

# Autoregressive Networks with Dependent Edges

Jinyuan Chang<sup>1,2</sup>, Qin Fang<sup>3</sup>, Eric D. Kolaczyk<sup>4</sup>, Peter W. MacDonald<sup>4</sup>, and Qiwei Yao<sup>5</sup>

<sup>1</sup>*Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics,  
Chengdu, Sichuan 611130, China*

<sup>2</sup>*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

<sup>3</sup>*Business School, University of Sydney, Sydney, 2008, Australia*

<sup>4</sup>*Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 0B9, Canada*

<sup>5</sup>*Department of Statistics, London School of Economics and Political Science, London, WC2A 2AE, U.K.*

## Abstract

We propose an autoregressive framework for modelling dynamic networks with dependent edges. It encompasses the models which accommodate, for example, transitivity, density-dependent and other stylized features often observed in real network data. By assuming the edges of network at each time are independent conditionally on their lagged values, the models, which exhibit a close connection with temporal ERGMs, facilitate both simulation and the maximum likelihood estimation in the straightforward manner. Due to the possible large number of parameters in the models, the initial MLEs may suffer from slow convergence rates. An improved estimator for each component parameter is proposed based on an iteration based on the projection which mitigates the impact of the other parameters (Chang et al., 2021, 2023). Based on a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the stationarity assumption. The limiting distribution is not normal in general, and it reduces to normal when the underlying process satisfies some mixing conditions. Illustration with a transitivity model was carried out in both simulation and a real network data set.

*Key words:* conditional independence, dynamic networks, maximum likelihood estimation, stylized features of network data, transitivity.

# 1 Introduction

Dynamic network modelling with dependent edges is practically important and relevant but technically challenging. Without dependence across different edges, it is impossible to incorporate into the models some stylized features often observed in real network data such as transitivity, density dependence. On the other hand, dependent edges make the dynamic structures of network processes complex and statistical inference challenging. Existing literature on modeling dynamic networks with dependent edges can be divided into two categories: latent process based models (Friel et al., 2016; Durante and Dunson, 2016; Matias and Miele, 2017), and temporal exponential family random-graph models (ERGMs) (Hanneke et al., 2010; Krivitsky and Handcock, 2014; Leifeld et al., 2018). Inference and simulation for the latent process based models rely on computationally intensive methods such as MCMC and EM algorithms, as their likelihood functions are not explicitly available. Furthermore the dependence depicted by latent process based models are implicit, and it cannot be configured easily to accommodate stylized features of real network data. On the other hand, it is well documented that ERGMs without a proper control on the level of dependence may suffer from lack of computational scalability, instable estimation algorithms, and concentration on extreme subspaces of graph space. See Schweinberger et al. (2020) and the references within. More recently Süveges and Olhede (2023) proposed a block logistic autoregressive network model with dependent edges, which was fitted using an EM algorithm.

In this paper, we propose a new autoregressive based Markov chain model with dependent edge. Following Jiang et al. (2023a,b), we specify the transition probabilities of forming a new edge or dissolving an existing edge between each pair of nodes explicitly depending on its history. Furthermore we allow those probabilities depending on the histories of other edge processes. This enlarged form of the transition probabilities make the model flexible enough to accommodate the stylized features such as transitivity and density dependence. The resulting network processes have dependent edges, which is radically different from those considered in Jiang et al. (2023a,b). Similar to Hanneke et al. (2010), Leifeld et al. (2018) and Süveges and Olhede (2023), we assume that the edges are conditionally independent given their joint histories. This makes both statistical inference and theoretical analysis more transparent. This conditional independence avoids the difficulties caused by the normalized constants in ERGMs; see Hanneke et al. (2010) and Leifeld et al. (2018).

Based on the conditional independence, we can build up a martingale difference structure which facilitates the asymptotic analysis for the maximum likelihood estimation (see Section 4 below). This, to our best knowledge, has never been done before in the context of dynamic networks with dependence edges.

The rest of the paper is organized as follows. Section 2 presents the general AR network framework with dependent edges. We also discuss the relationship between the proposed AR models and temporal ERGMs. Section 3 contains three concrete AR models which are designed to model, respectively, density-dependence, persistence, and transitivity — those are among the stylized features often observed in real network data. The two versions of the maximum likelihood estimation (MLE) for the parameters in the AR models and the associated asymptotic theory are presented in Section 4. We introduce the concepts of local parameters and global parameters, which need to be identified and estimated separately. They may also entertain different convergence rates. The initial MLE suffers from slower convergence rates due to the diverging number of local parameters. An improved MLE for each component parameter is obtained by projecting the score function onto the corresponding direction, which mitigates the impact of the other parameters (Chang et al., 2021, 2023). Based on a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the stationarity assumption. The limiting distribution is not normal in general. But it reduces to normal when the underlying process satisfies some mixing conditions which holds for many stationary processes. Section 5 presents an illustration via simulation for the proposed transitivity model. Further illustration with a real dynamic network data set is reported in Section 6. An online supplementary contains further more numerical results and all the technical proofs.

*Notation.* For any positive integer  $r$ , write  $[r] = \{1, \dots, r\}$  and  $\mathbb{R}_+^r = \{(x_1, \dots, x_r)^\top : x_i > 0 \text{ for any } i \in [r]\}$ . For any  $x, y \in \mathbb{R}$ , we write  $x \vee y = \max(x, y)$ . For two positive sequences  $\{a_n\}$  and  $\{b_n\}$ , we write  $a_n \ll b_n$  if  $\limsup_{n \rightarrow \infty} a_n/b_n = 0$ , and  $a_n \lesssim b_n$  if  $\limsup_{n \rightarrow \infty} a_n/b_n < \infty$ . For any vector  $\mathbf{b} = (b_1, \dots, b_r)^\top \in \mathbb{R}^r$ , we let  $\mathbf{b}_{-l}$  denote the sub-vector of  $\mathbf{b}$  by removing the  $l$ -th component  $b_l$ . Given an index set  $\mathcal{M} \subset [r]$ , we let  $\mathbf{b}_{\mathcal{M}}$  denote the sub-vector of  $\mathbf{b}$  that consists of the components of  $\mathbf{b}$  indexed by  $\mathcal{M}$ . For any  $r_1 \times r_2$  real matrix  $\mathbf{B}$ , denote by  $\mathbf{B}^\top$  its transpose. When  $r_1 = r_2$ , we use  $\lambda_{\min}(\mathbf{B})$  to denote the smallest eigenvalue of the square matrix  $\mathbf{B}$ . For any set  $\mathcal{U}$ ,  $|\mathcal{U}|$  denotes its cardinality.

## 2 AR( $m$ ) network framework

### 2.1 Model

Consider a dynamic network process defined on  $p$  nodes denoted by  $1, \dots, p$ . Let  $\mathbf{X}_t \equiv (X_{i,j}^t)_{p \times p}$  be the adjacency matrix at time  $t$ , where  $X_{i,j}^t = 1$  denotes the existence of an edge between nodes  $i$  and  $j$  at time  $t$ , and 0 otherwise. For simplicity, we only consider undirected networks without self-loops, i.e.  $X_{i,i}^t \equiv 0$  and  $X_{i,j}^t = X_{j,i}^t$ . The main idea can be applied to directed networks.

**Definition 1 (AR( $m$ ) networks)** *Conditionally on  $\{\mathbf{X}_s\}_{s \leq t-1}$ , the edges  $\{X_{i,j}^t\}_{1 \leq i < j \leq p}$  are mutually independent with*

$$\begin{aligned} \alpha_{i,j}^{t-1} &\equiv \mathbb{P}(X_{i,j}^t = 1 \mid X_{i,j}^{t-1} = 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-k} \text{ for } k \geq 2) \\ &= \mathbb{P}(X_{i,j}^t = 1 \mid X_{i,j}^{t-1} = 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}), \end{aligned} \quad (2.1)$$

$$\begin{aligned} \beta_{i,j}^{t-1} &\equiv \mathbb{P}(X_{i,j}^t = 0 \mid X_{i,j}^{t-1} = 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-k} \text{ for } k \geq 2) \\ &= \mathbb{P}(X_{i,j}^t = 0 \mid X_{i,j}^{t-1} = 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}), \end{aligned} \quad (2.2)$$

where  $m \geq 1$  is an integer.

An AR( $m$ ) network process defined above is a Markov chain with order  $m$ . Based on (2.1) and (2.2), we have

$$\mathbb{P}(X_{i,j}^t = 1 \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = \alpha_{i,j}^{t-1} + X_{i,j}^{t-1}(1 - \alpha_{i,j}^{t-1} - \beta_{i,j}^{t-1}) \equiv \gamma_{i,j}^{t-1}, \quad (2.3)$$

which implies that

$$X_{i,j}^t \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m} \sim \text{Bernoulli}(\gamma_{i,j}^{t-1}), \quad 1 \leq i < j \leq p.$$

Clearly edges  $X_{i,j}^t$ , for different  $(i, j)$ , are not independent with each other. We may impose various forms for the conditional probabilities  $\alpha_{i,j}^{t-1}$  and  $\beta_{i,j}^{t-1}$  to reflect different stylized features of network data. Put

$$\begin{aligned} \alpha_{i,j}^{t-1} &= f_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}_0), \\ \beta_{i,j}^{t-1} &= g_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}_0), \end{aligned} \quad (2.4)$$

where  $f_{i,j}$ 's and  $g_{i,j}$ 's are known functions, and  $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subset \mathbb{R}^q$  is a  $q$ -dimensional unknown true parameter vector. For any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , write

$$\begin{aligned}\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= f_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}), \\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= g_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}).\end{aligned}$$

Then  $\alpha_{i,j}^{t-1} = \alpha_{i,j}^{t-1}(\boldsymbol{\theta}_0)$  and  $\beta_{i,j}^{t-1} = \beta_{i,j}^{t-1}(\boldsymbol{\theta}_0)$ .

Modelling dynamic networks by Markov or/and AR models is not new. See, for example, [Snijders \(2005\)](#), [Ludkin et al. \(2018\)](#), [Yang et al. \(2011\)](#), [Yudovina et al. \(2015\)](#), and [Jiang et al. \(2023b\)](#). However, most available Markov models are designed for Erdős-Renyi networks with independent edges. Our setting provides a general framework to accommodate various dependence structures across different edges. Some practical network models satisfy this general framework are introduced in [Section 3](#).

For the special AR(1) processes (i.e.  $m = 1$ ), if both  $f_{i,j}$  and  $g_{i,j}$  in [\(2.4\)](#) are always positive and smaller than 1 for all  $1 \leq i < j \leq p$ ,  $\{\mathbf{X}_t\}_{t \geq 1}$  is an irreducible homogeneous Markov chain with  $2^{p(p-1)/2}$  states. Therefore when  $p$  is fixed, (i) there exists a unique stationary distribution, and (ii) if  $\mathbf{X}_0$  is activated according to this stationary distribution, the process  $\{\mathbf{X}_t\}_{t \geq 1}$  is strictly stationary and ergodic. See Theorems 3.1, 3.3 and 4.1 in Chapter 3 of [Brémaud \(1998\)](#). Hence the density-dependent model introduced in [Section 3.1](#) and the transitivity model introduced in [Section 3.3](#) are strictly stationary for any fixed constant  $p$  if all the transition probability functions  $\alpha_{i,j}^{t-1}$  and  $\beta_{i,j}^{t-1}$  are strictly between 0 and 1. It is worth pointing out that the ergodicity only holds for any fixed constant  $p$ . Hence we cannot take for granted that the sample means of  $\mathbf{X}_t$  and/or its summary statistics converge when  $p$  diverges together with the sample size, even when  $\mathbf{X}_t$  is stationary. Note that stationarity is not an asymptotic property while ergodicity is.

## 2.2 Relationship to temporal ERGMs

Similar to temporal ERGMs explored in [Hanneke et al. \(2010\)](#) and [Leifeld et al. \(2018\)](#), we assume that the edges are conditionally independent given their lagged values. But instead of specifying some exponential family distributions as the transition probabilities, we define, separately, the probability for forming a new edge in [\(2.1\)](#), and that for dissolving an existing edge in [\(2.2\)](#). Those two probability functions can be in any desirable forms as presented in [\(2.4\)](#). This allows us to

impose some closed-forms of parametric functions for  $\alpha_{i,j}^{t-1}$  and  $\beta_{i,j}^{t-1}$ , and those functions need only to be between 0 and 1. Hence, the likelihood functions are explicitly available, which allows us to depict more explicitly in our models some stylized features often observed in real network data. See Section 3 for details. The numerical analysis with both simulated and real data in Sections 5 and 6 indicates that the proposed AR models are capable to simulate and to reflect some observed interesting dynamic network phenomena.

The temporal ERGMs with conditional independent edges (Hanneke et al., 2010; Leifeld et al., 2018) can be expressed in our AR( $m$ ) framework as

$$\begin{aligned}\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}}{1 + \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}}, \\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\exp\{\boldsymbol{\psi}(\boldsymbol{\theta})^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}}{1 + \exp\{\boldsymbol{\psi}(\boldsymbol{\theta})^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}},\end{aligned}$$

where  $\mathbf{u}_{i,j}(\cdot)$  and  $\mathbf{v}_{i,j}(\cdot)$  have closed-form expressions based on the sufficient statistics of the original temporal ERGM. Furthermore, if  $\boldsymbol{\phi}(\boldsymbol{\theta})$  and  $\boldsymbol{\psi}(\boldsymbol{\theta})$  in the above expressions are replaced by, respectively,  $\boldsymbol{\phi}(\boldsymbol{\theta}_\alpha)$  and  $\boldsymbol{\psi}(\boldsymbol{\theta}_\beta)$ , where  $\boldsymbol{\theta}_\alpha$  and  $\boldsymbol{\theta}_\beta$  are two sets of different parameters,  $\mathbf{X}_t$  follows the separable temporal ERGM with conditional independent edges given the past networks (Krivitsky and Handcock, 2014). See Section A of the supplementary material for the detailed discussion on the relationship of the proposed AR models and temporal ERGMs.

### 3 Some interesting AR network models

To illustrate the usefulness of the AR( $m$ ) framework proposed above, we state three AR( $m$ ) network models which reflect various stylized features in real network data. In all three models, the parameters  $\{\xi_i\}_{i=1}^p$  and  $\{\eta_i\}_{i=1}^p$  reflect node heterogeneity in, respectively, forming a new edge and dissolving an existing edge. Specifically, the larger  $\xi_i$  is, the more likely node  $i$  will form new edges with other nodes, and the larger  $\eta_i$  is, the more likely the existing edges between node  $i$  and the others will be dissolved. Instances of these three models can be simulated using our development R package `arnetworks`, available at <https://github.com/peterwmacd/arnetworks>.

### 3.1 Density-dependent model

Let  $\vartheta_{i,j}^{t-1} = \exp\{a_0 D_{-i,-j}^{t-1} + a_1(D_i^{t-1} + D_j^{t-1})\}$  and  $\varpi_{i,j}^{t-1} = \exp\{b_0(1 - D_{-i,-j}^{t-1}) + b_1(2 - D_i^{t-1} - D_j^{t-1})\}$  with

$$D_{-i,-j}^{t-1} = \frac{1}{(p-2)(p-3)} \sum_{k,\ell: k,\ell \neq i,j, k \neq \ell} X_{k,\ell}^{t-1} \quad \text{and} \quad D_i^{t-1} = \frac{1}{p-1} \sum_{\ell: \ell \neq i} X_{i,\ell}^{t-1},$$

where  $D_i^{t-1}$  and  $D_j^{t-1}$  are, respectively, the densities of node  $i$  and node  $j$  at time  $t-1$ , and  $D_{-i,-j}^{t-1}$  is the network density excluding nodes  $i$  and  $j$  at time  $t-1$ . We specify the transition probabilities as follows:

$$\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\xi_i \xi_j \vartheta_{i,j}^{t-1}}{1 + \vartheta_{i,j}^{t-1} + \varpi_{i,j}^{t-1}}, \quad \beta_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\eta_i \eta_j \varpi_{i,j}^{t-1}}{1 + \vartheta_{i,j}^{t-1} + \varpi_{i,j}^{t-1}}. \quad (3.1)$$

This is an AR(1) model with parameter vector  $\boldsymbol{\theta} = (a_0, a_1, b_0, b_1, \xi_1, \dots, \xi_p, \eta_1, \dots, \eta_p)^\top \in \boldsymbol{\Theta} \subset \mathbb{R}_+^{2p+4}$ . Then the propensity to form a new edge between nodes  $i$  and  $j$  at time  $t$  is positively impacted by  $D_{-i,-j}^{t-1}$ ,  $D_i^{t-1}$  and  $D_j^{t-1}$ , and the propensity to dissolve an existing edge between nodes  $i$  and  $j$  at time  $t$  is negatively impacted by these three densities.

[Hanneke et al. \(2010\)](#) proposed an ERGM with network density in its index function. In (3.1) we explicitly specify the impact from the density functions on forming a new edge and dissolving an existing edge, while the model in Section 2.1 of [Hanneke et al. \(2010\)](#) does not differentiate the representations for these two types of impact. Within a separable ERGM framework, the edge counts model of [Krivitsky and Handcock \(2014\)](#) assumes that the collection of all newly formed edges is conditionally independent of the collection of all newly dissolved edges given their history, and the two conditional distributions are controlled by different parameters.

### 3.2 Persistence model

We define the transition probabilities

$$\begin{aligned} \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= \xi_i \xi_j \exp[-1 - a\{(1 - X_{i,j}^{t-2}) + (1 - X_{i,j}^{t-2})(1 - X_{i,j}^{t-3})\}], \\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= \eta_i \eta_j \exp\{-1 - b(X_{i,j}^{t-2} + X_{i,j}^{t-2} X_{i,j}^{t-3})\}. \end{aligned} \quad (3.2)$$

This is an AR(3) model with parameter vector  $\boldsymbol{\theta} = (a, b, \xi_1, \dots, \xi_p, \eta_1, \dots, \eta_p)^\top \in \boldsymbol{\Theta} \subset \mathbb{R}_+^{2p+2}$ . The probability to form a new edge between nodes  $i$  and  $j$  at time  $t$  is reduced if  $X_{i,j}^{t-2} = 0$ , and it is reduced further if, in addition,  $X_{i,j}^{t-3} = 0$ . The probability to dissolve an existing edge is reduced if

$X_{i,j}^{t-2} = 1$ , and it is reduced further if, in addition,  $X_{i,j}^{t-3} = 1$ . Hence if the edge status between two nodes is unchanged for 2 or 3 time periods, the probability for it remaining unchanged next time is larger than that otherwise.

Model (3.2) defines an AR(3) network process  $\mathbf{X}_t = (X_{i,j}^t)_{p \times p}$  with  $p(p-1)/2$  independent edge processes. Although the conclusion on the AR(1) stationarity in the last paragraph of Section 2.1 does not apply directly, this AR(3) network process is also strictly stationary, which is implied by the fact that  $\{X_{i,j}^t\}_{t \geq 1}$  is strictly stationary for each  $1 \leq i < j \leq p$ . Formally, for given  $(i, j)$  such that  $1 \leq i < j \leq p$ , let  $\mathbf{Y}_t = (X_{i,j}^t, X_{i,j}^{t-1}, X_{i,j}^{t-2})^\top$ . Then  $\{\mathbf{Y}_t\}_{t \geq 1}$  is a homogeneous Markov chain with  $2^3 = 8$  states. Let  $\mathbf{P}$  denote the transition probability matrix of  $\{\mathbf{Y}_t\}_{t \geq 1}$ . Then  $\mathbf{P}$  is a  $8 \times 8$  matrix with only 2 positive elements in each row and each column, provided that  $\xi_i \xi_j, \eta_i \eta_j \in (0, e)$ . It is straightforward to check that each row or column of  $\mathbf{P}^2$  has only 4 positive elements, and, more importantly, all the elements of  $\mathbf{P}^3$  is positive. Hence, the Markov chain  $\{\mathbf{Y}_t\}_{t \geq 1}$  is irreducible. By Theorems 3.1 and 3.3 in Chapter 3 of Brémaud (1998), the process  $\{\mathbf{Y}_t\}_{t \geq 1}$  is strictly stationary, and so is  $\{X_{i,j}^t\}_{t \geq 1}$ .

The persistent connectivity or non-connectivity is widely observed in, for example, brain networks, gene connections and social networks. The stability ERGM of Hanneke et al. (2010) does not differentiate between the propensity for retaining an existing edge and that for retaining a no-edge status.

### 3.3 Transitivity model

We propose an AR(1) model to reflect the feature of transitivity which refers to the phenomenon that nodes are more likely to link if they share links in common (i.e. ‘the friend of my friend is also my friend’). To this end, we specify the transition probabilities as follows:

$$\begin{aligned}\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\xi_i \xi_j \exp(aU_{i,j}^{t-1})}{1 + \exp(aU_{i,j}^{t-1}) + \exp(bV_{i,j}^{t-1})}, \\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\eta_i \eta_j \exp(bV_{i,j}^{t-1})}{1 + \exp(aU_{i,j}^{t-1}) + \exp(bV_{i,j}^{t-1})},\end{aligned}\tag{3.3}$$

where  $\boldsymbol{\theta} = (a, b, \xi_1, \dots, \xi_p, \eta_1, \dots, \eta_p)^\top \in \mathbb{R}_+^{2p+2}$ , and

$$\begin{aligned} U_{i,j}^{t-1} &= \frac{1}{p-2} \sum_{k: k \neq i,j} X_{i,k}^{t-1} X_{j,k}^{t-1}, \\ V_{i,j}^{t-1} &= \frac{1}{p-2} \sum_{k: k \neq i,j} \{X_{i,k}^{t-1}(1 - X_{j,k}^{t-1}) + (1 - X_{i,k}^{t-1})X_{j,k}^{t-1}\}. \end{aligned} \quad (3.4)$$

The pair  $(U_{i,j}^{t-1}, V_{i,j}^{t-1})$  characterizes the number of nodes with which both nodes  $i$  and  $j$  are connected, and the number of nodes with which only one of  $i$  and  $j$  is connected at time  $t-1$ . The larger  $U_{i,j}^{t-1}$  is (i.e. the more common friends  $i$  and  $j$  share at time  $t-1$ ), the more likely  $X_{i,j}^t = 1$ . The larger  $V_{i,j}^{t-1}$  is, the more likely  $X_{i,j}^t = 0$ . This reflects the transitivity of the networks. High levels of transitivity are found in various networks including friendship networks, industrial supply-chains, international trade flows, and alliances across firms and nations. Note that the quantity  $U_{i,j}^{t-1}$ , used in [Graham \(2016\)](#) to define the edge status of  $X_{i,j}^t$ , reflects the information based on which companies such as Facebook and LinkedIn have recommended new links to their customers. See also the transitivity ERGM of [Hanneke et al. \(2010\)](#).

We may use different parameters  $a$  and  $b$  in defining  $\alpha_{i,j}^{t-1}(\boldsymbol{\theta})$  and  $\beta_{i,j}^{t-1}(\boldsymbol{\theta})$  in (3.3). We do not pursue this more general form as (i) using different  $\xi_i$  and  $\eta_i$  reflects already the differences in the propensity between forming a new edge and dissolving an existing edge, and, perhaps more importantly, (ii) since most practical networks are sparse, the effective sample size for estimating the transition probability from the state of an existing edge is small. Therefore estimating the parameters only occurring in  $\beta_{i,j}^{t-1}(\boldsymbol{\theta})$  will be harder than those in  $\alpha_{i,j}^{t-1}(\boldsymbol{\theta})$ . Using the same  $a$  and  $b$  in both  $\alpha_{i,j}^{t-1}(\boldsymbol{\theta})$  and  $\beta_{i,j}^{t-1}(\boldsymbol{\theta})$  improves the estimation by pulling the information together. See also the relevant simulation results in Section C.2 in the online supplementary.

