A TWO-WAY HETEROGENEITY MODEL FOR DYNAMIC NETWORKS

By Binyan Jiang[1,a], Chenlei Leng[2,c], Ting Yan[3,d], Qiwei Yao[4,e], and Xinyang Yu[1,b]

[1]*Department of Applied Mathematics, The Hong Kong Polytechnic University,* [a]*by.jiang@polyu.edu.hk;* [b]*xinyang.yu@connect.polyu.hk*

[2]*Department of Statistics, University of Warwick* , [c]*C.Leng@warwick.ac.uk*

[3]*Department of Statistics, Central China Normal University* , [d]*tingyanty@mail.ccnu.edu.cn*

[4]*Department of Statistics, London School of Economics* , [e]*q.yao@lse.ac.uk*

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. This paper proposes a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks each having p nodes, we develop a new asymptotic theory for the maximum likelihood estimation of $2p$ parameters when $np \to \infty$. We overcome the global non-convexity of the negative log-likelihood function by the virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and the analysis of some real network data sets.

1. Introduction.

Network data featuring prominent interactions between subjects arise in various areas such as biology, economics, engineering, medicine, and social sciences [24, 19]. As a rapidly growing field of active research, statistical modelling of networks aims to capture and understand the linking patterns in these data. A large part of the literature has focused on examining these patterns for canonical, static networks that are observed at a single snapshot. Due to the increasing availability of networks that are observed multiple times, models for dynamic networks evolving in time are of increasing interest now. These models typically assume, among others, that networks observed at different time are independent [25, 1], independent conditionally on some latent processes [4, 21], or drawn sequentially from an exponential random graph model conditional on the previous networks [10, 9, 20].

One of the stylized facts of real-life networks is that their nodes often have different tendencies to form ties and may evolve differently over time. The former is manifested by the fact that the so-called hub nodes have many links while the peripheral nodes have small numbers of connections in, for example, a big social network. The latter becomes evident when some individuals are more active in seeking new ties/friends than the others. In this paper, we refer to these two kinds of heterogeneity as static

MSC2020 subject classifications: 62F12, 62J99, 62R99.

Keywords and phrases: Degree heterogeneity, Dynamic networks, Maximum likelihood estimation, Uniform deviation bound.

heterogeneity and dynamic heterogeneity respectively. Also known as degree heterogeneity in the static network literature, static heterogeneity has featured prominently in several popular models widely used in practice including the stochastic block model and its degree-corrected generalization [16]. See also [14, 26, 15, 18], and the references therein. Another common and natural approach to capture the static heterogeneity is to introduce node-specific parameters, one for each node. For single networks, this is often conducted via modelling the logit of the link probability between each pair of nodes as the sum of their heterogeneity parameters. Termed as the β-model [2], this model and its generalizations have been extensively studied when a single static network is observed [36, 17, 6, 32, 3, 28, 27].

The goal of this paper is two-fold: (i) We propose a dynamic network model named the two-way heterogeneity model that captures both static heterogeneity and dynamic heterogeneity, and develop the associate inference methodology; (ii) We establish new generic asymptotic results that can be applied or extended to different models with a large number of parameters (in relation to p). We focus on the scenario that the number of nodes p goes to infinity. Our asymptotic results hold when $np \to \infty$, though n may be fixed. The main contributions of our paper can be summarized as follows.

- We propose a parsimonious formulation of the general autoregressive network model [13] to accommodate heterogeneity in both node degree and dynamic fluctuation. Our new model can also be viewed as an extension of the β-model [2] to a dynamic setting – it contains two sets of heterogeneity parameters: one controls static heterogeneity, similar to the set of parameters in the standard β-model; the other facilitates dynamic heterogeneity. Different from the general model in [13], which requires the number of network observations n to be large (i.e., $n \to \infty$), we have shown that our formulation is valid under the small n large p scenario.
- The formulation of our model gives rise to a high-dimensional non-convex loss function based on likelihood. By establishing the local convexity of the loss function in a neighborhood of the true parameters, we compute the local MLE by a standard gradient descent algorithm using a newly proposed method of moment estimator (MME) as its initial value. To our best knowledge, this is the first result in network data analysis for solving such a non-convex optimization problem with algorithmic guarantees.
- Furthermore, to characterize the local MLE, we have derived its estimation error bounds in the ℓ_2 norm and the ℓ_∞ norm when $np \to \infty$ in which $n \geq 2$ can be finite. Due to the dynamic structure of the data, the Hessian matrix of the loss function exhibits a complex structure. As a result, existing analytical approaches, such as the interior point theorem [5, 34] developed for static networks, are no longer applicable; see Section 3.1 for further elaboration. We derive a novel locally uniform deviation bound in a neighborhood of the true parameters with a diverging radius. Based on this we first establish ℓ_2 norm consistency of the MLE, which paves the way for the uniform consistency in ℓ_∞ norm.
- In establishing the locally uniform deviation bound, we have provided a general result for functions of the form $L(\boldsymbol{\theta}) = \frac{1}{p}\sum_{1\leq i\neq j\leq p} l_{i,j}\left(\theta_i, \theta_j\right) Y_{i,j}$ as defined in (4.11) below. This result explores the sparsity structure of $L(\boldsymbol{\theta})$ in the sense that most of its higher order derivatives are zero – the condition which our model satisfies, and provides a new bound that substantially extends the scope of empirical processes for the M-estimators [30] for the models with a fixed number of parameters to those with a growing number of parameters. The result here is of independent interest as it can be applied to any model with an objective function taking the form of L.

The rest of the paper is organized as follows. We introduce in Section 2 the new two-way heterogeneity model and present its properties. The estimation of its local MLE in a neighborhood of the truth and the associated theoretical properties are presented in Section 3. The development of these properties relies on new local deviation bounds which are presented in Section 4. Simulation studies and an analysis of ants interaction data are reported in Section 5. We conclude the paper in Section 6. All technical proofs are relegated to Appendix A. Further numerical results on community detection under stochastic block structures and the application to 12 dynamic protein-protein interaction networks are presented in Appendix B.

2. Two-way Heterogeneity Model. Consider a dynamic network defined on p nodes which are unchanged over time. Denote by a $p \times p$ matrix $\mathbf{X}^t = (X^t_{i,j})$ its adjacency matrix at time t, i.e. $X^t_{i,j} = 1$ indicates the existence of a connection between nodes i and j at time t, and 0 otherwise. We focus on undirected networks without self-loops, i.e., $X^t_{i,j} = X^t_{j,i}$ for all $(i,j) \in \mathcal{J} \equiv \{(i,j) : 1 \le i < j \le p\}$, and $X^t_{i,i} = 0$ for $1 \le i \le p$, though our approach can be readily extended to directed networks.

To capture the autoregressive pattern in dynamic networks, [13] proposed to model the network process via the following stationary AR(1) framework:

$$X^t_{i,j} \;=\; X^{t-1}_{i,j} I(\varepsilon^t_{i,j} = 0) \;+\; I(\varepsilon^t_{i,j} = 1), \quad t \ge 1,$$

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

$$P(\varepsilon^t_{i,j} = 1) = \alpha_{i,j}, \quad P(\varepsilon^t_{i,j} = -1) = \beta_{i,j}, \quad P(\varepsilon^t_{i,j} = 0) = 1 - \alpha_{i,j} - \beta_{i,j},$$

for some positive parameters $\alpha_{i,j}, \beta_{i,j}$. This general model opts to neglect the inherent nature of the networks and chooses to estimate each pair $(\alpha_{i,j}, \beta_{i,j})$ independently. As a result, there are $p(p-1)$ parameters and consistent model estimation requires $n \to \infty$. On the opposite, in many real applications, it has been oftentimes observed that the number of network observations n is small while the number of nodes p can be much larger than n. The general model in [13] would no longer be appropriate under this small-n-large-p scenario. To capture the inherent node heterogeneity in dynamic networks and to be able to handle small-n-large-p networks we propose the following parsimonious formulation for the general AR(1) model above, which also reduces the number of parameters from $p(p-1)$ to $2p$.

Definition 1. **Two-way Heterogeneity Model (TWHM)**. The data generating process satisfies

$$X^t_{i,j} = I(\varepsilon^t_{i,j} = 0) + X^{t-1}_{i,j} I(\varepsilon^t_{i,j} = 1), \qquad (i,j) \in \mathcal{J}, \tag{2.1}$$

where the $\varepsilon^t_{i,j}$, for $(i,j) \in \mathcal{J}$ and $t \ge 1$ are independent innovations with their distributions satisfying

$$P(\varepsilon^t_{i,j} = r) = \frac{e^{\beta_{i,r}+\beta_{j,r}}}{1 + \sum_{k=0}^{1} e^{\beta_{i,k}+\beta_{j,k}}} \quad \text{for } r = 0, 1, \quad P(\varepsilon^t_{i,j} = -1) = \frac{1}{1 + \sum_{k=0}^{1} e^{\beta_{i,k}+\beta_{j,k}}}. \tag{2.2}$$

TWHM defined above is a parsimonious formulation of the AR(1) network model [13] as it reduces the total number of parameters from $2p^2$ therein to $2p$. By Proposition 1 of [13], the matrix process $\{\mathbf{X}^t, t \ge 1\}$ is strictly stationary with

$$P(X^t_{i,j} = 1) = \frac{e^{\beta_{i,0}+\beta_{j,0}}}{1 + e^{\beta_{i,0}+\beta_{j,0}}} = 1 - P(X^t_{i,j} = 0), \tag{2.3}$$

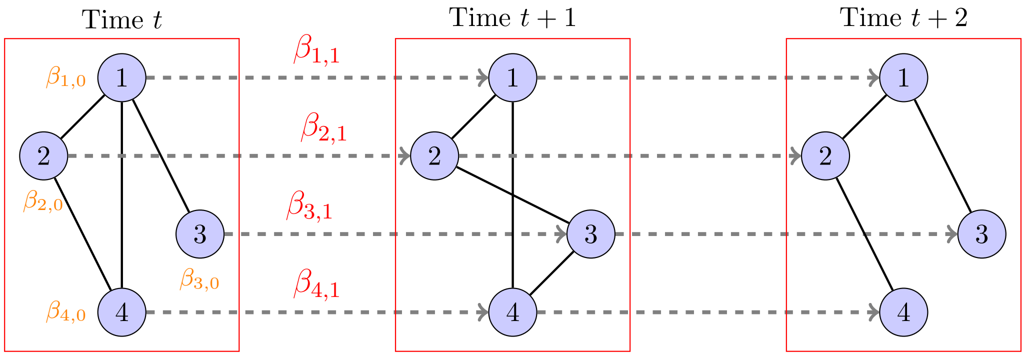

Fig 1: A schematic depiction of TWHM: $\beta_{i,0}, i=1,...,4$, are parameters to characterize the static heterogeneity of nodes, while $\beta_{i,1}$ characterize their dynamic heterogeneity.

provided that we activate the process with $\mathbf{X}^0 = (X^0_{i,j})$ also following this stationary marginal distribution.

Furthermore,

$$\mathrm{E}(X^t_{i,j}) = \frac{e^{\beta_{i,0}+\beta_{j,0}}}{1+e^{\beta_{i,0}+\beta_{j,0}}}, \qquad \mathrm{Var}(X^t_{i,j}) = \frac{e^{\beta_{i,0}+\beta_{j,0}}}{(1+e^{\beta_{i,0}+\beta_{j,0}})^2},$$

$$\rho_{i,j}(|t-s|) \equiv \mathrm{Corr}(X^t_{i,j}, X^s_{i,j}) = \left(\frac{e^{\beta_{i,1}+\beta_{j,1}}}{1+\sum_{r=0}^{1} e^{\beta_{i,r}+\beta_{j,r}}}\right)^{|t-s|}. \tag{2.4}$$

Note that the connection probabilities in (2.3) depend on $\boldsymbol{\beta}_0 = (\beta_{1,0}, \cdots, \beta_{p,0})^\top$ only, and are of the same form as the (static) β-model [2]. Hence we call $\boldsymbol{\beta}_0$ the static heterogeneity parameter. Proposition 2.1 below confirms that means and variances of node degrees in TWHM also depend on $\boldsymbol{\beta}_0$ only, and that different values of $\beta_{i,0}$ reflect the heterogeneity in the degrees of nodes.

Under TWHM, it holds that

$$P(X^t_{i,j}=1|X^{t-1}_{i,j}=0) = \frac{e^{\beta_{i,0}+\beta_{j,0}}}{1+\sum_{k=0}^{1} e^{\beta_{i,k}+\beta_{j,k}}}, P(X^t_{i,j}=0|X^{t-1}_{i,j}=1) = \frac{1}{1+\sum_{k=0}^{1} e^{\beta_{i,k}+\beta_{j,k}}}. \tag{2.5}$$

Hence the dynamic changes (over time) of network $\mathbf{X}^t$ depend on, in addition to $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1 \equiv (\beta_{1,1}, \cdots, \beta_{p,1})^\top$: the larger $\beta_{i,1}$ is, the more likely $X^t_{i,j}$ will retain the value of $X^{t-1}_{i,j}$ for all j. Thus we call $\boldsymbol{\beta}_1$ the dynamic heterogeneity parameter, as its components reflect the different dynamic behaviours of the p nodes. A schematic description of the model can be seen from Figure 1 where three snapshots of a dynamic network with four nodes are depicted.

From now on, let $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}}$ denote the stationary TWHM with parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top)^\top$, and $d^t_i = \sum_{j=1}^p X^t_{i,j}$ be the degree of node i at time t. The proposition below lists some properties of the node degrees.

