Original citation:
Yan, Ting, Jiang, Binyan, Fienberg, Stephen E. and Leng, Chenlei (2018) *Statistical inference in a directed network model with covariates.* Journal of the American Statistical Association. . doi:10.1080/01621459.2018.1448829

Permanent WRAP URL:
http://wrap.warwick.ac.uk/99216

Statistical Inference in a Directed Network Model with Covariates *

Ting Yan† Binyan Jiang‡ Stephen E. Fienberg§ Chenlei Leng¶

Abstract

Networks are often characterized by node heterogeneity for which nodes exhibit different degrees of interaction and link homophily for which nodes sharing common features tend to associate with each other. In this paper, we rigorously study a directed network model that captures the former via node-specific parametrization and the latter by incorporating covariates. In particular, this model quantifies the extent of heterogeneity in terms of outgoingness and incomingness of each node by different parameters, thus allowing the number of heterogeneity parameters to be twice the number of nodes. We study the maximum likelihood estimation of the model and establish the uniform consistency and asymptotic normality of the resulting estimators. Numerical studies demonstrate our theoretical findings and two data analyses confirm the usefulness of our model.

Key words: Asymptotic normality; Consistency; Degree heterogeneity; Directed network; Homophily; Increasing number of parameters; Maximum likelihood estimator.

1 Introduction

Most complex systems involve multiple entities that interact with each other. These interactions are often conveniently represented as networks in which nodes act as entities and a link between two nodes indicates an interaction of some form between the two corresponding entities. The study of networks has attracted increasing attention in a wide variety of fields including social networks (Burt et al., 2013; Lewisa et al., 2012), communication networks (Adamic and Glance, 2005; Diesner and Carley, 2005), biological networks (Bader and Hogue,

*Shortly after finishing the first draft of this paper, we were saddened to hear Steve Fienberg's death. We dedicate this work to his memory.

†Department of Statistics, Central China Normal University, Wuhan, 430079, China. `Email:` tingyanty@mail.ccnu.edu.cn.

‡Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong. `Email:`by.jiang@polyu.edu.hk.

§Department of Statistics, Heinz College, Machine Learning Department, Cylab, Carnegie Mellon University, Pittsburgh, PA 15213, USA. `Email:` fienberg@stat.cmu.edu.

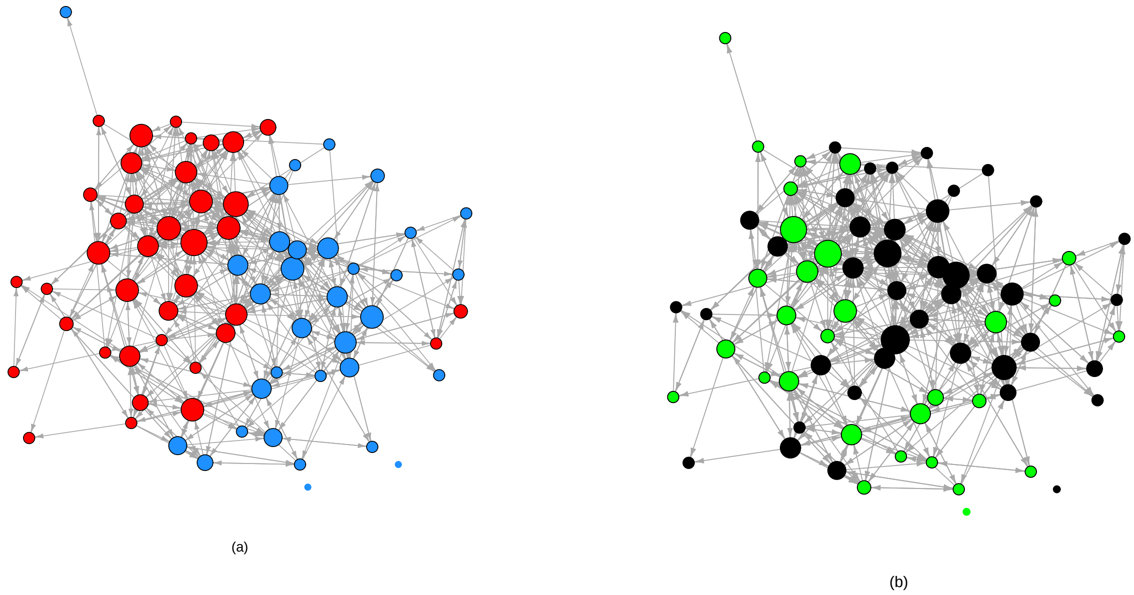
¶Corresponding author. Department of Statistics, University of Warwick and Alan Turing Institute, Coventry, CV4 7AL, UK. `Email:` C.Leng@warwick.ac.uk.

2003; Nepusz et al., 2012), disease transmission networks (Newman, 2002) and so on. Many statistical models have been developed for analyzing networks in the hope to understand their generative mechanism. However, it remains a unique challenge to understand the statistical properties of many network models; for surveys, see Goldenberg et al. (2009), Fienberg (2012), and a book long treatment of networks in Kolaczyk (2009).

Many networks are characterized by two distinctive features. The first is the so-called *degree heterogeneity* for which nodes exhibit different degrees of interaction. In the language of Barabási and Bonabau (2003), a typical network often includes a handful of high degree “hub” nodes having many edges and many low degree individuals having few edges. The second distinctive feature inherent in most natural and synthetic networks is the so-called *homophily* phenomenon for which links tend to form between nodes sharing common features such as age and sex; see, for example, McPherson et al. (2001). As the name suggests, homophily is best explained by node or link specific covariates used to define similarity between nodes. As a concrete example, we examine the directed friendship network between 71 lawyers studied in Lazega (2001) that motivated this paper. The detail of the data can be found in Section 4. As is typical for interactions of this sort, various members’ attributes, including formal status (partner or associate), practice (litigation or corporate) etc., are also collected. A major question of interest is whether and how these covariates influence how ties are formed. Towards this end, we plot the network in Figure 1 using red and blue colors to indicate different statuses in (a) and black and green colors to represent lawyers with different practices in (b). To appreciate the difference in the degrees of connectedness, we use node sizes to represent in-degrees in (a) and out-degrees in (b). This figure highlights a few interesting features. First, there is substantial degree heterogeneity. Different lawyers have different in-degrees and out-degrees, while the in-degrees and the out-degrees of the same lawyers can also be substantially different. This necessitates a model which can characterize the node-specific outgoingness and incomingness. Second, ties seem to form more frequently if the vertices share a common status or a common practice. As a result, a useful model should account for the covariates in order to explain the observed homophily phenomenon.

This paper concerns the study of a generative model for directed networks seen in Figure 1 that addresses node heterogeneity and link homophily simultaneously. Although this model is not entirely new, developing its inference tools is extremely challenging and we have only started to see similar tools for models much simpler when homophily is not considered (Yan et al., 2016). Let’s start by spelling out the model first. Consider a directed graph $\mathcal{G}_n$ on

Figure 1: Visualization of Lazega's friendship network among 71 lawyers. The vertex sizes are proportional to either nodal in-degrees in (a) or out-degrees in (b). The positions of the vertices are the same in (a) and (b). For nodes with degrees less than 5, we set their sizes the same (as a node with degrees 4). In (a), the colors indicate different statuses (red for partner and blue for associate), while in (b), the colors represent different practices (black for litigation and green for corporate).



$n \geq 2$ nodes labeled by $1, \ldots, n$. Let $a_{ij} \in \{0, 1\}$ be an indictor whether there is a directed edge from node i pointing to j. That is, if there is a directed edge from i to j, then $a_{ij} = 1$; otherwise, $a_{ij} = 0$. Denote $A = (a_{ij})_{n \times n}$ as the adjacency matrix of $\mathcal{G}_n$. We assume that there are no self-loops, i.e., $a_{ii} = 0$. Our model postulates that a_{ij}'s follow independent Bernoulli distributions such that a directed link exists from node i to node j with probability

$$P(a_{ij} = 1) = \frac{\exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)}{1 + \exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)}.$$

In this model, the degree heterogeneity of each node is parametrized by two scalar parameters, an incomingness parameter denoted by β_i characterizing how attractive the node is and an outgoingness parameter denoted by α_i illustrating the extent to which the node is attracted to others (Holland and Leinhardt, 1981). The covariate Z_{ij} is either a link dependent vector or a function of node-specific covariates. If X_i denotes a vector of node-level attributes, then these node-level attributes can be used to construct a p-dimensional vector $Z_{ij} = g(X_i, X_j)$, where $g(\cdot, \cdot)$ is a function of its arguments. For instance, if we let $g(X_i, X_j)$ equal to $\|X_i - X_j\|_1$, then it measures the similarity between node i and j features. The vector $\boldsymbol{\gamma}$ is an unknown parameter that characterizes the tendency of two nodes to make a connection. Apparently in our model, a larger $Z_{ij}^\top \boldsymbol{\gamma}$ implies a higher likelihood for node i and j to be connected. For the friendship network in Figure 1, for example, the covariate vector may include two covariates, one indicating whether the two nodes share a common status and the other indicating whether their practices are the same. Though similar models for capturing homophily and degree heterogeneity have been considered by Dzemski (2014) for a general distribution function and Graham (2017) in the undirected case, they focused on the homophily parameter and the inference problem for degree heterogeneity was not studied. Because the formation of networks is not only influenced by external factors (e.g., dyad covariates), but also affected by intrinsic factors (e.g., the strengths of nodes to form network connection), it is statistically interesting to conduct inference on the parameter associated with degree heterogeneity.

Model (1) assumes the independence of the network edges. As pointed out by Graham (2017), the independent assumption may hold in some settings where the drivers of link formation are predominately bilateral in nature, as may be true in some trade networks as well as in models of (some types of) conflict between nation-states.

Since the $n(n-1)$ random variables $a_{i,j}$, $i \neq j$, are mutually independent given the covariates, the probability of observing $\mathcal{G}_n$ is simply

$$\prod_{i,j=1;i\neq j}^{n} \frac{\exp\big((Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)a_{ij}\big)}{1+\exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)} = \exp\big(\sum_{i,j} a_{ij} Z_{ij}^\top \boldsymbol{\gamma} + \boldsymbol{\alpha}^\top \mathbf{d} + \boldsymbol{\beta}^\top \mathbf{b} - C(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\big), \quad (1)$$

where

$$C(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i\neq j} \log\big(1+\exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)\big)$$

is the normalizing constant. Here $d_i = \sum_{j\neq i} a_{ij}$ denotes the out-degree of vertex i and $\mathbf{d} = (d_1, \ldots, d_n)^\top$ is the out-degree sequence of the graph $\mathcal{G}_n$. Similarly, $b_j = \sum_{i\neq j} a_{ij}$ denotes the in-degree of vertex j and $\mathbf{b} = (b_1, \ldots, b_n)^\top$ is the in-degree sequence. The pair $\{\mathbf{b}, \mathbf{d}\}$ or $\{(b_1, d_1), \ldots, (b_n, d_n)\}$ is the so-called bi-degree sequence. As discussed before, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ is a parameter vector tied to the out-degree sequence, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$ is a parameter vector tied to the in-degree sequence, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ is a parameter vector tied to the information of node covariates. Since an out-edge from vertex i pointing to j is the in-edge of j coming from i, it is immediate that the sum of out-degrees is equal to that of in-degrees. If one transforms $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to $(\boldsymbol{\alpha} - c, \boldsymbol{\beta} + c)$, the likelihood does not change. Because of this, for the identifiability of the model, we set $\beta_n = 0$ as in Yan et al. (2016). Since we treat Z_{ij} as observed and the likelihood function (1) is conditional on Z_{ij}'s, we assume that all Z_{ij}'s are bounded. Therefore, the natural parameter space is

$$\Theta = \{(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_{1,\ldots,n-1}^\top, \boldsymbol{\gamma}^\top)^\top : (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_{1,\ldots,n-1}^\top, \boldsymbol{\gamma}^\top)^\top \in R^{2n+p-1}\},$$

under which the normalizing constant is finite.

Because of the form of the model and the independent assumption on the links, it appears that maximum likelihood estimation developed for logistic regression is all that is needed for inference. A major challenge of models of this kind is, however, that the number of parameters grows with the network size. In particular, the number of outgoingness and incomingness parameters needed by our model is already twice the size of the network, and the presence of the covariates poses additional challenges. See the literature review below. To a certain extent, our model can be seen as a special case of the exponential random graph model (ERGM) as discussed by Robins et, al. (2007a,b), as the sufficient statistics are the covariates and the bi-degree sequence. It is known, however, that fitting any nontrivial exponential random graph models is extremely challenging, not to mention developing valid procedures for their statistical inference (Goldenberg et al., 2009; Fienberg, 2012). Studying the asymptotic theory of the proposed directed network model is the main contribution of this paper.

