Bagging the Network*

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

[a]Department of Economics and Risk Management Institute, National University of Singapore

[b]Department of Economics, The Chinese University of Hong Kong

Abstract

This paper studies parametric estimation and inference in a dyadic network formation model with nontransferable utilities, incorporating observed covariates and unobservable individual fixed effects. We address both theoretical and computational challenges of maximum likelihood estimation in this complex network model by proposing a new bootstrap aggregating (bagging) estimator, which is asymptotically normal, unbiased, and efficient. We extend the approach to estimating average partial effects and analyzing link function misspecification. Simulations demonstrate strong finite-sample performance. Two empirical applications to Nyakatoke risk-sharing networks and Indian microfinance data find insignificant roles of wealth differences in link formation and the strong influence of caste in Indian villages, respectively.

Keywords: bootstrap aggregating, Cramér–Rao lower bound, dyadic network formation, nontransferable utilities, one-step approximation, unobserved heterogeneity.

*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 Donald Andrews, Xiaohong Chen, Chih-Sheng Hsieh, Bryan Graham, Yuichi Kitamura, Oliver Linton, Shuyang Sheng, Martin Weidner, Weichen Wang, Jun Yu, Yichong Zhang, and the conference/seminar participants of University of Sydney, University of Macau, Singapore Management University, and Econometric Society World Congress 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. 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 behaviors—an essential area of inquiry. By 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 (CRLB, Rao, 1992) asymptotically, and a computationally efficient algorithm is provided. By "flexible," we include both observed pairwise covariates for the homophily effect and the unobserved individual heterogeneity as fixed effects. 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., Chatterjee et al., 2011; Graham, 2017; Qu et al., 2025) that considers transferable utilities (TU), we model real-world social interactions by requiring bilateral consent via NTU. For instance, friendship is usually formed only when both individuals are willing to accept each other, or in other words, when both derive sufficiently high utilities from establishing the friendship. It is even more prominent in business networks since no firm would make a deal if it incurs a loss from the transaction when there is no mechanism to guarantee profit redistribution. Moreover, NTU can effectively incorporate homophily effects on unobserved heterogeneity (Gao et al., 2023). We provide a more in-depth comparison between TU and NTU in Remark 2.

The presence of fixed effects together with 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. Such nonseparability in the fixed effects makes inapplicable the arithmetic-differencing-based methods that cancels out the fixed effects (e.g., the innovative tetrad estimator of Graham,

2017 cannot be applied under NTU). Second, classical maximum likelihood estimation is subject to well-known theoretical difficulties, most notably the incidental parameter problem, which leads to nonexistence and nonuniqueness of estimators as well as asymptotic bias. It also entails substantial computational burdens when individual fixed effects are included alongside fixed-dimensional structural (homophily) parameters. To deal with these challenges, existing dyadic network formation literature with TU typically relies on specific distributional assumptions (e.g., the joint maximum likelihood (JML) estimator of Graham, 2017 uses logistic distribution while Dzemski, 2019 uses normal distribution) to obtain the fixed effects as functions of homophily parameters and subsequently maximize the composite or profile likelihood function with these estimated functions plugged in. However, the combination of NTU and a general link function renders the existing methods inapplicable. Third, our network formation model with NTU makes the asymptotic theory different from the current literature that focuses on TU. For example, the Jacobian matrix of the moment equations used to construct initial moment estimators is asymmetric due to NTU, invalidating the asymptotic analysis under TU.

To deal with those theoretical and computational challenges, this paper proposes a bootstrap aggregating (bagging) estimator for the homophily parameters β_0. We show its asymptotic normality centered at zero, and efficiency in the sense that the bagging estimator achieves the CRLB asymptotically. Our paper is the first one in the literature of dyadic network formation with NTU that has inference and efficiency results. A key step of our proposal is inspired by Le Cam (1969)'s one-step approximation to the maximum likelihood (ML) estimator, which effectively circumvents the computational difficulties associated with the classic ML estimator. The one-step approximation updates once, via a Newton-type procedure, an initial estimator that converges to the true parameter at the parametric rate:

$$\widehat{\beta}_{\text{OS}} = \widehat{\beta}_{\text{Initial}} + \mathrm{I}_n^{-1}(\widehat{\beta}_{\text{Initial}}) \cdot s_n(\widehat{\beta}_{\text{Initial}}),$$

