\tau_0) + \mathbb{M}(\tau_0)|$**.**
Notice that when $\widehat{\nu} = \nu$,

$$\begin{aligned}
& \mathbb{M}_n(\tau) - \mathbb{M}(\tau) - \mathbb{M}_n(\tau_0) + \mathbb{M}(\tau_0) \\
= \; & \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell}; \tau) + \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell}; \tau) \\
& -E \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell}; \tau) - E \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}; \tau) \\
& - \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\widehat{\theta}^{\tau_0}_{1,k,\ell}, \widehat{\eta}^{\tau_0}_{1,k,\ell}; \tau_0) - \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\widehat{\theta}^{\tau_0}_{2,k,\ell}, \widehat{\eta}^{\tau_0}_{2,k,\ell}; \tau_0) \\
& +E \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell}; \tau_0) + E \sum_{1\leq k\leq \ell\leq q} \sum_{(i,j)\in S_{k,\ell}} g_{2,i,j}(\theta_{2,k,\ell}, \eta_{2,k,\ell}; \tau_0)
\end{aligned}$$

Note that

$$\begin{aligned}
& g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell}, \widehat{\eta}^{\tau}_{1,k,\ell}; \tau) - g_{1,i,j}(\widehat{\theta}^{\tau_0}_{1,k,\ell}, \widehat{\eta}^{\tau_0}_{1,k,\ell}; \tau_0) - E[g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell}; \tau) - g_{1,i,j}(\theta_{1,k,\ell}, \eta_{1,k,\ell}; \tau_0)] \\
= \; & \sum_{t=1}^{\tau} \Bigg\{ X^t_{i,j}(1-X^{t-1}_{i,j}) \log \frac{\widehat{\theta}^{\tau}_{1,k,\ell}}{\widehat{\theta}^{\tau_0}_{1,k,\ell}} + (1-X^t_{i,j})(1-X^{t-1}_{i,j}) \log \frac{1-\widehat{\theta}^{\tau}_{1,k,\ell}}{1-\widehat{\theta}^{\tau_0}_{1,k,\ell}} \\
& +(1-X^t_{i,j})X^{t-1}_{i,j} \log \frac{\widehat{\eta}^{\tau}_{1,k,\ell}}{\widehat{\eta}^{\tau_0}_{1,k,\ell}} + X^t_{i,j}X^{t-1}_{i,j} \log \frac{1-\widehat{\eta}^{\tau}_{1,k,\ell}}{1-\widehat{\eta}^{\tau_0}_{1,k,\ell}} \Bigg\} - \sum_{t=\tau+1}^{\tau_0} \Bigg\{ X^t_{i,j}(1-X^{t-1}_{i,j}) \log \widehat{\theta}^{\tau_0}_{1,k,\ell} \\
& +(1-X^t_{i,j})(1-X^{t-1}_{i,j}) \log(1-\widehat{\theta}^{\tau_0}_{1,k,\ell}) + (1-X^t_{i,j})X^{t-1}_{i,j} \log \widehat{\eta}^{\tau_0}_{1,k,\ell} + X^t_{i,j}X^{t-1}_{i,j} \log(1-\widehat{\eta}^{\tau_0}_{1,k,\ell}) \Bigg\} \\
& +E \sum_{t=\tau+1}^{\tau_0} \Big\{ X^t_{i,j}(1-X^{t-1}_{i,j}) \log \theta_{1,k,\ell} + (1-X^t_{i,j})(1-X^{t-1}_{i,j}) \log(1-\theta_{1,k,\ell}) \\
& +(1-X^t_{i,j})X^{t-1}_{i,j} \log \eta_{1,k,\ell} + X^t_{i,j}X^{t-1}_{i,j} \log(1-\eta_{1,k,\ell}) \Big\}.
\end{aligned}$$

When sum over all $(i,j) \in S_{k,\ell}$ and $1 \leq k \leq \ell \leq q$, the last two terms in the above inequality can be bounded similar to (A.22) and (A.24), with τ replaced by $\tau_0 - \tau$. For the first term, with some calculations we have there exists a constant $c_1 > 0$ such that with probability larger than $1 - O(np)^{-B})$,

$$\begin{aligned}
\sup_{1\leq k\leq \ell\leq q} \left| \widehat{\theta}^{\tau}_{1,k,\ell} - \widehat{\theta}^{\tau_0}_{1,k,\ell} \right| &\leq c_1 \sqrt{\frac{\tau_0-\tau}{\tau_0}} \sqrt{\frac{\log(np)}{ns_{\min}^2}}, \\
\sup_{1\leq k\leq \ell\leq q} \left| \widehat{\eta}^{\tau}_{1,k,\ell} - \widehat{\eta}^{\tau_0}_{1,k,\ell} \right| &\leq c_1 \sqrt{\frac{\tau_0-\tau}{\tau_0}} \sqrt{\frac{\log(np)}{ns_{\min}^2}}.
\end{aligned} \tag{A.27}$$

Brief derivations of (A.27) are provided in Section A.10.3. Consequently, similar to (A.26), we have there exists a large enough constant $c_2 > 0$ such that

$$\begin{aligned}&\Bigg|\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}}\Big[g_{1,i,j}(\widehat{\theta}^{\tau}_{1,k,\ell},\widehat{\eta}^{\tau}_{1,k,\ell};\tau)-g_{1,i,j}(\widehat{\theta}^{\tau_0}_{1,k,\ell},\widehat{\eta}^{\tau_0}_{1,k,\ell};\tau_0)\Big]\\&\quad -E\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}}\Big[g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau)-g_{1,i,j}(\theta_{1,k,\ell},\eta_{1,k,\ell};\tau_0)\Big]\Bigg|\\ \le\ & c_2p^2\sqrt{\frac{(\tau_0-\tau)\log(np)}{s^2_{\min}}}. \qquad \text{(A.28)}\end{aligned}$$

Here in the last step we have used the fact that $\tau_0 \simeq O(n)$, $\sqrt{\frac{\log(np)}{p^2}} \le \sqrt{\frac{\log(np)}{s^2_{\min}}}$, and $\frac{(\tau_0-\tau)\log(np)}{ns^2_{\min}} = o\Big(\sqrt{\frac{(\tau_0-\tau)\log(np)}{s^2_{\min}}}\Big)$. Similarly, note that,