We empirically explore the asymptotic properties of the proposed estimators of the heterogeneity parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as well as the homophily parameter $\boldsymbol{\gamma}$. Our results demonstrate that the empirical study concur with our theoretical findings. Two real data examples are also provided for illustration.

1.1 Literature review

Many network characteristics or configurations can be easily modeled as exponential family distributions on graphs (Robins et, al., 2007a,b). For undirected networks, if we put the node degrees as the sufficient statistics, then the model explains the observed degree heterogeneity but not homophily. This model is referred to as the β-model by Chatterjee et al. (2011). Exploring the properties of the β-model and its generalizations, however, is nonstandard due to an increasing dimension of the parameter space and has attracted much recent interest (Chatterjee et al., 2011; Perry and Wolfe, 2012; Olhede and Wolfe, 2012; Hillar and Wibisono, 2013; Yan and Xu, 2013; Rinaldo et al., 2013; Graham, 2017; Karwa and Slavković, 2016). In particular, Chatterjee et al. (2011) proved the uniform consistency of the maximum likelihood estimator (MLE) and Yan and Xu (2013) derived the asymptotic normality of the MLE. In the directed case, Yan et al. (2016) studied the MLE of a directed version of the β-model which is a special case of the p_1 model by Holland and Leinhardt (1981). Yan et al. (2016) did not consider modelling homophily. By treating the node-specific parameters in the p_1 model as random effects, Van Duijn et al. (2004) proposed a random effects model incorporating nodal covariates. The theoretical properties of the MLE of this model are difficult to establish and thus have not been studied. Fellows and Handcock (2012) generalized exponential random graph models by modeling nodal attributes as random variates. However, the theoretical properties of their model are not explored. Hoff (2009) appears to be among the first to study the model in (1). However, the theoretical properties of Hoff's model are again unknown.

It is also worth noting that the consistency and asymptotic normality of the MLE have been derived for two related models: the Rasch model (Rasch, 1960) for item response experiments (Haberman, 1977) and the Bradley-Terry model (Bradley and Terry, 1952) for paired comparisons by Simons and Yao (1999) in which a growing number of parameters are modelled. The data for an item response experiment can be represented as a bipartite network and for a paired comparisons data as a weighted directed network. None of these papers discussed how to incorporate covariates. Finally, Model (1) can also be represented as a log-linear model (Fienberg and Rinaldo, 2012). Although the necessary and sufficient conditions for the existence of the MLE for log-linear models with arbitrary dimension have been established [e.g., Haberman (1974); Fienberg and Rinaldo (2012)], there is lack of general results on the asymptotic properties of the MLE for high dimensional log-linear models as the analysis would be challenging [Erosheva et al. (2007); Fienberg and Rinaldo (2007, 2012); Rinaldo et al. (2011)].

In the above mentioned network models, the dyads of network edges between two nodes are assumed to be mutually independent. If network configurations such as k-stars and triangles are included as sufficient statistics in the ERGMs, then edges are not independent and such models incur the problem of model degeneracy in the sense of Handcock (2003), in which almost all realized graphs essentially have no edges or are complete, completely skipping all intermediate structures. Chatterjee and Diaconis (2013) have shown that most realizations from many ERGMs look like the results of a simple Erdos-Renyi model and given a first rigorous proof of the degeneracy observed in the ERGM with the counts of edges and triangles as the exclusively sufficient statistics. Yin (2015) further gave an explicit characterization of the degenerate tendency as a function of the parameters. On the other hand, the MLE in ERGMs with dependent structures also incur problematic properties. Shalizi and Rinaldo

(2013) demonstrated that the MLE is not consistent. In order to overcome the mode degeneracy in ERGMs, Schweinberger and Handcock (2015) have proposed local dependent ERGMs by assuming that the graph nodes can be partitioned into K subsets (correspondingly, K subgraphs), in which dependence exists within subgraphs and edges are independence between subgraphs. Based on this assumption, they established a central limit theorem for a network statistic by referring to the Lindeberg–Feller central limit theorem when K goes to infinity and the number of nodes in subgraphs is fixed. The local dependency assumption essentially contains a sequence of independent networks. On the other hand, some refined network statistics such as "alternating k-stars", "alternating k-triangles" and so on in Robins et, al. (2007b) are proposed, but the theoretical properties of the model are still unknown. Moreover, Sadeghi and Rinaldo (2014) formalized the ERGM for the joint degree distributions and derived the condition under which the MLE exists.

The work close to our paper is Graham (2017) in which the β-model was generalized to incorporate covariates to explain the homophily phenomenon and degree heterogeneity for undirected networks. The asymptotic properties of a restricted version of the maximum likelihood estimator were derived under the assumptions that all parameters are bounded and that the estimators for all parameters are taken in one compact set. That is, his results are only applicable to dense networks as pointed out in Graham (2017). In this paper, our focus is on directed networks and our theory is established under more relaxed assumptions. In particular the boundedness assumption on the parameters of degree heterogeneity in Graham (2017) is not needed in our work. Hence our result covers more general networks. In addition, Graham (2017) has focused on the consistency and the asymptotic normality of the parameter estimator associated with covariates, while the asymptotic normality of the heterogeneity parameter estimator was not studied. In this paper, we derive these two properties for the covariate parameter and the heterogeneity parameters in model (1). It is worth remarking that establishing the asymptotic normality for estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is very challenging with the presence of the covariate Z. Graham (2016) further proposed a dynamic model to capture homophily and transitivity when an undirected network over multiple periods is observed. The setup is different from ours in that we only observe one network once. Moreover, Jochmans (2017) developed a conditional-likelihood based approach to estimate the homophily parameter by constructing a quadruple sufficient statistic to eliminate the degree heterogeneity parameter, and further established the consistency and asymptotic normality of the resulting estimator.

To some extent, our network model is connected to the longitudinal panel data model considered by Fernández-Val and Weidner (2016) and Cruz-Gonzalez et al. (2017) where time and individual fixed effects are both considered. They focused mainly on the homophily parameter. Dzemski (2017) applied the method in Fernández-Val and Weidner (2016) to a network model similar to ours by including a scalar parameter to characterize the correlation of dyads. A two-step approach was used for estimation and again the focus is on the homophily parameter. There are major differences between these papers and ours including the methods of proofs, the conditions required by the theorems and the attention to the degree parameters. We will clarify these points after stating our main results in Section 3.

For the remainder of the paper, we proceed as follows. In Section 2, we give the details on the model considered in this paper. In section 3, we establish asymptotic results. Numerical studies are presented in Section 4. We provide further discussion and future work in Section 5. All proofs are relegated to the appendix.

2 Maximum Likelihood Estimation

We first introduce some notations. Let $\mathbb{R} = (-\infty, \infty)$ be the real domain. For a subset $C \subset \mathbb{R}^n$, let C^0 and $\overline{C}$ denote the interior and closure of C, respectively. For convenience, let $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_{n-1})^\top$ and $\mathbf{g} = (d_1, \ldots, d_n, b_1, \ldots, b_{n-1})^\top$. Sometimes, we use $\boldsymbol{\theta}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ interchangeably. For a vector $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, denote by $\|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|$ the ℓ_∞-norm of $\mathbf{x}$. For an $n \times n$ matrix $J = (J_{ij})$, let $\|J\|_\infty$ denote the matrix norm induced by the ℓ_∞-norm on vectors in $\mathbb{R}^n$, i.e.

$$\|J\|_\infty = \max_{\mathbf{x}\neq 0} \frac{\|J\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1\le i\le n} \sum_{j=1}^n |J_{ij}|.$$

The notation $i < j < k$ is a shorthand for $\sum_{i=1}^n \sum_{j=i+1}^n \sum_{k=j+1}^n$. A "$*$" superscript on a parameter denotes its true value and may be omitted when doing so causes no confusion.

In what follows, it is convenient to define the notation:

$$p_{ij}(\boldsymbol{\gamma}, \alpha_i, \beta_j) = \frac{\exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)}{1 + \exp(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j)}.$$

The log-likelihood of observing a directed network $\mathcal{G}_n$ under model (1) is

$$\begin{aligned} \ell(\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i\neq j}\{a_{ij}\log p_{ij}(\boldsymbol{\gamma},\alpha_i,\beta_j)+(1-a_{ij})\log(1-p_{ij}(\boldsymbol{\gamma},\alpha_i,\beta_j))\} \\ &= \sum_{i\neq j}a_{ij}Z_{ij}^\top\boldsymbol{\gamma}+\sum_{i=1}^n\alpha_i d_i+\sum_{j=1}^n\beta_j b_j-\sum_{i\neq j}\log(1+e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_j}). \end{aligned} \tag{2}$$

The score equations for the vector parameters $\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta}$ are easily seen as

$$\begin{aligned} \sum_{i\neq j}a_{ij}Z_{ij} &= \sum_{i\neq j}\frac{Z_{ij}e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_j}}{1+e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_j}}, \\ d_i &= \sum_{k=1,k\neq i}^n\frac{e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_k}}{1+e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_k}}, \quad i=1,\ldots,n, \\ b_j &= \sum_{k=1,k\neq j}^n\frac{e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_k+\beta_j}}{1+e^{Z_{ij}^\top\boldsymbol{\gamma}+\alpha_k+\beta_j}}, \quad j=1,\ldots,n-1. \end{aligned} \tag{3}$$

The MLEs of the parameters are the solution of the above equations if they exist. Let $\mathcal{K}$ be the convex hull of the set $\{(\mathbf{d}^\top,\mathbf{b}_{1,\ldots,n-1}^\top,\sum_{i,j}a_{ij}Z_{ij}^\top)^\top : a_{ij}\in\{0,1\},1\le i\neq j\le n\}$. Since the function $C(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ is steep and regularly strictly convex, the MLE of $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ exists if and only if $(\mathbf{d}^\top,\mathbf{b}_{1,\ldots,n-1}^\top,\sum_{i,j}a_{ij}Z_{ij}^\top)^\top$ lies in the interior of $\mathcal{K}$ [see, e.g., Theorem 5.5 in Brown (1986) (p. 148)]. When the number of nodes n is small, we can simply use the R function "glm" to solve (3). For relatively large n, this is no longer feasible as it is memory demanding to store the design matrix needed for α and β. In this case, we recommend the use of a two-step iterative algorithm by alternating between solving the second and third equations in (3) via the fixed point method in Yan et al. (2016) and solving the first equation in (3) via some existing algorithm for generalized linear models.

In this paper, we assume that p, the dimension of Z, is fixed and that the support of Z_{ij} is $\mathbb{Z}^p$, where $\mathbb{Z}$ is a compact subset of $\mathbb{R}$. For example, if Z_{ij}'s are indictor variables such as sex, then the assumption holds. For the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we make no such assumption and allow them to diverge slowly with n, the network size. To be precise, as long as $\|\boldsymbol{\theta}^*\|_\infty$, the maximum entry of the true heterogeneity parameter, is bounded by a number proportional to $\log n$, our theory holds. See Theorem 1 for example. For technical reasons, it is more convenient to work with the following restricted maximum likelihood estimators of $\boldsymbol{\alpha},\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ defined as

$$(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}})=\arg\max_{\boldsymbol{\gamma}\in\Gamma,\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^{n-1}}\ell(\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{4}$$

where Γ is a compact subset of $\mathbb{R}^p$ and $\widehat{\boldsymbol{\gamma}}=(\hat{\gamma}_1,\ldots,\hat{\gamma}_p)^\top$, $\widehat{\boldsymbol{\alpha}}=(\hat{\alpha}_1,\ldots,\hat{\alpha}_n)^\top$, $\widehat{\boldsymbol{\beta}}=(\hat{\beta}_1,\ldots,\hat{\beta}_{n-1})^\top$ are the respective restricted MLEs of $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $\hat{\beta}_n=0$. Write $\widehat{\boldsymbol{\theta}}=(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}})^\top$. Let $\tilde{\mathcal{K}}$ be the convex hull of the set constructed by all graphical bi-degree sequence $(\mathbf{d}^\top, \mathbf{b}_{1,\ldots,n-1}^\top)^\top$ and write $(\widehat{\boldsymbol{\alpha}}(\boldsymbol{\gamma}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma})) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. For every fixed $\gamma \in \Gamma$, by Theorem 5.5 in Brown (1986) (p. 148), the MLE $(\widehat{\boldsymbol{\alpha}}(\boldsymbol{\gamma}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma}))$ exists if and only if $(\mathbf{d}^\top, \mathbf{b}_{1,\ldots,n-1}^\top)^\top$ lies in the interior of $\tilde{\mathcal{K}}$. Since Γ is a compact set, the restricted MLE exists if and only if $(\mathbf{d}^\top, \mathbf{b}_{1,\ldots,n-1}^\top)^\top$ lies in the interior of $\tilde{\mathcal{K}}$.

