Journal of the Royal Statistical Society Series B: Statistical Methodology, 2026, **00**, 1–20
https://doi.org/10.1093/jrsssb/qkag063

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Autoregressive networks with dependent edges

Jinyuan Chang[1,2], Qin Fang[3], Eric D. Kolaczyk[4], Peter W. MacDonald[5] and Qiwei Yao[6]

[1]Joint Laboratory of Data Science and Business Intelligence, Institute of Statistical Interdisciplinary Research, Southwestern University of Finance and Economics, Chengdu, China
[2]State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[3]Business School, The University of Sydney, Sydney, Australia
[4]Department of Mathematics and Statistics, McGill University, Montreal, Canada
[5]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada
[6]Department of Statistics, The London School of Economics and Political Science, London, UK

Address for correspondence: Eric D. Kolaczyk, Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St W, Montreal, Quebec, Canada H3A 0B9. Email: eric.kolaczyk@mcgill.ca

Abstract

We propose an autoregressive framework for modelling dynamic networks with dependent edges. It encompasses models that accommodate, for example, transitivity, degree heterogenenity, and other stylized features often observed in real network data. By assuming the edges of networks at each time are independent conditionally on their lagged values, the models, which exhibit a close connection with temporal exponential random graph models, facilitate both simulation and the maximum likelihood estimation (MLE) in a straightforward manner. Due to the possibly large number of parameters in the models, the natural MLEs may suffer from slow convergence rates. An improved estimator for each component parameter is proposed based on an iteration employing projection, which mitigates the impact of the other parameters. Leveraging a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the assumption of stationarity. The limiting distribution is not normal in general, although 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.

Keywords: 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. It is natural that the sequence of edges observed over time between two nodes will depend on the behavior of other edges in the network. This dependence represents known social phenomena such as reciprocity, transitivity, degree heterogeneity (popularity), which are thought to govern the evolution of real-world networks in diverse applications. On the other hand, dependent edges make the dynamic structures of network processes complex, and statistical inference challenging—not only the computation but also the accompanying theory.

In the growing research area of dynamic network modelling, work continues in three interrelated directions: model specification, computation, and theory. Our work contributes model specification, computation, and substantial and highly non-trivial supporting theory. Specifically, we develop a

Received: May 10, 2024. **Revised:** March 15, 2026. **Accepted:** March 22, 2026

class of dependent-edge autoregressive (dep. edge AR) models for network data observed in discrete time. This class is amenable to theoretical analysis, with formal asymptotic analysis for the maximum likelihood estimators. Our theoretical results in turn motivate a statistically and computationally efficient approach to parameter estimation, which we implement in an open-source software package.

We now briefly review some existing models for dynamic networks. We provide a more thorough review in Section A (online supplementary material). Under our discrete-time view, we assume that we observe time-indexed network *snapshots* at regular intervals, or can coerce relational event data into network snapshots sampled at a constant time resolution. When relational event data with continuous time stamps are available, classes like stochastic actor-oriented models (SAOM) for edge formation and dissolution events (Snijders, 2017), and relational event models (network-structured multivariate point process models) for instantaneous relational events (Butts et al., 2023; Matias et al., 2018; Perry & Wolfe, 2013) exist, and are well-developed, flexible, and computationally robust. The exponential random graph model (ERGM, Robins et al., 2007) has been extended to the dynamic setting in seminal work on (separable) temporal ERGMs ((S)TERGM, Hanneke et al., 2010; Krivitsky & Handcock, 2014). (S)TERGMs, along with a related class of logistic network regression models (LNR, Almquist & Butts, 2014) allow for flexible model specification through parametric models which specify the transition distributions of network snapshots. Previous works on AR network models instead specify transition distributions for individual edge variables, assuming a form of *edgewise dependence*: the only dependence between snapshots is between edge variables with the same (unordered) indices $\{i, j\}$ (Jiang et al., 2023, 2025). Finally, there is a large body of work which extends existing network models like the stochastic block model (SBM) and latent space model (LSM) to the dynamic setting based on a smooth or stochastically evolving sequence of latent variables (Durante & Dunson, 2016; Gallagher et al., 2021; Matias & Miele, 2017; Pensky & Zhang, 2019; Sewell & Chen, 2015; Süveges & Olhede, 2023; Zhang et al., 2024). While latent variable models do induce edge dependence both over time and within each snapshot, this dependence is implicit and cannot be configured easily to accommodate stylized features of real network data (e.g. transitivity).

Our new modelling framework directly extends the AR network models of Jiang et al. (2023, 2025). 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, and on the histories of other edge processes. This enlarged form of the transition probabilities makes the model flexible enough to accommodate stylized features such as transitivity and degree heterogeneity. The resulting network processes have dependent edges, which is radically different from previous AR network models. Nevertheless, based on a conditional independence assumption similar to Hanneke et al. (2010) and Almquist and Butts (2014), we can build up a martingale difference structure which facilitates asymptotic analysis for the maximum likelihood estimators (see Section 4). Additionally, we allow the number of parameters in the model to diverge together with the network size to accommodate, for example, heterogeneity in forming and dissolving edges modelled with node-specific parameters. To the best of our knowledge, high-dimensional asymptotic analysis of this kind has not been done for a discrete-time dynamic network model class as flexible as the one we specify in this work. We visualize the relative scope of this advancement in Table 1, which is explained in full detail in Section A (online supplementary material).

Table 1. Classification of marginal edge dependence structures in some existing static and dynamic network models (see Section A of the online supplementary material).

Between snapshot	*Full dep.*	**dep. edge AR**, LNR		(S)TERGM, SAOM
	Local dep.			
	Edgewise dep.	AR		
	Static/Full ind.	ER	LD-ERGM	ERGM
		Full ind.	*Local dep.*	*Full dep.*
			Within snapshot	

Note. Between snapshot dependence structures are contained in one another from low to high. Within snapshot dependence structures are contained in one another from left to right. Our model class, '**dep. edge AR**', can accommodate dependence structures in the entire shaded region.

Section 2 presents the general AR network framework with dependent edges. We also discuss the relationship between the proposed AR models, and TERGMs in Section 2. Section 3 contains three concrete AR models which are designed to model, respectively, degree heterogeneity, persistence, and transitivity—those are among the stylized features often observed in real network data. Section 4 presents the estimation procedures based on the maximum likelihood principle for the parameters in the AR models and the associated asymptotic theory. In so doing, we introduce the concepts of local parameters and global parameters, which need to be identified and estimated separately and may also entertain different convergence rates. An improved estimator 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). The limiting distribution of the improved estimator is derived without the assumption of stationarity. It is not normal in general, but it reduces to normal when the underlying process satisfies some mixing conditions which hold for many stationary processes. Numerical illustration with a real dynamic network data set is reported in Section 5. An online supplementary material contains simulation studies for the proposed transitivity model, additional real data results and all the technical proofs.