$$\begin{aligned}&g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell},\widehat{\eta}^{\tau}_{2,k,\ell};\tau)-g_{2,i,j}(\widehat{\theta}^{\tau_0}_{2,k,\ell},\widehat{\eta}^{\tau_0}_{2,k,\ell};\tau_0)-E[g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau)-g_{2,i,j}(\theta_{2,k,\ell},\eta_{2,k,\ell};\tau_0)] \qquad \text{(A.29)}\\ =\ &\sum_{t=\tau_0+1}^{n}\Bigg\{X^t_{i,j}(1-X^{t-1}_{i,j})\Bigg[\log\frac{\widehat{\theta}^{\tau}_{2,k,\ell}}{\widehat{\theta}^{\tau_0}_{2,k,\ell}}-\log\frac{\theta^{\tau}_{2,k,\ell}}{\theta_{2,k,\ell}}\Bigg]+(1-X^t_{i,j})(1-X^{t-1}_{i,j})\cdot\Bigg[\log\frac{1-\widehat{\theta}^{\tau}_{2,k,\ell}}{1-\widehat{\theta}^{\tau_0}_{2,k,\ell}}\\&-\log\frac{1-\theta^{\tau}_{2,k,\ell}}{1-\theta_{2,k,\ell}}\Bigg]+X^t_{i,j}(1-X^{t-1}_{i,j})\Bigg[\log\frac{\widehat{\eta}^{\tau}_{2,k,\ell}}{\widehat{\eta}^{\tau_0}_{2,k,\ell}}-\log\frac{\eta^{\tau}_{2,k,\ell}}{\eta_{2,k,\ell}}\Bigg]+X^t_{i,j}X^{t-1}_{i,j}\Bigg[\log\frac{1-\widehat{\eta}^{\tau}_{2,k,\ell}}{1-\widehat{\eta}^{\tau_0}_{2,k,\ell}}\\&-\log\frac{1-\eta^{\tau}_{2,k,\ell}}{1-\eta_{2,k,\ell}}\Bigg]\Bigg\}+[g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau_0)-g_{2,i,j}(\theta_{2,k,\ell},\eta_{2,k,\ell};\tau_0)]\\&-E[g_{2,i,j}(\theta^{\tau}_{2,k,\ell},\eta^{\tau}_{2,k,\ell};\tau_0)-g_{2,i,j}(\theta_{2,k,\ell},\eta_{2,k,\ell};\tau_0)]+\sum_{t=\tau+1}^{\tau_0}\Bigg\{X^t_{i,j}(1-X^{t-1}_{i,j})\log\widehat{\theta}^{\tau}_{2,k,\ell}\\&+(1-X^t_{i,j})(1-X^{t-1}_{i,j})\log(1-\widehat{\theta}^{\tau}_{2,k,\ell})+(1-X^t_{i,j})X^{t-1}_{i,j}\log\widehat{\eta}^{\tau}_{2,k,\ell}+X^t_{i,j}X^{t-1}_{i,j}\log(1-\widehat{\eta}^{\tau}_{2,k,\ell})\Bigg\}\\&-E\sum_{t=\tau+1}^{\tau_0}\Big\{X^t_{i,j}(1-X^{t-1}_{i,j})\log\theta^{\tau}_{2,k,\ell}+(1-X^t_{i,j})(1-X^{t-1}_{i,j})\log(1-\theta^{\tau}_{2,k,\ell})\\&+(1-X^t_{i,j})X^{t-1}_{i,j}\log\eta^{\tau}_{2,k,\ell}+X^t_{i,j}X^{t-1}_{i,j}\log(1-\eta^{\tau}_{2,k,\ell})\Big\}\\ :=\ & I+II-III+IV-V.\end{aligned}$$

For $II-III$, from Lemma 16 and the fact that $\Big|\theta^{\tau}_{2,k,\ell}-\theta_{2,k,\ell}\Big| \le \frac{c_3(\tau_0-\tau)}{n-\tau}$, and $\Big|\eta^{\tau}_{2,k,\ell}-\eta_{2,k,\ell}\Big| \le \frac{c_3(\tau_0-\tau)}{n-\tau}$ for some large enough constant c_3, we have there exists a large enough constant $c_4 > 0$ such that with probability greater than $1 - O((np)^{-B})$,

$$\Bigg|\sum_{1\le k\le \ell\le q}\sum_{(i,j)\in S_{k,\ell}}(II-III)\Bigg| \le c_4p^2\frac{\tau_0-\tau}{n-\tau}\sqrt{\frac{\log(np)}{\tau_0p^2}} = o\Bigg(p^2\sqrt{\frac{(\tau_0-\tau)\log(np)}{s^2_{\min}}}\Bigg). \qquad \text{(A.30)}$$

When sum over all $(i,j) \in S_{k,\ell}$ and $1 \le k \le \ell \le q$, the $IV - V$ term can be bounded similar to (A.22) and (A.24), with τ replaced by $\tau_0 - \tau$, i.e., there exist a constant $c_5 > 0$ such that with probability greater than $1 - O((np)^{-B})$,

$$\begin{aligned} \left| \sum_{1\le k\le \ell\le q} \sum_{(i,j)\in S_{k,\ell}} (IV - V) \right| &\le c_5 p^2 \left[\frac{(\tau_0 - \tau)\log(np)}{ns_{\min}^2} + \sqrt{\tau_0 - \tau}\sqrt{\frac{\log(np)}{p^2}} \right] \\ &= O\left(p^2 \sqrt{\frac{(\tau_0 - \tau)\log(np)}{s_{\min}^2}} \right). \end{aligned} \tag{A.31}$$

Lastly, similar to (A.27), we can show that there exists a constant $c_6 > 0$ such that with probability larger than $1 - O(np)^{-B})$,

$$\begin{aligned} \sup_{1\le k\le \ell\le q} \left| \log \frac{\widehat{\theta}^{\tau}_{2,k,\ell}}{\theta^{\tau}_{2,k,\ell}} - \log \frac{\widehat{\theta}^{\tau_0}_{2,k,\ell}}{\theta_{2,k,\ell}} \right| &\le c_6 \sqrt{\frac{\tau_0 - \tau}{n}} \sqrt{\frac{\log(np)}{ns_{\min}^2}}, \\ \sup_{1\le k\le \ell\le q} \left| \log \frac{\widehat{\eta}^{\tau}_{2,k,\ell}}{\eta^{\tau}_{2,k,\ell}} - \log \frac{\widehat{\eta}^{\tau_0}_{2,k,\ell}}{\eta_{2,k,\ell}} \right| &\le c_6 \sqrt{\frac{\tau_0 - \tau}{n}} \sqrt{\frac{\log(np)}{ns_{\min}^2}}. \end{aligned} \tag{A.32}$$

A brief proof of (A.32) is provided in Section A.10.3. Consequently, we can show that there exists a constant $c_7 > 0$ such that with probability larger than $1 - O(np)^{-B})$,

$$\begin{aligned} &\Bigg| \sum_{1\le k\le \ell\le q} \sum_{(i,j)\in S_{k,\ell}} \left[g_{2,i,j}(\widehat{\theta}^{\tau}_{2,k,\ell}, \widehat{\eta}^{\tau}_{2,k,\ell}; \tau) - g_{2,i,j}(\widehat{\theta}^{\tau_0}_{2,k,\ell}, \widehat{\eta}^{\tau_0}_{2,k,\ell}; \tau_0) \right] \\ &\quad - E \sum_{1\le k\le \ell\le q} \sum_{(i,j)\in S_{k,\ell}} \left[g_{2,i,j}(\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}; \tau) - g_{2,i,j}(\theta_{2,k,\ell}, \eta_{2,k,\ell}; \tau_0) \right] \Bigg| \\ \le\ & c_7 p^2 \sqrt{\frac{(\tau_0 - \tau)\log(np)}{s_{\min}^2}}. \end{aligned} \tag{A.33}$$