If $\widehat{\boldsymbol{\gamma}}$ lies in the interior of Γ, then it is also the global MLE of $\boldsymbol{\gamma}$. Since we assume the dimension of Z_{ij} is fixed and $\boldsymbol{\gamma}$ is one common parameter vector, it seems reasonable to assume that $\|\boldsymbol{\gamma}\|$ is bounded by a constant. If the restricted MLEs of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ exist, they would satisfy the second and third equations in (3). If $\widehat{\boldsymbol{\gamma}} \in \Gamma^0$, then it satisfies the first equation in (3). Hereafter, we will work with the MLE defined in (4) and use "MLE" to denote "restricted MLE" for shorthand.

3 Theoretical Properties

3.1 Characterization of the Fisher information matrix

The Fisher information matrix is a key quantity in the asymptotic analysis as it measures the amount of information that a random variable carries about an unknown parameter of a distribution that models the random variable. In order to characterize this matrix for the vector parameter $\boldsymbol{\theta}$ in our model (1), we introduce a general class of matrices that encompass the Fisher matrix. Given two positive numbers m and M with $M \geq m > 0$, we say the $(2n-1) \times (2n-1)$ matrix $V = (v_{i,j})$ belongs to the class $\mathcal{L}_n(m, M)$ if the following holds:

$$
\begin{aligned}
& m \leq v_{i,i} - \textstyle\sum_{j=n+1}^{2n-1} v_{i,j} \leq M, \quad i = 1, \ldots, n-1; \quad v_{n,n} = \textstyle\sum_{j=n+1}^{2n-1} v_{n,j}, \\
& v_{i,j} = 0, \quad i, j = 1, \ldots, n, \ i \neq j, \\
& v_{i,j} = 0, \quad i, j = n+1, \ldots, 2n-1, \ i \neq j, \\
& m \leq v_{i,j} = v_{j,i} \leq M, \quad i = 1, \ldots, n, \ j = n+1, \ldots, 2n-1, \ j \neq n+i, \\
& v_{i,n+i} = v_{n+i,i} = 0, \quad i = 1, \ldots, n-1, \\
& v_{i,i} = \textstyle\sum_{k=1}^{n} v_{k,i} = \textstyle\sum_{k=1}^{n} v_{i,k}, \quad i = n+1, \ldots, 2n-1.
\end{aligned}
\tag{5}
$$

Clearly, if $V \in \mathcal{L}_n(m, M)$, then V is a $(2n-1) \times (2n-1)$ diagonally dominant, symmetric nonnegative matrix and V has the following structure:

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^\top & V_{22} \end{pmatrix},$$

where $V_{11} \in \mathbb{R}^{n \times n}$ and $V_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$ are diagonal matrices, V_{12} is a nonnegative matrix whose non-diagonal elements are positive and diagonal elements equal to zero. One can easily show that the Fisher information matrix for the vector parameter $\boldsymbol{\theta}$ belongs to $\mathcal{L}_n(m, M)$ for any $\boldsymbol{\gamma} \in \Gamma$. The exact form of this matrix can be found after Theorem 3 in Section 3.2. Thus, with some abuse of notation, we use V to denote the Fisher information matrix for the vector parameter $\boldsymbol{\theta}$ in the model (1).

Define $v_{2n,i} = v_{i,2n} := v_{i,i} - \sum_{j=1;j\neq i}^{2n-1} v_{i,j}$ for $i = 1, \ldots, 2n-1$ and $v_{2n,2n} = \sum_{i=1}^{2n-1} v_{2n,i}$. Then $m \leq v_{2n,i} \leq M$ for $i = 1, \ldots, n-1$, $v_{2n,i} = 0$ for $i = n, n+1, \ldots, 2n-1$ and $v_{2n,2n} = \sum_{i=1}^{n} v_{i,2n} = \sum_{i=1}^{n} v_{2n,i}$. Because of the special structure of any matrix $V \in \mathcal{L}_n(m, M)$, Yan et al. (2016) proposed to approximate its inverse V^{-1} by the matrix $S = (s_{i,j})$, which is defined as

$$s_{i,j} = \begin{cases} \frac{\delta_{i,j}}{v_{i,i}} + \frac{1}{v_{2n,2n}}, & i, j = 1, \ldots, n, \\ -\frac{1}{v_{2n,2n}}, & i = 1, \ldots, n, \quad j = n+1, \ldots, 2n-1, \\ -\frac{1}{v_{2n,2n}}, & i = n+1, \ldots, 2n-1, \quad j = 1, \ldots, n, \\ \frac{\delta_{i,j}}{v_{i,i}} + \frac{1}{v_{2n,2n}}, & i, j = n+1, \ldots, 2n-1, \end{cases} \tag{6}$$

where $\delta_{i,j} = 1$ when $i = j$ and $\delta_{i,j} = 0$ when $i \neq j$. They established an upper bound on the approximation errors, stated in the lemma below.

Lemma 1. *If $V \in \mathcal{L}_n(m, M)$ with $M/m = o(n)$, then for large enough n,*

$$\|V^{-1} - S\| \leq \frac{c_1 M^2}{m^3 (n-1)^2},$$

where c_1 is a constant that does not depend on M, m and n, and $\|A\| := \max_{i,j} |a_{i,j}|$ for a general matrix $A = (a_{i,j})$.

This lemma provides an accurate approximation of the inverse of the Fisher information matrix of $\boldsymbol{\theta}$ that has a close-form expression. As used throughout our theoretical development, this close-form expression greatly facilitates analytical calculations and makes the covariance matrix in the limiting distribution of the MLE be explicit.

3.2 Asymptotic results

We first establish the existence and consistency of $\widehat{\boldsymbol{\theta}}$. The main idea of the proof is as follows. For every fixed $\boldsymbol{\gamma} \in \Gamma$, we define a system of functions

$$\begin{aligned} F_{\gamma,i}(\boldsymbol{\theta}) &= d_i - \sum_{k=1;k\neq i}^{n} \frac{e^{Z_{ij}^\top \boldsymbol{\gamma}+\alpha_i+\beta_k}}{1+e^{Z_{ij}^\top \boldsymbol{\gamma}+\alpha_i+\beta_k}}, \quad i=1,\ldots,n, \\ F_{\gamma,n+j}(\boldsymbol{\theta}) &= b_j - \sum_{k=1;k\neq j}^{n} \frac{e^{Z_{ij}^\top \boldsymbol{\gamma}+\alpha_k+\beta_j}}{1+e^{Z_{ij}^\top \boldsymbol{\gamma}+\alpha_k+\beta_j}}, \quad j=1,\ldots,n, \\ F_{\gamma}(\boldsymbol{\theta}) &= (F_{\gamma,1}(\boldsymbol{\theta}),\ldots,F_{\gamma,2n-1}(\boldsymbol{\theta}))^\top, \end{aligned} \tag{7}$$

which are just the score equations for $\boldsymbol{\theta}$ with $\boldsymbol{\gamma}$ fixed. Then we construct a Newton's iterative sequence $\{\boldsymbol{\theta}^{(k+1)}\}_{k=0}^{\infty}$ with initial value $\boldsymbol{\theta}^{(0)}$, where $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - [F'(\boldsymbol{\theta}^{(k)})]^{-1}F(\boldsymbol{\theta}^{(k)})$. If the iterative converges, then the solution lies in the neighborhood of $\boldsymbol{\theta}_0$. This is done by establishing a geometrically fast convergence rate of the algorithm with the initial value as the true value. This technique is also used in Yan et al. (2016). We first present the consistency of the MLE $\widehat{\boldsymbol{\theta}}$ for estimating $\boldsymbol{\theta}$ in the following theorem, whose proof is given in the supplementary material.

Theorem 1. *Assume that $\boldsymbol{\gamma}^* \in \Gamma^0$ and $\boldsymbol{\theta}^* \in \mathbb{R}^{2n-1}$ with $\|\boldsymbol{\theta}^*\|_\infty \le \tau \log n$, where $0<\tau<1/24$ is a constant, and that $A \sim \mathbb{P}_{\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*}$, where $\mathbb{P}_{\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*}$ denotes the probability distribution (1) on A under the parameters $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$. Then as n goes to infinity, with probability approaching one, the MLE $\widehat{\boldsymbol{\theta}}$ exists and satisfies*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty = O_p\left(\frac{(\log n)^{1/2} e^{8\|\boldsymbol{\theta}^*\|_\infty}}{n^{1/2}}\right) = o_p(1).$$

Further, if $\widehat{\boldsymbol{\theta}}$ exists, it is unique.

In order to prove the consistency of $\widehat{\boldsymbol{\gamma}}$, we define a profile likelihood

$$\ell^c(\boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}(\boldsymbol{\gamma})) = \sum_{i\neq j} a_{ij} Z_{ij}^\top \boldsymbol{\gamma} + \sum_{i=1}^{n} \alpha_i(\boldsymbol{\gamma}) d_i + \sum_{j=1}^{n} \beta_j(\boldsymbol{\gamma}) b_j + \sum_{i\neq j} \log(1+e^{Z_{ij}^\top \boldsymbol{\gamma}+\alpha_i(\boldsymbol{\gamma})+\beta_j(\boldsymbol{\gamma})}), \tag{8}$$

where $\widehat{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\gamma}, \boldsymbol{\theta})$. It is easy to show that

$$\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})] = -\sum_{i\neq j} D_{KL}(p_{ij} \| p_{ij}(\boldsymbol{\gamma}, \alpha_i, \beta_j)) - \sum_{i\neq j} S(p_{ij}), \tag{9}$$

where

$$D_{KL}(p_{ij} \| p_{ij}(\boldsymbol{\gamma}, \alpha_i, \beta_j)) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{ij}(\boldsymbol{\gamma}, \alpha_i, \beta_j)}$$

is the Kullback-Leibler divergence of $p_{ij}(\boldsymbol{\gamma}, \alpha_i, \beta_j)$ from $p_{ij} := p_{ij}(\boldsymbol{\gamma}^*, \alpha_i^*, \beta_j^*)$ and $S(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy function. Since the Kullback-Leibler distance is nonnegative, the function (9) attains its maximum value when $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. On the other hand, since p_{ij} is a monotonic function on its arguments, $(\boldsymbol{\gamma}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is a unique maximizer of the function $\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})]$. The main idea of proving the consistency of $\widehat{\boldsymbol{\gamma}}$ is to show that $n^{-2}|\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})]|$ is small in contrast with the magnitude of $n^{-2}\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})]$, then the MLE approximately attains at the maximum of the function $\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})]$. The consistency of $\widehat{\boldsymbol{\gamma}}$ is stated formally below, whose proof is given in Section 6.1.

Theorem 2. *Assume that $\boldsymbol{\gamma}^* \in \Gamma^0$ and $\|\boldsymbol{\theta}^*\|_\infty \le \tau \log n$, where $0 < \tau < 1/24$ is a constant, and that $A \sim \mathbb{P}_{\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*}$. Then as n goes to infinity, we have*

$$\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*.$$

Next, we establish asymptotic normality of $\widehat{\boldsymbol{\theta}}$, whose proof is given in the supplementary mateiral. This is done by approximately representing $\widehat{\boldsymbol{\theta}}$ as a function of $\mathbf{g} = (d_1, \ldots, d_n, b_1, \ldots, b_{n-1})^\top$ with an explicit expression.

Theorem 3. *Assume that $\boldsymbol{\gamma}^* \in \Gamma^0$ and $A \sim \mathbb{P}_{\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*}$. If $\|\boldsymbol{\theta}^*\|_\infty \le \tau \log n$, where $\tau \in (0, 1/44)$ is a constant, then for any fixed $k \ge 1$, as $n \to \infty$, the vector consisting of the first k elements of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically multivariate normal with mean $\mathbf{0}$ and covariance matrix given by the upper left $k \times k$ block of S defined in* (6).

Remark 1. By Theorem 3, for any fixed i, as $n \to \infty$, the convergence rate of $\hat{\theta}_i$ is $1/v_{i,i}^{1/2}$, whose magnitude is between $O(n^{-1/2} e^{\|\boldsymbol{\theta}^*\|_\infty})$ and $O(n^{-1/2})$ by inequality (6) in the supplementary material.