where $\mathrm{I}_n(\cdot)$ is the negative Hessian and $s_n(\cdot)$ is the score for the log likelihood function. The one-step estimator is significantly faster and more stable to compute than the classic ML estimator, and yet retains the same asymptotic rate and variance. It requires only a single updating step—making it ideal for large datasets, in particular for network data which includes links between all pairs of individuals[1]—and remains asymptotically optimal. For $\widehat{\beta}_{\text{Initial}}$, we propose a general method of moment estimator and characterize its asymptotic distribution. In this step, we also estimate the fixed effects and prove their convergence to the true parameters $\boldsymbol{\alpha}_0$ in the ℓ_∞ norm.

[1]For example, in a reasonably small network with $n = 100$ individuals, there are a total of $N = (n^2 - n)/2 = 4950$ undirected links.

Next, since the one-step estimator is asymptotically equivalent to the ML estimator, it carries asymptotic bias. We debias it via the bagging method from the machine learning literature (Breiman, 1996; Hirano and Wright, 2017) after split-network jackknife. As far as we know, bagging is novel in the network formation literature. The idea is, for each round $t \in \{1, ..., \tilde{T}_n\}$, we randomly split all nodes into two halves and estimate parameters only based on each subnetwork formed among the half nodes to obtain $\widehat{\beta}^{(t)}_{\text{OS},1}$ and $\widehat{\beta}^{(t)}_{\text{OS},2}$. Then, we take average over the $\tilde{T}_n$ splits and debias $\widehat{\beta}_{\text{OS}}$ by

$$\widehat{\beta}_{\text{BG}} = 2\widehat{\beta}_{\text{OS}} - \frac{1}{2\tilde{T}_n} \sum_{t=1}^{\tilde{T}_n} \left(\hat{\beta}^{(t)}_{\text{OS},1} + \hat{\beta}^{(t)}_{\text{OS},2} \right).$$

We show that, as n and $\tilde{T}_n$ go to infinity, the bagging estimator $\widehat{\beta}_{\text{BG}}$ is asymptotically unbiased and achieves the efficiency bound. Note that if we set $T_n = 1$, $\widehat{\beta}_{\text{BG}}$ reduces to the split-network jackknife estimator, which is asymptotically unbiased but inefficient because it doubles the asymptotic variance of the ML estimator. Bagging is essential in deflating the variance to the efficiency level, i.e., the CRLB. Moreover, bagging also makes the computation more stable and insensitive to the choice of random splits.

As two extensions, we provide a consistent estimator and prove its asymptotic normality for the average partial effects (APEs) and discuss how misspecification of the link function affects the analysis, the latter of which is much less considered in this literature. We show that the APEs can be consistently estimated and that $\widehat{\beta}_{\text{BG}}$ converges to the pseudo-true value under link function misspecification.

Simulation results confirm that the proposed estimators for the homophily parameters, individual fixed effects, and APEs perform as predicted by the theory. We present two empirical examples. First, in the risk-sharing network data of Nyakatoke (De Weerdt, 2004), our method indicates that wealth differences have no statistically significant effect on link formation. Second, we apply our method to the information and favor networks to each of the 75 villages from the India microfinance dataset (Banerjee et al., 2013, 2024). It is evident that belonging to the same caste strongly and uniformly boosts link formation among households in the sampled Indian villages, whereas the influence of other factors varies village by village. A demonstration of our proposed methods is available at the GitHub repository https://github.com/YapengZheng/network_formation_NTU.

Literature Review. Our paper contributes to the literature on dyadic network formation in a single large network. Most existing work studies TU, which allow individual fixed effects to be eliminated by arithmetic differencing (Chatterjee et al., 2011; Graham, 2017; Dzemski, 2019; Zeleneev, 2020; see Graham, 2020, for a review). Gao et al. (2023) study a semiparametric NTU model using logical differencing, but without inference for homophily

parameters or estimators of fixed effects. We complement their work by establishing inference for homophily parameters, delivering ℓ_∞-consistent estimators of fixed effects, and developing asymptotic results for the APEs and under link function misspecification.