Proposition 2.1. Let $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}}$. Then $\{(d^t_1, \ldots, d^t_p), t=0,1,2,\cdots\}$ is a strictly stationary process. Furthermore for any $1 \le i < j \le p$ and $t, s \ge 0$,

$$\mathrm{E}(d^t_i) = \sum_{k=1,\, k\ne i}^{p} \frac{e^{\beta_{i,0}+\beta_{k,0}}}{1+e^{\beta_{i,0}+\beta_{k,0}}}, \qquad \mathrm{Var}(d^t_i) = \sum_{k=1,\, k\ne i}^{p} \frac{e^{\beta_{i,0}+\beta_{k,0}}}{(1+e^{\beta_{i,0}+\beta_{k,0}})^2},$$

$$\begin{aligned}\rho^d_{i,j}(|t-s|) &\equiv \mathrm{Corr}(d_i^t, d_j^s)\\ &= \begin{cases} C_{i,\rho}\sum_{k=1,\,k\neq i}^{p}\left(\frac{e^{\beta_{i,1}+\beta_{k,1}}}{1+\sum_{r=0}^{1}e^{\beta_{i,r}+\beta_{k,r}}}\right)^{|t-s|}\frac{e^{\beta_{i,0}+\beta_{k,0}}}{(1+e^{\beta_{i,0}+\beta_{k,0}})^2} & \text{if } i=j,\\ 0 & \text{if } i\neq j,\end{cases}\end{aligned}$$

where $C_{i,\rho}=\left(\sum_{k=1,\,k\neq i}^{p}\frac{e^{\beta_{i,0}+\beta_{k,0}}}{(1+e^{\beta_{i,0}+\beta_{k,0}})^2}\right)^{-1}$.

Proposition 2.1 implies that when there exist constants β_0, β_1 such that $\beta_{i,0}\approx\beta_0$ and $\beta_{i,1}\approx\beta_1$ for all i, the degree sequence $\{d_i^t, t=1,\dots,n\}$ is approximately AR(1).

3. Parameter Estimation. We introduce some notation first. Denote by $\mathbf{I}_p$ the $p\times p$ identity matrix. For any $s\in\mathbb{R}$, $\mathbf{s}_p$ denotes the $p\times 1$ vector with all its elements equal to s. For $\mathbf{a}=(a_1,\dots,a_p)^\top\in\mathbb{R}^p$ and $\mathbf{A}=(A_{i,j})\in\mathbb{R}^{p\times p}$, let $\|\mathbf{a}\|_q=(a_i^q)^{1/q}$ for any $q\geq 1$, $\|\mathbf{a}\|_\infty=\max_i|a_i|$, and $\|\mathbf{A}\|_\infty=\max_i\sum_{j=1}^p|A_{i,j}|$. Furthermore, let $\|\mathbf{A}\|_2$ denote the spectral norm of $\mathbf{A}$ which equals its largest eigenvalue. For a random matrix $\mathbf{W}\in\mathbb{R}^{p\times p}$ with $\mathrm{E}(\mathbf{W})=\mathbf{0}$, define its matrix variance as $\mathrm{Var}(\mathbf{W})=\max\{\|\mathrm{E}(\mathbf{W}\mathbf{W}^\top)\|_2, \|\mathrm{E}(\mathbf{W}^\top\mathbf{W})\|_2\}$. The notation $x\lesssim y$ means that there exists a constant $c_1>0$ such that $|x|\leq c_1|y|$, while notation $x\gtrsim y$ means there exists a constant $c_2>0$ such that $|x|\geq c_2|y|$. Denote by $\mathbf{B}_\infty(\mathbf{x},r)=\{\mathbf{y}:\|\mathbf{y}-\mathbf{x}\|_\infty\leq r\}$ the ball centred at $\mathbf{x}$ with ℓ_∞ radius r. Let $c, c_0, c_1,\dots, C, C_0, C_1,\dots$ denote some generic constants that may be different in different places.

Let $\boldsymbol{\theta}^*=(\boldsymbol{\beta}_0^{*\top},\boldsymbol{\beta}_1^{*\top})^\top=(\beta_{1,0}^*,\cdots,\beta_{p,0}^*,\beta_{1,1}^*,\cdots,\beta_{p,1}^*)^\top$ be the true unknown parameters. We assume:

(A1) There exists a constant K such that for any $i=1,2,\cdots,p$ the true parameters satisfy $\beta_{i,1}^*-\max(\beta_{i,0}^*,0)<K$.

Condition (A1) above ensures that autocorrelation functions (ACFs) $\rho_{i,j}$ in (2.4) are bounded away from 1 for any $(i,j)\in\mathcal{J}$. We note in particular that both $\beta_{i,1}^*$ and $\beta_{i,0}^*$ are allowed to depend on p such that sparse networks are included in our exploration. In practice, the $\beta_{i,0}^*$ which captures the sparsity of the stationary network is usually very small for large networks, and (A1) would hold when $\beta_{i,1}^*$ is bounded from above.

3.1. *Maximum likelihood estimation.* With the available observations $\mathbf{X}^0,\cdots,\mathbf{X}^n$, the log-likelihood function conditionally on $\mathbf{X}^0$ is of the form $L(\boldsymbol{\theta};\mathbf{X}^n,\cdots,\mathbf{X}^1|\mathbf{X}^0)=\prod_{t=1}^n L(\boldsymbol{\theta};\mathbf{X}^t|\mathbf{X}^{t-1})$. Note $\{X_{i,j}^t\}$ for different $(i,j)\in\mathcal{J}$ are independent with each other. By (2.5), a (normalized) negative log-likelihood admits the following form:

$$\begin{aligned}(3.6)\quad l(\boldsymbol{\theta}) &= -\frac{1}{np}L(\boldsymbol{\theta};\mathbf{X}^n,\mathbf{X}^{n-1},\cdots,\mathbf{X}^1|\mathbf{X}^0)\\ &= -\frac{1}{p}\sum_{1\leq i<j\leq p}\log\left(1+e^{\beta_{i,0}+\beta_{j,0}}+e^{\beta_{i,1}+\beta_{j,1}}\right)+\frac{1}{np}\sum_{1\leq i<j\leq p}\Bigg\{(\beta_{i,0}+\beta_{j,0})\sum_{t=1}^n X_{i,j}^t\\ &\quad+\log\left(1+e^{\beta_{i,1}+\beta_{j,1}}\right)\sum_{t=1}^n\left(1-X_{i,j}^t\right)\left(1-X_{i,j}^{t-1}\right)\\ &\quad+\log\left(1+e^{\beta_{i,1}+\beta_{j,1}-\beta_{i,0}-\beta_{j,0}}\right)\sum_{t=1}^n X_{i,j}^t X_{i,j}^{t-1}\Bigg\}.\end{aligned}$$

For brevity, write

$$a_{i,j} = \sum_{t=1}^{n} X_{i,j}^{t}, \quad b_{i,j} = \sum_{t=1}^{n} X_{i,j}^{t} X_{i,j}^{t-1}, \quad d_{i,j} = \sum_{t=1}^{n} \Big(1 - X_{i,j}^{t}\Big)\Big(1 - X_{i,j}^{t-1}\Big). \tag{3.7}$$

Then the Hessian matrix of $l(\boldsymbol{\theta})$ is of the form

$$\mathbf{V}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{bmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0^\top} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_1^\top} \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_0^\top} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^\top} \end{bmatrix} := \begin{bmatrix} \mathbf{V}_1(\boldsymbol{\theta}) & \mathbf{V}_2(\boldsymbol{\theta}) \\ \mathbf{V}_2(\boldsymbol{\theta}) & \mathbf{V}_3(\boldsymbol{\theta}) \end{bmatrix},$$

where for $i \neq j$,

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_{i,0} \partial \beta_{j,0}} &= \frac{1}{p} \frac{e^{\beta_{i,0}+\beta_{j,0}}(1 + e^{\beta_{i,1}+\beta_{j,1}})}{(1 + e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2} - \frac{1}{np} b_{i,j} \frac{e^{\beta_{i,0}+\beta_{j,0}+\beta_{i,1}+\beta_{j,1}}}{(e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2}, \\
\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_{i,0} \partial \beta_{j,1}} &= -\frac{1}{p} \frac{e^{\beta_{i,0}+\beta_{j,0}+\beta_{i,1}+\beta_{j,1}}}{(1 + e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2} + \frac{1}{np} b_{i,j} \frac{e^{\beta_{i,0}+\beta_{j,0}+\beta_{i,1}+\beta_{j,1}}}{(e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2}, \\
\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_{i,1} \partial \beta_{j,1}} &= \frac{1}{p} \frac{e^{\beta_{i,1}+\beta_{j,1}}(1 + e^{\beta_{i,0}+\beta_{j,0}})}{(1 + e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2} - \frac{1}{np} d_{i,j} \frac{e^{\beta_{i,1}+\beta_{j,1}}}{(1 + e^{\beta_{i,1}+\beta_{j,1}})^2} \\
&\quad - \frac{1}{np} b_{i,j} \frac{e^{\beta_{i,0}+\beta_{j,0}+\beta_{i,1}+\beta_{j,1}}}{(e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}})^2}.
\end{aligned}$$

Note that matrix $\mathbf{V}_2(\boldsymbol{\theta})$ is symmetric. Furthermore, the three matrices $\mathbf{V}_1(\boldsymbol{\theta}), \mathbf{V}_2(\boldsymbol{\theta})$ and $\mathbf{V}_3(\boldsymbol{\theta})$ are diagonally balanced [11] in the sense that their diagonal elements are the sums of their respective rows, namely,

$$(\mathbf{V}_k(\boldsymbol{\theta}))_{i,i} = \sum_{j=1,\, j\neq i}^{p} (\mathbf{V}_k(\boldsymbol{\theta}))_{i,j}, \quad k = 1, 2, 3.$$

Unfortunately the Hessian matrix $\mathbf{V}(\boldsymbol{\theta})$ is not uniformly positive-definite. Hence $l(\boldsymbol{\theta})$ is not convex; see Section 5.1 for an example. Therefore, finding the global MLE by minimizing $l(\boldsymbol{\theta})$ would be infeasible, especially given the large dimensionality of $\boldsymbol{\theta}$. To overcome the obstacle, we propose the following roadmap to search for the local MLE over a neighbourhood of the true parameter values $\boldsymbol{\theta}^*$.

(1) First we show that $l(\boldsymbol{\theta})$ is locally convex in a neighbourhood of $\boldsymbol{\theta}^*$ (see Theorem 3.1 below). Towards this end, we first prove that $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ is positive definite in a neighborhood of $\boldsymbol{\theta}^*$. Leveraging on some newly proved concentration results, we show that $\mathbf{V}(\boldsymbol{\theta})$ converges to $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ uniformly over the neighborhood.

(2) Denote by $\widehat{\boldsymbol{\theta}}$ the local MLE in the neighbourhood identified above. We derive the bounds for $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ respectively in both ℓ_2 and ℓ_∞ norms (see Theorems 3.2 and 3.3 below). The ℓ_2 convergence is established by providing a uniform upper bound for the local deviation between $l(\boldsymbol{\theta}) - \mathrm{E}(l(\boldsymbol{\theta}))$ and $l(\boldsymbol{\theta}^*) - \mathrm{E}(l(\boldsymbol{\theta}^*))$ (see Corollary 4.1 in Section 4). The ℓ_∞ convergence of $\widehat{\boldsymbol{\theta}}$ is established by further exploiting the special structure of the objective function.

(3) We propose a new method of moment estimator (MME) which is proved to lie asymptotically in the neighbourhood specified in (1) above. With this MME as the initial value, the local MLE $\widehat{\boldsymbol{\theta}}$ can be simply obtained via a gradient decent algorithm.

The main technical challenges in the roadmap above can be summarized as follows.

Firstly, to establish the upper bounds as stated in (2) above, we need to evaluate the uniform local deviations of the loss function. While the theoretical framework for deriving similar deviations of M-estimators has been well established in, for example, [30, 29], classical techniques in empirical process for establishing uniform laws [31] are not applicable because the number of parameters in TWHM diverges.

Secondly, for the classical β-model, proving the existence and convergence of its MLE relies strongly on the interior point theorem [5]. In particular, this theorem is applicable only because the Hessian matrix of the β-model admits a nice structure, i.e. it is diagonally dominant and all its elements are positive depending on the parameters only [2, 36, 34, 7]. However the Hessian matrix of $l(\boldsymbol{\theta})$ for TWHM depends on random variables $X_{i,j}^t$'s in addition to the parameters, making it impossible to verify if the score function is uniformly Fréchet differentiable or not, a key assumption required by the interior point theorem.

Lastly, the higher order derivatives of $l(\boldsymbol{\theta})$ may diverge as the order increases. To see this, notice that for any integer k, the k-th order derivatives of $l(\boldsymbol{\theta})$ is closely related to the $(k-1)$-th order derivatives of the Sigmoid function $S(x) = \frac{1}{1+e^{-x}}$ in that $\frac{\partial^k S(x)}{\partial x^k} = \frac{\sum_{m=0}^{k-2} -A(k-1,m)(-e^x)^{m+1}}{(1+e^x)^k}$, where $A\left(k-1,m\right)$ is the Eulerian number [23]. Some of the coefficients $A\left(k-1,m\right)$ can diverge very quickly as k increases. Thus, loosely speaking, $l(\boldsymbol{\theta})$ is not smooth. This non-smoothness and the need to deal with a growing number of parameters make various local approximations based on Taylor expansions highly non-trivial; noting that the consistency of MLEs in many finite-dimensional models is often established via these approximations.

In our proofs, we have made great use of the special sparse structure of the loss function in the form (4.11) below. This sparsity structure stems from the fact that most of its higher order derivatives are zero. Based on the uniform local deviation bound obtained in Section 4, we have established an upper bound for the error of the local MLE under the l_2 norm. Utilizing the structure of the marginalized functions of the loss we have further established an upper bound for the estimation error under the l_∞ norm thanks to an iterative procedure stated in Section 3.3.

3.2. *Existence of the local MLE.* To establish the convexity of $l(\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}^*$, we first show that such a local convexity holds for $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$.

Proposition 3.1. Let $\mathbf{A}$ be a $2p \times 2p$ matrix defined as $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_2 & \mathbf{A}_3 \end{bmatrix}$, where $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$ are $p \times p$ symmetric matrices. Then $\mathbf{A}$ is positive (negative) definite if $-\mathbf{A}_2, \mathbf{A}_2 + \mathbf{A}_3, \mathbf{A}_2 + \mathbf{A}_1$ are all positive (negative) definite.