Now we provide the exact form of V, the Fisher information matrix of the vector parameter $\boldsymbol{\theta}$. For $i = 1, \ldots, n$,

$$v_{i,l} = 0, \quad l = 1, \ldots, n, \ l \neq i; \quad v_{i,i} = \sum_{k=1; k \neq i}^{n} \frac{e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_k}}{(1 + e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_k})^2},$$

$$v_{i,n+j} = \frac{e^{Z_{ij}^\top \gamma + \alpha_i + \beta_j}}{(1 + e^{Z_{ij}^\top \gamma + \alpha_i + \beta_j})^2}, \quad j = 1, \ldots, n-1, \ j \neq i; \quad v_{i,n+i} = 0$$

and for $j = 1, \ldots, n-1$,

$$v_{n+j,i} = \frac{e^{Z_{ij}^\top \gamma + \alpha_l + \beta_j}}{(1 + e^{Z_{ij}^\top \gamma + \alpha_l + \beta_j})^2}, \quad l = 1, \ldots, n, \ l \neq j; \quad v_{n+j,j} = 0,$$

$$v_{n+j,n+j} = \sum_{k=1;k\neq j}^{n} \frac{e^{Z_{ij}^\top \gamma + \alpha_k + \beta_j}}{(1 + e^{Z_{ij}^\top \gamma + \alpha_k + \beta_j})^2}, \quad v_{n+j,i} = 0, \quad i = 1, \ldots, n-1.$$

Let H be the Hessian matrix of the log-likelihood function $\ell(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ in (2) which can be represented as

$$H = \begin{pmatrix} H_{\gamma\gamma} & H_{\gamma\theta} \\ H_{\gamma\theta}^\top & -V \end{pmatrix}.$$

Following Amemiya (1985) (p. 126), the Hessian matrix of $\ell^c(\boldsymbol{\gamma}^*, \hat{\theta}(\boldsymbol{\gamma}^*))$ is $H_{\gamma\gamma} + H_{\gamma\theta}V^{-1}H_{\gamma\theta}^\top$. To state the form of the limit distribution of $\hat{\boldsymbol{\gamma}}$, define

$$I_n(\boldsymbol{\gamma}^*) = -\frac{1}{n(n-1)} \frac{\partial^2 \ell^c(\boldsymbol{\gamma}^*, \hat{\theta}(\boldsymbol{\gamma}^*))}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top} = \frac{1}{n(n-1)}(-H_{\gamma\gamma} - H_{\gamma\theta}V^{-1}H_{\gamma\theta}^\top), \tag{10}$$

whose approximate expression is given in (20), and $I_*(\boldsymbol{\gamma})$ as the limit of $I_n(\boldsymbol{\gamma}^*)$ as n goes to infinity.

Theorem 4. *Assume that $\boldsymbol{\gamma}^* \in \Gamma^0$ and $\boldsymbol{\theta}^* \in \mathbb{R}^{2n-1}$ with $\|\boldsymbol{\theta}^*\|_\infty \leq \tau \log n$, where $0 < \tau < 1/24$ is a constant, and that $A \sim \mathbb{P}_{\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*}$. Then as n goes to infinity, the p-dimensional vector $N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ is asymptotically multivariate normal distribution with mean $I_*^{-1}(\boldsymbol{\gamma})B_*$ and covariance matrix $I_*^{-1}(\boldsymbol{\gamma})$, where $N = n(n-1)$ and B_* is the bias term given in (24).*

Remark 2. The limiting distribution of $\widehat{\boldsymbol{\gamma}}$ is involved with a bias term

$$\mu_* = \frac{I_*^{-1}(\boldsymbol{\gamma})B_*}{\sqrt{n(n-1)}}.$$

If all parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are bounded, then $\mu_* = O(n^{-1/2})$. It follows that $B_* = O(1)$ and $(I_*)_{i,j} = O(1)$ according to their expressions. Since the MLE $\widehat{\boldsymbol{\gamma}}$ is not centered at the true parameter value, the confidence intervals and the p-values of hypothesis testing constructed from $\widehat{\boldsymbol{\gamma}}$ cannot achieve the nominal level without bias-correction under the null: $\boldsymbol{\gamma}^* = 0$. This is referred to as the so-called incidental parameter problem in econometric literature [Neyman and Scott (1984); Fernández-Val and Weidner (2016); Dzemski (2017)]. The produced bias is due to the appearance of additional parameters. Here, we propose to use the analytical bias correction formula: $\widehat{\gamma}_{bc} = \widehat{\gamma} - \hat{I}^{-1}\hat{B}/\sqrt{n(n-1)}$, where $\hat{I}$ and $\hat{B}$ are the estimates of I_* and B_* by replacing $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ in their expressions with their MLEs $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\theta}}$, respectively. Dzemski (2014) also used this bias correction procedure, but his expression depends on projected values of pair-wise covariates into the space spanned by degree parameters α_i and β_j under a weighted least square problem and is not explicit. In the simulation in next section, we can see that the correction formula offer dramatically improvements over uncorrected estimates and exhibit the corrected coverage probabilities, in which those for uncorrected estimates are below the nominal level evidently. See also Hahn and Newey (2004) and Fernández-Val and Weidner (2016) for Jackknife bias correction for nonlinear panel models. But as discussed in Dzemski (2014), this method is difficult to implement for network models. Moreover, Graham (2017) described an iterated bias correction procedure, which may be numerically unstable and is not guaranteed to converge as demonstrated in Juodis (2013).

Remark 3. There are three main differences between the results in Fernández-Val and Weidner (2016) and those in our paper. First, for proving their asymptotic results, Fernández-Val and Weidner (2016) used a projection method by projecting the pairwise covariates into the space spanned by degree parameters α_i and β_j as a weighted least squares problem, while we use an elementary method by approximating the inverse matrix of the Fisher information of the degree parameters via an analytical expression. As a result, the asymptotic variances of the estimators in Fernández-Val and Weidner (2016) depend on projected values not having closed form expressions, while ours are explicit and easier to compute. We also note that the matrix to approximate the inverse of the incidental parameter Hessian in Fernández-Val and Weidner (2016) is diagonal while ours is not. Second, the asymptotic distribution of the MLE of the incident parameters in α_i and β_j is not addressed in Fernández-Val and Weidner (2016). Note that the properties of the incidental parameter estimators are more challenging than the fixed dimensional parameter γ due to their increasing dimensions. Third, Fernández-Val and Weidner (2016) assumed that all parameters are bounded while we consider an asymptotic setting to allow the upper bound of the degree parameter to increase as the size of a network grows.

4 Numerical Studies

In this section, we evaluate the asymptotic results of the MLEs for model (1) through simulation studies and a real data example.

4.1 Simulation studies

Similar to Yan et al. (2016), the parameter values take a linear form. Specifically, we set $\alpha_{i+1}^* = (n-1-i)L/(n-1)$ for $i = 0, \ldots, n-1$ and let $\beta_i^* = \alpha_i^*$, $i = 1, \ldots, n-1$ for simplicity. By default, $\beta_n^* = 0$. We considered four different values for L as $L \in \{0, \log(\log n), (\log n)^{1/2}, \log n\}$. By allowing the true value of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to grow with n, we intended to assess the asymptotic properties under different asymptotic regimes. Similar to Graham (2017) and Dzemski (2014), each element of the p-dimensional node-specific covariate X_i is independently generated from a $Beta(2,2)$ distribution. The difference is that their papers considered $p = 1$ while in this paper we set $p = 2$ by letting $Z_{ij} = (|X_{i1} - X_{j1}|, |X_{i2} - X_{j2}|)^\top$. For the parameter $\boldsymbol{\gamma}^*$, we let it be $(1, 1.5)^\top$. Thus, the homophily effect of the network is determined by a weighted sum of the similarity measures of the two covariates between two nodes.

Note that by Theorems 3, $\hat{\xi}_{i,j} = [\hat{\alpha}_i - \hat{\alpha}_j - (\alpha_i^* - \alpha_j^*)]/(1/\hat{v}_{i,i} + 1/\hat{v}_{j,j})^{1/2}$, $\hat{\zeta}_{i,j} = (\hat{\alpha}_i + \hat{\beta}_j - \alpha_i^* - \beta_j^*)/(1/\hat{v}_{i,i} + 1/\hat{v}_{n+j,n+j})^{1/2}$, and $\hat{\eta}_{i,j} = [\hat{\beta}_i - \hat{\beta}_j - (\beta_i^* - \beta_j^*)]/(1/\hat{v}_{n+i,n+i} + 1/\hat{v}_{n+j,n+j})^{1/2}$ are all asymptotically distributed as standard normal random variables, where $\hat{v}_{i,i}$ is the estimate of $v_{i,i}$ by replacing $(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$ with $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}})$. Therefore, we assess the asymptotic normality of $\hat{\xi}_{i,j}$, $\hat{\zeta}_{i,j}$ and $\hat{\eta}_{i,j}$ using the quantile-quantile (QQ) plot. Further, we also record the coverage probability of the 95% confidence interval, the length of the confidence interval, and the frequency that the MLE does not exist. The results for $\hat{\xi}_{i,j}$, $\hat{\zeta}_{i,j}$ and $\hat{\eta}_{i,j}$ are similar, thus only the results of $\hat{\xi}_{i,j}$ are reported. The average and median values of $\widehat{\boldsymbol{\gamma}}$ are also reported. Finally, each simulation is repeated $10,000$ times.

We simulated networks with $n = 100$ or $n = 200$ and found that the QQ-plots for these two network sizes were similar. Therefore, we only show the QQ-plots for $n = 200$ in Figure 2 to save space. In this figure, the horizontal and vertical axes are the theoretical and empirical quantiles, respectively, and the straight lines correspond to the reference line $y = x$. In Figure 2, when $L = 0$ and $\log(\log n)$, the empirical quantiles coincide well with the theoretical ones, while there are notable deviations when $L = (\log n)^{1/2}$. When $L = \log n$, the MLE did not exist in all repetitions (see Table 1, thus the corresponding QQ plot could not be shown).

Table 1 reports the coverage probability of the 95% confidence interval for $\alpha_i - \alpha_j$, the length of the confidence interval as well as the frequency that the MLE did not exist. As we can see, the length of the confidence interval increases as L increases and decreases as n increases, which qualitatively agrees with the theory. The coverage frequencies are all close to the nominal level when $L = 0$ or $\log(\log n)$, while when $L = (\log n)^{1/2}$, the MLE often does

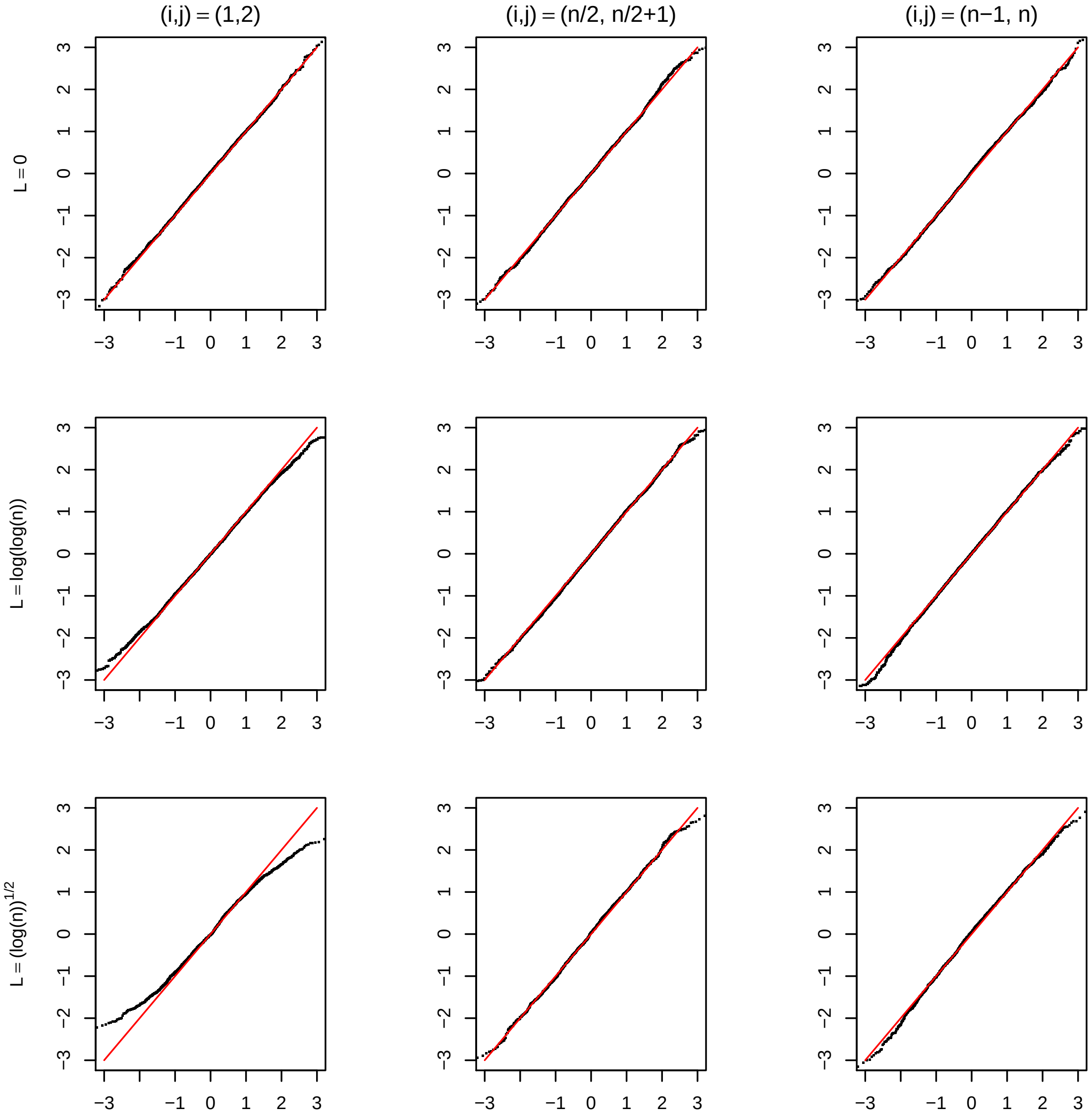


Figure 2: The QQ plots of $\hat{v}_{ii}^{1/2}(\hat{\theta}_i - \theta_i)$.

