Estimation and Inference in Dyadic Network Formation Models with Nontransferable Utilities*

Ming Li[a], Zhentao Shi[b], Yapeng Zheng[b]

[a]Department of Economics, National University of Singapore
[b]Department of Economics, Chinese University of Hong Kong

Abstract

This paper studies estimation and inference in a dyadic network formation model with observed covariates, unobserved heterogeneity, and nontransferable utilities. With the presence of the high dimensional fixed effects, the maximum likelihood estimator is numerically difficult to compute and suffers from the incidental parameter bias. We propose an easy-to-compute one-step estimator for the homophily parameter of interest, which is further refined to achieve $\sqrt{N}$-consistency via split-network jackknife and efficiency by the bootstrap aggregating (bagging) technique. We establish consistency for the estimator of the fixed effects and prove asymptotic normality for the unconditional average partial effects. Simulation studies show that our method works well with finite samples. We provide two empirical applications, one using the Nyakatoke risk-sharing network dataset and the other using the India micro-finance network dataset, and obtain economically meaningful results.

Keywords: dyadic network formation, unobserved heterogeneity, nontransferable utilities, bagging, machine learning, Cramér-Rao lower bound.

*Li: mli@nus.edu.sg. Shi (corresponding author): zhentao.shi@cuhk.edu.hk, 9F Esther Lee Building, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China. Zheng: yapengzheng@link.cuhk.edu.hk. We thank Chih-Sheng Hsieh, Oliver Linton, Bryan Graham, Shuyang Sheng, Martin Weidner, and the participants of USyd econometrics seminar and for the helpful comments.

1 Introduction

Our society and economy do not exist in isolation; they are inherently connected through complex networks of relationships and interactions (Banerjee, Chandrasekhar, Duflo, and Jackson (2013); Acemoglu, Ozdaglar, and Tahbaz-Salehi (2015); König, Rohner, Thoenig, and Zilibotti (2017); Battaglini, Patacchini, and Rainone (2022)). These networks play a pivotal role in shaping the decisions and behaviors of individuals, organizations, and institutions. For example, a consumer's purchasing decision may be swayed by the opinions of friends, or a company's strategic move could be shaped by the actions of competitors within its network. Understanding the structure and dynamics of these networks is therefore crucial for analyzing how decisions propagate through society and the economy. This makes the study of network formation—how these networks come into existence, evolve, and influence behavior—an essential area of inquiry. By estimating and understanding network formation, we can gain insights into the underlying mechanisms that drive social and economic phenomena, ultimately leading to more informed decisions and effective policies.

This paper studies efficient estimation and inference in a flexible dyadic network formation model with observed covariates, unobserved heterogeneity, and nontransferable utilities (NTU). We consider one single large network which is arguably the most common type of network data available in empirical studies. By "efficient," we mean that our proposed estimator achieves the Cramér-Rao lower bound asymptotically, and a computationally efficient algorithm is provided. By "flexible," we include both observed pairwise covariates for studying homophily effect and unobserved individual fixed effects that can be arbitrarily correlated with the observed covariates. Consequently, our model can capture rich forms of heterogeneity among agents in the network. Finally, in contrast to a large body of work (e.g., Graham (2017)) that considers transferable utilities (TU), we model real-world social interactions by requiring bilateral consent which is captured by NTU. For instance, friendship is usually formed only when both individuals in question are willing to accept each other as a friend, or in other words, when both individuals derive sufficiently high utilities from establishing the friendship. It is even more prominent in business networks since no firm would want to deal if it incurs a loss from the transaction when there is a lack of mechanism to guarantee profit redistribution. Moreover, as pointed out by Gao, Li, and Xu (2023), NTU can effectively incorporate homophily effects on unobserved heterogeneity, which is another advantage over TU.

The combination of unobserved individual fixed effects and NTU poses significant challenges for estimation and inference. First, the requirement of bilateral agreement to form a link under NTU breaks down the additivity in the fixed effects in the utility surplus function, i.e., the linking probability between two individuals is no longer additively separable in their fixed effects, which makes infeasible the arithmetic differencing based methods (e.g., Toth (2017); Candelaria (2024)) to cancel out the individual fixed effects. Second, including individual fixed effects leads to a high-dimensional optimization problem for maximum likelihood estimation that is known to suffer from computational inefficiency and instability. Third, in the presence of high-dimensional fixed effects, it is difficult to verify concavity of the log-likelihood function, which is required to ensure the existence of the global minimum of the criterion function. Moreover, with the fixed effects treated as incidental parameters, the maximum likelihood estimators for the homophily parameters are biased (Moreira (2009)). Fourth, the Jacobian matrix of the moment equations used to construct moment estimators is asymmetric due to NTU, making existing methods for asymptotic analysis that works under TU not directly applicable (e.g., Chatterjee, Diaconis, and Sly (2011); Graham (2017); Yan, Jiang, Fienberg, and Leng (2019); Candelaria (2024)). Last but not the least, many existing results are based on the assumption that the distribution of the idiosyncratic shock is logistic (e.g., Chatterjee, Diaconis, and Sly (2011); Graham (2017); Yan, Jiang, Fienberg, and Leng (2019)). It is unclear whether the conclusions still hold when a different distribution for the shock is used. Thus, a new method is called for to deal with these challenges.

In this paper, we propose an easy-to-compute bagging estimator for the homophily parameters that deals with the issues discussed above. We prove its asymptotic normality, $\sqrt{N}$-consistency[1], and efficiency. Our paper is the first one in the literature of dyadic network formation with NTU that has inference and efficiency results. The new bagging estimator involves three steps. First, we propose an initial joint method of moments (JMM) estimator that solves a high-dimensional system of moment equations. The JMM estimator is not $\sqrt{N}$-consistent, nor is it efficient. Nonetheless, we show that it satisfies the conditions for Le Cam (1969)'s one-step approximation to the MLE. In this step, we also estimate the high-dimensional fixed effects and prove their uniform consistency. Second, we plug the JMM estimator into Le Cam (1969)'s approximation step and obtain the one-step estimator that is asymptotically equivalent to the MLE. The one-step estimator is—similar to the MLE (Moreira (2009))—efficient but not $\sqrt{N}$-consistent. To correct for the bias while maintaining its efficiency, we use the bootstrap aggregating (also known as "bagging") method from the machine learning literature (Breiman (1996); Hirano and Wright (2017)) for split-network jackknife on the one-step estimator to obtain the bagging estimator. As far as we are aware, the application of the bagging method is novel in the context of network formation literature. As two extensions, we provide a consistent estimator and prove its asymptotic normality for

[1]By "$\sqrt{N}$-consistent," we mean the asymptotic distribution of the estimator after centering at the true parameter and multiplied by $\sqrt{N}$ has mean zero.

the average partial effects (APE, see Hughes (2023)) and discuss how misspecification of the link function affects the analysis, the latter of which is much less considered in this literature (Graham (2024)).

In simulation studies, we find our proposed estimators for the homophily parameters, individual fixed effects, and APEs all work well. We present two empirical applications. First, we apply our method to the risk-sharing network data of Nyakatoke (De Weerdt (2004); De Weerdt and Dercon (2006); De Weerdt and Fafchamps (2011)) and obtain economically meaningful results. Our empirical findings complement the results of Gao, Li, and Xu (2023) by showing that wealth differences do not have a statistically significant impact on the formation of links. Second, we use the well known microfinance network dataset (Banerjee et al. (2013, 2024)) to show how our method works in capturing important network features such as average degree, clustering effect, and number of isolates. By adding only one more parameter to capture triangles, we find that the extended model captures most network features well.

We discuss two important limitations of the paper. First, we require correct specification of the link functions. When the link function is misspecified, we prove theoretical properties of these estimators in Section 4.2 and provide their finite sample performance in Section 5.3. Second, we do not consider interdependent link preferences; instead, we focus on modeling individual heterogeneity via fixed effects and NTU. Following Graham (2017), we briefly discuss how to test the hypothesis of no interdependent link preferences in Section 7.

1.1 Literature Review

Our paper belongs to the literature that studies dyadic network formation in a single large network setting. An incomplete list of the papers in this category include Blitzstein and Diaconis (2011), Chatterjee, Diaconis, and Sly (2011), Yan and Xu (2013), Yan, Leng, and Zhu (2016a), Graham (2017), Charbonneau (2017), Jochmans (2017), Toth (2017), Dzemski (2019), Yan, Jiang, Fienberg, and Leng (2019), Gao (2020), Zeleneev (2020), Gao, Li, and Xu (2023), Hughes (2023), Candelaria (2024), Qu, Chen, Yan, and Chen (2024). See Graham (2020) for a comprehensive review. Most of the papers in this list except Gao, Li, and Xu (2023) consider link formation with TU, which generally allows one to cancel out the individual fixed effects by arithmetic differencing. In a semiparametric network formation model, Gao, Li, and Xu (2023) propose a logical differencing technique to cancel out the fixed effects without imposing any distributional assumption on the idiosyncratic shocks. However, they do not have inference results for the homophily parameters, nor do they have an estimator for the fixed effects. In contrast, we consider a flexible network formation model with known distribution of the idiosyncratic shocks to link formation and prove asymptotic normality for our estimators of the homophily parameters. We also provide ℓ_∞-consistent estimators for the high-dimensional fixed effects and prove asymptotic normality for the unconditional APEs. Therefore, we consider our paper to be complimentary to Gao, Li, and Xu (2023).

Our paper builds on the highly influential paper by Graham (2017) who considers network formation problem with TU. Graham (2017) proposes a novel tetrad logit estimator, which is not directly applicable to our setting with NTU even when the link functions are logistic. The joint MLE of Graham (2017) relies on both the link functions being logistic and TU to solve for the fixed effects as solutions to a system of fixed point equations. Our one-step estimator does not require the logistic assumption and we introduce a new stochastic gradient descent type algorithm to estimate the fixed effects. Furthermore, our method deals with several theoretical and computational challenges, such as non-concavity of criterion function, non-stability of optimization procedure, high-dimensionality of parameters, that are common with MLE's. Finally, we propose a bagging split-network jackknife estimator that achieves both $\sqrt{N}$-consistency and the Cramér-Rao lower bound. More recently, Graham (2024) studies sparse network asymptotics for logistic regression under possible model misspecification. He shows that the parameter that indexes the logit approximation solves a particular Kullback–Leibler Information Criterion minimization problem, and proves asymptotic normality of the logistic regression coefficients. This paper is different from Graham (2024) in that we consider network formation with NTU, which is effectively ruled out by his use of composite likelihood to cancel out the individual heterogeneity.

Our paper is also related to Hughes (2023), who considers a parametric link formation model with TU and fixed effects and proposes a jackknife estimator for bias-correction. His results are not directly applicable to NTU due to the asymmetry of the Hessian matrix. Moreover, we propose a one-step split-network jackknife bagging estimator that both achieves the Cramér-Rao lower bound and is easy to compute. Recently, Qu, Chen, Yan, and Chen (2024) study inference in semiparametric formation models for directed networks with TU and propose a projection approach to estimate the unknown homophily parameters. Their model of directed network formation with TU and method which relies on the existence of a special regressor are different from ours. There are also papers dedicated to semiparametric or nonparametric undirected network formation models with TU and unobserved individual heterogeneity, e.g., Toth (2017) (maximum rank based method), Gao (2020) (identification strategy based on an in-fill and out-expansion strategy), Zeleneev (2020) (pseudo-distance type argument to identify agents with same fixed effect), Candelaria (2024) (special regressor based method, which is also used in Qu, Chen, Yan, and Chen (2024)). Recognizing the possibility that the distribution of idiosyncratic shocks to link formation can be misspecified, we prove theoretical results under such misspecification, and provide simulation results to justify the use of our estimator.

The insights from large T panel data literature (e.g., Hahn and Newey (2004); Hahn and Kuersteiner (2011); Dhaene and Jochmans (2015); Mei, Sheng, and Shi (2024)) shed light upon important issues in network formation problems. See Fernández-Val and Weidner (2018) for an excellent review of this literature. Fernández-Val and Weidner (2016) study nonlinear panel models with large N and T in the context of a broad class of maximum likelihood models with fixed effects. Their general results are applicable to network data wherein N and T grow at the same rate in asymptotics. Their key assumptions include that the log-likelihood function is concave and that certain derivatives of functionals of the fixed effects satisfy a sparsity assumption, which are different from ours. As a result, their method differs from ours substantially. It is worth pointing out that in their footnote 8, Fernández-Val and Weidner (2016) discuss how to avoid ambiguity and arbitrariness in the choice of the random splitting by repeated sampling, an idea similar to bagging. In contrast, we formally introduce the bagging technique to achieve bias-correction while maintaining efficiency of the estimator simultaneously. Furthermore, we provide formal asymptotic results for the bagging estimator for inference. In another seminal paper, Chen, Fernández-Val, and Weidner (2021) study nonlinear factor models for panel and network data. They introduce iterative factor structures to network data, which enables one to capture important network features such as degree heterogeneity and homophily in latent factors in an unspecified or reduced-form fashion. They propose an EM-type algorithm for estimating structural parameters and a separate algorithm for estimating the number of factors. Their method requires that the covariates and unobserved effects enter conditional likelihood function through a single index, which rules out network formation with NTU. Moreover, they require concavity on the joint likelihood function in the single index. Thus, their results are not directly applicable to our problem.

Our paper is also related to papers that utilize dyadic link formation models to study structural social interaction models. A few examples include Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (2016), Johnsson and Moon (2021), and Auerbach (2022). In these papers, the social interaction models are the main focus of identification and estimation, while the link formation models are used mainly as a control function to deal with network endogeneity or unobserved heterogeneity problems in the social interaction model. We consider our work to be expanding the tool box for researchers interested in this line of research to use to control for the confounding factors outlined above. Additionally, since the homophily parameters and the fixed effects contain important information concerning the causal effects with peer effects or spillover effects (e.g., Lewbel, Qu, and Tang (2023); Jackson, Lin, and Yu (2024)), the method of this paper can be useful to studying such causal effects. Finally, our paper provides ℓ_∞-consistent estimators of the fixed effects, which are the central object of interest of Jochmans and Weidner (2019) who consider a linear regression model with network data and individual fixed effects. Their model and method are very different from ours.

It should be pointed out that in our paper we do not consider link interdependence in network formation, which is actively studied by the line of econometric literature on strategic network formation models and empirical games. This line of literature primarily uses pairwise stability (Jackson and Wolinsky (1996)) as the solution concept for network formation, and also often builds NTU into the econometric specification. See, for example, Miyauchi (2016), de Paula, Richards-Shubik, and Tamer (2018), Leung and Moon (2019); Leung (2019), Boucher and Mourifié (2017), Mele (2017, 2022), Sheng (2020), Ridder and Sheng (2020), Gualdani (2021), Chandrasekhar and Jackson (2023), and Menzel (2024). The authors in these papers study network formation models that account for network externalities, which generate interdependencies in the linking decisions that depend on the structure of the network. However, this type of models either do not feature or impose distributional assumption on unobserved individual heterogeneity. Instead, we adopt the "fixed effect" type of approach and allow them to be arbitrarily correlated with the observable covariates. Furthermore, many of the papers in this line of research require bounded degree (e.g., de Paula, Richards-Shubik, and Tamer (2018)) or appropriate "rate requirements" on the parameters governing the probabilities of subgraphs forming (e.g., Chandrasekhar and Jackson (2023)), which we do not need. Therefore, we consider the two lines of research to be very different. See de Paula, Áureo (2020) for a detailed comparison between these two lines of research.

The rest of the paper is organized as follows. Section 2 formally introduces a dyadic model of link formation with covariates, individual fixed effects, and NTU and presents a set of baseline assumptions. Section 3 presents the estimators for the structural parameters and fixed effects and derives their theoretical properties. Section 4 extends our theory to discuss estimation of APEs and model misspecification. Section 5 includes a simulation study. Section 6 provides two empirical applications. All proofs are relegated to the Appendix.

Notation. Let ":=" denote a definition and superscript "$\top$" denote the transpose of a vector or a matrix. For an $n \times 1$ vector $\mathbf{a} = (a_1, \dots, a_n)^\top$, its ℓ_1 norm is $\|\mathbf{a}\|_1 := \sum_{i=1}^n |a_i|$, ℓ_2 norm is $\|\mathbf{a}\|_2 := (\sum_{i=1}^n a_i^2)^{1/2}$, and ℓ_∞ norm is $\|\mathbf{a}\|_\infty := \max_{1 \le i \le n} |a_i|$. When $O(\cdot)$ (and other notations for order) is written for a vector (or matrix), it means that each element in the vector (or matrix) is of the order in $O(\cdot)$. "plim" refers to the limit in probability, "$\xrightarrow{p}$" stands for convergence in probability, and "$\xrightarrow{d}$" is the convergence in distribution. For an $n \times n$ matrix $\mathbf{A}$, we write $\|\mathbf{A}\|_1 := \max_{1\le i\le n}\|\mathbf{A}_{\cdot i}\|_1$, $\|\mathbf{A}\|_\infty := \max_{1\le i\le n}\|\mathbf{A}_{i\cdot}\|_1$ and $\|\mathbf{A}\|_{\max} := \max_{1\le i,j\le n}|\mathbf{A}_{ij}|$, where $\mathbf{A}_{\cdot i}$ and $\mathbf{A}_{i\cdot}$ are the ith column and row of $\mathbf{A}$, respectively. $\lfloor c\rfloor$ denotes the integer part of any number c. To simplify notation, we write $F_{ij}(\boldsymbol{\alpha},\beta) := F(\alpha_i + x_{ij}^\top\beta)$, $F_{ji}(\boldsymbol{\alpha},\beta) := F(\alpha_j + x_{ji}^\top\beta)$, and $p_{ij}(\boldsymbol{\alpha},\beta) := F_{ij}(\boldsymbol{\alpha},\beta)F_{ji}(\boldsymbol{\alpha},\beta)$. We use F_{ij}, F_{ji}, and p_{ij} when the corresponding functions are evaluated at the true values of $(\boldsymbol{\alpha}_0,\beta_0)$. Finally, the abbreviation "w.p.a.1" stands for "with probability approaching 1."

2 Model and Baseline Assumptions

We consider an undirected[2] network formed among agents $i \in \mathcal{I}_n = \{1,\ldots,n\}$. Hence, there are $N = \binom{n}{2}$ dyads to be linked. Agent i agrees to form a link with j if her utility from the connection is strictly positive. Let binary random variable Z_{ij} denotes agent i's decision on whether to link with j, then

$$Z_{ij} := 1(\alpha_{i0} + X_{ij}^\top\beta_0 - \epsilon_{ij} > 0), \qquad 1 \le i \ne j \le n. \tag{1}$$

We rule out self-loops, i.e., $Z_{ii} \equiv 0,\ i = 1,...,n$. There are three components that determine the value of Z_{ij}: (i) the unobserved fixed effect α_{i0}, which is agent i specific; (ii) dyad-specific index $X_{ij}^\top\beta_0$ that captures homophily effect in the observable characteristics between any ij pairs, where $X_{ij} \in \mathbb{R}^K$ is a symmetric function of agent-level characteristics X_i and X_j, i.e., $X_{ij} = g(X_i, X_j)$ for all $i \ne j$,[3] and (iii) an idiosyncratic component ϵ_{ij} with a known distribution, assumed to be independently and identically distributed across all directed dyads (i,j). [4]

Under NTU, a link Y_{ij} between i and j is formed by the following rule:

$$Y_{ij} := Z_{ij} \cdot Z_{ji}, \qquad 1 \le i \ne j \le n. \tag{2}$$

In this model, the utility of two agents are nontransferable, which is different from the network formation model considered by Graham (2017). Our model is similar to the one

[2]Our model can be extended to cover directed network by introducing two sets of heterogeneity that captures in-degree and out-degree separately as in Yan et al. (2019) and Hughes (2023).

[3]We follow the literature (e.g., Graham (2017)) to make X_{ij} a symmetric function of individual characteristics X_i and X_j. The theoretical results of this paper can be extended to cover the general case in which X_{ij} is some generic pairwise observable characteristics. In the simulations, we randomly draw the first coordinate of X_{ij} directly from Bernoulli distribution without generating X_i nor X_j and find the results to be satisfactory. Nevertheless, allowing for asymmetric $X_{ij} \ne X_{ji}$ could result in more complications and is out of scope of this paper. We leave it for future research.

[4]In principle, we can allow ϵ_{ij} and ϵ_{ji} to be correlated, e.g., bivariate normal distribution with a known correlation coefficient. The analysis goes through with minor changes.

studied by Gao, Li, and Xu (2023), however the focus is different. In this paper, we use the information on the distribution of ϵ_{ij} to estimate β_0 from realizations of a single large network $(y_{ij})_{1\leq i,j\leq n}$ and $(x_{ij})_{1\leq i,j\leq n}$, and conduct asymptotically valid statistical inference.

In what follows, we use bold-case symbols for variables to indicate that its dimension is dependent on n. For example, the dimension of $\boldsymbol{\alpha} = (\alpha_i)_{1\leq i\leq n}$ is $n \times 1$. Let $\mathbf{Y} = (Y_{ij})_{1\leq i,j\leq n}$ and $\mathbf{X} = (X_{ij})_{1\leq i,j\leq n}$ be the $n\times n$ adjacency matrix and $n\times n\times k$ random tensor of covariates, respectively. Denote their realizations by $\mathbf{y} = (y_{ij})_{1\leq i,j\leq n}$ and $\mathbf{x} = (x_{ij})_{1\leq i,j\leq n}$, respectively. Let $\boldsymbol{\alpha}_0 = (\alpha_{i0})_{1\leq i\leq n}$. Unless otherwise stated, we maintain the following assumption.

Assumption 1 (Correctly Specified Model). *The conditional likelihood of $\mathbf{Y} = \mathbf{y}$ given $\mathbf{X} = \mathbf{x}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ is*

$$\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, \boldsymbol{\alpha} = \boldsymbol{\alpha}_0) = \prod_{i=1}^{n}\prod_{j>i}\Pr(Y_{ij} = y_{ij}|x_i, x_j, \alpha_{i0}, \alpha_{j0}), \tag{3}$$

where

$$\begin{aligned}\Pr(Y_{ij} = y_{ij}|x_i, x_j, \alpha_{i0}, \alpha_{j0}) = &\left[F(\alpha_{i0} + x_{ij}^\top\beta_0)F(\alpha_{j0} + x_{ji}^\top\beta_0)\right]^{y_{ij}} \\ &\times\left[1 - F(\alpha_{i0} + x_{ij}^\top\beta_0)F(\alpha_{j0} + x_{ji}^\top\beta_0)\right]^{1-y_{ij}},\end{aligned} \tag{4}$$

for all $i \neq j$ where $F(\alpha_{i0} + x_{ij}^\top\beta_0) := \Pr(Z_{ij} = 1|x_{ij}, \alpha_{i0})$ and $F(\cdot)$ is known.

Assumption 1 is similar to Assumption 1 of Graham (2017) except for two important differences. First, under NTU $\alpha_{i0} + x_{ij}^\top\beta_0$ and $\alpha_{j0} + x_{ji}^\top\beta_0$ are not additively separable in the linking probability between i and j, thus the tetrad logit estimator of Graham (2017) does not apply in our setting. Instead, we propose a one-step split-network jackknife bagging estimator that works with NTU and achieves the Cramér-Rao efficiency bound asymptotically. Second, we do not specify the functional form of $F(\cdot)$ to be logistic as in the literature (e.g., Chatterjee, Diaconis, and Sly (2011); Graham (2017); Qu, Chen, Yan, and Chen (2024)). Instead, we impose mild conditions on $F(\cdot)$ for our asymptotic analysis in Assumption 3. Most common distributions including logistic and normal distribution satisfy Assumption 3. In this regard, we consider Assumption 1 to be more general.

Assumption 1 requires that the ϵ_{ij}'s are i.i.d. across dyads (i, j), i.e., links are formed independently of one another conditional on agent attributes. The assumption that links form independently could be plausible in certain settings, such as risk-sharing networks, online friendships, trade networks, and conflicts between nation-states. However, it rules out link interdependencies and thus may not be applicable to certain types of networks with explicit strategic interactions such as supply chain networks. See the discussion of Assumption 1 of Graham (2017) for more details on this issue.

Assumption 1 also requires the link function $F(\cdot)$ to be correctly specified. It is well-known that under regularity conditions MLE will converge to the point that minimizes the Kullback-Leibler information criterion between the true and the misspecified model (White (1982)). The issue is further complicated by the high-dimensional individual fixed effects and NTU of model (4). To our knowledge, the misspecification issue has not been investigated in the network formation literature. We discuss how model misspecification affects our theoretical results in Section 4.2 and provide simulation evidence for our estimator under misspecification in Section 5.

The next two assumptions are needed to facilitate our asymptotic analysis in the next section.

Assumption 2 (Bounded Support and Random Sampling)**.** *Suppose the following conditions hold:*

(a) *The true value of agent-level heterogeneity $\boldsymbol{\alpha}_0$ lies in the interior of a compact set $\mathbb{A} \subset \mathbb{R}^n$.*

(b) *The true value of structural parameter β_0 lies in the interior of a compact set $\mathbb{B} \subset \mathbb{R}^K$.*

(c) *Dyad-level characteristics X_{ij} satisfy $X_{ij} \in \mathbb{X} \subset \mathbb{R}^K$ for some compact set $\mathbb{X}$.*

Assumption 2 collects and combines Assumptions 2 and 5(i) of Graham (2017). Assumption 2(a)–(c) collectively imply that probability of linking between dyad (i, j) is uniformly bounded in $[\kappa, 1-\kappa]$ for some $\kappa \in (0, 1)$, which essentially requires the network to be dense.[5] The dense network condition is needed so that we can estimate α_{i0} consistently for each i.[6] It is similar to the estimation of individual fixed effect in large T panel data models. In simulations we find that asymptotic results hold when the network is fairly sparse. See Section 5.3 for the simulation results for networks with a density smaller than .1. Note that our theory in principle can allow the support of X_{ij} to be unbounded; however, it would introduce more technical complications in deriving the rates of convergence without adding much theoretical insight since it involves handling the complicated tail behaviors of the random variables. Thus, we impose Assumption 2(c), which is similarly assumed in Assumption 2(ii) of Graham (2017), to focus on the main idea of the paper.

[5]Density of an undirected network is defined as $\rho_n = N^{-1} \sum_{i=1}^{n} \sum_{j>i} y_{ij}$, where $N = \binom{n}{2}$. A network is dense if $\lim_{n\to\infty} \rho_n \geq c > 0$ for some constant c.

[6]The Beta models with logistic link functions studied by Chatterjee, Diaconis, and Sly (2011), Graham (2017), and Yan, Jiang, Fienberg, and Leng (2019) point out that Assumption 2(a) could be relaxed to be $\|\boldsymbol{\alpha}_0\|_\infty = O(\log(\log n))$, which allows for sparser networks (but nearly dense). However, this result does not extend to our setting directly due to NTU and general $F(\cdot)$. Thus, we require dense network in Assumption 2.

Assumption 3 (Restrictions on $F(\cdot)$)**.** *$F(\cdot)$ is three-times continuously differentiable with its first to third derivatives $f(\cdot)$, $f^{(1)}(\cdot)$, and $f^{(2)}(\cdot)$ satisfying*

$$\begin{aligned} F(\alpha_i + x_{ij}^\top \beta) &\in [c_1, 1 - c_1], \\ f(\alpha_i + x_{ij}^\top \beta) &\in [c_2, 1 - c_2], \\ |f^{(1)}(\alpha_i + x_{ij}^\top \beta)| &\leq c_3, \ \ \textit{and} \\ |f^{(2)}(\alpha_i + x_{ij}^\top \beta)| &\leq c_4, \end{aligned}$$

for some constants $c_1, c_2 \in (0, 1/2]$, $c_3, c_4 > 0$ and all $(\boldsymbol{\alpha}, \beta) \in \mathbb{A} \times \mathbb{B}$, $x_{ij} \in \mathbb{X}$, $1 \leq i \neq j \leq n$.

Assumption 3 puts bounds on $F(\cdot)$ and its derivatives. We consider Assumption 3 to be mild because it is generally satisfied for common distributions such as logistic and normal distributions when combined with Assumption 2. This Assumption is similar to Assumption 4.3(v) of Fernández-Val and Weidner (2016), which restricts the smoothness of the likelihood functions. We use Assumption 2 to bound the norm of the Jacobian matrix of the moment conditions to be presented in the next section.

3 Estimation and Large Sample Properties

In this section, we introduce the one-step split-jackknife bagging estimator for β_0 and prove that it is $\sqrt{N}$-consistent and achieves the Cramér-Rao lower bound asymptotically. The idea is based on the one–step approximation of the MLE of Le Cam (1969). However, there are three challenges in applying Le Cam (1969)'s idea to our setting. First, Le Cam (1969)'s one-step estimator requires an initial estimator of β_0. Second, with the inclusion of the high-dimensional fixed effects $\boldsymbol{\alpha}_0$ into the model (4), the one-step estimator proposed by Le Cam (1969)—even if successfully constructed—will be $\sqrt{N}$-inconsistent[7] with a complicated nonzero mean in the limit after normalization, a problem also shared by the original MLE of β_0 when incidental parameters exist in the model. Moreover, popular debiasing methods in the literature such as jackknife will inflate the variance of the estimator, making it asymptotically inefficient.

We provide intuition on how we deal with these challenges here. To address the initial estimator problem, we propose a joint method of moments (JMM) estimator for β_0 (and the high-dimensional fixed effects $\boldsymbol{\alpha}_0$). We first estimate the high-dimensional fixed effects $\boldsymbol{\alpha}$ as a function of β. The main deviation from the existing methods (e.g., Theorem 1.5 of Chatterjee, Diaconis, and Sly (2011) or the fixed point equation (17) of Graham (2017)) in this step is that we do not rely on ϵ_{ij} being logistic random variable nor the link formation

[7] By $\sqrt{N}$-inconsistent, we mean that $\sqrt{N}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(\mu, V)$ as $n \to \infty$ for some $\mu \neq 0$.

process being TU. Instead, we construct a set of moment conditions that involves the average degree sequence for each individual and use a new iterative algorithm to compute $\widehat{\boldsymbol{\alpha}}(\beta)$ as a function of β. The new iterative algorithm is inspired by the stochastic gradient descents method (SGD, Robbins and Monro (1951)).[8] We prove existence and uniqueness of $\widehat{\boldsymbol{\alpha}}(\beta)$, the solution to a high-dimensional system of moment equations. Our new method can handle a general class of distributions, not only logistic, as well as NTU.