Proof. Consider any nonzero $\mathbf{x} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top)^\top \in \mathbb{R}^{2p}$ where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$, we have:

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}_1^\top \mathbf{A}_1 \mathbf{x}_1 + \mathbf{x}_2^\top \mathbf{A}_3 \mathbf{x}_2 + 2\mathbf{x}_1^\top \mathbf{A}_2 \mathbf{x}_2 \\ &= \mathbf{x}_1^\top (\mathbf{A}_1 + \mathbf{A}_2)\mathbf{x}_1 + \mathbf{x}_2^\top (\mathbf{A}_3 + \mathbf{A}_2)\mathbf{x}_2 - (\mathbf{x}_1 - \mathbf{x}_2)^\top \mathbf{A}_2 (\mathbf{x}_1 - \mathbf{x}_2). \end{aligned}$$

This proves the proposition. □

Noting that $-\mathbf{V}_2(\boldsymbol{\theta}), \mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_3(\boldsymbol{\theta})$ and $\mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_1(\boldsymbol{\theta})$ are all diagonally balanced matrices, with some routine calculations it can be shown that $-\mathrm{E}\mathbf{V}_2(\boldsymbol{\theta}^*), \mathrm{E}(\mathbf{V}_2(\boldsymbol{\theta}^*) + \mathbf{V}_3(\boldsymbol{\theta}^*))$ and $\mathrm{E}(\mathbf{V}_2(\boldsymbol{\theta}^*) + \mathbf{V}_1(\boldsymbol{\theta}^*))$ have only positive elements, and thus are all positive definite. Therefore, $\mathrm{E}\mathbf{V}(\boldsymbol{\theta}^*)$ is positive definite by Proposition 3.1. By continuity, when $\boldsymbol{\theta}$ is close enough to $\boldsymbol{\theta}^*$, $\mathrm{E}\mathbf{V}(\boldsymbol{\theta})$ is also positive definite, and hence $\mathrm{E}l(\boldsymbol{\theta})$ is strongly convex in a neighborhood of $\boldsymbol{\theta}^*$. Next we want to show the local convexity of $l(\boldsymbol{\theta})$ whose second order derivatives depend on the sufficient statistics $b_{i,j} = \sum_{t=1}^n X_{i,j}^t X_{i,j}^{t-1}$, and

$d_{i,j} = \sum_{t=1}^{n}\left(1-X_{i,j}^{t}\right)\left(1-X_{i,j}^{t-1}\right)$. By noticing that the network process is α-mixing with an exponential decaying mixing coefficient, we first obtain the following concentration results for $b_{i,j}$ and $d_{i,j}$, which ensure element-wise convergence of $\mathbf{V}(\boldsymbol{\theta})$ to $\mathrm{E}\mathbf{V}(\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$ when $np \to \infty$.

Lemma 3.1. Suppose $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} = (\beta_{1,0}, \cdots, \beta_{p,0}, \beta_{1,1}, \cdots, \beta_{p,1})^\top$ satisfying condition (A1). Then for any $(i,j) \in \mathcal{J}$, $\{X_{i,j}^t, t \geq 1\}$ is α-mixing with exponential decaying rates. Moreover, for any positive constant $c > 0$, by choosing $c_1 > 0$ to be large enough, it holds with probability greater than $1-(np)^{-c}$ that

$$\max_{1\leq i<j\leq p}\left\{n^{-1}\left|\sum_{t=1}^{n}\left\{X_{i,j}^t - \mathrm{E}\left(X_{i,j}^t\right)\right\}\right|, n^{-1}\left|b_{i,j}-\mathrm{E}(b_{i,j})\right|, n^{-1}\left|d_{i,j}-\mathrm{E}(d_{i,j})\right|\right\} \leq c_1 r_{n,p},$$

where $r_{n,p} = \sqrt{n^{-1}\log(np)} + n^{-1}\log(n)\log\log(n)\log(np)$.

The following lemma provides a lower bound for the smallest eigenvalue of $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$.

Lemma 3.2. Let $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty \leq r\}$ and $\mathbf{B}(\kappa_0, \kappa_1) := \{(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) : \|\boldsymbol{\beta}_0\|_\infty \leq \kappa_0, \|\boldsymbol{\beta}_1\|_\infty \leq \kappa_1\}$. Under condition (A1), for any κ_0, κ_1 and $r = c_r e^{-4\kappa_0-4\kappa_1}$ with $c_r > 0$ being a small enough constant, there exists a constant $C > 0$ such that

$$\inf_{\boldsymbol{\theta}\in\mathbf{B}_\infty(\boldsymbol{\theta}^*,r)\cap\mathbf{B}(\kappa_0,\kappa_1);\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))\,\mathbf{a} \geq Ce^{-4\kappa_0-4\kappa_1}.$$

Examining the proof indicates that the lower bound in Lemma 3.2 is attained when $\boldsymbol{\beta}_0 = (\kappa_0, \ldots, \kappa_0)^\top$ and $\boldsymbol{\beta}_1 = (-\kappa_1, \ldots, -\kappa_1)^\top$. Hence the smallest eigenvalue of $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ can decay exponentially in κ_0 and κ_1. Consequently, an upper bound for the radius κ_0 and κ_1 must be imposed so as to ensure the positive definiteness of the sample analog $\mathbf{V}(\boldsymbol{\theta})$. Moreover, Lemma 3.2 also indicates that the positive definiteness of $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ can be guaranteed when $\boldsymbol{\theta}$ is within the ℓ_∞ ball $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$. To establish the existence of the local MLE in the neighborhood, we need to evaluate the closeness of $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ and $\mathbf{V}(\boldsymbol{\theta})$ in terms of the operator norm. Intuitively, for some appropriately chosen κ_0, κ_1, if $\|\mathrm{E}(\mathbf{V}(\boldsymbol{\theta})) - \mathbf{V}(\boldsymbol{\theta})\|_2$ has a smaller order than $e^{-4\kappa_0-4\kappa_1}$ uniformly over the parameter space $\{\boldsymbol{\theta} : \|\boldsymbol{\beta}_0\|_\infty \leq \kappa_0, \|\boldsymbol{\beta}_1\|_\infty \leq \kappa_1$ and $\boldsymbol{\theta} \in \mathbf{B}_\infty(\boldsymbol{\theta}^*, r)\}$, the positive definiteness of $\mathbf{V}(\boldsymbol{\theta})$ can be concluded.

Note that $\mathbf{V}_2(\boldsymbol{\theta}) - \mathrm{E}\mathbf{V}_2(\boldsymbol{\theta})$, $\mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_3(\boldsymbol{\theta}) - \mathrm{E}(\mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_3(\boldsymbol{\theta}))$ and $\mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_1(\boldsymbol{\theta}) - \mathrm{E}(\mathbf{V}_2(\boldsymbol{\theta}) + \mathbf{V}_1(\boldsymbol{\theta}))$ are all centered and diagonally balanced matrices which can be decomposed into sums of independent random matrices. The following lemma provides a bound for evaluating the moderate deviations of these centered matrices.

Lemma 3.3. Let $\mathbf{Z} = (Z_{i,j})_{1\leq i,j\leq p}$ be a symmetric $p \times p$ random matrix such that the off-diagonal elements $Z_{i,j}, 1 \leq i < j \leq p$ are independent of each other and satisfy

$$Z_{i,i} = \sum_{j=1,\, j\neq i}^{n} Z_{i,j}, \quad \mathrm{E}(Z_{i,j}) = 0, \quad \mathrm{Var}(Z_{i,j}) \leq \sigma^2, \quad \text{and} \quad Z_{i,j} \leq b \quad \text{almost surely.}$$

Then it holds that

$$P(\|\mathbf{Z}\|_2 > \epsilon) \leq 2p\ \exp\left(-\frac{\epsilon^2}{2\sigma^2(p-1)+4b\epsilon}\right).$$

Proposition 3.1, Lemma 3.2 and Lemma 3.3 imply the theorem below.

THEOREM 3.1. Let condition (A1) hold, assume $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, and $\kappa_r := \|\boldsymbol{\beta}_r^*\|_\infty$ where $r = 0, 1$ with $\kappa_0 + \kappa_1 \leq c \log(np)$ for some small enough constant $c > 0$. Then as $np \to \infty$ with $n \geq 2$, we have that, with probability tending to 1, there exists a unique MLE in the ℓ_∞ ball $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \leq r\}$ for some $r = c_r e^{-4\kappa_0 - 4\kappa_1}$, where $c_r > 0$ is a constant.

In the proof of Theorem 3.1, we have shown that with probability tending to 1, $l(\boldsymbol{\theta})$ is convex in the convex and closed set $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$. Consequently, we conclude that there exists a unique local MLE in $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$. From Theorem 3.1 we can also see that when $\kappa_0 + \kappa_1$ becomes larger, the radius r will be smaller, and when $\kappa_0 + \kappa_1$ is bounded away from infinity, r has a constant order. From the proof we can also see that the constant c_r can be larger if the smallest eigenvalue of the expected Hessian matrix $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ is larger. Further, by allowing the upper bound of $\|\boldsymbol{\beta}_0^*\|_\infty$ to grow to infinity, our theoretical analysis covers the case where networks are sparse. Specifically, under the condition that $\|\boldsymbol{\beta}_0^*\|_\infty \leq \kappa_0$, from (2) we can obtain the following lower bound (which is achievable when $\beta_{1,0}^* = \ldots = \beta_{p,0}^* = -\kappa_0$) for the density of the stationary network:

$$\rho := \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathbf{P}\left(X_{i,j}^t = 1\right) \geq \frac{e^{-2\kappa_0}}{1 + e^{-2\kappa_0}} = O\left(e^{-2\kappa_0}\right).$$

In particular, when $\kappa_0 \leq c \log(np)$ for some constant $c > 0$, we have $\rho \geq \frac{1}{[1+(np)^{2c}]}$. Thus, compared to full dense network processes where the total number of edges for each network is of the order p^2, TWHM allows the networks with much fewer edges.

3.3. *Consistency of the local MLE.* In the previous subsection, we have proved that with probability tending to one, $l(\boldsymbol{\theta})$ is convex in $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$, where $r = c_r e^{-4\kappa_0 - 4\kappa_1}$ is defined in Theorem 3.1. Denote by $\widehat{\boldsymbol{\theta}}$ the (local) MLE in $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$. We now evaluate the ℓ_2 and ℓ_∞ distances between $\widehat{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}^*$.

Based on Theorem 4.1 we obtain a local deviation bound for $l(\boldsymbol{\theta})$ as in Corollary 4.1 in Section 4, from which we establish the following upper bound for the estimation error of $\widehat{\boldsymbol{\theta}}$ under the ℓ_2 norm:

THEOREM 3.2. Let condition (A1) hold, assume $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, and $\kappa_r := \|\boldsymbol{\beta}_r^*\|_\infty$ where $r = 0, 1$ with $\kappa_0 + \kappa_1 \leq c \log(np)$ for some small enough constant $c > 0$. Then as $np \to \infty$ with $n \geq 2$, it holds with probability converging to 1 that

$$\frac{1}{\sqrt{p}} \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 \lesssim e^{4\kappa_0 + 4\kappa_1} \sqrt{\frac{\log(np)}{np}} \left(1 + \frac{\log(np)}{\sqrt{p}}\right).$$

We discuss the implication of this theorem. When $n \to \infty$ and p is finite, that is, when we have a fixed number of nodes but a growing number of network snapshots, Theorem 3.2 indicates that $\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 = O_p\left(\sqrt{\frac{\log^3 n}{n}} e^{4\kappa_0 + 4\kappa_1}\right) = o_p(1)$ when c is small enough. On the other hand, when n, κ_0 and κ_1 are finite, Theorem 3.2 indicates that as the number of parameters p increases, the ℓ_2 error bound of $\widehat{\boldsymbol{\theta}}$ increases at a much slower rate $O\left(\sqrt{\log p}\right)$.

Although Theorem 3.2 indicates that $\frac{1}{\sqrt{p}} \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 = o_p(1)$ as $np \to \infty$, it does not guarantee the uniform convergence of all the elements in $\widehat{\boldsymbol{\theta}}$. To prove the uniform convergence in the ℓ_∞ norm, we exploit a special structure of the loss function and the

ℓ_2 norm bound obtained in Theorem 3.2. Specifically, denote $l(\boldsymbol{\theta})$ in (3.6) as $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}_{(-i)})$ where $\boldsymbol{\theta}_{(i)} := (\beta_{i,0}, \beta_{i,1})^\top$, and $\boldsymbol{\theta}_{(-i)}$ contains the remaining elements of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_{(i)}$. Using this notation, we can analogously define $\boldsymbol{\theta}^*_{(i)}$ and $\boldsymbol{\theta}^*_{(-i)}$ for the true parameter $\boldsymbol{\theta}^*$, and $\widehat{\boldsymbol{\theta}}_{(i)}$ and $\widehat{\boldsymbol{\theta}}_{(-i)}$ for the local MLE $\widehat{\boldsymbol{\theta}}$. We then have that $\boldsymbol{\theta}^*_{(i)}$ is the mimizer of $\mathrm{E}l\left(\cdot, \boldsymbol{\theta}^*_{(-i)}\right)$ while $\widehat{\boldsymbol{\theta}}_{(i)}$ is the minimizer of $l\left(\cdot, \widehat{\boldsymbol{\theta}}_{(-i)}\right)$. The error of $\widehat{\boldsymbol{\theta}}_{(i)}$ in estimating $\boldsymbol{\theta}^*_{(i)}$ then relies on the distance between $\mathrm{E}l\left(\cdot, \boldsymbol{\theta}^*_{(-i)}\right)$ and $l\left(\cdot, \widehat{\boldsymbol{\theta}}_{(-i)}\right)$, which on the other hand depends on both the ℓ_2 bound of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ and the uniform local deviation bound of $l\left(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}_{(-i)}\right)$. Based on Theorem 3.2, Corollary 4.1 in Section 4, and a sequential approach (see equations (A.12) and (A.13) in the appendix), we obtain the following bound for the estimation error under the ℓ_∞ norm.