Notation. For any positive integer r, write $[r]=\{1,\ \ldots,\ r\}$ and $\mathbb{R}_+^r=\{(x_1,\ \ldots,\ x_r)^\top : x_i>0$ 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 $\limsup_{n\to\infty} a_n/b_n=0$, and $a_n\lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n<\infty$. For any vector $\mathbf{b}=(b_1,\ \ldots,\ b_r)^\top\in\mathbb{R}^r$, we let $\mathbf{b}_{-l}$ denote the sub-vector of $\mathbf{b}$ by removing the lth 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 $\mathbf{b}$ with the indices in $\mathcal{M}$. For any $r_1\times r_2$ real matrix $\mathbf{B}$, denote by $\mathbf{B}^\top$ its transpose. We use $\lambda_{\min}(\mathbf{B})$ to denote the smallest eigenvalue of a square matrix $\mathbf{B}$. For any countable 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 $1,\ \ldots,\ p$. 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 $X_{i,j}^t=0$ 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 $\{\mathbf{X}_s\}_{s\le t-1}$, the edges $\{X_{i,j}^t\}_{1\le i<j\le p}$ are mutually independent with

$$\begin{aligned}\alpha_{i,j}^{t-1}&\equiv\mathbb{P}(X_{i,j}^t=1\,|\,X_{i,j}^{t-1}=0,\mathbf{X}_{t-1}\setminus X_{i,j}^{t-1},\mathbf{X}_{t-k}\ \text{for}\ k\ge 2)\\&=\mathbb{P}(X_{i,j}^t=1\,|\,X_{i,j}^{t-1}=0,\mathbf{X}_{t-1}\setminus X_{i,j}^{t-1},\mathbf{X}_{t-2},\ \ldots,\mathbf{X}_{t-m}),\end{aligned}\tag{1}$$

$$\begin{aligned}\beta_{i,j}^{t-1}&\equiv\mathbb{P}(X_{i,j}^t=0\,|\,X_{i,j}^{t-1}=1,\mathbf{X}_{t-1}\setminus X_{i,j}^{t-1},\mathbf{X}_{t-k}\ \text{for}\ k\ge 2)\\&=\mathbb{P}(X_{i,j}^t=0\,|\,X_{i,j}^{t-1}=1,\mathbf{X}_{t-1}\setminus X_{i,j}^{t-1},\mathbf{X}_{t-2},\ \ldots,\mathbf{X}_{t-m}),\end{aligned}\tag{2}$$

where $m\ge 1$ is an integer.

An AR(m) network process defined above is a Markov chain with order m. Based on (1) and (2), it holds that for $x\in\{0, 1\}$,

$$\mathbb{P}(X_{i,j}^t=x\,|\,\mathbf{X}_{t-1},\ \ldots,\mathbf{X}_{t-m})=(1-\gamma_{i,j}^{t-1})^{1-x}(\gamma_{i,j}^{t-1})^x,$$

where

$$\gamma_{i,j}^{t-1}=\alpha_{i,j}^{t-1}+X_{i,j}^{t-1}(1-\alpha_{i,j}^{t-1}-\beta_{i,j}^{t-1})=\mathbb{P}(X_{i,j}^t=1\,|\,\mathbf{X}_{t-1},\ \ldots,\mathbf{X}_{t-m}),\tag{3}$$

i.e. $X_{i,j}^t\,|\,\mathbf{X}_{t-1},\ \ldots,\mathbf{X}_{t-m}\sim\text{Bernoulli}(\gamma_{i,j}^{t-1})$, $1\le i<j\le p$. Clearly edges $X_{i,j}^t$, for different (i, j), are

not necessarily independent of each other in practice. 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 that capture such dependency. Put

$$\begin{aligned}\alpha_{i,j}^{t-1} &= f_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}; \boldsymbol{\theta}_0),\\ \beta_{i,j}^{t-1} &= g_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}; \boldsymbol{\theta}_0),\end{aligned} \tag{4}$$

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

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

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. (2023). However, most available Markov models are designed for Erdös-Renyi networks with independent edges. Our setting provides a general framework to accommodate various dependency structures across different edges. Some practical network models satisfying 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 (4) are always positive and smaller than 1 for all $1 \le i < j \le p$, $\{\mathbf{X}_t\}_{t\ge 1}$ 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 $\{\mathbf{X}_t\}_{t\ge 1}$ is strictly stationary and ergodic. See Theorems 3.1, 3.3 and 4.1 in Chapter 3 of Brémaud (1998). Hence the degree heterogeneity model introduced below 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. This causes the major theoretical challenges. Note that stationarity is not an asymptotic property while ergodicity is.

2.2 Relationship to TERGMs

Our model assumes that the edges are conditionally independent given their lagged values. While general TERGMs can specify models with full within-snapshot dependence, a similar conditional independence assumption is made by Hanneke et al. (2010), to ensure non-degeneracy of the network transition distribution. Compared to this restricted class of TERGMs studied by Hanneke et al. (2010), instead of requiring the transitions to follow an exponential family distribution, we may flexibly define the probability for forming a new edge in (1), and dissolving an existing edge in (2). Those two probability functions can be in any desired forms as presented in (4), provided that their values lie between 0 and 1. See Section 3 for examples. The numerical analyses in Section 5 and Section C (online supplementary material) indicate that the proposed AR models are capable of simulating and reflecting some observed interesting dynamic network phenomena.

A TERGM (Hanneke et al., 2010; Krivitsky & Handcock, 2014) is specified by

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

where $\boldsymbol{\varsigma}$ maps the underlying unknown parameters to a natural parameter, and $\boldsymbol{\varrho}$ maps the present and past network configurations to a low-dimensional vector of sufficient statistics. Under the additional assumption of conditional edge independence, $\boldsymbol{\varrho}$ will decompose into a sum over the

edges, and any TERGM can be expressed in our AR(m) framework as

$$\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m})\}}{1 + \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m})\}},$$

$$\beta_{i,j}^{t-1}(\boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\psi}(\boldsymbol{\theta})^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m})\}}{1 + \exp\{\boldsymbol{\psi}(\boldsymbol{\theta})^\top \mathbf{v}_{i,j}(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m})\}}, \tag{5}$$

where $\mathbf{u}_{i,j}(\cdot)$ and $\mathbf{v}_{i,j}(\cdot)$ have closed-form expressions depending on the sufficient statistics of the original TERGM. Notice that the possible forms of the transition probabilities in (5) are restricted: one must take a linear combination of the natural parameter and sufficient statistics, and pass through an inverse logistic link function.