## 4 Estimation

### 4.1 General approach

The natural units of observation in our model are the  $X_{i,j}^t$ , indicating presence or absence of an edge between nodes  $i$  and  $j$  at time  $t$ . Intuitively, the extent to which these observations can contribute useful information to the estimation of a given element  $\theta_l$  of  $\boldsymbol{\theta} \in \Theta$  depends in turn on the extent to which that element plays a consistent role over time  $t$  in the corresponding probabilities

$$\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) = \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) + X_{i,j}^{t-1}\{1 - \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) - \beta_{i,j}^{t-1}(\boldsymbol{\theta})\}.$$

By (2.3), we have  $\gamma_{i,j}^{t-1} = \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)$ .

We formalize the above intuition as follows.

**Definition 2 (Global/local parameters)** Write  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$ , where  $q \geq 1$  is the total number of parameters. Let

$$\mathcal{G} = \{l \in [q] : \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \text{ involves } \theta_l \text{ for all } 1 \leq i < j \leq p \text{ and } t \in [n] \setminus [m]\}.$$

We call  $\boldsymbol{\theta}_{\mathcal{G}}$  and  $\boldsymbol{\theta}_{\mathcal{G}^c}$ , respectively, the global parameter vector and the local parameter vector.

In all three models presented in Section 3,  $\{\xi_i\}_{i=1}^p$  and  $\{\eta_i\}_{i=1}^p$  are local parameters, while all the other parameters in the models are global parameters. As we discuss below in Section 4.2, the global parameter vector  $\boldsymbol{\theta}_{0,\mathcal{G}}$  and the local parameter vector  $\boldsymbol{\theta}_{0,\mathcal{G}^c}$  need to be treated differently, which we accomplish via partial likelihoods. The resulting estimators may also entertain different convergence rates.

The asymptotic theory on the convergence rates and the limiting distributions of the proposed estimators will be developed here under the scenario where the sample size  $n \rightarrow \infty$  while the number of nodes  $p$  can be either fixed or diverge together with  $n$ . When  $p$  diverges with  $n$ , both the ergodicity and the central limit theorem for stationary Markov chains no longer apply even when  $\mathbf{X}_t$  is stationary (see the last paragraph in Section 2.1). Based on the conditional independence in our models, regardless of whether  $p$  is fixed or diverges with  $n$ , we can construct some martingale difference sequences, appropriate partial sums of which are amenable to the required asymptotic analysis without the stationarity assumption.

We develop the estimation theory for our models in three stages below. Sufficient conditions for identification of  $\boldsymbol{\theta}_0$  are established with respect to an expected partial log-likelihood  $\ell_{n,p}^{(l)}(\boldsymbol{\theta})$ , defined in (4.2). An initial estimator  $\tilde{\boldsymbol{\theta}}$  results from maximizing the corresponding partial log-likelihoods  $\tilde{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  defined in (4.6), for each  $l \in [q]$ . Finally, because of the potential high-dimensionality of our models (number of local parameters increasing with number of nodes), these estimators can suffer from slow rates of convergence. We offer estimators with improved rate of convergence, derived as a refinement of the initial estimator via the notion of projected score functions.

## 4.2 Identification of $\theta_0$

Let  $\mathcal{F}_t$  be the  $\sigma$ -field generated by  $\{\mathbf{X}_1, \dots, \mathbf{X}_t\}$ . For any  $l \in [q]$ , define

$$\mathcal{S}_l = \{(i, j) : 1 \leq i < j \leq p \text{ and } \gamma_{i,j}^{t-1}(\theta) \text{ involves } \theta_l \text{ for any } t \in [n] \setminus [m]\}. \quad (4.1)$$

If  $\theta_l$  is a global parameter,  $\mathcal{S}_l = \{(i, j) : 1 \leq i < j \leq p\}$ . For estimating  $\theta_l$  for  $l \in [q]$ , put

$$\ell_{n,p}^{(l)}(\theta) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \mathbb{E}_{\mathcal{F}_{t-1}} \{ \log [\{\gamma_{i,j}^{t-1}(\theta)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta)\}^{1-X_{i,j}^t}] \}, \quad (4.2)$$

where  $\mathbb{E}_{\mathcal{F}_{t-1}}(\cdot)$  denotes the conditional expectation given  $\mathcal{F}_{t-1}$  with the unknown true parameter vector  $\theta_0$ . For any  $t \in [n] \setminus [m]$  and  $1 \leq i < j \leq p$ , due to  $\log x \leq x - 1$  for any  $x > 0$ , we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_{t-1}} \{ \log [\{\gamma_{i,j}^{t-1}(\theta)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta)\}^{1-X_{i,j}^t}] \} \\ & \quad - \mathbb{E}_{\mathcal{F}_{t-1}} \{ \log [\{\gamma_{i,j}^{t-1}(\theta_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta_0)\}^{1-X_{i,j}^t}] \} \\ & \leq \mathbb{E}_{\mathcal{F}_{t-1}} \left[ \frac{\{\gamma_{i,j}^{t-1}(\theta)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta)\}^{1-X_{i,j}^t}}{\{\gamma_{i,j}^{t-1}(\theta_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta_0)\}^{1-X_{i,j}^t}} \right] - 1 \\ & = \frac{\gamma_{i,j}^{t-1}(\theta)}{\gamma_{i,j}^{t-1}(\theta_0)} \cdot \gamma_{i,j}^{t-1}(\theta_0) + \frac{1 - \gamma_{i,j}^{t-1}(\theta)}{1 - \gamma_{i,j}^{t-1}(\theta_0)} \cdot \{1 - \gamma_{i,j}^{t-1}(\theta_0)\} - 1 = 0, \end{aligned}$$

which implies  $\ell_{n,p}^{(l)}(\theta) \leq \ell_{n,p}^{(l)}(\theta_0)$  for any  $\theta \in \Theta$ . Notice that

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_{t-1}} \{ \log [\{\gamma_{i,j}^{t-1}(\theta)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta)\}^{1-X_{i,j}^t}] \} \\ & \quad - \mathbb{E}_{\mathcal{F}_{t-1}} \{ \log [\{\gamma_{i,j}^{t-1}(\theta_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta_0)\}^{1-X_{i,j}^t}] \} = 0 \end{aligned}$$

if and only if

$$\frac{\{\gamma_{i,j}^{t-1}(\theta)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta)\}^{1-X_{i,j}^t}}{\{\gamma_{i,j}^{t-1}(\theta_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\theta_0)\}^{1-X_{i,j}^t}} \equiv 1, \quad (4.3)$$

where (4.3) is equivalent to  $\gamma_{i,j}^{t-1}(\theta) = \gamma_{i,j}^{t-1}(\theta_0)$ . Hence, for any  $\theta \in \Theta \setminus \{\theta_0\}$ ,  $\ell_{n,p}^{(l)}(\theta) = \ell_{n,p}^{(l)}(\theta_0)$  if and only if  $\gamma_{i,j}^{t-1}(\theta) = \gamma_{i,j}^{t-1}(\theta_0)$  for any  $t \in [n] \setminus [m]$  and  $(i, j) \in \mathcal{S}_l$ . To guarantee the identification of  $\theta_0$ , we impose the following regularity conditions.

**Condition 1** (i) *There exists some universal constant  $C_1 > 0$  such that*

$$\min_{t \in [n] \setminus [m]} \min_{i,j: 1 \leq i < j \leq p} \inf_{\theta \in \Theta} \gamma_{i,j}^{t-1}(\theta) \{1 - \gamma_{i,j}^{t-1}(\theta)\} \geq C_1.$$

(ii) For any  $1 \leq i < j \leq p$  and  $t \in [n] \setminus [m]$ ,  $\gamma_{i,j}^{t-1}(\boldsymbol{\theta})$  is thrice continuously differentiable with respect to  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Furthermore, there exists some universal constant  $C_2 > 0$  such that

$$\max_{t \in [n] \setminus [m]} \max_{i,j: 1 \leq i < j \leq p} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial^k \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^k} \right|_{\infty} \leq C_2$$

for any  $k \in [3]$ .

Condition 1 specifies conditions for the parameter space  $\boldsymbol{\Theta}$ . Recall that  $\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) = \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) + X_{i,j}^{t-1} \{1 - \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) - \beta_{i,j}^{t-1}(\boldsymbol{\theta})\}$ . Due to  $X_{i,j}^{t-1} \in \{0, 1\}$ , Condition 1(i) holds if there exist two universal constants  $c_1, c_2 \in (0, 1)$  with  $c_1 < c_2$  such that

$$c_1 \leq \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) \leq c_2 \quad \text{and} \quad 1 - c_2 \leq \beta_{i,j}^{t-1}(\boldsymbol{\theta}) \leq 1 - c_1$$

for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,  $t \in [n] \setminus [m]$  and  $1 \leq i < j \leq p$ . Also, Condition 1(ii) holds provided that

$$\left| \frac{\partial^k \alpha_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^k} \right|_{\infty} \leq C_2 \quad \text{and} \quad \left| \frac{\partial^k \beta_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^k} \right|_{\infty} \leq C_2$$

for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,  $t \in [n] \setminus [m]$  and  $1 \leq i < j \leq p$ . Based on the explicit forms of  $\alpha_{i,j}^{t-1}(\boldsymbol{\theta})$  and  $\beta_{i,j}^{t-1}(\boldsymbol{\theta})$  in the specific models, we can identify the associated restrictions for the parameter space  $\boldsymbol{\Theta}$ .

For any  $1 \leq i < j \leq p$ , we define

$$\mathcal{I}_{i,j} = \{l \in [q] : \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \text{ involves } \theta_l \text{ for any } t \in [n] \setminus [m]\}.$$

**Condition 2** There exists a universal constant  $s \geq 1$  such that  $\max_{1 \leq i < j \leq p} |\mathcal{I}_{i,j}| \leq s$ .

Condition 2 requires that the dynamics of each edge process  $\{X_{i,j}^t\}_{t \geq 1}$  be driven by a finite number of parameters. Hence the number of global parameters is finite while the total number of local parameters may diverge together with  $p$ . For the density-dependent model introduced in Section 3.1, we have  $\boldsymbol{\theta}_{\mathcal{I}_{i,j}} = (a_0, a_1, b_0, b_1, \xi_i, \xi_j, \eta_i, \eta_j)^\top$  with  $s = 8$ . For both the persistence model and transitivity model introduced in Sections 3.2 and 3.3, we have  $\boldsymbol{\theta}_{\mathcal{I}_{i,j}} = (a, b, \xi_i, \xi_j, \eta_i, \eta_j)^\top$  with  $s = 6$ .

**Condition 3** There exists a universal constant  $C_3 > 0$  such that

$$\min_{i,j: 1 \leq i < j \leq p} \lambda_{\min} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_{\mathcal{I}_{i,j}}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_{\mathcal{I}_{i,j}}^\top} \right\} \geq C_3$$

with probability approaching one when  $n \rightarrow \infty$ .

**Proposition 1** *Let Conditions 1–3 hold, and  $C_* = 2(2C_1^{-2} + C_1^{-3})C_2^3 + 3(C_1^{-1} + C_1^{-2})C_2^2 + C_1^{-1}C_2$  with  $(C_1, C_2)$  specified in Condition 1. Assume  $\sup_{\boldsymbol{\theta} \in \Theta} |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty < 2C_3/(C_*s^3)$ . As  $n \rightarrow \infty$ , it holds with probability approaching one that*

$$\ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}) \geq \frac{\bar{C}}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\boldsymbol{\theta}_{\mathcal{I}_{i,j}} - \boldsymbol{\theta}_{0,\mathcal{I}_{i,j}}|_2^2$$

for any  $\boldsymbol{\theta} \in \Theta$  and  $l \in [q]$ , where  $\bar{C} > 0$  is a universal constant.

The proof of Proposition 1 is given in Section B.1 of the supplementary material. Notice that  $|\mathcal{S}_l \cap \mathcal{S}_{l'}| = |\mathcal{S}_l|$  for any  $l' \in \mathcal{G} \cup \{l\}$ . By Proposition 1, it holds with probability approaching one that for any  $\boldsymbol{\theta} \in \Theta$  and  $l \in [q]$ ,

$$\begin{aligned} \ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}) &\geq \frac{\bar{C}}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} \sum_{l' \in \mathcal{I}_{i,j}} |\theta_{l'} - \theta_{0,l'}|_2^2 = \frac{\bar{C}}{|\mathcal{S}_l|} \sum_{l'=1}^q \sum_{(i,j) \in \mathcal{S}_l \cap \mathcal{S}_{l'}} |\theta_{l'} - \theta_{0,l'}|^2 \\ &= \bar{C} \sum_{l' \in \mathcal{G} \cup \{l\}} |\theta_{l'} - \theta_{0,l'}|^2 + \bar{C} \sum_{l' \in \mathcal{G}^c \setminus \{l\}} \frac{|\mathcal{S}_l \cap \mathcal{S}_{l'}| |\theta_{l'} - \theta_{0,l'}|^2}{|\mathcal{S}_l|}. \end{aligned} \quad (4.4)$$

Hence, for any  $l \in [q]$ , the function  $\ell_{n,p}^{(l)}(\cdot)$  defined as (4.2) is a good candidate for identifying  $\theta_{0,l}$  and the global parameter vector  $\boldsymbol{\theta}_{0,\mathcal{G}}$  but is powerless in identifying  $\theta_{0,l'}$  with  $l' \in \mathcal{G}^c \setminus \{l\}$  if  $|\mathcal{S}_l \cap \mathcal{S}_{l'}| \ll |\mathcal{S}_l|$ .

### 4.3 Initial estimation for $\boldsymbol{\theta}_0$

With available observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , since  $\{\mathbf{X}_t\}_{t \geq 1}$  is a Markov chain with order  $m$ , the likelihood function for  $\boldsymbol{\theta}$ , conditionally on  $\mathbf{X}_1, \dots, \mathbf{X}_m$ , admits the form

$$\mathcal{L}_{n,p}(\mathbf{X}_n, \dots, \mathbf{X}_{m+1} | \mathbf{X}_m, \dots, \mathbf{X}_1; \boldsymbol{\theta}) = \prod_{t=m+1}^n L_{t,p}(\mathbf{X}_t | \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}),$$

where  $L_{t,p}(\mathbf{X}_t | \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta})$  is the transition probability of  $\mathbf{X}_t$  given  $\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}$ . By (2.3), the (normalized) log-likelihood admits the form

$$\begin{aligned} &\frac{2}{(n-m)p(p-1)} \log \mathcal{L}_{n,p}(\mathbf{X}_n, \dots, \mathbf{X}_{m+1} | \mathbf{X}_m, \dots, \mathbf{X}_1; \boldsymbol{\theta}) \\ &= \frac{2}{(n-m)p(p-1)} \sum_{t=m+1}^n \sum_{i,j: 1 \leq i < j \leq p} \log [\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}], \end{aligned} \quad (4.5)$$

which is the sample version of  $\ell_{n,p}^{(l)}(\boldsymbol{\theta})$  defined as (4.2) with  $l \in \mathcal{G}$ . As pointed out below (4.4), we should not estimate the local parameters based on this full log-likelihood. Therefore, for each

$l \in [q]$ , we define

$$\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \log [\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}], \quad (4.6)$$

which contains only the terms depending on  $\theta_l$  on the right-hand side of (4.5) (with a rescaled normalized constant).

For any  $l \in [q]$ , Lemma 1 in the supplementary material shows that  $\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  converges in probability to  $\ell_{n,p}^{(l)}(\boldsymbol{\theta})$  defined as (4.2) uniformly over  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Together with Proposition 1, we can estimate the global parameter vector  $\boldsymbol{\theta}_{0,\mathcal{G}}$  by maximizing the full log-likelihood  $\hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta})$  with some  $l' \in \mathcal{G}$ , and estimate the local parameter  $\theta_{0,l}$  with  $l \in \mathcal{G}^c$  by maximizing the corresponding (partial) log-likelihood  $\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$ . More specifically, letting  $(\hat{\theta}_{*,1}^{(l)}, \dots, \hat{\theta}_{*,q}^{(l)})^\top = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  for each  $l \in [q]$ , we define the initial estimator  $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_{\mathcal{G}}^\top, \tilde{\boldsymbol{\theta}}_{\mathcal{G}^c}^\top)^\top$  for  $\boldsymbol{\theta}_0$  as

$$\tilde{\boldsymbol{\theta}}_{\mathcal{G}} = (\hat{\theta}_{*,l}^{(l')})_{l \in \mathcal{G}} \quad \text{and} \quad \tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} = (\hat{\theta}_{*,l}^{(l)})_{l \in \mathcal{G}^c} \quad (4.7)$$

for some  $l' \in \mathcal{G}$ . Due to  $\mathcal{S}_l = \{(i, j) : 1 \leq i < j \leq p\}$  for any  $l \in \mathcal{G}$ , we know  $\hat{\ell}_{n,p}^{(l_1)}(\boldsymbol{\theta}) = \hat{\ell}_{n,p}^{(l_2)}(\boldsymbol{\theta})$  for any  $l_1, l_2 \in \mathcal{G}$ , which implies that the estimator  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  given in (4.7) does not depend on the selection of  $l' \in \mathcal{G}$ .

To investigate the theoretical properties of the estimator  $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_{\mathcal{G}}^\top, \tilde{\boldsymbol{\theta}}_{\mathcal{G}^c}^\top)^\top$ , we define

$$\begin{cases} c_{n,\mathcal{G}}^2 = \frac{q \log(np)}{\sqrt{np}} + \frac{q^{3/2} \log^{3/2}(np)}{\sqrt{np}^2}, \\ c_{n,\mathcal{G}^c}^2 = \frac{q \log(nS_{\mathcal{G}^c, \min})}{\sqrt{nS_{\mathcal{G}^c, \min}}} + \frac{q^{3/2} \log^{3/2}(nS_{\mathcal{G}^c, \min})}{\sqrt{nS_{\mathcal{G}^c, \min}}}, \end{cases} \quad (4.8)$$

where  $S_{\mathcal{G}^c, \min} = \min_{l \in \mathcal{G}^c} |\mathcal{S}_l|$ . Theorem 1 shows that the convergence rate of the initial estimator for the local parameters is slower than that of the global parameters if  $S_{\mathcal{G}^c, \min} \ll p^2$ . The proof of Theorem 1 is given in Section B.2 of the supplementary material.

**Theorem 1** *Let the conditions of Proposition 1 hold. Then  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2 = O_p(c_{n,\mathcal{G}})$  and  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_\infty = O_p(c_{n,\mathcal{G}^c})$ .*

**Remark 1** *By Theorem 1, the initial estimator  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  for the global parameters is consistent provided that*

$$q \ll \min \left\{ \frac{\sqrt{np}}{\log(np)}, \frac{n^{1/3} p^{4/3}}{\log(np)} \right\},$$

and the initial estimator  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c}$  for the local parameters is consistent provided that

$$q \ll \min \left\{ \frac{\sqrt{nS_{\mathcal{G}^c, \min}}}{\log(nS_{\mathcal{G}^c, \min})}, \frac{n^{1/3}S_{\mathcal{G}^c, \min}^{2/3}}{\log(nS_{\mathcal{G}^c, \min})} \right\}.$$

For the density-dependent model introduced in Section 3.1, we have  $q = 2p + 4$  and  $S_{\mathcal{G}^c, \min} = p - 1$ .

For both the persistence model and transitivity model introduced in Sections 3.2 and 3.3, we have

$q = 2p + 2$  and  $S_{\mathcal{G}^c, \min} = p - 1$ . Hence, for these three models, Theorem 1 gives the convergence

rates of  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  and  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c}$  as follows:

$$\begin{aligned} |\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0, \mathcal{G}}|_2 &= O_p \left\{ \frac{\log^{1/2}(np)}{n^{1/4}} \vee \frac{\log^{3/4}(np)}{(np)^{1/4}} \right\}, \\ |\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0, \mathcal{G}^c}|_\infty &= O_p \left\{ \frac{p^{1/4} \log^{1/2}(np)}{n^{1/4}} \vee \frac{p^{1/4} \log^{3/4}(np)}{n^{1/4}} \right\}, \end{aligned}$$

which implies the consistency of  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  provided that  $\log p \ll n^{1/2}$ , and the consistency of  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c}$  provided that  $p \ll n(\log n)^{-3}$ .

#### 4.4 Improved estimation for $\boldsymbol{\theta}_0$

Recall  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$ . The initial estimator  $\tilde{\boldsymbol{\theta}}$  specified in (4.7) suffers from slow convergence rates due to the high dimensionality of  $\boldsymbol{\theta}$ . In this section, we improve the estimation for each component  $\theta_{0, l}$  by projecting the score function onto certain direction. See (4.10) below for details.

An improved estimator for  $\theta_{0, l}$  is then obtained by solving the projected score function while letting  $\boldsymbol{\theta}_{-l} = \tilde{\boldsymbol{\theta}}_{-l}$ . The projection mitigates the impact of  $\tilde{\boldsymbol{\theta}}_{-l}$  in the improved estimation for  $\theta_{0, l}$ . This strategy was initially proposed by Chang et al. (2021) and Chang et al. (2023) for constructing the valid confidence regions of some low-dimensional subvector of the whole parameters in high-dimensional models with removing the impact of the high-dimensional nuisance parameter.

For  $(c_{n, \mathcal{G}}, c_{n, \mathcal{G}^c})$  defined as (4.8), put

$$\Delta_n = \max \{ |\mathcal{G}| c_{n, \mathcal{G}}^2, |\mathcal{G}^c|^2 c_{n, \mathcal{G}^c}^2 \}. \quad (4.9)$$

For any  $t \in [n] \setminus [m]$ ,  $l \in [q]$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , we define

$$\mathbf{g}_t^{(l)}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}_l|} \sum_{(i, j) \in \mathcal{S}_l} \frac{X_{i, j}^t - \gamma_{i, j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i, j}^{t-1}(\boldsymbol{\theta}) \{1 - \gamma_{i, j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial \gamma_{i, j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Then the score function can be written as

$$\frac{\partial \hat{\ell}_{n, p}^{(l)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{n - m} \sum_{t=m+1}^n \mathbf{g}_t^{(l)}(\boldsymbol{\theta}).$$

To estimate  $\theta_{0,l}$ ,  $\boldsymbol{\theta}_{-l}$  can be treated as a nuisance parameter vector. Following [Chang et al. \(2021\)](#) and [Chang et al. \(2023\)](#), we project  $\mathbf{g}_t^{(l)}(\boldsymbol{\theta})$  to form a new estimating function:

$$\hat{f}_t^{(l)}(\boldsymbol{\theta}) = \hat{\boldsymbol{\varphi}}_l^\top \mathbf{g}_t^{(l)}(\boldsymbol{\theta}),$$

where  $\hat{\boldsymbol{\varphi}}_l$  is defined as

$$\hat{\boldsymbol{\varphi}}_l = \arg \min_{\mathbf{u} \in \mathbb{R}^q} \|\mathbf{u}\|_1 \quad \text{s.t.} \quad \left\| \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \mathbf{g}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right\}^\top \mathbf{u} - \mathbf{e}_l \right\|_\infty \leq \tau. \quad (4.10)$$

In the above expression,  $\tau > 0$  is a tuning parameter satisfying  $\tau \lesssim \Delta_n^{1/2}$  with  $\Delta_n$  defined as (4.9),  $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_q)^\top$  is the initial estimator defined as (4.7), and  $\mathbf{e}_l$  is a  $q$ -dimensional vector with the  $l$ -th component being 1 and other components being 0. Then we can re-estimate  $\boldsymbol{\theta}_0$  by  $\check{\boldsymbol{\theta}} = (\check{\theta}_1, \dots, \check{\theta}_q)^\top$ , where

$$\check{\theta}_l = \arg \min_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l}) \right|^2$$

for some  $\tilde{r} > 0$  satisfying  $\max\{c_{n,\mathcal{G}}, c_{n,\mathcal{G}^c}\} \ll \tilde{r} \ll 1$ .