Theorem 3.3. Let condition (A1) hold, assume $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, and $\kappa_r := \|\boldsymbol{\beta}^*_r\|_\infty$ where $r = 0, 1$ with $\kappa_0 + \kappa_1 \le c\log(np)$ for some small enough constant $c > 0$. Then as $np \to \infty, n \ge 2$, it holds with probability converging to 1 that

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty \lesssim e^{8\kappa_0 + 8\kappa_1} \log\log(np) \sqrt{\frac{\log(np)}{np}} \left(1 + \frac{\log(np)}{\sqrt{p}}\right).$$

Theorem 3.3 indicates that $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty = o_p(1)$ as $np \to \infty$. Thus all the components of $\widehat{\boldsymbol{\theta}}$ converge uniformly. On the other hand, when $\kappa = c\log(np)$ for some small enough positive constant c, we have $e^{8\kappa_0 + 8\kappa_1} \log\log(np) \sqrt{\frac{\log(np)}{np}} \left(1 + \frac{\log(np)}{\sqrt{p}}\right) \le o(c_r e^{-4\kappa_0 - 4\kappa_1})$. Compared with Theorem 3.1, we observe that although the radius r in Theorem 3.1 already tends to zero when $\|\boldsymbol{\beta}^*_0\|_\infty \le \kappa_0, \|\boldsymbol{\beta}^*_1\|_\infty \le \kappa_1$ and $\kappa_0 + \kappa_1 \le c\log(np)$ for some small enough constant $c > 0$, the ℓ_∞ error bound of $\widehat{\boldsymbol{\theta}}$ has a smaller order asymptotically and thus gives a tighter convergence rate.

We remark that in the MLE, $\boldsymbol{\beta}^*_0$ and $\boldsymbol{\beta}^*_1$ are estimated jointly. As we can see from the log-likelihood function, the information related to $\beta_{i,0}$ is captured by $X^t_{i,j}$ and $X^t_{i,j}X^{t-1}_{i,j}$, $t = 1, \ldots, n, j \ne i$, while that related to $\beta_{i,1}$ is captured by $(1 - X^t_{i,j})(1 - X^{t-1}_{i,j})$ and $X^t_{i,j}X^{t-1}_{i,j}$, $t = 1, \ldots, n, j \ne i$. This indicates that the effective "sample sizes" for estimating $\beta_{i,0}$ and $\beta_{i,1}$ are both of the order $O(np)$. While the theorems we have established in this section is for $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_0^\top, \widehat{\boldsymbol{\beta}}_1^\top)^\top$ jointly, we would expect $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ to have the same rate of convergence.

3.4. *A method of moment estimator.* Having established the existence of a unique local MLE in $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$ and proved its convergence, we still need to specify how to find this local MLE. To this end, we propose an initial estimator lying in this neighborhood. Consequently we can adopt any convex optimization method such as the coordinate descent algorithm to locate the local MLE, thanks to the convexity of the loss function in this neighborhood. Based on (2.3), an initial estimator of $\boldsymbol{\beta}_0$ denoted as $\tilde{\boldsymbol{\beta}}_0$ can be found by solving the following method of moment equations

$$\frac{\sum_{t=1}^n \sum_{j=1,\, j\ne i}^p X^t_{i,j}}{n} - \sum_{j=1,\, j\ne i}^p \frac{e^{\beta_{i,0} + \beta_{j,0}}}{1 + e^{\beta_{i,0} + \beta_{j,0}}} = 0, \quad i = 1, \cdots, p. \tag{3.8}$$

These equations can be viewed as the score functions of the pseudo loss function $f(\boldsymbol{\beta}_0) := \sum_{1 \le i,j \le p} \log\{1 + e^{\beta_{i,0} + \beta_{j,0}}\} - n^{-1} \sum_{i=1}^p \{\beta_{i,0} \sum_{t=1}^n \sum_{j=1,\, j\ne i}^p X^t_{i,j}\}$. Since the Hessian matrix of $f(\boldsymbol{\beta}_0)$ is diagonally balanced with positive elements, the Hessian matrix is positive definite, and, hence, $f(\boldsymbol{\beta}_0)$ is strongly convex. With the strong convexity, the solution of (3.8) is the minimizer of $f(\cdot)$ which can be easily obtained using any standard algorithms such as the gradient descent. On the other hand, note that

$$\mathrm{E}(X_{i,j}^t X_{i,j}^{t-1}) = \frac{e^{\beta_{i,0}+\beta_{j,0}}}{1+e^{\beta_{i,0}+\beta_{j,0}}}\left(1-\frac{1}{1+e^{\beta_{i,0}+\beta_{j,0}}+e^{\beta_{i,1}+\beta_{j,1}}}\right),$$

which motivates the use of the following estimating equations to obtain $\tilde{\boldsymbol{\beta}}_1$, the initial estimator of $\boldsymbol{\beta}_1$,

$$\sum_{t=1}^{n}\sum_{j=1,\,j\neq i}^{p}\left\{X_{i,j}^t X_{i,j}^{t-1} - \frac{e^{\tilde{\beta}_{i,0}+\tilde{\beta}_{j,0}}}{1+e^{\tilde{\beta}_{i,0}+\tilde{\beta}_{j,0}}}\left(1-\frac{1}{1+e^{\tilde{\beta}_{i,0}+\tilde{\beta}_{j,0}}+e^{\beta_{i,1}+\beta_{j,1}}}\right)\right\}=0, \tag{3.9}$$

with $i=1,\cdots,p$. Similar to (3.8), we can formulate a pseudo loss function such that given $\tilde{\boldsymbol{\beta}}_0$, its Hessian matrix corresponding to the score equations (3.9) is positive definite, and hence (3.9) can also be solved via the standard gradient descent algorithm. Since $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}_0^\top, \tilde{\boldsymbol{\beta}}_1^\top)^\top$ is obtained by solving two sets of moment equations, we call it the method of moment estimator (MME). An interesting aspect of our construction of these moment equations is that the equations corresponding to the estimation of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are decoupled. While the estimator error in estimating $\boldsymbol{\beta}_0$ propagates clearly in that of estimating $\boldsymbol{\beta}_1$, we have the following existence, uniqueness, and a uniform upper bound for the estimation error of $\tilde{\boldsymbol{\theta}}$. Our results build on a novel application of the classical interior mapping theorem [5, 35, 34].

Theorem 3.4. Let condition (A1) hold, and $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$. The MME $\tilde{\boldsymbol{\theta}}$ defined by equations (3.8) and (3.9) exists and is unique in probability. Further, assume that $\kappa_r := \|\boldsymbol{\beta}_r^*\|_\infty$ where $r=0,1$ with $\kappa_0+\kappa_1 \le c\log(np)$ for some small enough constant $c>0$. Then as $np\to\infty$ and $n\ge 2$, it holds that

$$\left\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \le O_p\left(e^{14\kappa_0+6\kappa_1}\sqrt{\frac{\log(n)\log(p)}{np}}\right).$$

When $np\to\infty$ and κ_0,κ_1 are finite, Theorem 3.4 gives $\left\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty = O_p\left(\sqrt{\frac{\log(n)\log(p)}{np}}\right)$. When $\kappa_0+\kappa_1 \asymp \log(np)$, we see that the upper bound for the local MLE in Theorem 3.3 is dominated by the upper bound of the MME in Theorem 3.4. Moreover, when $\kappa_0+\kappa_1\le c\log(np)$ for some small enough constant $c>0$, we have $\tilde{\boldsymbol{\theta}}\in\mathbf{B}_\infty(\boldsymbol{\theta}^*,r)$, where r is defined in Theorem 3.1. Thus, $\tilde{\boldsymbol{\theta}}$ is in the small neighborhood of $\boldsymbol{\theta}^*$ as required.

3.5. *The sparse case.* Note that in the previous theoretical results, the estimation error bounds depend on both κ_0 and κ_1, i.e., the upper bounds for $\|\boldsymbol{\beta}_0^*\|_\infty$ and $\|\boldsymbol{\beta}_1^*\|_\infty$. Clearly, the larger κ_0 is, the more sparse the networks could be, and the larger κ_1 is, the lag-one correlations (c.f. equation (2.4)) could be closer to one, indicating fewer fluctuations in the network process. To further characterize the effect of network sparsity, in this section, we derive further properties under a relatively sparse scenario where $-\kappa_0\le\beta_{i,0}^*\le C_\kappa$ and $-\kappa_1\le\beta_{i,1}^*\le\kappa_1$ for all $i=1,\ldots,p$ and $C_\kappa>0$ here is a constant. Under this case we have, there exist constants $C>0$ and $C_1>0$, such that $Ce^{-2\kappa_0}\le \mathrm{E}\left(X_{i,j}^t\right)\le C_1<1$. In the most sparse case where $\beta_{0,i}=-\kappa_0, i=1,\ldots,p$, the density of the stationary network is of order $O(e^{-2\kappa_0})$. Similar to Lemma 3.2 and Theorem 3.1, the following corollary provides a lower bound for the smallest eigenvalue of $\mathrm{E}(\mathbf{V}(\boldsymbol{\theta}))$ and the existence of the MLE.

Corollary 3.1. Let $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \leq r\}$ for some $r = c_r e^{-2\kappa_0 - 4\kappa_1}$ where $c_r > 0$ is a small enough constant. and denote $\mathbf{B}'(\kappa_0, \kappa_1) := \Big\{(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) : -\kappa_0 \leq \boldsymbol{\beta}_{i,0} \leq C_\kappa, i = 1, \ldots, p, \|\boldsymbol{\beta}_1\|_\infty \leq \kappa_1\Big\}$ for some constant $C_\kappa > 0$. Then, under condition (A1), there exists a constant $C > 0$ such that:

$$\inf_{\boldsymbol{\theta}\in\mathbf{B}_\infty(\boldsymbol{\theta}^*,r)\cap\mathbf{B}'(\kappa_0,\kappa_1);\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathrm{E}\left(\mathbf{V}(\boldsymbol{\theta})\right)\mathbf{a} \geq C e^{-2\kappa_0 - 4\kappa_1}.$$

Further, assume that $\boldsymbol{\theta}^* \in \mathbf{B}'(\kappa_0, \kappa_1)$ and $\kappa_0 + 2\kappa_1 < c\log(np)$ for some positive constant $c < 1/6$. Then, as $np \to \infty$ with $n \geq 2$, we have, with probability tending to 1, there exists a unique MLE in $\mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$.

Corollary 3.2. Let condition (A1) hold, assume $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, $\|\boldsymbol{\beta}_1^*\|_\infty \leq \kappa_1$, and $-\kappa_0 \leq \beta_{i,0}^* \leq C_\kappa$ for $i = 1, \ldots, p$ and some constant $C_\kappa > 0$. Then as $np \to \infty$ with $n \geq 2$, it holds with probability converging to 1 that

$$\frac{1}{\sqrt{p}}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq C e^{2\kappa_0 + 4\kappa_1}\sqrt{\frac{\log(np)}{np}}\left(1 + \frac{\log(np)}{\sqrt{p}}\right),$$

$$\text{and} \quad \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty \leq C e^{4\kappa_0 + 8\kappa_1}\log\log(np)\sqrt{\frac{\log(np)}{np}}\left(1 + \frac{\log(np)}{\sqrt{p}}\right).$$

Corollary 3.3. Let condition (A1) hold, assume $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$, $\|\boldsymbol{\beta}_1^*\|_\infty \leq \kappa_1$, and $-\kappa_0 \leq \beta_{i,0}^* \leq C_\kappa$ for $i = 1, \ldots, p$ and some constant $C_\kappa > 0$. Then as $np \to \infty$ with $n \geq 2$, it holds with probability converging to 1 that the MME $\tilde{\boldsymbol{\theta}}$ defined by equations (3.8) and (3.9) exists uniquely, and when $\kappa_0 + 2\kappa_1 < c\log(np)$ for some constant $c < 1/12$, it holds that

$$\left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty \leq O_p\left(e^{4\kappa_0 + 6\kappa_1}\sqrt{\frac{\log(n)\log(p)}{np}}\right).$$

From Corollary 3.2, we can see that when $\kappa_1 \asymp O(1)$, the MLE is consistent when $\kappa_0 \leq c\log(np)$ for some positive constant $c < 1/8$, and the corresponding lower bound for density is $O(e^{-2c\log(np)}) \succ O((np)^{-1/4})$. Similarly, from Corollary 3.3 we can see that when $\kappa_1 \asymp O(1)$, the density of the networks can be as small as $O(e^{-2c\log(np)})$ for some constant $c < 1/12$, i.e., the density is having a larger order than $(np)^{-1/6}$ for the estimation of the MME. Further, when $6\kappa_0 + 10\kappa_1 \leq c_1\log(np)$ for some constant $c_1 < 1/2$, we have $\tilde{\boldsymbol{\theta}} \in \mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$, where r is defined in Corollary 3.1. This implies the validity of using $\tilde{\boldsymbol{\theta}}$ as an initial estimator for computing the local MLE.

4. A uniform local deviation bound under high dimensionality. As we have discussed, a key to establish the consistency of the local MLE is to evaluate the magnitude of $\left|[l(\boldsymbol{\theta}) - \mathrm{E}l(\boldsymbol{\theta})] - [l(\boldsymbol{\theta}^*) - \mathrm{E}l(\boldsymbol{\theta}^*)]\right|$ for all $\boldsymbol{\theta} \in \mathbf{B}_\infty(\boldsymbol{\theta}^*, r)$ with r specified in Theorem 3.1. Such local deviation bounds are important for establishing error bounds for general M-estimators in the empirical processes [30]. Note that

$$\begin{aligned}(4.10)\qquad l(\boldsymbol{\theta}) - \mathrm{E}l(\boldsymbol{\theta}) = -\frac{1}{p}\sum_{1\leq i<j\leq p}\Bigg[&(\beta_{i,0} + \beta_{j,0})\left(\frac{a_{i,j} - \mathrm{E}(a_{i,j})}{n}\right)\\ &+ \log\left(1 + e^{(\beta_{i,1} + \beta_{j,1})}\right)\left(\frac{d_{i,j} - \mathrm{E}(d_{i,j})}{n}\right)\\ &+ \log\left(1 + e^{(\beta_{i,1} - \beta_{i,0}) + (\beta_{j,1} - \beta_{j,0})}\right)\end{aligned}$$

where $a_{i,j}, b_{i,j}$ and $d_{i,j}$ are defined in (3.7). The three terms on the right-hand side all admit the following form

$$\mathbf{L}(\boldsymbol{\theta}) = \frac{1}{p} \sum_{1 \leq i \neq j \leq p} l_{i,j}(\theta_i, \theta_j) Y_{i,j}, \tag{4.11}$$

for some functions $\mathbf{L} : \mathbb{R}^p \to \mathbb{R}$, $l_{i,j} : \mathbb{R}^2 \to \mathbb{R}$, and centered random variables $Y_{i,j}$ $(1 \leq i, j \leq p)$. Instead of establishing the uniform bound for each term in (4.10) separately, below we will establish a unified result for bounding $|\mathbf{L}(\boldsymbol{\theta}) - \mathbf{L}(\boldsymbol{\theta}')|$ over a local ℓ_∞ ball defined as $\boldsymbol{\theta} \in \mathbf{B}_\infty(\boldsymbol{\theta}', \cdot)$ for a general $\mathbf{L}$ function as in (4.11). We remark that in general without further assumptions on $\mathbf{L}$, establishing uniform deviation bounds is impossible when the dimension of the problem diverges. For our TWHM however, the decomposition (4.10) is of a particularly appealing structure in the sense that only two-way interactions between parameters θ_i exist. Based on this "sparsity" structure, we develop a novel reformulation (c.f. equation (A.24)) for the main components of the Taylor series of $\mathbf{L}(\boldsymbol{\theta})$ satisfying the following two conditions.