If $\boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi}(\boldsymbol{\theta})$ in (5) 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 a STERGM with conditional independent edges given the past networks (Krivitsky & Handcock, 2014). See Section B (online supplementary material) for detailed discussion and derivations regarding the relationship of the proposed dependent-edge AR models and TERGMs.

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 ξ_i is, the more likely node i will form new edges with other nodes, and the larger η_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`. Our package provides an implementation of our maximum-likelihood-based parameter estimation and inference procedure for the transitivity model (Section 3.3), as described in Section 4 and Section C (online supplementary material), and a general method-of-moments-based estimation approach introduced in Section E (online supplementary material), which can be applied to a broad class of dynamic network models, including but not limited to the three models discussed below. More details are available at https://github.com/peterwmacd/arnetworks.

3.1 Degree heterogeneity model

For any $i \neq j$, let

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

with

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

where D_i^{t-1} and D_j^{t-1} are, respectively, the (normalized) degrees of node i and node j at time $t-1$, and $D_{-i,-j}^{t-1}$ is the (normalized) average degree excluding nodes i and j at time $t-1$. We specify the transition probabilities as follows:

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

This is an AR(1) model with parameter vector $\boldsymbol{\theta} = (a_0, a_1, b_0, b_1, \xi_1, \ldots, \xi_p, \eta_1, \ldots, \eta_p)^\top$

$\in \boldsymbol{\Theta} \subset \mathbb{R}_+^{2p+4}$. This model is able to capture a 'rich get richer' effect in network evolution, whereby nodes with higher degree at time $t-1$ are more likely to form new connections at time t, and vice versa. In particular, 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 quantities.

Hanneke et al. (2010) proposed a TERGM including a density statistic (equivalent to network average degree). In (1), we explicitly specify the impact from the previous snapshot's average degree on forming a new edge and dissolving an existing edge, while the model defined in Section 2.1 of Hanneke et al. (2010) depends on the current snapshot's average degree, and does not differentiate the representations for these two types of impact. Within the STERGM 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

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

This is an AR(3) model with parameter vector $\boldsymbol{\theta} = (a, b, \xi_1, \ldots, \xi_p, \eta_1, \ldots, \eta_p)^\top \in \boldsymbol{\Theta} \subset \mathbb{R}_+^{2p+2}$. 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 (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 8×8 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 $\{X_{i,j}^t\}_{t\geq 1}$.

Persistent connectivity or non-connectivity is widely observed in, for example, brain networks, gene connections and social networks. A related TERGM including a stability statistic defined in Hanneke et al. (2010), does not consider lags of order 2 and 3, and 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:

$$\begin{aligned}\alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\xi_i\xi_j \exp\left(aU_{i,j}^{t-1}\right)}{1 + \exp\left(aU_{i,j}^{t-1}\right) + \exp\left(bV_{i,j}^{t-1}\right)},\\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= \frac{\eta_i\eta_j \exp\left(bV_{i,j}^{t-1}\right)}{1 + \exp\left(aU_{i,j}^{t-1}\right) + \exp\left(bV_{i,j}^{t-1}\right)},\end{aligned} \tag{3}$$

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

$$U_{i,j}^{t-1} = \frac{1}{p-2} \sum_{k:\, k \neq i,j} X_{i,k}^{t-1} X_{j,k}^{t-1},$$
$$V_{i,j}^{t-1} = \frac{1}{p-2} \sum_{k:\, k \neq i,j} \{X_{i,k}^{t-1}(1 - X_{j,k}^{t-1}) + (1 - X_{i,k}^{t-1}) X_{j,k}^{t-1}\}. \tag{4}$$

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. Also see the TERGM including a transitivity statistic defined in 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). We do not pursue this more general form as (i) using different ξ_i and η_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 D.3 (online supplementary material).

4 Estimation

In this section, we present estimation procedures based on the maximum likelihood principle, and show they are valid if the number of nodes p and the sample size n satisfy the restriction $p \ll n^\delta$ for some constant $\delta > 0$, which allows the number of nodes p to either be 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). This causes the major theoretical challenges faced here. Based on the conditional independence in our setting, we can construct an appropriate martingale difference sequence, which facilitates the required asymptotic analysis without the stationarity assumption, regardless of whether p is fixed or diverges together with n.

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 of $\boldsymbol{\theta}_0$ 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 (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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$, where $q \geq 1$ is the total number of parameters. Let

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

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

In the degree heterogeneity model described in Section 3.1, the global parameter vector takes the form $\boldsymbol{\theta}_{\mathcal{G}} = (a_0, a_1, b_0, b_1)^\top$. For the persistence and transitivity models introduced in Sections 3.2 and 3.3, we have $\boldsymbol{\theta}_{\mathcal{G}} = (a, b)^\top$. In these three models, the local parameters are represented by $\boldsymbol{\theta}_{\mathcal{G}^c} = (\xi_1, \ldots, \xi_p, \eta_1, \ldots, \eta_p)^\top$. Recall that m denotes the order of the AR network process. We have $m = 1$ for both the degree heterogeneity model and transitivity model, and $m = 3$ for the persistence model. As we will discuss 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.

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 (1) below. An initial estimator $\tilde{\boldsymbol{\theta}}$ results from maximizing the corresponding partial log-likelihoods $\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$ defined in (5) below, 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 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. To study the convergence properties of the proposed estimation procedures, a key step is to investigate the convergence rate of $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) - \ell_{n,p}^{(l)}(\boldsymbol{\theta})|$. As standard techniques are not applicable due to the temporal dependence inherent in our dependent-edge AR network models, we construct a martingale difference structure for the network models and use such structure to address the theoretical challenge rigorously. See Lemma 2 and its proof in the online supplementary material for further details.