To construct the convergence rate of  $|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty$ , we need the following regularity condition, which is analogous to Condition 1 of [Chang et al. \(2021\)](#) and Condition 7 of [Chang et al. \(2023\)](#). See the discussion there for the validity of such condition.

**Condition 4** For each  $l \in [q]$ , there is a nonrandom vector  $\boldsymbol{\varphi}_l \in \mathbb{R}^q$  such that  $\|\boldsymbol{\varphi}_l\|_1 \leq C_4$  for some universal constant  $C_4 > 0$ , and  $\max_{l \in [q]} \|\hat{\boldsymbol{\varphi}}_l - \boldsymbol{\varphi}_l\|_1 = O_p(\omega_n)$  for some  $\omega_n \rightarrow 0$  satisfying  $\omega_n(\log q)^{1/2} \log(qn) = o(1)$ .

Proposition 2 shows that  $\check{\boldsymbol{\theta}}$  has faster convergence rate than the initial estimator  $\tilde{\boldsymbol{\theta}}$  given in (4.7). The proof of Proposition 2 is given in Section B.3 of the supplementary material.

**Proposition 2** Let the conditions of Proposition 1 and Condition 4 holds. Then  $|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty = O_p(\Delta_n)$ , where  $\Delta_n$  is defined as (4.9).

Based on the obtained  $\check{\boldsymbol{\theta}}$ , we consider the final estimate  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_q)^\top$  for  $\boldsymbol{\theta}_0$  defined as follows:

$$\hat{\theta}_l = \arg \min_{\theta_l \in B(\check{\theta}_l, \check{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_l, \check{\boldsymbol{\theta}}_{-l}) \right|^2 \quad (4.11)$$

for some  $\check{r} > 0$  satisfying  $q\Delta_n \ll \check{r} \ll 1$  with  $\Delta_n$  defined as (4.9).

**Remark 2** Given the initial estimate  $\tilde{\boldsymbol{\theta}}$ , there are three tuning parameters  $(\tau, \tilde{r}, \check{r})$  for deriving our final estimate  $\hat{\boldsymbol{\theta}}$ . For the density-dependent model introduced in Section 3.1, we have  $|\mathcal{G}| = 4$  and  $|\mathcal{G}^c| = 2p$ . For both the persistence model and transitivity model introduced in Sections 3.2 and 3.3, we have  $|\mathcal{G}| = 2$  and  $|\mathcal{G}^c| = 2p$ . Together with Remark 1, we have  $\Delta_n = n^{-1/2}p^{5/2}\log^{3/2}(np)$  for these three models. The improved estimation procedure thus requires  $\tau \lesssim n^{-1/4}p^{5/4}\log^{3/4}(np)$ ,  $n^{-1/4}p^{1/4}\log^{3/4}(np) \ll \tilde{r} \ll 1$  and  $n^{-1/2}p^{7/2}\log^{3/2}(np) \ll \check{r} \ll 1$ , which suggests  $p \ll n^{1/7}(\log n)^{-3/7}$ . In practice, for the three models introduced in Section 3, we compute the final estimate  $\hat{\boldsymbol{\theta}}$  with  $\tau$  proportional to  $n^{-1/4}p^{5/4}\log^{3/4}(np)$  and adopting reasonably large  $\tilde{r}$  and  $\check{r}$ . Numerical experiments in Sections 5 and 6 validate the robustness of our proposed estimation procedure regarding the selections of  $\tilde{r}$  and  $\check{r}$  as long as  $\boldsymbol{\theta}_0$  falls within the defined search range.

For any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and  $l \in [q]$ , define

$$\zeta_{n,l}(\boldsymbol{\theta}) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \left\{ \boldsymbol{\varphi}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^2, \quad (4.12)$$

where  $\boldsymbol{\varphi}_l$  is given in Condition 4. Under Conditions 1 and 4, we have  $|\zeta_{n,l}(\boldsymbol{\theta})| \leq C_1^{-1}C_2^2C_4^2$ , which implies that, for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,  $\zeta_{n,l}(\boldsymbol{\theta})$  is a bounded random variable. To construct the asymptotic distribution of each  $\hat{\boldsymbol{\theta}}_l$ , we require the following condition.

**Condition 5** For each  $l \in [q]$ , there exists some random variable  $\kappa_l \geq 0$  such that  $\zeta_{n,l}(\boldsymbol{\theta}_0) \rightarrow \kappa_l$  in probability as  $n \rightarrow \infty$ .

**Remark 3** For each  $l \in [q]$  and  $t \geq m+1$ , let

$$v_l^{t-1} = \frac{1}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left\{ \boldsymbol{\varphi}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\}^2.$$

As  $\{\zeta_{n,l}(\boldsymbol{\theta}_0)\}_{n \geq m+1}$  is a bounded sequence of random variables for each  $l \in [q]$ , Condition 5 is mild and  $\kappa_l$  is a random variable in general. Generally speaking, the asymptotic distribution of  $\hat{\boldsymbol{\theta}}_l$  is a mixture of normal distributions. See Theorem 2 below for details. However, if the long-run variance of  $\{v_l^{t-1}\}_{l=m+1}^n$  satisfies the condition

$$\text{Var}\left(\frac{1}{\sqrt{n-m}} \sum_{t=m+1}^n v_l^{t-1}\right) = o(\sqrt{n}), \quad (4.13)$$

$\kappa_l$  is reduced to a constant

$$\kappa_l = \lim_{n \rightarrow \infty} \mathbb{E}\left(\frac{1}{n-m} \sum_{t=m+1}^n v_l^{t-1}\right).$$

Then Theorem 2 implies that  $\hat{\theta}_l$  is asymptotically normal distributed. When the sequence  $\{v_l^t\}_{t \geq m}$  is  $\alpha$ -mixing with the mixing coefficients attaining certain convergence rates, (4.13) holds automatically.

**Theorem 2** Let the conditions of Proposition 1 and Conditions 4 and 5 hold. For each  $l \in [q]$ , if  $\sqrt{n|\mathcal{S}_l|} \max\{q\Delta_n^{3/2}, q^2\Delta_n^2\} = o(1)$  with  $\Delta_n$  defined as (4.9), it then holds that

$$\sqrt{n|\mathcal{S}_l|}(\hat{\theta}_l - \theta_{0,l}) \rightarrow \sqrt{\kappa_l} \cdot Z$$

in distribution as  $n \rightarrow \infty$ , where  $Z$  is a standard normally distributed random variable independent of  $\kappa_l$  specified in Condition 5.

The proof of Theorem 2 is given in Section B.4 of the supplementary material.

**Remark 4** (i) Theorem 2 shows that, for the global parameter  $\theta_l$  with  $l \in \mathcal{G}$ ,

$$|\hat{\theta}_l - \theta_{0,l}| = O_p\left(\frac{1}{\sqrt{np}}\right),$$

provided that

$$q \ll \min \left\{ \frac{n^{1/10} p^{1/5}}{|\mathcal{G}|^{3/5} \log^{3/5}(np)}, \frac{n^{1/13} p^{8/13}}{|\mathcal{G}|^{6/13} \log^{9/13}(np)}, \frac{n^{1/10} S_{\mathcal{G}^c, \min}^{3/10}}{|\mathcal{G}^c|^{6/5} p^{2/5} \log^{3/5}(nS_{\mathcal{G}^c, \min})}, \frac{n^{1/13} S_{\mathcal{G}^c, \min}^{6/13}}{|\mathcal{G}^c|^{12/13} p^{4/13} \log^{9/13}(nS_{\mathcal{G}^c, \min})} \right\},$$

and for the local parameter  $\theta_l$  with  $l \in \mathcal{G}^c$ ,

$$|\hat{\theta}_l - \theta_{0,l}| = O_p\left(\frac{1}{\sqrt{n|\mathcal{S}_l|}}\right),$$

provided that

$$q \ll \min \left\{ \frac{n^{1/10} p^{3/5}}{|\mathcal{G}|^{3/5} |\mathcal{S}_l|^{1/5} \log^{3/5}(np)}, \frac{n^{1/13} p^{12/13}}{|\mathcal{G}|^{6/13} |\mathcal{S}_l|^{2/13} \log^{9/13}(np)}, \frac{n^{1/8} p^{1/2}}{|\mathcal{G}|^{1/2} |\mathcal{S}_l|^{1/8} \log^{1/2}(np)}, \frac{n^{1/10} S_{\mathcal{G}^c, \min}^{3/10}}{|\mathcal{G}^c|^{6/5} |\mathcal{S}_l|^{1/5} \log^{3/5}(nS_{\mathcal{G}^c, \min})}, \frac{n^{1/13} S_{\mathcal{G}^c, \min}^{6/13}}{|\mathcal{G}^c|^{12/13} |\mathcal{S}_l|^{2/13} \log^{9/13}(nS_{\mathcal{G}^c, \min})} \right\}.$$

In particular, for the three models introduced in Section 3, the estimators satisfy  $|\hat{\theta}_l - \theta_{0,l}| = O_p(n^{-1/2} p^{-1})$  for  $l \in \mathcal{G}$  if  $p \ll n^{1/23} (\log n)^{-9/23}$ , and  $|\hat{\theta}_l - \theta_{0,l}| = O_p\{(np)^{-1/2}\}$  for  $l \in \mathcal{G}^c$  if  $p \ll n^{1/21} (\log n)^{-3/7}$ . Compared with the results in Theorem 1, the improved estimator  $\hat{\boldsymbol{\theta}}$  achieves a faster convergence rate than the initial estimator  $\tilde{\boldsymbol{\theta}}$ .

(ii) For each  $l \in [q]$ , write

$$\hat{\zeta}_{n,l}(\hat{\boldsymbol{\theta}}) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{1}{\gamma_{i,j}^{t-1}(\hat{\boldsymbol{\theta}})\{1 - \gamma_{i,j}^{t-1}(\hat{\boldsymbol{\theta}})\}} \left\{ \hat{\boldsymbol{\varphi}}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right\}^2$$

with  $\hat{\boldsymbol{\varphi}}_l$  defined as (4.10). Since  $\hat{\zeta}_{n,l}(\hat{\boldsymbol{\theta}}) - \zeta_{n,l}(\boldsymbol{\theta}_0) \rightarrow 0$  in probability as  $n \rightarrow \infty$ , by Corollary 3.2 of Hall and Heyde (1980), it holds that

$$\sqrt{\frac{n|\mathcal{S}_l|}{\hat{\zeta}_{n,l}(\hat{\boldsymbol{\theta}})}}(\hat{\theta}_l - \theta_{0,l}) \rightarrow \mathcal{N}(0, 1) \quad (4.14)$$

in distribution as  $n \rightarrow \infty$ , provided that  $\mathbb{P}(\kappa_l > 0) = 1$ . We can use (4.14) to construct the confidence interval for each  $\theta_l$ .

## 5 Simulation with transitivity models

In this section, we use the transitivity model introduced in Section 3.3 as an example to illustrate numerical behaviour of both the initial estimation proposed in Section 4.3 and the improved estimation suggested in Section 4.4.

### 5.1 Implementation details

Network data  $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  used in the experiments described below are generated according to (2.1), (2.2) and (3.3). For each sample, we generate a sequence of length  $n + 200$ , and discard the first 200 observations.

Regarding implementation of our estimation procedures, recall that  $\mathcal{G}$  and  $\mathcal{G}^c$  are, respectively, the index sets of the global parameters and the local parameters. For the transitivity model (3.3), we have  $\boldsymbol{\theta} = (a, b, \xi_1, \dots, \xi_p, \eta_1, \dots, \eta_p)^\top$ , where  $a$  and  $b$  are the global parameters and  $\{\xi_i\}_{i=1}^p$  and  $\{\eta_i\}_{i=1}^p$  are the local parameters. Hence, for this model we have  $|\mathcal{S}_l| = p(p-1)/2$  and  $|\mathcal{S}_l \cap \mathcal{S}_{l'}| = p-1$  when  $l \in \mathcal{G}$  and  $l' \in \mathcal{G}^c$ . By (4.4), for each given  $l \in \mathcal{G}$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , it holds with probability approaching one that  $\ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}) \geq \bar{C}|\boldsymbol{\theta}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2^2 + 2\bar{C}p^{-1}|\boldsymbol{\theta}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_2^2$  for some universal constant  $\bar{C} > 0$  independent of  $\boldsymbol{\theta}$ , which means that the function  $\ell_{n,p}^{(l)}(\cdot)$  defined as (4.2) exhibits robustness against fluctuations in the values of local parameters when  $l \in \mathcal{G}$  and  $p$  is large.

Motivated by this fact, when we compute the initial estimator  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  for the global parameter vector  $\boldsymbol{\theta}_{0,\mathcal{G}}$ , we can just approximate  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  by  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}^{(\text{app})} = \arg \max_{\boldsymbol{\theta}_{\mathcal{G}}} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_{\mathcal{G}}, \bar{\boldsymbol{\theta}}_{\mathcal{G}^c})$ , for some given  $\bar{\boldsymbol{\theta}}_{\mathcal{G}^c}$  and

$\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  defined as (4.6) for some  $l \in \mathcal{G}$ . This simple idea can significantly improve the computational efficiency. Specifically, note that computing the original  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  requires solving an optimization problem with  $2p + 2$  variables while this alternative approach only requires solving an optimization problem with two variables. Our above discussion guarantees  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}^{(\text{app})}$  can approximate  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$  well. Similarly, when we compute the initial estimator  $\tilde{\boldsymbol{\theta}}_l$  for the local parameter  $\theta_{0,l}$  with  $l \in \mathcal{G}^c$ , we can approximate it by  $\tilde{\boldsymbol{\theta}}_l^{(\text{app})} = \arg \max_{\boldsymbol{\theta}_l} \hat{\ell}_{n,p}^{(l)}(\tilde{\boldsymbol{\theta}}_{\mathcal{G}}^{(\text{app})}, \boldsymbol{\theta}_l, \bar{\boldsymbol{\theta}}_{\mathcal{G}^c \setminus \{l\}})$  with some given  $\bar{\boldsymbol{\theta}}_{\mathcal{G}^c \setminus \{l\}}$ .

In practice, we first estimate the global parameters  $a$  and  $b$  via the Quasi-Newton method, given certain initial values for the local parameters  $\{\xi_i\}_{i=1}^p$  and  $\{\eta_i\}_{i=1}^p$ . To be specific, we consider 9 different sets of the initial values between 0.5 and 0.9 for  $\{\xi_i\}_{i=1}^p$  and  $\{\eta_i\}_{i=1}^p$ , and compute  $\tilde{a}^{(\nu)}$  and  $\tilde{b}^{(\nu)}$  for the  $\nu$ -th initial setting. With  $a = \tilde{a}^{(\nu)}$  and  $b = \tilde{b}^{(\nu)}$ , we then compute  $\tilde{\xi}_1^{(\nu)}, \dots, \tilde{\xi}_p^{(\nu)}, \tilde{\eta}_1^{(\nu)}, \dots, \tilde{\eta}_p^{(\nu)}$  through maximizing each of the associated  $\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  with  $l \in \mathcal{G}^c$ . Subsequently, the improved estimates  $\hat{a}^{(\nu)}, \hat{b}^{(\nu)}, \hat{\xi}_1^{(\nu)}, \dots, \hat{\xi}_p^{(\nu)}, \hat{\eta}_1^{(\nu)}, \dots, \hat{\eta}_p^{(\nu)}$  are obtained according to (4.11) with  $(\tau, \tilde{r}, \tilde{r}) = (0.5\Delta_n^{1/2}, 0.5, 0.1)$  for the local parameters and  $(\tau, \tilde{r}, \tilde{r}) = (0.01\Delta_n^{1/2}, 10, 2)$  for the global parameters. The simulations in this section utilise our development R package `arnetworks`, which provides a user-friendly implementation of the practical estimation algorithm described above.

## 5.2 Estimation errors

Here we report results on experiments exploring the behavior of the initial estimator  $\tilde{\boldsymbol{\theta}}$  given in (4.7) and the improved estimator  $\hat{\boldsymbol{\theta}}$  given in (4.11). For simplicity, we again set all the true values for  $\{\xi_i\}_{i=1}^p$  to be the same, and those for  $\{\eta_i\}_{i=1}^p$  also to be the same. The same four sets of parameter values were used as in Section C.1. We set  $n \in \{100, 200\}$  and  $p \in \{50, 100, 150\}$ . For each setting, we replicate the estimation 400 times.

Table 1 presents the means and the standard errors, over the 400 replications, of the relative mean absolute errors (rMAE):

$$\text{rMAE}(\hat{\xi}_i) = \frac{1}{9} \sum_{\nu=1}^9 \frac{1}{p} \sum_{i=1}^p \left| \frac{\hat{\xi}_i^{(\nu)} - \xi_i}{\xi_i} \right| \quad \text{and} \quad \text{rMAE}(\hat{a}) = \frac{1}{9} \sum_{\nu=1}^9 \left| \frac{\hat{a}^{(\nu)} - a}{a} \right|,$$

where the sum over  $\nu$  corresponds to taking the average over the 9 initial values discussed in Section 5.1. As indicated in the table, the improved estimator (4.11) is significantly more accurate than the initial estimator (4.7). For example, for setting  $(0.6, 0.7, 15, 10)$ , we observe an approxi-

Table 1: The means and STDs (in parenthesis) of rMAEs for estimating parameters in transitivity model (3.3) with 400 replications.

| $(\xi_i, \eta_i, a, b)$ | $p$ | Estimation | $n = 100$     |               |               |               | $n = 200$     |               |               |               |
|-------------------------|-----|------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                         |     |            | $\xi_i$       | $\eta_i$      | $a$           | $b$           | $\xi_i$       | $\eta_i$      | $a$           | $b$           |
| (0.7, 0.8, 30, 15)      | 50  | Initial    | 0.161 (0.023) | 0.070 (0.028) | 0.207 (0.008) | 0.166 (0.003) | 0.157 (0.020) | 0.066 (0.026) | 0.206 (0.006) | 0.167 (0.002) |
|                         |     | Improved   | 0.093 (0.026) | 0.051 (0.029) | 0.062 (0.012) | 0.060 (0.013) | 0.085 (0.022) | 0.044 (0.026) | 0.058 (0.009) | 0.056 (0.006) |
|                         | 100 | Initial    | 0.172 (0.001) | 0.062 (0.002) | 0.293 (0.004) | 0.172 (0.001) | 0.171 (0.001) | 0.060 (0.001) | 0.293 (0.003) | 0.172 (0.001) |
|                         |     | Improved   | 0.126 (0.006) | 0.057 (0.003) | 0.196 (0.022) | 0.134 (0.011) | 0.123 (0.004) | 0.054 (0.003) | 0.194 (0.016) | 0.132 (0.006) |
|                         | 150 | Initial    | 0.177 (0.001) | 0.060 (0.002) | 0.371 (0.003) | 0.173 (0.001) | 0.177 (0.001) | 0.058 (0.001) | 0.371 (0.002) | 0.173 (0.001) |
|                         |     | Improved   | 0.141 (0.015) | 0.050 (0.004) | 0.166 (0.024) | 0.161 (0.017) | 0.135 (0.008) | 0.046 (0.003) | 0.159 (0.015) | 0.158 (0.015) |
| (0.6, 0.7, 20, 20)      | 50  | Initial    | 0.211 (0.004) | 0.093 (0.003) | 0.389 (0.033) | 0.207 (0.003) | 0.205 (0.002) | 0.085 (0.002) | 0.389 (0.023) | 0.207 (0.002) |
|                         |     | Improved   | 0.131 (0.006) | 0.063 (0.005) | 0.157 (0.026) | 0.089 (0.004) | 0.118 (0.003) | 0.051 (0.003) | 0.150 (0.014) | 0.087 (0.003) |
|                         | 100 | Initial    | 0.230 (0.001) | 0.084 (0.002) | 0.530 (0.029) | 0.210 (0.001) | 0.228 (0.001) | 0.080 (0.001) | 0.531 (0.021) | 0.210 (0.001) |
|                         |     | Improved   | 0.152 (0.004) | 0.057 (0.002) | 0.301 (0.028) | 0.114 (0.004) | 0.144 (0.003) | 0.051 (0.001) | 0.293 (0.020) | 0.113 (0.003) |
|                         | 150 | Initial    | 0.238 (0.001) | 0.081 (0.001) | 0.614 (0.016) | 0.212 (0.001) | 0.236 (0.001) | 0.078 (0.001) | 0.614 (0.012) | 0.212 (0.001) |
|                         |     | Improved   | 0.150 (0.002) | 0.055 (0.002) | 0.282 (0.018) | 0.137 (0.006) | 0.144 (0.001) | 0.052 (0.002) | 0.276 (0.015) | 0.135 (0.006) |
| (0.6, 0.7, 15, 10)      | 50  | Initial    | 0.217 (0.003) | 0.097 (0.002) | 0.444 (0.029) | 0.247 (0.003) | 0.213 (0.001) | 0.092 (0.001) | 0.446 (0.022) | 0.248 (0.002) |
|                         |     | Improved   | 0.147 (0.004) | 0.068 (0.005) | 0.243 (0.027) | 0.136 (0.017) | 0.138 (0.004) | 0.061 (0.004) | 0.232 (0.024) | 0.134 (0.017) |
|                         | 100 | Initial    | 0.226 (0.001) | 0.093 (0.001) | 0.582 (0.009) | 0.258 (0.003) | 0.224 (0.001) | 0.091 (0.001) | 0.581 (0.006) | 0.258 (0.002) |
|                         |     | Improved   | 0.142 (0.007) | 0.059 (0.003) | 0.177 (0.018) | 0.195 (0.012) | 0.137 (0.008) | 0.055 (0.003) | 0.169 (0.013) | 0.190 (0.014) |
|                         | 150 | Initial    | 0.230 (0.001) | 0.093 (0.001) | 0.687 (0.002) | 0.267 (0.002) | 0.229 (0.001) | 0.092 (0.001) | 0.687 (0.002) | 0.267 (0.001) |
|                         |     | Improved   | 0.169 (0.001) | 0.057 (0.001) | 0.234 (0.008) | 0.236 (0.007) | 0.166 (0.001) | 0.054 (0.001) | 0.229 (0.005) | 0.233 (0.004) |
| (0.6, 0.7, 10, 10)      | 50  | Initial    | 0.217 (0.003) | 0.098 (0.002) | 0.608 (0.031) | 0.261 (0.003) | 0.212 (0.002) | 0.092 (0.002) | 0.610 (0.022) | 0.261 (0.002) |
|                         |     | Improved   | 0.147 (0.004) | 0.068 (0.005) | 0.316 (0.049) | 0.137 (0.019) | 0.139 (0.004) | 0.059 (0.004) | 0.293 (0.039) | 0.133 (0.017) |
|                         | 100 | Initial    | 0.226 (0.001) | 0.094 (0.001) | 0.769 (0.006) | 0.266 (0.002) | 0.225 (0.001) | 0.092 (0.001) | 0.770 (0.004) | 0.266 (0.002) |
|                         |     | Improved   | 0.143 (0.007) | 0.060 (0.002) | 0.249 (0.032) | 0.201 (0.012) | 0.138 (0.006) | 0.057 (0.003) | 0.239 (0.028) | 0.193 (0.011) |
|                         | 150 | Initial    | 0.231 (0.001) | 0.093 (0.001) | 0.868 (0.005) | 0.271 (0.002) | 0.230 (0.001) | 0.091 (0.001) | 0.868 (0.003) | 0.271 (0.001) |
|                         |     | Improved   | 0.170 (0.003) | 0.058 (0.001) | 0.325 (0.015) | 0.238 (0.009) | 0.168 (0.002) | 0.055 (0.001) | 0.318 (0.004) | 0.235 (0.005) |

mate 70% improvement in the estimation accuracy of  $a$  when  $p = 150$ . Furthermore, the setting (0.7, 0.8, 30, 15) attains the lowest overall estimation errors. This is well-expected, as this is the most dynamic setting among the four settings considered.