not exist and the coverage frequencies for pair $(1, 2)$ are higher than the nominal level; when L is $\log n$, the MLE did not exist for all repetitions.

Table 1: The reported values are the coverage frequency ($\times 100\%$) for $\alpha_i - \alpha_j$ for a pair (i, j) / the length of the confidence interval / the frequency ($\times 100\%$) that the MLE did not exist.

n	(i, j)	$L = 0$	$L = \log(\log n)$	$L = (\log n)^{1/2}$	$L = \log n$
100	$(1, 2)$	94.82/1.20/0	97.02/2.62/0	99.80/3.80/90.04	$NA/NA/100$
	$(50, 51)$	94.76/1.20/0	95.79/1.86/0	96.98/2.37/90.04	$NA/NA/100$
	$(99, 100)$	94.84/1.20/0	95.21/1.44/0	96.38/1.57/90.04	$NA/NA/100$
200	$(1, 2)$	95.18/0.84/0	96.31/1.96/0	98.64/3.05/45.08	$NA/NA/100$
	$(100, 101)$	94.33/0.84/0	94.88/1.36/0	94.99/1.72/45.08	$NA/NA/100$
	$(199, 200)$	95.08/0.84/0	94.78/1.02/0	94.95/1.12/45.08	$NA/NA/100$

Table 2 reports the coverage probabilities for the estimate $\widehat{\gamma}$ and bias correction estimate

$\widehat{\gamma}_{bc}(=\widehat{\gamma}-\hat{I}^{-1}\hat{B}/\sqrt{n(n-1)})$ at the nominal level 95%, the average absolute bias as well as the standard error. As we can see, the coverage frequencies for the uncorrected estimate is visibly below the nominal level with at least 10 percentage points and the bias correction estimate dramatically improve the coverage frequencies, whose coverage frequencies are close to the nominal level when the MLE exists with a high frequency. On the other hand, when n is fixed, the average absolute bias of $\widehat{\gamma}$ increases as L becomes larger and so is the standard error.

Table 2: The reported values are the coverage frequency ($\times 100\%$) for γ_i for i / average bias / length of confidence interval /the frequency ($\times 100\%$) that the MLE did not exist ($\boldsymbol{\gamma}^* = (1, 1.5)^\top$).

n	$\widehat{\gamma}$	$L=0$	$L=\log(\log n)$	$L=(\log n)^{1/2}$	$L=\log n$
100	$\hat{\gamma}_1$	80.78/0.18/0.56/0	5.80/0.81/0.84/0	0.20/1.28/1.02/90.04	NA
	$\hat{\gamma}_{bc,1}$	94.28/0.18/0.56/0	94.56/0.81/0.84/0	94.76/1.28/1.02/90.04	NA
	$\hat{\gamma}_2$	81.14/0.19/0.57/0	7.31/0.80/0.85/0	1.41/1.26/1.04/90.04	NA
	$\hat{\gamma}_{bc,2}$	94.14/0.19/0.57/0	94.56/0.80/0.85/0	93.76/1.26/1.04/90.04	NA
200	$\hat{\gamma}_1$	81.23/0.04/0.28/0	3.69/0.22/0.43/0	0.34/0.34/0.52/45.08	NA
	$\hat{\gamma}_{bc,1}$	95.22/0.04/0.28/0	94.37/0.22/0.43/0	96.19/0.34/0.52/45.08	NA
	$\hat{\gamma}_2$	81.05/0.05/0.28/0	4.14/0.22/0.44/0	0.69/0.34/0.52/45.08	NA
	$\hat{\gamma}_{bc,2}$	94.38/0.05/0.28/0	94.75/0.22/0.44/0	95.33/0.34/0.53/45.08	NA

4.2 Two data examples

The analysis of a Lazega's dataset. We first analyze Lazega's datasets of lawyers (Lazega, 2001), downloaded from `https://www.stats.ox.ac.uk/~snijders/siena/Lazega_lawyers_data.htm`. This data set comes from a network study of corporate law partnership that was carried out in a Northeastern US corporate law firm between 1988 and 1991 in New England. We focus on the friendship network among the 71 attorneys including partners and associates of this firm. These attorneys were asked to name attorneys whom they socialized with outside work. Naturally for a network of this sort, many covariates of each attorney were collected. In particular, the collected covariates at the node level include formal status (partner or associate); gender (man or woman), location in which they worked (Boston, Hartford, or Providence), years with the firm, age, practice (litigation or corporate) and law school attended (harvard and yale, or ucon, or others). We define the covariate for each dyad as a 7 dimensional vector consisting of the differences between these 7 variables of the two individuals, where for categorical variables, the difference is defined as the indicator whether they are equal, and for continuous variable, the difference indicates their absolute distance. The directed graph of this data set is shown in Figure 1 where colors indicate either different status in (a) or different practice in (b). Although it may deem appropriate to treat the friendship relationship as undirected, from Figure 1, we can see that the numbers of outgoing and incoming connections for many individuals are dramatically different. As a result, we model the friendship network as a directed one.

Table 3: The estimators of α_i and β_j and their standard errors in the Lazega's data set.

Vertex	d_i	$\hat{\alpha}_i$	$\hat{\sigma}_i$	b_j	$\hat{\beta}_i$	$\hat{\sigma}_j$	Vertex	d_i	$\hat{\alpha}_i$	$\hat{\sigma}_i$	b_j	$\hat{\beta}_i$	$\hat{\sigma}_j$
1	4	−6.21	0.63	5	0.53	0.60	34	6	−5.54	0.47	11	1.18	0.38
2	4	−6.01	0.67	9	1.91	0.51	35	9	−4.25	0.47	10	1.55	0.49
4	14	−3.46	0.44	14	2.79	0.41	36	9	−5.4	0.4	11	0.77	0.37
5	3	−5.01	0.64	5	1.43	0.56	38	8	−5.21	0.43	13	1.42	0.37
7	1	−6.59	1.06	2	−0.04	0.77	39	8	−5.47	0.43	13	1.14	0.37
8	1	−8.32	1.06	7	0.56	0.53	40	10	−5.29	0.39	8	0.21	0.43
9	6	−5.98	0.55	14	2.1	0.41	41	12	−5.04	0.37	17	1.42	0.35
10	14	−4.17	0.44	4	−0.45	0.70	42	14	−4.55	0.35	9	0.54	0.41
11	5	−6.49	0.56	14	1.7	0.41	43	15	−4.4	0.35	13	1.21	0.37
12	22	−2.95	0.38	8	0.86	0.49	45	6	−5.8	0.46	4	−0.63	0.56
13	14	−4.35	0.42	19	2.56	0.36	46	3	−5.61	0.66	5	0.53	0.56
14	6	−4.27	0.51	6	1.21	0.54	48	7	−5.4	0.44	4	−0.39	0.57
15	3	−4.89	0.64	2	0.39	0.79	49	4	−6.7	0.55	6	−0.42	0.48
16	8	−5.66	0.48	10	0.94	0.44	50	8	−4.34	0.47	8	1.15	0.48
17	23	−2.85	0.37	18	2.5	0.37	51	6	−4.67	0.51	7	1.11	0.51
18	8	−4.62	0.46	5	0.33	0.58	52	11	−5.1	0.38	14	1.12	0.37
19	4	−6.85	0.59	4	−0.77	0.63	54	7	−5.78	0.45	11	0.68	0.40
20	12	−5.01	0.43	7	0.2	0.49	56	7	−5.91	0.44	10	0.39	0.40
21	8	−5.73	0.46	15	1.47	0.37	57	9	−5.42	0.41	12	0.87	0.38
22	8	−5.67	0.44	6	−0.1	0.48	58	13	−3.6	0.39	12	1.83	0.42
23	1	−8.65	1.05	7	−0.01	0.48	59	5	−5.04	0.57	4	0.12	0.64
24	23	−3.59	0.34	17	1.68	0.35	60	4	−6.2	0.56	8	0.47	0.44
25	11	−3.95	0.41	10	1.6	0.46	61	3	−6.57	0.63	3	−0.88	0.64
26	9	−5.45	0.43	22	2.24	0.33	62	4	−6.32	0.55	5	−0.38	0.52
27	13	−4.54	0.38	17	2.02	0.35	64	19	−3.71	0.33	14	1.55	0.35
28	11	−3.91	0.42	9	1.32	0.49	65	22	−3.68	0.33	8	0.32	0.43
29	10	−4.81	0.39	10	1.09	0.39	66	15	−4.56	0.35	3	−0.97	0.63
30	6	−5.26	0.53	5	−0.1	0.61	67	4	−6.5	0.55	3	−1.04	0.63
31	25	−2.21	0.33	14	2.21	0.42	68	6	−5.81	0.48	5	−0.32	0.53
32	4	−5.86	0.63	7	0.54	0.56	69	5	−6.13	0.5	4	−0.64	0.56
33	12	−4.03	0.42	2	−1.55	0.89	70	7	−5.5	0.44	5	−0.25	0.52
34	6	−5.54	0.47	11	1.18	0.38							

In the data set, individuals are labelled from 1 to 71. After removing those individuals whose in-degrees or out-degrees are zeros, we perform the analysis on the 63 vertices left. The minimum, 1/4 quantile, 3/4 quantile and maximum values of $\mathbf{d}$ are 1, 5, 8, 12 and 25, respectively; those of $\mathbf{b}$ are 2, 5, 8, 13 and 22, respectively.

The estimators of α_i and β_i with their estimated standard errors are given in Table 3, in which $\beta_{71}=0$ is set as a reference. The estimates of heterogeneity parameters for in-degrees and out-degrees vary widely: from the minimum -7.36 to maximum -1.68 for $\widehat{\alpha}_i$s and from -1.32 to 2.56 for $\widehat{\beta}_i$s. We then test three null hypotheses $\alpha_1=\alpha_4$, $\alpha_1=\beta_1$ and $\beta_1=\beta_4$, using the proposed homogeneity test statistics $\hat{\xi}_{i,j}=|\hat{\alpha}_i-\hat{\alpha}_j|/(1/\hat{v}_{i,i}+1/\hat{v}_{j,j})^{1/2}$, $\hat{\zeta}_{i,j}=|\hat{\alpha}_i-\hat{\beta}_j|/(1/\hat{v}_{i,i}+1/\hat{v}_{n+j,n+j})^{1/2}$, and $\hat{\eta}_{i,j}=|\hat{\beta}_i-\hat{\beta}_j|/(1/\hat{v}_{n+i,n+i}+1/\hat{v}_{n+j,n+j})^{1/2}$ respectively. The obtained p-values turn out to be 3.5×10^{-4}, 8.7×10^{-15} and 1.7×10^{-3}, respectively, confirming the need to use our model for parameterizing the in-degree and out-degree of each node differently to characterize the heterogeneity of bi-degrees. The estimated covariate effects, their bias corrected estimators, their standard errors, and their p-values under the null of having no effects are reported in Table 4. The five categorial variables status, gender, location and practice are all significant and positive, implying that a common value for any of these three variables increases the likelihood of two lawyers to have connection. This is consistent with Figure 1. On the other hand, the larger the difference between two lawyers' age or their years with the firm, the less likely they are friends. This makes sense intuitively.