(L-A1) There exists a constant $\alpha > 0$, such that for any $1 \leq i \neq j \leq p$, any positive integer k, and any non-negative integer $s \leq k$, we have:

$$\frac{\partial^k l_{i,j}(\theta_i, \theta_j)}{\partial \theta_i^s \partial \theta_j^{k-s}} \leq \frac{(k-1)!}{\alpha^k}.$$

(L-A2) Random variables $Y_{i,j}, 1 \leq i \neq j \leq p$ are independent satisfying $\mathrm{E}(Y_{i,j}) = 0$, $|Y_{i,j}| \leq b_{(p)}$ and $\mathrm{Var}(Y_{i,j}) \leq \sigma_{(p)}^2$ for any i and j, where $b_{(p)}$ and $\sigma_{(p)}^2$ are constants depending on n and p but independent of i and j.

Loosely speaking, Condition (L-A1) can be seen as a smoothness assumption on the higher order derivatives of $l_{i,j}(\theta_i, \theta_j)$ so that we can properly bound these derivatives when Taylor expansion is applied. On the other hand, the upper bound for these derivatives is mild as it can diverge very quickly as k increases. For our TWHM, it can be verified that (L-A1) holds for $l_{i,j}(\theta_i, \theta_j) = \theta_i + \theta_j$ and $l_{i,j}(\theta_i, \theta_j) = \log(1 + e^{\theta_i + \theta_j})$; see (3.6). For the latter, note that the first derivative of function $l(x) = \log(1 + e^x)$ is seen as the Sigmoid function:

$$S(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}.$$

By the expression of the higher order derivatives of the Sigmoid function [23], the k-th order derivative of l is

$$\frac{\partial^k l(x)}{\partial x^k} = \frac{\sum_{m=0}^{k-2} -A(k-1, m)(-e^x)^{m+1}}{(1 + e^x)^k},$$

where $k \geq 2$ and $A(k-1, m)$ is the Eulerian number. Now for any x, we have

$$\left| \frac{\sum_{m=0}^{k-2} -A(k-1, m)(-e^x)^{m+1}}{(1 + e^x)^k} \right| \leq \sum_{m=0}^{k-2} A(k-1, m) = (k-1)!.$$

Therefore,

$$\left| \frac{\partial^k l(x)}{\partial x^k} \right| \leq (k-1)!$$

holds for all $x \in \mathbb{R}$ and $k \geq 2$. With extra arguments using the chain rule, this in return implies that (L-A1) is satisfied with $\alpha = 1$ when $l_{i,j}(\boldsymbol{\theta}) = \log\left(1 + e^{\theta_i + \theta_j}\right)$.

Condition (L-A2) is a regularization assumption for the random variables $Y_{i,j}, 1 \leq i, j \leq p$, and the bounds on their moments are imposed to ensure point-wise concentration. For our TWHM, from Lemma 1 and Lemma A.2, we have that there exist large enough constants $C > 0$ and $c > 0$ such that with probability greater than $1 - (np)^{-c}$, the random variables $\frac{a_{i,j} - \mathrm{E}(a_{i,j})}{n}$, $\frac{b_{i,j} - \mathrm{E}(b_{i,j})}{n}$ and $\frac{d_{i,j} - \mathrm{E}(d_{i,j})}{n}$ all satisfy condition (L-A2) with $b_{(p)} = C\sqrt{n^{-1}\log(np)} + Cn^{-1}\log(n)\log\log(n)\log(np)$ and $\sigma^2_{(p)} = Cn^{-1}$.

We present the uniform upper bound on the deviation of $\mathbf{L}(\boldsymbol{\theta})$ below.

Theorem 4.1. Assume conditions (L-A1) and (L-A2). For any given $\boldsymbol{\theta}' \in \mathbb{R}^p$ and $\alpha_0 \in (0, \alpha/2)$, there exist large enough constants $C > 0$ and $c > 0$ which are independent of $\boldsymbol{\theta}'$, such that, as $np \to \infty$, with probability greater than $1 - (np)^{-c}$,

$$\left|\mathbf{L}(\boldsymbol{\theta}) - \mathbf{L}(\boldsymbol{\theta}')\right| \leq C\frac{b_{(p)}\log(np) + \sigma_{(p)}\sqrt{p\log(np)}}{p}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_1$$

holds uniformly for all $\boldsymbol{\theta} \in \mathbf{B}_\infty(\boldsymbol{\theta}', \alpha_0)$.

One of the main difficulties in analyzing $\mathbf{L}(\boldsymbol{\theta})$ defined in (4.11) is that $l_{i,j}(\theta_i, \theta_j)$ and $Y_{i,j}$ are coupled, giving rise to complex terms involving both in the Taylor expansion of $\mathbf{L}(\boldsymbol{\theta})$. When Taylor expansion with order K is used, condition (L-A1) can reduce the number of higher order terms from $O(p^K)$ to $O(p^2 2^K)$. On the other hand, by formulating the main terms in the Taylor series into a matrix form in (A.24), the uniform convergence of the sum of these terms is equivalent to that of the spectral norm of a centered random matrix, which is independent of the parameters. Further details can be found in the proofs of Theorem 4.1.

Define the marginal functions of $\mathbf{L}(\boldsymbol{\theta})$ as

$$\mathbf{L}_i(\boldsymbol{\theta}) = \frac{1}{p}\sum_{j=1,\, j\neq i}^{p} l_{i,j}(\theta_i, \theta_j) Y_{i,j}, \quad i = 1, \ldots, p,$$

by retaining only those terms related to θ_i. Similar to Theorem 4.1, we state the following upper bound for these marginal functions. With some abuse of notation, let $\boldsymbol{\theta}_{-i} := (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_p)^\top$ be the vector containing all the elements in $\boldsymbol{\theta}$ except θ_i.

Theorem 4.2. If conditions (L-A1) and (L-A2) hold, then for any given $\boldsymbol{\theta}' \in \mathbb{R}^p$ and $\alpha_0 \in (0, \alpha/2)$, there exist large enough constants $C > 0$ and $c > 0$ which are independent of $\boldsymbol{\theta}'$, such that, as $np \to \infty$, with probability greater than $1 - (np)^{-c}$,

$$\begin{aligned}&\left|\mathbf{L}_i(\boldsymbol{\theta}) - \mathbf{L}_i(\boldsymbol{\theta}')\right| \\ &\leq C\frac{b_{(p)}}{p}\left\|\boldsymbol{\theta}_{-i} - \boldsymbol{\theta}'_{-i}\right\|_1 + C\left(\left\|\boldsymbol{\theta}_{-i} - \boldsymbol{\theta}'_{-i}\right\|_1 + 1\right)\left|\theta_i - \theta'_i\right|\frac{b_{(p)}\log(np) + \sigma_{(p)}\sqrt{p\log(np)}}{p}\end{aligned}$$

holds uniformly for all $\boldsymbol{\theta} \in \mathbf{B}_\infty(\boldsymbol{\theta}', \alpha_0)$, and $i = 1, \cdots, p$.

Similar to (4.10), we can also decompose $l\left(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}_{(-i)}\right) - \mathrm{E}l\left(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}_{(-i)}\right)$ into the sum of three components taking the form (4.11). Consequently, by setting $\boldsymbol{\theta}'$ in Theorems 4.1 and 4.2 to be the true parameter $\boldsymbol{\theta}^*$, we can obtain the following upper bounds.

Corollary 4.1. For any given $0 < \alpha_0 < 1/4$, there exist large enough positive constants c_1, c_2, and C such that

(i) with probability greater than $1-(np)^{-c_1}$,

$$\text{(4.12)}\quad \left|\left(l(\boldsymbol{\theta})-l(\boldsymbol{\theta}^*)\right)-\left(\mathrm{E}l(\boldsymbol{\theta})-\mathrm{E}l(\boldsymbol{\theta}^*)\right)\right| \le C_1\left(1+\frac{\log(np)}{\sqrt{p}}\right)\sqrt{\frac{\log(np)}{n}}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\right\|_2$$

holds uniformly for all $\boldsymbol{\theta}\in\mathbf{B}_\infty(\boldsymbol{\theta}^*,\alpha_0)$ with some constant $\alpha_0<1/2$;

(ii) with probability greater than $1-(np)^{-c_2}$,

$$\text{(4.13)}\quad \left|l\left(\boldsymbol{\theta}_{(i)},\boldsymbol{\theta}^*_{(-i)}\right)-l\left(\boldsymbol{\theta}^*_{(i)},\boldsymbol{\theta}^*_{(-i)}\right)-\left[\mathrm{E}l\left(\boldsymbol{\theta}_{(i)},\boldsymbol{\theta}^*_{(-i)}\right)-\mathrm{E}l\left(\boldsymbol{\theta}^*_{(i)},\boldsymbol{\theta}^*_{(-i)}\right)\right]\right|$$
$$\le C_2\left(1+\frac{\log(np)}{\sqrt{p}}\right)\sqrt{\frac{\log(np)}{n}}\left\|\boldsymbol{\theta}_{(i)}-\boldsymbol{\theta}^*_{(i)}\right\|_2$$

holds uniformly for all $\boldsymbol{\theta}_{(i)}\in\mathbf{B}_\infty\left(\boldsymbol{\theta}^*_{(i)},\alpha_0\right)$ with some constant $\alpha_0<1/2$.

In (4.12) and (4.13) we have replaced the ℓ_1 norm based upper bounds in Theorems 4.1 and 4.2 with ℓ_2 norm based upper bounds using the fact that for all $\mathbf{x}\in\mathbb{R}^p$, $\|\mathbf{x}\|_1\le\sqrt{p}\|\mathbf{x}\|_2$. We remark that networks are generally stylized by different features such as dynamic change, node heterogeneity, homophily, transitivity, among others. In this paper, we are mainly focusing on dealing with node heterogeneity in dynamic networks. When other stylized features are considered together with node heterogeneity, the objective function can also take a similar form as the $\mathbf{L}(\boldsymbol{\theta})$ defined in (4.12). On the other hand, the log-likelihood functions of many other models accounting for node heterogeneity can be written in a form similar to (4.11). For example, the general class of network models with the edge formation probabilities in the form of $f(\alpha_i,\beta_j)$, where $f(\cdot)$ is a density or probability mass function and (α_i,β_i) are the node-specific parameters of node i. This includes for example, the p_1 model [12], the directed β-model [33], and the bivariate gamma model [7]). Further, in the analysis of ranking data, a common formulation is also to introduce individual-specific parameters/scores for ranking; see for example the classical Bradley-Terry model and its variations [8]. Our results here can potentially be applied to the theoretical analysis of these models or their variations when other stylized features are simultaneously considered with node heterogeneity.

5. Numerical study.

In this section, we assess the performance of the local MLE. We compute a regularized MME, which demonstrates enhanced numerical stability compared to the vanilla MME presented in (3.9), for the purpose of comparative analysis. The regularized MME can be viewed as a special case of [27] with a shrinkage towards $\mathbf{0}$ for the parameter $\boldsymbol{\beta}_1$. Specifically, for the former, we solve

$$\text{(5.14)}\quad \frac{-1}{np}\sum_{t=1}^{n}\sum_{j=1,\,j\neq i}^{p}\left\{X_{i,j}^tX_{i,j}^{t-1}-\frac{e^{\tilde\beta_{i,0}+\tilde\beta_{j,0}}}{1+e^{\tilde\beta_{i,0}+\tilde\beta_{j,0}}}\left(1-\frac{1}{1+e^{\tilde\beta_{i,0}+\tilde\beta_{j,0}}+e^{\beta_{i,1}+\beta_{j,1}}}\right)\right\}+\lambda\beta_{i,1}=0,$$

with $i=1,\cdots,p$, where $\lambda\beta_{i,1}$ can be seen as a ridge penalty with $\lambda>0$ as the regularization parameter. Denote the regularized MME as $\tilde{\boldsymbol{\theta}}_\lambda$. Similar to Theorem 3.4, by choosing $\lambda=C_\lambda e^{2\kappa}\sqrt{\frac{\log(np)}{np}}$ for some constant C_λ, we can show that $\left\|\tilde{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_\infty\le O_p\left(e^{26\kappa}\sqrt{\frac{\log(n)\log(p)}{np}}\right)$. In our implementation we take $\lambda=\sqrt{\frac{\log(np)}{np}}$. Following [27], one may also apply a ridge penalty to the equation (3.8). This could result in a tighter error bound for the estimation of $\boldsymbol{\beta}_0$. The MLE of TWHM is obtained via gradient descent using $\tilde{\boldsymbol{\theta}}_\lambda$ as the initial value.

TABLE 1

Signs of the smallest eigenvalues of the Hessian matrices of $l(\boldsymbol{\theta})$ and $\mathrm{E}l(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ or $\mathbf{0}_{2p}$ when different values of $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_0^{*\top}, \boldsymbol{\beta}_1^{*\top})^\top$ are used to generate data.