## 6 Real data analysis: Email interactions

In this section, we apply the transitivity model (3.3) to a dynamic network dataset of email interactions in a medium-sized Polish manufacturing company, from January to September 2010 (Michalski et al., 2014). We analyze a subset of the data among  $p = 106$  of the most active participants out of an original 167 employees. The organizational tree of direct reports in the company is also available for these employees. Each of the  $n = 39$  network snapshots corresponds to a non-overlapping time window, with  $X_{i,j}^t = 1$  if participants  $i$  and  $j$  exchanged at least one email in the previous seven days. This accounts for periodic weekly effects.

We first present some preliminary summaries of the data to inspect the stationarity of the

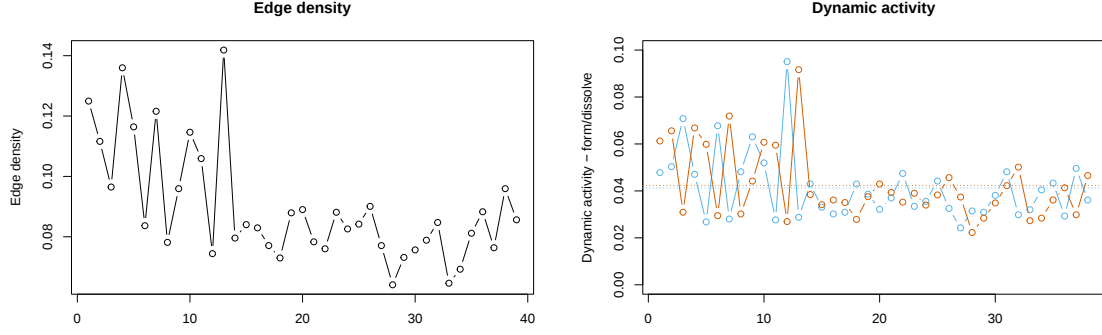


Figure 1: Evolution of edge density  $D_t$  (left panel), percentage of grown  $D_{1,t}$  (blue) and dissolved  $D_{0,t}$  (orange) edges (right panel), manufacturing email networks.

network and the effective sample size. The behavior shown in Figure 1 suggests a change point in the network behavior, in terms of both edge density  $D_t$  and two dynamics density measures  $D_{1,t}$  and  $D_{0,t}$  (see (C.1)). Hence in the following analysis, we fit the model separately to the first 13 and last 26 snapshots, referred to as “period 1” and “period 2”. In the right panel, about 4% of node pairs see a grown edge or a dissolved edge between consecutive snapshots. After accounting for the low edge density, the relative frequency of growing a new edge is about 5%, while relative frequency of an existing edge to dissolve is only about 45%, clear evidence of temporal edge dependence.

We also identify empirical evidence in the data for transitivity effects. This is demonstrated in Figure 2. To construct these plots, we partition the edge variables as follows: for each integer  $\ell \geq 0$ , define

$$\begin{aligned}\mathcal{U}_\ell &= \{(i, j, t) : 1 \leq i < j \leq p, t \in [n] \setminus \{1\}, X_{i,j}^{t-1} = 0, U_{i,j}^{t-1} = \ell/(p-2)\}, \\ \mathcal{V}_\ell &= \{(i, j, t) : 1 \leq i < j \leq p, t \in [n] \setminus \{1\}, X_{i,j}^{t-1} = 1, V_{i,j}^{t-1} = \ell/(p-2)\}, \\ \mathcal{U}_\ell^1 &= \{(i, j, t) \in \mathcal{U}_\ell, X_{i,j}^t = 1\}, \quad \mathcal{V}_\ell^0 = \{(i, j, t) \in \mathcal{V}_\ell, X_{i,j}^t = 0\},\end{aligned}$$

where  $U_{i,j}^{t-1}$  and  $V_{i,j}^{t-1}$  are given in (3.4).

The left panel plots the relative frequency  $|\mathcal{U}_\ell^1|/|\mathcal{U}_\ell|$  against  $\ell$  for  $\ell = 0, 1, \dots$ , showing that this frequency of grown edges tends to be higher for node pairs with more common neighbours in the previous snapshot. The right panel analogously plots the relative frequency  $|\mathcal{V}_\ell^0|/|\mathcal{V}_\ell|$  against  $\ell$ , and shows a similar increasing relationship between disjoint neighbours and frequency of dissolved edges.

This is confirmed by the fit of our model parameters, using the estimation algorithm described

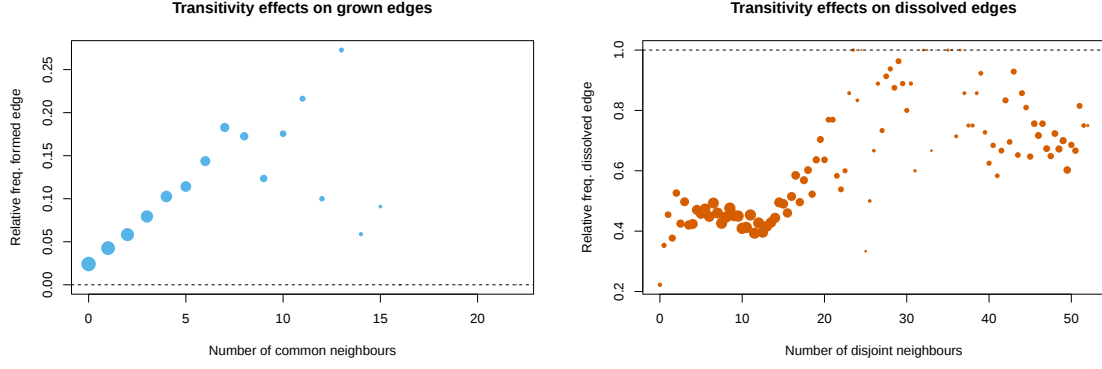


Figure 2: Left panel: the plot of relative edge frequency  $|\mathcal{U}_\ell^1|/|\mathcal{U}_\ell|$  against  $\ell$ . Right panel: the plot of relative non-edge frequency  $|\mathcal{V}_\ell^0|/|\mathcal{V}_\ell|$  against  $\ell$ . In both panels, point size is proportional to  $\log|\mathcal{U}_\ell|$  and  $\log|\mathcal{V}_\ell|$  respectively.

in Section 4, and implemented in our development R package **arnetworks**. For period 1, we estimate the global parameters  $\hat{a} = 13.12$  and  $\hat{b} = 9.34$ , suggesting a tendency towards edge growth given more common neighbours, and edge dissolution given more distinct neighbours, which agrees with the empirical evidence in Figure 2. We interpret the estimates of the local parameters  $\{\xi_i\}_{i=1}^{106}$  and  $\{\eta_i\}_{i=1}^{106}$  in the left panel of Figure 3. The estimates  $\{\hat{\xi}_i\}_{i=1}^{106}$  have mean 0.61 and skew towards the right, implying degree heterogeneity in the edge growth. Conversely, the estimates  $\{\hat{\eta}_i\}_{i=1}^{106}$  have mean 0.89 and skew towards the left. There is a decreasing relationship between the paired parameters: employees who tend to grow new edges also tend to maintain existing edges. Finally, there is an observed relationship between email behavior and company hierarchy: managers (non-leaf nodes in the organizational tree) tend to have larger estimates  $\{\hat{\xi}_i\}_{i=1}^{106}$  compared to non-managers (means 0.74 and 0.57 respectively), implying that managers are more likely to grow edges. However, this increasing pattern does not continue at higher levels of the organizational tree.

The model fit to period 2 shows many of the same patterns. We estimate  $\hat{a} = 21.69$  and  $\hat{b} = 9.84$  and summarize the estimates  $\{\hat{\xi}_i\}_{i=1}^{106}$  and  $\{\hat{\eta}_i\}_{i=1}^{106}$  in the right panel of Figure 3. Relative to period 1, the larger estimate of  $a$  implies a stronger transitivity effect in this time period. The estimates  $\{\hat{\xi}_i\}_{i=1}^{106}$  now have mean 0.49 and the estimates  $\{\hat{\eta}_i\}_{i=1}^{106}$  have mean 0.92, to model overall lower edge density. The decreasing relationship between the paired parameters is stronger, and the means of  $\hat{\xi}_i$  for managers and non-managers are, respectively, 0.68 and 0.43. Along with the stronger transitivity effect, we interpret that the decreased network density in period 2 has led to a concentration of email activity among a smaller group of employees, many of them managers.

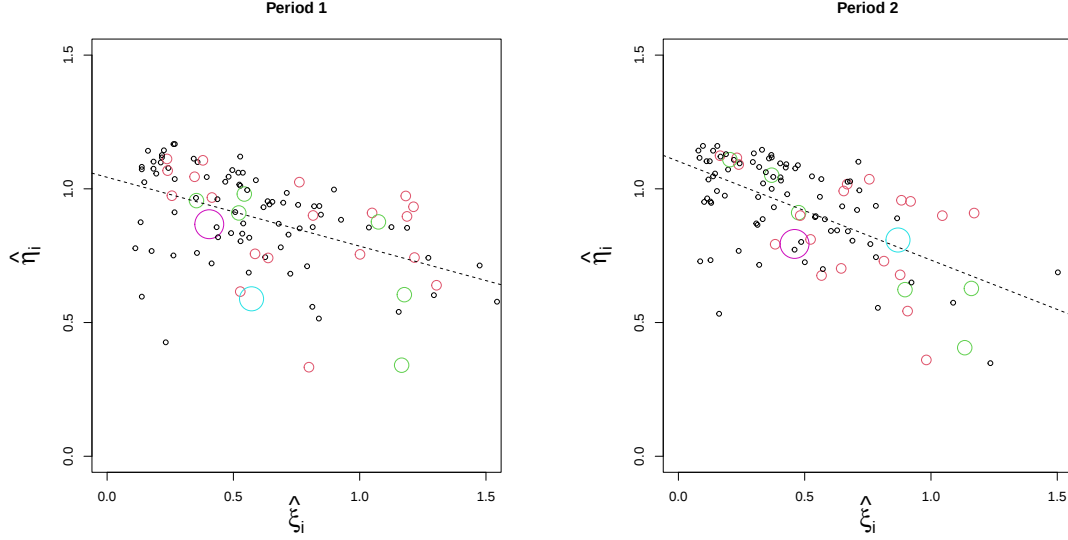


Figure 3: Scatter plots of estimates  $\{\hat{\xi}_i\}_{i=1}^{106}$  and  $\{\hat{\eta}_i\}_{i=1}^{106}$  for periods 1 and 2. Circles are sized and coloured according to their level in the company organizational tree. The smallest black circles have no direct reports, while the largest purple circle is the CEO.

We compare our model to some competing models from the literature in terms of Akaike and Bayesian information criteria (AIC, BIC). To briefly describe these competitors: the “global AR model” and “edgewise AR model” fit the model of [Jiang et al. \(2023b\)](#), with either two global switching parameters or two parameters for each edge. The “edgewise mean model” assumes

$$X_{i,j}^t \stackrel{\text{iid}}{\sim} \text{Bernoulli}(P_{i,j})$$

with no temporal dependence, and estimates the edge probability  $\{P_{i,j}\}_{i,j:i < j}$  for each node pair by its relative frequency in the training set; and the “degree parameter mean model” assumes

$$X_{i,j}^t \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\nu_i \nu_j)$$

and estimates the degree parameters  $\{\nu_i\}_{i=1}^{106}$  by fitting 1-dimensional adjacency spectral embedding ([Athreya et al., 2017](#)) to the mean adjacency matrix over the training set. Note that the edgewise mean model has  $O(p^2)$  parameters, while the degree parameter model has  $O(p)$  parameters like our AR network model with transitivity. All of these models can be directly compared using the AR network model likelihood, although only our AR network model with transitivity incorporates edge dependence, and the final two models do not incorporate any temporal dependence. Results for both regimes are reported in [Table 2](#).

|                             | Period 1     |              | Period 2     |              |
|-----------------------------|--------------|--------------|--------------|--------------|
| Model                       | AIC          | BIC          | AIC          | BIC          |
| Transitivity AR model       | <b>33226</b> | <b>35175</b> | 52547        | <b>54654</b> |
| Global AR model             | 36309        | 36327        | 58267        | 58287        |
| Edgewise AR model           | 42717        | 144102       | 55840        | 165394       |
| Edgewise mean model         | 33248        | 83941        | <b>47133</b> | 101910       |
| Degree parameter mean model | 41730        | 42695        | 68969        | 70013        |

Table 2: AIC and BIC performance for email interaction data, periods 1 and 2.

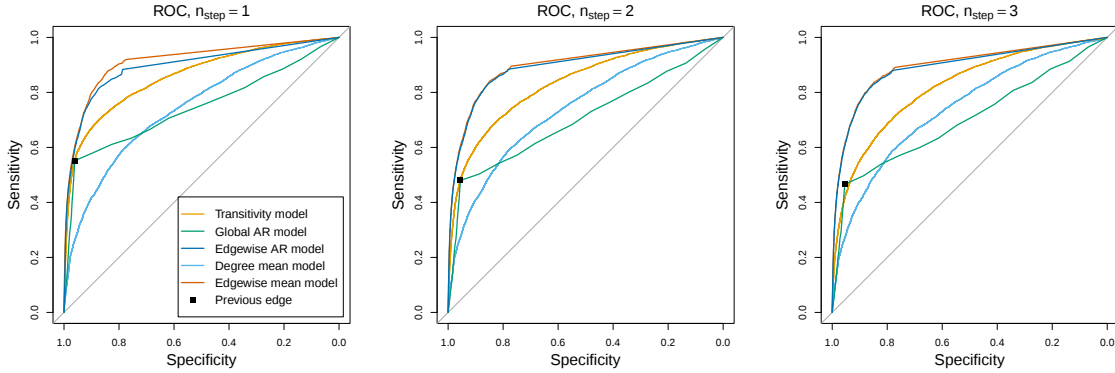


Figure 4: ROC curves for link prediction performance, email interaction data.

In period 1, our AR network model with transitivity achieves the lowest AIC and BIC, while in period 2 it is outperformed slightly by the edgewise mean model in terms of AIC, but achieves a lower BIC as it uses fewer parameters. This reduction of the parameter space is important for modeling sparse dynamic network data: although there is clear temporal edge dependence in this data, the edgewise mean model outperforms the edgewise AR model, as there is low effective sample size to estimate the edge dissolution parameters.

Finally, we compare the performances of those models in an edge forecasting task on the final 26 network snapshots (period 2). For  $n_{\text{train}} = 10, \dots, 23$ , we train these models on the first  $n_{\text{train}}$  snapshots of period 2, then forecast the state of each edge  $n_{\text{step}}$  steps forward, for  $n_{\text{step}} = 1, 2, 3$ . The combined results are presented in Figure 4 as receiver operating characteristic (ROC) curves. We also include a single point summarizing the naive forecasting performance using the most recent observation of that edge in the training set.

Our AR network model with transitivity dominates or is competitive with all the models besides these highly parameterized edgewise models for all choices of  $n_{\text{step}}$ . The good performance of the

edgewise models suggests the presence of higher order structure in this network that cannot be modeled with only two parameters per node. Note that the edgewise mean and edgewise AR models give very similar, but not identical edge predictions; as mentioned above, due to network sparsity the edgewise AR model has a low effective sample size to estimate the dissolution parameters, leading to slightly worse link prediction performance.

## References

- Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y. and Sussman, D. L. (2017). Statistical inference on random dot product graphs: a survey, *Journal of Machine Learning Research (JMLR)* **18**: Paper No. 226, 92.
- Azuma, K. (1967). Weighted sums of certain dependent random variables, *The Tohoku Mathematical Journal. Second Series* **19**: 357–367.
- Brémaud, P. (1998). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer.
- Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F. and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks, *PloS one* **5**(7): e11596.
- Chang, J., Chen, S. X., Tang, C. Y. and Wu, T. (2021). High-dimensional empirical likelihood inference, *Biometrika* **108**(1): 127–147.
- Chang, J., Shi, Z. and Zhang, J. (2023). Culling the herd of moments with penalized empirical likelihood, *Journal of Business & Economic Statistics* **41**(3): 791–805.
- Durante, D. and Dunson, D. B. (2016). Locally adaptive dynamic networks, *The Annals of Applied Statistics* **10**(4): 2203–2232.
- Friel, N., Rastelli, R., Wyse, J. and Raftery, A. (2016). Interlocking directorates in irish companies using a latent space model for bipartite networks, *Proceedings of the national academy of sciences* **113**(24): 6629–6634.
- Génois, M. and Barrat, A. (2018). Can co-location be used as a proxy for face-to-face contacts?, *EPJ Data Science* **7**(1): 1–18.
- Graham, B. (2016). Homophily and transitivity in dynamic network formation, *London: Institute for Fiscal Studies*.
- Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*, Probability and Mathematical Statistics, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.
- Hanneke, S., Fu, W. and Xing, E. (2010). Discrete temporal models of social networks, *Electronic Journal of Statistics* **4**: 585–605.

- Jiang, B., Leng, C., Yan, T., Yao, Q. and Yu, X. (2023a). A two-way heterogeneity model for dynamic networks, [arXiv:2305.12643](#) .
- Jiang, B., Li, J. and Yao, Q. (2023b). Autoregressive networks, *Journal of Machine Learning Research* **24**(227): 1–69.
- Krivitsky, N. and Handcock, M. (2014). A separable model for dynamic networks, *Journal of the Royal Statistical Society, Series B* **76**: 29–48.
- Leifeld, P., Cranmer, S. and Desmarais, B. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals., *Journal of Statistical Software* **83**: 1–36.
- Lesigne, E. and Volný, D. (2001). Large deviations for martingales, *Stochastic Processes and their Applications* **96**(1): 143–159.
- Ludkin, M., Eckley, I. and Neal, P. (2018). Dynamic stochastic block models: parameter estimation and detection of changes in community structure, *Statistics and Computing* **28**(6): 1201–1213.
- Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model, *Journal of the Royal Statistical Society, Series B* **79**: 1119–1141.
- Michalski, R., Kajdanowicz, T., Bródka, P. and Kazienko, P. (2014). Seed selection for spread of influence in social networks: Temporal vs. static approach, *New Generation Computing* **32**(3-4): 213–235.
- Schweinberger, M., Krivitsky, P., Butts, C. and Stewart, J. (2020). Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios, *Statistical Science* **35**: 627–662.
- Snijders, T. A. B. (2005). Models for longitudinal network data, in P. Carrington, J. Scott and S. S. Wasserman (eds), *Models and Methods in Social Network Analysis*, Cambridge University Press, New York, chapter 11.
- Süveges, M. and Olhede, S. C. (2023). Networks with correlated edge processes, *Journal of the Royal Statistical Society, Series A* **186**: 441–462.
- Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks? a bayesian approach, *Machine Learning* **82**: 157–189.
- Yudovina, E., Banerjee, M. and Michailidis, G. (2015). Changepoint inference for erdős-rényi random graphs, *Stochastic Models, Statistics and Their Applications*, Springer, pp. 197–205.

## Supplementary material to “Autoregressive Networks with Dependent Edges”

Jinyuan Chang, Qin Fang, Eric D. Kolaczyk, Peter W. MacDonald, and Qiwei Yao

This supplementary material contains a detailed analysis on the relationship between the proposed AR models and temporal ERGMs (Section A), all the technical proofs (Section B), additional numerical simulation results (Section C), and the analysis of an additional dynamic network dataset (Section D).

### A Relationship to temporal ERGMs

A dynamic network sequence follows a temporal ERGM of order  $m$  if it satisfies

$$\mathbb{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}) \propto \exp\{\boldsymbol{\varsigma}(\boldsymbol{\theta})^\top \boldsymbol{\varrho}(\mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m})\}, \quad (\text{A.1})$$

where  $\boldsymbol{\varsigma} : \mathbb{R}^q \rightarrow \mathbb{R}^p$  maps the parameter vector  $\boldsymbol{\theta}$  to the vector of natural parameters, and  $\boldsymbol{\varrho}$  maps the data, including the past network snapshots, to the corresponding sufficient statistics.

As in Equation (2) of [Hanneke et al. \(2010\)](#), suppose  $\boldsymbol{\varrho}$  factors over the edges of the present snapshot,

$$\boldsymbol{\varrho}(\mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = \sum_{i,j: i < j} \boldsymbol{\varrho}_{i,j}(X_{i,j}^t; \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}).$$

Then

$$\begin{aligned} & \mathbb{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}) \\ & \propto \prod_{i,j: i < j} \exp\{\boldsymbol{\varsigma}(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}, \end{aligned} \quad (\text{A.2})$$

which implies  $\mathbf{X}_t$  will have mutually independent edges conditional on the past snapshots. We refer to this as the edge conditional independence assumption, which is a property of AR network models defined in Definition 1 of the main document. We will show that any edge conditionally independent temporal ERGM can be rewritten as an AR network model.

Denote the logit function by  $\sigma(x) = \log\{x/(1-x)\}$ , and specify an AR network model defined

in Definition 1 by setting

$$\begin{aligned}\alpha_{i,j}^{t-1} &= \sigma^{-1}[\varsigma(\boldsymbol{\theta})^\top \{\boldsymbol{\varrho}_{i,j}(1; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &\quad - \boldsymbol{\varrho}_{i,j}(0; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}] \\ &:= \sigma^{-1}[\varsigma(\boldsymbol{\theta})^\top (\boldsymbol{\varrho}_{i,j,10}^{t-1} - \boldsymbol{\varrho}_{i,j,00}^{t-1})],\end{aligned}\tag{A.3}$$

$$\begin{aligned}\beta_{i,j}^{t-1} &= \sigma^{-1}[\varsigma(\boldsymbol{\theta})^\top \{\boldsymbol{\varrho}_{i,j}(0; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &\quad - \boldsymbol{\varrho}_{i,j}(1; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\}] \\ &:= \sigma^{-1}[\varsigma(\boldsymbol{\theta})^\top (\boldsymbol{\varrho}_{i,j,01}^{t-1} - \boldsymbol{\varrho}_{i,j,11}^{t-1})].\end{aligned}\tag{A.4}$$

With renormalizing, we have

$$\begin{aligned}\alpha_{i,j}^{t-1} &= \frac{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,10}^{t-1}\}}{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,10}^{t-1}\} + \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,00}^{t-1}\}}, \\ 1 - \alpha_{i,j}^{t-1} &= \frac{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,00}^{t-1}\}}{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,10}^{t-1}\} + \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,00}^{t-1}\}}, \\ \beta_{i,j}^{t-1} &= \frac{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,01}^{t-1}\}}{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,01}^{t-1}\} + \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,11}^{t-1}\}}, \\ 1 - \beta_{i,j}^{t-1} &= \frac{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,11}^{t-1}\}}{\exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,01}^{t-1}\} + \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,11}^{t-1}\}}.\end{aligned}$$

For the AR network model with  $(\alpha_{i,j}^{t-1}, \beta_{i,j}^{t-1})$  specified in (A.3) and (A.4), it holds that

$$\begin{aligned}\mathbb{P}(\mathbf{X}_t | \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}) &= \prod_{i,j: i < j} \mathbb{P}(X_{i,j}^t | \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}) \\ &= \prod_{i,j: i < j, X_{i,j}^t=0, X_{i,j}^{t-1}=0} (1 - \alpha_{i,j}^{t-1}) \cdot \prod_{i,j: i < j, X_{i,j}^t=1, X_{i,j}^{t-1}=0} \alpha_{i,j}^{t-1} \\ &\quad \cdot \prod_{i,j: i < j, X_{i,j}^t=0, X_{i,j}^{t-1}=1} \beta_{i,j}^{t-1} \cdot \prod_{i,j: i < j, X_{i,j}^t=1, X_{i,j}^{t-1}=1} (1 - \beta_{i,j}^{t-1}) \\ &\propto \prod_{i,j: i < j, X_{i,j}^t=0, X_{i,j}^{t-1}=0} \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,00}^{t-1}\} \cdot \prod_{i,j: i < j, X_{i,j}^t=1, X_{i,j}^{t-1}=0} \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,10}^{t-1}\} \\ &\quad \cdot \prod_{i,j: i < j, X_{i,j}^t=0, X_{i,j}^{t-1}=1} \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,01}^{t-1}\} \cdot \prod_{i,j: i < j, X_{i,j}^t=1, X_{i,j}^{t-1}=1} \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j,11}^{t-1}\} \\ &= \prod_{i,j: i < j} \exp\{\varsigma(\boldsymbol{\theta})^\top \boldsymbol{\varrho}_{i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\},\end{aligned}$$

which shares the same form as (A.2). Thus for any edge conditionally independent temporal ERGM, we can specify an AR network model with the same distribution.