Table 4: The estimators of γ_i, the corresponding bias corrected estimators, the standard errors, and the p-values under the null $\gamma_i=0$ $(i=1,\ldots,7)$ for the Lazega's friendship data.

Covariate	$\hat{\gamma}_i$	$\hat{\gamma}_{bc,i}$	$\hat{\sigma}_i$	p-value
status	1.066	1.760	0.155	<0.001
gender	0.580	0.962	0.142	<0.001
location	2.600	3.225	0.176	<0.001
years	-0.108	-0.064	0.014	<0.001
age	-0.040	-0.027	0.011	0.015
practice	0.834	1.112	0.124	<0.001
school	0.267	-0.479	0.123	<0.001

The analysis of Sina Weibo data. We now analyze the Sina Weibo data collected by Cai et al. (2018). Sina Weibo is the largest Twitter-type social media in China. The original data contains 4077 nodes in an official MBA program with directed edges representing who follows who. For our analysis, we first remove those nodes with zero in-degrees or out-degrees since in this case the MLEs of the corresponding degree parameters do not exist. The largest strong connected subgraph of the remaining data set is then examined. This leaves a connected network with 2242 nodes. The minimum, 1/4 quantile, 3/4 quantile and maximum values of d are 1, 2, 5, 19 and 715, respectively; those of b are 1, 4, 9, 22, 253, respectively. It exhibits a strong degree heterogeneity.

Associated with each node are three variables: the number of characters in personal labels self-created by the users to describe their lifestyles (CHAR), the cumulated number of Weibo posts (POST), and the time length since Weibo registration measured in months (TIME). Before our analysis, these node attributes are normalized by subtracting the average and dividing their standard error. Then the covariates of edges are formed by using the absolute difference distance.

The two-step iterative algorithm in Section 2 is used to find the MLEs. The fitted values of the homophily parameters using model (1) are summarized in Table 5. From this table, we can see that all the node attributes are significant. In Figure 1 in the supplementary material, the histograms of the fitted values of the degree parameters are provided. We can see that the estimates of the heterogeneity parameters vary widely: from the minimum of -2.03 to the maximum of 4.13 for $\hat{\beta}_j$'s and from -8.87 to -1.28 for $\hat{\alpha}_i$'s. The histogram of $\hat{\beta}_j$'s indicates that β_j may follow a normal distribution while that of $\hat{\alpha}_i$'s clearly indicates a skewed distribution.

Table 5: The estimators of γ_i, the corresponding bias corrected estimators, the standard errors, and the p-values under the null $\gamma_i = 0$ ($i = 1, 2, 3$) for the Sina Weibo data.

Covariate	$\hat{\gamma}_i$	$\hat{\gamma}_{bc,i}$	$\hat{\sigma}_i$	p-value
CHAR	0.004	-0.391	0.018	$< 10^{-3}$
POST	0.015	-0.143	0.008	$< 10^{-3}$
TIME	-0.010	-0.158	0.008	$< 10^{-3}$

5 Discussion

In this paper, we have derived the consistency and asymptotic normality of the MLEs for estimating the parameters in model (1) when the number of vertices goes to infinity. By allowing $\|\boldsymbol{\theta}^*\|_\infty$ to diverge to infinity, our model can handle networks with the number of edges growing with the number of node at a slow rate [Krivitsky et al. (2011)]. If the growth rate on the degree parameters increases too fast, however, the MLE fails to exist with a positive frequency as demonstrated in the simulation. Note that the conditions imposed on $\|\boldsymbol{\theta}^*\|_\infty$ in Theorems 1–4 may not be the best possible. In particular, the conditions guaranteeing the asymptotic normality seem stronger than those guaranteeing the consistency. For example, the consistency requires $\|\boldsymbol{\theta}^*\|_\infty \le \frac{1}{24}\log n$ while the asymptotic normality requires $\|\boldsymbol{\theta}^*\|_\infty \le \frac{1}{44}\log n$. It would be interesting to investigate whether these bounds can be improved.

There is an implicit yet strong assumption for our model that the reciprocity parameter corresponding to the p_1-model in Holland and Leinhardt (1981) is zero. However, if similarity terms are included in the model, then there is a tendency toward reciprocity among nodes sharing similar node features. That would alleviate the lack of a reciprocity term to some extent, although it would not induce reciprocity between dissimilar nodes. To measure the reciprocity of dyads, it is natural to incorporate the model term $\rho\sum_{i<j} a_{ij}a_{ji}$ of the p_1 model into (1). In Yan and Leng (2015), encouraging empirical results were reported regarding the distribution of the MLE in the p_1 model without covariates. Nevertheless, although only one new parameter is added, the problem of investigating the asymptotic theory of the MLEs becomes more challenging. In particular, the Fisher information matrix for the parameter vector $(\rho, \alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_{n-1})$ is not diagonally dominant and thus does not belong to the class $\mathcal{L}_n(m, M)$. In order to apply the method of proofs here, a new approximate matrix with high accuracy of the inverse of the Fisher information matrix is needed. On the other hand, various extensions of the p_1 model have been developed to allow the reciprocity parameters to depend in a linear fashion on individuals i and j [Fienberg and Wasserman (1981)] and block structures [Holland, Laskey and Leinhardt (1983); Wang and Wong (1987)]. Though these models may be more realistic, their Fisher information matrices are no longer diagonally dominant. As a result, investigating their asymptotic theory becomes much more involved and we plan to do it in a future work.

Acknowledgements

We are very grateful to three referees, an associate editor, and the Editor Prof. Regina Liu for their valuable comments that have greatly improved the manuscript. Our simulation code is available on request. The authors thank Wei Cai at Northeast China Normal University for sharing the Sina Weibo data. Yan's research is partially supported by the National Natural Science Foundation of China (No. 11771171). Jiang's research is partially supported by the Hong Kong RGC grant (PolyU 253023/16P). Leng's research is partially supported by a Turing Fellowship under the EPSRC grant EP/N510129/1.

References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US Election. *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA. Harvard University Press.

Bader, G. D. and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, doi:10.1186/1471-2105-4-2.

Barabási, A. L. and Bonabau, E. (2003). Scale-free networks. *Scientific American*, 50–59.

Bradley, R. A. and Terry, M. E. (1952). The rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika*, **39**, 324–345.

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*. Hayward, California.

Burt, R. S., Kilduff, M., and Tasselli, S. (2013). Social Network Analysis: Foundations and Frontiers on Advantage. *Annual Review of Psychology*, **64**, 527–547.

Cai, W., Guan, G., Pan, R., Zhu, X., and Wang, H. (2018). Network linear discriminant analysis. *Computational Statistics & Data Analysis*, **117**, 32–44.

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, **41**, 2428–2461.

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

Cruz-Gonzalez, M., Fernández-Val, I., and Weidner, M. (2017). Probitfe and logitfe: Bias corrections for probit and logit models with two-way fixed effects. *The Stata Journal*, To appear

Diesner, J. and Carley, K. M. (2005). Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, SIAM International Conference on Data Mining, 3–14.

Dzemski, A. (2014). An empirical model of dyadic link formation in a network with unobserved heterogeneity. Preprint. Available at `http://pseweb.eu/ydepot/semin/texte1314/JMP\%20ANDREAS\%20DZEMKI.pdf`.

Dzemski, A. (2017). An empirical model of dyadic link formation in a network with unobserved heterogeneity. Working Papers in Economics, No 698.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, **1**, 502–537.

Fellows, I. and Handcock, M. S. (2012). Exponential-family Random Network Models. Available at `http://arxiv.org/abs/1208.0121`.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, **21**, 825–839.

Fienberg, S. E. and Wasserman, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociological Methodology*, **12**, 156–192.

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, **137**, 3430–3445.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, **40**, 996–1023.

Fernández-Vál, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics*, **192**, 291–312.

Goldenberg, A., Zheng, A. X., Feinberg, S. E., and Airoldi, E. M. (2009). A survey of statistical network models. *Foundations and Trends in Machine Learning*, **2**, 129–233.

Graham, B. S. (2017). An econometric model of link formation with degree heterogeneity.

Econometrica, **85**, 1033–1063.

Graham B. S. (2016). Homophily and transitivity in dynamic network formation. NBER Working Paper, No. 22186. Available at `http://www.nber.org/papers/w22186`.

Haberman, S. J. (1974). *The Analysis of Frequency Data.* Univ. Chicago Press, Chicago, IL.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, **5**, 815–841.

Hahn, J. and Newey W. (2004). Jackknife and analytical bias reduction for nonlinear panel data models. *Econometrica*, **72**, 1295–1319.

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks, Working Paper 39, Techenical report, Center for Statistics and the Social Sciences, University of Washington.

Hillar, C. and Wibisono, A. (2013). Maximum entropy distributions on graphs. Available at `http://arxiv.org/abs/1301.3321`.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13–30.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, **15**, 261–272.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, **76**, 33–65.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks.* **5**, 109-137.

Jochmans, K. (2017). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics*, To appear.

Juodis, A. (2013). A note on bias-corrected estimation in dynamic panel data models. *Economics Letters*, **118**, 435–438.

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

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models.* New York, Springer.

Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8, 319–339.

Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership.* Oxford University Press, Oxford.

Lang, S. (1993). *Real and Functional Analysis.* Springer.

Lewisa, K., Gonzaleza, M., and Kaufmanb, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 68–72.

Loéve, M. (1977). *Probability theory I.* 4th ed. Springer, New York.

McPherson, M., Lynn, S. L., and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, **27**, 415–444.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, **18**, 471–472.

Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Physics Review E*, **66**, 016128.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1–32.

Olhede, S. C. and Wolfe, P. J. (2012). Degree-based network models. Available at `http://arxiv.org/abs/1211.6537`.

Perry, P. O. and Wolfe, P. J. (2012). Null models for network data. Available at `http://arxiv.org/abs/1201.5871`.

Rasch, G. (1960). *Probabilistic Models For Some Intelligence And Attainment Tests.* Copenhagen: Paedagogiske Institut.

Rinaldo, A., Petrović, S., and Fienberg, S. (2011). Maximum likelihood estimation in network models. Technical report. Available at `http://arxiv.org/abs/1105.6145`.

Rinaldo, A., Petrović, S., and Fienberg, S. E. (2013). Maximum likelihood estimation in the β-model. *The Annals of Statistics*, **41**, 1085–1110.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, **29**, 173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, **29**, 192–215.

Sadeghi, K. and Rinaldo, A. (2014). Statistical models for degree distributions of networks, NIPS 2014 Workshop "From Graphs to Rich Data". Available at `http://arxiv.org/abs/1411.3825`.

Schweinberger, M. and Handcock, M. S. (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B*, **77**, 647–676.

Shalizi, C. R. and Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *The Annals of Statistics*, **41**, 508–535.

Simons, G. and Yao, Y. C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, **27**, 1041–1060.

Van Duijn, M. A. J., Snijders, T. A. B., and Zijlstra, B. J. H. (2004). p_2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, **58**, 234–254.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 9–19.

Wu, N. (1997). *The Maximum Entropy Method.* New York, Springer.

Yan, T. and Leng, C. (2015). A simulation study of the p_1 model. *Statistics and Its Interface*, **8**, 255–266.

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

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.

Yin, M. (2015). A detailed investigation into near degenerate exponential random graphs.

Preprint. Available at `http://arxiv.org/abs/1512.06563`.

6 Appendix: Proofs for theorems

In this section we give the proofs for Theorems 2 and 4 in Section 3, and the proofs for Theorems 1 and 3 are put in the online supplementary material.