Sign of the smallest eigenvalue of $l(\boldsymbol{\theta}^*)$				Sign of the smallest eigenvalue of $\mathrm{E}l(\boldsymbol{\theta}^*)$			
	$\boldsymbol{\beta}_0^*= \mathbf{0.2}_p$	$\boldsymbol{\beta}_0^*= \mathbf{0.5}_p$	$\boldsymbol{\beta}_0^*= \mathbf{1}_p$		$\boldsymbol{\beta}_0^*= \mathbf{0.2}_p$	$\boldsymbol{\beta}_0^*= \mathbf{0.5}_p$	$\boldsymbol{\beta}_0^*= \mathbf{1}_p$
$\boldsymbol{\beta}_1^*= \mathbf{0.2}_p$	+	+	+	$\boldsymbol{\beta}_1^*= \mathbf{0.2}_p$	+	+	+
$\boldsymbol{\beta}_1^*= \mathbf{0.5}_p$	+	+	+	$\boldsymbol{\beta}_1^*= \mathbf{0.5}_p$	+	+	+
$\boldsymbol{\beta}_1^*= \mathbf{1}_p$	+	+	+	$\boldsymbol{\beta}_1^*= \mathbf{1}_p$	+	+	+
Sign of the smallest eigenvalue of $l(\mathbf{0}_{2p})$				Sign of the smallest eigenvalue of $\mathrm{E}l(\mathbf{0}_{2p})$			
	$\boldsymbol{\beta}_0^*= \mathbf{0.2}_p$	$\boldsymbol{\beta}_0^*= \mathbf{0.5}_p$	$\boldsymbol{\beta}_0^*= \mathbf{1}_p$		$\boldsymbol{\beta}_0^*= \mathbf{0.2}_p$	$\boldsymbol{\beta}_0^*= \mathbf{0.5}_p$	$\boldsymbol{\beta}_0^*= \mathbf{1}_p$
$\boldsymbol{\beta}_1^*= \mathbf{0.2}_p$	+	+	−	$\boldsymbol{\beta}_1^*= \mathbf{0.2}_p$	+	+	−
$\boldsymbol{\beta}_1^*= \mathbf{0.5}_p$	−	−	−	$\boldsymbol{\beta}_1^*= \mathbf{0.5}_p$	+	+	−
$\boldsymbol{\beta}_1^*= \mathbf{1}_p$	−	−	−	$\boldsymbol{\beta}_1^*= \mathbf{1}_p$	−	−	−

5.1. *Non-convexity of $l(\boldsymbol{\theta})$ and $\mathrm{E}l(\boldsymbol{\theta})$.* Given the form of $l(\boldsymbol{\theta})$, it is intuitively true that it may not be convex everywhere. We confirm this via a simple example. Take $(n,p) = (2,1000)$ and set $\boldsymbol{\beta}_0^*, \boldsymbol{\beta}_1^*$ to be $\mathbf{0.2}_p$, $\mathbf{0.5}_p$ or $\mathbf{1}_p$. We evaluate the smallest eigenvalue of the Hessian matrix of $l(\boldsymbol{\theta})$ and its expectation $\mathrm{E}l(\boldsymbol{\theta})$ at the true parameter value $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_0^{*\top}, \boldsymbol{\beta}_1^{*\top})^\top$, or at $\boldsymbol{\theta} = \mathbf{0}_{2p}$ in one experiment. From the top half of Table 1 we can see that, when evaluated at $\boldsymbol{\theta}^*$, the Hessian matrices are all positive definite. However, when evaluated at $\boldsymbol{\theta} = \mathbf{0}_{2p}$, from the bottom half of the table we can see that the Hessian matrices are no longer positive definite when $\boldsymbol{\theta}^*$ is far away from $\mathbf{0}_{2p}$. Even when the Hessian matrix of $\mathrm{E}l(\boldsymbol{\theta})$ is so at $\boldsymbol{\theta} = \mathbf{0}_{2p}$ with $\boldsymbol{\theta}^* = \mathbf{0.5}_{2p}$, the corresponding Hessian matrix of $l(\boldsymbol{\theta})$ at this point has a negative eigenvalue. Thus, $\mathrm{E}l(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta})$ are not globally convex.

5.2. *Parameter estimation.* We first evaluate the error rates of the MLE and MME under different combinations of n and p. We set $n = 2, 5, 10$, or 20 and $p \in \lfloor 200 \times 1.2^{0:6} \rfloor = \{200, 240, 288, 346, 415, 498, 598\}$, which results in a total of 28 different combinations of (n,p). For each (n,p), the data are generated such that $\{\mathbf{X}^t\} \sim P_{\boldsymbol{\theta}^*}$ where the parameters $\beta_{i,0}^*$ and $\beta_{i,1}^* (1 \le i \le p)$ are drawn independently from the uniform distribution with parameters in $(-1,1)$. Each experiment is repeated 100 times under each setting. Denote the estimator (which is either the MLE or the MME) as $\widehat{\boldsymbol{\theta}}$, and the true parameter value as $\boldsymbol{\theta}^*$. We report the average ℓ_2 error $\frac{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2}{\sqrt{p}}$ and the average ℓ_∞ error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty$ in Figure 2. From this figure, we can see that the errors in terms of the ℓ_∞ norm and the ℓ_2 norm decrease for MME and MLE as n or p increases, while the errors of MLE are smaller across all settings. These observations are consistent with our findings in the main theory.

Next, we provide more numerical simulation to evaluate the performance of MLE and MME by imposing different structures on $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$. In particular, we want to evaluate how the estimation accuracy changes by varying the sparsity of the networks as well as varying the correlations of the network sequence. Note that the expected density of the stationary distribution of the network process is simply

$$\frac{1}{p(p-1)} \mathrm{E}\left(\sum_{1 \le i \ne j \le p} X_{i,j}^t \right) = \frac{1}{p(p-1)} \left(\sum_{1 \le i \ne j \le p} \frac{e^{\beta_{i,0}^* + \beta_{j,0}^*}}{1 + e^{\beta_{i,0}^* + \beta_{j,0}^*}} \right).$$

In the sequel, we will use two parameters a and b to generate $\boldsymbol{\beta}_r^*$, $r = 0, 1$, according to the following four settings:

Setting 1. $\{a\}$: all the elements in $\boldsymbol{\beta}_r^*$ are set to be equal to a.

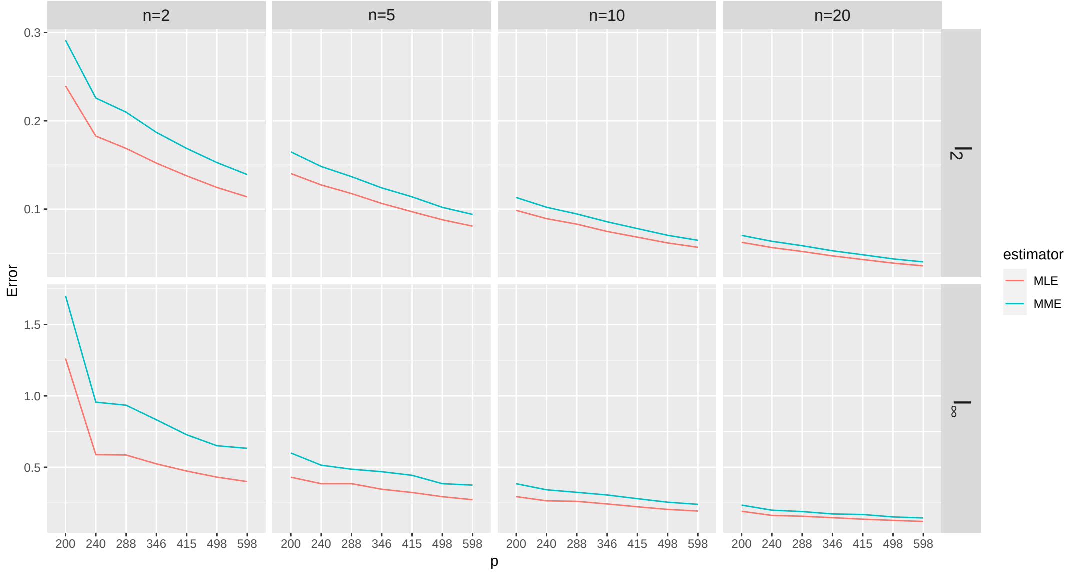

Fig 2: Mean errors of MME and MLE in terms of the ℓ_2 and ℓ_∞ norm.

Setting 2. $\{a, b\}$: the first 10% elements of $\boldsymbol{\beta}_r^*$ are set to be equal to a, while the other elements are set to be equal to b.

Setting 3. $\mathcal{L}_{(a,b)}$: the parameters take values in a linear form as $\beta_{i,r}^* = a + (a-b) * (i-1)/(p-1)$, $i = 1, \cdots, p$.

Setting 4. $U_{(a,b)}$: the p elements in $\boldsymbol{\beta}_r^*$ are generated independently from the uniform distribution with parameters a and b.

In Table 2, we generate $\boldsymbol{\beta}_1^*$ using Setting 1 with $a = 0$, and generate $\boldsymbol{\beta}_0^*$ using Setting 2 with different choices for a and b to obtain networks with different expected density. In Table 3, we generate $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$ using combinations of these four settings with different parameters such that the resulting networks have expected density either around 0.05 (sparse) or 0.5 (dense). The number of networks in each process and the number of nodes in each network are set as $(n, p) = (20, 200), (20, 500), (50, 200)$ or $(50, 500)$. The errors for estimating $\boldsymbol{\theta}^*$ in terms of the ℓ_∞ and ℓ_2 norms are reported via 100 replications. To further compare the accuracy for estimating $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$, in Table 4, we have conducted experiments under Settings 3 and 4, and reported the estimation errors for $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$ separately. We summarize the simulation results below:

- The effect of (n, p). Similar to what we have observed in Figure 2, the estimation errors become smaller when n or p becomes larger. Interestingly, from Tables 2–3 we can observe that, under the same setting, the errors in ℓ_2 norm when $(n, p) = (50, 200)$ are very close to those when $(n, p) = (20, 500)$. This is to some degree consistent with our finding in Theorem 3.2 where the upper bound depends on (n, p) through their product np.
- The effect of sparsity. From Table 2 we can see that, as the expected density decreases, the estimation errors increase in almost all the cases. On the other hand, even though the parameters take different values in Table 3, the errors in the sparse cases are in general larger than those in the dense cases.
- The effect of $\kappa_0 := \|\boldsymbol{\beta}_0^*\|_\infty$. In general, the estimation errors become larger when κ_0 is larger as observed from Table 2. With the same overall sparsity level, larger κ_0 is associated with larger estimation errors as can be seen in Table 3.

Table 2

The estimation errors of MME and MLE under Setting 1 and Setting 2 for $\boldsymbol{\beta}_0^*$ by setting $\boldsymbol{\beta}_1^* = \mathbf{0}_p$.

n	p	$\boldsymbol{\beta}_0^*$	MME, ℓ_2	MME, ℓ_∞	MLE, ℓ_2	MLE, ℓ_∞
20	200	$\{0\}$	0.074	0.219	0.071	0.212
50	200	$\{0\}$	0.046	0.138	0.045	0.136
20	500	$\{0\}$	0.046	0.150	0.045	0.146
50	500	$\{0\}$	0.029	0.093	0.028	0.092
20	200	$\{0.5, -0.5\}$	0.092	0.222	0.091	0.217
50	200	$\{0.5, -0.5\}$	0.058	0.140	0.058	0.139
20	500	$\{0.5, -0.5\}$	0.058	0.154	0.057	0.148
50	500	$\{0.5, -0.5\}$	0.036	0.095	0.036	0.093
20	200	$\{1, -1\}$	0.120	0.305	0.117	0.284
50	200	$\{1, -1\}$	0.074	0.186	0.074	0.177
20	500	$\{1, -1\}$	0.075	0.200	0.073	0.190
50	500	$\{1, -1\}$	0.038	0.125	0.036	0.119
20	200	$\{1.5, -1.5\}$	0.164	0.436	0.156	0.397
50	200	$\{1.5, -1.5\}$	0.102	0.255	0.097	0.236
20	500	$\{1.5, -1.5\}$	0.103	0.287	0.097	0.262
50	500	$\{1.5, -1.5\}$	0.065	0.178	0.061	0.164

- MLE vs MME. In general, the estimation errors of the MLE are smaller than those of the MME in most cases as can be seen in Tables 2 and Table 3. In Table 4 where the estimation errors for $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$ are reported separately, we can see that the estimation errors of the MME of $\boldsymbol{\beta}_1^*$ are generally larger than those of the MLE of $\boldsymbol{\beta}_1^*$, especially when n is large.

5.3. *Real data.* In this section, we apply our TWHM to a real dataset to examine an insect interaction network process [22]. We focus on a subset of the data named insecta-ant-colony4 that contains the social interactions of 102 ants in 41 days. In this dataset, the position and orientation of all the ants were recorded twice per second to infer their movements and interactions, based on which 41 daily networks were constructed. More specifically, $X_{i,j}^t$ is 1 if there is an interaction between ants i and j during day t, and 0 otherwise. In the ACF and PACF plots of the degree sequences of selected ants (c.f. Figure **??** in Appendix **??**), we can observe patterns similar to those of a first-order autoregressive model with long memory. This motivates the use of TWHM for the analysis of this dataset.

In [22], the 41 daily networks were split into four periods with 11, 10, 10, and 10 days respectively, because the corresponding days separating these periods were identified as change-points. By excluding ants that did not interact with others, we are left with $p = 102$ nodes in period one, $p = 73$ nodes in period two, $p = 55$ nodes in period three and $p = 35$ nodes in period four. Thus we take the networks on day 1, day 12, day 22 and day 32 as the initial networks and fit four different TWHMs, one for each of the four periods.