Conversely, suppose we have specified an AR network model such that

$$\begin{aligned}\sigma(\alpha_{i,j}^{t-1}) &= \phi(\boldsymbol{\theta})^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}), \\ \sigma(\beta_{i,j}^{t-1}) &= \psi(\boldsymbol{\theta})^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}),\end{aligned}$$

for some functions of the parameter vector  $\boldsymbol{\theta}$ , and the past network behavior. We claim that this AR network model can be written as an edge conditionally independent temporal ERGM (A.1) with  $\boldsymbol{\varsigma}(\boldsymbol{\theta}) = (\phi(\boldsymbol{\theta})^\top, \psi(\boldsymbol{\theta})^\top)^\top$  and  $\boldsymbol{\varrho}(\mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = (\boldsymbol{\varrho}_\alpha^\top, \boldsymbol{\varrho}_\beta^\top)^\top$ , where  $\boldsymbol{\varrho}_\alpha = \sum_{i,j: i < j} \boldsymbol{\varrho}_{\alpha,i,j}$  and  $\boldsymbol{\varrho}_\beta = \sum_{i,j: i < j} \boldsymbol{\varrho}_{\beta,i,j}$  with

$$\begin{aligned}\boldsymbol{\varrho}_{\alpha,i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &= \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^t = 1, X_{i,j}^{t-1} = 0), \\ \boldsymbol{\varrho}_{\beta,i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &= \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^t = 0, X_{i,j}^{t-1} = 1).\end{aligned}$$

Krivitsky and Handcock (2014) define the concept of separability of a dynamic model for binary networks. Define two subnetworks  $\mathbf{X}_t^+$  and  $\mathbf{X}_t^-$ , where

$$\begin{aligned}X_{i,j}^{t,+} &= 1 - I(X_{i,j}^t = 0, X_{i,j}^{t-1} = 0), \\ X_{i,j}^{t,-} &= I(X_{i,j}^t = 1, X_{i,j}^{t-1} = 1),\end{aligned}$$

for all  $i, j$ . A dynamic network model is said to be separable if  $\mathbf{X}_t^+$  and  $\mathbf{X}_t^-$  are independent conditional on the past, and do not share any parameters. In particular, a separable temporal ERGM (STERGM) can be specified by a product of a formation model and a dissolution model:

$$\begin{aligned}\mathbb{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}; \boldsymbol{\theta}) \\ \propto \exp \left\{ \boldsymbol{\varsigma}^+(\boldsymbol{\theta}^+)^\top \boldsymbol{\varrho}^+(\mathbf{X}_t^+, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) + \boldsymbol{\varsigma}^-(\boldsymbol{\theta}^-)^\top \boldsymbol{\varrho}^-(\mathbf{X}_t^-, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) \right\},\end{aligned}\quad (\text{A.5})$$

where  $\boldsymbol{\varsigma}^+$  and  $\boldsymbol{\varsigma}^-$  map parameter vectors  $\boldsymbol{\theta}^+$  and  $\boldsymbol{\theta}^-$  to the vectors of natural parameters, and  $\boldsymbol{\varrho}^+$  and  $\boldsymbol{\varrho}^-$  map the data, including the past network snapshots, to the corresponding sufficient statistics.

Under the edge conditional independence assumption, we can write

$$\boldsymbol{\varrho}^+(\mathbf{X}_t^+, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = \sum_{i < j} \boldsymbol{\varrho}_{i,j}^+(X_{i,j}^{t,+}; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \dots, \mathbf{X}_{t-m}),$$

or equivalently

$$\boldsymbol{\varrho}^+(\mathbf{X}_t^+, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = \sum_{i < j} \tilde{\boldsymbol{\varrho}}_{i,j}^+(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \dots, \mathbf{X}_{t-m}),$$

since for all  $i < j$ ,  $X_{i,j}^{t,+}$  can be recovered from  $X_{i,j}^t$  and  $X_{i,j}^{t-1}$ . Define  $\tilde{\boldsymbol{\varrho}}_{i,j}^-$  analogously for the dissolution model.

In this way  $X_t$  is an edge conditionally independent TERGM with parameter vector  $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ , natural parameter

$$(\varsigma^+(\boldsymbol{\theta}^+), \varsigma^-(\boldsymbol{\theta}^-)),$$

and sufficient statistic

$$\left( \sum_{i < j} \tilde{\boldsymbol{\varrho}}_{i,j}^+, \sum_{i < j} \tilde{\boldsymbol{\varrho}}_{i,j}^- \right).$$

Following the above construction to rewrite this as an AR network model, we can write

$$\begin{aligned} \alpha_{i,j}^{t-1} = & \sigma^{-1} [\varsigma^+(\boldsymbol{\theta}^+)^\top \{ \tilde{\boldsymbol{\varrho}}_{i,j}^+(1; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ & - \tilde{\boldsymbol{\varrho}}_{i,j}^+(0; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \} \\ & + \varsigma^-(\boldsymbol{\theta}^-)^\top \{ \tilde{\boldsymbol{\varrho}}_{i,j}^-(1; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ & - \tilde{\boldsymbol{\varrho}}_{i,j}^-(0; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \} ], \end{aligned}$$

and note that

$$\begin{aligned} \tilde{\boldsymbol{\varrho}}_{i,j}^-(1; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) &= \tilde{\boldsymbol{\varrho}}_{i,j}^-(0; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &= \boldsymbol{\varrho}_{i,j}^-(0; 0, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \end{aligned}$$

for all  $i < j$  and  $t$ , since  $X_{i,j}^{t,-} = 0$  whenever  $X_{i,j}^{t-1} = 0$ . Thus  $\alpha_{i,j}^{t-1}$  is free of  $\boldsymbol{\theta}^-$ . Similarly, we can write

$$\begin{aligned} \beta_{i,j}^{t-1} = & \sigma^{-1} [\varsigma^+(\boldsymbol{\theta}^+)^\top \{ \tilde{\boldsymbol{\varrho}}_{i,j}^+(0; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ & - \tilde{\boldsymbol{\varrho}}_{i,j}^+(1; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \} \\ & + \varsigma^-(\boldsymbol{\theta}^-)^\top \{ \tilde{\boldsymbol{\varrho}}_{i,j}^-(0; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ & - \tilde{\boldsymbol{\varrho}}_{i,j}^-(1; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \} ], \end{aligned}$$

and

$$\begin{aligned} \tilde{\boldsymbol{\varrho}}_{i,j}^+(0; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) &= \tilde{\boldsymbol{\varrho}}_{i,j}^+(1; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ &= \boldsymbol{\varrho}_{i,j}^+(1; 1, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \end{aligned}$$

for all  $i < j$  and  $t$ , since  $X_{i,j}^{t,+} = 1$  whenever  $X_{i,j}^{t-1} = 1$ . Then  $\beta_{i,j}^{t-1}$  is free of  $\boldsymbol{\theta}^+$ . Hence any edge conditionally independent STERGM can be written as an AR network model with separable parameters.

Conversely, suppose we have specified an AR network model such that

$$\begin{aligned}\sigma(\alpha_{i,j}^{t-1}) &= \phi(\boldsymbol{\theta}_\alpha)^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ \sigma(\beta_{i,j}^{t-1}) &= \psi(\boldsymbol{\theta}_\beta)^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m})\end{aligned}$$

with separable parameters  $\boldsymbol{\theta}_\alpha$  and  $\boldsymbol{\theta}_\beta$ .

We follow the same construction as above to rewrite this model as an edge conditionally independent TERGM, with parameter vector  $\boldsymbol{\varsigma}(\boldsymbol{\theta}) = (\phi(\boldsymbol{\theta}_\alpha)^\top, \psi(\boldsymbol{\theta}_\beta)^\top)^\top$  and sufficient statistics  $\boldsymbol{\varrho}(\mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-m}) = (\boldsymbol{\varrho}_\alpha^\top, \boldsymbol{\varrho}_\beta^\top)^\top$ , where  $\boldsymbol{\varrho}_\alpha = \sum_{i,j: i < j} \boldsymbol{\varrho}_{\alpha,i,j}$  and  $\boldsymbol{\varrho}_\beta = \sum_{i,j: i < j} \boldsymbol{\varrho}_{\beta,i,j}$  with

$$\begin{aligned}\boldsymbol{\varrho}_{\alpha,i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ = \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^t = 1, X_{i,j}^{t-1} = 0), \\ = \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^{t,+} = 1, X_{i,j}^{t-1} = 0),\end{aligned}\tag{A.6}$$

$$\begin{aligned}\boldsymbol{\varrho}_{\beta,i,j}(X_{i,j}^t; X_{i,j}^{t-1}, \mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) \\ = \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^t = 0, X_{i,j}^{t-1} = 1). \\ = \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-m}) I(X_{i,j}^{t,-} = 0, X_{i,j}^{t-1} = 1).\end{aligned}\tag{A.7}$$

Note that in (A.6) and (A.7), the sufficient statistics depend on  $X_{i,j}^t$  only through  $X_{i,j}^{t,+}$  and  $X_{i,j}^{t,-}$  respectively. It follows that when written as a conditionally independent TERGM, the distribution factors into a product of a formation model and dissolution model, as in (A.5). Thus this AR network model is an edge conditionally independent STERGM.

## B Technical proofs

### B.1 Proof of Proposition 1

Recall

$$\ell_{n,p}^{(l)}(\boldsymbol{\theta}) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} h_{i,j}^{t-1}(\boldsymbol{\theta})$$

with

$$h_{i,j}^{t-1}(\boldsymbol{\theta}) = \log\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\} + \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \log\left\{\frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}\right\}.$$

Then

$$\frac{\partial h_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} = \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}}, \quad (\text{B.1})$$

$$\begin{aligned} \frac{\partial^2 h_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2}} = & - \left[ \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} + \frac{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}\{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} \right] \\ & \times \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2}} \\ & + \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2}}, \end{aligned} \quad (\text{B.2})$$

$$\begin{aligned} \frac{\partial^3 h_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}} = & 2 \left[ \frac{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} + \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} \right. \\ & \left. + \frac{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}\{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^3\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^3} \right] \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3}} \\ & - \left[ \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} + \frac{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}\{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} \right] \\ & \times \left\{ \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3}} + \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2} \partial \theta_{l_3}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} + \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3} \partial \theta_{l_1}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2}} \right\} \\ & + \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial^3 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}}. \end{aligned} \quad (\text{B.3})$$

By the triangle inequality and Condition 1,

$$\begin{aligned} \left| \frac{\partial^3 h_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}} \right| & \leq 2(2C_1^{-2} + C_1^{-3}) \left| \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_\infty^3 + C_1^{-1} \left| \frac{\partial^3 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^3} \right|_\infty \\ & \quad + 3(C_1^{-1} + C_1^{-2}) \left| \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right|_\infty \left| \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_\infty \\ & \leq 2(2C_1^{-2} + C_1^{-3})C_2^3 + 3(C_1^{-1} + C_1^{-2})C_2^2 + C_1^{-1}C_2 =: C_*. \end{aligned} \quad (\text{B.4})$$

Write  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$  and  $\boldsymbol{\theta}_0 = (\theta_{0,1}, \dots, \theta_{0,q})^\top$ . By Taylor's theorem, (B.1) and (B.2),

$$\begin{aligned} h_{i,j}^{t-1}(\boldsymbol{\theta}) - h_{i,j}^{t-1}(\boldsymbol{\theta}_0) & = \frac{\partial h_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \frac{\partial^2 h_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + R_{i,j}^{t-1}(\boldsymbol{\theta}) \\ & = - \frac{1}{2\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left| \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} (\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \right|^2 + R_{i,j}^{t-1}(\boldsymbol{\theta}), \end{aligned}$$

where

$$\begin{aligned} R_{i,j}^{t-1}(\boldsymbol{\theta}) & = \frac{1}{2} \sum_{l_1=1}^q (\theta_{l_1} - \theta_{0,l_1})^3 \int_0^1 (1-v)^2 \frac{\partial^3 h_{i,j}^{t-1}(\boldsymbol{\theta}_v)}{\partial \theta_{l_1}^3} dv \\ & \quad + \frac{3}{2} \sum_{l_1 \neq l_2} (\theta_{l_1} - \theta_{0,l_1})^2 (\theta_{l_2} - \theta_{0,l_2}) \int_0^1 (1-v)^2 \frac{\partial^3 h_{i,j}^{t-1}(\boldsymbol{\theta}_v)}{\partial \theta_{l_1}^2 \partial \theta_{l_2}} dv \end{aligned}$$

$$+ 3 \sum_{l_1 \neq l_2 \neq l_3} (\theta_{l_1} - \theta_{0,l_1})(\theta_{l_2} - \theta_{0,l_2})(\theta_{l_3} - \theta_{0,l_3}) \int_0^1 (1-v)^2 \frac{\partial^3 h_{i,j}^{t-1}(\theta_v)}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}} dv$$

with  $\theta_v = \theta_0 + v(\theta - \theta_0)$ . Recall  $\mathcal{I}_{i,j} = \{l \subset [q] : \gamma_{i,j}^{t-1}(\theta) \text{ involves } \theta_l \text{ for any } t \in [n] \setminus [m]\}$ . We have  $\partial \gamma_{i,j}^{t-1}(\theta) / \partial \theta_{l_1} \equiv 0$  if  $l_1 \notin \mathcal{I}_{i,j}$ ,  $\partial^2 \gamma_{i,j}^{t-1}(\theta) / \partial \theta_{l_1} \partial \theta_{l_2} \equiv 0$  if  $l_1 \notin \mathcal{I}_{i,j}$  or  $l_2 \notin \mathcal{I}_{i,j}$ , and  $\partial^3 \gamma_{i,j}^{t-1}(\theta) / \partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3} \equiv 0$  if  $l_1 \notin \mathcal{I}_{i,j}$  or  $l_2 \notin \mathcal{I}_{i,j}$  or  $l_3 \notin \mathcal{I}_{i,j}$ . By (B.3), it holds that

$$\frac{\partial^3 h_{i,j}^{t-1}(\theta)}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}} \equiv 0 \text{ if } l_1 \notin \mathcal{I}_{i,j} \text{ or } l_2 \notin \mathcal{I}_{i,j} \text{ or } l_3 \notin \mathcal{I}_{i,j}.$$

By Condition 2, it holds that  $|R_{i,j}^{t-1}(\theta)| \leq C_* s^3 |\theta_{\mathcal{I}_{i,j}} - \theta_{0,\mathcal{I}_{i,j}}|_\infty^3$ , which implies

$$\begin{aligned} \ell_{n,p}^{(l)}(\theta_0) - \ell_{n,p}^{(l)}(\theta) &= \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \{h_{i,j}^{t-1}(\theta_0) - h_{i,j}^{t-1}(\theta)\} \\ &\geq \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{1}{2\gamma_{i,j}^{t-1}(\theta_0)\{1 - \gamma_{i,j}^{t-1}(\theta_0)\}} \left| \frac{\partial \gamma_{i,j}^{t-1}(\theta_0)}{\partial \theta^\top} (\theta_0 - \theta) \right|^2 \\ &\quad - \frac{C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{\mathcal{I}_{i,j}} - \theta_{0,\mathcal{I}_{i,j}}|_\infty^3 \\ &\geq \frac{2}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \left| \frac{\partial \gamma_{i,j}^{t-1}(\theta_0)}{\partial \theta_{\mathcal{I}_{i,j}}^\top} (\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}) \right|^2 - \frac{C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{\mathcal{I}_{i,j}} - \theta_{0,\mathcal{I}_{i,j}}|_\infty^3 \end{aligned} \quad (\text{B.5})$$

for any  $\theta \in \Theta$ . By Condition 3, it holds with probability approaching one that

$$\frac{1}{n-m} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \left| \frac{\partial \gamma_{i,j}^{t-1}(\theta_0)}{\partial \theta_{\mathcal{I}_{i,j}}^\top} (\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}) \right|^2 \geq C_3 \sum_{(i,j) \in \mathcal{S}_l} |\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}|_2^2$$

for any  $l \in [q]$ , which implies

$$\ell_{n,p}^{(l)}(\theta_0) - \ell_{n,p}^{(l)}(\theta) \geq \frac{2C_3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}|_2^2 - \frac{C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{\mathcal{I}_{i,j}} - \theta_{0,\mathcal{I}_{i,j}}|_\infty^3$$

for any  $\theta \in \Theta$  and  $l \in [q]$  with probability approaching one. Since  $d := \sup_{\theta \in \Theta} |\theta - \theta_0|_\infty < 2C_3/(C_* s^3)$ , there exists a universal constant  $\tilde{C} > 0$  such that  $d < \tilde{C} < 2C_3/(C_* s^3)$ . Hence, it holds with probability approaching one that

$$\begin{aligned} \ell_{n,p}^{(l)}(\theta_0) - \ell_{n,p}^{(l)}(\theta) &\geq \frac{2C_3 - \tilde{C}C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}|_2^2 \\ &\quad + \frac{\tilde{C}C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}|_2^2 - \frac{C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{\mathcal{I}_{i,j}} - \theta_{0,\mathcal{I}_{i,j}}|_\infty^3 \\ &\geq \frac{2C_3 - \tilde{C}C_* s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\theta_{0,\mathcal{I}_{i,j}} - \theta_{\mathcal{I}_{i,j}}|_2^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{\tilde{C}C_*s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\boldsymbol{\theta}_{0,\mathcal{I}_{i,j}} - \boldsymbol{\theta}_{\mathcal{I}_{i,j}}|_2^2 - \frac{C_*s^3d}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\boldsymbol{\theta}_{\mathcal{I}_{i,j}} - \boldsymbol{\theta}_{0,\mathcal{I}_{i,j}}|_\infty^2 \\
& \geq \frac{2C_3 - \tilde{C}C_*s^3}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} |\boldsymbol{\theta}_{0,\mathcal{I}_{i,j}} - \boldsymbol{\theta}_{\mathcal{I}_{i,j}}|_2^2
\end{aligned}$$

for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and  $l \in [q]$ . We then have Proposition 1 by selecting  $\bar{C} = 2C_3 - \tilde{C}C_*s^3$ .  $\square$

## B.2 Proof of Theorem 1

Recall  $S_{\mathcal{H},\min} = \min_{l \in \mathcal{H}} |\mathcal{S}_l|$  for any  $\mathcal{H} \subset [q]$ , and

$$\begin{cases} c_{n,\mathcal{G}}^2 = \frac{q \log(n|\mathcal{S}_{l'}|)}{\sqrt{n}|\mathcal{S}_{l'}|} + \frac{q^{3/2} \log^{3/2}(n|\mathcal{S}_{l'}|)}{\sqrt{n}|\mathcal{S}_{l'}|}, \\ c_{n,\mathcal{G}^c}^2 = \frac{q \log(nS_{\mathcal{G}^c,\min})}{\sqrt{n}S_{\mathcal{G}^c,\min}} + \frac{q^{3/2} \log^{3/2}(nS_{\mathcal{G}^c,\min})}{\sqrt{n}S_{\mathcal{G}^c,\min}}, \end{cases}$$

where  $l' \in \mathcal{G}$ . To show Theorem 1, we first present the following lemma whose proof is given in Section B.5.1.

**Lemma 1** *Assume Conditions 1 and 2 hold. Then*

$$\max_{l \in \mathcal{H}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})| = O_p \left\{ \frac{q \log(nS_{\mathcal{H},\min})}{\sqrt{n}S_{\mathcal{H},\min}} \right\} + O_p \left\{ \frac{q^{3/2} \log^{3/2}(nS_{\mathcal{H},\min})}{\sqrt{n}S_{\mathcal{H},\min}} \right\}$$

for any  $\mathcal{H} \subset [q]$ .

Notice that  $\hat{\boldsymbol{\theta}}_*^{(l)} = (\hat{\theta}_{*,1}^{(l)}, \dots, \hat{\theta}_{*,q}^{(l)})^\top = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$  for any  $l \in [q]$ . Then

$$\begin{aligned}
\ell_{n,p}^{(l')}(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta}) - \ell_{n,p}^{(l')}(\boldsymbol{\theta})| & \leq \hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta}_0) \\
& \leq \hat{\ell}_{n,p}^{(l')}(\hat{\boldsymbol{\theta}}_*^{(l')}) \leq \ell_{n,p}^{(l')}(\hat{\boldsymbol{\theta}}_*^{(l')}) + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta}) - \ell_{n,p}^{(l')}(\boldsymbol{\theta})|,
\end{aligned}$$

which implies

$$0 \leq \ell_{n,p}^{(l')}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l')}(\hat{\boldsymbol{\theta}}_*^{(l')}) \leq 2 \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta}) - \ell_{n,p}^{(l')}(\boldsymbol{\theta})|.$$

Recall  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}} = \hat{\boldsymbol{\theta}}_{*,\mathcal{G}}^{(l')}$ . Selecting  $\mathcal{H} = \{l'\}$  in Lemma 1, we have  $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l')}(\boldsymbol{\theta}) - \ell_{n,p}^{(l')}(\boldsymbol{\theta})| = O_p(c_{n,\mathcal{G}}^2)$ , which implies

$$\ell_{n,p}^{(l')}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l')}(\hat{\boldsymbol{\theta}}_*^{(l')}) = O_p(c_{n,\mathcal{G}}^2). \tag{B.6}$$

For any diverging  $\epsilon_{n,p} > 0$ , if  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2 \geq \epsilon_{n,p}c_{n,\mathcal{G}}$ , Proposition 1 yields that

$$\ell_{n,p}^{(l')}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l')}(\hat{\boldsymbol{\theta}}_*^{(l')}) \geq \bar{C}\epsilon_{n,p}^2c_{n,\mathcal{G}}^2$$

with probability approaching one, which contradicts with (B.6) and then implies  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2 = O_p(\epsilon_{n,p}c_{n,\mathcal{G}})$ . Notice that we can select arbitrary slowly diverging  $\epsilon_{n,p}$ . Following a standard result from probability theory, we have  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2 = O_p(c_{n,\mathcal{G}})$ .