Our paper also builds on Graham (2017), who introduces a tetrad logit estimator and a joint ML estimator under TU and logistic link functions. These methods, as well as functional differencing (Bonhomme, 2012; Bonhomme and Dano, 2024), do not extend to NTU due to the fixed effects being not additively separable in the specification of the utility surplus from a link. Recent contributions also remain within TU: Hughes (2023) develops a jackknife bias correction, Qu et al. (2025) propose a projection approach for directed networks, and Candelaria and Zhang (2024) study robust inference in bipartite networks. Semiparametric and nonparametric TU approaches include Toth (2017), Gao (2020), and Candelaria (2024). In contrast, our estimator applies under NTU.

Methodologically, our work relates to the large-T panel literature on nonlinear fixed-effect models (Hahn and Newey, 2004; Hahn and Kuersteiner, 2011; Dhaene and Jochmans, 2015; Fernández-Val and Weidner, 2016, 2018). These methods rely on concavity of log-likelihood functions and/or sparsity of certain derivatives of functionals of fixed effects assumptions that are hard to verify in our setting. Instead, we adapt the sample-splitting idea (see Mei et al., 2023; Liao et al., 2024 for tackling Nickell-type biases in panel predictive regressions using split-sample strategies) and establish formally that bagging delivers unbiased and efficient estimation. Related work on nonlinear factor models and orthogonalized estimators (Chen et al., 2021; Bonhomme et al., 2024) requires conditions that preclude NTU.

Two adjacent literature are worth noting. First, dyadic formation models are often used to control for endogeneity in structural models of social interactions (Goldsmith-Pinkham and Imbens, 2013; Hsieh and Lee, 2016; Johnsson and Moon, 2021; Auerbach, 2022). Our contribution differs in focusing directly on efficient estimation of the formation process, though our results may be useful for future studies of spillovers (Jackson et al., 2024). Second, there is a line of work on strategic network formation and empirical games based on pairwise stability (Jackson and Wolinsky, 1996; Mele, 2017; de Paula et al., 2018; Sheng, 2020; Chandrasekhar and Jackson, 2025). These models incorporate externalities but typically impose restrictions on heterogeneity or the degree distribution and often require TU (e.g., Pelican and Graham, 2024). Our fixed-effects NTU approach, which permits arbitrary correlation between observables and the fixed effects, is therefore complementary to and methodologically distinct from the existing literature (for a comprehensive review of the two approaches, see de Paula, 2020).

Organization. The rest of the paper is organized as follows. Section 2 formally introduces a dyadic model of link formation and presents our estimation algorithm. Section 3

develops the main theoretical results for our proposed estimators. Section 4 develops asymptotic theory for the APEs and analyzes link function misspecification. Section 5 carries out simulation studies. Section 6 provides two empirical applications. All proofs are relegated to the Appendix.

Notation. Let ":=" denote a definition, and let the superscript "$\top$" denote the transpose of a vector or a matrix. We use bold-case for variables of increasing dimension with n. For example, the true fixed effects $\boldsymbol{\alpha}_0 = (\alpha_{i0})_{1\le i\le n}$ is $n\times 1$. For an $n\times 1$ vector $\mathbf{a} = (a_1, \ldots, a_n)^\top$, 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 notation for order) is written for a vector (or matrix), it means that each element in the vector (or matrix) is of the order in $O(\cdot)$. Here, "plim" denotes the probability limit, "$\xrightarrow{p}$" convergence in probability, and "$\xrightarrow{d}$" convergence in distribution. Unless otherwise noted, for all convergence results we pass $n \to \infty$. 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. Let $F : \mathbb{R} \to (0, 1)$ be a link function. 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 the shorthand 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 one."

2 Model and Computation

We consider an undirected 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. We use a binary random variable Z_{ij} to denote agent i's decision on whether to link with j, that is,

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

We rule out self-loops, i.e., $Z_{ii} = 0$, $i \in \mathcal{I}_n$. Three components determine the value of Z_{ij}: (i) the unobserved fixed effect α_{i0}, which is specific to agent i; (ii) dyad-specific index $X_{ij}^\top \beta_0$ that captures the homophily effect in the observable characteristics of each pair (i, j), where $X_{ij} \in \mathbb{R}^K$ denotes the symmetric dyad-level covariates for all $i \ne j$, and (iii) an idiosyncratic component ϵ_{ij} with a known distribution, assumed to be independently and identically distributed across all dyads (i, j).