6.1 Proof of Theorem 2

Recall that $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. In what follows, the calculations are based on the condition that $\boldsymbol{\gamma} \in \Gamma$, $\|\boldsymbol{\theta}\|_\infty \le n^\tau$, where $\tau \in (0, 1/2)$ is a positive constant. By calculations, we have

$$\begin{aligned} \ell(\boldsymbol{\gamma}, \boldsymbol{\theta}) &= \ell(\boldsymbol{\gamma}, \boldsymbol{\theta}) - \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})] + \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})] \\ &= \sum_{i \ne j} (a_{ij} - p_{ij})(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j) + \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})], \end{aligned}$$

where $\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})]$ is given in (9) and $p_{ij} = p_{ij}(\boldsymbol{\gamma}^*, \alpha_i^*, \beta_j^*)$. By the triangle inequality, we have

$$\Big| \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \ne i} (a_{ij} - p_{ij}) Z_{ij}^\top \boldsymbol{\gamma} \Big| \le \frac{1}{n} \sum_{i=1}^n \Big| \frac{1}{n-1} \sum_{j \ne i} (a_{ij} - p_{ij}) Z_{ij}^\top \boldsymbol{\gamma} \Big|. \tag{11}$$

Since we assume that Z_{ij}'s lie in a compact subset of $\mathbb{R}^p$ and the parameter space Θ of covariate parameters is compact, we have for all $i \ne j$,

$$\max_{\boldsymbol{\gamma} \in \Theta} |Z_{ij}^\top \boldsymbol{\gamma}| \le \kappa, \tag{12}$$

where κ is a constant. By inequality (12), $a_{ij} Z_{ij}^\top \boldsymbol{\gamma}$ is a bounded random variable with the upper bound κ. By Hoeffding's (1963) inequality, we have

$$P\Big(\Big| \frac{1}{n-1} \sum_{j \ne i} (a_{ij} - p_{ij}) Z_{ij}^\top \boldsymbol{\gamma} \Big| \ge \epsilon \Big) \le 2 \exp\left(-\frac{(n-1)\epsilon^2}{2\kappa^2} \right).$$

By taking $\epsilon = 2\kappa[\log(n-1)/(n-1)]^{1/2}$, we have

$$P\left(\Big| \frac{1}{n-1} \sum_{j \ne i} (a_{ij} - p_{ij}) Z_{ij}^\top \boldsymbol{\gamma} \Big| \ge 2\kappa \sqrt{\frac{\log(n-1)}{(n-1)}} \right) \le \frac{4}{(n-1)^2}.$$

Therefore, we have

$$\begin{aligned}
& P\left(\Big|\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}(a_{ij}-p_{ij})Z_{ij}^\top\boldsymbol{\gamma}\Big| \ge 2\kappa\sqrt{\frac{\log(n-1)}{(n-1)}}\right) \\
\le\ & P\left(\frac{1}{n}\sum_{i=1}^n\Big|\frac{1}{n-1}\sum_{j\neq i}(a_{ij}-p_{ij})Z_{ij}^\top\boldsymbol{\gamma}\Big| \ge 2\kappa\sqrt{\frac{\log(n-1)}{(n-1)}}\right) \\
\le\ & P\left(\bigcup_{i=1}^n\Big|\frac{1}{n-1}\sum_{j\neq i}(a_{ij}-p_{ij})Z_{ij}^\top\boldsymbol{\gamma}\Big| \ge 2\kappa\sqrt{\frac{\log(n-1)}{(n-1)}}\right) \\
\le\ & \frac{n}{(n-1)^2}.
\end{aligned}$$

In the above, the first inequality is due to (11). Note that $\|\boldsymbol{\alpha}\| \le n^\tau$ and $\|\boldsymbol{\beta}\| \le n^\tau$. Similarly, with probability at most $n/(n-1)^2$, we have

$$\begin{aligned}
\Big|\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}(a_{ij}-p_{ij})\alpha_i\Big| &\ge \frac{1}{n(n-1)}\sum_{i=1}^n\Big|\sum_{j\neq i}\frac{\alpha_i}{n-1}(a_{ij}-p_{ij})\Big| \\
&\ge \frac{1}{n(n-1)}\cdot n\cdot n^\tau\sqrt{\frac{\log(n-1)}{n-1}} = \frac{(\log n)^{1/2}}{n^{1/2-\tau}},
\end{aligned}$$

and

$$\Big|\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}(a_{ij}-p_{ij})\beta_j\Big| \ge \frac{(\log n)^{1/2}}{n^{1/2-\tau}}.$$

Hence, with probability at least $1-3n/(n-1)^2$, we have

$$\max_{\boldsymbol{\gamma}\le\Gamma,\|\boldsymbol{\theta}\|_\infty\le n^\tau}\Big|\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}(a_{ij}-p_{ij})(Z_{ij}^\top\boldsymbol{\gamma}+\alpha_i+\beta_j)\Big| < \frac{(\log n)^{1/2}}{n^{1/2-\tau}},$$

or equivalently,

$$\max_{\boldsymbol{\gamma}\le\Gamma,\|\boldsymbol{\theta}\|_\infty\le n^\tau}\Big|\frac{1}{n(n-1)}\left\{\ell(\boldsymbol{\gamma},\boldsymbol{\theta})-\mathbb{E}[\ell(\boldsymbol{\gamma},\boldsymbol{\theta})]\right\}\Big| < \frac{(\log n)^{1/2}}{n^{1/2-\tau}}. \tag{13}$$

Let $B_n(\rho)=\{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_\infty<\rho\}$ be an open ball in Γ with $\boldsymbol{\gamma}^*$ as its center and ρ as its radius, and $B_n^c(\rho)$ be its complement in Γ. Define

$$\epsilon_n(\rho)=\frac{1}{n(n-1)}\left\{\max_{\|\boldsymbol{\theta}\|_\infty\le n^\tau}\mathbb{E}[\ell(\boldsymbol{\gamma}^*,\boldsymbol{\theta}]-\max_{\boldsymbol{\gamma}\in B_n^c(\rho),\|\boldsymbol{\theta}\|_\infty\le n^\tau}\mathbb{E}[\ell(\boldsymbol{\gamma},\boldsymbol{\theta})]\right\},$$

and

$$\epsilon_n(\rho_n)=\arg\min_\rho\epsilon_n(\rho) > \frac{2(\log n)^{1/2}}{n^{1/2-\tau}}.$$

Recall that $\mathbb{E}[\ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta})] = \sum_{i<j} D_{KL}(p_{ij} \| p_{ij}(\boldsymbol{\gamma}^*, \alpha_i, \beta_j)) - \sum_{i<j} S(p_{ij})$. Therefore,

$$
\begin{aligned}
& \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \mathbb{E}[\ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta}] - \max_{\boldsymbol{\gamma} \in B_n^c(\rho), \|\boldsymbol{\theta}\|_\infty \le n^\tau} \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})] \\
= & \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \sum_{i<j} D_{KL}(p_{ij} \| p_{ij}(\boldsymbol{\gamma}^*, \alpha_i, \beta_j)) - \max_{\boldsymbol{\gamma} \in B_n^c(\rho), \|\boldsymbol{\theta}\|_\infty \le n^\tau} \sum_{i<j} D_{KL}(p_{ij} \| p_{ij}(\boldsymbol{\gamma}^*, \alpha_i, \beta_j)).
\end{aligned}
$$

By the property of the Kullback-Leibler divergence and noticing that p_{ij} is a monotonous function on γ_k, α_i and β_j, $\mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})]$ is uniquely maximized at $(\boldsymbol{\gamma}^*,\ \boldsymbol{\theta}^*)$. Therefore, ϵ_n will be strictly greater than zero for each fixed n. Further, since $\epsilon_n(\rho)$ is a continuous increasing function on ρ as ρ increases, we have

$$
\rho_n \to 0, \quad \text{as} \quad n \to \infty. \tag{14}
$$

Let E_n be the event

$$
\frac{1}{n(n-1)} \Big| \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \ell(\boldsymbol{\gamma}, \boldsymbol{\theta}) - \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})] \Big| < \frac{\epsilon_n(\rho_n)}{2}.
$$

for all $\boldsymbol{\gamma} \in \Gamma$. Under event E_n, we get the inequalities

$$
\max_{\|\boldsymbol{\theta}_\infty\| \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta})] > \frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}}) - \frac{\epsilon_n(\rho_n)}{2}, \tag{15}
$$

$$
\max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta}) > \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta})] - \frac{\epsilon_n(\rho_n)}{2}. \tag{16}
$$

According to the definition of the restricted MLE, we have that

$$
\frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}}) \ge \max_{\|\boldsymbol{\theta}\| \le n^\tau} \frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}).
$$

Then, by inequality (15), we have

$$
\max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta})] > \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}) - \frac{\epsilon_n}{2}. \tag{17}
$$

Adding both sides of (16) and (17) gives

$$
\begin{aligned}
& \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta})] - \Big[\max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}) - \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta}) \Big] \\
> & \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta})] - \epsilon_n(\rho_n) \\
= & \max_{\boldsymbol{\gamma} \in B_n^c, \|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})],
\end{aligned}
$$

where the equality follows the definition of ϵ_n. By noting that

$$\max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}) \ge \max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta}),$$

we have

$$\max_{\|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta})] > \max_{\boldsymbol{\gamma} \in B_n^c, \|\boldsymbol{\theta}\|_\infty \le n^\tau} \frac{1}{n(n-1)} \mathbb{E}[\ell(\boldsymbol{\gamma}, \boldsymbol{\theta})].$$

From the above equation, we have that $E_n \Rightarrow \widehat{\boldsymbol{\gamma}} \in B_n(\rho_n)$. Therefore $P(E_n) \le P(\widehat{\boldsymbol{\gamma}} \in B_n(\rho_n))$. Inequality (13) implies that $\lim_{n\to\infty} P(E_n) = 1$ according to the definition of ρ_n. By (14), it follows that $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$.

6.2 Derivation of approximate expression for $I_*(\boldsymbol{\gamma})$

Recall that H is the Hessian matrix of the log-likelihood function (2):

$$H = \begin{pmatrix} H_{\gamma\gamma} & H_{\gamma\theta} \\ H_{\gamma\theta}^\top & -V \end{pmatrix},$$

where

$$-H_{\gamma\gamma} = \sum_{i \ne j} p_{ij}(1-p_{ij}) Z_{ij} Z_{ij}^\top, \tag{18}$$

and

$$-H_{\gamma\theta}^\top = \begin{pmatrix} \sum_{j\ne 1} p_{1j}(1-p_{1j}) Z_{1j}^\top \\ \vdots \\ \sum_{j \ne n} p_{nj}(1-p_{nj}) Z_{nj}^\top \\ \sum_{i \ne 1} p_{i1}(1-p_{i1}) Z_{i1}^\top \\ \vdots \\ \sum_{i \ne n-1} p_{i,n-1}(1-p_{i,n-1}) Z_{i,n-1}^\top \end{pmatrix}.$$

In what follows, we will derive the approximate expression of $I_*(\boldsymbol{\gamma})$. Let $(1)_{m\times n}$ be an $m \times n$ matrix whose elements all are 1. By calculations, we have

$$SH_{\gamma\theta}^\top = DH_{\gamma\theta}^\top + \frac{1}{v_{2n,2n}} \begin{pmatrix} (1)_{n\times n} & (-1)_{n\times(n-1)} \\ (-1)_{(n-1)\times n} & (1)_{(n-1)\times(n-1)} \end{pmatrix} H_{\gamma\theta}^\top,$$

where $D = \mathrm{diag}(1/v_{11}, \ldots, 1/v_{2n-1,2n-1})$. By noting that

$$\sum_{i=1}^n \sum_{j\ne i} p_{ij}(1-p_{ij}) Z_{ij}^\top - \sum_{j=1}^{n-1} \sum_{i \ne j} p_{ij}(1-p_{ij}) Z_{ij}^\top = \sum_{i \ne n} p_{in}(1-p_{in}) Z_{in}^\top,$$

we have

$$
\begin{aligned}
H_{\gamma\theta} S H_{\gamma\theta}^\top &= H_{\gamma\theta} D H_{\gamma\theta}^\top + \frac{1}{v_{2n,2n}} H_{\gamma\theta} \begin{pmatrix} (1)_{n\times 1} \\ (-1)_{(n-1)\times 1} \end{pmatrix} \sum_{i\neq n} p_{in}(1-p_{in}) Z_{in}^\top \\
&= \sum_{i=1}^{n} \frac{1}{v_{ii}} \Big(\sum_{j\neq i} p_{ij}(1-p_{ij}) Z_{ij} \Big) \Big(\sum_{j\neq i} p_{ij}(1-p_{ij}) Z_{ij}^\top \Big) \\
&\quad + \sum_{j=1}^{n} \frac{1}{v_{n+j,n+j}} \Big(\sum_{i\neq j} p_{ij}(1-p_{ij}) Z_{ij} \Big) \Big(\sum_{i\neq j} p_{ij}(1-p_{ij}) Z_{ij}^\top \Big).
\end{aligned} \tag{19}
$$

By Lemma 1, we have

$$
\|V^{-1} - S\| \le \frac{c_1 M^2}{m^3(n-1)} \le \frac{c_1}{(n-1)^2} \times \left(\frac{1}{4}\right)^2 \times \frac{(1+e^{2\|\theta^*\|_\infty + \kappa})^6}{(e^{2\|\theta^*\|_\infty+\kappa})^3} = O\Big(\frac{e^{6\|\theta^*\|_\infty}}{n^2} \Big).
$$

Therefore,

$$
\|H_{\gamma\theta}(V^{-1} - S) H_{\gamma\theta}^\top\|_\infty \le \|H_{\gamma\theta}\|_\infty^2 \|V^{-1} - S\|_\infty \le O(n^2) \times O\bigg(n \frac{e^{6\|\theta^*\|_\infty}}{n^2} \bigg) = O(n e^{6\|\theta^*\|_\infty}).
$$

Recall that $N = n(n-1)$ and note that

$$
(H_{\gamma\gamma} + H_{\gamma\theta} V^{-1} H_{\gamma\theta}^\top) = H_{\gamma\gamma} + H_{\gamma\theta} S H_{\gamma\theta}^\top + H_{\gamma\theta}(V^{-1} - S) H_{\gamma\theta}^\top.
$$

Therefore, we have

$$
-N^{-1}(H_{\gamma\gamma} + H_{\gamma\theta} V^{-1} H_{\gamma\theta}^\top) = -N^{-1}(H_{\gamma\gamma} + H_{\gamma\theta} S H_{\gamma\theta}^\top) + o(1), \tag{20}
$$

where $H_{\gamma\gamma}$ and $H_{\gamma\theta} S H_{\gamma\theta}^\top$ are given in (18) and (19), respectively. It shows that the limit of $-N^{-1}(H_{\gamma\gamma} + H_{\gamma\theta} S H_{\gamma\theta}^\top)$ is $I_*(\boldsymbol{\gamma})$ defined in (10).