4.2 Identification of $\boldsymbol{\theta}_0$

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

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

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

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

where $\mathbb{E}_{\mathcal{F}_{t-1}}(\cdot)$ denotes the conditional expectation given $\mathcal{F}_{t-1}$ with the unknown true parameter vector $\boldsymbol{\theta}_0$. For any $t \in [n]\backslash[m]$ and $1 \le i < j \le p$, due to $\log x \le x - 1$ for any $x > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_{t-1}} \big\{ \log\big[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}\big]\big\} \\
&\quad - \mathbb{E}_{\mathcal{F}_{t-1}} \big\{ \log\big[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{1-X_{i,j}^t}\big]\big\} \\
&\le \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{1-X_{i,j}^t}} \right] - 1 \\
&= \frac{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)} \cdot \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0) + \frac{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})}{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)} \cdot \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\} - 1 = 0,
\end{aligned}$$

which implies $\ell_{n,p}^{(l)}(\boldsymbol{\theta}) \le \ell_{n,p}^{(l)}(\boldsymbol{\theta}_0)$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Notice that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_{t-1}} \big\{ \log\big[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}\big]\big\} \\
&\quad - \mathbb{E}_{\mathcal{F}_{t-1}} \big\{ \log\big[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{X_{i,j}^t} \{1 - \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{1-X_{i,j}^t}\big]\big\} = 0
\end{aligned}$$

if and only if

$$\frac{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t}\{1-\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}}{\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{X_{i,j}^t}\{1-\gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)\}^{1-X_{i,j}^t}} \equiv 1, \tag{2}$$

where (2) is equivalent to $\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) = \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)$. Hence, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \{\boldsymbol{\theta}_0\}$, $\ell_{n,p}^{(l)}(\boldsymbol{\theta}) = \ell_{n,p}^{(l)}(\boldsymbol{\theta}_0)$ if and only if $\gamma_{i,j}^{t-1}(\boldsymbol{\theta}) = \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)$ for any $t \in [n]\setminus[m]$ and $(i, j) \in \mathcal{S}_l$. To guarantee the identification of $\boldsymbol{\theta}_0$, we impose the following regularity conditions.

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

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

(ii) For any $1 \le i < j \le p$ 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

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

for any $k \in [3]$.

Condition 1 specifies conditions for the parameter space $\boldsymbol{\Theta}$. We assume C_1 is a universal constant for simplifying the presentation. Our proposed methods also work when C_1 converges slowly to zero as $p \to \infty$. 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 four universal constants c_1, c_2, c_3, $c_4 \in (0, 1)$ with $c_1 < c_2$ and $c_3 < c_4$ such that

$$c_1 \le \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) \le c_2 \quad \text{and} \quad 1-c_4 \le \beta_{i,j}^{t-1}(\boldsymbol{\theta}) \le 1-c_3$$

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

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

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $t \in [n]\setminus[m]$ and $1 \le i < j \le p$. 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 $\boldsymbol{\Theta}$. See Section G (online supplementary material) for further discussion on Condition 1.

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

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

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

Condition 2 requires that the dynamics of each edge process $\{X_{i,j}^t\}_{t\ge1}$ 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 degree heterogeneity 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

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

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

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 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty < 2C_3/(C_* s^3)$. As $n \to \infty$, it holds with probability approaching one that

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

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

The proof of Proposition 1 is given in Section H.1 (online 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 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $l \in [q]$,

$$\begin{aligned}\ell_{n,p}^{(l)}(\boldsymbol{\theta}_0) - \ell_{n,p}^{(l)}(\boldsymbol{\theta}) &\ge \frac{\bar{C}}{|\mathcal{S}_l|}\sum_{(i,j)\in\mathcal{S}_l}\sum_{l'\in\mathcal{I}_{i,j}}|\theta_{l'} - \theta_{0,l'}|_2^2 = \frac{\bar{C}}{|\mathcal{S}_l|}\sum_{l'=1}^{q}\sum_{(i,j)\in\mathcal{S}_l\cap\mathcal{S}_{l'}}|\theta_{l'} - \theta_{0,l'}|^2 \\ &= \bar{C}\sum_{l'\in\mathcal{G}\cup\{l\}}|\theta_{l'} - \theta_{0,l'}|^2 + \bar{C}\sum_{l'\in\mathcal{G}^{\mathrm{c}}\setminus\{l\}}\frac{|\mathcal{S}_l\cap\mathcal{S}_{l'}|\theta_{l'} - \theta_{0,l'}|^2}{|\mathcal{S}_l|}. \end{aligned} \tag{3}$$

Hence, for any $l \in [q]$, the function $\ell_{n,p}^{(l)}(\cdot)$ defined as (1) 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 $l' \in \mathcal{G}^{\mathrm{c}} \setminus \{l\}$ if $|\mathcal{S}_l \cap \mathcal{S}_{l'}| \ll |\mathcal{S}_l|$. In the degree heterogeneity model, persistence model, and transitivity model introduced in Section 3, given $l \in [q]$ and $l' \in \mathcal{G}^{\mathrm{c}} \setminus \{l\}$, we have (i) $|\mathcal{S}_l \cap \mathcal{S}_{l'}| = 1$ and $|\mathcal{S}_l| = p - 1$ if $l \in \mathcal{G}^{\mathrm{c}}$, and (ii) $|\mathcal{S}_l \cap \mathcal{S}_{l'}| = p - 1$ and $|\mathcal{S}_l| = p(p-1)/2$ if $l \in \mathcal{G}$, reaffirming that $\ell_{n,p}^{(l)}(\cdot)$ is informative for the identification of $\theta_{0,l}$ and the global parameters, yet ineffective for identifying $\theta_{0,l'}$ with $l' \in \mathcal{G}^{\mathrm{c}} \setminus \{l\}$. Therefore, for the three models introduced in Section 3, we need to employ $\ell_{n,p}^{(l)}(\cdot)$ to identify each $\theta_{0,l}$ for $l \in [q]$.

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

With available observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, since $\{\mathbf{X}_t\}_{t\ge1}$ is a Markov chain with order m, the likelihood function for $\boldsymbol{\theta}$, conditionally on $\mathbf{X}_1, \ldots, \mathbf{X}_m$, admits the form

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

where $L_{t,p}(\mathbf{X}_t \,|\, \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-m}; \boldsymbol{\theta})$ is the transition probability of $\mathbf{X}_t$ given $\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-m}$. By (3), the (normalized) log-likelihood admits the form

$$\begin{aligned}&\frac{2}{(n-m)p(p-1)}\log\mathcal{L}_{n,p}(\mathbf{X}_n, \ldots, \mathbf{X}_{m+1} \,|\, \mathbf{X}_m, \ldots, \mathbf{X}_1; \boldsymbol{\theta}) \\ &= \frac{2}{(n-m)p(p-1)}\sum_{t=m+1}^{n}\sum_{i,j:\,1\le i<j\le p}\log\left[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t}\{1-\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}\right],\end{aligned} \tag{4}$$

which is the sample version of $\ell_{n,p}^{(l)}(\boldsymbol{\theta})$ defined as (1) with $l \in \mathcal{G}$. As pointed out below (3), we should not estimate the local parameters based on this full log-likelihood. Therefore, for each $l \in [q]$, we define

$$\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) = \frac{1}{(n-m)|\mathcal{S}_l|} \sum_{t=m+1}^{n} \sum_{(i,j)\in\mathcal{S}_l} \log\left[\{\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{X_{i,j}^t}\{1-\gamma_{i,j}^{t-1}(\boldsymbol{\theta})\}^{1-X_{i,j}^t}\right], \tag{5}$$

which contains only the terms depending on θ_l on the right-hand side of (4) (with a rescaled normalized constant).