Under NTU, an observed link Y_{ij} between i and j is formed as

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

The user specifies a link function F, which is the distribution function of the idiosyncratic error ϵ_{ij}. The log-likelihood is

$$\ell_n(\boldsymbol{\alpha}, \beta) := \sum_{i=1}^{n} \sum_{j>i} \{y_{ij} \log p_{ij}(\boldsymbol{\alpha}, \beta) + (1 - y_{ij}) \log (1 - p_{ij}(\boldsymbol{\alpha}, \beta))\}.$$

Remark 1. While we deal with the stylized model (1)–(2) to fix ideas, multiple generalizations are possible. First, our model can be extended to cover directed network with NTU by introducing two sets of heterogeneity that captures in-degree and out-degree separately as in Yan et al. (2019) and Hughes (2023). Specifically, we may include an additional unobserved fixed effect γ_{j0} in (1). It would require estimating $2n$ fixed effects via moment restrictions defined in Module 1 below. Second, we consider the general case in which X_{ij} represents generic symmetric pairwise observable characteristics such as distance between two households in a village. As a special case, our model allows X_{ij} to be generated by a symmetric function of individual characteristics X_i and X_j (e.g., Graham, 2017). While accommodating asymmetric X_{ij} is feasible and supported by simulations, it introduces additional technical complications in the proofs and would obscure the main focus of the paper.

Remark 2. Our network formation model (2) with NTU is different from that with TU in the following form:

$$Y_{ij} = \mathbb{1}\left\{\alpha_{i0} + \alpha_{j0} + X_{ij}^{\top}\beta_0 - \epsilon_{ij} > 0\right\}. \tag{3}$$

The TU model (3) essentially asserts that, if the joint surplus generated by a bilateral link $\alpha_{i0} + \alpha_{j0} + X_{ij}^{\top}\beta_0 - \epsilon_{ij}$ is positive, then the link between i and j is formed. An important assumption behind the TU model (3) is that the link surplus can be freely distributed between i and j, and that bargaining efficiency is always achieved, which can be strong in many networks (e.g., risk sharing network, friendship network). Our NTU model (2), on the other hand, requires that the utility surplus from the link for both i and j to be strictly positive in order to form a link, which is arguably more realistic in the aforementioned networks. Furthermore, the NTU model (2) reflects the fact that the party with relatively lower utility is the pivotal one in link formation. Finally, it can be shown (see Gao et al., 2023) that the NTU model (2) can accommodate homophily effect in both observable and unobservable covariates.

Given the model, we introduce the algorithm to estimate the homophily coefficient β_0. There are three sequential modules—Joint Method of Moments (JMM), One-Step (OS), and

Bagging (BG)—that lead to $\widehat{\beta}_{\text{BG}}$. Specifically, Module JMM provides an initial consistent estimator, which, however, does not reach the CRLB and is biased. We refine the JMM estimator with the one-step adjustment to achieve the CRLB. Finally, we apply the bagged split-network jackknife to debias the one-step estimator while preserving its efficiency.

We define a few objects before each module. Let $\mathbf{Y} = (Y_{ij})_{1\le i,j\le n}$ be the $n \times n$ adjacency matrix and $\mathbf{X} = (X_{ij})_{1\le i,j\le n}$ be the $n \times n \times k$ random tensor of covariates. Denote their realizations by $\mathbf{y} = (y_{ij})_{1\le i,j\le n}$ and $\mathbf{x} = (x_{ij})_{1\le i,j\le n}$, respectively. The *degree* $d_i := \sum_{j\neq i} Y_{ij}$ is defined for each $i \in \mathcal{I}_n$ of the observed network $\mathbf{Y}$. Define a vector of moment functions $\mathbf{m}(\boldsymbol{\alpha}, \beta) := (\mathbf{m}_1^\top(\boldsymbol{\alpha}, \beta), m_2^\top(\boldsymbol{\alpha}, \beta))^\top$, where $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) := (d_i - \sum_{j\neq 1} p_{ij}(\boldsymbol{\alpha}, \beta))_{i=1}^n$ is an n-dimensional function that concerns the average degree of each i, and $m_2(\boldsymbol{\alpha}, \beta) := \sum_{i=1}^{n} \sum_{j>i} [y_{ij} - p_{ij}(\boldsymbol{\alpha}, \beta)]x_{ij}$ is a K-dimensional function.