Once $\widehat{\boldsymbol{\alpha}}(\beta)$ is obtained, we do not maximize the concentrated log-likelihood function to estimate β_0 because it is hard to verify concavity of the concentrated criterion function with $\widehat{\boldsymbol{\alpha}}(\beta)$ plugged in and the maximization can be computationally unstable. Instead, we construct a new set of finite-dimensional moment conditions to compute $\hat{\beta}_{\mathrm{JMM}}$, the JMM estimator for β_0. We obtain asymptotic distribution for $\hat{\beta}_{\mathrm{JMM}}$ and prove uniform consistency for $\widehat{\boldsymbol{\alpha}}\left(\hat{\beta}_{\mathrm{JMM}}\right)$. Then, we plug $\hat{\beta}_{\mathrm{JMM}}$ into the one-step estimator of Le Cam (1969) to obtain $\hat{\beta}_{\mathrm{OS}}$ and prove its asymptotic normality. Since we have an analytical expression for $\hat{\beta}_{\mathrm{OS}}$, there is no need to maximize the log-likelihood function and consequently the computation is stable and very fast.

To deal with the second issue of the asymptotic bias in $\hat{\beta}_{\mathrm{OS}}$, we use the bagging method with split-network jackknife to debiase $\hat{\beta}_{\mathrm{OS}}$ and maintain its efficiency simultaneously. We prove that the new one-step split-network jackknife bagging estimator, denoted by $\widehat{\beta}_{\mathrm{BG}}$, is asymptotically normal, $\sqrt{N}$-consistent, and attains the Cramér-Rao lower bound asymptotically. A graphical illustration of the bagging procedure is provided in Figure 1.

3.1 Joint Method of Moments Estimator

We propose a new joint method of moments estimators for both $\boldsymbol{\alpha}_0$ and β_0. There are $n+K$ unknown parameters in our model (4), which requires at least $n+K$ moment equations for identification. To deal with the challenges caused by the high-dimensionality of $\boldsymbol{\alpha}_0$, we construct two sets of moment conditions and estimate the parameters sequentially. For each candidate β, we use the first set of n moment conditions to estimate $\boldsymbol{\alpha}_0$ as a function of β. Then, we plug the estimated $\widehat{\boldsymbol{\alpha}}(\beta)$ into the second set of K moment equations to estimate β_0, obtaining $\hat{\beta}_{\mathrm{JMM}}$.[9]

Define $d_i := \sum_{j \neq i} Y_{ij}$ as the degree sequence for $i = 1, ..., n$ of the observed network $\mathbf{Y}$.

[8]Strictly speaking, the algorithm does not comply with the standard definition of SGD as the Jacobian matrix of the moment conditions is not symmetric. Nonetheless, we use it to motivate the construction of the moment estimators of the high-dimensional $\boldsymbol{\alpha}_0$.

[9]In what follows, we use $\widehat{\beta}$ to denote $\hat{\beta}_{\mathrm{JMM}}$ for notational simplicity whenever there is no confusion.

The first set of moment conditions concerns the average degree for each individual i:

$$\mathbb{E}\left[\left(d_1 - \sum_{j\neq 1} p_{1j}\left(\boldsymbol{\alpha}_0, \beta_0\right), \cdots, d_n - \sum_{j\neq n} p_{nj}\left(\boldsymbol{\alpha}_0, \beta_0\right)\right)^\top \middle| \mathbf{X} = \mathbf{x}\right] = 0. \tag{5}$$

The moment condition (5) is used to obtain $\widehat{\boldsymbol{\alpha}}$ as a function of β, i.e., $\widehat{\boldsymbol{\alpha}}\left(\beta\right)$, via a SGD type algorithm.[10] Specifically, we use the following high-dimensional system of equations

$$\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = \left(d_1 - \sum_{j\neq 1} p_{1j}(\boldsymbol{\alpha}, \beta), \cdots, d_n - \sum_{j\neq n} p_{nj}(\boldsymbol{\alpha}, \beta)\right)^\top, \tag{6}$$

and provide an algorithm to find $\widehat{\boldsymbol{\alpha}}(\beta)$, the root of (6) as a function of β. Let

$$r_i(\boldsymbol{\alpha}, \beta) = \alpha_i + (n-1)^{-1}\left(d_i - \sum_{j\neq i} p_{ij}(\boldsymbol{\alpha}, \beta)\right), \qquad i = 1, \ldots, n \tag{7}$$

and $\mathbf{r}(\boldsymbol{\alpha}, \beta) = \left(r_1(\boldsymbol{\alpha}, \beta), \ldots, r_n(\boldsymbol{\alpha}, \beta)\right)^\top$. The intuition is, for any i when d_i is strictly larger than $\sum_{j\neq i} p_{ij}(\boldsymbol{\alpha}, \beta)$, we would like to increase α_i such that each $p_{ij}(\boldsymbol{\alpha}, \beta)$ for $j \neq i$ is larger, and vice versa. The validity of the argument is guaranteed by the definition $p_{ij}(\boldsymbol{\alpha}, \beta) = F(\alpha_i + x_{ij}^\top \beta) F(\alpha_j + x_{ji}^\top \beta)$ and the monotonicity of $F\left(\cdot\right)$ by Assumption 3. Then, starting with an arbitrary initial value $\boldsymbol{\alpha}^0$, we update $\boldsymbol{\alpha}^{k+1}(\beta) = \mathbf{r}(\boldsymbol{\alpha}^k(\beta), \beta)$ into the next iteration.

Observe that $\mathbf{m}_1(\boldsymbol{\alpha}, \beta)$ is a high-dimensional system of equations that features NTU, covariates, and possibly non-logistic link functions. In the next theorem, we prove that $\widehat{\boldsymbol{\alpha}}(\beta)$ exists and is unique for each β around β_0 with high probability as n grows.

Theorem 1. *Under Assumptions 1–3, as $n \to \infty$, w.p.a.1 $\widehat{\boldsymbol{\alpha}}(\beta)$ exists and is unique for each $\beta \in \{\beta \in \mathbb{B} | \left\|\beta - \beta_0\right\|_2 < c\}$ for some constant $c > 0$. Moreover, uniformly across all k, we have*

$$\|\boldsymbol{\alpha}^{k+2}(\beta) - \boldsymbol{\alpha}^{k+1}(\beta)\|_1 \leq \delta \|\boldsymbol{\alpha}^k(\beta) - \boldsymbol{\alpha}^{k-1}(\beta)\|_1 \quad \text{and}$$
$$\|\boldsymbol{\alpha}^{k+2}(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1 \leq \delta \|\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1,$$

for some fixed constant $\delta \in (0, 1)$.

We present the proofs of all the theoretical results of this paper in Appendix B. Theorem 1 guarantees that $\widehat{\boldsymbol{\alpha}}(\beta) = \lim_{k\to\infty} \boldsymbol{\alpha}^k(\beta)$ and that the ℓ_1-distance between $\widehat{\boldsymbol{\alpha}}(\beta)$ and $\boldsymbol{\alpha}^k(\beta)$ decreases geometrically after each two iterates. We find that computing $\widehat{\boldsymbol{\alpha}}(\beta)$ is fast in the simulations, which is another advantage of our SGD algorithm.

[10] It is worth emphasizing that (5) holds without assuming the distribution of ϵ_{ij} to be logistic, a condition that is needed to obtain the fixed point equation of (17)–(18) of Graham (2017) or Theorem 1.5 of Chatterjee, Diaconis, and Sly (2011). Therefore, we consider (5) to be more general. Moreover, even if ϵ_{ij} is logistic random variable, the nonseparability of p_{ij} in α_i and α_j brought about by NTU makes the fixed point equation of (17)–(18) of Graham (2017) or Theorem 1.5 of Chatterjee, Diaconis, and Sly (2011) not applicable.

Next, we impose a finite-dimensional orthogonality condition between X_{ij} and $(Y_{ij} - p_{ij})$ over the population of (i, j) dyads:

$$\mathbb{E}\left[(Y_{ij} - p_{ij}(\boldsymbol{\alpha}_0, \beta_0)) X_{ij} \middle| \mathbf{X} = \mathbf{x}\right] = 0. \tag{8}$$

Let

$$m_2(\boldsymbol{\alpha}, \beta) = \sum_{i=1}^{n} \sum_{j>i} \left[y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)\right] x_{ij}. \tag{9}$$

Given $\widehat{\boldsymbol{\alpha}}(\beta)$, we solve a concentrated moment equation for $\hat{\beta}$:

$$m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta) = 0. \tag{10}$$

Notice that (10) is finite-dimensional, hence it is easy to compute. It is clear that $(\widehat{\boldsymbol{\alpha}}, \hat{\beta}) = \left(\widehat{\boldsymbol{\alpha}}(\hat{\beta}), \hat{\beta}\right)$.

Before we state the consistency and asymptotic normality results for $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$, we combine $\mathbf{m}_1$ and m_2 into $\mathbf{m}$ for easier exposition of the results:

$$\mathbf{m}(\boldsymbol{\alpha}, \beta) := \left(d_1 - \sum_{j \neq 1} p_{1j}(\boldsymbol{\alpha}, \beta), \cdots, d_n - \sum_{j \neq n} p_{nj}(\boldsymbol{\alpha}, \beta), \sum_{i=1}^{n} \sum_{j>i} \left[y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)\right] x_{ij}\right)^{\top}. \tag{11}$$

Then $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ is one root of (11). Although $\widehat{\boldsymbol{\alpha}}(\beta)$ is unique by Theorem 1 for any β around β_0, in principle there could be multiple $\hat{\beta}$ that solves (10). However, the next identification condition guarantees that any $\hat{\beta}$ that solves (10) is consistent for β_0.

Assumption 4 (Identification)**.** *Define the concentrated moment equation*

$$\bar{S}_n(\beta) := \binom{n}{2}^{-1} \mathbb{E}[m_2(\boldsymbol{\alpha}(\beta), \beta) | \mathbf{x}, \boldsymbol{\alpha}_0],$$

where $\boldsymbol{\alpha}(\beta)$ is the unique root of $\mathbb{E}[\mathbf{m}_1(\boldsymbol{\alpha}, \beta)|\mathbf{x}, \boldsymbol{\alpha}_0] = 0$ for each $\beta \in \{\beta \in \mathbb{B} | \, \|\beta - \beta_0\|_2 < c\}$. Suppose that for all $\delta > 0$ and for n large enough

$$\inf_{\beta \in \mathbb{B}: \|\beta - \beta_0\|_2 \geq \delta} \left\|\bar{S}_n(\beta)\right\|_2 > 0. \tag{12}$$

Recall that Theorem 1 guarantees that the high-dimensional moment equation $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = 0$ always has a unique solution if $\|\beta - \beta_0\|_2 < c$ for some constant c. Assumption 4 is a local identification condition for the low-dimensional parameter β_0, as is extensively discussed in Chen, Chernozhukov, Lee, and Newey (2014) for nonlinear models with high-dimensional nuisance parameters. Condition (12) is equivalent to assuming that β_0 is the unique solution to $\bar{S}_n(\beta) = 0$, which is similar to the widely imposed "unique minimizer" condition in M-estimators literature, see Chapter 5 (pg. 45) of Van der Vaart (2000).

To better understand Assumption (4), consider a (low-dimensional) linear panel data model with individual fixed effects, $y_{it} = \alpha_{i0} + x_{it}^\top \beta_0 + \epsilon_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T$. Suppose $\mathbb{E}[(1, x_{it}^\top)^\top \epsilon_{it}] = 0$ in this model. Then, the expected concentrated moment function is $\bar{S}_n(\beta) = (nT)^{-1}\mathbb{E}\left\{\sum_{i,t}\left[y_{it} - \alpha_i(\beta) - x_{it}^\top\beta\right] x_{it}\right\}$, where $\alpha_i(\beta) = \alpha_{i0} + T^{-1}\sum_t x_{it}^\top(\beta_0 - \beta)$ is the solution to $\mathbb{E}\left[\sum_t (y_{it} - x_{it}^\top\beta - \alpha_i)\right] = 0,\ i = 1, \ldots, n$. Then $\bar{S}_n(\beta) = (nT)^{-1}\sum_{i,t}(x_{it} - \bar{x}_i)\left(x_{it} - \bar{x}_i\right)^\top(\beta - \beta_0)$ with $\bar{x}_i = T^{-1}\sum_t x_{it}$. Consequently, a sufficient condition for (12) in this linear panel model example is that the smallest eigenvalue of $(nT)^{-1}\sum_{i,t}(x_{it} - \bar{x}_i)\left(x_{it} - \bar{x}_i\right)^\top$ (which is also the concentrated Jacobian matrix for β) is strictly large than 0, which is a quite weak identification condition on the design matrix.

We prove that the JMM estimator $\hat{\beta}$ converges to β_0 in probability as $n \to \infty$. For the high-dimensional fixed effects, $\widehat{\boldsymbol{\alpha}}$ is consistent in the sense that $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_\infty \xrightarrow{p} 0$. We summarize these results in the next theorem.

Theorem 2. *If Assumptions 1–4 hold, then*

$$\hat{\beta} \xrightarrow{p} \beta_0, \quad \text{and} \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_\infty \xrightarrow{p} 0, \quad \text{as } n \to \infty.$$

To state the limit distribution of $\hat{\beta}$, we introduce more definitions on the Jacobian and covariance matrices of moment equations (11). Let

$$\mathbf{J}(\boldsymbol{\alpha}, \beta) := \nabla \mathbf{m}(\boldsymbol{\alpha}, \beta) = \begin{pmatrix} \nabla_{\boldsymbol{\alpha}^\top}\mathbf{m}_1(\boldsymbol{\alpha}, \beta) & \nabla_{\beta^\top}\mathbf{m}_1(\boldsymbol{\alpha}, \beta) \\ \nabla_{\boldsymbol{\alpha}^\top} m_2(\boldsymbol{\alpha}, \beta) & \nabla_{\beta^\top} m_2(\boldsymbol{\alpha}, \beta) \end{pmatrix} =: \begin{pmatrix} \mathbf{J}_{11}(\boldsymbol{\alpha}, \beta) & \mathbf{J}_{12}(\boldsymbol{\alpha}, \beta) \\ \mathbf{J}_{21}(\boldsymbol{\alpha}, \beta) & \mathrm{J}_{22}(\boldsymbol{\alpha}, \beta) \end{pmatrix}.$$

be the Jacobian matrix of $\mathbf{m}(\boldsymbol{\alpha}, \beta)$, where we separate it into four blocks according to the variables of differentiation. In Appendix A, we give explicit expressions of these blocks. It is worth emphasizing that $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta) \neq \mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)^\top$ and $\mathbf{J}_{12}(\boldsymbol{\alpha}, \beta) \neq \mathbf{J}_{21}(\boldsymbol{\alpha}, \beta)^\top$, thus $\mathbf{J}(\boldsymbol{\alpha}, \beta)$ is asymmetric. The asymmetry implies that $\mathbf{m}(\boldsymbol{\alpha}, \beta)$ can not be written as a gradient function of any scalar-valued criterion function.

The concentrated Jacobian matrix for β is defined as

$$\mathrm{J}_n(\beta) := \frac{\partial m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta)}{\partial \beta} = \mathrm{J}_{22}(\widehat{\boldsymbol{\alpha}}(\beta), \beta) - \mathbf{J}_{21}(\widehat{\boldsymbol{\alpha}}(\beta), \beta)\mathbf{J}_{11}^{-1}(\widehat{\boldsymbol{\alpha}}(\beta), \beta)\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\beta), \beta).$$

Then, we let

$$\mathbf{V} := \mathrm{Var}\big(\mathbf{m}(\boldsymbol{\alpha}, \beta)|\mathbf{x}, \boldsymbol{\alpha}_0\big) = \begin{pmatrix} \mathrm{Var}(\mathbf{m}_1) & \mathrm{Cov}(\mathbf{m}_1, m_2) \\ \mathrm{Cov}(\mathbf{m}_1, m_2)^\top & \mathrm{Var}(m_2) \end{pmatrix} := \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{12}^\top & \mathrm{V}_{22} \end{pmatrix}$$

be the covariance matrix of $\mathbf{m}(\boldsymbol{\alpha}, \beta)$. As we show in Appendix A, $\mathbf{V}$ does not depend on unknown parameter $(\boldsymbol{\alpha}, \beta)$ because it is a covariance matrix which cancels out the unknown parameters by demeaning. Define

$$B_{k0} = \lim_{n\to\infty} \frac{1}{2\sqrt{N}} \mathrm{Tr}\left[\mathbf{J}_{11}^{-1}\mathbf{V}_{11}\left(\mathbf{J}_{11}^{-1}\right)^\top \mathbf{R}_k\right] \tag{13}$$

where $\mathbf{R}_k$ is defined by (68) in the Appendix B. Let $B_0 = (B_{10}, \ldots, B_{K0})^\top$ and the limiting variance matrix be

$$\Omega_0 := \lim_{n\to\infty} N^{-1}\mathrm{J}_0^{-1}\left[\mathrm{V}_{22} + \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{11}(\mathbf{J}_{21}\mathbf{J}_{11}^{-1})^\top - \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{12} - (\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{12})^\top\right](\mathrm{J}_0^{-1})^\top, \tag{14}$$

where J_0 is the probability limit of $N^{-1}\mathrm{J}_n(\beta_0)$. We discuss J_0 in more details in Appendix B. If $\boldsymbol{\alpha}_0$ is known, the asymptotic variance of $\hat{\beta} - \beta_0$ becomes $\mathrm{J}_{22}^{-1}\mathrm{V}_{22}\mathrm{J}_{22}^{-1}$ and the additional terms shown in (14) are caused by estimating $\boldsymbol{\alpha}_0$.

Theorem 3. *If Assumptions 1–4 are satisfied, then*

$$\sqrt{N}(\hat{\beta} - \beta_0) - \mathrm{J}_0^{-1}B_0 \xrightarrow{d} \mathcal{N}(0, \Omega_0).$$

Theorem 3 shows that the JMM estimator $\hat{\beta}$ is normal asymptotically, however the limiting distribution does not center around zero. The bias term $\mathrm{J}_0^{-1}B_0$ arises from estimating $\boldsymbol{\alpha}_0$. Incidental parameter problem is common in the literature of non-linear panel fixed effects regression with large N and T (e.g., Neyman and Scott (1948); Hahn and Newey (2004)).

Remark 1. It is possible to correct for the bias for $\hat{\beta}$ by the split-network jackknife (SJ), which is inspired by the split-panel jackknife proposed by Dhaene and Jochmans (2015). The idea is to split the set of agents $\{1, ..., n\}$ randomly and equally into two disjoint subsets and estimate β_0 twice (denoted by $\hat{\beta}_1$ and $\hat{\beta}_2$) using each sub-network formed by the agents from each subset. Then, the JMM SJ estimator is $\hat{\beta}_{\mathrm{SJ}} := 2\hat{\beta} - \frac{1}{2}\left(\hat{\beta}_1 + \hat{\beta}_2\right)$. We prove in Appendix B.4 that $\hat{\beta}_{\mathrm{SJ}}$ is $\sqrt{N}$-consistent with $2\Omega_0$ being the asymptotic variance. The inflated variance arises because we ignore the links between agents from different subsets when computing $\hat{\beta}_1$ and $\hat{\beta}_2$.

3.2 One-Step Estimator

Under Assumption 1, the log-likelihood function of $(\boldsymbol{\alpha}, \beta)$ is

$$\begin{aligned}\ell_n(\boldsymbol{\alpha}, \beta) := \sum_{i=1}^{n}\sum_{j>i}\{&y_{ij}\log\left[F(\alpha_i + x_{ij}^\top\beta)F(\alpha_j + x_{ji}^\top\beta)\right] \\ &+ (1 - y_{ij})\log\left[1 - F(\alpha_i + x_{ij}^\top\beta)F(\alpha_j + x_{ji}^\top\beta)\right]\}.\end{aligned} \tag{15}$$

By maximizing (15) with respect to $\boldsymbol{\alpha}$ and β simultaneously, the maximum likelihood estimator (MLE) is

$$(\widehat{\boldsymbol{\alpha}}_{\text{MLE}}, \hat{\beta}_{\text{MLE}}) := \arg \max_{(\boldsymbol{\alpha},\beta)\in\mathbb{A}\times\mathbb{B}} \ell_n(\boldsymbol{\alpha}, \beta),$$

which can be equivalently defined via maximizing a concentrated log-likelihood function

$$\hat{\beta}_{\text{MLE}} := \arg\max_{\beta\in\mathbb{B}} \ell_n(\widehat{\boldsymbol{\alpha}}_{\text{MLE}}(\beta), \beta), \qquad \text{where } \widehat{\boldsymbol{\alpha}}_{\text{MLE}}(\beta) := \arg\max_{\boldsymbol{\alpha}\in\mathbb{A}} \ell_n(\boldsymbol{\alpha}, \beta).$$

Moreover, $(\widehat{\boldsymbol{\alpha}}_{\text{MLE}}, \hat{\beta}_{\text{MLE}})$ is also the root of the score function of $\ell_n(\boldsymbol{\alpha}, \beta)$, i.e., $\mathbf{s}(\boldsymbol{\alpha}, \beta) = \left(\mathbf{s}_1^\top(\boldsymbol{\alpha}, \beta), s_2^\top(\boldsymbol{\alpha}, \beta)\right)^\top$, where

$$\begin{aligned}
\mathbf{s}_1(\boldsymbol{\alpha}, \beta) &:= \frac{\partial \ell_n(\boldsymbol{\alpha}, \beta)}{\partial \boldsymbol{\alpha}} = \left(\sum_{j\neq 1} \tfrac{f_{1j}(\boldsymbol{\alpha},\beta)[y_{1j}-p_{1j}(\boldsymbol{\alpha},\beta)]}{F_{1j}(\boldsymbol{\alpha},\beta)[1-p_{1j}(\boldsymbol{\alpha},\beta)]}, \cdots, \sum_{j\neq n} \tfrac{f_{nj}(\boldsymbol{\alpha},\beta)[y_{nj}-p_{nj}(\boldsymbol{\alpha},\beta)]}{F_{nj}(\boldsymbol{\alpha},\beta)[1-p_{nj}(\boldsymbol{\alpha},\beta)]}\right)^\top, \\
s_2(\boldsymbol{\alpha}, \beta) &:= \frac{\partial \ell_n(\boldsymbol{\alpha}, \beta)}{\partial \beta} = \sum_{i=1}^n \sum_{j\neq i} \frac{f_{ij}(\boldsymbol{\alpha}, \beta)[y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)]}{F_{ij}(\boldsymbol{\alpha}, \beta)[1 - p_{ij}(\boldsymbol{\alpha}, \beta)]} x_{ij}.
\end{aligned}$$

It has been well documented in the literature that log-likelihood function can be non-concave over the parameter space. The non-concavity issue is further exacerbated by the high dimensionality of $\boldsymbol{\alpha}_0$ of this paper. For example, to compute $\widehat{\boldsymbol{\alpha}}_{\text{MLE}}(\beta)$, we need to maximize $\ell_n(\boldsymbol{\alpha}, \beta)$ with respect to $\boldsymbol{\alpha}$, an n-dimensional object, for each fixed β. This is numerically challenging. Instead, we propose a new estimator based on Le Cam's one-step approximation (Le Cam (1969)) which does not require concavity of the log-likelihood function, is easy to compute, and achieves the Cramér-Rao lower bound asymptotically.

To define the one-step estimator, we introduce more definitions on the Hessian matrix and the information matrix. The Hessian of $\ell_n(\boldsymbol{\alpha}, \beta)$ is defined as

$$\mathbf{H}(\boldsymbol{\alpha}, \beta) := \begin{pmatrix} \nabla_{\boldsymbol{\alpha}^\top} \mathbf{s}_1(\boldsymbol{\alpha}, \beta) & \nabla_{\beta^\top} \mathbf{s}_1(\boldsymbol{\alpha}, \beta) \\ \nabla_{\boldsymbol{\alpha}^\top} s_2(\boldsymbol{\alpha}, \beta) & \nabla_{\beta^\top} s_2(\boldsymbol{\alpha}, \beta) \end{pmatrix} := \begin{pmatrix} \mathbf{H}_{11}(\boldsymbol{\alpha}, \beta) & \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta) \\ \mathbf{H}_{12}^\top(\boldsymbol{\alpha}, \beta) & \mathrm{H}_{22}(\boldsymbol{\alpha}, \beta) \end{pmatrix}.$$

Define the information matrix as

$$\mathbf{I}(\boldsymbol{\alpha}, \beta) = -\mathbb{E}_\beta[\mathbf{H}(\boldsymbol{\alpha}, \beta)|\mathbf{x}, \boldsymbol{\alpha}] = \mathbb{E}_\beta[\mathbf{s}(\boldsymbol{\alpha}, \beta)\mathbf{s}(\boldsymbol{\alpha}, \beta)^\top|\mathbf{x}, \boldsymbol{\alpha}],$$

where $\mathbb{E}_\beta$ means taking expectation conditional on the population parameter being equal to β. We partition $\mathbf{I}(\boldsymbol{\alpha}, \beta)$ into four submatrices similarly as before. Following Chapter 4.2 of Amemiya (1985), we define the concentrated information matrix of β,

$$\mathrm{I}_n(\boldsymbol{\alpha}, \beta) = \mathrm{I}_{22}(\boldsymbol{\alpha}, \beta) - \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta). \tag{16}$$

The concentrated score function is defined as

$$s_n(\boldsymbol{\alpha}, \beta) = s_2(\boldsymbol{\alpha}, \beta) - \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{s}_1(\boldsymbol{\alpha}, \beta). \tag{17}$$

Finally, we define the one-step estimator $\hat{\beta}_{\mathrm{OS}}$ as

$$\hat{\beta}_{\mathrm{OS}} = \hat{\beta} + \mathrm{I}_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta})^{-1} s_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta}), \tag{18}$$

with the joint moment estimator $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ from Section 3.1.

Algebra shows

$$\mathbb{E}\left[\frac{\partial s_n(\boldsymbol{\alpha}, \beta_0)}{\partial \boldsymbol{\alpha}}\Big|\mathbf{x}, \boldsymbol{\alpha}_0\right] = \mathbf{0}_n, \qquad \mathbb{E}\left[\frac{\partial s_n(\boldsymbol{\alpha}, \beta_0)}{\partial \beta}\Big|\mathbf{x}, \boldsymbol{\alpha}_0\right] = -\mathrm{I}_n.$$

Therefore, by a Taylor expansion on the right hand side of (18), we have

$$\hat{\beta}_{\mathrm{OS}} - \beta_0 \approx \mathrm{I}_n(\boldsymbol{\alpha}_0, \beta_0)^{-1} s_n(\boldsymbol{\alpha}_0, \beta_0) \tag{19}$$

in large samples. To establish (19) rigorously and hence the asymptotic normality of $\hat{\beta}_{\mathrm{OS}}$, we need an additional assumption on the behavior of the information matrix. Let $w_{ki}(\boldsymbol{\alpha}, \beta) = \left[\mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1}\right]_{ki}$, the ki^{th} element of $\mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1}$.

Assumption 5. *For $(\boldsymbol{\alpha}, \beta) \in \mathbb{A} \times \mathbb{B}$, $1 \le k \le K$, and $1 \le i \ne j \le n$, $\sup_{k,i} |w_{ki}(\boldsymbol{\alpha}, \beta)|$ is $O(1)$ and continuously differentiable. Furthermore, the following conditions on $w_{ki}(\boldsymbol{\alpha}, \beta)$ are satisfied:*

(a) $\sup_{k,i} \|\partial w_{ki}(\boldsymbol{\alpha}, \beta)/\partial \beta\| = O(1)$,

(b) $\sup_{k,i} |\partial w_{ki}(\boldsymbol{\alpha}, \beta)/\partial \alpha_i| = O(1)$,

(c) $\sup_{k,i,j} |\partial w_{ki}(\boldsymbol{\alpha}, \beta)/\partial \alpha_j| = O(n^{-1})$, $\quad i \ne j$.

Assumption 5 is quite mild given that $\sup_{k,i} |w_{ki}(\boldsymbol{\alpha}, \beta)| = O(1)$ is assumed, which concerns the ki^{th} element of $\mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1}$, a well-defined object. To gain some intuition about Assumption 5(c), consider a classical linear panel data model with additive individual fixed effects. If there is no interaction between i and j, $w_{ki}(\boldsymbol{\alpha}, \beta)$ will only dependent on α_i and β, hence $|\partial w_{ki}(\boldsymbol{\alpha}, \beta)/\partial \alpha_j| = 0$, satisfying Assumption 5(c). Therefore, Assumption 5(c) is used to control for how much the function $w_{ki}(\boldsymbol{\alpha}, \beta)$ depends on α_j for $j \ne i$.

Let $\mathbf{W}_k$ be the $n \times n$ matrix of these derivatives with $(\mathbf{W}_k)_{ij} = \partial w_{ki}(\boldsymbol{\alpha}_0, \beta_0)/\partial \alpha_j$. Since we have estimated $\boldsymbol{\alpha}_0$ by the method of moments, a direct plug-in of $\widehat{\boldsymbol{\alpha}}$ introduces bias for $\hat{\beta}_{\mathrm{OS}}$ asymptotically and we need to specify this bias term. This asymptotic bias is characterized by $\mathbf{W}_k$ and a covariance matrix between $\mathbf{m}_1$ and $\mathbf{s}_1$, i.e., $\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)$, whose entries are

$$[\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)]_{ij} = F_{ij} f_{ji}, \quad \text{and} \quad [\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)]_{ii} = \sum_{k \ne i} f_{ik} F_{ki}.$$

More details on how to obtain these results are provided in (79) in Appendix B.5. The asymptotic bias $b_0 := (b_{10}, \dots, b_{K0})^\top$ for the one-step estimator is defined as

$$b_{k0} = \lim_{n\to\infty} \frac{1}{\sqrt{N}} \mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k], \quad k = 1, \dots, K. \tag{20}$$

With Assumption 5 in position, we prove the limit distribution of $\hat{\beta}_{\mathrm{OS}}$ in the next theorem.