For any  $l \in \mathcal{G}^c$ , we also have

$$0 \leq \ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\hat{\boldsymbol{\theta}}_*^{(l)}) \leq 2 \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})|.$$

Recall  $\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} = (\hat{\boldsymbol{\theta}}_{*,l}^{(l)})_{l \in \mathcal{G}^c}$ . Selecting  $\mathcal{H} = \mathcal{G}^c$  in Lemma 1, we have  $\max_{l \in \mathcal{G}^c} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})| = O_p(c_{n,\mathcal{G}^c}^2)$ , which implies

$$\max_{l \in \mathcal{G}^c} \{\ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\hat{\boldsymbol{\theta}}_*^{(l)})\} = O_p(c_{n,\mathcal{G}^c}^2). \quad (\text{B.7})$$

For any diverging  $\epsilon_{n,p} > 0$ , if  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_\infty \geq \epsilon_{n,p}c_{n,\mathcal{G}^c}$ , Proposition 1 yields that

$$\max_{l \in \mathcal{G}^c} \{\ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\hat{\boldsymbol{\theta}}_*^{(l)})\} \geq \bar{C}\epsilon_{n,p}^2c_{n,\mathcal{G}^c}^2$$

with probability approaching one, which contradicts with (B.7) and then implies  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_\infty = O_p(\epsilon_{n,p}c_{n,\mathcal{G}^c})$ . Notice that we can select arbitrary slowly diverging  $\epsilon_{n,p}$ . Following a standard result from probability theory, we have  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_\infty = O_p(c_{n,\mathcal{G}^c})$ . We complete the proof of Theorem 1.

□

### B.3 Proof of Proposition 2

Given  $\boldsymbol{\varphi}_l$  specified in Condition 4, define

$$f_t^{(l)}(\boldsymbol{\theta}) = \boldsymbol{\varphi}_l^\top \mathbf{g}_t^{(l)}(\boldsymbol{\theta})$$

for any  $t \in [n] \setminus [m]$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . To show Proposition 2, we need the following lemmas whose proofs are given in Sections B.5.2–B.5.4.

**Lemma 2** *Assume Condition 1 holds. Then*

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\boldsymbol{\theta}_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) \} \right| \leq \tau |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 + C_* |\hat{\boldsymbol{\varphi}}_l|_1 |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1^2$$

for any  $l \in [q]$ , where  $C_*$  is a universal constant specified in (B.4).

**Lemma 3** *Assume Conditions 1 and 4 hold. Then*

$$\begin{aligned} & \max_{l \in \mathcal{H}} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) \right| \\ &= O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{H}|) \log(n|\mathcal{H}|)}}{\sqrt{n}S_{\mathcal{H},\min}} \right\} + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{H}|) \log(n|\mathcal{H}|)}}{\sqrt{n}S_{\mathcal{H},\min}} \right\} \end{aligned}$$

for any  $\mathcal{H} \subset [q]$ .

**Lemma 4** Assume Condition 1 holds. Then

$$\sup_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} - 1 \right| \leq \tau + C_* \tilde{r} |\hat{\varphi}_l|_1$$

for any  $l \in [q]$ , where  $C_*$  is a universal constant specified in (B.4).

Recall  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_2 = O_p(c_{n,\mathcal{G}})$  and  $|\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_\infty = O_p(c_{n,\mathcal{G}^c})$  with

$$\begin{cases} c_{n,\mathcal{G}}^2 = \frac{q \log(n|\mathcal{S}_{l'}|)}{\sqrt{n}|\mathcal{S}_{l'}|} + \frac{q^{3/2} \log^{3/2}(n|\mathcal{S}_{l'}|)}{\sqrt{n}|\mathcal{S}_{l'}|}, \\ c_{n,\mathcal{G}^c}^2 = \frac{q \log(nS_{\mathcal{G}^c, \min})}{\sqrt{n}S_{\mathcal{G}^c, \min}} + \frac{q^{3/2} \log^{3/2}(nS_{\mathcal{G}^c, \min})}{\sqrt{n}S_{\mathcal{G}^c, \min}}, \end{cases}$$

where  $l' \in \mathcal{G}$ . Then  $|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 = |\tilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}_{0,\mathcal{G}}|_1 + |\tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} - \boldsymbol{\theta}_{0,\mathcal{G}^c}|_1 = O_p(|\mathcal{G}|^{1/2}c_{n,\mathcal{G}}) + O_p(|\mathcal{G}^c|c_{n,\mathcal{G}^c}) = O_p(\Delta_n^{1/2})$ . Based on Lemmas 2 and 3, due to  $\tau \lesssim \Delta_n^{1/2}$  and  $|\mathcal{G}| + |\mathcal{G}^c| = q$ , we have

$$\begin{aligned} & \max_{l \in \mathcal{G}} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) \right| \\ &= O_p(\Delta_n) + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{G}|)} \log(n|\mathcal{G}|)}{\sqrt{n}|\mathcal{S}_{l'}|} \right\} + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{G}|)} \log(n|\mathcal{G}|)}{\sqrt{n}|\mathcal{S}_{l'}|} \right\} \\ &= O_p(\Delta_n) \end{aligned}$$

for some  $l' \in \mathcal{G}$ , and

$$\begin{aligned} & \max_{l \in \mathcal{G}^c} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) \right| \\ &= O_p(\Delta_n) + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{G}^c|)} \log(n|\mathcal{G}^c|)}{\sqrt{n}S_{\mathcal{G}^c, \min}} \right\} + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{G}^c|)} \log(n|\mathcal{G}^c|)}{\sqrt{n}S_{\mathcal{G}^c, \min}} \right\} \\ &= O_p(\Delta_n). \end{aligned}$$

Due to  $\check{\theta}_l = \arg \min_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} |(n-m)^{-1} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})|^2$  and  $\theta_{0,l} \in B(\tilde{\theta}_l, \tilde{r})$  with probability approaching one, then

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\check{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l}) \right| \leq \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) \right|$$

with probability approaching one, which implies

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\check{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) \} \right| = O_p(\Delta_n). \quad (\text{B.8})$$

Notice that

$$\frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\check{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) \} = \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} \right\} (\check{\theta}_l - \theta_{0,l})$$

for some  $\bar{\theta}_l$  on the joint line between  $\check{\theta}_l$  and  $\theta_{0,l}$ . By Lemma 4 and Condition 4, due to  $\tau = o(1)$  and  $\tilde{r} = o(1)$ , we know

$$\min_{l \in [q]} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\theta}_{-l})}{\partial \theta_l} \right\} > \frac{1}{2}$$

with probability approaching one. Hence, by (B.8), we have  $|\check{\theta} - \theta_0|_\infty = O_p(\Delta_n)$ . We complete the proof of Proposition 2.  $\square$

#### B.4 Proof of Theorem 2

Recall  $\hat{f}_t^{(l)}(\theta) = \hat{\varphi}_l^\top \mathbf{g}_t^{(l)}(\theta)$ . By the definition of  $\hat{\varphi}_l$  given in (4.10), we have

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\theta})}{\partial \theta_{-l}} \right|_\infty \leq \tau.$$

It follows from the Taylor expansion that

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\check{\theta})}{\partial \theta_{-l}} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\theta})}{\partial \theta_{-l}} \right\} \right|_\infty \leq \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\dot{\theta})}{\partial \theta_{-l} \partial \theta^\top} \right|_\infty |\check{\theta} - \tilde{\theta}|_1,$$

where  $\dot{\theta}$  is on the joint line between  $\check{\theta}$  and  $\tilde{\theta}$ . Notice that  $\dot{\theta} \in B(\tilde{\theta}_1, \tilde{r}) \times \cdots \times B(\tilde{\theta}_q, \tilde{r})$ . Following the same arguments for deriving (B.28) in Section B.5.2 for the proof of Lemma 2, we know

$$\sup_{\theta \in B(\tilde{\theta}_1, \tilde{r}) \times \cdots \times B(\tilde{\theta}_q, \tilde{r})} \max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\theta)}{\partial \theta_{-l} \partial \theta^\top} \right|_\infty \leq C_* \max_{l \in [q]} |\hat{\varphi}_l|_1$$

for some universal constant  $C_*$  specified in (B.4), which implies

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\check{\theta})}{\partial \theta_{-l}} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\theta})}{\partial \theta_{-l}} \right\} \right|_\infty \leq C_* |\check{\theta} - \tilde{\theta}|_1 \max_{l \in [q]} |\hat{\varphi}_l|_1.$$

By Condition 4, we have  $\max_{l \in [q]} |\hat{\varphi}_l|_1 = O_p(1)$ , which implies

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\check{\theta})}{\partial \theta_{-l}} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\theta})}{\partial \theta_{-l}} \right\} \right|_\infty \leq |\check{\theta} - \tilde{\theta}|_1 \cdot O_p(1).$$

Hence, it holds that

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\check{\theta})}{\partial \theta_{-l}} \right|_\infty \leq \tau + |\check{\theta} - \tilde{\theta}|_1 \cdot O_p(1). \quad (\text{B.9})$$

Repeating the arguments for deriving Lemma 2 in Section B.5.2, we can also show

$$\begin{aligned} \max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\theta_{0,l}, \check{\theta}_{-l}) - \hat{f}_t^{(l)}(\theta_0) \} \right| \\ \leq \{ \tau + |\check{\theta} - \tilde{\theta}|_1 \cdot O_p(1) \} |\check{\theta} - \theta_0|_1 + |\check{\theta} - \theta_0|_1^2 \cdot O_p(1) \\ \leq \{ \tau + |\check{\theta} - \theta_0|_1 \cdot O_p(1) \} |\check{\theta} - \theta_0|_1 + |\check{\theta} - \theta_0|_1^2 \cdot O_p(1). \end{aligned} \quad (\text{B.10})$$

Together with Lemma 3, it holds that

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \check{\boldsymbol{\theta}}_{-l}) \right| \leq O_p(q\Delta_n).$$

Since  $\hat{\boldsymbol{\theta}}_l = \arg \min_{\boldsymbol{\theta}_l \in B(\check{\boldsymbol{\theta}}_l, \check{r})} |(n-m)^{-1} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\boldsymbol{\theta}_l, \check{\boldsymbol{\theta}}_{-l})|^2$  and  $\theta_{0,l} \in B(\check{\boldsymbol{\theta}}_l, \check{r})$  with probability approaching one, then

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l}) \right| \leq \left| \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \check{\boldsymbol{\theta}}_{-l}) \right|$$

with probability approaching one, which implies

$$\max_{l \in [q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\theta_{0,l}, \check{\boldsymbol{\theta}}_{-l}) \} \right| = O_p(q\Delta_n). \quad (\text{B.11})$$

Following the same arguments for deriving Lemma 4 in Section B.5.4 and noting (B.9), we can also have

$$\sup_{\boldsymbol{\theta}_l \in B(\check{\boldsymbol{\theta}}_l, \check{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\hat{f}_t^{(l)}(\boldsymbol{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_l} - 1 \right| \leq \tau + o_p(1) + C_* \check{r} |\hat{\boldsymbol{\varphi}}_l|_1 \quad (\text{B.12})$$

for any  $l \in [q]$ . Due to  $\tau = o(1)$ ,  $\check{r} = o(1)$  and  $\max_{l \in [q]} |\hat{\boldsymbol{\varphi}}_l|_1 = O_p(1)$ , we have

$$\max_{l \in [q]} \sup_{\boldsymbol{\theta}_l \in B(\check{\boldsymbol{\theta}}_l, \check{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\hat{f}_t^{(l)}(\boldsymbol{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_l} - 1 \right| = o_p(1). \quad (\text{B.13})$$

Due to

$$\frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\theta_{0,l}, \check{\boldsymbol{\theta}}_{-l}) \} = \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_l} \right\} (\hat{\boldsymbol{\theta}}_l - \theta_{0,l}) \quad (\text{B.14})$$

for some  $\bar{\boldsymbol{\theta}}_l$  on the joint line between  $\hat{\boldsymbol{\theta}}_l$  and  $\theta_{0,l}$ , by (B.11) and (B.12), we have  $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty = O_p(q\Delta_n)$ , which implies  $|\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}|_\infty \leq |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty + |\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty = O_p(q\Delta_n)$ . Hence,  $|\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}|_\infty \ll \check{r}$  with probability approaching one. Therefore,

$$0 = \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l}) \right\} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_l} \right\}$$

with probability approaching one. Together with (B.13), we have

$$\frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\hat{\boldsymbol{\theta}}_l, \check{\boldsymbol{\theta}}_{-l}) = 0$$

with probability approaching one.

By (B.14) and (B.10), due to  $|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 = O_p(q\Delta_n)$ ,  $|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 = O_p(\Delta_n^{1/2})$  and  $\tau \lesssim \Delta_n^{1/2}$ , then

$$\begin{aligned}\hat{\theta}_l - \theta_{0,l} &= - \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} \right\}^{-1} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\theta_{0,l}, \check{\boldsymbol{\theta}}_{-l}) \right\} \\ &= - \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} \right\}^{-1} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) \right\} \\ &\quad + O_p(q\Delta_n^{3/2}) + O_p(q^2\Delta_n^2)\end{aligned}$$

with probability approaching one. As shown in (B.30) in Section B.5.3 for the proof of Lemma 3,

$$\begin{aligned}\left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) - f_t^{(l)}(\boldsymbol{\theta}_0) \} \right| \\ = O_p \left\{ \frac{\omega_n \sqrt{(\log q) \log(qn)}}{\sqrt{n}|\mathcal{S}_l|} \right\} + O_p \left\{ \frac{\omega_n (\log q)^{1/2} \log(qn)}{\sqrt{n}|\mathcal{S}_l|} \right\} = o_p \left( \frac{1}{\sqrt{n}|\mathcal{S}_l|} \right),\end{aligned}$$

where the last step is based on Condition 4. Hence,

$$\begin{aligned}\hat{\theta}_l - \theta_{0,l} &= - \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} \right\}^{-1} \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) \right\} \\ &\quad + O_p(q\Delta_n^{3/2}) + O_p(q^2\Delta_n^2) + o_p \left( \frac{1}{\sqrt{n}|\mathcal{S}_l|} \right)\end{aligned}\tag{B.15}$$

with probability approaching one.

Write

$$\mathring{Q}_{l,t} = \frac{1}{|\mathcal{S}_l|^{1/2}} \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left\{ \boldsymbol{\varphi}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\}.$$

Then

$$\frac{1}{n-m} \sum_{t=m+1}^n \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) = \frac{1}{(n-m)|\mathcal{S}_l|^{1/2}} \sum_{t=m+1}^n \mathring{Q}_{l,t}.\tag{B.16}$$

In the sequel, we will use the martingale central limit theorem to establish the asymptotic distribution of  $(n-m)^{-1/2} \sum_{t=m+1}^n \mathring{Q}_{l,t}$ . Denote by  $\mathbb{P}_{\mathcal{F}_{t-1}}(\cdot)$  and  $\mathbb{E}_{\mathcal{F}_{t-1}}(\cdot)$ , respectively, the conditional probability measure and the conditional expectation given  $\mathcal{F}_{t-1}$  with the unknown true parameter vector  $\boldsymbol{\theta}_0$ . By Conditions 1 and 4, it holds that

$$\max_{(i,j) \in \mathcal{S}_l} \left| \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left\{ \boldsymbol{\varphi}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\} \right| \leq C_1^{-1} C_2 C_4 =: C_{**}.$$

It follows from the Bernstein inequality that

$$\mathbb{P}_{\mathcal{F}_{t-1}}(|\mathring{Q}_{l,t}| \geq x) \leq 2 \exp \left( - \frac{3|\mathcal{S}_l|^{1/2} x^2}{6|\mathcal{S}_l|^{1/2} C_{**}^2 + 2x C_{**}} \right)$$

for any  $x > 0$ , which implies, for any  $\delta > 0$ ,

$$\begin{aligned} & \sum_{t=m+1}^n \mathbb{E}_{\mathcal{F}_{t-1}} \left\{ \left( \frac{\dot{Q}_{l,t}}{\sqrt{n-m}} \right)^2 I \left( \frac{|\dot{Q}_{l,t}|}{\sqrt{n-m}} \geq \delta \right) \right\} \\ &= \frac{1}{n-m} \sum_{t=m+1}^n \mathbb{E}_{\mathcal{F}_{t-1}} \{ \dot{Q}_{l,t}^2 I(|\dot{Q}_{l,t}| \geq \delta \sqrt{n-m}) \} \rightarrow 0 \end{aligned}$$

in probability as  $n \rightarrow \infty$ . Meanwhile, by Condition 5, we also have

$$\begin{aligned} & \frac{1}{n-m} \sum_{t=m+1}^n \mathbb{E}_{\mathcal{F}_{t-1}} (\dot{Q}_{l,t}^2) \\ &= \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left\{ \boldsymbol{\varphi}_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\}^2 \rightarrow \kappa_l \end{aligned}$$

in probability as  $n \rightarrow \infty$ . By Conditions 1 and 4, we know  $\kappa_l$  is a almost surely bounded random variable. Corollary 3.1 of Hall and Heyde (1980) implies

$$\frac{1}{\sqrt{n-m}} \sum_{t=m+1}^n \dot{Q}_{l,t} \rightarrow \sqrt{\kappa_l} \cdot Z$$

in distribution as  $n \rightarrow \infty$ , where  $Z$  is a standard normally distributed random variable independent of  $\kappa_l$ . By (B.15) and (B.16), due to

$$\frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\bar{\theta}_l, \check{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} \rightarrow 1$$

in probability which is obtained in (B.13), it holds that

$$\sqrt{n|\mathcal{S}_l|}(\hat{\theta}_l - \theta_{0,l}) \rightarrow \sqrt{\kappa_l} \cdot Z$$

in distribution as  $n \rightarrow \infty$ , provided that  $\sqrt{n|\mathcal{S}_l|} \max\{q\Delta_n^{3/2}, q^2\Delta_n^2\} = o(1)$ . We complete the proof of Theorem 2.  $\square$

## B.5 Proofs of auxiliary lemmas

### B.5.1 Proof of Lemma 1

Without loss of generality, we assume  $\boldsymbol{\Theta} = [-C, C]^q$  for some constant  $C > 0$ . For given  $\epsilon > 0$  which will be specified later, we partition  $[-C, C]$  into  $K = \lceil 2C/\epsilon \rceil$  sub-intervals  $B_1, \dots, B_K$  with equal length, where the length of each  $B_k$  does not exceed  $\epsilon$ . Based on such defined  $B_1, \dots, B_K$ , we can partition  $\boldsymbol{\Theta}$  as follows:

$$\boldsymbol{\Theta} = \bigcup_{k_1=1}^K \cdots \bigcup_{k_q=1}^K B_{k_1} \times \cdots \times B_{k_q},$$

which includes  $K^q$  hyper-rectangles  $\mathcal{B}_1, \dots, \mathcal{B}_{K^q}$ . For each given  $u \in [K^q]$ , there exists  $(k_{1,u}, \dots, k_{q,u}) \in [K]^q$  such that  $\mathcal{B}_u = B_{k_{1,u}} \times \dots \times B_{k_{q,u}}$ . Let  $\boldsymbol{\theta}_u$  be the center of  $\mathcal{B}_u$ .

For each  $\boldsymbol{\theta} \in \mathcal{B}_u$ , since  $\gamma_{i,j}^{t-1}(\boldsymbol{\theta})$  only depends on  $\theta_l$  with  $l \in \mathcal{I}_{i,j}$ , it follows from the Taylor expansion that

$$\begin{aligned} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u) &= \left[ \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u) \{1 - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)\}} \frac{\partial \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\partial \boldsymbol{\theta}^\top} \right] (\boldsymbol{\theta} - \boldsymbol{\theta}_u) \\ &= \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u) \{1 - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)\}} \frac{\partial \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\partial \boldsymbol{\theta}_{\mathcal{I}_{i,j}}^\top} (\boldsymbol{\theta}_{\mathcal{I}_{i,j}} - \boldsymbol{\theta}_{u,\mathcal{I}_{i,j}}), \end{aligned}$$

where  $\tilde{\boldsymbol{\theta}}_u \in \mathcal{B}_u$  is on the joint line between  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}_u$ . Write  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$  and  $\boldsymbol{\theta}_u = (\theta_{u,1}, \dots, \theta_{u,q})^\top$ . By Conditions 1 and 2, it holds that

$$\begin{aligned} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u)| &\leq \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \sum_{l \in \mathcal{I}_{i,j}} \left\| \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u) \{1 - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)\}} \right\| \left\| \frac{\partial \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)}{\partial \theta_l} \right\| |\theta_l - \theta_{u,l}| \\ &\leq \frac{C_1^{-1} C_2 \epsilon}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \sum_{l \in \mathcal{I}_{i,j}} |X_{i,j}^t - \gamma_{i,j}^{t-1}(\tilde{\boldsymbol{\theta}}_u)| \leq s C_1^{-1} C_2 \epsilon \end{aligned}$$

for any  $\boldsymbol{\theta} \in \mathcal{B}_u$  and  $l \in [q]$ . Analogously, we also have

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_u} |\ell_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}_u)| \leq s C_1^{-1} C_2 \epsilon.$$

Therefore, by the triangle inequality, it holds that

$$\begin{aligned} \max_{l \in \mathcal{H}} \sup_{\boldsymbol{\theta} \in \Theta} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})| &= \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \sup_{\boldsymbol{\theta} \in \mathcal{B}_u} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})| \\ &\leq \max_{l \in \mathcal{H}} \max_{u \in [K^q]} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}_u)| + \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \sup_{\boldsymbol{\theta} \in \mathcal{B}_u} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u)| \\ &\quad + \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \sup_{\boldsymbol{\theta} \in \mathcal{B}_u} |\ell_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}_u)| \\ &\leq \max_{l \in \mathcal{H}} \max_{u \in [K^q]} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}_u)| + 2s C_1^{-1} C_2 \epsilon. \end{aligned} \tag{B.17}$$

For each  $l \in [q]$ ,  $u \in [K^q]$  and  $t \in [n] \setminus [m]$ , define

$$Q_{u,t}^{(l)} = \frac{1}{|\mathcal{S}_l|^{1/2}} \sum_{(i,j) \in \mathcal{S}_l} \left[ \{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\} \log \left\{ \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)} \right\} \right].$$

Then

$$\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}_u) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}_u) = \frac{1}{(n-m)|\mathcal{S}_l|^{1/2}} \sum_{t=m+1}^n Q_{u,t}^{(l)}. \tag{B.18}$$

Notice that

$$|X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)| \left| \log \left\{ \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)} \right\} \right| \leq \max_{i \neq j} \left| \log \left\{ \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)} \right\} \right|.$$

Due to

$$\log [\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)\}] < \log \left\{ \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)} \right\} < -\log [\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)\}],$$

by Condition 1, we have

$$\begin{aligned} & \max_{u \in [K^q]} \max_{t \in [n] \setminus [m]} |X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)| \left| \log \left\{ \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)} \right\} \right| \\ & \leq \max_{u \in [K^q]} \max_{t \in [n] \setminus [m]} \max_{i \neq j} |\log [\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_u)\}]| \leq \log C_1^{-1}. \end{aligned} \quad (\text{B.19})$$