For any $l \in [q]$, Lemma 2 (online supplementary material) shows that $\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$ converges in probability to $\ell_{n,p}^{(l)}(\boldsymbol{\theta})$ defined as (1) 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

$$(\hat{\theta}_{*,1}^{(l)}, \ldots, \hat{\theta}_{*,q}^{(l)})^\top = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$$

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

$$\tilde{\boldsymbol{\theta}}_{\mathcal{G}} = (\hat{\theta}_{*,l}^{(l')})_{l\in\mathcal{G}} \quad \text{and} \quad \tilde{\boldsymbol{\theta}}_{\mathcal{G}^c} = (\hat{\theta}_{*,l}^{(l)})_{l\in\mathcal{G}^c} \tag{6}$$

for some $l' \in \mathcal{G}$. Due to $\mathcal{S}_l = \{(i,j) : 1 \le i < j \le p\}$ 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 (6) 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

$$\begin{cases} c_{n,\mathcal{G}}^2 = \dfrac{q\log(np)}{\sqrt{np}} + \dfrac{q^{3/2}\log^{3/2}(np)}{\sqrt{np^2}}, \\ c_{n,\mathcal{G}^c}^2 = \dfrac{q\log(nS_{\mathcal{G}^c,\min})}{\sqrt{nS_{\mathcal{G}^c,\min}}} + \dfrac{q^{3/2}\log^{3/2}(nS_{\mathcal{G}^c,\min})}{\sqrt{nS_{\mathcal{G}^c,\min}}}, \end{cases} \tag{7}$$

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 H.2 (online 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 degree heterogeneity 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:

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

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 $p \ll n(\log n)^{-3}$.

Remark 2 Motivated by (3), we can approximate $\tilde{\boldsymbol{\theta}}_{\mathcal{G}}$ well by solving an alternative optimization problem involving $|\mathcal{G}|$ parameters, and each $\tilde{\theta}_l$ for $l \in \mathcal{G}^c$ through a univariate optimization. See Section C.1 (online supplementary material) for a detailed discussion. Since each optimization is of finite dimension, standard numerical methods such as the Newton–Raphson or a Quasi-Newton algorithm can be efficiently applied. For each $l \in [q]$, evaluating $\hat{\ell}^{(l)}_{n,p}(\boldsymbol{\theta})$ in (5), together with its gradient and Hessian, requires $O(n|\mathcal{S}_l|)$ operations per iteration. The subsequent step, i.e. solving a small linear system in Newton–Raphson or updating the inverse-Hessian approximation in Quasi-Newton, does not introduce any higher-order computational cost. Therefore, for the three models introduced in Section 3, the per-iteration complexity is $O(np^2)$ for the global parameters and $O(np)$ for each local parameter.

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

Recall $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$. The initial estimator $\tilde{\boldsymbol{\theta}}$ specified in (6) 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 (9) 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, 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 parameters.

For $(c_{n,\mathcal{G}}, c_{n,\mathcal{G}^c})$ defined as (7), put

$$\Delta_n = \max\left\{|\mathcal{G}|c^2_{n,\mathcal{G}}, |\mathcal{G}^c|^2 c^2_{n,\mathcal{G}^c}\right\}. \tag{8}$$

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

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

Then the score function can be written as

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

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

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

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

$$\hat{\boldsymbol{\varphi}}_l = \arg\min_{\mathbf{u}\in\mathbb{R}^q} |\mathbf{u}|_1 \quad \text{s.t.} \left| \left\{ \frac{1}{n-m} \sum_{t=m+1}^{n} \frac{\partial \mathbf{g}_t^{(l)}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right\}^\top \mathbf{u} - \mathbf{e}_l \right|_\infty \le \tau. \tag{9}$$

In the above expression, $\tau > 0$ is a tuning parameter satisfying $\tau \lesssim \Delta_n^{1/2}$ with Δ_n defined as (8), $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_q)^\top$ is the initial estimator defined as (6), and $\mathbf{e}_l$ is a q-dimensional vector with the lth component being 1 and other components being 0. Then we can re-estimate $\boldsymbol{\theta}_0$ by $\check{\boldsymbol{\theta}} = (\check{\theta}_1, \ldots, \check{\theta}_q)^\top$, where

$$\check{\theta}_l = \arg\min_{\theta_l \in B(\tilde{\theta}_l, \tilde{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^{n} \hat{f}_t^{(l)}(\theta_l, \tilde{\boldsymbol{\theta}}_{-l}) \right|^2 \tag{10}$$

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 $|\boldsymbol{\varphi}_l|_1 \le C_4$ 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 $\omega_n \to 0$ 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 (6). The proof of Proposition 2 is given in Section H.3 (online supplementary material).

Proposition 2 Let the conditions of Proposition 1 and Condition 4 hold. Then $|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty = O_p(\Delta_n)$, where Δ_n is defined as (8).

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

$$\hat{\theta}_l = \arg\min_{\theta_l \in B(\check{\theta}_l, \check{r})} \left| \frac{1}{n-m} \sum_{t=m+1}^{n} \hat{f}_t^{(l)}(\theta_l, \check{\boldsymbol{\theta}}_{-l}) \right|^2 \tag{11}$$

for some $\check{r} > 0$ satisfying $q\Delta_n \ll \check{r} \ll 1$ with Δ_n defined as (8).

Remark 3 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 degree heterogeneity 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 τ proportional to $n^{-1/4} p^{5/4} \log^{3/4}(np)$ and adopting reasonably large $\tilde{r}$ and $\check{r}$. Numerical experiments in Section 5 and Section C (online supplementary material) 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.