Now combining (A.28) and (A.33) and the probability for $\widehat{\nu} \ne \nu$ in (ii), we conclude that there exists a constant $C_0 > 0$ such that

$$\begin{aligned} & E \sup_{\tau\in[\tau_{n,p},\tau_0]} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau) - \mathbb{M}_n(\tau_0) + \mathbb{M}(\tau_0)| \\ \le\ & C_0 np^2 \left\{ \sqrt{\frac{(\tau_0 - \tau_{n,p})\log(np)}{n^2 s_{\min}^2}} + o\left(\sqrt{\frac{(\tau_0 - \tau_{n,p})\log(np)}{n^2 s_{\min}^2}} \right) \right\} \\ \le\ & 2C_0 p^2 \sqrt{\frac{(\tau_0 - \tau_{n,p})\log(np)}{s_{\min}^2}}. \end{aligned} \tag{A.34}$$

(v) Evaluating $\sup_{\tau\in[n_0,\tau_{n,p}]}[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)]$**.**
In this part we consider the case when $\tau \in [n - n_0, \tau_{n,p}]$. We shall see that $\sup_{\tau\in[n_0,\tau_{n,p}]}[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)] < 0$ in probability and hence $\arg\max_{\tau\in[n_0,\tau_0]} \mathbb{M}_n(\tau) = \arg\max_{\tau\in[\tau_{n,p},\tau_0]} \mathbb{M}_n(\tau)$ holds in probability. Note that for any $\tau \in [n - n_0, \tau_{n,p}]$,

$$\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0) = \mathbb{M}_n(\tau) - \mathbb{M}(\tau) - \mathbb{M}_n(\tau_0) + \mathbb{M}(\tau_0) - [\mathbb{M}(\tau_0) - \mathbb{M}(\tau)]. \tag{A.35}$$

Given $\widehat{\nu}^{1,\tau}$ and $\widehat{\nu}^{\tau+1,n}$, we define an intermediate term

$$\mathbb{M}_n^*(\tau) := l(\{\theta^-_{\tau,k,\ell}, \eta^-_{\tau,k,\ell}\};\ \widehat{\nu}^{1,\tau}) + l(\{\theta^*_{\tau,k,\ell}, \eta^*_{\tau,k,\ell}\};\ \widehat{\nu}^{\tau+1,n}).$$

where

$$\theta^-_{\tau,k,\ell} = \frac{\sum_{(i,j)\in\widehat{S}^\tau_{1,k,\ell}} \frac{\theta_{1,\nu(i),\nu(j)}\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}}}{\sum_{(i,j)\in\widehat{S}^\tau_{1,k,\ell}} \frac{\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}}}, \quad \eta^-_{\tau,k,\ell} = \frac{\sum_{(i,j)\in\widehat{S}^\tau_{1,k,\ell}} \frac{\theta_{1,\nu(i),\nu(j)}\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}}}{\sum_{(i,j)\in\widehat{S}^\tau_{1,k,\ell}} \frac{\theta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}}},$$

and

$$\theta^*_{\tau,k,\ell} = \frac{\sum_{(i,j)\in\widehat{S}^\tau_{2,k,\ell}} \left[\frac{(\tau_0-\tau)\theta_{1,\nu(i),\nu(j)}\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}} + \frac{(n-\tau_0)\theta_{2,\nu(i),\nu(j)}\eta_{2,\nu(i),\nu(j)}}{\theta_{2,\nu(i),\nu(j)}+\eta_{2,\nu(i),\nu(j)}}\right]}{\sum_{(i,j)\in\widehat{S}^\tau_{2,k,\ell}} \left[\frac{(\tau_0-\tau)\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}} + \frac{(n-\tau_0)\eta_{2,\nu(i),\nu(j)}}{\theta_{2,\nu(i),\nu(j)}+\eta_{2,\nu(i),\nu(j)}}\right]},$$

$$\eta^*_{\tau,k,\ell} = \frac{\sum_{(i,j)\in\widehat{S}^\tau_{2,k,\ell}} \left[\frac{(\tau_0-\tau)\theta_{1,\nu(i),\nu(j)}\eta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}} + \frac{(n-\tau_0)\theta_{2,\nu(i),\nu(j)}\eta_{2,\nu(i),\nu(j)}}{\theta_{2,\nu(i),\nu(j)}+\eta_{2,\nu(i),\nu(j)}}\right]}{\sum_{(i,j)\in\widehat{S}^\tau_{2,k,\ell}} \left[\frac{(\tau_0-\tau)\theta_{1,\nu(i),\nu(j)}}{\theta_{1,\nu(i),\nu(j)}+\eta_{1,\nu(i),\nu(j)}} + \frac{(n-\tau_0)\theta_{2,\nu(i),\nu(j)}}{\theta_{2,\nu(i),\nu(j)}+\eta_{2,\nu(i),\nu(j)}}\right]}.$$

We have

$$\mathbb{M}_n(\tau) - \mathbb{M}(\tau) = \mathbb{M}_n(\tau) - E\mathbb{M}_n^*(\tau) + E\mathbb{M}_n^*(\tau) - \mathbb{M}(\tau).$$

Note that the expected log-likelihood $E\sum_{1\le i\le j\le p} g_{1,i,j}(\alpha_{1,i,j}, \beta_{1,i,j}, \tau)$ is maximized at $\alpha_{1,i,j} = \theta_{1,\nu(i),\nu(j)}, \beta_{1,i,j} = \eta_{1,\nu(i),\nu(j)}$, and $E\sum_{1\le i\le j\le p} g_{2,i,j}(\alpha_{2,i,j}, \beta_{2,i,j}, \tau)$ is maximized at $\alpha_{2,i,j} = \theta_{\tau,\nu(i),\nu(j)}$, $\beta_{2,i,j} = \eta_{\tau,\nu(i),\nu(j)}$, we have

$$E\mathbb{M}_n^*(\tau) - \mathbb{M}(\tau) \le 0.$$

On the other hand, notice that given $\widehat{\nu}$, $\{\theta^-_{\tau,k,\ell}, \eta^-_{\tau,k,\ell}\}$ is the maximizer of $El(\{\theta_{k,\ell}, \eta_{k,\ell}\};\ \widehat{\nu}^{1,\tau})$ and $\{\theta^*_{\tau,k,\ell}, \eta^*_{\tau,k,\ell}\}$ is the maximizer of $El(\{\theta_{k,\ell}, \eta_{k,\ell}\};\ \widehat{\nu}^{\tau+1,n})$. Similar to (A.26), there exists a large enough constant $C_7 > 0$ such that with probability greater than $1 - O((np)^{-B})$,

$$\sup_{\tau\in[n_0,\tau_{n,p}]} |\mathbb{M}_n(\tau) - E\mathbb{M}_n^*(\tau)| \le C_7 np^2 \left\{\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right\}.$$

Consequently we have, with probability greater than $1 - O((np)^{-B})$,

$$\sup_{\tau\in[n_0,\tau_{n,p}]} \left[\mathbb{M}_n(\tau) - \mathbb{M}(\tau)\right] \le C_7 np^2 \left\{\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right\}. \tag{A.36}$$

We remark that since the membership structure $\widehat{\nu}^{\tau+1,n}$ can be very different from the original ν, the $s_{\min}$ in (A.26) is simply replaced by the lower bound 1, and hence the upper bound in (A.36) is independent of $\widehat{\nu}^{1;\tau}$ and $\widehat{\nu}^{\tau+1,n}$.