Module 1 (JMM)**.** *The* ***JMM estimator*** $(\widehat{\boldsymbol{\alpha}}, \widehat{\beta})$ *is the solution to the* $(n + K)$*-equation system* $\mathbf{m}(\boldsymbol{\alpha}, \beta) = 0$.

To find the solution to $\mathbf{m}(\boldsymbol{\alpha}, \beta) = 0$, for each β we let

$$r_i(\boldsymbol{\alpha}, \beta) = \alpha_i + (n-1)^{-1}\bigg(d_i - \sum_{j\neq i} p_{ij}(\boldsymbol{\alpha}, \beta)\bigg), \qquad i \in \mathcal{I}_n \tag{4}$$

and $\mathbf{r}(\boldsymbol{\alpha}, \beta) = \big(r_1(\boldsymbol{\alpha}, \beta), \ldots, r_n(\boldsymbol{\alpha}, \beta)\big)^\top$. 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. Starting with an initial value $\boldsymbol{\alpha}^0$, we iterate $\boldsymbol{\alpha}^{k+1}(\beta) = \mathbf{r}(\boldsymbol{\alpha}^k(\beta), \beta)$ until convergence to obtain $\widehat{\boldsymbol{\alpha}}(\beta)$, and then we solve the finite dimensional equations $m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta) = 0$.

Next, the OS module involves the score and information matrix. Define the score of ℓ_n as $\mathbf{s}(\boldsymbol{\alpha}, \beta) = \big(\mathbf{s}_1^\top(\boldsymbol{\alpha}, \beta), s_2^\top(\boldsymbol{\alpha}, \beta)\big)^\top = \big(\partial\ell_n/\partial\boldsymbol{\alpha}^\top, \partial\ell_n/\partial\beta^\top\big)^\top$, and partition the information matrix

$$\mathbf{I}(\boldsymbol{\alpha}, \beta) = \mathbb{E}[\mathbf{s}(\boldsymbol{\alpha}, \beta)\mathbf{s}(\boldsymbol{\alpha}, \beta)^\top | \mathbf{x}, \boldsymbol{\alpha}] := \begin{pmatrix} \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta) & \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta) \\ \mathbf{I}_{12}^\top(\boldsymbol{\alpha}, \beta) & \mathrm{I}_{22}(\boldsymbol{\alpha}, \beta) \end{pmatrix} \tag{5}$$

into the four compatible blocks. Define the concentrated score function and information matrix of β as

$$\begin{aligned} s_n(\boldsymbol{\alpha}, \beta) &= s_2(\boldsymbol{\alpha}, \beta) - \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{s}_1(\boldsymbol{\alpha}, \beta), \text{ and} \\ \mathrm{I}_n(\boldsymbol{\alpha}, \beta) &= \mathrm{I}_{22}(\boldsymbol{\alpha}, \beta) - \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{I}_{12}(\boldsymbol{\alpha}, \beta), \end{aligned}$$

respectively.

Module 2 (OS)**.** *Substitute the JMM estimator* $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ *into*

$$\hat{\beta}_{\mathrm{OS}} := \hat{\beta} + \mathrm{I}_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta})^{-1} s_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta}). \tag{6}$$

The variance of the **OS estimator** $\hat{\beta}_{\mathrm{OS}}$ achieves the CRLB asymptotically. To illustrate the idea, suppose β is a scalar. When $s_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ is positive, implying that an increase in β raises the log-likelihood, it is desirable to increase the initial estimator. When the $\mathrm{I}_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ is large, which means there is sufficient information to identify the parameters, one may want to make the adjustment smaller to avoid excessive correction. This explains why the inverse of information matrix also plays a role.