Remark 4 In terms of computational complexity, for each $l \in [q]$, $\hat{\boldsymbol{\varphi}}_l$ in (9) is obtained by solving a well-studied and efficiently solvable linear programming (LP) problem that involves computing a Hessian matrix. Recall that Condition 2 implies

that the dynamics of each edge process $\{X_{i,j}^t\}_{t\geq 1}$ depend on a finite number of parameters. Evaluating each Hessian matrix within the LP thus requires $O(n|\mathcal{S}_l|)$ operations. Similar to Remark 2, we use the Newton–Raphson or a Quasi-Newton algorithm to obtain $\check{\theta}_l$ in (10) and $\hat{\theta}_l$ in (11) with $O(nq|\mathcal{S}_l|)$ computations incurred per iteration for $l \in [q]$. Hence, for the three models introduced in Section 3, each iteration has a computational cost of $O(np^3)$ for each global parameter and $O(np^2)$ for each local parameter. Compared with the detailed computational cost of the initial estimation for these three models as reported in Remark 2, the additional factor of p for both the global and local parameters arises from the extra projection operation required for $\hat{f}_t^{(l)}(\boldsymbol{\theta})$, which eventually leads to faster convergence rates of $\hat{\boldsymbol{\theta}}$ and substantially improved empirical performance. See Theorem 2, Remark 6, and Section C.2 (online supplementary material) for further details. Importantly, the sequence of computations in (9), (10), and (11) can be performed in parallel across l within each step to improve computational efficiency. Further discussion on scalability is provided in Section G (online supplementary material).

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

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

where $\boldsymbol{\varphi}_l$ is given in Condition 4. Under Conditions 1 and 4, we have $|\zeta_{n,l}(\boldsymbol{\theta})| \leq C_1^{-1} C_2^2 C_4^2$, 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 $\hat{\theta}_l$, we require the following condition.

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

Remark 5 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 κ_l is a random variable in general. Generally speaking, the asymptotic distribution of $\hat{\theta}_l$ is a mixture of normal distributions. See Theorem 2 below for details. However, if the long-run variance of $\{v_l^{t-1}\}_{t=m+1}^n$ satisfies the condition

$$\mathrm{Var}\left(\frac{1}{\sqrt{n-m}} \sum_{t=m+1}^{n} v_l^{t-1} \right) = o(\sqrt{n}), \tag{12}$$

κ_l is reduced to a constant

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

Then Theorem 2 implies that $\hat{\theta}_l$ is asymptotically normal distributed. When the sequence $\{v_l^t\}_{t\geq m}$ is α-mixing with the mixing coefficients attaining certain convergence rates, (12) 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 Δ_n defined as (8), it then holds that

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

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

The proof of Theorem 2 is given in Section H.4 (online supplementary material).

Remark 6 (i) Theorem 2 shows that, for the global parameter θ_l with $l \in \mathcal{G}$,

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

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 θ_l with $l \in \mathcal{G}^c$,

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

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 $|\hat{\theta}_l - \theta_{0,l}| = \mathrm{O_p}(n^{-1/2}p^{-1})$ for $l \in \mathcal{G}$ if $p \ll n^{1/23}(\log n)^{-9/23}$, and $|\hat{\theta}_l - \theta_{0,l}| = \mathrm{O_p}\{(np)^{-1/2}\}$ 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 (9). Since $\hat{\zeta}_{n,l}(\hat{\boldsymbol{\theta}}) - \zeta_{n,l}(\boldsymbol{\theta}_0) \to 0$ in probability as $n \to \infty$, by Corollary 3.2 of Hall and Heyde (1980), it holds that

$$\sqrt{\frac{n|\mathcal{S}_l|}{\hat{\zeta}_{n,l}(\hat{\boldsymbol{\theta}})}}(\hat{\theta}_l - \theta_{0,l}) \to \mathcal{N}(0, 1) \tag{13}$$

in distribution as $n \to \infty$, provided that $\mathbb{P}(\kappa_l > 0) = 1$. We can use (13) to construct the confidence interval for each θ_l.

5 Real data analysis: email interactions

We studied estimation and coverage properties in the transitivity model (3) and an extended version of that model. See Sections C and D (online supplementary material) for a summary of results. In this section, we apply the transitivity model (3) to a dynamic network dataset of email interactions in a medium-sized Polish manufacturing company, from January to September 2010 (Michalski et al., 2011). 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 $X^t_{i,j} = 1$ if participants i and j exchanged at least one email in week t. Binarizing the communications at a weekly scale removes periodic effects and irregular behaviours which are present at higher frequencies.

In Section F.1 (online supplementary material), we inspect the stationarity of the network and the effective sample size. Our preliminary plots of edge density and dynamic activity suggest a change point in the network sequence, hence we fit our model separately to the first 13 and last 26 snapshots, referred to as 'period 1' and 'period 2'. Overall, the proportion of non-edges that form in the next snapshot is about 5%, while the proportion of existing edges which persist in the next snapshot is about 55%, clear evidence of temporal edge dependence.

Basic summaries also identify empirical evidence for transitivity effects, demonstrated in Figure 1. To construct these plots, we partition the edge variables as follows: for each integer $\ell \geq 0$, define

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

where $U^{t-1}_{i,j}$ and $V^{t-1}_{i,j}$ are given in (4). The left panel of Figure 1 plots the relative frequency $|\mathcal{U}^1_\ell|/|\mathcal{U}_\ell|$ against ℓ for $\ell = 0, 1, \ldots$, showing that this frequency of grown edges tends to be higher

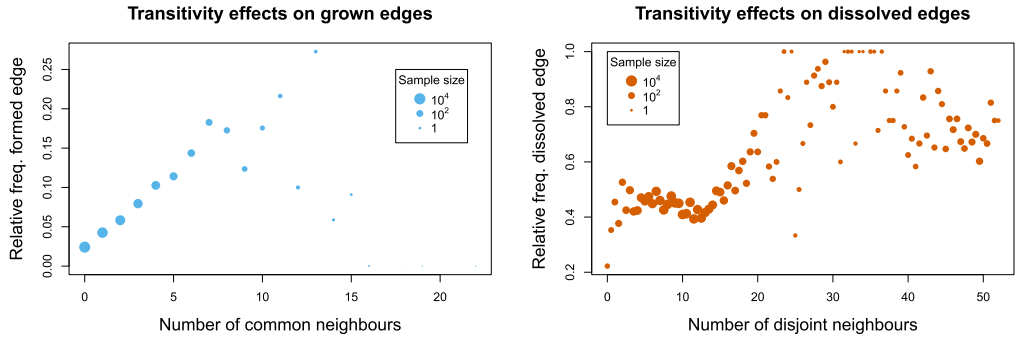

Figure 1. Left panel: the plot of relative edge frequency $|\mathcal{U}^1_\ell|/|\mathcal{U}_\ell|$ against ℓ, email interaction networks. Right panel: the plot of relative non-edge frequency $|\mathcal{V}^0_\ell|/|\mathcal{V}_\ell|$ against ℓ, email interaction networks. In both panels, point size is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