Denote by  $\mathbb{P}_{\mathcal{F}_{t-1}}(\cdot)$  the conditional probability measure given  $\mathcal{F}_{t-1}$  with the unknown true parameter vector  $\boldsymbol{\theta}_0$ . It follows from the Bernstein inequality that

$$\begin{aligned} & \max_{u \in [K^q]} \max_{t \in [n] \setminus [m]} \mathbb{P}_{\mathcal{F}_{t-1}} \{ |Q_{u,t}^{(l)}| \geq x \} \\ & \leq 2 \exp \left\{ - \frac{3|\mathcal{S}_l|^{1/2} x^2}{6|\mathcal{S}_l|^{1/2} (\log C_1^{-1})^2 + 2x \log C_1^{-1}} \right\} \end{aligned} \quad (\text{B.20})$$

for any  $x > 0$ . Furthermore, due to  $\mathbb{P}\{|Q_{u,t}^{(l)}| \geq x\} = \mathbb{E}[\mathbb{P}_{\mathcal{F}_{t-1}}\{|Q_{u,t}^{(l)}| \geq x\}]$ , we also have

$$\max_{u \in [K^q]} \max_{t \in [n] \setminus [m]} \mathbb{P}\{|Q_{u,t}^{(l)}| \geq x\} \leq 2 \exp \left\{ - \frac{3|\mathcal{S}_l|^{1/2} x^2}{6|\mathcal{S}_l|^{1/2} (\log C_1^{-1})^2 + 2x \log C_1^{-1}} \right\} \quad (\text{B.21})$$

for any  $x > 0$ . Let

$$\tilde{Q}_{u,t}^{(l)} = Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| \leq M\} - \mathbb{E}_{\mathcal{F}_{t-1}}[Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| \leq M\}]$$

for some diverging  $M > 0$  specified later. Notice that  $\{\tilde{Q}_{u,t}^{(l)}\}_{t \in [n] \setminus [m]}$  is a martingale difference sequence with  $\max_{t \in [n] \setminus [m]} |\tilde{Q}_{u,t}^{(l)}| \leq 2M$ . By the Azuma's inequality (Azuma, 1967; see also Theorem 3.1 of Lesigne and Volný (2001)), we have

$$\max_{l \in [q]} \max_{u \in [K^q]} \mathbb{P} \left\{ \left| \frac{1}{n-m} \sum_{t=m+1}^n \tilde{Q}_{u,t}^{(l)} \right| \geq x \right\} \leq 2 \exp \left\{ - \frac{(n-m)x^2}{8M^2} \right\} \quad (\text{B.22})$$

for any  $x > 0$ . By (B.21),

$$\begin{aligned} & \max_{u \in [K^q]} \mathbb{P} \left[ \left| \frac{1}{n-m} \sum_{t=m+1}^n Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\} \right| > 0 \right] \leq \max_{u \in [K^q]} \mathbb{P} \left\{ \max_{t \in [n] \setminus [m]} |Q_{u,t}^{(l)}| > M \right\} \\ & \leq 2n \exp \left\{ - \frac{3|\mathcal{S}_l|^{1/2} M^2}{6|\mathcal{S}_l|^{1/2} (\log C_1^{-1})^2 + 2M \log C_1^{-1}} \right\}. \end{aligned}$$

Together with (B.22), by the Bonferroni inequality, we have

$$\begin{aligned} & \mathbb{P}\left(\max_{l \in \mathcal{H}} \max_{u \in [K^q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n [\tilde{Q}_{u,t}^{(l)} + Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\}] \right| > x\right) \\ & \leq 2|\mathcal{H}|K^q \exp\left\{-\frac{(n-m)x^2}{8M^2}\right\} \\ & \quad + 2|\mathcal{H}|K^q n \max_{l \in \mathcal{H}} \exp\left\{-\frac{3|\mathcal{S}_l|^{1/2}M^2}{6|\mathcal{S}_l|^{1/2}(\log C_1^{-1})^2 + 2M \log C_1^{-1}}\right\} \end{aligned}$$

for any  $x > 0$ . Recall  $S_{\mathcal{H},\min} = \min_{l \in \mathcal{H}} |\mathcal{S}_l|$  and  $|\mathcal{H}| \leq q$ . Selecting

$$M = C \max\left(\sqrt{q \log K}, \frac{q \log K}{\sqrt{S_{\mathcal{H},\min}}}, \sqrt{\log n}, \frac{\log n}{\sqrt{S_{\mathcal{H},\min}}}\right) \quad (\text{B.23})$$

for some sufficiently large constant  $C > 0$ , it holds that

$$\begin{aligned} & \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n [\tilde{Q}_{u,t}^{(l)} + Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\}] \right| = O_p\left(M \sqrt{\frac{q \log K}{n}}\right) \\ & = O_p\left(\frac{q \log K}{\sqrt{n}}\right) + O_p\left\{\frac{(q \log K)^{3/2}}{\sqrt{n S_{\mathcal{H},\min}}}\right\} \\ & \quad + O_p\left\{\frac{\sqrt{q(\log K)(\log n)}}{\sqrt{n}}\right\} + O_p\left\{\frac{(q \log K)^{1/2}(\log n)}{\sqrt{n S_{\mathcal{H},\min}}}\right\}. \end{aligned} \quad (\text{B.24})$$

On the other hand, by (B.20), it holds that

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_{t-1}}[|Q_{u,t}^{(l)}| I\{|Q_{u,t}^{(l)}| > M\}] = M \mathbb{P}_{\mathcal{F}_{t-1}}\{|Q_{u,t}^{(l)}| > M\} + \int_M^\infty \mathbb{P}_{\mathcal{F}_{t-1}}(|Q_{u,t}^{(l)}| > x) dx \\ & \leq 2M \exp\left\{-\frac{3|\mathcal{S}_l|^{1/2}M^2}{6|\mathcal{S}_l|^{1/2}(\log C_1^{-1})^2 + 2M \log C_1^{-1}}\right\} \\ & \quad + 2 \int_M^\infty \exp\left\{-\frac{x^2}{4(\log C_1^{-1})^2}\right\} dx + 2 \int_M^\infty \exp\left(-\frac{3|\mathcal{S}_l|^{1/2}x}{4 \log C_1^{-1}}\right) dx \\ & \leq 2M \exp\left\{-\frac{3|\mathcal{S}_l|^{1/2}M^2}{6|\mathcal{S}_l|^{1/2}(\log C_1^{-1})^2 + 2M \log C_1^{-1}}\right\} \\ & \quad + 4(\log C_1^{-1})^2 M^{-1} \exp\left\{-\frac{M^2}{4(\log C_1^{-1})^2}\right\} + \frac{8 \log C_1^{-1}}{3|\mathcal{S}_l|^{1/2}} \exp\left(-\frac{3|\mathcal{S}_l|^{1/2}M}{4 \log C_1^{-1}}\right), \end{aligned}$$

which implies

$$\begin{aligned} & \max_{u \in [K^q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \mathbb{E}_{\mathcal{F}_{t-1}}[Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\}] \right| \\ & \leq 2M \exp\left\{-\frac{3|\mathcal{S}_l|^{1/2}M^2}{6|\mathcal{S}_l|^{1/2}(\log C_1^{-1})^2 + 2M \log C_1^{-1}}\right\} \\ & \quad + 4(\log C_1^{-1})^2 \exp\left\{-\frac{M^2}{4(\log C_1^{-1})^2}\right\} + \frac{8 \log C_1^{-1}}{3|\mathcal{S}_l|^{1/2}} \exp\left(-\frac{3|\mathcal{S}_l|^{1/2}M}{4 \log C_1^{-1}}\right). \end{aligned} \quad (\text{B.25})$$

Due to  $S_{\mathcal{H},\min} = \min_{l \in \mathcal{H}} |S_l| \rightarrow \infty$  and  $M \rightarrow \infty$  satisfying (B.23), by (B.25), we have

$$\begin{aligned} \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n \mathbb{E}_{\mathcal{F}_{t-1}} [Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\}] \right| \\ \lesssim \exp(-\tilde{C}_1 M^2) + \exp(-\tilde{C}_2 M S_{\mathcal{H},\min}^{1/2}) = o(n^{-1/2}), \end{aligned}$$

where  $\tilde{C}_1 > 0$  and  $\tilde{C}_2 > 0$  are two universal constants. Due to  $Q_{u,t}^{(l)} = \tilde{Q}_{u,t}^{(l)} + Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\} - \mathbb{E}_{\mathcal{F}_{t-1}, \theta_0} [Q_{u,t}^{(l)} I\{|Q_{u,t}^{(l)}| > M\}]$ , together with (B.24), it holds that

$$\begin{aligned} \max_{l \in \mathcal{H}} \max_{u \in [K^q]} \left| \frac{1}{n-m} \sum_{t=m+1}^n Q_{u,t}^{(l)} \right| &= O_p \left( \frac{q \log K}{\sqrt{n}} \right) + O_p \left\{ \frac{(q \log K)^{3/2}}{\sqrt{n S_{\mathcal{H},\min}}} \right\} \\ &+ O_p \left\{ \frac{\sqrt{q(\log K)(\log n)}}{\sqrt{n}} \right\} + O_p \left\{ \frac{(q \log K)^{1/2}(\log n)}{\sqrt{n S_{\mathcal{H},\min}}} \right\}. \end{aligned} \quad (\text{B.26})$$

By (B.18), we have

$$\begin{aligned} \max_{l \in \mathcal{H}} \max_{u \in [K^q]} |\hat{\ell}_{n,p}^{(l)}(\theta_u) - \ell_{n,p}^{(l)}(\theta_u)| &= O_p \left( \frac{q \log K}{\sqrt{n S_{\mathcal{H},\min}}} \right) + O_p \left\{ \frac{(q \log K)^{3/2}}{\sqrt{n S_{\mathcal{H},\min}}} \right\} \\ &+ O_p \left\{ \frac{\sqrt{q(\log K)(\log n)}}{\sqrt{n S_{\mathcal{H},\min}}} \right\} + O_p \left\{ \frac{(q \log K)^{1/2}(\log n)}{\sqrt{n S_{\mathcal{H},\min}}} \right\}. \end{aligned}$$

Recall  $K \asymp \epsilon^{-1}$ . Together with (B.17), it holds that

$$\begin{aligned} \max_{l \in \mathcal{H}} \sup_{\theta \in \Theta} |\hat{\ell}_{n,p}^{(l)}(\theta) - \ell_{n,p}^{(l)}(\theta)| &\leq 2sC_1^{-1}C_2\epsilon + O_p \left( \frac{q \log K}{\sqrt{n S_{\mathcal{H},\min}}} \right) + O_p \left\{ \frac{(q \log K)^{3/2}}{\sqrt{n S_{\mathcal{H},\min}}} \right\} \\ &+ O_p \left\{ \frac{\sqrt{q(\log K)(\log n)}}{\sqrt{n S_{\mathcal{H},\min}}} \right\} + O_p \left\{ \frac{(q \log K)^{1/2}(\log n)}{\sqrt{n S_{\mathcal{H},\min}}} \right\}. \end{aligned}$$

Due to  $s \leq q$ , with selecting  $\epsilon \asymp n^{-1/2} S_{\mathcal{H},\min}^{-1/2}$ , we have

$$\max_{l \in \mathcal{H}} \sup_{\theta \in \Theta} |\hat{\ell}_{n,p}^{(l)}(\theta) - \ell_{n,p}^{(l)}(\theta)| = O_p \left\{ \frac{q \log(n S_{\mathcal{H},\min})}{\sqrt{n S_{\mathcal{H},\min}}} \right\} + O_p \left\{ \frac{q^{3/2} \log^{3/2}(n S_{\mathcal{H},\min})}{\sqrt{n S_{\mathcal{H},\min}}} \right\}.$$

We complete the proof of Lemma 1. □

### B.5.2 Proof of Lemma 2

By the Taylor expansion, we have

$$\begin{aligned} \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\theta}_{-l}) - \hat{f}_t^{(l)}(\theta_0) \} \\ = \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\theta_{0,l}, \bar{\theta}_{-l})}{\partial \theta_{-l}^\top} (\tilde{\theta}_{-l} - \theta_{0,-l}) \end{aligned}$$

$$\begin{aligned}
&= \underbrace{\frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l}^\top} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l})}_{R_{1,l}} \\
&\quad + \underbrace{\frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\theta_{0,l}, \bar{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l}^\top} - \frac{\partial \hat{f}_t^{(l)}(\theta_{0,l}, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l}^\top} \right\} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l})}_{R_{2,l}},
\end{aligned}$$

where  $\bar{\boldsymbol{\theta}}_{-l}$  is on the joint line between  $\boldsymbol{\theta}_{0,-l}$  and  $\tilde{\boldsymbol{\theta}}_{-l}$ .

For  $R_{1,l}$ , by the Taylor expansion, it holds that

$$\begin{aligned}
R_{1,l} &= \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{-l}^\top} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}) \\
&\quad + (\theta_{0,l} - \tilde{\theta}_l) \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \boldsymbol{\theta}_{-l}^\top} \right\} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}),
\end{aligned}$$

where  $\bar{\theta}_l$  is on the joint line between  $\theta_{0,l}$  and  $\tilde{\theta}_l$ . Recall  $\hat{f}_t^{(l)}(\boldsymbol{\theta}) = \hat{\boldsymbol{\varphi}}_l^\top \mathbf{g}_t^{(l)}(\boldsymbol{\theta})$ . By the definition of  $\hat{\boldsymbol{\varphi}}_l$  given in (4.10), we have

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{-l}} \right|_\infty \leq \tau,$$

which implies

$$\begin{aligned}
\left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{-l}^\top} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}) \right| &\leq \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{-l}} \right|_\infty |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 \\
&\leq \tau |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1.
\end{aligned} \tag{B.27}$$

For any  $k \in [q]$ , due to

$$\frac{\partial^2 \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \theta_k} = \hat{\boldsymbol{\varphi}}_l^\top \frac{\partial^2 \mathbf{g}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \theta_k},$$

we then have

$$\begin{aligned}
\left| \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \theta_k} \right| &\leq |\hat{\boldsymbol{\varphi}}_l|_1 \left| \frac{\partial^2 \mathbf{g}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \theta_k} \right|_\infty \\
&\leq \frac{|\hat{\boldsymbol{\varphi}}_l|_1}{|S_l|} \sum_{(i,j) \in S_l} \left| \frac{\partial^2}{\partial \theta_l \partial \theta_k} \left[ \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \right|_\infty
\end{aligned}$$

Notice that

$$\begin{aligned}
& \frac{\partial^2}{\partial \theta_{l_1} \partial \theta_{l_2}} \left[ \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3}} \right] \\
&= 2 \left[ \frac{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2 \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} + \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2 \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} \right. \\
&\quad \left. + \frac{\{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\} \{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^3 \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^3} \right] \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3}} \\
&\quad - \left[ \frac{1}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} + \frac{\{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\} \{1 - 2\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2 \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^2} \right] \\
&\quad \times \left\{ \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3}} + \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2} \partial \theta_{l_3}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1}} + \frac{\partial^2 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_3} \partial \theta_{l_1}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_2}} \right\} \\
&\quad + \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}} \frac{\partial^3 \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\partial \theta_{l_1} \partial \theta_{l_2} \partial \theta_{l_3}}.
\end{aligned}$$

By the triangle inequality and Condition 1, we know

$$\max_{k \in [q]} \left| \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\boldsymbol{\theta}}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \theta_k} \right| \leq C_* |\hat{\boldsymbol{\varphi}}_l|_1 \quad (\text{B.28})$$

for some universal constant  $C_*$  specified in (B.4), which implies

$$\begin{aligned}
& \left| (\theta_{0,l} - \tilde{\theta}_l) \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\boldsymbol{\theta}}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \boldsymbol{\theta}_{-l}^\top} \right\} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}) \right| \\
& \leq |\theta_{0,l} - \tilde{\theta}_l| \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\boldsymbol{\theta}}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l \partial \boldsymbol{\theta}_{-l}^\top} \right|_\infty |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 \\
& \leq C_* |\hat{\boldsymbol{\varphi}}_l|_1 |\theta_{0,l} - \tilde{\theta}_l| |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1.
\end{aligned}$$

Together with (B.27), it holds that  $|R_{1,l}| \leq \tau |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 + C_* |\hat{\boldsymbol{\varphi}}_l|_1 |\theta_{0,l} - \tilde{\theta}_l| |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1$ .

For  $R_{2,l}$ , by the Taylor expansion, we have

$$R_{2,l} = (\bar{\boldsymbol{\theta}}_{-l} - \tilde{\boldsymbol{\theta}}_{-l})^\top \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\theta_{0,l}, \dot{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l} \partial \boldsymbol{\theta}_{-l}^\top} \right\} (\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l})$$

for some  $\dot{\boldsymbol{\theta}}_{-l}$  on the joint line between  $\bar{\boldsymbol{\theta}}_{-l}$  and  $\tilde{\boldsymbol{\theta}}_{-l}$ , which implies

$$|R_{2,l}| \leq |\bar{\boldsymbol{\theta}}_{-l} - \tilde{\boldsymbol{\theta}}_{-l}|_1 |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\theta_{0,l}, \dot{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l} \partial \boldsymbol{\theta}_{-l}^\top} \right|_\infty.$$

Since  $\bar{\boldsymbol{\theta}}_{-l}$  is on the joint line between  $\boldsymbol{\theta}_{0,-l}$  and  $\tilde{\boldsymbol{\theta}}_{-l}$ , then  $|\bar{\boldsymbol{\theta}}_{-l} - \tilde{\boldsymbol{\theta}}_{-l}|_1 \leq |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1$ . Parallel to (B.28), we can also show

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\theta_{0,l}, \dot{\boldsymbol{\theta}}_{-l})}{\partial \boldsymbol{\theta}_{-l} \partial \boldsymbol{\theta}_{-l}^\top} \right|_\infty \leq C_* |\hat{\boldsymbol{\varphi}}_l|.$$

Hence,  $|R_{2,l}| \leq C_* |\hat{\varphi}_l|_1 |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1^2$ . Then

$$\begin{aligned}
& \left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\boldsymbol{\theta}_{0,l}, \tilde{\boldsymbol{\theta}}_{-l}) - \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) \} \right| \leq |R_{1,l}| + |R_{2,l}| \\
& \leq \tau |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 + C_* |\hat{\varphi}_l|_1 |\boldsymbol{\theta}_{0,l} - \tilde{\boldsymbol{\theta}}_{-l}| |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 + C_* |\hat{\varphi}_l|_1 |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1^2 \\
& = \tau |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 + C_* |\hat{\varphi}_l|_1 |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 |\tilde{\boldsymbol{\theta}}_{-l} - \boldsymbol{\theta}_{0,-l}|_1 \\
& \leq \tau |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1 + C_* |\hat{\varphi}_l|_1 |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_1^2.
\end{aligned}$$

We complete the proof of Lemma 2.  $\square$

### B.5.3 Proof of Lemma 3

Due to  $\hat{f}_t^{(l)}(\boldsymbol{\theta}) = \hat{\varphi}_l^\top \mathbf{g}_t^{(l)}(\boldsymbol{\theta})$  and  $f_t^{(l)}(\boldsymbol{\theta}) = \varphi_l^\top \mathbf{g}_t^{(l)}(\boldsymbol{\theta})$ , then

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) - f_t^{(l)}(\boldsymbol{\theta}_0) \} \right| \leq |\hat{\varphi}_l - \varphi_l|_1 \left| \frac{1}{n-m} \sum_{t=m+1}^n \mathbf{g}_t^{(l)}(\boldsymbol{\theta}_0) \right|_\infty.$$

Write  $(G_1^{(l)}, \dots, G_q^{(l)})^\top = (n-m)^{-1} \sum_{t=m+1}^n \mathbf{g}_t^{(l)}(\boldsymbol{\theta}_0)$ . Due to

$$\mathbf{g}_t^{(l)}(\boldsymbol{\theta}_0) = \frac{1}{|\mathcal{S}_l|} \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}},$$

we then have

$$\begin{aligned}
G_k^{(l)} &= \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \theta_k} \\
&:= \frac{1}{(n-m)|\mathcal{S}_l|^{1/2}} \sum_{t=m+1}^n \tilde{Q}_{k,t}^{(l)}
\end{aligned}$$

for any  $k \in [q]$ . Using the same arguments for deriving (B.26), we can also show

$$\max_{l \in \mathcal{H}} \max_{k \in [q]} |G_k^{(l)}| = O_p \left\{ \frac{\sqrt{(\log q) \log(qn)}}{\sqrt{n} S_{\mathcal{H}, \min}} \right\} + O_p \left\{ \frac{(\log q)^{1/2} \log(qn)}{\sqrt{n} S_{\mathcal{H}, \min}} \right\}, \quad (\text{B.29})$$

where  $S_{\mathcal{H}, \min} = \min_{l \in \mathcal{H}} |\mathcal{S}_l|$ . Together with Condition 4, it holds that

$$\begin{aligned}
& \max_{l \in \mathcal{H}} \left| \frac{1}{n-m} \sum_{t=m+1}^n \{ \hat{f}_t^{(l)}(\boldsymbol{\theta}_0) - f_t^{(l)}(\boldsymbol{\theta}_0) \} \right| \\
& = O_p \left\{ \frac{\omega_n \sqrt{(\log q) \log(qn)}}{\sqrt{n} S_{\mathcal{H}, \min}} \right\} + O_p \left\{ \frac{\omega_n (\log q)^{1/2} \log(qn)}{\sqrt{n} S_{\mathcal{H}, \min}} \right\}. \quad (\text{B.30})
\end{aligned}$$

Notice that

$$\frac{1}{n-m} \sum_{t=m+1}^n f_t^{(l)}(\boldsymbol{\theta}_0) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^n \sum_{(i,j) \in \mathcal{S}_l} \frac{X_{i,j}^t - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}} \left\{ \varphi_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\}.$$

By Conditions 1 and 4, we know

$$\max_{l \in [q]} \max_{t \in [n] \setminus [m]} \max_{(i,j) \in \mathcal{S}_l} \left| \varphi_l^\top \frac{\partial \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right| \leq C_2 C_4.$$

Using the same arguments for deriving (B.29), we can also show

$$\begin{aligned} \max_{l \in \mathcal{H}} \left| \frac{1}{n-m} \sum_{t=m+1}^n f_t^{(l)}(\boldsymbol{\theta}_0) \right| \\ = O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{H}|) \log(n|\mathcal{H}|)}}{\sqrt{n} S_{\mathcal{H}, \min}} \right\} + O_p \left\{ \frac{\sqrt{\log(1+|\mathcal{H}|) \log(n|\mathcal{H}|)}}{\sqrt{n} S_{\mathcal{H}, \min}} \right\}. \end{aligned}$$

Together with (B.30), due to  $\omega_n(\log q)^{1/2} \log(qn) = o(1)$ , we have Lemma 3.  $\square$

#### B.5.4 Proof of Lemma 4

For any  $\theta_l \in B(\tilde{\theta}_l, \tilde{r})$ , by the Taylor expansion, we have

$$\frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l} \right\} = \left\{ \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l^2} \right\} (\theta_l - \tilde{\theta}_l)$$

for some  $\bar{\theta}_l$  on the joint line between  $\theta_l$  and  $\tilde{\theta}_l$ , which implies

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l} \right\} \right| = \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial^2 \hat{f}_t^{(l)}(\bar{\theta}_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l^2} \right| |\theta_l - \tilde{\theta}_l|.$$

Due to  $\theta_l \in B(\tilde{\theta}_l, \tilde{r})$ , by (B.28), it holds that

$$\sup_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \left\{ \frac{\partial \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} - \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l} \right\} \right| \leq C_* \tilde{r} |\hat{\varphi}_l|_1.$$

Since

$$\frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l} = \frac{\hat{\varphi}_l^\top}{n-m} \sum_{t=m+1}^n \frac{\partial \mathbf{g}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l},$$

by (4.10), we know

$$\left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \theta_l} - 1 \right| \leq \tau,$$

which implies

$$\sup_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^n \frac{\partial \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l})}{\partial \theta_l} - 1 \right| \leq \tau + C_* \tilde{r} |\hat{\varphi}_l|_1.$$

We complete the proof of Lemma 4.  $\square$

## C Additional simulation results

Further to Section 5, we report more simulation results on the transitivity model.