Finally, bagging is featured by randomization. Assume an even integer n for convenience. Let $t = 1, 2, \ldots, \tilde{T}_n$, for some $\tilde{T}_n \leq \binom{n}{n/2}$, index an equal-sized random partition of $\mathcal{I}_n$ into $\mathcal{I}_{1,n}^{(t)}$ and $\mathcal{I}_{2,n}^{(t)}$ such that $\mathcal{I}_{1,n}^{(t)} \cup \mathcal{I}_{2,n}^{(t)} = \mathcal{I}_n$, $\mathcal{I}_{1,n}^{(t)} \cap \mathcal{I}_{2,n}^{(t)} = \emptyset$, and the splits are independent over t.

Module 3 (BG)**.** *For each* $t = 1, ..., \tilde{T}_n$*, apply the JMM and OS modules on the subnetwork indexed by* $\mathcal{I}_{1,n}^{(t)}$ *to obtain* $\hat{\beta}_{\mathrm{OS},1}^{(t)}$*. Repeat the same procedure on* $\mathcal{I}_{2,n}^{(t)}$ *to obtain* $\hat{\beta}_{\mathrm{OS},2}^{(t)}$*. Apply the bagged jackknife to obtain the* ***BG estimator*** $\hat{\beta}_{\mathrm{BG}} := 2\hat{\beta}_{\mathrm{OS}} - (2\tilde{T}_n)^{-1} \sum_{t=1}^{\tilde{T}_n} (\hat{\beta}_{\mathrm{OS},1}^{(t)} + \hat{\beta}_{\mathrm{OS},2}^{(t)})$.

We employ split-network jackknife to debias $\hat{\beta}_{\mathrm{OS}}$. Due to the equal splits of nodes, each of $\hat{\beta}_{\mathrm{OS},1}^{(t)}$ and $\hat{\beta}_{\mathrm{OS},2}^{(t)}$ incurs twice the leading bias in the asymptotic expansion. If we apply the split-network jackknife only once, then

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

self-cancels the leading bias. However, the variance $\hat{\beta}_{\mathrm{OS-SJ}}^{(t)}$ is doubled, since splitting the network in half causes the links between nodes belonging to different subnetworks to be ignored. Furthermore, splitting the whole network randomly makes the estimator computationally unstable. To deal with these issues, we let $\tilde{T}_n \to \infty$ and indeed $\hat{\beta}_{\mathrm{BG}} = \tilde{T}_n^{-1} \sum_{t=1}^{\tilde{T}_n} \hat{\beta}_{\mathrm{OS-SJ}}^{(t)}$ averages $\hat{\beta}_{\mathrm{OS-SJ}}^{(t)}$ over $\tilde{T}_n$ independent splits.

When computing $\hat{\beta}_{\mathrm{OS},1}^{(t)}$ and $\hat{\beta}_{\mathrm{OS},2}^{(t)}$ for each random split t, we do not re-estimate the initial $\widehat{\beta}$ by applying JMM to the corresponding subnetworks. Instead, the full-sample JMM estimator $\widehat{\beta}$ is retained, and only $\widehat{\boldsymbol{\alpha}}$ is updated via (4). Consequently, the procedure remains computationally efficient for moderate values of $\tilde{T}_n$, such as 200 or 400 in our simulations.