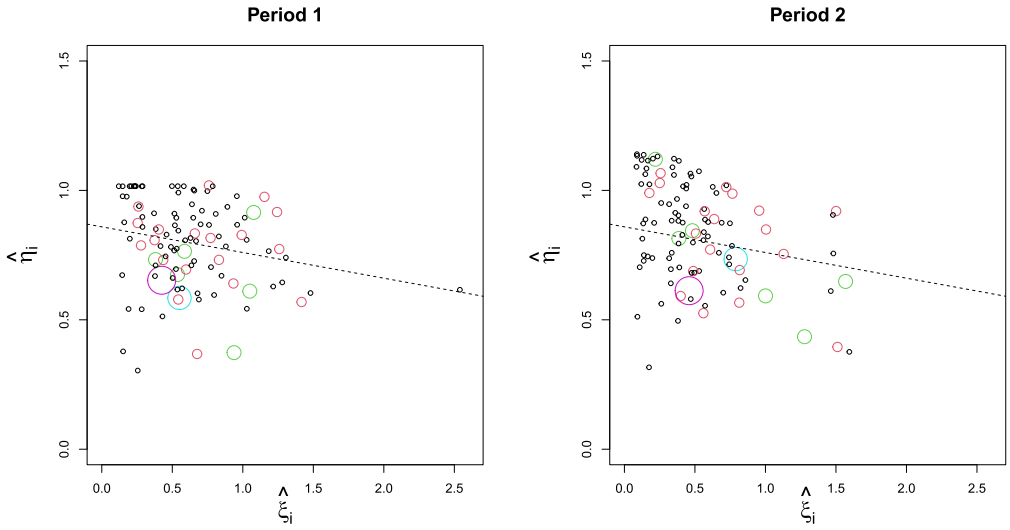

Figure 2. Scatter plots of estimates $\{\hat{\xi}_i\}_{i=1}^{106}$ and $\{\hat{\eta}_i\}_{i=1}^{106}$ for periods 1 and 2, email interaction data. Circles are sized and coloured according to their level in the company organizational tree from 1 (no direct reports) to 5 (CEO). Level 1: black, smallest; level 2: red; level 3: green; level 4: cyan; level 5: purple, largest.

for node pairs with more common neighbours in the previous snapshot. The right panel of Figure 1 analogously plots the relative frequency $|\mathcal{V}_\ell^0|/|\mathcal{V}_\ell|$ against ℓ, and shows a similar increasing relationship between disjoint neighbours and frequency of dissolved edges.

The presence of transitivity effects is confirmed by the fit of our model parameters, using the estimation algorithm described in Section C (online supplementary material). This algorithm applies the estimation procedure introduced in Section 4 and is implemented in our development R package `arnetworks`. For period 1, we estimate the global parameters $\hat{a} = 13.64$ and $\hat{b} = 9.51$, suggesting a tendency towards edge growth given more common neighbours, and edge dissolution given more distinct neighbours. 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 2. The estimates $\{\hat{\xi}_i\}_{i=1}^{106}$ have mean 0.62 and skew towards the right, implying node heterogeneity in the edge growth. Conversely, the estimates $\{\hat{\eta}_i\}_{i=1}^{106}$ have mean 0.80 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 0.72 and 0.59 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 $\hat{a} = 22.01$ and $\hat{b} = 11.31$ and summarize the estimates $\{\hat{\xi}_i\}_{i=1}^{106}$ and $\{\hat{\eta}_i\}_{i=1}^{106}$ in the right panel of Figure 2. 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 0.52 and the estimates $\{\hat{\eta}_i\}_{i=1}^{106}$ have mean 0.84, to model overall lower edge density than in period 1. The decreasing relationship between the paired parameters is stronger, and the means of $\hat{\xi}_i$ for managers and non-managers are, respectively, 0.74 and 0.45. Along with the stronger transitivity effect, we interpret that the decreased edge density in period 2 has led to a concentration of email activity among a smaller group of employees, many of whom are managers.

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. (2023), with two global switching parameters or two parameters for each edge, respectively. The 'edgewise mean model' assumes $X_{i,j}^t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(P_{i,j})$ with no temporal dependence, and estimates the edge probability $\{P_{i,j}\}_{i,j\,:\,i<j}$ for each node pair by its relative frequency of having an edge in the observed network

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

	Period 1		Period 2	
Model	**AIC**	**BIC**	**AIC**	**BIC**
Transitivity AR model	33,462	**35,412**	53,221	**55,327**
Global AR model	36,309	36,327	58,267	58,287
Edgewise AR model	42,717	144,102	55,840	165,394
Edgewise mean model	**33,248**	83,941	**47,133**	101,910
Degree parameter mean model	41,730	42,695	68,969	70,013

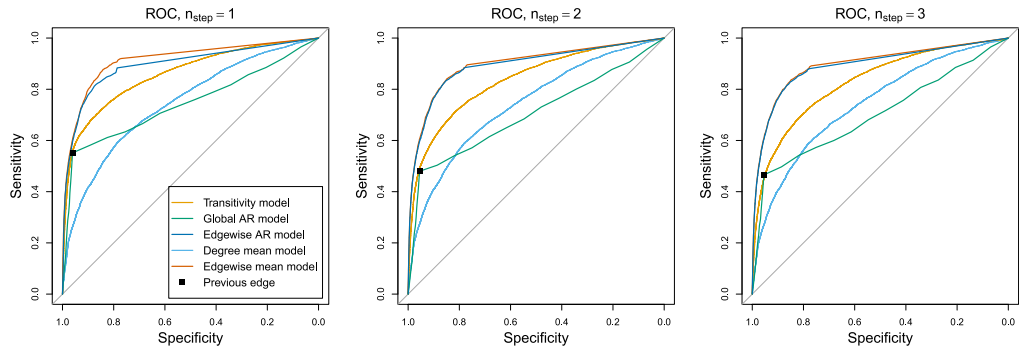

Figure 3. ROC curves for link prediction performance, email interaction data.

snapshots; and the 'degree parameter mean model' assumes $X_{i,j}^t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(v_i v_j)$ and estimates the degree parameters $\{v_i\}_{i=1}^{106}$ by fitting 1-dimensional adjacency spectral embedding (Athreya et al., 2018) to the mean adjacency matrix over the observed network snapshots. 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 likelihoods under their respective AR network formulations, 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 periods are reported in Table 2.

In both periods, our AR network model with transitivity is outperformed only by the edgewise mean model in terms of AIC, and achieves the lowest BIC, as it uses fewer parameters. This reduction of the parameter space is important for modelling 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 local 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 $n_{\text{train}} = 10, \ldots, 23$, we train these models on the first n_{train} snapshots of period 2, then forecast the state of each edge n_{step} steps forward, for $n_{\text{step}} = 1, 2, 3$. The combined results are presented in Figure 3 as receiver operating characteristic (ROC) curves. We also include a single point summarizing the performance of naively forecasting the state of each edge by its state in the n_{train}th (last available observed) network snapshot ('previous edge').