Theorem 4. *If Assumptions 1–5 hold, then*

$$\sqrt{N}(\hat{\beta}_{\mathrm{OS}} - \beta_0) - \mathrm{I}_0^{-1} b_0 \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1}). \tag{21}$$

Theorem 4 shows $\hat{\beta}_{\mathrm{OS}}$ achieves the Cramér-Rao lower bound asymptotically. In the proof of Theorem 4, we show that b_0 is $O(1)$ and depends on the covariance matrix between $\mathbf{m}_1$ and $\mathbf{s}_1$. It is because our plug-in estimator for $\boldsymbol{\alpha}$ is obtained from the moment estimating equation $\mathbf{m}_1$, and the one-step estimator (18) uses information from $\mathbf{s}_1$ to concentrate out $\boldsymbol{\alpha}$. The more similar between $\mathbf{m}_1$ and $\mathbf{s}_1$, the harder it is to estimate and control for the fixed effects simultaneously. This highlights the key difference between our one-step estimator with high-dimensional fixed effects and many other estimators in the literature with low-dimensional nuisance parameters.

While being efficient, $\hat{\beta}_{\mathrm{OS}}$ is not $\sqrt{N}$-consistent with an asymptotic bias of $\mathrm{I}_0^{-1} b_0$. In the next subsection, we use the bagging of split-network jackknife to achieve bias correction and maintain its efficiency simultaneously.

3.3 Bagging for Split-Network Jackknife Estimators

To debiase $\hat{\beta}_{\mathrm{OS}}$, one may use the split-network jackknife as discussed in Remark 1. Specifically, let

$$\hat{\beta}_{\mathrm{OS-SJ}} = 2\hat{\beta}_{\mathrm{OS}} - \frac{1}{2}\left(\hat{\beta}_{\mathrm{OS},1} + \hat{\beta}_{\mathrm{OS},2}\right), \tag{22}$$

where $\hat{\beta}_{\mathrm{OS},1}$ and $\hat{\beta}_{\mathrm{OS},2}$ are obtained from estimating β_0 using two randomly-split half networks, respectively. Then, it is straightforward to show that

$$\sqrt{N}(\hat{\beta}_{\mathrm{OS-SJ}} - \beta_0) \xrightarrow{d} \mathcal{N}(0, 2\mathrm{I}_0^{-1}). \tag{23}$$

The intuition for the doubled covariance matrix $2\mathrm{I}_0^{-1}$ in (23) is that by splitting the network just once we effectively ignore the links formed between agents from each subset. Furthermore, splitting the whole network randomly may make the SJ estimator computationally unstable. The inflated covariance matrix is not satisfactory given that the original one-step estimator achieves the Cramér-Rao lower bound of I_0^{-1}. Thus, we propose a bagging method to recover the efficiency of our estimator.

To motivate the bagging method, in theory there are a total of $T_n := \binom{n}{n/2}$ (suppose n is even for notational simplicity) possible ways to divide the network. However, T_n can be very large for a moderate choice of n. For example, when $n = 100$, $T_n = \binom{100}{50} \simeq 1.009 \times 10^{29}$, which is an astronomical number that beyond the capacity of most modern personal computers. The bagging method solves this problem by randomly selecting $T_n^{'} \ll T_n$ splits of the network and averaging over all the SJ estimators from these $T_n^{'}$ splits.

We show that when $T_n^{'} \to \infty$ and $n \to \infty$, the bagging estimator achieves the Cramér-Rao lower bound. To prove this claim, first consider the computationally infeasible SJ estimator based on all T_n splits of the network. Let $t = 1, \ldots, T_n$ index all different splits, where each t creates two equally separated random sets of agents $\mathcal{I}_{1,n}^{(t)}$ and $\mathcal{I}_{2,n}^{(t)}$ that satisfies $\mathcal{I}_{1,n}^{(t)} \cup \mathcal{I}_{2,n}^{(t)} = \mathcal{I}_n$ and $\mathcal{I}_{1,n}^{(t)} \cap \mathcal{I}_{2,n}^{(t)} = \emptyset$. Suppose the one-step estimators based on sub-networks $\mathcal{I}_{1,n}^{(t)} \times \mathcal{I}_{1,n}^{(t)}$ and $\mathcal{I}_{2,n}^{(t)} \times \mathcal{I}_{2,n}^{(t)}$ are $\hat{\beta}_{\mathrm{OS},1}^{(t)}$ and $\hat{\beta}_{\mathrm{OS},2}^{(t)}$, respectively. Then, the t-th one-step SJ estimator is

$$\hat{\beta}_{\mathrm{OS-SJ}}^{(t)} := 2\hat{\beta}_{\mathrm{OS}} - \frac{1}{2}\left(\hat{\beta}_{\mathrm{OS},1}^{(t)} + \hat{\beta}_{\mathrm{OS},2}^{(t)}\right). \tag{24}$$

Define the average of $\hat{\beta}_{\mathrm{OS-SJ}}^{(t)}$ over all $t \in \{1, ..., T_n\}$ as

$$\hat{\beta}_{T_n} := \frac{1}{T_n}\sum_{t=1}^{T_n} \hat{\beta}_{\mathrm{OS-SJ}}^{(t)}.$$

To implement the bagging method, we randomly select $T_n' \ll T_n$ estimators from $\left\{\hat{\beta}_{\mathrm{OS-SJ}}^{(t)}\right\}_{t=1}^{T_n}$ and take average of them to obtain the one-step split network jackknife bagging (BG) estimator,

$$\hat{\beta}_{\mathrm{BG}} := \frac{1}{T_n'}\sum_{t=1}^{T_n'} \hat{\beta}_{\mathrm{OS-SJ}}^{(t)}. \tag{25}$$

Denote the σ-algebra generated by $(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0)$ as $\mathcal{F}_n$. We have

$$\mathbb{E}\left(\hat{\beta}_{\mathrm{BG}} | \mathcal{F}_n\right) = \frac{1}{T_n}\sum_{t=1}^{T_n} \hat{\beta}_{\mathrm{OS-SJ}}^{(t)} = \hat{\beta}_{T_n}. \tag{26}$$

We provide a graphical illustration of the bagging method in Figure 1.

In the next theorem, we show that the oracle average $\hat{\beta}_{T_n}$ attains the Cramér-Rao lower bound, and the bagging estimator $\hat{\beta}_{\mathrm{BG}}$ approximates $\hat{\beta}_{T_n}$ sufficiently well as $T_n' \to \infty$. In Section 5 and 6, we find the finite-sample performance of our BG estimator satisfactory when $T_n^{'}$ is set to be around 100, which is significantly smaller than T_n.

Theorem 5. *If Assumptions 1–5 are satisfied, then, as $n \to \infty$,*

$$\sqrt{N}(\hat{\beta}_{T_n} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1}), \tag{27}$$

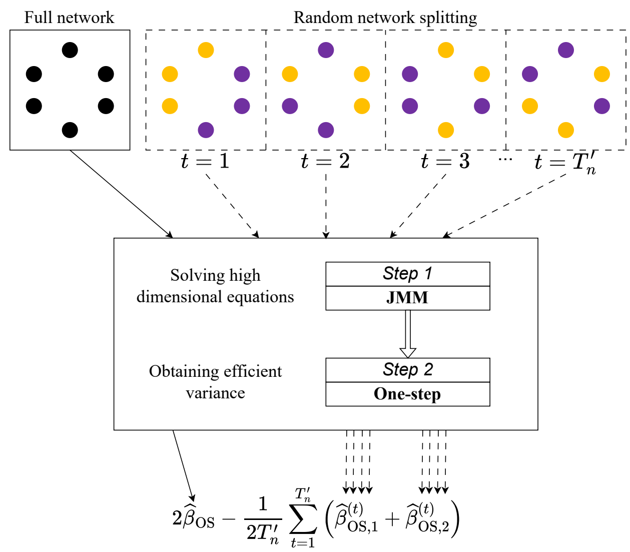


Figure 1: Summary of bagging procedure

$$\sqrt{N}(\hat{\beta}_{\mathrm{BG}} - \hat{\beta}_{T_n}) \xrightarrow{p} 0, \quad \text{as } T_n' \to \infty. \tag{28}$$

Hence, as $T_n' \to \infty$ *and* $n \to \infty$*,*

$$\sqrt{N}(\hat{\beta}_{\mathrm{BG}} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1}). \tag{29}$$

Remark 2. Although our BG estimator is inspired by half-panel jackknife from the literature on panel data (Dhaene and Jochmans (2015)) and can correct for the bias, it is different from SJ because directly applying SJ doubles the asymptotic variance of the estimator as shown in Corollary 1. The BG estimator, on the other hand, does not affect the variance as proved in (29) and is thus efficient. Alternatively, one may be inclined to apply BG to the JMM estimator directly and bypass Le Cam's approximation step. Indeed, BG can correct for the bias of the JMM estimator. However, it cannot make the JMM estimator efficient as Le Cam's approximation does. To sum up, applying both Le Cam's approximation (efficiency) and BG (bias-correction) to the JMM estimator leads to $\sqrt{N}$-consistent and efficient estimation of β_0.

Remark 3. In a nonlinear panel data model with time and individual fixed effects, Fernández-Val and Weidner (2016) propose splits along both individual and time dimensions for their split-sample jackknife estimator. For the individual dimension, Fernández-Val and Weidner (2016) suggest in their footnote 8 to use the average of all possible T_n partitions and point out that the average over $T_n' \ll T_n$ splits is enough. The objective of their repeated sampling is to avoid ambiguity and arbitrariness in the choice of the division. In contrast, we propose the BG estimator for achieving efficiency and bias-correction simultaneously. Furthermore, we prove formal asymptotic results for the BG estimator in Theorem 5 for inference.

4 Extensions

4.1 Average Partial Effects

In addition to $\boldsymbol{\alpha}_0$ and β_0, researchers and policy makers may be interested in estimating certain averages over the distribution of exogenous regressors and fixed effects. One leading example concerns the conditional mean of the outcome given covariates and individual fixed effects

$$\mathbb{E}\left[Y_{ij} \middle| X_{ij}, \boldsymbol{\alpha}, \beta\right] = F(x_{ij}^\top \beta + \alpha_i) F(x_{ij}^\top \beta + \alpha_j), \tag{30}$$

where the partial effects are differences or derivatives of (30) with respect to components of X_{ij}, say $X_{ij,k}$, the k^{th} coordinate of X_{ij}. We suppress its dependence on Y and X and define the partial effect of $x_{ij,k}$ for dyad (i,j) as

$$\Delta_{ij,k}(\alpha_i, \alpha_j, \beta) = \begin{cases} p_{ij}(\alpha_i, \alpha_j, \beta_k + x_{ij,-k}^\top \beta_{-k}) - p_{ij}(\alpha_i, \alpha_j, x_{ij,-k}^\top \beta_{-k}) & (b) \\ \beta_k \left[f(x_{ij}^\top \beta + \alpha_i) F(x_{ij}^\top \beta + \alpha_j) + F(x_{ij}^\top \beta + \alpha_i) f(x_{ij}^\top \beta + \alpha_i) \right] & (c) \end{cases}$$

where "(b)" corresponds to binary $x_{ij,k}$ while "(c)" refers to continuous $x_{ij,k}$. Define $\Delta_{ij} = (\Delta_{ij,1}, \ldots, \Delta_{ij,K})^\top$. Then, the unconditional APE is

$$\delta_0 = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{n} \sum_{j>i} \Delta_{ij}(\alpha_i, \alpha_j, \beta_0) \right], \tag{31}$$

where the expectation is taken over $(X_i, \alpha_i)_{i=1}^n$. Plugging the method of moments estimator of $(\boldsymbol{\alpha}, \beta)$ into (31) yields the estimator for the APE

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j>i} \Delta_{ij}(\hat{\alpha}_i, \hat{\alpha}_j, \hat{\beta}). \tag{32}$$

Define an (infeasible)

$$\bar{\Delta}_n = \frac{1}{N} \sum_{i=1}^{n} \sum_{j>i} \Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0). \tag{33}$$

Let the SJ estimator and bagging estimator of the APE be

$$\hat{\delta}_{\text{SJ}} := 2\hat{\delta} - \frac{1}{2}(\hat{\delta}_1 + \hat{\delta}_2) \quad \text{and} \quad \hat{\delta}_{\text{BG}} := \frac{1}{T_n'} \sum_{t=1}^{T_n'} \hat{\delta}_{\text{SJ}}^{(t)}, \tag{34}$$

respectively. Here, $\left(\hat{\delta}_1, \hat{\delta}_2\right)$ are the plug-in estimators based on two sub-networks after a random split of the nodes and $\left\{\hat{\delta}_{\text{SJ}}^{(t)}\right\}_{t=1}^{T_n'}$ are split-network jackknife estimators based on T_n' random splits. The next theorem shows that $\hat{\delta}$ is $\sqrt{n}$-consistent. We use a central limit theorem for U-statistics (Theorem 12.3 of Van der Vaart (2000)) to prove it. To make the result precise, we include asymptotically vanishing bias terms similarly to Theorem 4.2 of Fernández-Val and Weidner (2016). Additionally, since the bias terms are asymptotically negligible, the two bias-corrected estimators in (34) do not provide meaningful improvement over the original estimator $\hat{\delta}$. We provide numerical evidence in Section 5 and 6 that substantiates this claim.

Theorem 6. *Define $\sigma_{\delta,n} := \frac{\Sigma_\Delta}{N} + \frac{4\Sigma_\delta}{n}$. If Assumptions 1–4 hold, $(X_i, \alpha_i)_{1\le i\le n}$ are i.i.d. across i, and $\bar{\Delta}_n$ is a non-degenerate U-statistic, then*

$$\sigma_{\delta,n}^{-1/2}\left(\hat{\delta} - \delta_0 - \frac{1}{\sqrt{N}}B_\beta - \frac{1}{\sqrt{N}}B_\alpha\right) \xrightarrow{d} \mathcal{N}(0, I_K), \tag{35}$$

with

$$B_\alpha = \lim_{n\to\infty}\frac{1}{2\sqrt{N}}\text{Tr}\left[\mathbf{J}_{11}^{-1}\mathbf{V}_{11}\left(\mathbf{J}_{11}^{-1}\right)^\top \mathbf{R}_k^\mu\right], \quad B_\beta := \lim_{n\to\infty}(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\text{J}_0^{-1}B_0, \tag{36}$$

where $\mathbf{R}_k^\mu$, $k = 1,\ldots,K$ and $(\Delta_{\boldsymbol{\alpha}}, \Delta_\beta)$ are characterized in (90) and (87), respectively. The variance term Σ_Δ is defined in (94). The sampling variance $\Sigma_\delta = \mathbb{E}\left[\Delta_{ij}(\alpha_i,\alpha_j,\beta_0)\Delta_{ik}(\alpha_i,\alpha_k,\beta_0)\right]$, where the expectation is taken over $(\alpha_i, X_i)_{i=1}^n$.

Furthermore, the bagging estimator $\hat{\delta}_{\text{BG}}$ satisfies

$$\sigma_{\delta,n}^{-1/2}\left(\hat{\delta}_{\text{BG}} - \delta_0\right) \xrightarrow{d} \mathcal{N}(0, I_K).$$

In Theorem 6, B_β stems from the bias of the plug-in estimator $\hat{\beta}$ while B_α stems from the incidental parameter bias of plug-in estimator $\widehat{\boldsymbol{\alpha}}$. For the components of $\sigma_{\delta,n}$, Σ_Δ is the asymptotic variance of $\sqrt{N}(\hat{\delta} - \bar{\Delta}_n)$ and Σ_δ is the asymptotic variance of $\sqrt{n}(\bar{\Delta}_n - \delta_0)$. Note that in Theorem 6 the rate of convergence of $\hat{\delta}$ is $\sqrt{n}$ instead of $\sqrt{N}$. The slower convergence rate in (35) makes the bias terms introduced by estimating the individual fixed effects asymptotically negligible.

It is worth pointing out that if one is interested in $\bar{\Delta}_n$ (e.g., Chen, Fernández-Val, and Weidner (2021)), the asymptotic result will become

$$\sqrt{N}\left(\hat{\delta} - \bar{\Delta}_n\right) - B_\beta - B_\alpha \xrightarrow{d} \mathcal{N}(0, \Sigma_\Delta),$$

which generalizes Theorem 2 of Chen, Fernández-Val, and Weidner (2021) to the NTU setting.

The asymptotic variance Σ_δ can be estimated by

$$\hat{\Sigma}_\delta = \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j>i} \sum_{k>j} \left[\Delta_{ij}(\hat{\alpha}_i, \hat{\alpha}_j, \hat{\beta}) - \hat{\delta}\right] \left[\Delta_{ik}(\hat{\alpha}_i, \hat{\alpha}_k, \hat{\beta}) - \hat{\delta}\right],$$

which is consistent by the law of large numbers for U-statistics. Although the variance term Σ_Δ/N is dominated asymptotically by $4\Sigma_\delta/n$ in (35), in simulations we find that including it improves the coverage probabilities.

4.2 Model Misspecification

The Cramér-Rao lower bound is only defined for correctly specified models. A natural question is what if the distribution function $F(\cdot)$ of ϵ_{ij} is misspecified? Graham (2024) provides an insightful analysis for sparse network formation models with TU. However, this question has not been studied yet in the literature of network formation with NTU. In this section, we discuss theoretical properties of our estimators under such misspecification. If the model used for estimation is misspecified, the estimator of structural parameter in nonlinear models in general converges to a pseudo value defined as the minimizer of certain criterion function (for MLE, see White (1982, 1996)) or the solution of a moment equation (for method of moments).

First, we analyze the pseudo values that our estimators $\hat{\beta}$ and $\hat{\beta}_{\text{OS}}$ converge to under model misspecification. Suppose researchers specify the distribution function of ϵ_{ij} as $G(\cdot)$ while the true distribution $F(\cdot)$ differs from $G(\cdot)$ at points with strictly positive probability measure. For fixed n, we impose the following identification assumption for model misspecification. Let $q_{ij}(\boldsymbol{\alpha}, \beta) := G(\alpha_i + x_{ij}^\top \beta) G(\alpha_j + x_{ij}^\top \beta)$ be the misspecified probability of linking between i and j.

Assumption 6 (Identification under Model Misspecification)**.** *For a fixed n, the following nonlinear function of β:*

$$\tilde{S}_n(\beta) := \sum_{i=1}^{n} \sum_{j>i} \left[p_{ij}(\boldsymbol{\alpha}_0, \beta_0) - q_{ij}(\boldsymbol{\alpha}(\beta), \beta)\right] x_{ij} = 0 \tag{37}$$

has a unique solution β_{n*}, and satisfy for any $\delta > 0$ and n large enough,

$$\inf_{\beta \in \mathbb{B}: \|\beta - \beta_{n*}\|_2 \geq \delta} \left\| \tilde{S}_n(\beta) \right\|_2 > 0,$$

where $\boldsymbol{\alpha}(\beta)$ is the unique solution to the following system of equations

$$\left(\sum_{j\neq 1} p_{1j}(\boldsymbol{\alpha}_0, \beta_0) - \sum_{j\neq 1} q_{1j}(\boldsymbol{\alpha}, \beta), \cdots, \sum_{j\neq n} p_{nj}(\boldsymbol{\alpha}_0, \beta_0) - \sum_{j\neq n} q_{nj}(\boldsymbol{\alpha}, \beta)\right)^\top = 0. \tag{38}$$

Assumption 6 is the counterpart of Assumption 4 under model misspecification. Similarly to Theorem 1, equation (38) has a unique solution with high probability under mild conditions on $(\boldsymbol{\alpha}_0, \beta_0)$ and β. Thus, Assumption 6 essentially imposes a local identification condition for the common parameter. Notice that β_{n*} depends on the true link function $F(\cdot)$, misspecified link function $G(\cdot)$, true parameter values $(\boldsymbol{\alpha}_0, \beta_0)$, and the covariates $\{X_{ij}\}_{i \neq j}$. As a result, β_{n*} may vary with n. The following theorem demonstrates that the JMM estimator based on the misspecified link function $G(\cdot)$ will center around β_{n*} with a bias term, and the split-network jackknife procedure removes the bias asymptotically. Let $\boldsymbol{\alpha}_* := \boldsymbol{\alpha}(\beta_{n*})$ with $\boldsymbol{\alpha}(\cdot)$ satisfying (38).

Theorem 7 (JMM Estimation under Model Misspecification)**.** *If Assumptions 1–3 and 6 hold, then as $n \to \infty$*

$$\sqrt{N}\left(\hat{\beta} - \beta_{n*}\right) - \mathrm{J}_*^{-1} B_* \xrightarrow{d} \mathcal{N}(0, \Omega_*).$$

where J_*, B_* and Ω_* are defined similarly to J_0, B_0 and Ω_0 in Section 3.1 except that the pseudo value $(\boldsymbol{\alpha}_*, \beta_{n*})$ and the misspecified link function $G(\cdot)$ are used in the place of $(\boldsymbol{\alpha}_0, \beta_0)$ and $F(\cdot)$.

Theorem 7 shows that if the researcher believes that the moment equations hold in population, the JMM estimator is robust under the model misspecification in the sense that $\hat{\beta}$ is consistent for β_{n*}, which is the unique solution to the (pseudo) population moment equations (37).

To conduct statistical inference for the JMM estimator under model misspecification, we need an estimator for the limit covariance matrix Ω_*. Let $\mathbf{m}_{ij}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ be an $(n+K) \times 1$ vector where: (i) the ith and jth elements are both $y_{ij} - q_{ij}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$; (ii) the $(n+1)$th to $(n+K)$th elements are the vector of $[y_{ij} - q_{ij}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})] x_{ij}^{\top}$; and (iii) the rest of the coordinates are zero. Then, we use the plug-in estimator

$$\hat{\mathbf{V}}_* := \sum_{i=1}^{n} \sum_{j>i} \mathbf{m}_{ij}(\widehat{\boldsymbol{\alpha}}, \hat{\beta}) \mathbf{m}_{ij}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})^{\top}.$$

Further write submatrices of $\hat{\mathbf{V}}_*$ as $\hat{\mathbf{V}}_{11*}$, $\hat{\mathbf{V}}_{12*}$, $\hat{\mathbf{V}}_{21*}$, and $\hat{\mathrm{V}}_{22*}$, and similarly for $\hat{\mathbf{J}}_*$. Note that J_* is the concentrated Jacobian matrix for β_{n*}, while $\mathbf{J}_*$ is the Jacobian matrix for both $\boldsymbol{\alpha}_*$ and β_{n*}. Then, we propose to estimate Ω_* by

$$\hat{\Omega}_* := N^{-1} \hat{\mathrm{J}}_*^{-1} \left[\hat{\mathrm{V}}_{22*} + \hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{11*} (\hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{J}}_{21*})^{\top} - \hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{12*} - (\hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{12*})^{\top} \right] (\hat{\mathrm{J}}_*^{-1})^{\top}, \tag{39}$$

which is consistent for Ω_* by the law of large numbers.

Under the possible model misspecification, our one-step estimator can be expressed as

$$\hat{\beta}_{\mathrm{OS}} := \hat{\beta} + \mathrm{H}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})^{-1} s_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta}), \tag{40}$$

with the pilot moment estimator $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ as before. Note that

$$\mathrm{H}(\boldsymbol{\alpha}, \beta) := \mathrm{H}_{22}(\boldsymbol{\alpha}, \beta) - \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{H}_{22}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta), \tag{41}$$

is the concentrated Hessian matrix and

$$s_n(\boldsymbol{\alpha}, \beta) := s_2(\boldsymbol{\alpha}, \beta) - \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{H}_{11}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{s}_1(\boldsymbol{\alpha}, \beta),$$

is the concentrated score function. Under model misspecification, $\hat{\beta}_{\mathrm{OS}}$ in (40) centers around

$$\beta_{n\star} := \beta_{n*} + \mathrm{H}(\boldsymbol{\alpha}_*, \beta_{n*})^{-1} \mathbb{E} s_n(\boldsymbol{\alpha}_*, \beta_{n*}), \tag{42}$$

which can be seen as a projection of β_{n*} by concentrating out the fixed effects. When the model is correctly specified, $(\boldsymbol{\alpha}_*, \beta_{n*}) \equiv (\boldsymbol{\alpha}_0, \beta_0)$, thus $\beta_{n\star} \equiv \beta_{n*} \equiv \beta_0$ because $\mathbb{E} s_n(\boldsymbol{\alpha}_*, \beta_{n*}) = \mathbb{E} s_n(\boldsymbol{\alpha}_0, \beta_0) \equiv 0$. Furthermore, our one-step estimator and bias-corrected estimators in the misspecified case share similar asymptotic properties from their counterparts when the model is correctly specified, except that they now center around the projected pseudo value $\beta_{n\star}$ instead of β_0. These results are summarized in the next theorem.

Theorem 8. *Suppose all the bounds in Assumption 5 still hold for each element of* $[\mathbf{H}_{12}^\top \mathbf{H}_{11}^{-1}]$*. If Assumptions 1–3 and 6 are satisfied, the one-step estimator* $\hat{\beta}_{\mathrm{OS}}$ *and the bagging estimator* $\hat{\beta}_{\mathrm{BG}}$ *satisfy*

$$\sqrt{N}\left(\hat{\beta}_{\mathrm{OS}} - \beta_{n\star}\right) - \mathrm{H}_*^{-1} b_* \xrightarrow{d} \mathcal{N}(0, \Gamma_*) \text{ and}$$
$$\sqrt{N}\left(\hat{\beta}_{\mathrm{BG}} - \beta_{n\star}\right) \xrightarrow{d} \mathcal{N}(0, \Gamma_*),$$

respectively, where H_* *and* b_* *are defined similarly to* H_0 *and* b_0*, but under* $(\boldsymbol{\alpha}_*, \beta_*)$ *and misspecified link* $G(\cdot)$*. The asymptotic covariance matrix* Γ_* *is*

$$\Gamma_* := \lim_{n\to\infty} N^{-1} \mathrm{H}_*^{-1} \left[\begin{array}{c} \mathrm{I}_{22*} + \mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{11*} (\mathbf{H}_{11*}^{-1} \mathbf{H}_{12*}^\top)^\top \\ -\mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{12*} - (\mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{12*})^\top \end{array} \right] (\mathrm{H}_*^{-1})^\top. \tag{43}$$

We point out that the limits of the variance term and Hessian term are functions of $(\boldsymbol{\alpha}_*, \beta_{n*})$ because $\beta_{n\star}$ is a function of $(\boldsymbol{\alpha}_*, \beta_{n*})$ by (42). Theorem (8) demonstrates that $\hat{\beta}_{\mathrm{BG}}$ serves as a robust estimator for common parameters under model misspecification in the following sense. If the model is correctly specified, $\hat{\beta}_{\mathrm{BG}}$ centers around β_0 without bias and achieves the Cramér-Rao lower bound asymptotically. If the model is misspecified, $\hat{\beta}_{\mathrm{BG}}$ centers around a projected pseudo value with no asymptotic bias and achieves a lower variance than $\hat{\beta}_{\mathrm{SJ}}$. Finally, we can estimate $\mathbf{I}_*$ by $\hat{\mathbf{I}} = \sum_{i=1}^n \sum_{j>i} \mathbf{s}(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ and $\mathbf{H}_*$ by plugging

$(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ into (41), which together give a consistent estimator for Γ_* by (43).

5 Monte Carlo Simulation

We investigate the finite sample performance of our estimators through a comprehensive set of Monte Carlo experiments. Specifically, we (i) examine the performance of the proposed estimators for β_0, which include the joint method of moments estimator (JMM), Le Cam's one-step estimator (OS), and one-step estimator with bagging of split-network jackknife (BG); (ii) show how these estimators perform when the model is misspecified; (iii) investigate the performance of these estimators when the network is reasonably sparse; (iv) check how well the method can estimate individual fixed effect $\boldsymbol{\alpha}_0$; and (v) study the performance of the estimators for the APEs.

The data generating process (DGP) is as follows. We set $\beta_0 = (1, -1)^\top$, and draw the first covariate of X_{ij} as $X_{1,ij} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.3), $X_{1,ij} = X_{1,ji}$. This way, we allow for discrete variable in X_{ij}. For the second covariate of X_{ij}, we draw $X_i \overset{\text{i.i.d.}}{\sim} U(-0.5, 0.5)$ and let $X_{2,ij} = |X_i - X_j|$. Next, we generate the individual fixed effects by setting $\alpha_i = 0.75 \times X_i + 0.25 \times \xi_i$, where $\xi_i \overset{\text{i.i.d.}}{\sim} U(-0.5, 0.5)$ and is independent of everything else. Thus, α_i and X_{ij} are correlated via X_i. The idiosyncratic shock to each dyad, ϵ_{ij}, is randomly drawn from standard logistic distribution, independently of X_{ij}, α_i, and ξ_i. Finally, we obtain each ij pair of the network $\mathbf{Y}$ by

$$Y_{ij} = 1(\alpha_i + X_{ij}^\top \beta_0 + \epsilon_{ij} > 0) \cdot 1(\alpha_j + X_{ji}^\top \beta_0 + \epsilon_{ji} > 0).$$

For all the simulations in this paper, we run $R = 1,000$ replications. For the baseline results, we set $n = 100$ and 200, which is comparable to the size of the dataset we use for the empirical application. We also conduct simulations for $n = 50$ to investigate the small sample performance of the estimator and present these results in Appendix C. To further investigate the performance of the estimator of the high-dimensional fixed effects, we also let $n = 500$ and $1,000$ in Section 5.2. We report mean bias, median bias, standard deviation, mean absolute bias, median absolute bias, and root mean squared error (RMSE) across R runs of simulations.