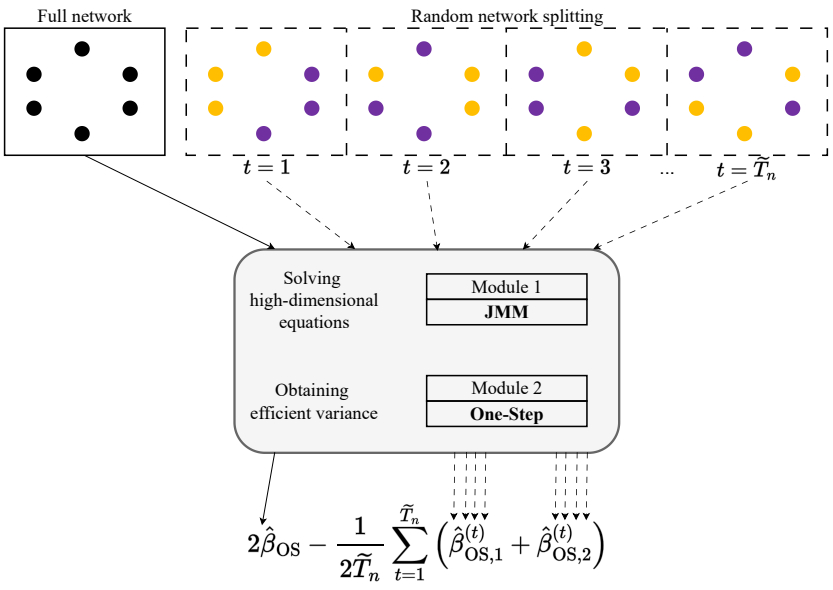


Figure 1: Flowchart of the estimation procedure

To summarize, Figure 1 illustrates how each module is stacked together to produce $\widehat{\beta}_{\mathrm{BG}}$. We begin by applying modules JMM and OS to the full network, whose nodes are shown as black dots, yielding $\widehat{\beta}_{\mathrm{OS}}$. Next, for each random split $t = 1, ..., \tilde{T}_n$, we divide the nodes into two halves. For instance, in split $t = 1$, the yellow dots in the top-left represent $\mathcal{I}^{(1)}_{1,n}$ and the purple dots represent $\mathcal{I}^{(1)}_{2,n}$. Running JMM (with $\widehat{\beta}$ from full sample and only updating $\widehat{\boldsymbol{\alpha}}$ via (4)) and OS on each subnetwork produces $\hat{\beta}^{(1)}_{\mathrm{OS},1}$ and $\hat{\beta}^{(1)}_{\mathrm{OS},2}$. Repeating this procedure over $t = 1, ..., \tilde{T}_n$ and applying the bagged jackknife yields the bagging estimator $\hat{\beta}_{\mathrm{BG}}$.

3 Large Sample Theory

Given the algorithm of $\hat{\beta}_{\mathrm{BG}}$, in this section we prove that $\hat{\beta}_{\mathrm{BG}}$ is asymptotically normal, unbiased, and efficient. We first estimate the high-dimensional fixed effects $\boldsymbol{\alpha}$ as a function of β by solving $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = 0$. We establish the existence and uniqueness of $\widehat{\boldsymbol{\alpha}}(\beta)$—the solution to $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = 0$—for each β that is local to β_0. Our method accommodates NTU and encompasses a broad class of distributions beyond the logistic. Once $\widehat{\boldsymbol{\alpha}}(\beta)$ is obtained, we solve a different set of finite-dimensional moment conditions $m_2(\boldsymbol{\alpha}, \beta) = 0$ to compute the JMM estimator $\hat{\beta}$. We show that $\hat{\beta}$ is $\sqrt{N}$-consistent for β_0, which suffices for the purpose of the one-step adjustment. As a by-product, we also demonstrate that $\widehat{\boldsymbol{\alpha}}(\hat{\beta})$ converges to $\boldsymbol{\alpha}_0$ in the ℓ_∞ norm. Then the JMM estimator is substituted into Module 2 to get $\hat{\beta}_{\mathrm{OS}}$. We prove that $\hat{\beta}_{\mathrm{OS}}$ achieves the CRLB but remains asymptotically biased. Finally, we show that the debias-once $\hat{\beta}^{(1)}_{\mathrm{OS-SJ}}$ is properly centered at zero but has a variance twice as large as that

of $\widehat{\beta}_{\text{OS}}$, whereas averaging over repeated splits t yields $\widehat{\beta}_{\text{BG}}$ with the desired properties.

We begin by stating three baseline assumptions that underlie the theoretical results.

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*

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

where

$$\Pr(Y_{ij} = y_{ij}|x_{ij}, \alpha_{i0}, \alpha_{j0}) = p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{1-y_{ij}}, \tag{7}$$

for all $i \neq j$.