6.3 Proofs for Theorem 4

Let $\hat{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\gamma}^*, \boldsymbol{\theta})$. Similar to the proofs of Theorems 1 and 2 in Yan et al. (2016), we have two lemmas below, which will be used in the proof of Theorem 4.

Lemma 2. *Assume that $\boldsymbol{\theta}^* \in \mathbb{R}^{2n-1}$ with $\|\boldsymbol{\theta}^*\|_\infty \le \tau \log n$, where $0 < \tau < 1/24$ is a constant, and that $A \sim \mathbb{P}_{\boldsymbol{\theta}^*}$. Then as n goes to infinity, with probability approaching one, the $\hat{\boldsymbol{\theta}}^*$ exists and satisfies*

$$
\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_\infty = O_p\left(\frac{(\log n)^{1/2} e^{8\|\boldsymbol{\theta}^*\|_\infty}}{n^{1/2}} \right) = o_p(1).
$$

Lemma 3. *If* $\|\boldsymbol{\theta}^*\|_\infty \leq \tau \log n$ *and* $\tau < 1/40$, *then for any* i,

$$\hat{\theta}_i^* - \theta_i^* = [S\{\mathbf{g} - \mathbb{E}(\mathbf{g})\}]_i + o_p(n^{-1/2}).$$

For convenience, define $\ell_{ij}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ by the $(i,j)^{th}$ dyad's contributions to the log-likelihood function in (2), i.e.,

$$\ell_{ij}(\boldsymbol{\gamma}, \boldsymbol{\theta}) = a_{ij}(Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j) - \log(1 + e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j}).$$

Let T_{ij} be a $2n-1$ dimensional vector with ones in its ith and $n+j$th elements and zeros otherwise. Let $s_{\gamma_{ij}}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ and $s_{\theta_{ij}}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ denote the score of $\ell_{ij}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ associated with the vector parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, respectively:

$$s_{\gamma_{ij}}(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{\partial \ell_{ij}}{\partial \boldsymbol{\gamma}} = a_{ij} Z_{ij} - \frac{Z_{ij} e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j}}{1 + e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j}},$$

$$s_{\theta_{ij}}(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{\partial \ell_{ij}}{\partial \boldsymbol{\theta}} = a_{ij} T_{ij} - \frac{e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j}}{1 + e^{Z_{ij}^\top \boldsymbol{\gamma} + \alpha_i + \beta_j}} T_{ij}.$$

Then we have the following lemma, whose proof is given in online supplementary material.

Lemma 4. *Let* $H_{\boldsymbol{\theta}\boldsymbol{\theta}} = -V$ *and*

$$s_{\gamma_{ij}}^*(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) := s_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) - H_{\boldsymbol{\gamma}\boldsymbol{\theta}} H_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} s_{\theta_{ij}}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*). \tag{21}$$

Then $\frac{1}{\sqrt{N}}[I_n(\boldsymbol{\gamma}^*)]^{-1/2} \sum_{i=1}^n \sum_{j \neq i} s_{\gamma_{ij}}^*(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$ *follows asymptotically a p-dimensional multivariate standard normal distribution.*

Proof of Theorem 4. Recall that $\widehat{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\gamma}, \boldsymbol{\theta})$. A mean value expansion gives

$$\sum_{i=1}^n \sum_{j \neq i} s_{\gamma_{ij}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}}) - \sum_{i=1}^n \sum_{j \neq i} s_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \widehat{\boldsymbol{\theta}}(\boldsymbol{\gamma}^*)) = \sum_{i=1}^n \sum_{j \neq i} \frac{\partial}{\partial \boldsymbol{\gamma}^\top} s_{\gamma_{ij}}(\bar{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{\gamma}}))(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*),$$

where $\bar{\boldsymbol{\gamma}} = t\boldsymbol{\gamma}^* + (1-t)\widehat{\boldsymbol{\gamma}}$ for some $t \in (0,1)$. By noting that $\sum_{i=1}^n \sum_{j \neq i} s_{\gamma_{ij}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}}) = 0$, we have

$$\sqrt{N}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = -\Big[\frac{1}{N} \sum_{i=1}^n \sum_{j \neq i} \frac{\partial}{\partial \boldsymbol{\gamma}^\top} s_{\gamma_{ij}}(\bar{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}(\bar{\boldsymbol{\gamma}}))\Big]^{-1} \times \Big[\frac{1}{\sqrt{N}} \sum_{i=1}^n \sum_{j \neq i} s_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \hat{\boldsymbol{\theta}}(\boldsymbol{\gamma}^*))\Big].$$

Since the dimension p of $\boldsymbol{\gamma}$ is fixed, by Theorem 2, we have

$$-\frac{1}{N}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\partial}{\partial\boldsymbol{\gamma}^\top}s_{\gamma_{ij}}(\bar{\gamma},\widehat{\boldsymbol{\theta}}(\bar{\gamma}))\xrightarrow{p} I_*(\boldsymbol{\gamma}).$$

Let $\hat{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}}(\boldsymbol{\gamma}^*)$. Therefore,

$$\sqrt{N}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) = I_*^{-1}(\boldsymbol{\gamma})\times\Big[\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j\neq i}s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\hat{\boldsymbol{\theta}}^*)\Big]+o_p(1). \tag{22}$$

By applying a third order Taylor expansion to the summation in brackets in (22), it yields

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j\neq i}s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\hat{\boldsymbol{\theta}}^*) = S_1+S_2+S_3, \tag{23}$$

where

$$\begin{aligned}
S_1 &= \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{n}\sum_{j\neq i}s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)+\tfrac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j\neq i}\Big[\tfrac{\partial}{\partial\boldsymbol{\theta}^\top}s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)\Big](\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*),\\
S_2 &= \tfrac{1}{2\sqrt{N}}\textstyle\sum_{k=1}^{2n-1}\Big[(\hat{\theta}_k^*-\theta_k^*)\sum_{i=1}^{n}\sum_{j\neq i}\tfrac{\partial^2}{\partial\theta_k\partial\boldsymbol{\theta}^\top}s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)\times(\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*)\Big],\\
S_3 &= \tfrac{1}{6\sqrt{N}}\textstyle\sum_{k=1}^{2n-1}\sum_{l=1}^{2n-1}\{(\hat{\theta}_k^*-\theta_k^*)(\hat{\theta}_l^*-\theta_l^*)\Big[\sum_{i=1}^{n}\sum_{j\neq i}\tfrac{\partial^3 s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\bar{\theta}^*)}{\partial\theta_k\partial\theta_l\partial\boldsymbol{\theta}^\top}\Big](\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*)\}.
\end{aligned}$$

Similar to the proof of Theorem 4 in Graham (2017), we will show that (1) S_1 is asymptotically normal distribution; (2) S_2 is the bias term having a non-zero probability limit; (3)S_3 is an asymptotically negligible remainder term.

We work with S_1, S_2 and S_3 in reverse order. We first evaluate the term S_3. We calculate $g_{klh}^{ij} = \frac{\partial^3 s_{\gamma_{ij}}(\gamma,\theta)}{\partial\theta_k\partial\theta_l\partial\theta_h}$ as follows.
(1) For different k,l,h, $g_{klh}^{ij}=0$.
(2) Only two values are equal. If $k=l=i\le n; h\ge n+1$, $g_{klh}^{ij}=p_{ij}(1-p_{ij})(1-6p_{ij}+6p_{ij}^2)Z_{ij}$; for other cases, the results are similar.
(3) Three values are equal. $g_{klh}^{ij}=p_{ij}(1-p_{ij})(1-6p_{ij}+6p_{ij}^2)Z_{ij}$ if $k=l=h=i\le n$; $g_{klh}^{ij}=p_{ji}(1-p_{ji})(1-6p_{ji}+6p_{ji}^2)Z_{ji}$ if $k=l=h=j\ge n+1$.
Therefore, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k,l,h}\frac{\partial^3 s_{\gamma_{ij}}(\boldsymbol{\gamma}^*,\bar{\theta}^*)}{\partial\theta_k\partial\theta_l\partial\theta_h}\\
=\;&\frac{1}{2}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j=1}^{n-1}Z_{ij}[p_{ij}(1-p_{ij})(1-6p_{ij}+6p_{ij}^2)(\hat{\alpha}_i-\alpha_i^*)^2(\hat{\beta}_j-\beta_j^*)+\\
&p_{ji}(1-p_{ji})(1-6p_{ji}+6p_{ji}^2)(\hat{\alpha}_i-\alpha_i^*)(\hat{\beta}_j-\beta_j^*)^2].
\end{aligned}$$

Let $\lambda_n = \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_\infty$. Note that Z_{ij} lies in a compact set $\mathbb{Z}$, and $p_{ij}(1-p_{ij}) \leq 1/4$, and $|(1-6p_{ij}+6p_{ij}^2)| \leq 6$. By Lemma 2, any element of S_3 is bounded above by

$$\begin{aligned}\frac{n(n-1)}{\sqrt{N}} \times \frac{6}{4}\lambda_n^3 \times \sup_{z\in\mathbb{Z}}|z| &= 3\frac{n(n-1)}{\sqrt{n(n-1)}} \times \frac{C^3(\log n)^{3/2}e^{24\|\theta^*\|_\infty}}{n^{3/2}} \times \sup_{z\in Z}|z| \\ &= O\left(\frac{(\log n)^{3/2}e^{24\|\theta^*\|_\infty}}{\sqrt{n}}\right) = o(1).\end{aligned}$$

Similar to the calculation of deriving the asymptotic bias in Theorem 4 in Graham (2017), we have $S_2 = B_* + o_p(1)$, where

$$B_* = \lim_{n\to\infty}\frac{1}{2\sqrt{N}}\left[\sum_{i=1}^n \frac{\sum_{j\neq i} p_{ij}(1-p_{ij})(1-2p_{ij})Z_{ij}}{\sum_{j\neq i} p_{ij}(1-p_{ij})} + \sum_{j=1}^n \frac{\sum_{i\neq j} p_{ij}(1-p_{ij})(1-2p_{ij})Z_{ij}}{\sum_{i\neq j} p_{ij}(1-p_{ij})}\right]. \tag{24}$$

By Lemma 3, similar to deriving the asymptotic expression of S_1 in Graham (2017), we have

$$S_1 = \frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j\neq i} s^*_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) + o_p(1),$$

Therefore, it shows that equation (23) equal to

$$\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j\neq i} s_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \hat{\boldsymbol{\theta}}^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j\neq i} s^*_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) + B_* + o_p(1), \tag{25}$$

with $\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j\neq i} s^*_{\gamma_{ij}}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$ equivalent to the first two terms in (23) and B_* the probability limit of the third term in (23).

Substituting (25) into (22) then gives

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = I_*^{-1}(\gamma)B_* + I_*^{-1}(\gamma)\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j\neq i} s^*_{\gamma_{ij}}(\gamma^*, \theta^*) + o_p(1).$$

Then Theorem 4 immediately follows from Lemma 4. □