5.1 Common Parameter

Table 1 summarizes the results for the proposed six estimators of the common parameter β_0 for $n = 100$ and 200 when the link function $F(\cdot)$ is correctly specified. The main conclusions when $n = 100$ are as follows. First, in terms of mean bias, the bias-corrected BG estimator performs significantly better than JMM and OS. The median bias shows a similar pattern. These results are consistent with theory prediction because BG uses bagging of split-network jackknife to achieve bias-correction. Second, BG works very well in simultaneously achieving bias-correction and low standard deviation, leading to the lowest RMSE among all estimators. Moreover, in line with the theory, we find that SJ without BG (not reported in the tables) inflates the variances of JMM and OS estimators by a constant factor of two in the simulations. Third, the coverage probabilities of the confidence intervals constructed based on the asymptotic distribution of each estimator are close to the nominal rate. This result shows that the asymptotic distributions for the proposed estimators of this paper are good approximation to the true underlying distribution. Fourth, the performance of $\widehat{\beta}_2$ is worse than that of $\widehat{\beta}_1$ across all estimators in terms of RMSE, likely caused by the correlation between $X_{2,ij}$ and (α_i, α_j). Finally, the mean standard errors estimated from the asymptotic normality results are close to the standard deviations computed from Monte Carlo simulations across all estimators. We also find that the quantiles of empirical distributions of all estimators across simulations are well approximated by the quantiles of corresponding normal distributions. These results provide support for our theory.

When $n = 200$, the performance of all the estimators improve. RMSE's, for example, are about half the size of those when $n = 100$, which is expected given the $\sqrt{N}$-convergence rate and $\sqrt{N} = O(n)$. The coverage probabilities are also closer to the nominal level when $n = 200$ than when $n = 100$. The main conclusions remain unchanged as when $n = 100$.

5.2 Individual Fixed Effects

Given the large number of individual fixed effects, we plot histogram of the estimation error, $\widehat{\alpha}_i - \alpha_{i0},\ i = 1, \ldots, n$. We find that, consistent with our theory, centers of histograms are around 0 for small $n = 100$ and 200. As we increase n, the performance of $\widehat{\alpha}_i$ improves, as can be seen from Figure 2(c) and (d) for $n = 500$ and $1,000$, respectively. Moreover, the range of estimation error shrinks toward 0 as sample size increases, hence having a large n further reduces the occurrence of extreme values of $\widehat{\alpha}_i$ and improves uniformity of the estimates.

5.3 Average Partial Effects, Model Misspecification, and Sparser Network

Table 2 summarizes the estimation results for the APEs defined in (31). We provide results for APE for each coordinate of X_{ij}. Across the board we find our plug-in estimator performs quite well in terms of RMSE and coverage probabilities. We also apply the bagging

Table 1: Estimation Results of β_0

$n=100,\ density$=25%	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0304	-0.0288	0.0291	-0.0282	-0.0026	0.0028
Median Bias	0.0302	-0.0345	0.0287	-0.0361	-0.0022	-0.0034
Standard Deviation	0.0594	0.1349	0.0591	0.1352	0.0573	0.1318
Mean Standard Error	0.0568	0.1296	0.0568	0.1293	0.0568	0.1293
Mean Absolute Bias	0.0538	0.1115	0.0532	0.1116	0.0459	0.1058
Median Absolute Bias	0.0453	0.0931	0.0437	0.0938	0.0401	0.0925
RMSE	0.0667	0.1380	0.0659	0.1381	0.0574	0.1318
90% Coverage Rate	84.5	87.8	85.2	87.7	90.1	88.8
95% Coverage Rate	91.0	93.7	91.3	93.6	94.8	94.5
$n=200,\ density$=25%	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0142	-0.0171	0.0137	-0.0166	-0.0017	-0.0014
Median Bias	0.0147	-0.0203	0.0137	-0.0181	-0.0019	-0.0029
Standard Deviation	0.0289	0.0652	0.0288	0.0650	0.0284	0.0640
Mean Standard Error	0.0278	0.0636	0.0278	0.0634	0.0278	0.0634
Mean Absolute Bias	0.0259	0.0540	0.0256	0.0536	0.0229	0.0508
Median Absolute Bias	0.0226	0.0460	0.0218	0.0450	0.0193	0.0421
RMSE	0.0322	0.0674	0.0319	0.0670	0.0285	0.0640
90% Coverage Rate	84.4	88.0	84.6	88.2	89.7	90.0
95% Coverage Rate	90.6	94.1	90.8	94.1	95.2	95.7

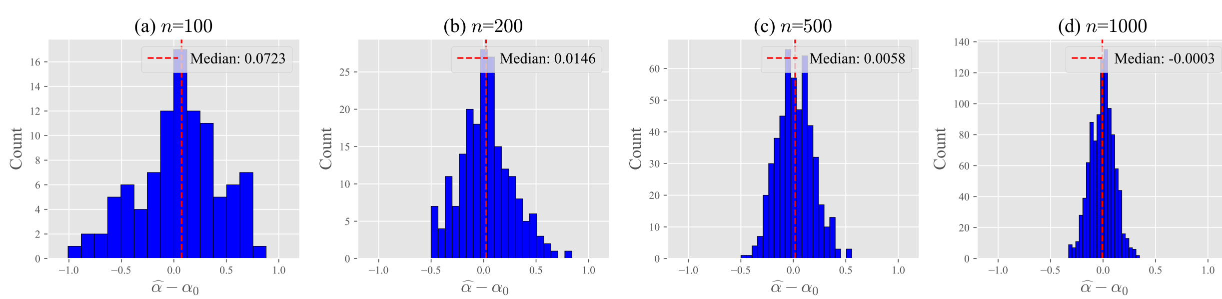


Figure 2: Histograms of $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0$

Table 2: Estimation Results of APE

	$n = 100$, *Density=25%*				$n = 200$, *Density=25%*			
	Plug-in		Bagging		Plug-in		Bagging	
	$X_{ij,1}$	$X_{ij,2}$	$X_{ij,1}$	$X_{ij,2}$	$X_{ij,1}$	$X_{ij,2}$	$X_{ij,1}$	$X_{ij,2}$
Mean Bias	-0.0028	0.0014	-0.0006	-0.0016	-0.0014	0.0007	-0.0002	-0.0006
Median Bias	-0.0032	0.0007	-0.0006	-0.0015	-0.0014	0.0014	-0.0002	-0.0001
Standard Deviation	0.0141	0.0273	0.0144	0.0279	0.0075	0.0142	0.0077	0.0145
Mean Standard Error	0.0148	0.0285	0.0148	0.0285	0.0075	0.0142	0.0075	0.0142
Mean Absolute Bias	0.0114	0.0221	0.0114	0.0226	0.0061	0.0112	0.0062	0.0114
Median Absolute Bias	0.0097	0.0187	0.0091	0.0192	0.0052	0.0093	0.0052	0.0097
RMSE	0.0144	0.0274	0.0144	0.0280	0.0077	0.0142	0.0077	0.0145
90% Coverage Rate	91.0	92.1	91.3	91.3	90.0	90.6	88.7	89.1
95% Coverage Rate	95.3	96.5	95.9	96.2	94.8	94.8	94.9	94.5

Note: true values of APEs are calibrated by a simulation with $n = 10,000$ agents for all simulations.

method of split-network jackknife to the estimators of APE and find that it does not achieve meaningful improvement. This is expected because by Theorem 6, the asymptotic bias in estimating the APEs is asymptotically negligible and of an order smaller than that in estimating β.

Table 3 presents the results for estimating the homophily coefficients under misspecification of the distribution of ϵ_{ij}. We draw ϵ_{ij} from the standard normal distribution, but "mistakenly" specify the distribution of ϵ_{ij} as logistic in the estimation. We compare $\widehat{\beta}$ to the pseudo true value $\beta_{n\star}$ defined in (42) and find that the results are satisfactory. The performance of our BG estimator dominates other estimators in terms of bias, variance, and coverage probabilities, highlighting the efficacy and importance of employing proper bias-correction procedures.

Finally, we investigate how the method works for sparser networks. Specifically, we lower the density of the network to be less than 9% by setting $\alpha_i = 0.75 \times X_i + 0.25\xi_i - 1$, so that the network formation process is driven more by the homophily effect. Table 4 summarizes the results. It is clear that the performance of our method is negatively affected by the sparsity of the network. For example, the RMSE's of the BG estimators of β_1 and β_2 are 0.0365 and 0.0851, respectively, when $n = 200$ and the density of network is 8.63%. In comparison, when the density of network is 25.4% in Table 1 for $n = 200$, the RMSE's are 0.0285 and 0.0640, respectively. With that said, the way we introduce sparsity actually also increases the correlation between α and X, which in theory also makes it more challenging to estimate β_0. So, we consider the magnitude of the change in performance to be an upper bound of the effect from a sparser network. We also find that our BG estimator performs better than other candidates in almost all metrics.

Table 3: Estimation Results under Model Misspecification

$n = 100$, *density=27%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0611	-0.0570	0.0647	-0.0550	-0.0025	0.0087
Median Bias	0.0608	-0.0568	0.0648	-0.0556	-0.0025	0.0083
Standard Deviation	0.0634	0.1510	0.0636	0.1500	0.0609	0.1441
Mean Standard Error	0.0621	0.1457	0.0636	0.1482	0.0636	0.1482
Mean Absolute Bias	0.0722	0.1279	0.0746	0.1266	0.0484	0.1149
Median Absolute Bias	0.0643	0.1102	0.0662	0.1060	0.0411	0.0980
RMSE	0.0881	0.1614	0.0908	0.1597	0.0609	0.1443
90% Coverage Rate	74.0	87.3	73.1	88.3	90.8	91.7
95% Coverage Rate	84.1	93.0	84.8	93.5	96.9	95.5
$n = 200$, *density=27%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0281	-0.0294	0.0296	-0.0273	-0.0017	0.0014
Median Bias	0.0290	-0.0311	0.0310	-0.0286	-0.0001	0.0006
Standard Deviation	0.0305	0.0766	0.0304	0.0761	0.0298	0.0748
Mean Standard Error	0.0304	0.0748	0.0306	0.0751	0.0306	0.0751
Mean Absolute Bias	0.0345	0.0662	0.0354	0.0650	0.0236	0.0594
Median Absolute Bias	0.0311	0.0589	0.0322	0.0563	0.0197	0.0498
RMSE	0.0415	0.0820	0.0424	0.0809	0.0298	0.0748
90% Coverage Rate	75.7	86.1	75.2	87.5	90.2	90.6
95% Coverage Rate	84.7	92.9	83.9	92.9	95.6	95.7

Table 4: Estimation Results of β_0 under Sparser Network

$n = 100$, *density=8.6%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0447	-0.0825	0.0517	-0.0515	-0.0023	0.0018
Median Bias	0.0446	-0.0802	0.0506	-0.0494	-0.0031	0.0050
Standard Deviation	0.0747	0.1780	0.0749	0.1795	0.0708	0.1700
Mean Standard Error	0.0740	0.1791	0.0742	0.1793	0.0742	0.1793
Mean Absolute Bias	0.0701	0.1582	0.0735	0.1503	0.0564	0.1370
Median Absolute Bias	0.0602	0.1346	0.0622	0.1305	0.0472	0.1179
RMSE	0.0870	0.1962	0.0910	0.1867	0.0709	0.1700
90% Coverage Rate	84.5	85.3	82.3	88.1	91.4	91.2
95% Coverage Rate	91.1	93.4	90.0	94.7	96.4	96.8
$n = 200$, *density=8.6%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0189	-0.0360	0.0215	-0.0221	-0.0035	0.0029
Median Bias	0.0197	-0.0364	0.0221	-0.0242	-0.0025	0.0007
Standard Deviation	0.0373	0.0869	0.0371	0.0871	0.0363	0.0850
Mean Standard Error	0.0359	0.0873	0.0360	0.0872	0.0360	0.0872
Mean Absolute Bias	0.0335	0.0757	0.0346	0.0722	0.0290	0.0679
Median Absolute Bias	0.0288	0.0651	0.0301	0.0628	0.0248	0.0582
RMSE	0.0418	0.0940	0.0429	0.0899	0.0365	0.0851
90% Coverage Rate	83.4	87.0	82.6	89.3	89.4	90.9
95% Coverage Rate	90.4	93.1	89.9	94.1	94.5	96.0

6 Empirical Illustration

In this section, we provide two empirical applications. First, we apply our method to the risk-sharing network data of Nyakatoke and obtain economically meaningful results. Our empirical findings complement the results of Gao, Li, and Xu (2023) by showing that wealth differences do not have a statistically significant impact on the formation of links. Second, we use the India microfinance network dataset (Banerjee et al. (2013, 2024)) to show how our method works in capturing important network features such as average degree, clustering effect, and percentage of isolates. By adding only one more parameter for modeling triangles, we find that the extended model captures most network features well.

6.1 Nyakatoke Risk-Sharing Network

As an empirical illustration, we apply our estimation methods to the risk-sharing network data of Nyakatoke, a small Haya community of 119 households in 2000 located in the Kagera Region of Tanzania. We investigate how important factors, such as wealth, distance, and blood or religious ties, are relative to each other in deciding the formation of risk-sharing links among local residents. A similar exercise has been conducted in Gao, Li, and Xu (2023), however they only provide point estimates of the homophily coefficients β_0. In this paper, we estimate β_0, individual fixed effects α, and the APEs. We also provide confidence intervals for our estimates based on the asymptotic results provided in Section 3.

6.1.1 Data

The Nyakatoke risk-sharing network data, collected by Joachim De Weerdt in 2000, cover all of the 119 households in the community. It includes the following: (i) whether or not two households are linked in the insurance network, (ii) total USD assets and religion of each household, (iii) kinship and distance between households. See De Weerdt (2004), De Weerdt and Dercon (2006), and De Weerdt and Fafchamps (2011) for more details. To define the dependent variable *link*, the interviewer asks each household the following question:

"*Can you give a list of people from inside or outside of Nyakatoke, who you can personally rely on for help and/or that can rely on you for help in cash, kind or labor?*"

The data contains three answers of "bilaterally mentioned", "unilaterally mentioned", and "not mentioned" between each pair of households. Considering the question is about whether one can rely on the other for help, we interpret both "bilaterally mentioned" and "unilaterally mentioned" as they are connected in this undirected network, meaning that the dependent variable Y_{ij} *link* equals 1.[11]

We estimate the coefficients for 3 regressors: *wealth difference*, *distance* and *tie* between households. *Wealth* is defined as the total assets in USD owned by each household in 2000, including livestocks, durables and land. *Distance* measures how far away two households are located in kilometers. *Tie* is a discrete variable, defined to be 3 if members of one household are parents, children and/or siblings of members of the other household, 2 if nephews, nieces, aunts, cousins, grandparents and grandchildren, 1 if any other blood relation applies or if two households share the same religion, and 0 if no blood religious tie exists[12]. Following the literature we take natural log on *wealth* and *distance*, and we construct the *wealth difference* variable as the absolute difference in *wealth*, i.e.,

$$w\left(X_i, X_j\right) = \left(\left|\ln \text{wealth}_i - \ln \text{wealth}_j\right|,\ \text{ln-distance}_{ij},\ \text{tie}_{ij}\right)^\top.$$

Figure 3 shows the structure of the insurance network in Nyakatoke. In the left sub-figure, each node in the graph represents a household. The solid line between two nodes indicates they are connected, i.e., *link* equals 1. The numbers inside each circle represent the number of links the household has, which is also positively correlated with the size of the circle. The right sub-figure summarizes the degree distribution for the network. Most of the households in this network have links between 5 and 13. The maximum degree is 32, while the minimum is 1.

In the dataset there are 5 households that lack information on *wealth* and/or *distance*. We drop these observations, resulting in a sample size N of 114. The total number of ordered household pairs is 12,882. Table 5 provides summary statistics for the dataset we use.

6.1.2 Results and Discussion

Table 6 presents the estimation results for the three homophily coefficients using method of moments (JMM), one-step estimation (OS), and one-step estimation with split-network jackknife and bagging technique (BG). Since JMM and OS estimators are $\sqrt{N}$-inconsistent with asymptotic bias, we do not report their p-values. The estimated coefficient for wealth difference is negative using all three methods. However, it is statistically insignificant based

[11]In the context of the village economies in our application, we think, at the time of link formation, the risk-sharing links are less likely (in comparison with the contexts of business or financial networks) to be driven by efficient arrangements of side-payment transfers, thus satisfying NTU.

[12]Notice that *distance* and *tie* are dyadic characteristics that may not be constructed by individual level covariates, however, our theory continues to work if we treat them as fixed.

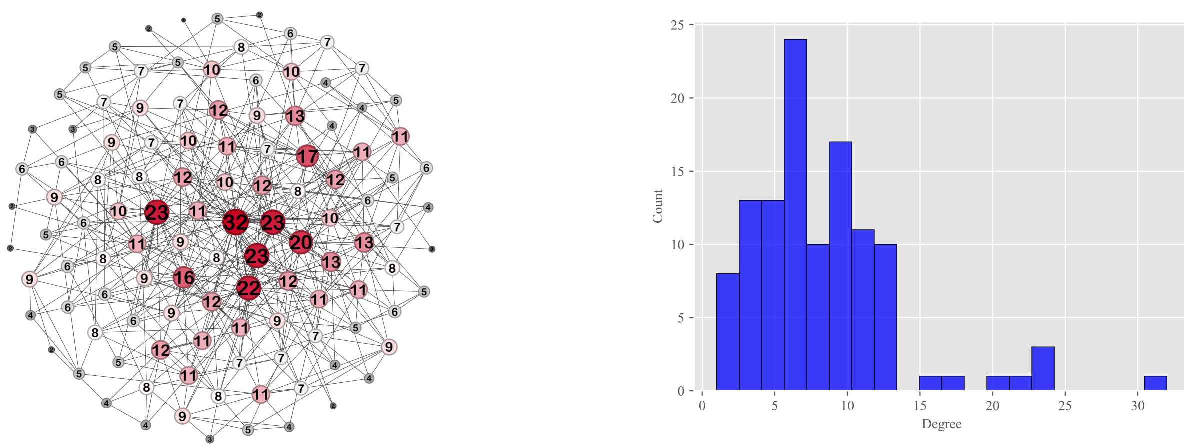


Figure 3: Nyakatoke Network Structure and Degree Heterogeneity

Table 5: Empirical Application: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
link	12,882	0.0732	0.2606	0	1
|(ln) wealth difference|	12,882	1.0365	0.8228	0.0004	5.8898
(ln) distance	12,882	6.0553	0.7092	2.6672	7.4603
tie	12,882	0.4260	0.6123	0.0000	3.0000

Table 6: Estimation Results for Nyakatoke Network

	JMM	OS	BG
|(ln) wealth difference|	-0.0882	-0.0974	-0.0777
	–	–	(0.2257)
(ln) distance	-0.7824	-0.8636	-0.8187
	–	–	(0.0000)
Tie	0.6714	0.6287	0.5817
	–	–	(0.0000)

Note: p-values of two-side t-tests for BG estimator are reported in the parentheses.

on BG, suggesting that wealth difference does not matter statistically to form a link. To interpret this result, consider two scenarios. One is when two households possess similar amount of wealth. Then, everything else being equal, they may not be willing to form a link because the other household may not have the capacity to insure themselves against unpredictable shocks such as natural disaster or severe diseases. The other scenario is when there is a huge wealth difference between the two households. Then, by link formation rule (4) under NTU, the linking decision is likely to be driven by the household with a larger amount of wealth. It is again unlikely to form a risk-sharing link because the surplus of the richer household is likely to be negative from the link. Therefore, the net effect of absolute wealth difference on forming a link is expected to be close to zero. It is clear that bilateral agreement required to form a link in a model with NTU plays a central role in both scenarios. Our estimates of the homophily coefficient for the wealth difference are consistent with these explanations. Note that the estimated coefficient for wealth difference in Gao, Li, and Xu (2023) is also negative. However, they do not have inference results for the common parameters, hence cannot evaluate statistical significance of these estimates.

In addition to the wealth difference, we find the estimated coefficient for distance is significantly negative at -0.8187, while for tie it is significantly positive at 0.5817, both using BG estimates. The results are economically intuitive. We also estimate the individual fixed effects α_i and plot their distribution in Figure 4. We find that most estimated fixed effects are in the range of $[2, 4]$, although the maximum $\widehat{\alpha}_i$ can be as large as 9.5, demonstrating significant heterogeneity in unobserved characteristics among the households. Finally, we compute the APEs and summarize the results in Table 7. We find that the APEs of wealth difference are not significant based on both plug-in and bagging estimators. Distance between households and social ties, on the other hand, matter more significantly on average to form a link.

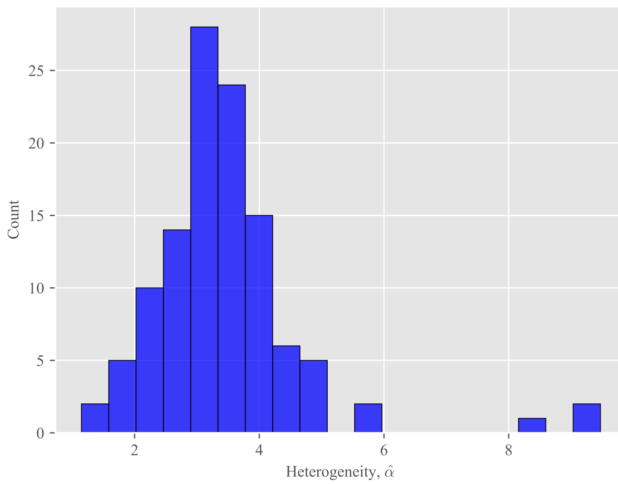


Figure 4: Histogram of the Estimated Individual Heterogeneity

Table 7: Average Partial Effects for Nyakatoke Network

	Plug-in	Bagging
|(ln) wealth difference|	-0.0065	-0.0083
	(0.2124)	(0.1126)
(ln) distance	-0.0576	-0.0641
	(0.0000)	(0.0000)
Tie	0.0514	0.0501
	(0.0000)	(0.0000)

Note: p-values of two-side t-tests are reported in the parentheses.

6.2 India Microfinance Network

As a second application, we apply our method to the India microfinance network dataset (Banerjee et al. (2013, 2024)), which contains detailed demographic information about every household surveyed and their links within each of the 75 villages. A brief discussion of the data is first provided. Then, we present estimation results for both the homophily parameters β_0 and the individual fixed effects $\boldsymbol{\alpha}_0$ for each village. Finally, we investigate via a simulation study how our method can capture important network features such as average degree, clustering effect, and average path lengths that are present in the observed India microfinance network.

6.2.1 Data

Banerjee et al. (2013) conduct a detailed survey among villagers in India, asking them about their daily interactions as well as demographic information such as caste, family size, and wealth with various measures. The survey covers 89.14% of the 16,476 households across 75 Indian villages in the sample. On average, there are $n = 220$ households in each village.

As for the dependent variable, we follow Chandrasekhar and Jackson (2023) to consider two types of links, one defined as "information link" if two households exchange advice with each other and the other as "favor link" if they borrow or lend material goods from each other. As for the covariates, we use six dyadic variables that are constructed based on the demographics of each household. The first set of covariates are binary, defined to be 1 if two households share the same characteristics and 0 otherwise. These binary characteristics include (1) what caste group the household belongs to, (2) whether the household has access to electricity, (3) what type of latrine the household uses, and (4) whether the household owns or rents a house. The second set of covariates include the absolute difference of the number of beds and the number of rooms between any pair of households. In Table 8, we present the summary statistics of these variables.

6.2.2 Estimation Results

We estimate $\boldsymbol{\alpha}_0$ and β_0 for each of the 75 villages for both the information network and the favor network based on our BG estimator. Figure 5 and 6 summarize the distribution of the t-statistics calculated based on the estimated $\widehat{\beta}$'s for the information network and the favor network, respectively. Figure 7a and 7b show the distribution of the estimated $\widehat{\alpha}_i$ for each network.

We draw the following conclusions. First, for the first four binary covariates with "same" in the name which capture whether two households share the same characteristics, the esti-

Table 8: Summary statistics of Indian networks

Variable	Obs	Mean	Std. Dev.	Min	Max
Information link	1,238,970	0.0330	0.1787	0	1
Favor link	1,238,970	0.0388	0.1932	0	1
Same caste	1,238,970	0.4828	0.4997	0	1
Same electricity	1,238,970	0.5244	0.4994	0	1
Same latrine	1,238,970	0.6201	0.4854	0	1
Same ownrent	1,238,970	0.8488	0.3583	0	1
Bed number difference	1,238,970	1.0371	1.4439	0	50
Room number difference	1,238,970	1.2789	1.2898	0	18

mated $\widehat{\beta}$'s are generally significantly positive for both networks with the majority of probability mass of the t-statistics lying to the right of 1.645, the 95^{th} percentile of standard normal distribution. The implication is that two households with the same caste, access to electricity, latrine, own or rent a house are more likely to be linked, which is intuitive. Second, the opposite pattern is observed for the $\widehat{\beta}$'s for the last two discrete covariates that capture how two households differ in the number of beds and rooms for both networks. For example, in Figure 5, most of the t-statistics are negative, suggesting a negative correlation between the difference in the number of beds and rooms and the likelihood to be connected for two households. Third, the t-statistics for the estimated $\widehat{\beta}$ for "same caste" is much larger in absolute value than those for the other covariates such as "same electricity" or "room number difference." It shows that caste plays a crucial role in determining link formation in Indian villages. Finally, from Figure 7a and 7b, we observe significant heterogeneity in the distribution of $\widehat{\alpha}_i$ across individuals among all the villages, which highlights the importance of allowing for unobserved heterogeneity in the model.

6.2.3 Matching Features of Empirical Network Data

As pointed out by Chandrasekhar and Jackson (2023), a challenge for many network formation models has been to capture multiple observed features of networks simultaneously. They show via a simulation study that many network formation models, including stochastic block model, network formation model with degree heterogeneity (Graham (2017)), latent space model (Hoff et al. (2002)), and exponential random graph model, struggle to fit those features that are present in the observed networks in the data. The same rationale applies to our link formation model with NTU that rules out link interdependencies, which would naturally produce low clustering coefficient. Nonetheless, we show that it is possible to resolve this issue by combining it with other independent subgraph formation processes

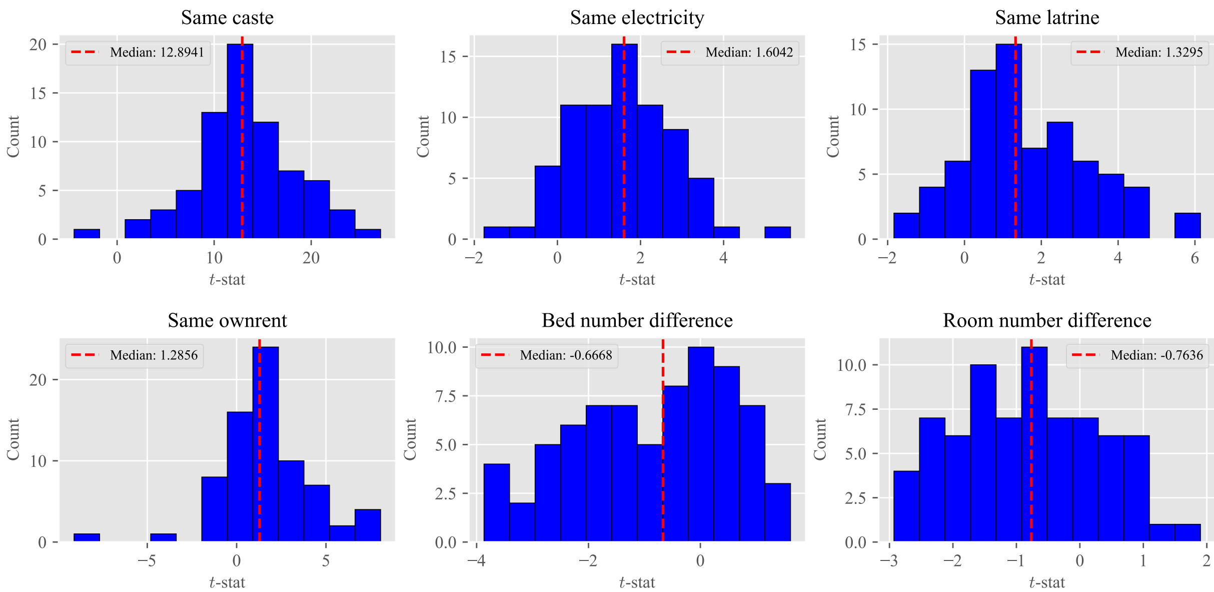


Figure 5: Histograms of t-statistics for the Information Network

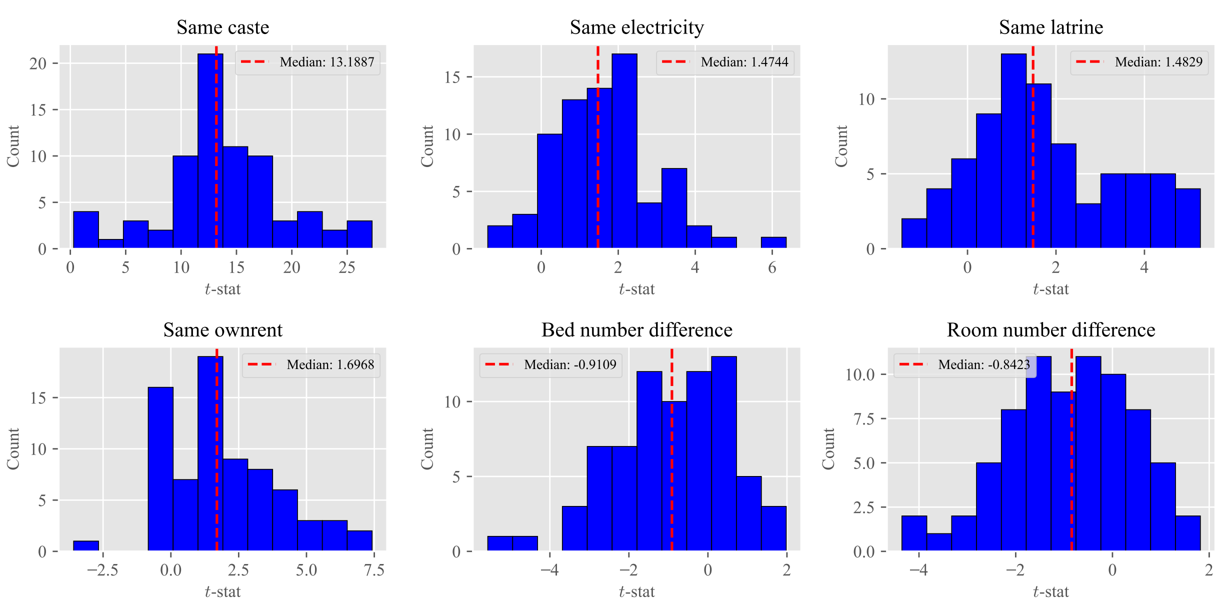
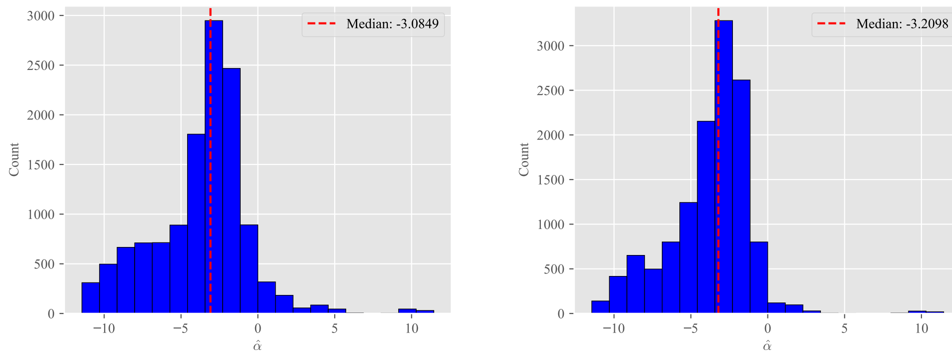


Figure 6: Histograms of t-statistics for the Favor Network



(a) Histogram of $\widehat{\alpha}_i$ for the Information Network

(b) Histogram of $\widehat{\alpha}_i$ for the Favor Network

Figure 7: Fitted heterogeneity

such as triangle formation that introduces only one more unknown parameter (see Newman (2009); Karrer and Newman (2010) for this class of models and Chandrasekhar and Jackson (2023) for an important application and adaptation to econometrics).