Combining (A.35), (A.36), (A.20), (A.26) (with $\tau = \tau_0$), and choosing $\kappa > 0$ to be large enough, we have with probability greater than $1 - O((np)^{-B})$,

$$\begin{aligned} & \sup_{\tau \in [n_0, \tau_{n,p}]} \big[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)\big] \\ \le\ & C_7 np^2 \left\{ \frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}} \right\} + C_0 np^2 \left\{ \frac{\log(np)}{ns_{\min}^2} + \sqrt{\frac{\log(np)}{np^2}} \right\} \\ & -C_3(\tau_0 - \tau_{n,p})\big[\|\mathbf{W}_{1,1} - \mathbf{W}_{2,1}\|_F^2 + \|\mathbf{W}_{1,2} - \mathbf{W}_{2,2}\|_F^2\big] \\ <\ & 0. \end{aligned}$$

Consequently we have,

$$P\left(\arg\max_{\tau \in [n_0, \tau_0]} \big[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)\big] = \arg\max_{\tau \in [\tau_{n,p}, \tau_0]} \big[\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0)\big] \right) \ge 1 - O((np)^{-B}). \tag{A.37}$$

(vi) Error bound for $\tau_0 - \widehat{\tau}$.

One of the key steps in the proof of (v) is to compare $\mathbb{M}_n(\tau)$, the estimated log-likelihood evaluated under the MLEs at a searching time point τ, with $\mathbb{M}(\tau)$, the maximized expected log-likelihood at time τ. The error between $\mathbb{M}_n(\tau)$ and $\mathbb{M}(\tau)$, which is of order $O_p\left(np^2\left(\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right)\right)$ reflects the noise level. On the other hand, the signal is captured by $\mathbb{M}(\tau_0) - \mathbb{M}(\tau) = O(|\tau_0 - \tau|p^2\Delta_F^2)$, i.e., the difference between the maximized expected log-likelihood evaluated at the true change point τ_0 and the maximized expected log-likelihood evaluated at the searching time point τ. Consequently, when $|\tau_0 - \tau|p^2\Delta_F^2 > \kappa\left[np^2\left(\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right)\right]$ for some large enough constant $\kappa > 0$, we are able to claim that $|\tau_0 - \widehat{\tau}| \le |\tau_0 - \tau| = O_p\left(n\Delta_F^{-2}\left[\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right]\right)$. By further deriving the estimation errors for any τ in the neighborhood of τ_0 with radius $O\left(\Delta_F^{-2}\left[\frac{\log(np)}{n} + \sqrt{\frac{\log(np)}{np^2}}\right]\right)$, we obtained a better bound based on Markov's inequality (see (A.38) below).

From (A.37) we have for any $0 < \epsilon \le \tau_0 - \tau_{n,p}$,

$$P(\tau_0 - \widehat{\tau} > \epsilon) \le P\left(\sup_{\tau \in [\tau_{n,p}, \tau_0 - \epsilon]} \mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0) \ge 0 \right) + O((np)^{-B}).$$

Note that from (i) and (iv) we have

$$\begin{aligned}
& P\Big(\sup_{\tau\in[\tau_{n,p},\tau_0-\epsilon]} \mathbb{M}_n(\tau)-\mathbb{M}_n(\tau_0)\geq 0\Big) \qquad \text{(A.38)}\\
\leq\ & P\Big(\sup_{\tau\in[\tau_{n,p},\tau_0-\epsilon]} \big[(\mathbb{M}_n(\tau)-\mathbb{M}(\tau)-\mathbb{M}_n(\tau_0)+\mathbb{M}(\tau_0))-(\mathbb{M}(\tau_0)-\mathbb{M}(\tau))\big]\geq 0\Big)\\
\leq\ & P\Big(\sup_{\tau\in[\tau_{n,p},\tau_0-\epsilon]} |\mathbb{M}_n(\tau)-\mathbb{M}(\tau)-\mathbb{M}_n(\tau_0)+\mathbb{M}(\tau_0)|\geq C_3\epsilon p^2\Delta_F^2\Big)\\
\leq\ & \frac{E\sup_{\tau\in[\tau_{n,p},\tau_0-\epsilon]}\big|\mathbb{M}_n(\tau)-\mathbb{M}(\tau)-\mathbb{M}_n(\tau_0)+\mathbb{M}(\tau_0)\big|}{C_3\epsilon p^2\Delta_F^2}\\
\leq\ & \frac{2C_0p^2\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{s_{\min}^2}}}{C_3\epsilon p^2\Delta_F^2}.
\end{aligned}$$

We thus conclude that $\tau_0-\widehat{\tau}=O_p\left(\Delta_F^{-2}\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{s_{\min}^2}}\right)$. By the definition of $\tau_{n,p}$ and condition C5 we have,

$$\Delta_F^{-2}\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{s_{\min}^2}} = O\Bigg(\frac{\tau_0-\tau_{n,p}}{\Delta_F}\sqrt{\frac{\log(np)}{ns_{\min}^2}}\left[\frac{\log(np)}{n}+\sqrt{\frac{\log(np)}{np^2}}\right]^{-1/2}\Bigg).$$

Consequently, we conclude that

$$\tau_0-\widehat{\tau}=O_p\left((\tau_0-\tau_{n,p})\min\left\{1,\frac{\min\left\{1,(n^{-1}p^2\log(np))^{\frac{1}{4}}\right\}}{\Delta_F s_{\min}}\right\}\right).$$

A.10.2 CHANGE POINT ESTIMATION WITH $\nu^{1,\tau_0}\neq\nu^{\tau_0+1,n}$.

We modify steps (i)-(v) to the case where $\nu^{1,\tau_0}\neq\nu^{\tau_0+1,n}$.