For all choices of n_{step}, the ROC curve for our AR network model with transitivity dominates or is competitive with the global AR model, degree mean model, and the naive previous edge prediction. However, the two highly parameterized edgewise models achieve better performance in terms of area under the ROC, which suggests the presence of higher order structure in this network that cannot be modeled with only two parameters per node. The edgewise mean and edgewise AR models give very similar, but not identical edge predictions; due to network sparsity the edgewise AR model has a low effective sample size to estimate the dissolution parameters, leading to slightly worse forecasting performance.

Last but not least, the analysis above is based on the combined information for each pair (i, j) in each week into a simple binary $X_{i,j}^t$, though the original data set contains more information including the precise time stamp on each email exchange. While we did not make the full use of the available data, the information accumulation over one week makes data more stationary. See Figure S3 (online supplementary material) in comparison with Figure S2 in Section F.1 (online supplementary material). Filtering out some irregular behaviours at higher frequencies also reveals more clearly the transitivity patterns in this data set. See Figure S4 (online supplementary material) in comparison with Figure 1.

Acknowledgments

The authors are grateful to the Co-Editor, an Associate Editor and two referees for their helpful suggestions.

Conflicts of interest: None declared.

Funding

Jinyuan Chang was supported in part by the National Natural Science Foundation of China (grant nos. 72495122 and 72125008). Eric D. Kolaczyk was supported in part by the U.S. National Science Foundation (grant no. SES-2120115) and the Natural Sciences and Engineering Research Council of Canada (grant nos. RGPIN-2023-03566 and DGDND-2023-03566). Peter W. MacDonald was partially supported by a Postdoctoral Fellowship from the Natural Sciences and Engineering Research Council of Canada. Qiwei Yao was partially supported by the U.K. Engineering and Physical Sciences Research Council (grants nos. EP/V007556/1 and EP/X002195/1).

Data availability

The R package `arnetworks`, which implements the proposed estimation and inference procedure, along with the data used in this paper, is publicly available at: https://github.com/peterwmacd/arnetworks.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Almquist Z. W., & Butts C. T. (2014). Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological Methodology*, *44*, 273–321. https://doi.org/10.1177/0081175013520159

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., & Sussman, D. L. (2018). Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research*, *18*, 1–92. http://jmlr.org/papers/v18/17-448.html

Brémaud P. (1998). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer.

Butts C. T., Lomi A., Snijders T. A., & Stadtfeld C. (2023). Relational event models in network science. *Network Science*, *11*, 175–183. https://doi.org/10.1017/nws.2023.9

Chang J., Chen S. X., Tang C. Y., & Wu T. (2021). High-dimensional empirical likelihood inference. *Biometrika*, *108*, 127–147. https://doi.org/10.1093/biomet/asaa051

Chang J., Shi Z., & Zhang J. (2023). Culling the herd of moments with penalized empirical likelihood. *Journal of Business & Economic Statistics*, *41*, 791–805. https://doi.org/10.1080/07350015.2022.2071903

Durante D., & Dunson D. B. (2016). Locally adaptive dynamic networks. *The Annals of Applied Statistics*, *10*, 2203–2232. https://doi.org/10.1214/16-AOAS971

Gallagher, I., Jones, A., & Rubin-Delanchy, P. (2021). Spectral embedding for dynamic networks with stability guarantees. *Advances in Neural Information Processing Systems*, *34*, 10158–10170. https://openreview.net/forum?id=9-ArDPYbUZG

Graham B. S. (2016). Homophily and transitivity in dynamic network formation. NBER Working Paper 22186, National Bureau of Economic Research.

Hall P., & Heyde C. C. (1980). *Martingale limit theory and its application*. Academic Press.

Hanneke S., Fu W., & Xing E. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, *4*, 585–605. https://doi.org/10.1214/09-EJS548

Jiang, B., Leng, C., Yan, T., Yao, Q., & Yu, X. (2025). A two-way heterogeneity model for dynamic networks. *Annals of Statistics*, *53*, 2617–2641. https://doi.org/10.1214/25-AOS2557

Jiang, B., Li, J., & Yao, Q. (2023). Autoregressive networks. *Journal of Machine Learning Research*, *24*, 1–69. http://jmlr.org/papers/v24/22-0845.html

Krivitsky P. N., & Handcock M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society, Series B*, *76*, 29–48. https://doi.org/10.1111/rssb.12014

Ludkin M., Eckley I., & Neal P. (2018). Dynamic stochastic block models: Parameter estimation and detection of changes in community structure. *Statistics and Computing*, *28*, 1201–1213. https://doi.org/10.1007/s11222-017-9788-9

Matias C., & 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. https://doi.org/10.1111/rssb.12200

Matias C., Rebafka T., & Villers F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, *105*, 665–680. https://doi.org/10.1093/biomet/asy016

Michalski R., Palus S., & Kazienko P. (2011). *Matching organizational structure and social network extracted from email communication*. In *International Conference on Business Information Systems* (pp. 197–206). Springer.

Pensky M., & Zhang T. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, *13*, 678–709. https://doi.org/10.1214/19-EJS1533

Perry P. O., & Wolfe P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society, Series B*, *75*, 821–849. https://doi.org/10.1111/rssb.12013

Robins G., Pattison P., Kalish Y., & Lusher D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, *29*, 173–191. https://doi.org/10.1016/j.socnet.2006.08.002

Sewell D. K., & Chen Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, *110*, 1646–1657. https://doi.org/10.1080/01621459.2014.988214

Snijders T. A. (2005). Models for longitudinal network data. In *Models and methods in social network analysis* (chapter 11). Cambridge University Press.

Snijders T. A. (2017). Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, *4*, 343–363. https://doi.org/10.1146/annurev-statistics-060116-054035

Süveges M., & Olhede S. C. (2023). Networks with correlated edge processes. *Journal of the Royal Statistical Society, Series A*, *186*, 441–462. https://doi.org/10.1093/jrsssa/qnad028

Yang T., Chi Y., Zhu S., Gong Y., & Jin R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine Learning*, *82*, 157–189. https://doi.org/10.1007/s10994-010-5214-7

Yudovina E., Banerjee M., & Michailidis G. (2015). *Changepoint inference for erdös-rényi random graphs*. In *Stochastic models, statistics and their applications* (pp. 197–205). Springer.

Zhang Y., Zhang J., Sun Y., & Wang J. (2024). Change point detection in dynamic networks via regularized tensor decomposition. *Journal of Computational and Graphical Statistics*, *33*, 515–524. https://doi.org/10.1080/10618600.2023.2240864