We briefly review the features of the graph structure that are included in Table 9. These features are also used by Chandrasekhar and Jackson (2023) in their Table 1. The first one is the average degree which is the average number of links for each node. The second is the clustering coefficient defined by

$$\frac{1}{n}\sum_{i=1}^{n} c_i, \quad \text{with } c_i := \frac{2\sum_{j\neq i}\sum_{k>j,k\neq i} Y_{ij}Y_{jk}Y_{ki}}{d_i(d_i-1)},$$

where $d_i = \sum_{j\neq i} Y_{ij}$ is the degree of node i. It captures the number of triangles that a node lies in on average. The third and fourth measures include the number of isolated nodes and the fraction of the nodes that belongs to the giant component of the network, respectively. The fifth feature is the first eigenvalue of the adjacency matrix, which is a measure of diffusiveness of a network under a percolation process (Jackson (2008)). The sixth measure is the second eigenvalue of the stochasticized adjacency matrix[13], a concept closely related to homophily (Golub and Jackson (2012)) and is similarly labeled in Table 9. Lastly, we consider average path length in the largest component of the graph.

We provide details on how to extend the baseline link formation model (4) to allow for more clustering effect in Appendix D. The idea is to introduce a separate link formation mechanism where the outcome binary variable for whether a triangle is formed or not is i.i.d.

[13]The stochasticized adjacency matrix $\widetilde{Y}$ is defined as $\widetilde{Y}_{ij} = \frac{Y_{ij}}{\sum_k Y_{ik}}$, where either $Y_{ii} = 1$, or $Y_{ik} > 0$ for some $k \neq i$.

Bernoulli across the triads (Newman (2009); Karrer and Newman (2010)) with unknown probability $1 - \lambda_{n0}$. Then, the link between each pair of nodes is determined by the max of the baseline model (4) and the triangle model. For each village, we estimate $(\boldsymbol{\alpha}_0, \lambda_{n0}, \beta_0)$ using the method outlined in Appendix D. Then, for each village we generate 100 simulated networks based on the extended network formation model. We calculate the seven measures of network structure discussed above for each simulated network and average it over the 100 replications to get the network statistics of 75 villages. Finally, we present the average measures and standard errors across the 75 villages. For transparency purpose, we also run the same exercise with our baseline model (4) only and present the results in Table 9.

Table 9 presents the main results. First, the baseline model (4) can capture most of the network features except clustering reasonably well. For example, for the information network, the true average degree is 7.4382 while the simulated average degree is 7.4434 based on the baseline model. The clustering effect, on the other hand, shows a larger discrepancy between the truth of 0.2202 and the simulated value of 0.0951 for the information network, which reveals that dyadic network formation models tend to underestimate the number of triangles in the network, a challenge that is also shared by classic methods of stochastic block models with unobserved heterogeneity (e.g., Graham (2017)), latent space model(e.g., Hoff et al. (2002)), and exponential random graph model. Second, by extending our model (4) to include triangles with only one more unknown parameter, the extended model is able to capture all seven network features very well. Notably, the clustering coefficient simulated from the extended model (0.1868) is significantly closer to the truth (0.2202) than what the baseline model generates (0.0951). A similar pattern is observed for the favor network. Meanwhile, including triangles into the model does not change other network measures, especially the average degree, by much.

It should be pointed out that we do not have theoretical results for estimating the extended model. Thus, the results presented in Table 9 is mainly to illustrate the flexibility of our baseline mode (4) and how to adapt it to capture triangles of a graph. We leave a thorough investigation of the theoretical properties of the extended model for future research.

7 Conclusion

In this paper, we propose an easy-to-compute bagging estimator for the homophily coefficients in a dyadic network formation model with NTU. We show that the proposed bagging estimator has desirable theoretical properties of being asymptotically normal and $\sqrt{N}$-consistent, as well as achieving the Cramér-Rao lower bound. We also propose uniformly consistent estimators for the high-dimensional individual fixed effects. Two extensions to

Table 9: Summary statistics of fitted networks with covariates

	Information network			*Favor network*		
	True	Baseline	Extended	True	Baseline	Extended
Degree	7.4382	7.4434	7.6757	6.3670	6.3751	6.5948
	(2.0282)	(2.0118)	(2.1508)	(2.0063)	(1.9955)	(2.1624)
Clustering	0.2202	0.0951	0.1868	0.2846	0.0804	0.1728
	(0.0517)	(0.0426)	(0.0578)	(0.0514)	(0.0427)	(0.0655)
Isolates	7.2933	10.5461	9.4551	6.8933	10.9647	10.4915
	(4.8049)	(6.7639)	(6.3471)	(4.2851)	(6.2719)	(6.7905)
% Giant	0.9560	0.9387	0.9455	0.9510	0.9361	0.9398
	(0.0260)	(0.0340)	(0.0327)	(0.0519)	(0.0287)	(0.0322)
Max Eigenvalue	10.7867	11.1938	11.3919	9.0273	9.3499	9.5249
	(2.8043)	(2.5858)	(2.5951)	(2.4847)	(2.3669)	(2.3875)
Homophily	0.8703	0.7527	0.7348	0.9396	0.8135	0.7711
	(0.0795)	(0.0837)	(0.0800)	(0.0497)	(0.0869)	(0.0881)
Ave Path Length	3.0099	2.7826	2.7612	3.5436	3.0037	2.9363
	(0.4185)	(0.3048)	(0.2958)	(0.6299)	(0.3543)	(0.3403)

consider APEs and model specification are presented. The efficacy of all our estimators is examined via extensive simulations. We find that the estimators perform well under small sample size, various DGP configurations, relatively sparser network, and model misspecification. Finally, two applications of the method to the Nyakatoke risk-sharing network data and India microfinance network data highlight its usefulness and relevance.

There are several research questions worthy of further examining related to the findings of this paper. Being the first paper providing estimation and inference results for the homophily coefficients in a dyadic network formation model with covariates, individual fixed effects, and NTU, our theoretical analysis relies on the additivity among the index term $X'\beta$, fixed effects α, and idiosyncratic shocks ϵ in specifying the utility surplus from the link for each individual. Additionally, we require the knowledge of the distribution of ϵ. Relaxing these restrictions is desirable to make the result more robust and reliable, however at the cost of more complications in deriving the theory and implementation of the methods. We consider applying the theory of sieve MLE method (Shen (1997); Chen (2007)) to generalize the results of this paper to be potentially useful and promising. Second, as noted in the introduction, one limitation of our model is that we exclude interdependencies in link preferences. So, a natural question to ask is whether it is possible to test for the assumption of no interdependencies in dyadic network formation models with NTU? Essentially, this boils down to testing the "neglected transitivity" in link formation process (see Graham (2017); Dzemski (2019) for the case of TU). It would be useful to adapt existing tests (e.g., the LM test for neglected transitivity by Hahn, Moon, and Snider (2017)) to the NTU setting. Perhaps an even more challenging question is how to do estimation and inference in a dyadic network formation model with NTU and link dependency. We leave these research questions to future work.

References

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2015): "Systemic risk and stability in financial networks," *American Economic Review*, 105, 564–608.

Amemiya, T. (1985): *Advanced econometrics*, Harvard university press.

Auerbach, E. (2022): "Identification and estimation of a partially linear regression model using network data," *Econometrica*, 90, 347–365.

Banerjee, A., E. Breza, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, and C. Kinnan (2024): "Changes in social network structure in response to exposure to formal credit markets," *Review of Economic Studies*, 91, 1331–1372.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498.

Battaglini, M., E. Patacchini, and E. Rainone (2022): "Endogenous social interactions with unobserved networks," *The Review of Economic Studies*, 89, 1694–1747.

Blitzstein, J. and P. Diaconis (2011): "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Mathematics*, 6, 489–522.

Boucher, V. and I. Mourifié (2017): "My friend far, far away: a random field approach to exponential random graph models," *The Econometrics Journal*, 20, S14–S46.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration inequalities: a nonasymptotic theory of independence*, Oxford University Press.

Breiman, L. (1996): "Bagging predictors," *Machine learning*, 24, 123–140.

Candelaria, L. E. (2024): "A semiparametric network formation model with unobserved linear heterogeneity," *arXiv preprint arXiv:2007.05403*.

Chandrasekhar, A. G. and M. O. Jackson (2023): "A network formation model based on subgraphs," *arXiv preprint arXiv:1611.07658*.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

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

Chen, M., I. Fernández-Val, and M. Weidner (2021): "Nonlinear factor models for network and panel data," *Journal of Econometrics*, 220, 296–324.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): "Local identification of nonparametric and semiparametric models," *Econometrica*, 82, 785–809.

de Paula, Á., S. Richards-Shubik, and E. Tamer (2018): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288.

de Paula, Áureo (2020): "Strategic network formation," in *The Econometric Analysis of Network Data*, Elsevier, 41–61.

De Weerdt, J. (2004): "Risk-sharing and endogenous group formation," in *Insurance against Poverty*, ed. by S. Dercon, Oxford University Press, chap. 10.

De Weerdt, J. and S. Dercon (2006): "Risk-sharing networks and insurance against illness," *Journal of Development Economics*, 81, 337–356.

De Weerdt, J. and M. Fafchamps (2011): "Social identity and the formation of health insurance networks," *Journal of Development Studies*, 47, 1152–1177.

Dhaene, G. and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82, 991–1030.

Dzemski, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

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

——— (2018): "Fixed effects estimation of large-T panel data models," *Annual Review of Economics*, 10, 109–138.

Gao, W. Y. (2020): "Nonparametric identification in index models of link formation," *Journal of Econometrics*, 215, 399–413.

Gao, W. Y., M. Li, and S. Xu (2023): "Logical differencing in dyadic network formation models with nontransferable utilities," *Journal of Econometrics*, 235, 302–324.

Goldsmith-Pinkham, P. and G. W. Imbens (2013): "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 31, 253–264.

Golub, B. and M. O. Jackson (2012): "How homophily affects the speed of learning and best-response dynamics," *The Quarterly Journal of Economics*, 127, 1287–1338.

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

——— (2020): "Network data," in *Handbook of Econometrics*, Elsevier, vol. 7, 111–218.

——— (2024): "Sparse network asymptotics for logistic regression under possible misspecification," *Econometrica*, 92, 1837–1868.

Gualdani, C. (2021): "An econometric model of network formation with an application to board interlocks between firms," *Journal of Econometrics*, 224, 345–370.

Hahn, J. and G. Kuersteiner (2011): "Bias reduction for dynamic nonlinear panel models with fixed effects," *Econometric Theory*, 27, 1152–1191.

Hahn, J., H. R. Moon, and C. Snider (2017): "LM test of neglected correlated random effects and its application," *Journal of Business & Economic Statistics*, 35, 359–370.

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

Hirano, K. and J. H. Wright (2017): "Forecasting with model uncertainty: Representations and risk reduction," *Econometrica*, 85, 617–643.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Sssociation*, 97, 1090–1098.

Hsieh, C.-S. and L. F. Lee (2016): "A social interactions model with endogenous friendship formation and selectivity," *Journal of Applied Econometrics*, 31, 301–319.

Hughes, D. W. (2023): "Estimating nonlinear network data models with fixed effects," *arXiv preprint arXiv:2203.15603*.

Jackson, M. (2008): *Social and Economic Networks*, Princeton University Press.

Jackson, M. O., Z. Lin, and N. N. Yu (2024): "Adjusting for peer-influence in propensity scoring when estimating treatment effects," *Working Paper*.

Jackson, M. O. and A. Wolinsky (1996): "A strategic model of social and economic networks," *Journal of Economic Theory*, 71, 44–74.

Jochmans, K. (2017): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 1–9.

Jochmans, K. and M. Weidner (2019): "Fixed-effect regressions on network data," *Econometrica*, 87, 1543–1560.

Johnsson, I. and H. R. Moon (2021): "Estimation of peer effects in endogenous social networks: Control function approach," *Review of Economics and Statistics*, 103, 328–345.

Karrer, B. and M. E. Newman (2010): "Random graphs containing arbitrary distributions of subgraphs," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82, 066118.

König, M. D., D. Rohner, M. Thoenig, and F. Zilibotti (2017): "Networks in conflict: Theory and evidence from the great war of africa," *Econometrica*, 85, 1093–1132.

Le Cam, L. M. (1969): "Théorie asymptotique de la décision statistique," *Les Presses de l'Universitede Montreal, Montreal.*

Leung, M. P. (2019): "A weak law for moments of pairwise stable networks," *Journal of Econometrics*, 210, 310–326.

Leung, M. P. and H. R. Moon (2019): "Normal approximation in large network models," *arXiv preprint arXiv:1904.11060.*

Lewbel, A., X. Qu, and X. Tang (2023): "Social networks with unobserved links," *Journal of Political Economy*, 131, 898–946.

Mei, Z., L. Sheng, and Z. Shi (2024): "Nickell bias in panel local projection: Financial crises are worse than you think," *arXiv preprint arXiv:2302.13455.*

Mele, A. (2017): "A structural model of dense network formation," *Econometrica*, 85, 825–850.

——— (2022): "A structural model of homophily and clustering in social networks," *Journal of Business & Economic Statistics*, 40, 1377–1389.

Menzel, K. (2016): "Inference for games with many players," *The Review of Economic Studies*, 83, 306–337.

——— (2024): "Strategic network formation with many agents," Working Paper.

Miyauchi, Y. (2016): "Structural estimation of pairwise stable networks with nonnegative externality," *Journal of Econometrics*, 195, 224–235.

Moreira, M. J. (2009): "A maximum likelihood method for the incidental parameter problem," *The Annals of Statistics*, 37, 3660 – 3696.

Newman, M. E. (2009): "Random graphs with clustering," *Physical Review Letters*, 103, 058701.

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

Qu, L., L. Chen, T. Yan, and Y. Chen (2024): "Inference in semiparametric formation models for directed networks," *arXiv preprint arXiv:2405.19637.*

Ridder, G. and S. Sheng (2020): "Two-step estimation of a strategic network formation model with clustering," *arXiv preprint arXiv:2001.03838.*

Robbins, H. and S. Monro (1951): "A stochastic approximation method," *The Annals of Mathematical Statistics*, 400–407.

Shen, X. (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 25, 2555–2591.

Sheng, S. (2020): "A structural econometric analysis of network formation games through subnetworks," *Econometrica*, 88, 1829–1858.

Toth, P. (2017): "Semiparametric estimation in network formation models with homophily and degree heterogeneity," SSRN 2988698.

Van der Vaart, A. W. (2000): *Asymptotic Statistics*, vol. 3, Cambridge university press.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 1–25.

——— (1996): *Estimation, inference and specification analysis*, 22, Cambridge university press.

Yan, T. (2019): "Approximating the inverse of a diagonally dominant matrix with positive elements," *arXiv preprint arXiv:1902.00668*.

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

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

Yan, T., H. Qin, and H. Wang (2016b): "Asymptotics in undirected random graph models parameterized by the strengths of vertices," *Statistica Sinica*, 273–293.

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

Zeleneev, A. (2020): "Identification and estimation of network models with nonparametric unobserved heterogeneity," *Department of Economics, Princeton University*.

Appendix

A Matrices and Lemmas

In this Appendix, we first give explicit formulas for the various matrices used in the main text. Then, we present several lemmas that are used in the following proofs. Recall some notations: for an $n \times n$ matrix $\mathbf{A}$, we write $\|\mathbf{A}\|_1 := \max_{1\le i\le n} \|\mathbf{A}_{\cdot i}\|_1$, $\|\mathbf{A}\|_\infty := \max_{1\le i\le n} \|\mathbf{A}_{i\cdot}\|_1$ and $\|\mathbf{A}\|_{\max} := \max_{1\le i,j\le n} |\mathbf{A}_{ij}|$, where $\mathbf{A}_{\cdot i}$ and $\mathbf{A}_{i\cdot}$ are the ith column and row of $\mathbf{A}$, respectively. In the appendix, we write "$a_n \asymp b_n$" to denote $a_n = O(b_n)$ and $b_n = O(a_n)$ simultaneously, use $C_1, C_2, \ldots$ to denote strictly positive and finite constants.

A.1 Definitions of Matrices in the Main Text

Jacobian matrix. First, $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ is an $n \times n$ matrix with its off-diagonal and diagonal elements equal to

$$\begin{aligned} [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ij} &= -F_{ij}(\boldsymbol{\alpha},\beta) f_{ji}(\boldsymbol{\alpha},\beta), \qquad 1 \le i \ne j \le n \text{ and} \\ [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ii} &= -\sum_{j\ne i} f_{ij}(\boldsymbol{\alpha},\beta) F_{ji}(\boldsymbol{\alpha},\beta), \qquad i = 1,\ldots,n, \end{aligned}$$

respectively. Clearly, $[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ij} \ne [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ji}$. Moreover, there is a specific relationship between diagonal and off-diagonal elements, i.e.,

$$[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ii} = \sum_{j\ne i} [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ji}, \qquad i = 1,\ldots,n.$$

Hence, $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)^\top$ is asymmetric and diagonally dominant with strictly positive entries by Assumption 3. We prove that $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ is invertible under Assumptions 2–3 below.
Next, $\mathbf{J}_{12}(\boldsymbol{\alpha},\beta)$ is an $n \times K$ matrix with its ith row written as

$$-\sum_{j\ne i} \left[f_{ij}(\boldsymbol{\alpha},\beta) F_{ji}(\boldsymbol{\alpha},\beta) + F_{ij}(\boldsymbol{\alpha},\beta) f_{ji}(\boldsymbol{\alpha},\beta) \right] x_{ij}^\top.$$

Similarly, $\mathbf{J}_{21}(\boldsymbol{\alpha},\beta)$ is a $K \times n$ matrix and its ith column is $-\sum_{j\ne i} f_{ij}(\boldsymbol{\alpha},\beta) F_{ji}(\boldsymbol{\alpha},\beta) x_{ij}$. Finally,

$$\mathrm{J}_{22}(\boldsymbol{\alpha},\beta) = -\sum_{i=1}^{n}\sum_{j\ne i} f_{ij}(\boldsymbol{\alpha},\beta) F_{ji}(\boldsymbol{\alpha},\beta) x_{ij} x_{ij}^\top$$

is a $K \times K$ matrix.

Variance matrix of moment equations. $\mathbf{V}_{11}(\boldsymbol{\alpha},\beta)$ is an $n \times n$ matrix. The off-diagonal and diagonal elements of $\mathbf{V}_{11}(\boldsymbol{\alpha}, \beta)$ are

$$
\begin{aligned}
[\mathbf{V}_{11}(\boldsymbol{\alpha}, \beta)]_{ij} =& p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)), \qquad 1 \le i \ne j \le n \text{ and} \\
[\mathbf{V}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} =& \sum_{j \ne i} p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ji}(\boldsymbol{\alpha}, \beta)), \qquad i = 1, \ldots, n,
\end{aligned}
$$

respectively. $\mathbf{V}_{12}$ is an $n \times K$ matrix with its ith row equal to $\sum_{j \ne i} p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) x_{ij}^\top$. Finally, $\mathbf{V}_{22} = \sum_{i=1}^n \sum_{j > i} p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) x_{ij} x_{ij}^\top$.

Hessian matrix. For $\mathbf{H}_{11}(\boldsymbol{\alpha}, \beta)$, an $n \times n$ matrix, it has entries:

$$
\begin{aligned}
[\mathbf{H}_{11}(\boldsymbol{\alpha}, \beta)]_{ij} =& -\frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{(1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2}, \qquad 1 \le i \ne j \le n, \\
[\mathbf{H}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} =& \sum_{j \ne i} \Bigg[-\frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))} + (y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)) \\
& \times \frac{f_{ij}^{(1)}(\boldsymbol{\alpha}, \beta) F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) - f_{ij}^2(\boldsymbol{\alpha}, \beta)(1 - 2p_{ij}(\boldsymbol{\alpha}, \beta))}{F_{ij}^2(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2} \Bigg], \ i = 1, \ldots, n.
\end{aligned}
$$

Next, $\mathbf{H}_{12}(\boldsymbol{\alpha}, \beta)$ is an $n \times K$ matrix and its ith row can be written as

$$
\begin{aligned}
& \sum_{j \ne i} \Bigg[-(1 - y_{ij}) \frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{(1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2} - \frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))} \\
& \quad + (y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)) \frac{f_{ij}^{(1)}(\boldsymbol{\alpha}, \beta) F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) - f_{ij}^2(\boldsymbol{\alpha}, \beta)(1 - 2p_{ij}(\boldsymbol{\alpha}, \beta))}{F_{ij}(\boldsymbol{\alpha}, \beta)^2 (1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2} x_{ij}^\top \Bigg].
\end{aligned}
$$

Finally, $\mathbf{H}_{22}(\boldsymbol{\alpha}, \beta)$ equals

$$
\begin{aligned}
& \sum_{i=1}^n \sum_{j \ne i} \Bigg[-(1 - y_{ij}) \frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{(1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2} - \frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))} \\
& \quad + (y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)) \times \frac{f_{ij}^{(1)}(\boldsymbol{\alpha}, \beta) F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) - f_{ij}^2(\boldsymbol{\alpha}, \beta)(1 - 2p_{ij}(\boldsymbol{\alpha}, \beta))}{F_{ij}^2(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))^2} x_{ij} x_{ij}^\top \Bigg].
\end{aligned}
$$

Information matrix. First, $\mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)$ is an $n \times n$ matrix with off-diagonal elements and diagonal elements equal to

$$
\begin{aligned}
[\mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)]_{ij} =& \frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{1 - p_{ij}(\boldsymbol{\alpha}, \beta)}, \qquad 1 \le i \ne j \le n, \text{ and} \\
[\mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} =& \sum_{j \ne i} \frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))}, \qquad i = 1, \ldots, n,
\end{aligned}
$$

respectively. Next, $\mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)$ is an $n \times K$ matrix with its ith row written as

$$\sum_{j \neq i} \left[\frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{1 - p_{ij}(\boldsymbol{\alpha}, \beta)} + \frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))} \right] x_{ij}^{\top}.$$

Finally, $\mathrm{I}_{22}(\boldsymbol{\alpha}, \beta)$ equals

$$\sum_{i=1}^{n} \sum_{j \neq i} \left[\frac{f_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta)}{1 - p_{ij}(\boldsymbol{\alpha}, \beta)} + \frac{f_{ij}^2(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta)}{F_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta))} \right] x_{ij} x_{ij}^{\top}.$$

In what follows, we use the mean value theorem for vector-valued functions in its integral form, which is also used in Chatterjee et al. (2011). For example,

$$\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}, \beta) - \mathbf{m}_1(\boldsymbol{\alpha}, \beta) = \left[\int_0^1 \mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta) dt \right] (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) =: \mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}; \beta) \left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \right).$$

We write $\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}; \beta)$ as $\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha})$ whenever there is no confusion, other integral form Jacobian matrices are defined similarly. Notice that for each fixed $t \in (0, 1)$, we have $[\mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta)]_{ii} = \sum_{j \neq i} [\mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta)]_{ji}$, so

$$[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha})]_{ii} = \int_0^1 \sum_{j \neq i} [\mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta)]_{ji} \, dt = \sum_{j \neq i} [\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha})]_{ji},$$

which implies $\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha})$ inherits the diagonally dominant property from $\mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta)$.

A.2 Analytic Approximation of $\mathbf{J}_{11}^{-1}(\boldsymbol{\alpha}, \beta)$

We adapt Theorem 1 of Yan (2019) to the NTU framework here to provide an analytic approximation for the inverse of the Jacobian matrix $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ and bound the approximation errors. Similar approximation techniques have been used in proving asymptotic normality for network estimation problems, e.g., Yan and Xu (2013), Yan, Jiang, Fienberg, and Leng (2019), Graham (2017). We prove that $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ is non-singular for n large enough and $\mathbf{J}_{11}^{-1}(\boldsymbol{\alpha}, \beta)$ is well approximated by a diagonal matrix.

Lemma 1. *(Yan (2019)) Suppose an $n \times n$ matrix $\mathbf{A} = (a_{ij})_{n \times n}$ is invertible with its entries all positive and $a_{ii} \geq \sum_{j \neq i} a_{ji}$. Let $\mathbf{B} = [\mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})]^{-1}$, $\Delta_i = a_{ii} - \sum_{j \neq i} a_{ji}$, $M \equiv \max\{\max_{1 \leq i \neq j \leq n} a_{ij}, \max_{i=1,\ldots,n} \Delta_i\}$, and $m \equiv \min_{1 \leq i \neq j \leq n} a_{ij}$. If $M \asymp 1$ and $m \asymp 1$ we have*

$$\|\mathbf{A}^{-1} - \mathbf{B}\|_{\max} = O(n^{-2}). \tag{44}$$

Proof. The proof is adapted from Yan (2019), for completeness, we present what is different from his proof here. Let I_n be the $n \times n$ identity matrix. Define

$$\mathbf{F} = (f_{ij})_{n\times n} = \mathbf{A}^{-1} - \mathbf{B}, \quad \mathbf{U} = (u_{ij})_{n\times n} = I_n - \mathbf{AB}, \quad \mathbf{W} = (w_{ij})_{n\times n} = \mathbf{BU}.$$

Then, we have

$$\mathbf{F} = \mathbf{A}^{-1} - \mathbf{B} = (\mathbf{A}^{-1} - \mathbf{B})(I_n - \mathbf{AB}) + \mathbf{B}(I_n - \mathbf{AB}) = \mathbf{FU} + \mathbf{W}. \tag{45}$$

Some algebra leads to

$$u_{ij} = \delta_{ij} - \sum_{k=1}^{n} a_{ik} b_{kj} = \delta_{ij} - \sum_{k=1}^{n} a_{ik} \frac{\delta_{kj}}{a_{jj}} = \delta_{ij} - \frac{a_{ij}}{a_{jj}} = (\delta_{ij} - 1)\frac{a_{ij}}{a_{jj}}, \tag{46}$$

and

$$w_{ij} = \sum_{k=1}^{n} b_{ik} u_{kj} = \sum_{k=1}^{n} \frac{\delta_{ik}}{a_{ii}} (\delta_{kj} - 1) \frac{a_{kj}}{a_{jj}} = \frac{(\delta_{ij} - 1) a_{ij}}{a_{ii} a_{jj}}. \tag{47}$$

Recall that $m \le a_{ij} \le M$ and $(n-1)m \le a_{ii} \le (n-1)M$. When $i \ne j$, we have

$$0 < \frac{a_{ij}}{a_{ii} a_{jj}} \le \frac{M}{m^2 (n-1)^2},$$

such that for $i \ne j \ne k$, the following bounds hold

$$w_{ii} = 0, \quad |w_{ij}| \le \frac{M}{m^2(n-1)^2}, \quad |w_{ii} - w_{ik}| = |w_{ik}| \le \frac{M}{m^2(n-1)^2},$$

$$|w_{ij} - w_{ik}| \le \max(w_{ij}, w_{ik}) \le \frac{M}{m^2(n-1)^2}$$

It follows that

$$\max(|w_{ij}|, |w_{ij} - w_{ik}|) \le \frac{M}{m^2(n-1)^2}, \qquad \text{for all } i, j, k. \tag{48}$$

We use (45) to obtain a bound for the approximate error $\|\mathbf{F}\|_{\max}$. By (45) and (46), for any $i \le n$, we have

$$f_{ij} = \sum_{k=1}^{n} f_{ik} u_{kj} + w_{ij} = \sum_{k=1}^{n} f_{ik} (\delta_{kj} - 1) \frac{a_{kj}}{a_{jj}} + w_{ij}. \tag{49}$$

Define $f_{i\theta} = \max_{1\le k\le n} f_{ik}$ and $f_{i\xi} = \min_{1\le k \le n} f_{ik}$. First, we show that $f_{i\xi} < 0$. Since for any fixed i, we have

$$\sum_{k=1}^{n} f_{ik} a_{ki} = \sum_{k=1}^{n} \left([\mathbf{A}^{-1}]_{ik} - \frac{\delta_{ik}}{a_{ii}} \right) a_{ki} = 1 - 1 = 0.$$

Hence, $f_{i\xi} \sum_{k=1}^{n} a_{ki} \le \sum_{k=1}^{n} f_{ik} a_{ki} = 0$. So, we have $f_{i\xi} < 0$. Similarly, we have that $f_{i\theta} > 0$.