With some abuse of notations, we put $\mathbf{W}_{1,1}=(\alpha_{1,i,j})_{p\times p}$ with $\alpha_{1,i,j}=\theta_{1,\nu^{1,\tau_0}(i),\nu^{1,\tau_0}(j)}$, $\mathbf{W}_{1,2}=(1-\beta_{1,i,j})_{p\times p}$ with $\beta_{1,i,j}=\eta_{1,\nu^{1,\tau_0}(i),\nu^{1,\tau_0}(j)}$, $\mathbf{W}_{2,1}=(\alpha_{2,i,j})_{p\times p}$ with $\alpha_{2,i,j}=\theta_{2,\nu^{\tau_0+1,n}(i),\nu^{\tau_0+1,n}(j)}$, and $\mathbf{W}_{2,2}=(1-\beta_{2,i,j})_{p\times p}$ with $\beta_{2,i,j}=\eta_{2,\nu^{\tau_0+1,n}(i),\nu^{\tau_0+1,n}(j)}$. Similar to previous proofs we define

$$\begin{aligned}
\mathbb{M}_n(\tau) &:= \sum_{1\leq i\leq j\leq p} g_{1,i,j}(\widehat{\alpha}_{1,i,j}^{\tau},\widehat{\beta}_{1,i,j}^{\tau},\tau)+\sum_{1\leq i\leq j\leq p} g_{2,i,j}(\widehat{\alpha}_{2,i,j}^{\tau},\widehat{\beta}_{2,i,j}^{\tau},\tau),\\
\mathbb{M}(\tau) &:= E\sum_{1\leq i\leq j\leq p} g_{1,i,j}(\alpha_{1,i,j},\beta_{1,i,j},\tau)+E\sum_{1\leq i\leq j\leq p} g_{2,i,j}(\alpha_{2,i,j}^{\tau},\beta_{2,i,j}^{\tau},\tau),
\end{aligned}$$

where

$$\alpha_{2,i,j}^{\tau}=\frac{\frac{\tau_0-\tau}{n-\tau}\frac{\alpha_{1,i,j}\beta_{1,i,j}}{\alpha_{1,i,j}+\beta_{1,i,j}}+\frac{n-\tau_0}{n-\tau}\frac{\alpha_{2,i,j}\beta_{2,i,j}}{\alpha_{2,i,j}+\beta_{2,i,j}}}{\frac{\tau_0-\tau}{n-\tau}\frac{\beta_{1,i,j}}{\alpha_{1,i,j}+\beta_{1,i,j}}+\frac{n-\tau_0}{n-\tau}\frac{\beta_{2,i,j}}{\alpha_{2,i,j}+\beta_{2,i,j}}},$$

$$\beta_{2,i,j}^{\tau}=\frac{\frac{\tau_0-\tau}{n-\tau}\frac{\alpha_{1,i,j}\beta_{1,i,j}}{\alpha_{1,i,j}+\beta_{1,i,j}}+\frac{n-\tau_0}{n-\tau}\frac{\alpha_{2,i,j}\beta_{2,i,j}}{\alpha_{2,i,j}+\beta_{2,i,j}}}{\frac{\tau_0-\tau}{n-\tau}\frac{\alpha_{1,i,j}}{\alpha_{1,i,j}+\beta_{1,i,j}}+\frac{n-\tau_0}{n-\tau}\frac{\alpha_{2,i,j}}{\alpha_{2,i,j}+\beta_{2,i,j}}},$$

and

$$\widehat{\alpha}^{\tau}_{1,i,j} = \widehat{\theta}^{\tau}_{1,\widehat{\nu}^{1,\tau}(i),\widehat{\nu}^{1,\tau}(j)}, \quad \widehat{\beta}^{\tau}_{1,i,j} = \widehat{\eta}^{\tau}_{1,\widehat{\nu}^{1,\tau}(i),\widehat{\nu}^{1,\tau}(j)},$$
$$\widehat{\alpha}^{\tau}_{2,i,j} = \widehat{\theta}^{\tau}_{2,\widehat{\nu}^{\tau+1,n}(i),\widehat{\nu}^{\tau+1,n}(j)}, \quad \widehat{\beta}^{\tau}_{2,i,j} = \widehat{\eta}^{2}_{\tau,\widehat{\nu}^{\tau+1,n}(i),\widehat{\nu}^{\tau+1,n}(j)}.$$

Note that the definition of $M(\tau)$ here is now slightly different from the previous definition in that the $\alpha^{\tau}_{2,i,j}$ and $\beta^{\tau}_{2,i,j}$ will generally be different from $\theta^{\tau}_{2,\nu^{\tau_0+1,n}(i),\nu^{\tau_0+1,n}(j)}$ and $\eta^{\tau}_{2,\nu^{\tau_0+1,n}(i),\nu^{\tau_0+1,n}(j)}$, unless $\nu^{1,\tau_0} = \nu^{\tau_0+1,n}$. We first of all point out the main difference we are facing in the case where $\nu^{1,\tau_0} \neq \nu^{\tau_0+1,n}$. Consider a detection time $\tau \in [\tau_{n,p}, \tau_0]$. In the case where $\widehat{\nu}^{1,\tau} = \widehat{\nu}^{\tau+1,n} = \nu$, we have $\alpha^{\tau}_{2,i,j} = \theta^{\tau}_{2,k,\ell}$ for all $(i,j) \in S_{k,\ell}$, and we have $|\widehat{\theta}^{\tau}_{2,k,\ell} - \theta^{\tau}_{2,k,\ell}| = O_p\Big(\sqrt{\frac{\log(np)}{ns^2_{\min}}}\Big)$ for all $1 \le k \le \ell \le q$, or equivalently, $|\widehat{\alpha}^{\tau}_{2,i,j} - \theta^{\tau}_{2,\nu(i),\nu(j)}| = O_p\Big(\sqrt{\frac{\log(np)}{ns^2_{\min}}}\Big)$ for all $1 \le i \le j \le p$. However, when $\widehat{\nu}^{1,\tau} = \nu^{1,\tau_0}$ $\widehat{\nu}^{\tau+1,n} = \nu^{\tau_0+1,n}$ but $\nu^{1,\tau_0} \neq \nu^{\tau_0+1,n}$, the order of the estimation error becomes $O_p\Big(\sqrt{\frac{\log(np)}{ns^2_{\min}}} + \frac{\tau_0-\tau}{n}\Big)$. Here $\frac{\tau_0-\tau}{n}$ is a bias terms brought by the fact that $\widehat{\nu}^{1,\tau} \neq \widehat{\nu}^{\tau+1,n}$. The main issue is that the the following terms from the definition of $\widehat{\theta}^{\tau}_{2,k,\ell}$:

$$\sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{\tau_0} X^t_{i,j}(1 - X^{t-1}_{i,j}), \qquad \sum_{(i,j)\in\widehat{S}^{\tau}_{2,k,\ell}} \sum_{t=\tau+1}^{\tau_0} (1 - X^{t-1}_{i,j}),$$

are no longer unbiased estimators (subject to a normalization) of the following corresponding terms in the definition of $\theta^{\tau}_{2,k,\ell}$:

$$\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell} + \eta_{1,k,\ell}}, \quad \frac{\theta_{1,k,\ell},}{\theta_{1,k,\ell} + \eta_{1,k,\ell}}.$$

The proof of (i) does not involve any parameter estimators and hence can be established similarly.