## C.1 Stationarity and ergodicity

As stated in the last paragraph of Section 2.1, for each fixed constant  $p$ ,  $\{\mathbf{X}_t\}_{t \geq 1}$  defined by the transitivity model (3.3) is stationary and ergodic as long as  $\alpha_{i,j}^{t-1}$  and  $\beta_{i,j}^{t-1}$  are strictly between 0 and 1. Nevertheless it is a Markov chain with  $2^{p(p-1)/2}$  states. When  $p$  is a fixed constant, the ergodicity of  $\mathbf{X}_t$  (i.e. the average in time converges to the average over the state space) may take a long time to be observed; see Theorem 4.1 in Chapter 3 of Brémaud (1998). However the ergodicity of some scalar summary statistics of  $\mathbf{X}_t$  can be observed in much short time spans, as indicated in the simulation reported below.

We consider the following three network density measures at each time  $t$ :

$$D_t = \frac{\sum_{i,j: i < j} X_{i,j}^t}{p(p-1)/2}, \quad D_{1,t} = \frac{\sum_{i,j: i < j} (1 - X_{i,j}^{t-1}) X_{i,j}^t}{p(p-1)/2}, \quad D_{0,t} = \frac{\sum_{i,j: i < j} X_{i,j}^{t-1} (1 - X_{i,j}^t)}{p(p-1)/2}, \quad (\text{C.1})$$

where  $D_t$  is the network density at time  $t$ , and  $D_{1,t}$  and  $D_{0,t}$  are, respectively, the densities of newly formed edges and newly dissolved edges at time  $t$ . If  $\{\mathbf{X}_t\}_{t \geq 1}$  is stationary, all three density sequences  $\{D_t\}_{t \geq 1}$ ,  $\{D_{1,t}\}_{t \geq 2}$  and  $\{D_{0,t}\}_{t \geq 2}$  are also stationary. We also plot

$$\bar{D}_t = \frac{1}{t} \sum_{u=1}^t D_u, \quad \bar{D}_{1,t} = \frac{1}{t} \sum_{u=1}^t D_{1,u}, \quad \bar{D}_{0,t} = \frac{1}{t} \sum_{u=1}^t D_{0,u},$$

against  $t$  for  $t \geq 2$ , to see how quickly the ergodicity can be observed. These are sample means of one-dimensional network summaries. We expect that their convergences are much faster than that for the sample mean of  $p \times p$  network  $\mathbf{X}_t$  itself.

Setting  $\xi_1 = \dots = \xi_p$  and  $\eta_1 = \dots = \eta_p$ , we let  $(\xi_i, \eta_i, a, b)$  take four different sets of values:  $(0.7, 0.8, 30, 15)$ ,  $(0.6, 0.7, 20, 20)$ ,  $(0.6, 0.7, 15, 10)$  and  $(0.6, 0.7, 10, 10)$ . Figure S1 displays the time series plots of simulated  $\{D_t\}_{t=2}^{200}$ ,  $\{D_{1,t}\}_{t=2}^{200}$ ,  $\{D_{0,t}\}_{t=2}^{200}$ ,  $\{\bar{D}_t\}_{t=2}^{200}$ ,  $\{\bar{D}_{1,t}\}_{t=2}^{200}$  and  $\{\bar{D}_{0,t}\}_{t=2}^{200}$  when  $p = 50$ . As expected, all simulated series  $\{D_t\}_{t \geq 2}$ ,  $\{D_{1,t}\}_{t \geq 2}$  and  $\{D_{0,t}\}_{t \geq 2}$  exhibit patterns in line with stationarity. The convergence of their sample means is observed with the sample sizes greater than 50. In particular,  $(\xi_i, \eta_i, a, b) = (0.7, 0.8, 30, 15)$  displays the most dynamic edge changing behaviour, while  $(\xi_i, \eta_i, a, b) = (0.6, 0.7, 20, 20)$  is the least dynamic among the four settings.

## C.2 A more general model

As stated towards the end of Section 3.3, a more general transitivity model admits the form:

$$\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\xi_i \xi_j e^{a_1 U_{i,j}^{t-1}}}{1 + e^{a_1 U_{i,j}^{t-1}} + e^{b_1 V_{i,j}^{t-1}}}, \quad \beta_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\eta_i \eta_j e^{b_2 V_{i,j}^{t-1}}}{1 + e^{a_2 U_{i,j}^{t-1}} + e^{b_2 V_{i,j}^{t-1}}}, \quad (\text{C.2})$$

with  $\boldsymbol{\theta} = (a_1, b_1, a_2, b_2, \xi_1, \dots, \xi_p, \eta_1, \dots, \eta_p)^\top \in \mathbb{R}_+^{2p+4} > 0$ . Different from the transitivity model (3.3) introduced in Section 3.3, we allow  $a_1 \neq a_2$  and  $b_1 \neq b_2$  in (C.2). We adopt the same sim-

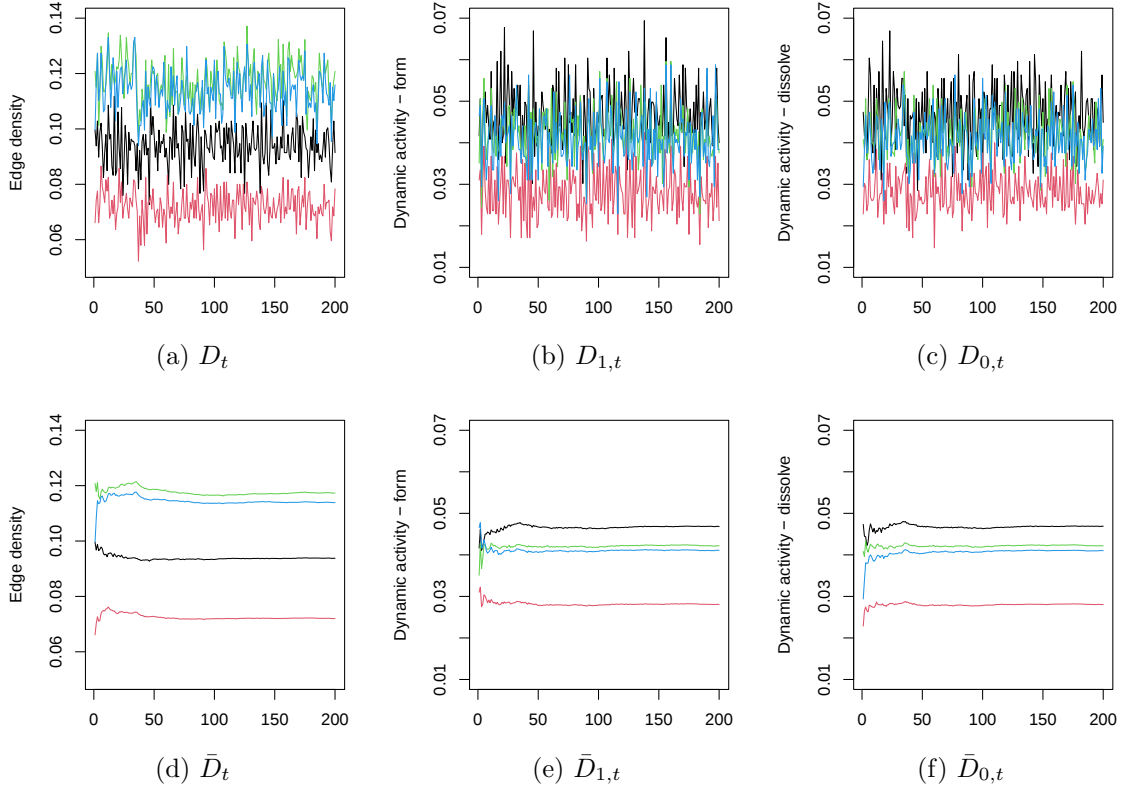


Figure S1: Time series plots of  $\{D_t\}_{t=2}^{200}$ ,  $\{D_{1,t}\}_{t=2}^{200}$ ,  $\{D_{0,t}\}_{t=2}^{200}$ ,  $\{\bar{D}_t\}_{t=2}^{200}$ ,  $\{\bar{D}_{1,t}\}_{t=2}^{200}$  and  $\{\bar{D}_{0,t}\}_{t=2}^{200}$  for the four simulated settings with  $p = 50$ . The black, red, green and blue curves correspond to the settings  $(\xi_i, \eta_i, a, b) = (0.7, 0.8, 30, 15)$ ,  $(0.6, 0.7, 20, 20)$ ,  $(0.6, 0.7, 15, 10)$  and  $(0.6, 0.7, 10, 10)$ , respectively.

Table S1: The means and STDs (in parenthesis) of rMAEs for estimating parameters in transitivity model (C.2) with 400 replications.

| $(\xi_i, a_1, b_1, \eta_i, a_2, b_2)$ | $n$ | Estimation | $\alpha_{i,j}^{t-1}(\theta)$ |               |               | $\beta_{i,j}^{t-1}(\theta)$ |                 |                 |
|---------------------------------------|-----|------------|------------------------------|---------------|---------------|-----------------------------|-----------------|-----------------|
|                                       |     |            | $\xi_i$                      | $a_1$         | $b_1$         | $\eta_i$                    | $a_2$           | $b_2$           |
| (0.7, 30, 15, 0.8, 30, 15)            | 100 | Initial    | 0.178 (0.026)                | 0.250 (0.023) | 0.171 (0.005) | 0.121 (0.019)               | 5.275 (8.602)   | 24.141 (24.917) |
|                                       |     | Improved   | 0.103 (0.030)                | 0.117 (0.013) | 0.066 (0.004) | 0.112 (0.025)               | 5.687 (8.593)   | 26.259 (24.462) |
|                                       | 200 | Initial    | 0.172 (0.022)                | 0.247 (0.021) | 0.170 (0.004) | 0.119 (0.023)               | 2.904 (6.050)   | 19.186 (22.146) |
|                                       |     | Improved   | 0.093 (0.027)                | 0.113 (0.023) | 0.064 (0.007) | 0.111 (0.025)               | 3.356 (6.067)   | 21.646 (21.935) |
|                                       | 400 | Initial    | 0.168 (0.016)                | 0.245 (0.019) | 0.170 (0.004) | 0.117 (0.021)               | 1.777 (3.273)   | 15.318 (18.832) |
|                                       |     | Improved   | 0.088 (0.019)                | 0.109 (0.010) | 0.063 (0.003) | 0.109 (0.027)               | 2.180 (3.401)   | 17.636 (18.904) |
| (0.6, 20, 20, 0.7, 20, 20)            | 100 | Initial    | 0.224 (0.004)                | 0.403 (0.023) | 0.218 (0.005) | 0.147 (0.007)               | 3.397 (8.745)   | 12.343 (19.473) |
|                                       |     | Improved   | 0.138 (0.006)                | 0.135 (0.034) | 0.081 (0.009) | 0.130 (0.009)               | 3.585 (8.585)   | 13.754 (19.465) |
|                                       | 200 | Initial    | 0.219 (0.002)                | 0.404 (0.017) | 0.219 (0.003) | 0.140 (0.008)               | 1.266 (3.404)   | 7.684 (11.048)  |
|                                       |     | Improved   | 0.123 (0.004)                | 0.124 (0.021) | 0.079 (0.007) | 0.122 (0.009)               | 1.479 (3.488)   | 9.146 (11.735)  |
|                                       | 400 | Initial    | 0.217 (0.002)                | 0.405 (0.012) | 0.219 (0.002) | 0.135 (0.008)               | 0.687 (0.316)   | 5.567 (8.182)   |
|                                       |     | Improved   | 0.114 (0.002)                | 0.117 (0.015) | 0.075 (0.005) | 0.117 (0.009)               | 0.787 (0.606)   | 6.830 (9.161)   |
| (0.6, 15, 10, 0.7, 15, 10)            | 100 | Initial    | 0.243 (0.003)                | 0.378 (0.015) | 0.266 (0.005) | 0.160 (0.019)               | 4.163 (10.012)  | 16.690 (24.226) |
|                                       |     | Improved   | 0.136 (0.004)                | 0.178 (0.047) | 0.098 (0.014) | 0.146 (0.023)               | 4.322 (10.035)  | 18.638 (24.586) |
|                                       | 200 | Initial    | 0.241 (0.002)                | 0.378 (0.011) | 0.266 (0.004) | 0.151 (0.019)               | 1.708 (4.439)   | 10.105 (17.250) |
|                                       |     | Improved   | 0.124 (0.003)                | 0.162 (0.026) | 0.090 (0.008) | 0.138 (0.022)               | 1.785 (4.572)   | 11.984 (18.233) |
|                                       | 400 | Initial    | 0.240 (0.001)                | 0.379 (0.008) | 0.266 (0.003) | 0.146 (0.018)               | 1.312 (3.015)   | 7.461 (14.351)  |
|                                       |     | Improved   | 0.117 (0.002)                | 0.159 (0.018) | 0.088 (0.005) | 0.133 (0.022)               | 1.202 (3.158)   | 9.039 (15.240)  |
| (0.6, 10, 10, 0.7, 10, 10)            | 100 | Initial    | 0.242 (0.003)                | 0.534 (0.026) | 0.279 (0.005) | 0.165 (0.020)               | 10.003 (20.550) | 25.992 (28.225) |
|                                       |     | Improved   | 0.136 (0.005)                | 0.254 (0.073) | 0.101 (0.013) | 0.152 (0.025)               | 10.210 (20.584) | 28.117 (28.413) |
|                                       | 200 | Initial    | 0.240 (0.002)                | 0.536 (0.019) | 0.279 (0.004) | 0.162 (0.022)               | 5.987 (15.780)  | 21.763 (26.073) |
|                                       |     | Improved   | 0.125 (0.003)                | 0.238 (0.045) | 0.095 (0.007) | 0.149 (0.027)               | 6.113 (15.673)  | 24.344 (26.729) |
|                                       | 400 | Initial    | 0.239 (0.001)                | 0.535 (0.013) | 0.279 (0.003) | 0.157 (0.023)               | 2.956 (6.909)   | 16.885 (21.903) |
|                                       |     | Improved   | 0.117 (0.002)                | 0.231 (0.028) | 0.094 (0.005) | 0.144 (0.028)               | 2.817 (7.196)   | 19.332 (23.305) |

ulation settings as above, i.e.  $(\xi_i, a_1, b_1, \eta_i, a_2, b_2) \in \{(0.7, 30, 15, 0.8, 30, 15), (0.6, 20, 20, 0.7, 20, 20), (0.6, 15, 10, 0.7, 15, 10), (0.6, 10, 10, 0.7, 10, 10)\}$ .

The two estimation methods are implemented in the same manner as in Section 5.2. Table S1 reports the resulting rMAEs over 400 replications with  $p = 50$  and  $n \in \{100, 200, 400\}$ . The estimation for the parameters in  $\alpha_{i,j}^{t-1}(\theta)$ , namely  $\xi_i, a_1$  and  $b_1$ , exhibits the similar patterns as in Table 1, where  $(\xi_i, a_1, b_1, \eta_i, a_2, b_2) = (0.7, 30, 15, 0.8, 30, 15)$  achieves the best estimation accuracy. In contrast, the estimation for parameters in  $\beta_{i,j}^{t-1}(\theta)$  deteriorates significantly, and especially for  $a_2$  and  $b_2$ . Note that only some components of  $\mathbf{X}_{t-1}$  with  $X_{i,j}^{t-1} = 1$ ,  $t \in [n] \setminus [m]$  and  $j \neq i$ , were used in estimating parameters  $a_2$  and  $b_2$ . For sparse networks, the total number of those data points is small. This is the intrinsic difficulty in estimating the parameters in  $\beta_{i,j}^{t-1}(\theta)$ . See also the relevant discussion at the end of Section 3.3.

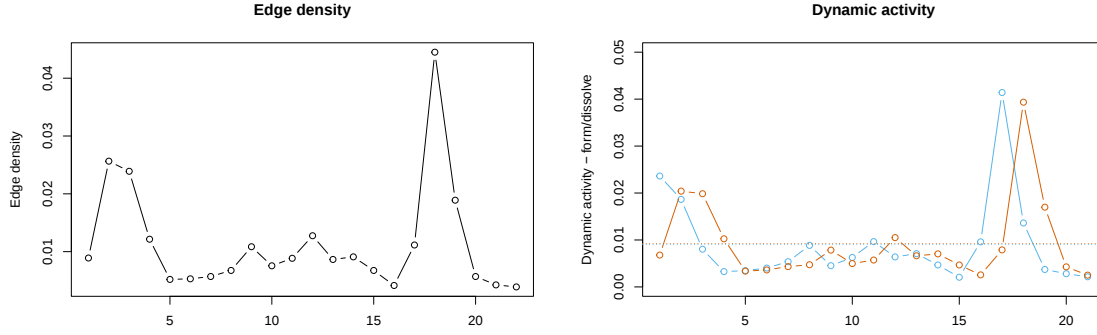


Figure S2: Evolution of edge density (left panel), percentage of grown (blue) and dissolved (orange) edges (right panel), SFHH participant networks.

## D Conference interactions

We apply our model to an additional dynamic network dataset, in this case of face-to-face interactions among attendees of an academic conference. The conference in question was the 2009 congress of the Société Française d’Hygiène Hospitalière (SFHH) (Cattuto et al., 2010; Génois and Barrat, 2018). The original data was collected automatically by RFID badges worn by the conference participants. We analyze a subset of the data corresponding to an active portion of the first day of the congress (June 4, 2009) from about 11:00AM to 6:00PM among the  $p = 200$  most active participants out of the total 403. Each of the  $n = 22$  network snapshots corresponds to a non-overlapping time window, with  $X_{i,j}^t = 1$  if participants were in close proximity at any time during the prior 20 minutes.

Similar to Section 6, we summarize some key features of this dataset. Figure S2 (left panel) shows that while there are some spikes in edge density, there is no clear increasing or decreasing pattern, so we choose to model this dataset with a single AR network model. Figures S2 and S3 show empirical evidence of temporal edge dependence, as well as transitivity effects: after accounting for edge density, edges persist at a higher rate than they grow, they more often grow for node pairs which had more common neighbours, and they more often dissolve for node pairs which had more disjoint neighbours.

Fitting our AR network model with transitivity, we estimate  $\hat{a} = 26.75$  and  $\hat{b} = 15.14$ , confirming these empirical dynamic effects of common and disjoint neighbours. We summarize the estimates of the local parameters  $\{\xi_i\}_{i=1}^{200}$  and  $\{\eta_i\}_{i=1}^{200}$  in Figure S4.

The estimates  $\{\hat{\xi}_i\}_{i=1}^{200}$  have mean 0.18 and a longer right tail, while the estimates  $\{\hat{\eta}_i\}_{i=1}^{200}$  have mean 1.16 and a longer left tail. Moreover, their scatter plot shows that there is a negative relation-

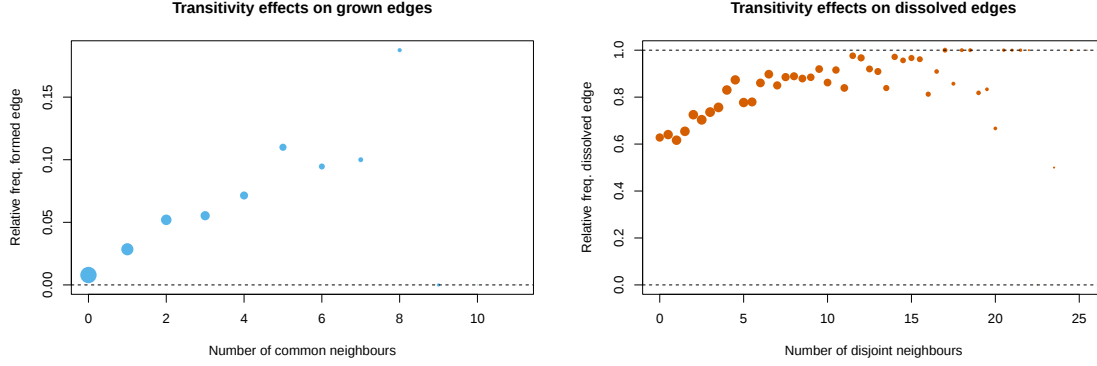


Figure S3: Left panel: relative edge frequencies,  $\{\mathcal{U}_c\}_{c \geq 0}$ . Right panel: relative non-edge frequencies,  $\{\mathcal{V}_c\}_{c \geq 0}$ . In both panels, point size is proportional to log of sample size.

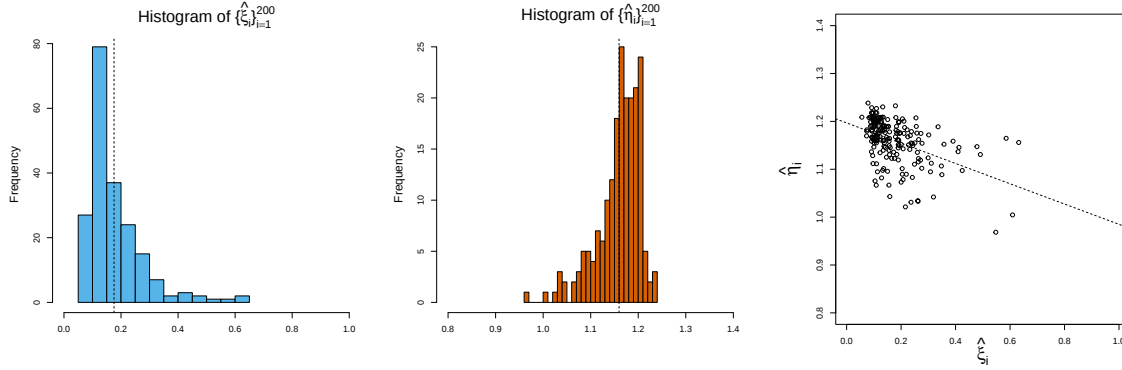


Figure S4: Histograms and scatter plot of estimates  $\{\hat{\xi}_i\}_{i=1}^{200}$  and  $\{\hat{\eta}_i\}_{i=1}^{200}$ .

ship between these estimates for a given node. All of these observations are consistent with overall degree heterogeneity of the network. The estimates  $\{\hat{\xi}_i\}_{i=1}^{200}$  have a wide range from 0.05 to 0.64, while the estimates  $\{\hat{\eta}_i\}_{i=1}^{200}$  are all between 0.97 and 1.24. This implies that conference attendees are more heterogeneous in their propensity to form new connections, than in their propensity to extend the length of existing ones.

Finally, we compare our model to the same competing models described in Section 6 in terms of AIC and BIC. These results are reported in Table S2. Our AR network model with transitivity achieves the smallest AIC, followed closely by the global AR model, which only has 2 parameters. The global AR model achieves the smallest BIC, thus in both cases the best model incorporates temporal edge dependence. Under either criterion, the two edgewise models require  $O(p^2)$  parameters and thus perform poorly, as there are relatively many nodes ( $p = 200$ ) compared to network samples ( $n = 22$ ). Although the transitivity model has  $O(p)$  parameters, it achieves the smallest AIC and the 2nd smallest BIC, suggesting that degree heterogeneity parameters and the imposed

| Model                       | AIC          | BIC          |
|-----------------------------|--------------|--------------|
| Transitivity AR model       | <b>48013</b> | 52412        |
| Global AR model             | 48037        | <b>48059</b> |
| Edgewise AR model           | 109284       | 544815       |
| Edgewise mean model         | 71205        | 288970       |
| Degree parameter mean model | 54061        | 56250        |

Table S2: AIC and BIC performance for conference interaction data

transitivity form are an effective parameterization to summarize the structure in this dataset.