Recall that

$$a_{\theta\theta} = \sum_{k\neq\theta} a_{k\theta} + \Delta_\theta = \sum_{k=1}^{n}(1-\delta_{k\theta})a_{k\theta} + \Delta_\theta,\ \text{hence, } 1 \equiv \sum_{k=1}^{n}(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}} + \frac{\Delta_\theta}{a_{\theta\theta}} \tag{50}$$

for any θ, which yields the following identities

$$\begin{aligned} f_{i\xi} =& f_{i\xi}\left[\sum_{k=1}^{n}(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}} + \frac{\Delta_\theta}{a_{\theta\theta}}\right] = \sum_{k=1}^{n} f_{i\xi}(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}} + \frac{f_{i\xi}\Delta_\theta}{a_{\theta\theta}}, \\ f_{i\xi} =& f_{i\xi}\left[\sum_{k=1}^{n}(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} + \frac{\Delta_\xi}{a_{\xi\xi}}\right] = \sum_{k=1}^{n} f_{i\xi}(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} + \frac{f_{i\xi}\Delta_\xi}{a_{\xi\xi}}, \end{aligned} \tag{51}$$

where the first and second part of this equation use (50) for $a_{\theta\theta}$ and $a_{\xi\xi}$, respectively.

By combining (49) with the first part of (51) where we set $j=\theta$ in (49), we have

$$f_{i\theta} + f_{i\xi} = \sum_{k=1}^{n}(f_{i\xi} - f_{ik})(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}} + w_{i\theta} + \frac{f_{i\xi}\Delta_\theta}{a_{\theta\theta}}. \tag{52}$$

Similarly, we have

$$2f_{i\xi} = \sum_{k=1}^{n}(f_{i\xi} - f_{ik})(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} + w_{i\xi} + \frac{f_{i\xi}\Delta_\xi}{a_{\xi\xi}}. \tag{53}$$

Subtracting (53) from (52), we have

$$f_{i\theta} - f_{i\xi} = \sum_{k=1}^{n}(f_{ik} - f_{i\xi})\left[(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} - (1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}}\right] + w_{i\theta} - w_{i\xi} + f_{i\xi}\left(\frac{\Delta_\theta}{a_{\theta\theta}} - \frac{\Delta_\xi}{a_{\xi\xi}}\right). \tag{54}$$

Let $\Omega = \{k : (1-\delta_{k\xi})a_{k\xi}/a_{\xi\xi} \geq (1-\delta_{k\theta})a_{k\theta}/a_{\theta\theta}\}$ and define λ as the cardinality of Ω. Notice that $1-\delta_{\theta\theta}=0$ and $1-\delta_{\xi\xi}=0$, we have $\theta\in\Omega$ and $\xi\notin\Omega$ (here we assume that $\theta\neq\xi$. Otherwise, when $\theta=\xi$ we have $f_{i\theta}=f_{i\xi}=0$, which is trivial.) Consequently, the cardinality satisfies $1\leq\lambda\leq n-1$, and then

$$\begin{aligned} &\sum_{k=1}^{n}(f_{ik} - f_{i\xi})\left[(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} - (1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}}\right] \\ \leq\ & \sum_{k\in\Omega}(f_{ik} - f_{i\xi})\left[(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}} - (1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}}\right] \\ \leq\ & (f_{i\theta} - f_{i\xi})\left[\frac{\sum_{k\in\Omega} a_{k\xi}}{a_{\xi\xi}} - \frac{\sum_{k\in\Omega}(1-\delta_{k\theta})a_{k\theta}}{a_{\theta\theta}}\right] \\ \leq\ & (f_{i\theta} - f_{i\xi})\left[\frac{\lambda M}{\lambda M + (n-1-\lambda)m} - \frac{(\lambda-1)m}{(\lambda-1)m + (n-\lambda+1)M}\right] \\ \leq\ & (f_{i\theta} - f_{i\xi})\left\{\frac{nM-(n-2)m}{nM+(n-2)m} + \frac{(n-2)Mm}{[(n-2)m+M][(n-2)m+2M]}\right\}, \end{aligned} \tag{55}$$

where the last inequality comes from equations (15)-(17) of Yan (2019), which is obtained by a maximization with respect to λ. Because

$$f_{i\xi}\left(\frac{\Delta_\theta}{a_{\theta\theta}}-\frac{\Delta_\xi}{a_{\xi\xi}}\right)\le (f_{i\theta}-f_{i\xi})\frac{2M}{m(n-1)}. \tag{56}$$

Combining (54), (55), and (56), we have

$$f_{i\theta}-f_{i\xi}\le\frac{\max_{i,j,k}|w_{ik}-w_{i\xi}|}{C(n,m,M)}\le\frac{M}{m^2(n-1)^2C(n,m,M)},$$

with

$$\begin{aligned}C(n,m,M)&=1-\frac{nM-(n-2)m}{nM+(n-2)m}-\frac{(n-2)Mm}{[(n-2)m+M][(n-2)m+2M]}-\frac{2M}{m(n-1)}\\&=\frac{2(n-2)m}{nM+(n-2)m}-\frac{(n-2)Mm}{[(n-2)m+M][(n-2)m+2M]}-\frac{2M}{m(n-1)}\\&\asymp 1.\end{aligned}$$

provided that $m/M\asymp 1$. This proves that for each i, we have $\max_{k=1,\dots,n}|f_{ik}|\le f_{i\theta}-f_{i\xi}=O(n^{-2})$ as $m,M\asymp 1$. Hence, we have shown $\|\mathbf{A}^{-1}-\mathbf{B}\|_{\max}=\|\mathbf{F}\|_{\max}=O(n^{-2})$. $\square$

Based on this lemma, we prove that $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ is non-singular for $(\boldsymbol{\alpha},\beta)\in\mathbb{A}\times\mathbb{B}$ and large n.

Lemma 2. *Under Assumptions 2 and 3, for n large enough, the Jacobian matrix $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ is invertible for all $(\boldsymbol{\alpha},\beta)\in\mathbb{A}\times\mathbb{B}$.*

Proof of Lemma 2. We partition $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ into a block matrix as

$$\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)=\begin{pmatrix}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times(1:n-1)} & [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\\ [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)} & [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{nn}\end{pmatrix},$$

where the subscript denotes the specific rows/columns that each sub-matrix includes. Recall that $[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ii}=\sum_{j\neq i}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ji}$, the first sub-matrix $[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times(1:n-1)}$ is strictly diagonally dominant with all negative entries, hence it is non-singular. Lemma 1 demonstrates that its inverse can be approximated by $\operatorname{diag}\left([\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{11}^{-1},\dots,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n-1n-1}^{-1}\right)$ with maximum entry-wise error of $O(n^{-2})$. Under Assumptions 2 and 3, $[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ii}\asymp -n$, $[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ij}\asymp -1$, $j\neq i$, and

$$\begin{aligned}&[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]^{-1}_{(1:n-1)\times(1:n-1)}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\\=\;&[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)}\operatorname{diag}\left([\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{11}^{-1},\dots,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n-1n-1}^{-1}\right)[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\\&+O(n^{-2})\times[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)}\mathbf{1}\mathbf{1}^\top[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\neq n} \frac{[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ni}\,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{in}}{\sum_{j\neq i}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ji}} + O(n^{-2})\left\{\sum_{i\neq n}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{ni}\right\}\times\left\{\sum_{i\neq n}[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{in}\right\}\\
&= O(1).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{nn} - [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)}\,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]^{-1}_{(1:n-1)\times(1:n-1)}\,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\\
&\qquad\qquad \asymp -n - O(1) \neq 0
\end{aligned}
$$

for n large enough. Finally, by the formula for the determinants of block matrices, we have

$$
\begin{aligned}
&\det[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]\\
&= \det[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times(1:n-1)}\\
&\quad\times\left\{[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{nn} - [\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n\times(1:n-1)}\,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]^{-1}_{(1:n-1)\times(1:n-1)}\,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{(1:n-1)\times n}\right\}\\
&\neq 0.
\end{aligned}
$$

for n large enough. Hence, $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ is invertible for large n. $\square$

For the inverse of $\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$, it is straightforward to verify that $-\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$ satisfies conditions in 1. Let $\mathbf{T}(\boldsymbol{\alpha},\beta) = [\operatorname{diag}(\mathbf{J}_{11}(\boldsymbol{\alpha},\beta))]^{-1}$. Applying Lemma 1 to $-\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)$, we have $\|[-\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]^{-1} + \mathbf{T}(\boldsymbol{\alpha},\beta)\|_{\max} = O(n^{-2})$ under Assumptions 2 and 3. All of these results could also be applied to $\mathbf{J}^{\circ}_{11}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})$, we use $\mathbf{T}^{\circ}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})$ to denote the diagonal approximation for $[\mathbf{J}^{\circ}_{11}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})]^{-1}$.

A.3 Deviation Bound

We give some non-asymptotic deviation bounds in this subsection. The following probabilities are defined conditional on $\boldsymbol{\alpha}$ and $\mathbf{x}$ and we suppress such conditioning whenever there is no confusion. Lemma 3 below controls the deviation of the weighted sum of centered Bernoulli random variables, i.e., $\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})$. This result will be used extensively in the proof.

Lemma 3. *Under Assumptions 2 and 3, and bounded constants* $\max_{i,j}|\lambda_{ij}| < C_1$, *we have*

$$
\Pr\left\{\max_{1\le i\le n}\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right| > C_1\sqrt{\frac{6\log n}{n-1}}\right\} \le 2n^{-2}. \tag{57}
$$

Proof. First, notice that $|\lambda_{ij}(y_{ij}-p_{ij})| < 2C_1$ because $y_{ij}-p_{ij}\in(-1,1)$; in addition, y_{ij}'s are independent Bernoulli random variables with expectations p_{ij}. By Hoeffding's inequality (see

Theorem 2.8 of Boucheron, Lugosi, and Massart (2013)) for sum of bounded and independent random variables, we have

$$\Pr\left(\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right|>t\right)\leq 2\exp\left(-\frac{(n-1)t^2}{2C_1^2}\right).$$

Letting $t=C_1\sqrt{6(n-1)^{-1}\log n}$, we obtain

$$\Pr\left(\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right|>C_1\sqrt{\frac{6\log n}{n-1}}\right)\leq 2n^{-\frac{3(n-1)}{n-1}}=2n^{-3}.$$

By Boole's inequality,

$$\Pr\left(\max_{1\leq i\leq n}\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right|>C_1\sqrt{\frac{6\log n}{n-1}}\right)\leq n\cdot 2n^{-3}=2n^{-2}. \tag{58}$$

We complete the proof. □

Based on Lemma 3, we can bound the estimation error of $\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0$, which guarantees that our moment estimator for $\boldsymbol{\alpha}_0$ would be uniformly consistent if β_0 were known. This result can be strengthened to prove the second part of Theorem 2, which we will do in Appendix B.

Lemma 4. *Under Assumptions 2 and 3, for bounded constants* $\max_{i,j}|\lambda_{ij}|<C_1$, *we have*

$$\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{n}\sum_{j>i}\lambda_{ij}(y_{ij}-p_{ij})\right|>C_1\sqrt{\frac{2\log N}{N}}\right\}\leq (n(n-1))^{-1}. \tag{59}$$

Proof. Similar to the proof of Lemma 3, by Hoeffding's inequality, we have

$$\Pr\left(\frac{1}{N}\left|\sum_{i=1}^{n}\sum_{j>i}\lambda_{ij}(y_{ij}-p_{ij})\right|>t\right)\leq 2\exp\left(-\frac{Nt^2}{2C_1^2}\right).$$

Letting $t=C_1\sqrt{\frac{2\log N}{N}}$, we obtain

$$\Pr\left(\frac{1}{N}\left|\sum_{i=1}^{n}\sum_{j>i}\lambda_{ij}(y_{ij}-p_{ij})\right|>C_1\sqrt{\frac{2\log N}{N}}\right)\leq 2N^{-1}=(n(n-1))^{-1}. \ \square$$

Lemma 5. *Under Assumptions 2 and 3, with probability at least* $1-2n^{-2}$, *we have*

$$\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\|_\infty=O\left(\sqrt{\frac{\log n}{n}}\right),$$

and

$$\left\| \sqrt{n}[\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0] + \left(\frac{\mathbf{J}_{11}}{n}\right)^{-1} \frac{\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)}{\sqrt{n}} \right\|_\infty = O\left(\frac{\log n}{\sqrt{n}}\right). \tag{60}$$

Proof. The rest of proof is conditional on the following event, which happens with probability at least $1 - 2n^{-2}$ by Lemma (3):

$$\mathcal{E}_n := \left\{ \max_{1 \le i \le n} \frac{1}{n-1} \left| \sum_{j \ne i} (y_{ij} - p_{ij}) \right| \le \sqrt{\frac{6 \log n}{n-1}} = O\left(\sqrt{\frac{\log n}{n}}\right) \right\}.$$

For any finite n, a first-order Taylor expansion of the estimating equation for $\widehat{\boldsymbol{\alpha}}(\beta_0)$, $\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) = 0$, around $\boldsymbol{\alpha}_0$ gives

$$\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) = \mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\left(\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0\right)$$

which implies that

$$\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0 = -\left[\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) \tag{61}$$

because $\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) = 0$ by the definition of $\widehat{\boldsymbol{\alpha}}(\beta_0)$. Recall the diagonal approximation of $\left[\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\right]^{-1}$ is $\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)$. By Lemma 1, we decompose $\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0$ into two parts and apply the triangle inequality:

$$\begin{aligned}
&\|\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0\|_\infty \\
&= \|\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) + [\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)]\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty \\
&\le \|\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty + \|[\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)]\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty \\
&\le \|\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty + \|\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty,
\end{aligned}$$

Let's analyze the two parts on the right hand side of the last line separately. For the first part, notice that $\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)$ is a diagonal matrix and each diagonal element is of order $O(n^{-1})$ uniformly, hence $\|\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty = O(n^{-1})$. Recall the definition of $\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)$ and by Lemma 3, we obtain

$$\|\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty = O(n^{-1}) \cdot \max_{1 \le i \le n} \left| \sum_{j \ne i} (y_{ij} - p_{ij}) \right| = O\left(\sqrt{\frac{\log n}{n}}\right).$$

For the second part, by Lemma 1, we know that

$$\|\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty \le n \|\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_{\max} = O(n^{-1}).$$

Hence we have

$$\|\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0) - \mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)\|_\infty \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty$$

$$= O(n^{-1}) \cdot \max_{1\le i\le n} \left| \sum_{j\neq i} (y_{ij} - p_{ij}) \right| = O\left(\sqrt{\frac{\log n}{n}} \right).$$

Combining these two results, we have

$$\|\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0\|_\infty = O\left(\sqrt{\frac{\log n}{n}} \right).$$

We turn to the proof of (60). By a second-order Taylor expansion, which is also used in the proof of Lemma 6 of Graham (2017),

$$\begin{aligned} &\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) \\ &= \mathbf{J}_{11}(\boldsymbol{\alpha}_0, \beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0] + \frac{1}{2}\left[\sum_{k=1}^n (\hat{\alpha}_k(\beta_0) - \alpha_{k0}) \frac{\partial \mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}}^k, \beta_0)}{\partial \alpha_k} \right] [\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0] \end{aligned} \tag{62}$$

with mean value $\tilde{\boldsymbol{\alpha}}^k$ lies between $\widehat{\boldsymbol{\alpha}}(\beta_0)$ and $\boldsymbol{\alpha}_0$ and it may vary with different k. With a slight abuse of notation, we write all $\tilde{\boldsymbol{\alpha}}^k$ as $\tilde{\boldsymbol{\alpha}}$. Because only the kth row and the kth column of $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ contain functions of α_k, by a direct calculation we summarize the entries of $\Lambda_k := \frac{\partial \mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}}, \beta_0)}{\partial \alpha_k}$ as

$$\begin{aligned} (\Lambda_k)_{pq} &= 0, \quad p \neq k \text{ and } q \neq k, \\ (\Lambda_k)_{kl} &= -f_{kl}(\tilde{\boldsymbol{\alpha}}, \beta_0) f_{lk}(\tilde{\boldsymbol{\alpha}}, \beta_0), \quad l \neq k, \\ (\Lambda_k)_{lk} &= -f_{kl}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{lk}(\tilde{\boldsymbol{\alpha}}, \beta_0), \quad l \neq k, \\ (\Lambda_k)_{kk} &= -\sum_{p\neq k} f_{kp}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{pk}(\tilde{\boldsymbol{\alpha}}, \beta_0). \end{aligned}$$

Hence, let $\Lambda = \sum_{k=1}^n (\hat{\alpha}_k(\beta_0) - \alpha_{k0}) \frac{\partial \mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}}, \beta_0)}{\partial \alpha_k}$, where its entries are

$$\begin{aligned} \Lambda_{ij} &= -(\hat{\alpha}_i(\beta_0) - \alpha_{i0}) f_{ij}(\tilde{\boldsymbol{\alpha}}, \beta_0) f_{ji}(\tilde{\boldsymbol{\alpha}}, \beta_0) - (\hat{\alpha}_j(\beta_0) - \alpha_{j0}) f_{ji}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{ij}(\tilde{\boldsymbol{\alpha}}, \beta_0), \ i \neq j \\ \Lambda_{ii} &= -(\hat{\alpha}_i(\beta_0) - \alpha_{i0}) \sum_{j\neq i} f_{ij}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{ji}(\tilde{\boldsymbol{\alpha}}, \beta_0) - \sum_{k\neq i} (\hat{\alpha}_k(\beta_0) - \alpha_{k0}) f_{ik}(\tilde{\boldsymbol{\alpha}}, \beta_0) f_{ki}(\tilde{\boldsymbol{\alpha}}, \beta_0). \end{aligned}$$

Define the $n \times 1$ vector

$$\boldsymbol{\eta} := \sum_{k=1}^n (\hat{\alpha}_k(\beta_0) - \alpha_{k0}) \frac{\partial \mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}}, \beta_0)}{\partial \alpha_k} [\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0],$$

its ith element η_i can be calculated as

$$\begin{aligned}\eta_i &= \Lambda_{ii} \cdot (\hat{\alpha}_i(\beta_0) - \alpha_{i0}) + \sum_{j \neq i} \Lambda_{ij} \cdot (\hat{\alpha}_j(\beta_0) - \alpha_{j0}) \\ &= -\sum_{j \neq i} f_{ij}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{ji}(\tilde{\boldsymbol{\alpha}}, \beta_0)(\hat{\alpha}_i(\beta_0) - \alpha_{i0})^2 \\ &\quad - \sum_{j \neq i} f_{ij}(\tilde{\boldsymbol{\alpha}}, \beta_0) f_{ji}(\tilde{\boldsymbol{\alpha}}, \beta_0)(\hat{\alpha}_i(\beta_0) - \alpha_{i0})(\hat{\alpha}_j(\beta_0) - \alpha_{j0}) \\ &\quad - \sum_{j \neq i} f_{ji}^{(1)}(\tilde{\boldsymbol{\alpha}}, \beta_0) F_{ij}(\tilde{\boldsymbol{\alpha}}, \beta_0)(\hat{\alpha}_j(\beta_0) - \alpha_{j0})^2.\end{aligned}$$

By Assumption 3, $F_{ij}, f_{ij}, f_{ij}^{(1)}$ are all bounded by some constants. So, we have

$$|\eta_i| \leq 3(n-1) \cdot O(1) \cdot \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_\infty^2,$$

uniformly for $i = 1, \ldots, n$, which implies that

$$\|\boldsymbol{\eta}\|_\infty \leq 3(n-1) \cdot O(1) \cdot O\left(\frac{\log n}{n}\right) = O(\log n)$$

because $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_\infty = O(\sqrt{(\log n)/n})$. By the triangle inequality,

$$\|\mathbf{J}_{11}^{-1}\boldsymbol{\eta}\|_\infty = \|\mathbf{T}\boldsymbol{\eta} + (\mathbf{J}_{11}^{-1} - \mathbf{T})\boldsymbol{\eta}\|_\infty \leq (\|\mathbf{T}\|_\infty + \|\mathbf{J}_{11}^{-1} - \mathbf{T}\|_\infty)\|\boldsymbol{\eta}\|_\infty = O\left(\frac{\log n}{n}\right).$$

Finally, from (62), we have

$$\left\| \sqrt{n}(\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0) + \left(\frac{\mathbf{J}_{11}}{n}\right)^{-1} \frac{\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)}{\sqrt{n}} \right\|_\infty = \left\| \frac{1}{2}\sqrt{n}\mathbf{J}_{11}^{-1}\boldsymbol{\eta} \right\|_\infty = O\left(\frac{\log n}{\sqrt{n}}\right).$$

This completes the proof. □

B Proofs of Main Results

In this Appendix, we present proofs of Theorems 1-3, Corollary 1 and Theorems 4-8.

B.1 Proof of Theorem 1

Before the proof of this theorem, we state a different version of Lemma 2.1 of Chatterjee, Diaconis, and Sly (2011). Given $\delta > 0$, we say an $n \times n$ matrix $\mathbf{A}$ belongs to the class $\mathcal{G}_n(\delta)$ if $\|\mathbf{A}\|_1 \leq 1$, and for each $1 \leq i \neq j \leq n$,

$$\mathbf{A}_{ii} \leq \delta, \text{ and } \mathbf{A}_{ij} \geq -\frac{\delta}{n-1}.$$

Lemma 6. *If* $\mathbf{A}, \mathbf{B} \in \mathcal{G}_n(\delta)$*, we have*

$$\|\mathbf{AB}\|_1 \le 1 - \frac{2(n-2)}{n-1}\delta^2.$$

Proof. This is equivalent to proving if $A, B \in \mathcal{G}_n(\delta)$, then

$$\|\mathbf{B}^\top \mathbf{A}^\top\|_\infty \le 1 - \frac{2(n-2)}{n-1}\delta^2,$$

which is a direct application of Lemma 2.1 of Chatterjee, Diaconis, and Sly (2011). □

We prove Theorem 1 based on this lemma.

Proof of Theorem 1. First, we suppose there exists a solution to $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = 0$. Let $\mathbf{G}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})$ be the matrix whose (i,j)th element is

$$[\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})]_{ij} = \int_0^1 \frac{\partial r_i}{\partial \alpha_j}(t\boldsymbol{\alpha} + (1-t)\widehat{\boldsymbol{\alpha}})dt.$$

Then by an integral type of mean value theorem, we have

$$\mathbf{r}(\boldsymbol{\alpha}) - \mathbf{r}(\widehat{\boldsymbol{\alpha}}) = \mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}).$$

Notice that for $i \ne j$, $\partial r_j / \partial \alpha_i = -(n-1) f_{ij}(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta) < 0$; while for each i, $\partial r_i / \partial \alpha_i = 1 - (n-1)\sum_{j \ne i} f_{ij}(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta) > 0$. Moreover, for each i,

$$\sum_{j=1}^n \left|\frac{\partial r_j}{\partial \alpha_i}\right| = \frac{\partial r_i}{\partial \alpha_i} - \sum_{j \ne i} \frac{\partial r_j}{\partial \alpha_i} \equiv 1.$$

For each i and any $\boldsymbol{\alpha}$, this proves $\sum_{j=1}^n |[\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})]_{ji}| = 1$, i.e., $\|\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})\|_1 = 1$. By Assumptions 2 and 3, we know that $f_{ij}(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta) \in [c_1 c_2, (1-c_1)(1-c_2)]$, this demonstrates that

$$\frac{\partial r_j}{\partial \alpha_i} \le -\frac{c_1 c_2}{n-1}, \quad \text{and} \quad \frac{\partial r_i}{\partial \alpha_i} \ge c_1 + c_2 - c_1 c_2 \ge c_1 c_2,$$

where the last inequality is because that $c_1 + c_2 \ge 2\sqrt{c_1 c_2} \ge 2c_2 c_2$ provided that $c_1, c_2 \le 1/2$. So if we choose $\delta_1 = c_1 c_2$ $(\le 1/4)$, it is obvious that $[\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})]_{ii} < \delta_1$ and $[\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})]_{ij} > -\frac{\delta_1}{n-1}$. Therefore, we have proved $\mathbf{G}^\circ(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}}) \in \mathcal{G}_n(\delta)$.

By the updating algorithm specified in Section 3.1, after every two updates, we have

$$\begin{aligned}
\|\boldsymbol{\alpha}^{k+2}(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1 &= \|\mathbf{r}(\mathbf{r}(\boldsymbol{\alpha}^k(\beta))) - \mathbf{r}(\mathbf{r}(\widehat{\boldsymbol{\alpha}}(\beta)))\|_1 \\
&= \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)) - \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \\
&\le \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))\mathbf{G}^\circ(\boldsymbol{\alpha}^k(\beta), \widehat{\boldsymbol{\alpha}}(\beta))(\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \\
&\le \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))\mathbf{G}^\circ(\boldsymbol{\alpha}^k(\beta), \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \|\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1
\end{aligned}$$

$$\leq \left(1 - \frac{2(n-2)}{n-1}\delta_1^2\right) \|\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1,$$

where the first equality holds by the fact that $\widehat{\boldsymbol{\alpha}}(\beta) = \mathbf{r}(\widehat{\boldsymbol{\alpha}}(\beta))$ which implies $\widehat{\boldsymbol{\alpha}}(\beta)$ is the fixed point of the updating function, and the last inequality holds by Lemma 6. We write $\delta := 1 - \frac{2(n-2)}{n-1}\delta_1^2$, the second inequality of Theorem 1 is proved.

The proof of the first inequality is the same as above, thereby omitted for conciseness.

By this result, $\mathbf{r}(\boldsymbol{\alpha})$ is a contraction mapping for $(\boldsymbol{\alpha}, \beta) \in \mathbb{A} \times \mathbb{B}$. So, if there exists a solution $\widehat{\boldsymbol{\alpha}}(\beta) \in \mathbb{A}$, then the solution is unique. Now we prove of the existence of the solution, where the main technique is adapted from Yan, Qin, and Wang (2016b) and Yan, Jiang, Fienberg, and Leng (2019). Define a sequence of Newton iterations $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} - \mathbf{J}_{11}^{-1}(\boldsymbol{\alpha}^{(k)}, \beta)\mathbf{m}_1(\boldsymbol{\alpha}^{(k)}, \beta)$, and choose the initial value as $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}_0$. Following Proposition A.1 of Yan, Qin, and Wang (2016b), in a convex subset $\mathbb{D} \subset \mathbb{A}$ that contains $\boldsymbol{\alpha}_0$ it is sufficient to establish three facts: (1) $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ is Lipschitz continuous with Lipschitz constant of order $O(n)$, (2) $\left\|\mathbf{J}_{11}^{-1}(\boldsymbol{\alpha}_0, \beta)\right\|_\infty = O(n^{-1})$, and (3) $\left\|\mathbf{J}_{11}^{-1}(\boldsymbol{\alpha}_0, \beta)\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta)\right\|_\infty = O(\|\beta - \beta_0\|_2)$.

For the first fact, we calculate the derivative of $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ with respect to $\boldsymbol{\alpha}$:

$$\frac{\partial \mathbf{J}_{11,ij}}{\partial \alpha_k} = \begin{cases} -\sum_{j\neq i} \frac{\partial^2 p_{ij}}{\partial \alpha_i^2} & i = j = k, \\ -\frac{\partial^2 p_{ij}}{\partial \alpha_i} & i \neq j, k = i, \\ -\frac{\partial^2 p_{ij}}{\partial \alpha_j} & i \neq j, k = j, \\ 0 & \text{ortherise.} \end{cases}$$

which implies that $\max_i \sum_{j,k} \left| \int_0^1 \frac{\partial \mathbf{J}_{11,ij}(t\boldsymbol{\alpha}_1 + (1-t)\boldsymbol{\alpha}_2)}{\partial \alpha_k} \right| = O(n)$. Hence $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ is Lipschitz continuous with Lipschitz constant $O(n)$. The second fact is a direct application of the inverse approximation Lemma 1. Finally, the third result follows from

$$\begin{aligned} &\left\|[\mathbf{J}_{11}(\boldsymbol{\alpha}_0, \beta)]^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta)\right\|_\infty \\ \leq\ &\left\|[\mathbf{J}_{11}(\boldsymbol{\alpha}_0, \beta)]^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\right\|_\infty + \left\|[\mathbf{J}_{11}(\boldsymbol{\alpha}_0, \beta)]^{-1}\left[\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta)\right]\right\|_\infty \\ \leq\ &o_p(1) + O\left(\|\beta - \beta_0\|_2\right) \\ =\ &O_p\left(\|\beta - \beta_0\|_2\right), \end{aligned}$$

where the first inequality holds by the triangular inequality and the second inequality is true by Lemma 3 and the Lipschitz continuity of $F(\cdot)$ under Assumption 3. Then, by an application of Proposition A.1 of Yan, Qin, and Wang (2016b), we have $\lim_{k\to\infty} \boldsymbol{\alpha}^{(k)}$ exists and the limit equals to $\widehat{\boldsymbol{\alpha}}(\beta)$ if $\|\beta - \beta_0\|_2 < c$ for some constant $c > 0$. □

B.2 Proof of Theorem 2

Recall the concentrated moment equation and its population counterpart are

$$S_n(\beta) := \binom{n}{2}^{-1} m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta) \text{ and } \bar{S}_n(\beta) := \binom{n}{2}^{-1} \mathbb{E}[m_2(\boldsymbol{\alpha}(\beta), \beta)|\mathbf{x}, \boldsymbol{\alpha}_0],$$

respectively, where $\boldsymbol{\alpha}(\beta)$ is the unique solution to $\mathbb{E}[\mathbf{m}_1(\boldsymbol{\alpha}, \beta)|\mathbf{x}, \boldsymbol{\alpha}_0] = 0$. By Assumption 4, $\hat{\beta}$ and β_0 are unique solutions of $S_n(\beta) = 0$ and $\bar{S}_n(\beta) = 0$, respectively.