For (ii), note that $|\widehat{\alpha}^{\tau}_{2,i,j} - \alpha_{2,i,j}| \le |\widehat{\alpha}^{\tau}_{2,i,j} - \alpha^{\tau}_{2,i,j}| + O\big(\frac{\tau_0-\tau}{n}\big)$ holds for all $1 \le i < j \le p$, where the $O\big(\frac{\tau_0-\tau}{n}\big)$ is independent of i, j. This implies that when estimating the $\alpha_{2,i,j}$, we have introduced a bias term $O\big(\frac{\tau_0-\tau}{n}\big)$ by including the $\tau_0 - \tau$ samples before the change point. From the proofs of Lemma 19, and condition C4, we conclude that (ii) hold for $\widehat{\nu}^{\tau+1,n}$.

For (iii), replacing the order of the error bound for $\widehat{\theta}^{+}_{\tau,k,\ell}$ and $\widehat{\theta}^{+}_{\tau,k,\ell}$ from $\sqrt{\frac{\log(np)}{ns^2_{\min}}}$ to $\sqrt{\frac{\log(np)}{ns^2_{\min}}} + \frac{\tau_0-\tau}{n}$, we have there exists a large enough constant $C_0 > 0$ such that

$$\begin{aligned}\sup_{\tau\in[\tau_{n,p},\tau_0]} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau)| &\le C_0 np^2 \left\{ \frac{\log(np)}{ns^2_{\min}} + \sqrt{\frac{\log(np)}{np^2}} + \frac{(\tau_0 - \tau_{n,p})^2}{n^2} \right\} \\ &= O\left(np^2 \left\{ \sqrt{\frac{\log(np)}{ns^2_{\min}}} + \frac{(\tau_0 - \tau_{n,p})^2}{n^2} \right\} \right).\end{aligned}$$

For (iv), the error bounds related to $g_{1,i,j}(\cdot,\cdot;\cdot)$ remain unchanged. Note that the decomposition (A.29) still holds with $\theta^{\tau}_{2,k,\ell}, \eta^{\tau}_{2,k,\ell}$ replaced be $\alpha^{\tau}_{2,i,j}, \beta^{\tau}_{2,i,j}$ and $\widehat{\theta}^{\tau}_{2,k,\ell}, \widehat{\eta}^{\tau}_{2,k,\ell}$

replaced be $\widehat{\alpha}^{\tau}_{2,i,j}, \widehat{\beta}^{\tau}_{2,i,j}$. The bound for (A.30) still holds owing to the fact that $|\alpha^{\tau}_{2,i,j} - \alpha_{2,i,j}| = O\Big(\frac{\tau_0-\tau}{n}\Big)$ and $|\beta^{\tau}_{2,i,j} - \beta_{2,i,j}| = O\Big(\frac{\tau_0-\tau}{n}\Big)$. The bound for (A.31) would become $O\left(p^2\sqrt{\frac{(\tau_0-\tau)\log(np)}{s^2_{\min}}} + \frac{(\tau_0-\tau_{n,p})^2}{n^2}\right)$. Notice that similar to (A.32), we have with probability larger than $1 - O((np)^{-B})$,

$$\sup_{1\le i\le j\le p}\left|\log\frac{\widehat{\alpha}^{\tau}_{2,i,j}}{\alpha^{\tau}_{2,i,j}} - \log\frac{\widehat{\alpha}^{\tau_0}_{2,k,\ell}}{\alpha_{2,k,\ell}}\right| = O\left(\sqrt{\frac{\tau_0-\tau}{n}}\sqrt{\frac{\log(np)}{ns^2_{\min}}} + \frac{\tau_0-\tau}{n}\right),$$

$$\sup_{1\le i\le j\le p}\left|\log\frac{\widehat{\beta}^{\tau}_{2,i,j}}{\beta^{\tau}_{2,i,j}} - \log\frac{\widehat{\beta}^{\tau_0}_{2,k,\ell}}{\beta 2,k,\ell}\right| = O\left(\sqrt{\frac{\tau_0-\tau}{n}}\sqrt{\frac{\log(np)}{ns^2_{\min}}} + \frac{\tau_0-\tau}{n}\right).$$

Consequently, we have

$$E\sup_{\tau\in[\tau_{n,p},\tau_0]}|\mathbb{M}_n(\tau) - \mathbb{M}(\tau) - \mathbb{M}_n(\tau_0) + \mathbb{M}(\tau_0)| \le C_0p^2\left\{\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{s^2_{\min}}} + (\tau_0-\tau_{n,p})\right\}.$$

By noticing that $\{\alpha_{1,i,j}, \beta_{1,i,j}, \alpha^{\tau}_{2,i,j}, \beta^{\tau}_{2,i,j}\}$ is the maximizer of $\mathbb{M}(\tau)$, we conclude that (v) also holds. Consequently, for (vi), we have

$$P\bigg(\sup_{\tau\in[\tau_{n,p},\tau_0-\epsilon]}\mathbb{M}_n(\tau) - \mathbb{M}_n(\tau_0) \ge 0\bigg) \le \frac{C_0p^2\sqrt{\frac{(\tau_0-\tau_{n,p})\log(np)}{s^2_{\min}}} + C_0p^2(\tau_0-\tau_{n,p})}{C_3\epsilon p^2\Delta^2_F}.$$

Consequently, we conclude that

$$\tau_0 - \widehat{\tau} = O_p\left((\tau_0-\tau_{n,p})\min\left\{1, \frac{\min\left\{1, (n^{-1}p^2\log(np))^{\frac{1}{4}}\right\}}{\Delta_F s_{\min}} + \frac{1}{\Delta^2_F}\right\}\right).$$

A.10.3 Proofs of (A.27) and (A.32) when $\widehat{\nu} = \nu$

For (A.27), note that

$$\left|\widehat{\theta}^{\tau}_{1,k,\ell} - \widehat{\theta}^{\tau_0}_{1,k,\ell}\right| = \left|\frac{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}X^t_{i,j}(1-X^{t-1}_{i,j})}{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}(1-X^{t-1}_{i,j})} - \frac{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau_0}X^t_{i,j}(1-X^{t-1}_{i,j})}{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau_0}(1-X^{t-1}_{i,j})}\right|. \tag{A.39}$$

Similar to Lemma 16, we can show that for any constant $B > 0$, there exists a large enough constant B_1 such that with probability larger than $1 - O((np)^{-(B+2)})$,

$$\left|\frac{1}{\tau n_{k,\ell}}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}(1-X^{t-1}_{i,j}) - \frac{\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\right| \le B_1\sqrt{\frac{\log(np)}{\tau n_{k,\ell}}},$$

$$\left|\frac{1}{\tau n_{k,\ell}}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}X^t_{i,j}(1-X^{t-1}_{i,j}) - \frac{\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\right| \le B_1\sqrt{\frac{\log(np)}{\tau n_{k,\ell}}},$$

and

$$\frac{1}{\tau(\tau_0-\tau)n_{k,\ell}^2}\Bigg|\Big[\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}X_{i,j}^t(1-X_{i,j}^{t-1})\Big]\Big[\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{\tau_0}(1-X_{i,j}^{t-1})\Big]$$
$$-\Big[\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{\tau_0}X_{i,j}^t(1-X_{i,j}^{t-1})\Big]\Big[\sum_{(i,j)\in S_{k,\ell}}\sum_{t=1}^{\tau}(1-X_{i,j}^{t-1})\Big]\Bigg|\le B_1\sqrt{\frac{\log(np)}{(\tau_0-\tau)n_{k,\ell}}}.$$