To appreciate how TWHM captures static heterogeneity, we present a subgraph of 10 nodes during the fourth period ($t = 32$–41), 5 of which have the largest and 5 have the smallest fitted $\beta_{i,0}$ values. The edges of this subgraph are drawn to represent aggregated static connections defined as $(\mathbf{X}^{32} + \cdots + \mathbf{X}^{42})/10$ between these ants. We can see from the left panel of Figure 3 that the magnitudes of the fitted static heterogeneity parameters agree in principle with the activeness of each ant making connections. On the other hand, we examine how TWHM can capture dynamic heterogeneity. Towards this, we plot a subgraph of the 10 nodes having the smallest fitted $\beta_{i,0}$ values in Figure 3(b), where edges represent the magnitude of $\sum_{t=33}^{41} I\left(X_{i,j}^t = X_{i,j}^{t-1}\right)/9$ which is a measure

TABLE 3

The average estimation errors of MME and MLE under combinations of different settings

Density = 0.05							
n	p	$\boldsymbol{\beta}_0^*$	$\boldsymbol{\beta}_1^*$	MME, ℓ_2	MME, ℓ_∞	MLE, ℓ_2	MLE, ℓ_∞
20	200	$\mathcal{L}_{(-4,0)}$	$U_{(-1,1)}$	0.419	1.833	0.392	1.8
50	200	$\mathcal{L}_{(-4,0)}$	$U_{(-1,1)}$	0.253	0.913	0.227	0.82
20	500	$\mathcal{L}_{(-4,0)}$	$U_{(-1,1)}$	0.246	1.119	0.218	0.9
50	500	$\mathcal{L}_{(-4,0)}$	$U_{(-1,1)}$	0.170	0.626	0.148	0.621
20	200	$\mathcal{L}_{(-4,0)}$	$\{0\}$	0.275	1.452	0.280	1.516
50	200	$\mathcal{L}_{(-4,0)}$	$\{0\}$	0.161	0.771	0.162	0.774
20	500	$\mathcal{L}_{(-4,0)}$	$\{0\}$	0.160	0.892	0.162	0.904
50	500	$\mathcal{L}_{(-4,0)}$	$\{0\}$	0.098	0.506	0.099	0.507
20	200	$\{-1.47\}$	$U_{(-1,1)}$	0.187	0.588	0.161	0.514
50	200	$\{-1.47\}$	$U_{(-1,1)}$	0.116	0.351	0.099	0.305
20	500	$\{-1.47\}$	$U_{(-1,1)}$	0.114	0.387	0.099	0.339
50	500	$\{-1.47\}$	$U_{(-1,1)}$	0.073	0.246	0.062	0.208
20	200	$\{-1.47\}$	$\{0\}$	0.150	0.482	0.151	0.484
50	200	$\{-1.47\}$	$\{0\}$	0.93	0.289	0.093	0.29
20	500	$\{-1.47\}$	$\{0\}$	0.93	0.309	0.093	0.311
50	500	$\{-1.47\}$	$\{0\}$	0.058	0.195	0.058	0.195
Density = 0.5							
20	200	$\mathcal{L}_{(-2,2)}$	$U_{(-0.1,0.1)}$	0.132	0.415	0.012	0.318
50	200	$\mathcal{L}_{(-2,2)}$	$U_{(-0.1,0.1)}$	0.080	0.238	0.069	0.194
20	500	$\mathcal{L}_{(-2,2)}$	$U_{(-0.1,0.1)}$	0.080	0.272	0.068	0.217
50	500	$\mathcal{L}_{(-2,2)}$	$U_{(-0.1,0.1)}$	0.050	0.168	0.043	0.135
20	200	$\mathcal{L}_{(-1,1)}$	$U_{(-1,1)}$	0.107	0.324	0.095	0.264
50	200	$\mathcal{L}_{(-1,1)}$	$U_{(-1,1)}$	0.067	0.194	0.060	0.163
20	500	$\mathcal{L}_{(-1,1)}$	$U_{(-1,1)}$	0.071	0.267	0.061	0.205
50	500	$\mathcal{L}_{(-1,1)}$	$U_{(-1,1)}$	0.044	0.156	0.039	0.130
20	200	$\mathcal{L}_{(-2,2)}$	$U_{(-1,1)}$	0.137	0.478	0.112	0.329
50	200	$\mathcal{L}_{(-2,2)}$	$U_{(-1,1)}$	0.084	0.274	0.070	0.205
20	500	$\mathcal{L}_{(-2,2)}$	$U_{(-1,1)}$	0.087	0.352	0.071	0.250
50	500	$\mathcal{L}_{(-2,2)}$	$U_{(-1,1)}$	0.054	0.211	0.044	0.150

of the extent that an edge is preserved across the whole period and hence dynamic heterogeneity. Again, we can see an agreement between the fitted $\boldsymbol{\beta}_1^*$ and how likely these nodes will preserve their ties.

To evaluate how TWHM performs when it comes to making prediction, we further carry out the following experiments:

(i) From (1), given the MLE $\{\widehat{\beta}_{i,r}, i=1,\dots,p, r=0,1\}$ and the network at time $t-1$, we can estimate the conditional expectation of node i's degree as

$$\tilde{d}_i^t := \sum_{j=1,\, j\neq i}^{p} \mathrm{E}\left(X_{i,j}^t \middle| X_{i,j}^{t-1}, \widehat{\boldsymbol{\theta}}\right)$$

$$= \sum_{j=1,\, j\neq i}^{p} \left(\frac{e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}}{1+e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}+e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}} + \frac{e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}}{1+e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,}}+e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}} X_{i,j}^{t-1} \right).$$

We can then compare the density of the estimated degree sequence $\{\tilde{d}_i^t, i=1,\dots,p\}$ with that of the observed degree sequence $\{d_i^t, i=1,\dots,p\}$ at time t. As a comparison, we treat the network observations from the same period as i.i.d. observations

Table 4

The means and standard deviations of the errors of MME and MLE for estimating β_0^* and β_1^*.

n		20	100	20	100
p		200	200	500	500
$\beta_0^* \sim \mathcal{L}_{(-1,1)}$ and $\beta_1^* \sim U_{(0,2)}$					
MME, ℓ_2	β_0^*	0.163(0.010)	0.096(0.006)	0.099(0.004)	0.057(0.002)
	β_1^*	0.177(0.010)	0.084(0.005)	0.104(0.004)	0.050(0.002)
MME, ℓ_∞	β_0^*	0.570(0.103)	0.367(0.085)	0.395(0.070)	0.241(0.042)
	β_1^*	0.658(0.137)	0.421(0.079)	0.438(0.076)	0.214(0.037)
MLE, ℓ_2	β_0^*	0.211(0.013)	0.091(0.006)	0.121(0.005)	0.054(0.002)
	β_1^*	0.166(0.011)	0.072(0.005)	0.096(0.004)	0.043(0.002)
MLE, ℓ_∞	β_0^*	0.809(0.180)	0.354(0.076)	0.532(0.098)	0.232(0.041)
	β_1^*	0.617(0.116)	0.265(0.052)	0.399(0.065)	0.172(0.028)
$\beta_0^* \sim \mathcal{L}_{(-2,0)}$ and $\beta_1^* \sim U_{(0,2)}$					
MME, ℓ_2	β_0^*	0.133(0.012)	0.080(0.007)	0.081(0.004)	0.047(0.002)
	β_1^*	0.093(0.006)	0.053(0.004)	0.056(0.003)	0.032(0.002)
MME, ℓ_∞	β_0^*	0.568(0.104)	0.365(0.087)	0.394(0.071)	0.241(0.042)
	β_1^*	0.387(0.069)	0.236(0.043)	0.258(0.037)	0.162(0.025)
MLE, ℓ_2	β_0^*	0.176(0.016)	0.076(0.007)	0.100(0.006)	0.044(0.002)
	β_1^*	0.116(0.009)	0.051(0.004)	0.068(0.003)	0.031(0.002)
MLE, ℓ_∞	β_0^*	0.809(0.181)	0.351(0.078)	0.531(0.099)	0.232(0.041)
	β_1^*	0.513(0.088)	0.227(0.047)	0.348(0.058)	0.158(0.024)

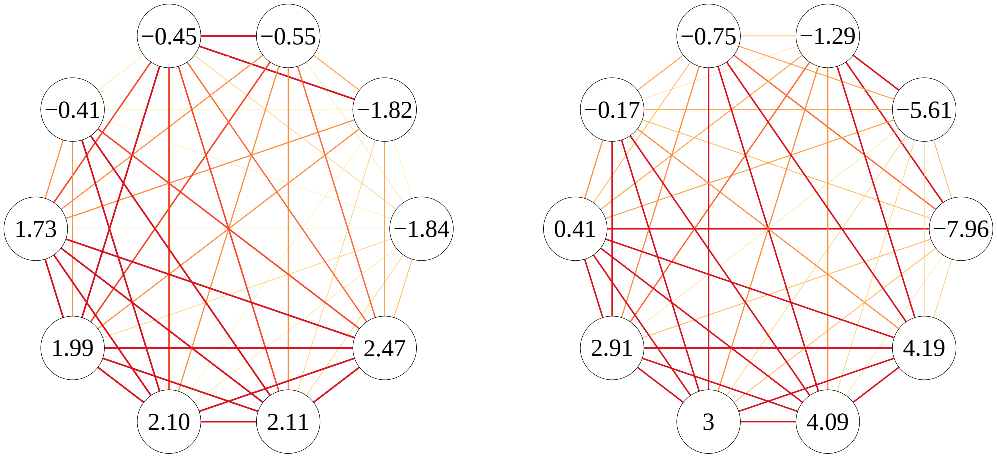

Fig 3: The aggregated networks of 10 selected ants during the fourth period reflect static heterogeneity (Left) and dynamic heterogeneity (Right) respectively. The thickness of each edge is proportional to the aggregation. The number in the nodes are the fitted $\beta_{i,0}$ (Left) and $\beta_{i,1}$ (Right).

and modeled them using the classical β-model. This yielded four static static β-model estimates, one for each of the four periods. The blue curves in Fig. 4 represent the smoothed degree distributions derived from the degree sequences $\{\check{\mathbf{d}}^t\}$ of these estimated β-models in the four periods.

The fitted degree distributions are presented in Figure 4, from which we can see that the estimated densities follow the observed densities closely. This suggests that the TWHM performs well for one-step-ahead prediction. To quantitatively evaluate the closeness between the estimated degree sequences $\{\tilde{\mathbf{d}}^t\}$, $\{\check{\mathbf{d}}^t\}$ and the true degree sequence $\{\mathbf{d}^t\}$, we compute the Kolmogorov-Smirnov (KS) distance and conduct the KS test at $t = 2, \cdots, 41$. The mean and standard deviation of the KS distances, the p-values of the KS test, and the rejection rate are summarized in Table 5. In particular, with a significant level 0.05, out of the 40 KS tests, there are 38 times we do not reject the null hypothesis that $\{\tilde{\mathbf{d}}^t\}$ and $\{\mathbf{d}^t\}$ are from the same distribution, resulting in a rejection rate of 0.05 which is identical to the significance level. For the β-model based degree sequence estimators $\{\check{\mathbf{d}}^t\}$, 8 out of the 40 tests were rejected. These results suggest that our model has very promising performance in recovering the degree sequences.

Table 5

The mean and standard deviation of the KS distances, the p-values of KS test, and the rejection rates between the true degree sequence $\{\mathbf{d}^t\}$ and the estimators (i.e., $\{\tilde{\mathbf{d}}^t\}$ based on the THWM, and $\{\check{\mathbf{d}}^t\}$ the β-model for each period). These metrics are evaluated across the 40 networks ($t = 2, \ldots, 41$) in the ant dataset.

	KS distance	KS test p-value	Rejection rate
$\tilde{\mathbf{d}}^t$ vs $\mathbf{d}^t$	0.179(0.058)	0.361(0.267)	0.05
$\check{\mathbf{d}}^t$ vs $\mathbf{d}^t$	0.192(0.061)	0.298(0.246)	0.20

(ii) By incorporating network dynamics, TWHM naturally enables one-step-ahead link prediction via

$$\mathbf{P}\left(\widehat{X}_{i,j}^t = 1 \middle| X_{i,j}^{t-1}\right) = \frac{e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}}{1+e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}+e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}} + \frac{e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}}{1+e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}+e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}}} X_{i,j}^{t-1}. \tag{5.15}$$

To transform these probabilities into links, we threshold them by setting $\widehat{X}_{i,j}^t = 1$ when $\mathbf{P}\left(\widehat{X}_{i,j}^t = 1\right) \geq c_{i,j}$ and $\widehat{X}_{i,j}^t = 0$ when $\mathbf{P}\left(\widehat{X}_{i,j}^t = 1\right) < c_{i,j}$ for some cut-off constants $c_{i,j}$. As an illustration, we first consider simply setting $c_{i,j} = 0.5$ for all $1 \leq i < j \leq p$ for predicting links. We shall denote this approach as TWHM$_{0.5}$.

As an alternative, owing to the fact that networks may change slowly, for a given parameter ω, we also consider the following adaptive approach for choosing $c_{i,j}$:

$$\tilde{X}_{i,j}^t := I\{\omega\mathbf{P}\left(\widehat{X}_{i,j}^t = 1\right) + (1-\omega)\, X_{i,j}^{t-1} > 0.5\}. \tag{5.16}$$

It can be shown that the above estimator is equivalent to the prediction rule $I\left\{\mathbf{P}\left(\widehat{X}_{i,j}^t = 1\right) > c_{i,j}\right\}$ with cut-off values specified as

$$c_{i,j} = \frac{0.5e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}} + (1-w)\, e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}}{(1-w) + e^{\widehat{\beta}_{i,1}+\widehat{\beta}_{j,1}} + (1-w)\, e^{\widehat{\beta}_{i,0}+\widehat{\beta}_{j,0}}}, \quad 1 \leq i < j \leq p.$$

This method is denoted as TWHM$_{adaptive}$. Lastly, as a benchmark, we have also considered a naive approach that simply predicts $\mathbf{X}^t$ as $\mathbf{X}^{t-1}$.