First, we present a lemma to bound the difference between $S_n(\beta)$ and $\bar{S}_n(\beta)$ for $\beta \in \mathbb{B}$.

Lemma 7. *Under Assumptions 1–4, we have*

$$\sup_{\beta \in \mathbb{B}} \left\| S_n(\beta) - \bar{S}_n(\beta) \right\|_2 \xrightarrow{p} 0.$$

Proof. By the definitions of $\widehat{\boldsymbol{\alpha}}(\beta)$ and $\boldsymbol{\alpha}(\beta)$, we have $\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta), \beta) = 0$ and $\mathbb{E}[\mathbf{m}_1(\boldsymbol{\alpha}(\beta), \beta)|\mathbf{x}, \boldsymbol{\alpha}_0] = 0$. Thus,

$$\sum_{j \neq i} (y_{ij} - p_{ij}) - (p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta), \beta) - p_{ij}(\boldsymbol{\alpha}(\beta), \beta)) = 0, \qquad i = 1, \dots, n.$$

By an integral type mean-value theorem, we have

$$\widehat{\boldsymbol{\alpha}}(\beta) - \boldsymbol{\alpha}(\beta) = -\left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0),$$

and recall that $\mathbf{J}_{21}(\boldsymbol{\alpha}, \beta) := \frac{\partial m_2(\boldsymbol{\alpha}, \beta)}{\partial \boldsymbol{\alpha}^\top}$,

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j>i} \left(p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta), \beta) - p_{ij}(\boldsymbol{\alpha}(\beta), \beta)\right) x_{ij} =& m_2(\boldsymbol{\alpha}(\beta), \beta) - m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta) \\
=& -\mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \left(\widehat{\boldsymbol{\alpha}}(\beta) - \boldsymbol{\alpha}(\beta)\right) \\
=& \mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0).
\end{aligned}$$

Straightforward algebra then shows

$$\begin{aligned}
S_n(\beta) - \bar{S}_n(\beta) =& \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} \left[y_{ij} - p_{ij} - (p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta), \beta) - p_{ij}(\boldsymbol{\alpha}(\beta), \beta))\right] x_{ij} \\
=& \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} (y_{ij} - p_{ij}) x_{ij} - \binom{n}{2}^{-1} \mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) \\
=& \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} (y_{ij} - p_{ij}) x_{ij} - \binom{n}{2}^{-1} \mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) \\
&+ \binom{n}{2}^{-1} \mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) \left[\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) - \mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))^{-1}\right] \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) \\
=:& R_1 + R_2 + R_3,
\end{aligned}$$

where $\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) = [\mathrm{diag}(\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)))]^{-1}$ is the analytic approximation for $\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))$ by Lemma 1.

For R_1, by Lemma 4 and the fact that x_{ij} is bounded, it is of order $O_p(\sqrt{(\log N)/N})$. For R_2, notice that $\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))$ is diagonal with $[\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))]_{ii} = O(n^{-1})$ and each element in $\mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))$ is of order $O(n)$ uniformly. Thus, by Lemma 3 we have

$$\|R_2\|_\infty = O_p\left(n^{-2} \cdot n \cdot n \cdot \sqrt{\frac{\log n}{n}}\right) = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

Finally for R_3, we use Lemma 1 to bound it as

$$\begin{aligned}\|R_3\|_\infty \leq & \binom{n}{2}^{-1} \cdot n^2 \cdot \|\mathbf{J}_{21}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))\|_{\max} \\ & \cdot \|\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta)) - \mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta), \boldsymbol{\alpha}(\beta))^{-1}\|_{\max} \cdot \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty \\ = & O_p\left(n^{-2} \cdot n^2 \cdot n \cdot n^{-2} \cdot n \cdot \sqrt{\frac{\log n}{n}}\right) = O_p\left(\sqrt{\frac{\log n}{n}}\right).\end{aligned}$$

Further notice that these bounds hold uniformly in β, we have completed the proof. □

Proof of Theorem 2. By the definitions of $\hat{\beta}$ and β_0, we have $S_n(\hat{\beta}) = 0$ and $\bar{S}_n(\beta_0) = 0$, combine this fact with Lemma 7, we have

$$\left\|\bar{S}_n(\hat{\beta})\right\|_2 = \left\|\bar{S}_n(\hat{\beta}) - S_n(\hat{\beta})\right\|_2 \leq \sup_{\beta \in \mathbb{B}} \left\|S_n(\beta) - \bar{S}_n(\beta)\right\|_2 \xrightarrow{p} 0. \tag{63}$$

By Assumption 4, fix $\delta > 0$, there exists an $\epsilon > 0$ such that $\|\beta - \beta_0\|_2 \geq \delta$ implies $\left\|\bar{S}_n(\beta)\right\|_2 \geq \epsilon$, hence

$$\Pr\left(\left\|\hat{\beta} - \beta_0\right\|_2 \geq \delta\right) \leq \Pr\left(\left\|\bar{S}_n(\hat{\beta})\right\|_2 \geq \epsilon\right) \leq \Pr\left(\sup_{\beta \in \mathbb{B}} \left\|\bar{S}_n(\beta)\right\|_2 \geq \epsilon\right) \to 0$$

by (63).

We turn to the proof of the uniform consistency of $\widehat{\boldsymbol{\alpha}}$. By the integral type mean-value theorem, we have

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = & - [\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) \\ = & - [\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) - [\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)]^{-1} \left(\mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\right)\end{aligned} \tag{64}$$

Following the proof of Lemma 5, we have $\left\|[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)]^{-1}\right\|_\infty = O\left(n^{-1}\right)$ and $\|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty = O_p\left(\sqrt{n \log n}\right)$, hence $\left\|[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\right\|_\infty \xrightarrow{p} 0$. Thus, we only need to show that $O(n^{-1}) \cdot \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty \xrightarrow{p} 0$. Notice that

$$\|\mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_\infty$$

$$
\begin{aligned}
&= \max_{1\le i\le n} \left| \sum_{j\neq i} \left[p_{ij}(\boldsymbol{\alpha}_0, \hat{\beta}) - p_{ij}(\boldsymbol{\alpha}_0, \beta_0) \right] \right| \\
&\le \max_{1\le i\le n} \left\| \sum_{j\neq i} \left[f_{ij}(\boldsymbol{\alpha}_0, \bar{\beta}) F_{ji}(\boldsymbol{\alpha}_0, \bar{\beta}) + F_{ij}(\boldsymbol{\alpha}_0, \bar{\beta}) f_{ji}(\boldsymbol{\alpha}_0, \bar{\beta}) \right] x_{ij} \right\|_2 \times \|\hat{\beta} - \beta_0\|_2 \\
&= O(n) \times o_p(1) = o_p(n),
\end{aligned}
$$

where we use a Taylor expansion of $p_{ij}(\boldsymbol{\alpha}_0, \beta)$ around β_0 ($\bar{\beta}$ is the mean value which may vary with i) and the fact that f_{ij} and F_{ij} are bounded by Assumption 3. The proof is completed. □

B.3 Proof of Theorem 3

Before we prove Theorem 3, we characterize the limit of the concentrated Jacobian matrices. By Lemma 1,

$$
\begin{aligned}
N^{-1}\mathrm{J}_n(\beta) &= N^{-1}\left[\mathrm{J}_{22}(\widehat{\boldsymbol{\alpha}}(\beta),\beta) - \mathbf{J}_{21}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)^{-1}\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\right] \\
&= N^{-1}\mathrm{J}_{22}(\widehat{\boldsymbol{\alpha}}(\beta),\beta) - N^{-1}\mathbf{J}_{21}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\mathbf{T}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\beta),\beta) \\
&\quad - N^{-1}\mathbf{J}_{21}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\left(\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)^{-1} - \mathbf{T}(\widehat{\boldsymbol{\alpha}}(\beta),\beta)\right)\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\beta),\beta) = O_p(1).
\end{aligned}
$$

Since $\hat{\beta} \xrightarrow{p} \beta_0$ and $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_\infty = o_p(1)$ by Theorem 2, we have

$$
N^{-1}\mathrm{J}_n(\bar{\beta}) \xrightarrow{p} \mathrm{J}_0 := \mathrm{plim}_{n\to\infty} N^{-1}\mathrm{J}_n(\beta_0) \tag{65}
$$

for any $\bar{\beta}$ lies between $\hat{\beta}$ and β_0. The existence of J_0 is guaranteed by the identification Assumption 3.

Now, we turn to the proof of asymptotic normality of our moment estimators.

Proof of Theorem 3. By a first-order Taylor expansion of $m_n(\hat{\beta}) = m_2(\widehat{\boldsymbol{\alpha}}(\hat{\beta}), \hat{\beta})$ around β_0, we have

$$
m_n(\hat{\beta}) - m_n(\beta_0) = \mathrm{J}_n(\bar{\beta})(\hat{\beta} - \beta_0),
$$

where $\bar{\beta}$ is a mean-value between $\hat{\beta}$ and β_0. By $m_n(\hat{\beta}) = 0$, we obtain

$$
\begin{aligned}
\sqrt{N}(\hat{\beta} - \beta_0) &= -\left[N^{-1}\mathrm{J}_n(\bar{\beta})\right]^{-1}\frac{1}{\sqrt{N}} m_2(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) \\
&= -\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\left[y_{ij} - p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0)\right] x_{ij}\right\} + o_p(1)
\end{aligned} \tag{66}
$$

in view of (65). Thus, Note that we cannot directly apply standard central limit theorem (CLT) to the term in the curly bracket of (66) because of the existence of $\widehat{\boldsymbol{\alpha}}(\beta_0)$. By a third-order Taylor expansion for $\widehat{\boldsymbol{\alpha}}(\beta_0)$ around the true value $\boldsymbol{\alpha}_0$, we have

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\left[y_{ij}-p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right]x_{ij}\\
=&\ \frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0)+\frac{1}{\sqrt{N}}\mathbf{J}_{21}(\boldsymbol{\alpha}_0,\beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\\
&+\frac{1}{2}\left\{-\frac{1}{\sqrt{N}}\sum_{k=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_k\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}_n(\beta_0)-\boldsymbol{\alpha}_0]x_{ij}\right\}\\
&+\frac{1}{6}\left\{-\frac{1}{\sqrt{N}}\sum_{k=1}^{n}\sum_{l=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}][\hat{\alpha}_l(\beta_0)-\alpha_{l0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}}_n,\beta_0)}{\partial\alpha_k\partial\alpha_l\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}_n(\beta_0)-\boldsymbol{\alpha}_{n0}]x_{ij}\right\}\\
=:&\ (I)+(II)+(III)+(IV).
\end{aligned}
\tag{67}
$$

Let's handle the last term (IV) first. Since $p_{ij}(\boldsymbol{\alpha},\beta)$ only contains α_i and α_j, the last term (IV) equals to

$$
(IV)=-\frac{1}{6\sqrt{N}}\sum_{i=1}^{N}\sum_{j>i}\left[\begin{array}{r}(\hat{\alpha}_i-\alpha_{i0})^3\dfrac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_i^3}+(\hat{\alpha}_j-\alpha_{j0})^3\dfrac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_j^3}\\ +3\left((\hat{\alpha}_i-\alpha_{i0})^2(\hat{\alpha}_j-\alpha_{j0})\dfrac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_i^2\partial\alpha_j}\right)\\ +3\left((\hat{\alpha}_i-\alpha_{i0})(\hat{\alpha}_j-\alpha_{j0})^2\dfrac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_j^2\partial\alpha_i}\right)\end{array}\right]x_{ij}.
$$

By Lemma 5, $\sup_i|\hat{\alpha}_i(\beta_0)-\alpha_{i0}|=O_p(\sqrt{(\log n)/n})$. Notice that $\frac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}}_n,\beta_0)}{\partial\alpha_i^2\partial\alpha_j}x_{ij}$ is bounded under Assumptions 2 and 3. Thus, we have

$$
(IV)=O_p\left(\frac{1}{\sqrt{N}}\cdot\frac{n(n-1)}{2}\cdot\left(\frac{\log n}{n}\right)^{3/2}\right)=O_p\left(\frac{(\log n)^{3/2}}{n^{1/2}}\right)=o_p(1).
$$

For (III), we substitute the asymptotic linear approximation of $\sqrt{n}[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]$ into it. After some lengthy algebra, its kth entry, which involves many terms of third order derivatives, equals to

$$
-\frac{1}{2\sqrt{N}}\mathrm{Tr}\left[\mathbf{J}_{11}^{-1}\mathbf{V}_{11}\left(\mathbf{J}_{11}^{-1}\right)^\top\mathbf{R}_k\right]+o_p(1),
$$

where elements of $\mathbf{R}_k$ for $k=1,\ldots,K$ are

$$
\begin{aligned}
(\mathbf{R}_k)_{ij}&=\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_i\partial\alpha_j}x_{ij,k},\ 1<i\neq j<n,\\
(\mathbf{R}_k)_{ii}&=\sum_{j\neq i}\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial^2\alpha_i}x_{ij,k},\ i=1,\ldots,n.
\end{aligned}
\tag{68}
$$

Recall the definition of B_{k0} in (13), we find that $(III) = -B_0 + o_p(1)$.

We directly substitute the rest of terms, (I) and (II), into (66) and obtain,

$$
\begin{aligned}
&\sqrt{N}(\hat{\beta}-\beta_0) - \mathrm{J}_0^{-1}B_0 \\
=&-\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0) + \frac{1}{\sqrt{N}}\mathbf{J}_{21}(\boldsymbol{\alpha}_0,\beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\right\} + o_p(1) \\
=&-\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0) - \frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\} + o_p(1).
\end{aligned}
$$

We check the Lindeberg condition to apply CLT to the first two terms of the last equation. Define

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\xi_{ij} :=& -\mathrm{J}_0^{-1}\frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{n}\sum_{j>i}(y_{ij}-p_{ij})x_{ij} - \frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\} \\
=& -\mathrm{J}_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}(y_{ij}-p_{ij})\tilde{x}_{ij},
\end{aligned} \tag{69}
$$

where $\tilde{x}_{ij}$ absorbs x_{ij} and coefficient of $y_{ij}-p_{ij}$ from the contribution of $\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)$. Hence (69) is a weighted sum of $y_{ij}-p_{ij}$ by the definitions of each component of $\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)$ at the beginning of Section A. By Assumptions 2 and 3, we have $\|\tilde{x}_{ij}\|_\infty < \infty$. For $y_{ij}-p_{ij}$, they are independent across dyads (i,j), $1 \le i < j \le n$ conditional on $(\mathbf{x},\boldsymbol{\alpha})$ and are bounded by $[-1,1]$. Thus, the Lindeberg condition is satisfied.

Further notice that the variance of $m_2(\boldsymbol{\alpha}_0,\beta_0) - \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)$ is

$$
\mathrm{V}_{22} + \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{11}(\mathbf{J}_{11}^{-1})^\top\mathbf{J}_{21}^\top - \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{12} - (\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{V}_{12})^\top.
$$

By the Lindeberg-Feller CLT, we have

$$
\sqrt{N}(\hat{\beta}-\beta_0) - \mathrm{J}_0^{-1}B_0 \xrightarrow{d} \mathcal{N}(0,\Omega_0),
$$

where Ω_0 is defined in (14). □

B.4 Proof of Asymptotic Distribution of JMM SJ Estimator

Corollary 1. *If Assumptions 1–4 hold, then*

$$
\sqrt{N}(\hat{\beta}_{\mathrm{SJ}}-\beta_0) \xrightarrow{d} \mathcal{N}(0,2\Omega_0). \tag{70}
$$

Proof of Corollary 1 . Using the asymptotic representation for $\hat{\beta}$ provided in the proof of Theorem 3, we have

$$
\hat{\beta}-\beta_0 = \frac{\mathrm{J}_0^{-1}B_0}{\sqrt{N}} + \frac{\sum_{(i,j)\in\mathcal{I}_n\times\mathcal{I}_n;j>i}\xi_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{N} + o_p\left(\frac{1}{\sqrt{N}}\right), \tag{71}
$$

where ξ_{ij}'s are independent random variables with $\sum_{(i,j)\in\mathcal{I}_n\times\mathcal{I}_n;j>i}\xi_{ij}(\boldsymbol{\alpha}_0,\beta_0)/\sqrt{N} \xrightarrow{d} \mathcal{N}(0,\Omega_0)$. Without loss of generality, suppose n is even and $\mathcal{I}_{1,n} = \{1,\ldots,n/2\}$, $\mathcal{I}_{2,n} = \{n/2+1,\ldots,n\}$. Then, for $\hat{\beta}_1$ and $\hat{\beta}_2$, we have

$$\begin{aligned}\hat{\beta}_1-\beta_0 &= \frac{\mathrm{J}_0^{-1}B_0}{\sqrt{N/4}} + \frac{\sum_{(i,j)\in\mathcal{I}_{1,n}\times\mathcal{I}_{1,n};j>i}\xi_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{N/4} + o_p\left(\frac{1}{\sqrt{N}}\right), \quad \text{and} \\ \hat{\beta}_2-\beta_0 &= \frac{\mathrm{J}_0^{-1}B_0}{\sqrt{N/4}} + \frac{\sum_{(i,j)\in\mathcal{I}_{2,n}\times\mathcal{I}_{2,n};j>i}\xi_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{N/4} + o_p\left(\frac{1}{\sqrt{N}}\right),\end{aligned} \tag{72}$$

where each summation has $N/4$ independent random variables. Combining (71) and (72), we have

$$\sqrt{N}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\beta}_1-\beta_0\\ \hat{\beta}_2-\beta_0\end{pmatrix} - \begin{pmatrix}\mathrm{J}_0^{-1}B_0\\ 2\mathrm{J}_0^{-1}B_0\\ 2\mathrm{J}_0^{-1}B_0\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix}\Omega_0 & \Omega_0 & \Omega_0\\ \Omega_0 & 4\Omega_0 & 0\\ \Omega_0 & 0 & 4\Omega_0\end{pmatrix}\right).$$

Hence

$$\sqrt{N}(\hat{\beta}_{\mathrm{SJ}}-\beta_0) = 2\sqrt{N}(\hat{\beta}-\beta_0) - \frac{1}{2}\sqrt{N}(\hat{\beta}_1-\beta_0) - \frac{1}{2}\sqrt{N}(\hat{\beta}_2-\beta_0) \xrightarrow{d} \mathcal{N}\left(0, 2\Omega_0\right),$$

because the linear transformation of jointly normal vectors keeps normality. This completes the proof. □

B.5 Proof of Theorem 4

First, we characterize the partial derivatives of $s_n(\boldsymbol{\alpha},\beta)$ with respect to $\boldsymbol{\alpha}$ and β. We rewrite $s_n(\boldsymbol{\alpha},\beta)$ as

$$s_n(\boldsymbol{\alpha},\beta) = s_2(\boldsymbol{\alpha},\beta) - \sum_{i=1}^{n} s_{1i}(\boldsymbol{\alpha},\beta)w_i(\boldsymbol{\alpha},\beta), \tag{73}$$

where $w_i(\boldsymbol{\alpha},\beta)$ is the ith column of $\mathbf{I}_{12}(\boldsymbol{\alpha},\beta)^\top\mathbf{I}_{11}(\boldsymbol{\alpha},\beta)^{-1}$. Taking derivatives, we have

$$\nabla_{\boldsymbol{\alpha}^\top}s_n(\boldsymbol{\alpha},\beta) = \mathbf{H}_{12}(\boldsymbol{\alpha},\beta)^\top - \mathbf{I}_{12}(\boldsymbol{\alpha},\beta)^\top\mathbf{I}_{11}(\boldsymbol{\alpha},\beta)^{-1}\mathbf{H}_{11}(\boldsymbol{\alpha},\beta) - \sum_{i=1}^{n} s_{1i}(\boldsymbol{\alpha},\beta)\frac{\partial w_i(\boldsymbol{\alpha},\beta)}{\partial\boldsymbol{\alpha}^\top},$$

and

$$\nabla_{\beta^\top}s_n(\boldsymbol{\alpha},\beta) = \mathrm{H}_{22}(\boldsymbol{\alpha},\beta) - \mathbf{I}_{12}(\boldsymbol{\alpha},\beta)^\top\mathbf{I}_{11}(\boldsymbol{\alpha},\beta)^{-1}\mathbf{H}_{12}(\boldsymbol{\alpha},\beta) - \sum_{i=1}^{n} s_{1i}(\boldsymbol{\alpha},\beta)\frac{\partial w_i(\boldsymbol{\alpha},\beta)}{\partial\beta^\top}.$$

We prove the following lemma.

Lemma 8. *Under Assumptions 1–5, for any $\tilde{\beta}$ such that $\|\tilde{\beta}-\beta_0\|_2 = O_p(N^{-1/2})$, we have*

$$\frac{1}{\sqrt{N}}\nabla_{\boldsymbol{\alpha}^\top}s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\left[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\right] \xrightarrow{p} b_0, \tag{74}$$

and

$$\frac{1}{N}\nabla_{\beta^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial \widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial \beta^\top} + \mathrm{I}_0 \xrightarrow{p} 0, \tag{75}$$

where $\bar{\beta}$ *lies in the segment between* $\tilde{\beta}$ *and* β_0*;* b_0 *is a* $K \times 1$ *vector of bias terms with its* k*th element equal to* $b_{k0} = \lim_{n\to\infty} \frac{1}{\sqrt{N}} Tr[\mathbf{J}_{11}^{-1} Cov(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k]$.

Proof. By the definitions of $\mathbf{H}_{12}(\boldsymbol{\alpha},\beta)$ and $\mathbf{I}_{12}(\boldsymbol{\alpha},\beta)$, we have

$$\mathbf{H}_{12}(\boldsymbol{\alpha},\beta) + \mathbf{I}_{12}(\boldsymbol{\alpha},\beta) = \begin{pmatrix} \sum_{j\neq 1}[y_{1j} - p_{1j}(\boldsymbol{\alpha},\beta)]z_{1j}(\boldsymbol{\alpha},\beta)x_{1j}^\top \\ \vdots \\ \sum_{j\neq n}[y_{nj} - p_{nj}(\boldsymbol{\alpha},\beta)]z_{nj}(\boldsymbol{\alpha},\beta)x_{nj}^\top \end{pmatrix},$$

where $z_{ij} = \frac{f_{ij}^{(1)}F_{ij}(1-p_{ij}) - f_{ij}^2(1-2p_{ij}) + F_{ij}^2 f_{ij} f_{ji}}{F_{ij}^2(1-p_{ij})^2}$ and we omit the dependence of z_{ij} on $\boldsymbol{\alpha},\beta$ for simplicity. Clearly, $N^{-1/2}\left\|\mathbf{H}_{12}(\boldsymbol{\alpha}_0,\beta_0) + \mathbf{I}_{12}(\boldsymbol{\alpha}_0,\beta_0)\right\|_{\max} = o_p(1)$ by Lemma 5. Hence, by continuous mapping theorem (CMT) and the fact that $\|\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0\|_\infty = o_p(1)$, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) + \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right\|_{\max} \\
&\leq \frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) - \mathbf{H}_{12}(\boldsymbol{\alpha}_0,\beta_0)\right\|_{\max} + \frac{1}{\sqrt{N}}\left\|\mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) - \mathbf{I}_{12}(\boldsymbol{\alpha}_0,\beta_0)\right\|_{\max} \\
&\quad + \frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\boldsymbol{\alpha}_0,\beta_0) + \mathbf{I}_{12}(\boldsymbol{\alpha}_0,\beta_0)\right\|_{\max} \\
&= o_p(1).
\end{aligned}$$

Similarly, we can obtain

$$\frac{1}{\sqrt{N}}\left\|\mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top \mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}[\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) + \mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)]\right\|_{\max} = o_p(1).$$

Combining these two bounds, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top \mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\beta_0)^{-1}\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right\|_{\max} \\
&\leq \frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) + \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\right\|_{\max} \\
&\quad + \frac{1}{\sqrt{N}}\left\|\mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top \mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}\left[\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) + \mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right]\right\|_{\max} \\
&= o_p(1).
\end{aligned} \tag{76}$$

By Lemma 5, we have $\left\|\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0 + \mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\|_\infty = o_p(n^{-1/2})$, which implies that for any deterministic vector $\|\mathbf{c}\|_2 = 1$,

$$\sqrt{n}\mathbf{c}^\top(\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0) = O_p(1). \tag{77}$$

Combining (76) and (77) yields

$$\frac{1}{\sqrt{N}}[\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)](\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0) = o_p(1). \tag{78}$$

Next, similar to the process of finding the bias term in the proof of Theorem 3, we have:

$$\begin{aligned}
&-\frac{1}{\sqrt{N}}\sum_{i=1}^n s_{1i}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\frac{\partial w_{ki}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)}{\partial\boldsymbol{\alpha}^\top}(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\\
&=\frac{1}{\sqrt{N}}\sum_{i=1}^n s_{1i}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\frac{\partial w_{ki}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)}{\partial\boldsymbol{\alpha}^\top}\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\boldsymbol{\alpha}_0)^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\\
&=\frac{1}{\sqrt{N}}\mathbf{s}_1(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\mathbf{W}_k(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\boldsymbol{\alpha}_0)^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\\
&=\frac{1}{\sqrt{N}}\mathbf{s}_1^\top\mathbf{W}_k\mathbf{J}_{11}^{-1}\mathbf{m}_1 + o_p(1)
\end{aligned}$$

where $[\mathbf{W}_k(\boldsymbol{\alpha},\beta)]_{ij} = \frac{\partial w_{ki}(\boldsymbol{\alpha},\beta)}{\partial\alpha_j}$ and the last equality holds because $\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\|_\infty = o_p(1)$ and by CMT.

Recall the definition of b_{k0} in (20), where the entries of the $n\times n$ covariance matrix $\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)$ can be specified as

$$\begin{aligned}
[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ij} =&\mathbb{E}\left[\left(\sum_{k\neq i}(y_{ik}-p_{ik})\right)\left(\sum_{k\neq j}\frac{f_{jk}(y_{jk}-p_{jk})}{F_{jk}(1-p_{jk})}\right)\right]\\
=&\frac{f_{ji}\mathrm{Var}(y_{ij})}{F_{ji}(1-p_{ij})} = f_{ji}F_{ij}, \quad 1\le i\neq j\le n,\\
[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ii} =&\mathbb{E}\left[\left(\sum_{k\neq i}(y_{ik}-p_{ik})\right)\left(\sum_{k\neq i}\frac{f_{ik}(y_{ik}-p_{ik})}{F_{ik}(1-p_{ik})}\right)\right]\\
=&\sum_{k\neq i}\frac{f_{ik}\mathrm{Var}(y_{ik})}{F_{ik}(1-p_{ik})} = \sum_{k\neq i} f_{ik}F_{ki}, \quad 1\le i\le n.
\end{aligned} \tag{79}$$

We show that $b_{k0}=O(1)$ for all k. Notice that $[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ij}\asymp 1$ and $[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ii}\asymp n$ uniformly by (79). By Assumption 5, $[\mathbf{W}_k(\boldsymbol{\alpha},\beta)]_{ij} = O(n^{-1})$ and $[\mathbf{W}_k(\boldsymbol{\alpha},\beta)]_{ii} = O(1)$ uniformly. By 1, $\mathbf{J}_{11}^{-1}$ can be approximated by the diagonal matrix $\mathbf{T}=[\mathrm{diag}(\mathbf{J}_{11})]^{-1}$ with $\|\mathbf{J}_{11}^{-1}-\mathbf{T}\|_{\max} = O(n^{-2})$. Thus, we have uniformly for all i,

$$\begin{aligned}
&\frac{1}{\sqrt{N}}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii}\\
=&\frac{1}{\sqrt{N}}[\mathbf{T}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii} + \frac{1}{\sqrt{N}}[(\mathbf{J}_{11}^{-1}-\mathbf{T})\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_\mathrm{k}]_{ii}\\
=&O(n^{-1}\cdot n^{-1})\cdot[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii} + O(n^{-1}\cdot n^{-2})\cdot[\mathbf{1}_n\mathbf{1}_n^\top\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii}\\
=&O(n^{-2})\cdot O(n) + O(n^{-3})\cdot O(n^2) = O(n^{-1}).
\end{aligned}$$

Taking the trace on both sides, we have

$$\frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k] = O(n^{-1})\cdot n = O(1).$$

Because b_{k0} is the limit of $\frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]$, it is also $O(1)$. Next, we have

$$\begin{aligned}
&-\frac{1}{\sqrt{N}}\mathbf{s}_1^\top\mathbf{W}_k\mathbf{J}_{11}^{-1}\mathbf{m}_1 = \frac{1}{\sqrt{N}}\mathrm{Tr}(\mathbf{J}_{11}^{-1}\mathbf{m}_1\mathbf{s}_1^\top\mathbf{W}_k)\\
&=\frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k] + \left\{\frac{1}{\sqrt{N}}\mathrm{Tr}(\mathbf{J}_{11}^{-1}\mathbf{m}_1\mathbf{s}_1^\top\mathbf{W}_k) - \frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]\right\} \qquad (80)\\
&= R_1 + R_2.
\end{aligned}$$

Notice that $R_1 \to b_{k0}$ by definition. By the law of large numbers for U-statistics, $R_2 \xrightarrow{p} 0$ under Assumption 5. By (78) and (80), we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\nabla_{\boldsymbol{\alpha}^\top}s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\\
=&\frac{1}{\sqrt{N}}[\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)](\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\\
&-\frac{1}{\sqrt{N}}\sum_{i=1}^n s_{1i}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\frac{\partial w_i(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)}{\partial\boldsymbol{\alpha}^\top}(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\xrightarrow{p} b_0,
\end{aligned}$$

which proves (74).