Plug these into (A.39) we have with probability larger than $1-O((np)^{-(B+2)})$,

$$\left|\widehat{\theta}_{1,k,\ell}^{\tau}-\widehat{\theta}_{1,k,\ell}^{\tau_0}\right|\le\frac{c_0\tau(\tau_0-\tau)n_{k,\ell}^2}{\tau_0\tau n_{k,\ell}^2}\sqrt{\frac{\log(np)}{(\tau_0-\tau)n_{k,\ell}}}\le\frac{c_0\sqrt{\tau_0-\tau}}{\tau_0}\sqrt{\frac{\log(np)}{n_{k,\ell}}},$$

for some constant $c_0>0$. Since $\tau_0\simeq O(n)$, and $n_{k,\ell}\ge s_{\min}^2$, we conclude that there exists a constant $c_1>0$ such that with probability larger than $1-O(np)^{-B})$,

$$\sup_{1\le k\le\ell\le q}\left|\widehat{\theta}_{1,k,\ell}^{\tau}-\widehat{\theta}_{1,k,\ell}^{\tau_0}\right|\le c_1\sqrt{\frac{\tau_0-\tau}{\tau_0}}\sqrt{\frac{\log(np)}{ns_{\min}^2}}.$$

For (A.32), note that

$$\begin{aligned}
&\log\frac{\widehat{\theta}_{2,k,\ell}^{\tau}}{\theta_{2,k,\ell}^{\tau}}-\log\frac{\widehat{\theta}_{2,k,\ell}^{\tau_0}}{\theta_{2,k,\ell}}\\
=\ &\log\frac{\frac{1}{n_{k,\ell}(n-\tau)}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{n}X_{i,j}^t(1-X_{i,j}^{t-1})}{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}+\frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}-\log\frac{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau_0+1}^{n}X_{i,j}^t(1-X_{i,j}^{t-1})}{n_{k,\ell}(n-\tau_0)\cdot\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}\\
&-\log\frac{\frac{1}{n_{k,\ell}(n-\tau)}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{n}(1-X_{i,j}^{t-1})}{\frac{\tau_0-\tau}{n-\tau}\frac{\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}+\frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}+\log\frac{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau_0+1}^{n}(1-X_{i,j}^{t-1})}{n_{k,\ell}(n-\tau_0)\cdot\frac{\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}.
\end{aligned}$$

It suffices to establish a bound for

$$\Delta_{\tau_0,\tau}:=\frac{\frac{1}{n_{k,\ell}(n-\tau)}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{n}X_{i,j}^t(1-X_{i,j}^{t-1})}{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}+\frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}-\frac{\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau_0+1}^{n}X_{i,j}^t(1-X_{i,j}^{t-1})}{n_{k,\ell}(n-\tau_0)\cdot\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}.$$

Note that for any $B>0$, there exists a large enough constant B_2 such that with probability greater than $1-O((np)^{-(B+2)})$,

$$\begin{aligned}
|\Delta_{\tau_0,\tau}|\ \le\ &\left|\frac{\frac{1}{n_{k,\ell}(n-\tau)}\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau+1}^{\tau_0}\left[X_{i,j}^t(1-X_{i,j}^{t-1})-\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}\right]}{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}+\frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}\right|\\
&+\Bigg|\left(\frac{\frac{1}{n_{k,\ell}(n-\tau)}}{\frac{\tau_0-\tau}{n-\tau}\frac{\theta_{1,k,\ell}\eta_{1,k,\ell}}{\theta_{1,k,\ell}+\eta_{1,k,\ell}}+\frac{n-\tau_0}{n-\tau}\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}-\frac{\frac{1}{n_{k,\ell}(n-\tau_0)}}{\frac{\eta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}}\right)\sum_{(i,j)\in S_{k,\ell}}\sum_{t=\tau_0+1}^{n}\Big[X_{i,j}^t(1-X_{i,j}^{t-1})\\
&-\frac{\theta_{2,k,\ell}\eta_{2,k,\ell}}{\theta_{2,k,\ell}+\eta_{2,k,\ell}}\Big]\Bigg|\\
\le\ &B_2\frac{\tau_0-\tau}{n-\tau}\sqrt{\frac{\log(np)}{(\tau_0-\tau)n_{k,\ell}}}+B_2\frac{\tau_0-\tau}{n-\tau}\sqrt{\frac{\log(np)}{(n-\tau_0)n_{k,\ell}}}.
\end{aligned}$$

(A.32) then follows by noticing that $\frac{\tau_0-\tau}{n-\tau}\sqrt{\frac{\log(np)}{(n-\tau_0)n_{k,\ell}}} = o\Big(\frac{\tau_0-\tau}{n-\tau}\sqrt{\frac{\log(np)}{(\tau_0-\tau)n_{k,\ell}}}\Big)$.

Appendix B: Real data analysis

B.1 French high school contact data (cont.)

The more details of the data analysis in Section 5.2 are presented below.

We now compare the dynamic stochastic block model method in Matias and Miele (2017) which is implemented in an `R` package `dynsbm`. We note that the model in Matias and Miele (2017) allows the membership probabilities and the transition parameters to vary over time.

We use the `dynsbm` package to analyze the same French high school contact data as we reported in the main text. The function `selection.dynsbm` can automatically select the number of clusters by maximizing the so-called integrated classification likelihood (ICL) criterion. Figure 6 shows that the optimal cluster number is selected to be 9 using ICL for this data set. This agrees with our findings reported in the main text when using the BIC selection criterion. We then compare the detected clusters from `dynsbm` with the actual class types in Table 9 at all the five time points.

We may notice that `dynsbm` method reserves one group as “0” for subjects with no edges (the absence nodes). Our algorithm, in comparison, is not affected by those subjects.

The grouping results from `dynsbm` method is quite stable over the five time points. Such results lend support to our method which assumes the constant cluster structure over time. Furthermore, our clustering results, shown in Table 10, appear to be more accurate and agree more closely to the true grouping (class types) for this data analysis.

Another practical advantage of our method is its relatively short computing time. Using a computer with Intel(R) Core(TM) i7-10875H CPU and 32.0 GB RAM, we need to spend 0.36 and 252.39 seconds to obtain the community detection results with our method and `dynsbm` method, respectively.

B.2 Enron email data

The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron’s collapse. The data file was published at `https://www.cs.cmu.edu/~./enron/`.

Rastelli et al. (2018) developed a latent stochastic block model for directed dynamic network data. They applied their methods for the Enron email data. To compare with their results, we used data for all the emails exchanged from January 2000 to March 2002 ($n = 27$) between the Enron members. The number of nodes in our analysis is 184 (which is different from Rastelli et al. (2018) where they have kept 148 subjects).