In this experiment, we set the number of training samples to be $n_{train} = 2, 5$ or 8. For a given training sample size n_{train} and a period with n networks, we predict the graph $\mathbf{X}^{n_{train}+i}$ based on the previous n_{train} networks $\{\mathbf{X}^t, t = i, \ldots, n_{train} + i - 1\}$ for $i = 1, \ldots, n - n_{train}$. That is, over the four periods in the data, we have predicted 33, 21 and 9 networks, with 5151 edges in each network in the first period, 2628 in the second period, 1485 in the third period, and 595 in the fourth period for

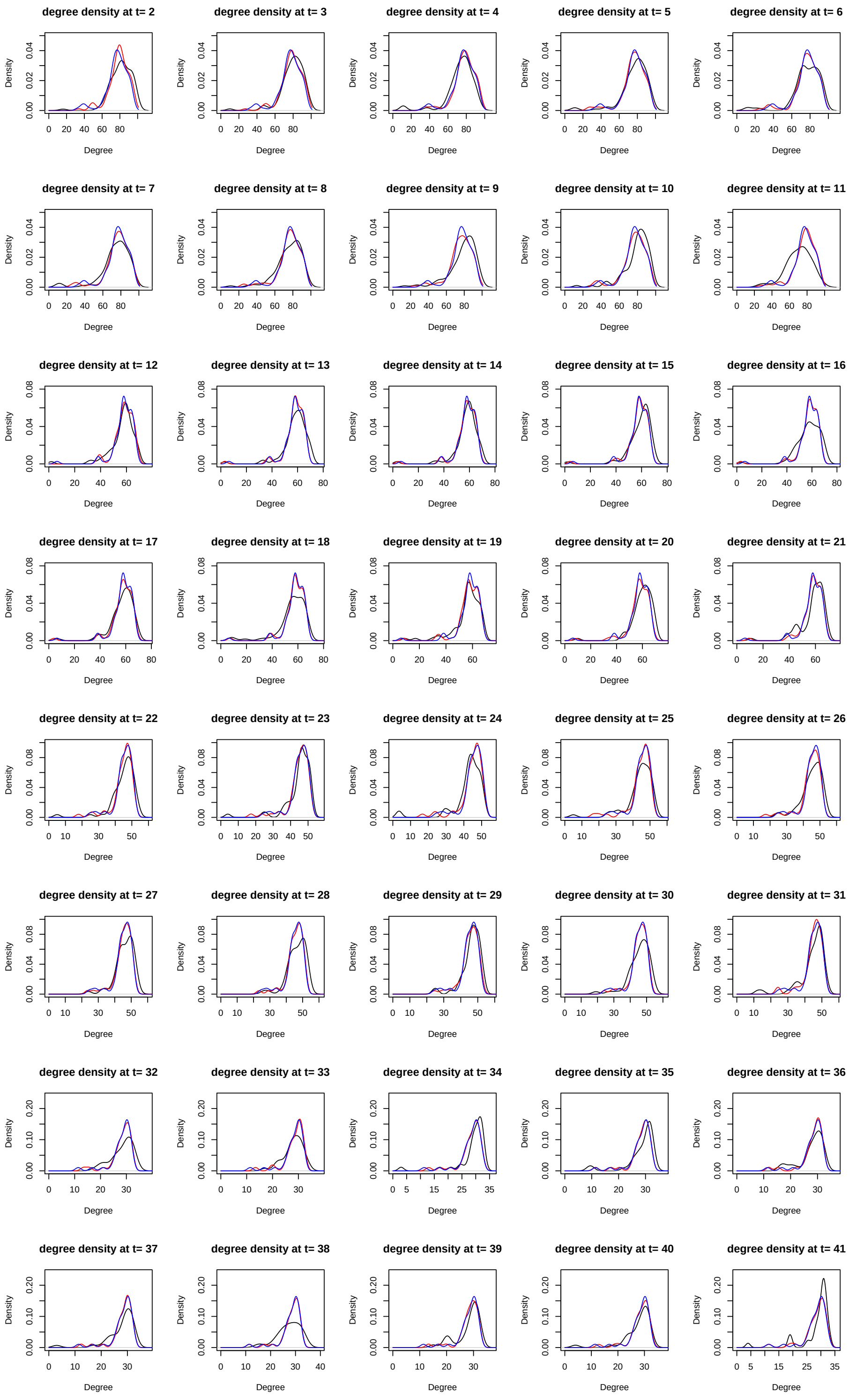

Fig 4: The observed and estimated degree distributions. X-axis: the node degrees; Red curves: the smoothed degree distributions of the estimated degree sequences from TWHM; Blue curves: smoothed degree distributions of the degree sequences from the estimated classical β-model for each of the four periods;

Black curves: the smoothed degree distributions of the observed degree sequences.

Table 6
The prediction accuracy of TWHM with 0.5 *as a cut-off point, TWHM with adaptive cut-off points, and the naive estimator* $\mathbf{X}^{t-1}$.

n_{train}	Period	TWHM$_{0.5}$	TWHM$_{adaptive}$	Naive
2	One	0.773	0.800	0.749
	Two	0.817	0.817	0.780
	Three	0.837	0.837	0.806
	Four	0.824	0.831	0.807
	Overall	0.811	0.822	0.784
5	One	0.789	0.807	0.759
	Two	0.826	0.823	0.779
	Three	0.846	0.849	0.805
	Four	0.833	0.842	0.805
	Overall	0.822	0.829	0.786
8	One	0.795	0.800	0.759
	Two	0.832	0.832	0.778
	Three	0.855	0.845	0.823
	Four	0.831	0.863	0.779
	Overall	0.825	0.831	0.782

our choices of n_{train}. The ω parameter employed in TWHM$_{adaptive}$ is selected as follows. For prediction in each period, we choose the value in a sequence of ω values that produces the highest prediction accuracy in predicting $\mathbf{X}^{n_{train}+i-1}$ for predicting $\mathbf{X}^{n_{train}+i}$. For example, in the first period with $n = 11$ networks, when $n_{train} = 8$, we used $\{\mathbf{X}^t, t = i, \cdots, i+7\}$ to predict $\mathbf{X}^{i+8}$ for $i = 1, 2, 3$. For each i, let $\tilde{\mathbf{X}}^{i+7}$ be defined as in (5.16). A set of candidate values for ω were used to compute $\tilde{\mathbf{X}}^{i+7}$, and the one that returns the smallest misclassification rate (in predicting $\mathbf{X}^{i+7}$) was used in TWHM$_{adaptive}$ for predicting $\mathbf{X}^{i+8}$. The mean of the chosen ω is 0.936 when $n = 2$, 0.895 when $n = 5$, and 0.905 when $n = 8$. The prediction accuracy of the above-mentioned methods, defined as the percentages of correctly predicted links, are reported in Table 6. We can see that TWHM$_{0.5}$ and TWHM$_{adaptive}$ both perform better than the naive approach in all the cases. On the other hand, TWHM coupled with adaptive cut-off points can improve the prediction accuracy of TWHM with a cur-off value 0.5 in most periods.

6. Summary and Discussion.

We have proposed a novel two-way heterogeneity model that utilizes two sets of parameters to explicitly capture static heterogeneity and dynamic heterogeneity. In a high-dimension setup, we have provided the existence and the rate of convergence of its local MLE, and proposed a novel method of moment estimator as an initial value to find this local MLE. To the best of our knowledge, this is the first model in the network literature that the local MLE is obtained for a non-convex loss function. The theory of our model is established by developing new uniform upper bounds for the deviation of the loss function.

While we have focused on the estimation of the parameters in this paper, how to conduct statistical inference for the local MLE is a natural next step for research. In our setup, we assume that the parameters are time invariant but this need not be the case. A future direction is to allow the static heterogeneity parameter $\boldsymbol{\beta}_0$ and/or the dynamic heterogeneity parameter $\boldsymbol{\beta}_1$ to depend on time, giving rise to non-stationary network processes. In case when these parameters change smoothly over time, we may consider estimating the parameters $\beta_{i,0}^{\tau}, \beta_{i,1}^{\tau}$ at time τ by kernel smoothing, that is, by maximizing the following smoothed log-likelihood:

$$
\begin{aligned}
&\tilde{L}(\tau, \mathbf{X}^n, \mathbf{X}^{n-1}, \cdots, \mathbf{X}^1 | \mathbf{X}^0) \\
&= \sum_{t=1}^{n} w_t \sum_{1 \le i < j \le p} \Bigg\{ -\log\left(1 + e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}}\right) + \left(\beta_{i,0} + \beta_{j,0}\right) X_{i,j}^t \left(1 - X_{i,j}^{t-1}\right) \\
&+ \left(1 - X_{i,j}^t\right)\left(1 - X_{i,j}^{t-1}\right) \log\left(1 + e^{\beta_{i,1}+\beta_{j,1}}\right) + X_{i,j}^t X_{i,j}^{t-1} \log\left(e^{\beta_{i,0}+\beta_{j,0}} + e^{\beta_{i,1}+\beta_{j,1}}\right) \Bigg\},
\end{aligned}
$$

with $w_t = \frac{K(h^{-1}|t-\tau|)}{\sum_{t=1}^{n} K(h^{-1}|t-\tau|)}$, where $K(\cdot)$ is a kernel function and h is the bandwidth parameter. As another line of research, note that TWHM is formulated as an AR(1) process. We can extend it by including more time lags. For example, we can extend TWHM to include lag-k dependence by writing

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

where the innovations $\varepsilon_{i,j}^t$ are independent such that

$$
P(\varepsilon_{i,j}^t = r) = \frac{e^{\beta_{i,r}+\beta_{j,r}}}{1 + \sum_{s=0}^{k} e^{\beta_{i,s}+\beta_{j,s}}} \quad \text{for } r = 0, \cdots, k; \quad P(\varepsilon_{i,j}^t = -1) = \frac{1}{1 + \sum_{s=0}^{k} e^{\beta_{i,s}+\beta_{j,s}}},
$$

with parameter $\boldsymbol{\beta}_0 = (\beta_{1,0}, \ldots, \beta_{p,0})^\top$ denoting node-specific static heterogeneity and $\boldsymbol{\beta} = (\beta_{i,r})_{1 \le i \le p; 1 \le r \le k} \in \mathbb{R}^{p \times k}$ denoting lag-k dynamic fluctuation. Other future lines of research include adding covariates to model the tendency of nodes making connections [32], exploring additional structures such as sparsity by adding regularizations to the negative likelihood function [3, 27].

Acknowledgments. Jiang's research is funded by the Hong Kong RGC/GRF grant 15302722. Leng's research is partially supported by EPSRC grant EP/X009505/1. Yao's research is partially supported by EPSRC grants EP/V0075561/1 and EP/X002195/1.

SUPPLEMENTARY MATERIAL

Supplement to "A two-way heterogeneity model for dynamic networks". In this supplemental material, we present the proofs of lemmas, propositions and theorems.

REFERENCES

[1] Bhattacharjee, M., Banerjee, M. and Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *Journal of Machine Learning Research* **21** 1–59.

[2] Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *The Annals of Applied Probability* **21** 1400–1435.

[3] Chen, M., Kato, K. and Leng, C. (2021). Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**.

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

[5] Gragg, W. and Tapia, R. (1974). Optimal error bounds for the Newton–Kantorovich theorem. *SIAM Journal on Numerical Analysis* **11** 10–13.

[6] Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica* **85** 1033–1063.

[7] Han, R., Chen, K. and Tan, C. (2020). Bivariate gamma model. *Journal of Multivariate Analysis* **180** 104666.

[8] Han, R., Xu, Y. and Chen, K. (2023). A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association* **118** 2422–2432.

[9] Hanneke, S., Fu, W. and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic journal of statistics* **4** 585–605.

[10] Hanneke, S. and Xing, E. P. (2007). Discrete temporal models of social networks. In *Statistical network analysis: models, issues, and new directions: ICML 2006 workshop on statistical network analysis, Pittsburgh, PA, USA, June 29, 2006, Revised Selected Papers* 115–125. Springer.

[11] Hillar, C. J., Lin, S. and Wibisono, A. (2012). Inverses of symmetric, diagonally dominant positive matrices and applications. *arXiv preprint arXiv:1203.6812*.

[12] Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76** 33–50.

[13] Jiang, B., Li, J. and Yao, Q. (2023). Autoregressive networks. *Journal of Machine Learning Research* **24** 1–69.

[14] Jin, J. (2015). Fast community detection by score. *The Annals of Statistics* **43** 57–89.

[15] Jin, J., Ke, Z. T., Luo, S. and Wang, M. (2022). Optimal estimation of the number of network communities. *Journal of the American Statistical Association* 1–16.

[16] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E* **83** 016107.

[17] Karwa, V., Slavković, A. et al. (2016). Inference using noisy degrees: Differentially private *beta*-model and synthetic graphs. *The Annals of Statistics* **44** 87–112.

[18] Ke, Z. T. and Jin, J. (2022). The SCORE normalization, especially for highly heterogeneous network and text data. *arXiv preprint arXiv:2204.11097*.

[19] Kolaczyk, E. D. and Csárdi, G. (2020). *Statistical analysis of network data with R* **65**, 2 ed. Springer.

[20] Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **76** 29.

[21] Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society,* **B 79** 1119–1141.

[22] Mersch, D. P., Crespi, A. and Keller, L. (2013). Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science* **340** 1090–1093.

[23] Minai, A. A. and Williams, R. D. (1993). On the derivatives of the sigmoid. *Neural Networks* **6** 845–853.

[24] Newman, M. (2018). *Networks*. Oxford university press.

[25] Pensky, M. (2019). Dynamic network models and graphon estimation. *Annals of Statistics* **47** 2378–2403.

[26] Sengupta, S. and Chen, Y. (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 365–386.

[27] Shao, M., Zhang, Y., Wang, Q., Zhang, Y., Luo, J. and Yan, T. (2021). L-2 Regularized maximum likelihood forβ-model in large and sparse networks. *arXiv preprint arXiv:2110.11856*.

[28] Stein, S. and Leng, C. (2020). A sparse β-model with covariates for networks. *arXiv preprint arXiv:2010.13604*.

[29] Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

[30] Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* 16–28. Springer.

[31] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

[32] Yan, T., Jiang, B., Fienberg, S. E. and Leng, C. (2019). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association* **114** 857–868.

[33] Yan, T., Leng, C. and Zhu, J. (2015). Supplement to "Asymptotics in directed exponential random graph models with an increasing bi-degree sequence". *Annals of Statistics.*

[34] Yan, T., Leng, C., Zhu, J. et al. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Annals of Statistics* **44** 31–57.

[35] Yan, T. and Xu, J. (2012). Approximating the inverse of a balanced symmetric matrix with positive elements. *arXiv preprint arXiv:1202.1058*.

[36] Yan, T. and Xu, J. (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika* **100** 519–524.