We now turn to prove (75). Similarly to the characterization of the probability limit of $N^{-1}\mathrm{I}_n(\widehat{\boldsymbol{\alpha}},\hat{\beta}_{\mathrm{SJ}})$, we have

$$\frac{1}{N}\left[\mathrm{H}_{22}(\boldsymbol{\alpha},\beta) - \mathbf{I}_{12}(\boldsymbol{\alpha},\beta)^\top\mathbf{I}_{11}(\boldsymbol{\alpha},\beta)^{-1}\mathbf{H}_{12}(\boldsymbol{\alpha},\beta)\right] = O_p(1).$$

Then, by the law of large numbers,

$$\frac{1}{N}\left[\mathrm{H}_{22}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^\top\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^{-1}\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\right] \xrightarrow{p} -\mathrm{I}_0.$$

By (64) with $\widehat{\beta}$ replaced by β_0, we have $\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\|_\infty = O_p(\sqrt{(\log n)/n})$. Combine this with Lemma 4 and we have

$$\left\|\frac{1}{N}\sum_{i=1}^n s_{1i}(\bar{\boldsymbol{\alpha}},\bar{\beta})\frac{\partial w_i(\bar{\boldsymbol{\alpha}},\bar{\beta})}{\partial\beta^\top}\right\|_\infty \le \left|\frac{1}{N}\sum_{i=1}^n s_{1i}(\bar{\boldsymbol{\alpha}},\bar{\beta})\right| \times \left\|\frac{\partial w_i(\bar{\boldsymbol{\alpha}},\bar{\beta})}{\partial\beta}\right\|_\infty = o_p(1).$$

Hence,

$$\frac{1}{N}\nabla_{\beta^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \mathrm{I}_0 = o_p(1). \qquad (81)$$

Finally, by (76) and the parallel steps to the proof of (74), we have

$$\frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial \widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial \beta^\top} = \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^{-1}\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) = o_p(1). \tag{82}$$

Combining (81) and (82) completes the proof of (75). □

With a direct application of this lemma, we prove the asymptotic normality of $\hat{\beta}_{\mathrm{OS}}$.

Proof of Theorem 4. First, by the definition of $\hat{\beta}_{\mathrm{OS}}$, we have

$$\hat{\beta}_{\mathrm{OS}} = \hat{\beta} + \mathrm{I}_n(\widehat{\boldsymbol{\alpha}},\hat{\beta})^{-1}s_n(\widehat{\boldsymbol{\alpha}},\hat{\beta})$$

with the joint moment estimator $\widehat{\boldsymbol{\alpha}} := \widehat{\boldsymbol{\alpha}}(\hat{\beta})$. Similarly to the proof of Theorem 3, by a first-order Taylor expansion of $s_n(\widehat{\boldsymbol{\alpha}},\hat{\beta})$ around β_0 followed by a first-order Taylor expansion of $s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)$ around $\boldsymbol{\alpha}_0$ as well as the fact that $N^{-1}\mathrm{I}_n(\widehat{\boldsymbol{\alpha}},\hat{\beta}) \xrightarrow{p} \mathrm{I}_0$, we have

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}_{\mathrm{OS}}-\beta_0)\\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\\
&\quad + \mathrm{I}_0^{-1}\left[\frac{1}{N}\nabla_{\beta^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial \widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0\right]\sqrt{N}(\hat{\beta}-\beta_0) + o_p(1)\\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}s_n(\boldsymbol{\alpha}_0,\beta_0) + \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}\nabla_{\boldsymbol{\alpha}^\top}s_n(\bar{\boldsymbol{\alpha}},\beta_0)(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\\
&\quad + \mathrm{I}_0^{-1}\left[\frac{1}{N}\nabla_{\beta^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial \widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0\right]\sqrt{N}(\hat{\beta}-\beta_0) + o_p(1).
\end{aligned} \tag{83}$$

By (74) of Lemma 8, we have $\frac{1}{\sqrt{N}}\nabla_{\boldsymbol{\alpha}^\top}s_n(\bar{\boldsymbol{\alpha}},\beta_0)(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0) \xrightarrow{p} b_0$. By (75) of Lemma 8,

$$\frac{1}{N}\nabla_{\beta^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top}s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial \widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0 = o_p(1).$$

Hence, using the result that $\sqrt{N}(\hat{\beta}-\beta_0) = O_p(1)$ by Theorem 3, we simplify (83) as

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{\mathrm{OS}}-\beta_0) &= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}s_n(\boldsymbol{\alpha}_0,\beta_0) + \mathrm{I}_0^{-1}b_0 + o_p(1)\\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}\sum_{i=1}^{n}\sum_{j>i}s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \mathrm{I}_0^{-1}b_0 + o_p(1),
\end{aligned} \tag{84}$$

where $s_{ij}(\boldsymbol{\alpha}_0,\beta_0)$ is dyad (i,j)'s contribution to the asymptotic representation. Then, by the Lindeberg-Feller CLT, as in the proof of Lemma 3, we have the stated asymptotic normality.
□

B.6 Proof of Theorem 5

Proof. By (84), we have,

$$\sqrt{N}(\hat{\beta}_{\mathrm{OS}} - \beta_0) = \mathrm{I}_0^{-1} b_0 + \frac{1}{\sqrt{N}} \mathrm{I}_0^{-1} \sum_{(i,j) \in \mathcal{I}_n \times \mathcal{I}_n; j > i} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + \mathcal{R}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0),$$

where $\mathcal{R}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0)$ is the residual term which has been shown to be $o_p(1)$ in the proof of Lemma 8. Thus, the one-step estimators based on sub-networks can be written as

$$\sqrt{N/4}(\hat{\beta}_{\mathrm{OS},1}^{(t)} - \beta_0) = \mathrm{I}_0^{-1} b_0 + \frac{1}{\sqrt{N/4}} \mathrm{I}_0^{-1} \sum_{(i,j) \in \mathcal{I}_{1,n}^{(t)} \times \mathcal{I}_{1,n}^{(t)}; j > i} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + \mathcal{R}(\mathbf{y}_1^{(t)}, \mathbf{x}_1^{(t)}, \boldsymbol{\alpha}_{0,1}^{(t)}),$$

$$\sqrt{N/4}(\hat{\beta}_{\mathrm{OS},2}^{(t)} - \beta_0) = \mathrm{I}_0^{-1} b_0 + \frac{1}{\sqrt{N/4}} \mathrm{I}_0^{-1} \sum_{(i,j) \in \mathcal{I}_{2,n}^{(t)} \times \mathcal{I}_{2,n}^{(t)}; j > i} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + \mathcal{R}(\mathbf{y}_2^{(t)}, \mathbf{x}_2^{(t)}, \boldsymbol{\alpha}_{0,2}^{(t)}),$$

where $\boldsymbol{\alpha}_{0,1}^{(t)}$ is the sub-vector of fixed effects indexed by $\mathcal{I}_{1,n}^{(t)}$ and the other notations are defined similarly. Hence, we have

$$\sqrt{N}(\hat{\beta}_{\mathrm{OS-SJ}}^{(t)} - \beta_0) = \frac{2}{\sqrt{N}} \mathrm{I}_0^{-1} \sum_{(i,j) \in \mathcal{I}_{1,n}^{(t)} \times \mathcal{I}_{2,n}^{(t)}} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + \mathcal{R}^{(t)}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0), \tag{85}$$

with $\mathcal{R}^{(t)}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0) := 2\mathcal{R}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0) - \left[\mathcal{R}(\mathbf{y}_1^{(t)}, \mathbf{x}_1^{(t)}, \boldsymbol{\alpha}_{0,1}^{(t)}) + \mathcal{R}(\mathbf{y}_2^{(t)}, \mathbf{x}_2^{(t)}, \boldsymbol{\alpha}_{0,2}^{(t)})\right]/2$, which is also $o_p(1)$. Notice that $\mathcal{R}^{(t)}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0)$ is a continuous and bounded function of its arguments, which means that $\max_{1 \le t \le T_n} \left\| \mathcal{R}^{(t)}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0) \right\|_2 \le \mathcal{R}^*(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0)$ for some function $\mathcal{R}^*(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0) = o_p(1)$. Hence, we have $\left\| T_n^{-1} \sum_{t=1}^{T_n} \mathcal{R}^{(t)} \right\|_2 \le \mathcal{R}^* = o_p(1)$. Then, taking average of (85) over all $1 \le t \le T_n$ yields

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{T_n} - \beta_0) &= \frac{2}{\sqrt{N}} \mathrm{I}_0^{-1} \frac{1}{T_n} \sum_{t=1}^{T_n} \sum_{(i,j) \in \mathcal{I}_{1,n}^{(t)} \times \mathcal{I}_{2,n}^{(t)}} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + \frac{1}{T_n} \sum_{t=1}^{T_n} \mathcal{R}^{(t)}(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}_0) \\
&= \frac{2}{\sqrt{N}} \mathrm{I}_0^{-1} \sum_{i=1}^{n} \sum_{j \ne i} \frac{\binom{n-2}{n/2-1}}{\binom{n}{n/2}} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + o_p(1) \\
&= \frac{1}{\sqrt{N}} \times \frac{n}{n-1} \mathrm{I}_0^{-1} \sum_{i=1}^{n} \sum_{j > i} s_{ij}(\boldsymbol{\alpha}_0, \beta_0) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1}),
\end{aligned}$$

where the second equality holds because for each $(i,j), i \ne j$, there are $\binom{n-2}{n/2-1}$ different splits among them. This proves the first result.

Conditional $\mathcal{F}_n$, random draws $\sqrt{N}\hat{\beta}_{\mathrm{OS-SJ}}^{(t)}$, $t = 1, \ldots, T_n'$ are independent and uniformly distributed over $\{\sqrt{N}\hat{\beta}_{\mathrm{OS-SJ}}^{(s)}\}_{s=1}^{T_n}$, which is a finite set for any fixed n. Hence, the law of large numbers implies that $\sqrt{N}(\hat{\beta}_{\mathrm{BG}} - \hat{\beta}_{T_n}) = o_p(1)$ as $T_n' \to \infty$, where the randomness comes from the independent draws. This proves the second result. Combing it with the first result gives

$$\sqrt{N}(\hat{\beta}_{\mathrm{BG}} - \beta_0) = \sqrt{N}(\hat{\beta}_{\mathrm{BG}} - \hat{\beta}_{T_n}) + \sqrt{N}(\hat{\beta}_{T_n} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1})$$

as $T_n' \to \infty$ and $n \to \infty$. This completes the proof. □

B.7 Proofs of Results in Section 4

Proof of Theorem 6. We decompose $\hat{\delta} - \delta_0$ as

$$\hat{\delta} - \delta_0 = \left(\hat{\delta} - \bar{\Delta}_n\right) + \left(\bar{\Delta}_n - \delta_0\right).$$

The second term is a U-statistics

$$\bar{\Delta}_n - \delta_0 = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} \left[\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0) - \mathbb{E}\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0)\right]$$

with the kernel $\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0) - \mathbb{E}\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0)$. So, if

$$\Sigma_\delta = \mathrm{Cov}(\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0), \Delta_{ik}(\alpha_{i0}, \alpha_{k0}, \beta_0))$$

exists, by Theorem 12.3 of Van der Vaart (2000), we have

$$\sqrt{n}(\bar{\Delta}_n - \delta_0) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_\delta). \tag{86}$$

Next, for the first term, notice that $\widehat{\boldsymbol{\alpha}} \equiv \widehat{\boldsymbol{\alpha}}(\hat{\beta})$ and we can decompose it as

$$\begin{aligned}
\sqrt{N}\left(\hat{\delta} - \bar{\Delta}_n\right) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{n} \sum_{j>i} \left[\Delta_{ij}(\hat{\alpha}_i(\hat{\beta}), \hat{\alpha}_j(\hat{\beta}), \hat{\beta}) - \Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0)\right] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{n} \sum_{j>i} \left[\Delta_{ij}(\hat{\alpha}_i(\hat{\beta}), \hat{\alpha}_j(\hat{\beta}), \hat{\beta}) - \Delta_{ij}(\hat{\alpha}_{i0}(\beta_0), \hat{\alpha}_{j0}(\beta_0), \beta_0)\right] \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{n} \sum_{j>i} \left[\Delta_{ij}(\hat{\alpha}_{i0}(\beta_0), \hat{\alpha}_{j0}(\beta_0), \beta_0) - \Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0)\right] \\
&:= U_1 + U_2,
\end{aligned}$$

where U_1 captures the variation from $\hat{\beta}$, and U_2 captures the variation from $\widehat{\boldsymbol{\alpha}}(\beta_0)$.

Let us define

$$\Delta_\beta(\boldsymbol{\alpha}, \beta) := \frac{1}{N} \sum_{i=1}^{n} \sum_{j>i} \frac{\partial \Delta_{ij}}{\partial \beta}(\alpha_i, \alpha_j, \beta), \qquad \Delta_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \beta) := \frac{1}{N} \begin{pmatrix} \sum_{j \neq 1} \frac{\partial \Delta_{1j}}{\partial \alpha_1}(\alpha_1, \alpha_j, \beta) \\ \vdots \\ \sum_{j \neq n} \frac{\partial \Delta_{nj}}{\partial \alpha_n}(\alpha_n, \alpha_j, \beta) \end{pmatrix}. \tag{87}$$

For U_1, a first-order Taylor's expansion around β_0 yields

$$\begin{aligned}
U_1 &= \frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{n}\sum_{j>i}\frac{\partial \Delta_{ij}}{\partial \beta^\top}(\hat{\alpha}_i(\bar{\beta}),\hat{\alpha}_j(\bar{\beta}),\bar{\beta}) + \sum_{i=1}^{n}\sum_{j\neq i}\frac{\partial \Delta_{ij}}{\partial \alpha_i}(\hat{\alpha}_i(\bar{\beta}),\hat{\alpha}_j(\bar{\beta}),\bar{\beta})\frac{\partial \hat{\alpha}_i}{\partial \beta^\top}(\bar{\beta})\right\}(\hat{\beta}-\beta_0)\\
&= \left\{\Delta_\beta(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^\top - \Delta_{\boldsymbol{\alpha}}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^\top \mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^{-1}\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\right\}\sqrt{N}(\hat{\beta}-\beta_0)\\
&= (\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\sqrt{N}(\hat{\beta}-\beta_0) + o_p(1), \qquad (88)
\end{aligned}$$

where $\bar{\beta}$ lies in the segment between $\hat{\beta}$ and β_0; the last equality uses the fact that $\bar{\beta} \xrightarrow{p} \beta_0$ and $\left\|\widehat{\boldsymbol{\alpha}}(\bar{\beta}) - \boldsymbol{\alpha}_0\right\|_\infty \xrightarrow{p} 0$.

For U_2, a third-order Taylor expansion yields

$$\begin{aligned}
U_2 = {} & \sqrt{N}\Delta_{\boldsymbol{\alpha}}^\top(\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0)\\
& + \frac{1}{2}\left\{\frac{1}{\sqrt{N}}\sum_{k=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^2 \Delta_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial \alpha_k \partial \boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\right\}\\
& + \frac{1}{6}\left\{\frac{1}{\sqrt{N}}\sum_{k=1}^{n}\sum_{l=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}][\hat{\alpha}_l(\beta_0)-\alpha_{l0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^3 \Delta_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial \alpha_k \partial \alpha_l \partial \boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_{n0}]\right\}.
\end{aligned} \qquad (89)$$

Similarly to the proof of Theorem 3, we can show that the second term of (89) converges in probability to a bias term B_α defined by (36) and

$$\begin{aligned}
(\mathbf{R}_k^\mu)_{ij} &= \frac{\partial^2 \Delta_{ij,k}(\boldsymbol{\alpha}_0,\beta_0)}{\partial \alpha_i \partial \alpha_j},\ 1 < i \neq j < n,\\
(\mathbf{R}_k^\mu)_{ii} &= \sum_{j\neq i}\frac{\partial^2 \Delta_{ij,k}(\boldsymbol{\alpha}_0,\beta_0)}{\partial^2 \alpha_i},\ i = 1,\ldots,n.
\end{aligned} \qquad (90)$$

The last term of (89) is $o_p(1)$ (equivalent to the limit of part (IV) of (67)). Additionally, from the proofs of Theorems 2 and 3, we have

$$\sqrt{N}(\hat{\beta}-\beta_0) - \mathrm{J}_0^{-1}B_0 = -\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0) - \frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\} + o_p(1) \qquad (91)$$

and

$$\left\|\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0 + \mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\|_\infty = o_p(n^{-1/2}). \qquad (92)$$

Substituting (91) and (92) into (88) and (89) respectively, we have

$$\begin{aligned}
\sqrt{N}(\hat{\delta}-\bar{\Delta}_n) - B_\beta - B_\alpha = {} & -(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0) - \frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\}\\
& - \sqrt{N}\Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0) + o_p(1)\\
= {} & -\frac{1}{\sqrt{N}}(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}m_2(\boldsymbol{\alpha}_0,\beta_0)
\end{aligned}$$

$$+\frac{1}{\sqrt{N}}\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21} - N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0) + o_p(1)$$

where B_β is defined by (36) in the main text.

Finally, by the Lindeberg-Feller CLT, we have

$$\sqrt{N}(\hat{\delta} - \bar{\Delta}_n) - B_\beta - B_\alpha \xrightarrow{d} \mathcal{N}(0, \Sigma_\Delta) \tag{93}$$

with

$$\begin{aligned}\Sigma_\Delta = \lim_{n\to\infty}\frac{1}{N}\Big\{&(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathrm{V}_{22}\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top\\
&+\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21} - N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{11}\left\{\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21} - N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\right\}^\top\\
&-\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21} - N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{12}\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top\\
&-\left\{\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21} - N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{12}\left[(\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top \mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top\right\}^\top\Big\}.\end{aligned} \tag{94}$$

Combining (86) and (93), and notice that $\hat{\delta} - \bar{\Delta}_n$ is uncorrelated with $\bar{\Delta}_n - \delta_0$ asymptotically, we have

$$\left(\frac{\Sigma_\Delta}{N} + \frac{4\Sigma_\delta}{n}\right)^{-1/2}\left(\hat{\delta} - \delta_0 - \frac{1}{\sqrt{N}}B_\beta - \frac{1}{\sqrt{N}}B_\alpha\right) \xrightarrow{d} \mathcal{N}(0, I_K).$$

The proof of asymptotic normality of $\hat{\delta}_{\mathrm{BG}}$ is similar to the proof of Theorem 1 and Theorem 5 since we have proved the asymptotic normality of the plug-in estimator $\hat{\delta}$. So, we omit it here. □

Proof of Theorem 7. Most of steps are basically same with what we have done in the proofs of Theorems 2–3 and Corollary 1, the difference is that we need to replace $(\boldsymbol{\alpha}_0, \beta_0)$ with $(\boldsymbol{\alpha}_*, \beta_{n*})$. For conciseness, we omit these repetitive steps. □

Proof of Theorem 8. Similarly, most of steps are same with the proofs of Theorem 4 and Theorem 5, we just need to replace $(\boldsymbol{\alpha}_0, \beta_0)$ with $(\boldsymbol{\alpha}_*, \beta_{n\star})$. We also omit these repetitive steps. □

C Additional simulation results

In Tables 10–12, we present the simulation results of β_0 and APEs for smaller networks with $n = 50$. The main conclusions drawn in Section 5.1–5.3 remain roughly unchanged, although the smaller sample size does increase the standard deviations of the estimates as

Table 10: Estimation Results of β_0

$n = 50,$ *density=25%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0599	-0.0713	0.0570	-0.0724	-0.0104	-0.0027
Median Bias	0.0568	-0.0670	0.0534	-0.0712	-0.0137	-0.0003
Standard Deviation	0.1262	0.2632	0.1258	0.2650	0.1185	0.2527
Mean Standard Error	0.1180	0.2720	0.1183	0.2712	0.1183	0.2712
Mean Absolute Bias	0.1116	0.2158	0.1100	0.2178	0.0960	0.1999
Median Absolute Bias	0.0946	0.1818	0.0914	0.1811	0.0842	0.1722
RMSE	0.1397	0.2727	0.1381	0.2748	0.1189	0.2527
90% Coverage Rate	83.3	90.1	83.4	90.5	90.5	92.9
95% Coverage Rate	90.1	94.9	90.6	94.3	95.2	96.4
$n = 50,$ *density=8.5%*	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.0940	-0.1948	0.1131	-0.1279	-0.0116	-0.0013
Median Bias	0.0915	-0.2097	0.1145	-0.1396	-0.0110	-0.0116
Standard Deviation	0.1651	0.4027	0.1641	0.4140	0.1475	0.3716
Mean Standard Error	0.1576	0.3833	0.1587	0.3842	0.1587	0.3842
Mean Absolute Bias	0.1528	0.3567	0.1614	0.3434	0.1186	0.2913
Median Absolute Bias	0.1324	0.2990	0.1439	0.2772	0.0985	0.2423
RMSE	0.1900	0.4473	0.1993	0.4333	0.1480	0.3716
90% Coverage Rate	83.2	85.9	80.8	86.4	91.5	92.1
95% Coverage Rate	89.5	91.5	89.4	92.8	96.5	95.8

expected. Overall, we find the performance of our BG estimators to be satisfactory under a very small sample size of $n = 50$.

D Additional Details of India Microfinance Network Application

D.1 Combining triangle and fitting empirical networks

In this section, we show how to extend our baseline model (4) to capture the formation of triangles in a graph explicitly, using a method similar to what Chandrasekhar and Jackson (2023) propose. Then, we provide a minimum distance estimator based on the moment conditions from the extended model.

Consider two independent subgraph formation processes. The first one is our basic

Table 11: Estimation Results of β_0 under Misspecification

$n = 50,\ density$=25%	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	0.1388	-0.1288	0.1470	-0.1264	-0.0056	0.0268
Median Bias	0.1401	-0.1291	0.1453	-0.1204	-0.0038	0.0361
Standard Deviation	0.1365	0.2927	0.1388	0.3239	0.1399	0.3033
Mean Standard Error	0.1273	0.2766	0.1377	0.3067	0.1377	0.3067
Mean Absolute Bias	0.1611	0.2541	0.1668	0.2601	0.1048	0.2240
Median Absolute Bias	0.1470	0.2105	0.1494	0.2138	0.0909	0.1819
RMSE	0.1946	0.3198	0.2022	0.3476	0.1400	0.3045
90% Coverage Rate	68.8	84.9	71.5	87.1	92.4	92.1
95% Coverage Rate	79.2	90.6	82.4	93.0	96.7	96.2

Table 12: Estimation Results of APE

$n = 50,\ density$=25%	Plug-in		Bagging	
	$X_{ij,1}$	$X_{ij,2}$	$X_{ij,1}$	$X_{ij,2}$
Mean Bias	-0.0063	0.0046	-0.0023	-0.0048
Median Bias	-0.0047	0.0065	-0.0019	-0.0043
Standard Deviation	0.0295	0.0578	0.0304	0.0600
Mean Standard Error	0.0296	0.0576	0.0296	0.0576
Mean Absolute Bias	0.0240	0.0457	0.0243	0.0471
Median Absolute Bias	0.0197	0.0371	0.0209	0.0382
RMSE	0.0302	0.0580	0.0305	0.0602
90% Coverage Rate	89.9	90.1	89.7	88.9
95% Coverage Rate	94.3	94.9	94.3	93.5

link-based model (4) with individual heterogeneity to generate $\{Y_{ij}\}_{1\le i<j\le n}$. The second process—independent of the first one—is the formation of triangles $\{T_{ijk}\}_{1\le i<j<k\le n}$, where $T_{ijk}=1$ if the links between nodes i,j,k are all connected and $T_{ijk}=0$ otherwise. Suppose

$$T_{ijk} \overset{i.i.d.}{\sim} \text{Bernoulli}(1-\lambda_{n0}), \quad 1\le i<j<k\le n,$$

where $\lambda_{n0}\to 1$ (hence $1-\lambda_{n0}\to 0$) as $n\to\infty$, which holds true in general for most networks as argued by Chandrasekhar and Jackson (2023). The observed network, denoted by $\{Y_{ij}^{\text{ob}}\}_{1\le i<j\le n}$, is then generated by

$$Y_{ij}^{\text{ob}} = 1\left(Y_{ij}=1 \text{ or } T_{ijk}=1 \text{ for some } k\neq i,j\right). \tag{95}$$

In the extended model (95), there are a total of $n+K+1$ unknown parameters, consisting of the heterogeneity vector $\boldsymbol{\alpha}_0$, common parameter β_0, and triangle formation parameter λ_{n0}.

The observed links in model (95) are dependent across dyads because of the existence of triangle formation. Due to the dependence among the links, it is generally impossible to write the joint likelihood function of the observed network directly. Therefore, we propose a minimum distance estimator based on certain moment conditions (Menzel (2016) and Chandrasekhar and Jackson (2023)). The idea is to first solve for $\widehat{\alpha}$ as a function of β and λ. Then, we estimate β and λ jointly by minimizing the weighted average of the Euclidean distance between the observed number of triangles and its expectation and the norm of the sample analogue of the moment conditions for β.

Specifically, for each i, the expectation of degree $d_i^{\text{ob}} := \sum_{j\neq i} y_{ij}^{\text{ob}}$ is

$$(n-1)^{-1}\mathbb{E}\left[d_i^{\text{ob}}|\mathbf{X}=\mathbf{x}\right] = 1-\left(1-\frac{1}{n-1}\sum_{j\neq i} p_{ij}(\boldsymbol{\alpha}_0,\beta_0)\right)\lambda_{n0}^{n-2}. \tag{96}$$

Similar to we show in Section 3.1, the corresponding moment conditions (96) induce a SGD algorithm for solving $\widehat{\boldsymbol{\alpha}}(\lambda_n,\beta)$ as a function of (λ_n,β), i.e., we may modify (7) to be

$$r_i(\boldsymbol{\alpha},\lambda_n,\beta) = \alpha_i + \lambda_n^{-(n-2)}\left((n-1)^{-1}d_i^{\text{ob}}-1\right) + 1 - \frac{1}{n-1}\sum_{j\neq i} p_{ij}(\boldsymbol{\alpha},\beta),$$

and iterate it until convergence.

Next, the expectation of the average number of triangles in the observed network is

$$\begin{aligned} &h_n(\boldsymbol{\alpha}_0,\lambda_{n0},\beta_0)\\ :=&\binom{n}{3}^{-1}\mathbb{E}\left[\sum_{1\le i<j<k\le n} Y_{ij}^{\text{ob}}Y_{jk}^{\text{ob}}Y_{ki}^{\text{ob}}|\mathbf{X}=\mathbf{x}\right]\end{aligned}$$

$$= 1 - \left(1 - \binom{n}{3}^{-1} \sum_{1\le i<j<k\le n} p_{ij}(\boldsymbol{\alpha}_0, \beta_0) p_{jk}(\boldsymbol{\alpha}_0, \beta_0) p_{ki}(\boldsymbol{\alpha}_0, \beta_0)\right) \lambda_{n0} \left[1 - (1 - \lambda_{n0}^{n-3})^3\right].$$

Finally, the moment condition for the common parameter β is

$$m_2(\boldsymbol{\alpha}_0, \beta_0) = \sum_{i=1}^{n} \sum_{j>i} [y_{ij}^{\text{ob}} - (1 - (1 - p_{ij}(\boldsymbol{\alpha}_0, \beta_0))\lambda_{n0}^{n-2}] x_{ij}.$$

With $\widehat{\boldsymbol{\alpha}}(\lambda_n, \beta)$, we estimate $(\hat{\lambda}_n, \hat{\beta})$ by a minimum distance estimator:

$$(\hat{\lambda}_n, \hat{\beta}) := \arg \min_{\lambda_n \in (0,1), \beta} w_T \left| \binom{n}{3}^{-1} \sum_{1\le i<j<k\le n} y_{ij}^{\text{ob}} y_{jk}^{\text{ob}} y_{ki}^{\text{ob}} - h_n\left(\widehat{\boldsymbol{\alpha}}(\lambda_n, \beta), \lambda_n, \beta\right) \right| + \|m_2(\widehat{\boldsymbol{\alpha}}(\lambda_n, \beta), \beta)\|_1,$$

where w_T is the weight of moment condition for triangle count. In the empirical application, we set it to be 10. Moreover, to improve computational efficiency, we set $\lambda'_n = \lambda_n^{n-2}$ and use

$$\lambda_n[1 - (1 - \lambda_n^{n-3})^3] \approx \lambda'_n(3 - 3\lambda'_n + {\lambda'_n}^2)$$

because λ_{n0} is close to 1 in theory. In the implementation of this algorithm, we estimate $\hat{\lambda'_n}$ first and then let $\hat{\lambda}_n = (\hat{\lambda'_n})^{1/(n-2)}$. The resulting estimator is $(\widehat{\boldsymbol{\alpha}}, \hat{\lambda}_n, \hat{\beta}) := (\widehat{\boldsymbol{\alpha}}(\hat{\lambda}_n, \hat{\beta}), \hat{\lambda}_n, \hat{\beta})$, which is used to simulate the networks for the India microfinance network application. Finally, we compute the characteristics of the simulated networks and summarize the results in column "Extended" of Table 9.