Using our spectral clustering algorithm for the whole network data and using the BIC, we obtain the best number of cluster is 13. Rastelli et al. (2018) used the exact ICL criterion and found 17 clusters but 4 groups seem to be extremely small or just contain inactive nodes. Their algorithm is for directed graph and also assume time-varying membership. It seems our results are still quite close to Rastelli et al. (2018). Similar to their analysis, we can see that there is a high degree of heterogeneity with 13 different clusters for this data set. When applying `dynsbm` package on this data the optimal number of cluster is only 6.

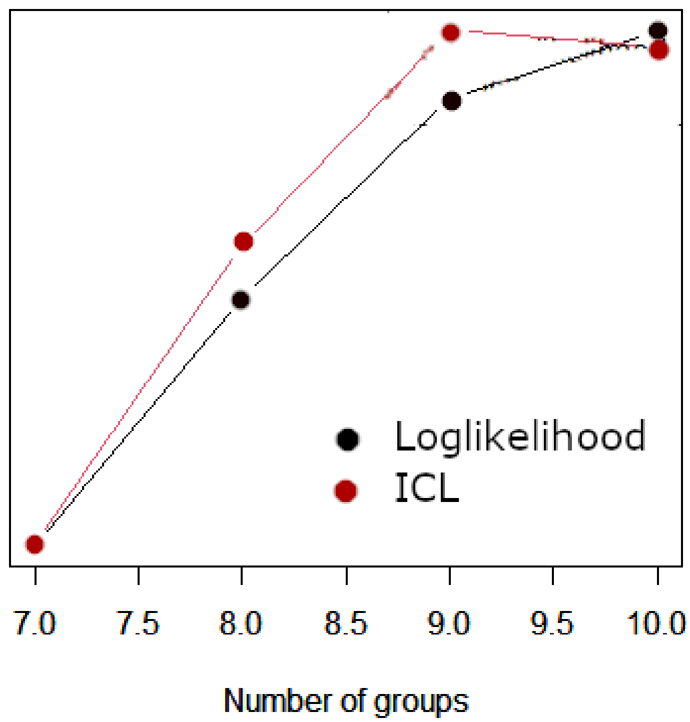

Figure 6: Integrated classification likelihood (ICL) and log-likelihood corresponding to different cluster numbers for French high school data.

Table 9: Clusters for French high school data by using the model in Matias and Miele (2017).

	Detected clusters									
Class types	0	1	2	3	4	5	6	7	8	9
$t = 1$										
BIO1	1	0	35	0	0	0	0	0	1	0
BIO2	1	0	0	0	0	1	0	0	31	0
BIO3	0	0	0	0	0	0	40	0	0	0
MP1	3	30	0	0	0	0	0	0	0	0
MP2	3	0	0	0	0	26	0	0	0	0
MP3	2	0	0	31	0	5	0	0	0	0
PC1	4	0	0	0	0	5	0	0	0	35
PC2	0	0	0	0	0	1	0	38	0	0
EGI	1	0	0	0	32	1	0	0	0	0
$t = 2$										
BIO1	0	0	36	0	0	0	0	0	1	0
BIO2	4	0	0	0	0	0	0	0	29	0
BIO3	1	0	0	0	0	0	39	0	0	0
MP1	0	31	0	0	0	2	0	0	0	0
MP2	2	0	0	0	0	27	0	0	0	0
MP3	1	0	0	34	0	3	0	0	0	0
PC1	3	0	0	0	0	2	0	0	0	39
PC2	3	0	0	0	0	1	0	35	0	0
EGI	3	0	0	0	31	0	0	0	0	0
$t = 3$										
BIO1	5	0	31	0	0	0	0	0	1	0
BIO2	2	0	0	0	0	0	0	0	31	0
BIO3	3	0	0	0	0	0	37	0	0	0
MP1	0	27	0	0	0	6	0	0	0	0
MP2	1	0	0	0	0	28	0	0	0	0
MP3	0	0	0	36	0	2	0	0	0	0
PC1	6	0	0	0	0	2	0	0	0	36
PC2	4	0	0	0	0	0	0	35	0	0
EGI	3	0	0	0	30	1	0	0	0	0
$t = 4$										
BIO1	2	0	35	0	0	0	0	0	0	0
BIO2	3	0	0	0	0	0	0	0	30	0
BIO3	7	0	0	0	0	0	33	0	0	0
MP1	1	28	0	0	0	4	0	0	0	0
MP2	1	1	0	0	0	27	0	0	0	0
MP3	4	0	0	34	0	0	0	0	0	0
PC1	4	0	0	0	0	3	0	0	0	37
PC2	4	0	0	0	0	0	0	35	0	0
EGI	6	0	0	0	26	2	0	0	0	0
$t = 5$										
BIO1	3	0	33	0	0	0	0	0	1	0
BIO2	2	0	0	0	0	0	0	0	31	0
BIO3	5	0	0	0	0	0	35	0	0	0
MP1	1	26	0	0	0	6	0	0	0	0
MP2	4	0	0	0	0	25	0	0	0	0
MP3	3	0	0	33	0	2	0	0	0	0
PC1	4	0	0	0	0	2	0	0	0	38
PC2	3	0	0	0	0	0	0	36	0	0
EGI	3	0	0	0	30	1	0	0	0	0

Table 10: Detected clusters for the French high school data by using our method.

	Detected clusters								
Class types	1	2	3	4	5	6	7	8	9
BIO1	0	0	1	0	0	0	0	0	36
BIO2	0	1	32	0	0	0	0	0	0
BIO3	0	1	0	0	39	0	0	0	0
MP1	33	0	0	0	0	0	0	0	0
MP2	0	1	0	28	0	0	0	0	0
MP3	0	0	0	0	0	38	0	0	0
PC1	0	44	0	0	0	0	0	0	0
PC2	0	0	0	0	0	0	0	39	0
EGI	0	0	0	0	0	0	34	0	0

Next we considered the change point analysis. Using our binary segmentation methods, we detect one change point at October 2001 which is exactly the month corresponding to the disclosure of Enron bankruptcy (see the log-likelihood functions plotted in Figure 7). This again agrees with the empirical findings in Rastelli et al. (2018).

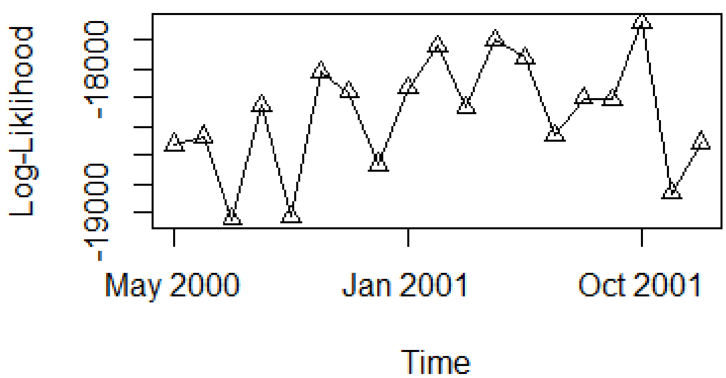

Figure 7: Log-likelihood functions corresponding to different change points in time for Enron email data.