Assumption 1 is similar to Assumption 1 of Graham (2017), except for two important differences. First, under NTU α_{i0} and α_{j0} are not additively separable in the linking probability between i and j, and thus the tetrad logit estimator of Graham (2017) does not apply in our setting. Second, the functional form of $F(\cdot)$ is general (See Assumption 3 below for mild restrictions) and includes the commonly used logistic (e.g., Chatterjee et al., 2011; Graham, 2017; Qu et al., 2025) and probit as special cases. The multiplicative form of the joint likelihood implies that the idiosyncratic ϵ_{ij}'s are i.i.d. across the dyads (i, j), i.e., links are formed independently of one another conditional on the agent attributes. It is suitable in settings such as risk-sharing networks, online friendships, trade networks, and conflicts between nation-states. The lack of interdependence, however, rules out 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, the ML estimator converges to the parameter value that minimizes the Kullback-Leibler divergence between the true and the misspecified model (White, 1982). The issue is complicated by the high-dimensional individual fixed effects and NTU. To our knowledge, the misspecification issue has not been investigated in the network formation literature. We discuss the impact of link function misspecification on the theoretical results in Section 4.2 and provide supporting simulation evidence in Section 5.

The next two assumptions facilitate our asymptotic analysis.

Assumption 2 (Bounded Support and Random Sampling)**.** *Suppose the following conditions hold: (a)* $\boldsymbol{\alpha}_0$ *lies in the interior of a compact set* $\mathbb{A} \subset \mathbb{R}^n$*, (b)* β_0 *lies in the interior of a compact set* $\mathbb{B} \subset \mathbb{R}^K$*, and (c)* X_{ij} *satisfy* $X_{ij} \in \mathbb{X} \subset \mathbb{R}^K$ *for some compact set* $\mathbb{X}$*.*

Assumption 2 collects and combines Graham (2017)'s Assumptions 2 and 5(i). It implies that the probability of a link forming between dyad (i,j) is uniformly bounded within $[\kappa, 1-\kappa]$ for some $\kappa \in (0, 1/2)$, which requires the network to be dense.[2] The dense network makes it possible to estimate α_{i0} consistently for each i.

Note that our theory in principle can allow the support of X_{ij} to be unbounded; however, it would add little theoretical insight but incur more technical complexity in the rates of convergence via the Bernstein inequalities to bound the tail probabilities of random variables. Assumption 2(c), which is similar to Graham (2017, Assumption 2(ii)), allows us to focus on the main idea.

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

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

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. Assumption 3 is regarded as mild, since in conjunction with Assumption 2 it is typically satisfied under common distributions, including the logistic and normal. This Assumption is similar to Fernández-Val and Weidner (2016, Assumption 4.3(v)), which restricts the smoothness of the likelihood functions.

3.1 JMM

Recall that the JMM module gives $(\widehat{\boldsymbol{\alpha}}, \widehat{\beta})$ to start with. The next lemma concerns the existence and uniqueness of $\widehat{\boldsymbol{\alpha}}(\beta)$, as well as the convergence of $\boldsymbol{\alpha}^k(\beta)$ to $\widehat{\boldsymbol{\alpha}}(\beta)$ via (4).

Lemma 1. *If Assumptions 1–3 hold, then there exists a unique $\widehat{\boldsymbol{\alpha}}(\beta)$ w.p.a.1 for each $\beta \in \{\beta \in \mathbb{B} | \, \|\beta - \beta_0\|_2 < c\}$ in a neighborhood around β_0, where $c > 0$ is small but fixed. Moreover, uniformly across all k, we have*

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

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

[2] 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 $\lim_{n\to\infty} \rho_n \in [c_1, c_2]$ for some constant $0 < c_1 \leq c_2 < 1$.

Lemma 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 every two iterations. Computing $\widehat{\boldsymbol{\alpha}}(\beta)$ is fast in the simulations, which is another advantage of our iterative algorithm. It is worth mentioning that we deviate from the existing methods (e.g., Theorem 1.5 of Chatterjee et al., 2011 or the fixed point equation (17) of Graham, 2017) in this step by not requiring ϵ_{ij} to be a logistic random variable, nor the link formation process being TU. Instead, we use a gradient-descent-type iterative algorithm (4) to compute $\widehat{\boldsymbol{\alpha}}(\beta)$ as a function of β and show that it is a contraction mapping. As a result, it can accommodate general non-logistic link functions and NTU.

Although $\widehat{\boldsymbol{\alpha}}(\beta)$ is unique by Lemma 1 for any β that lies within a distance of c of β_0, in principle there could be multiple solutions to $m_2(\widehat{\boldsymbol{\alpha}}(\beta), \beta) = 0$. The next identification condition guarantees that any such $\widehat{\beta}$ is consistent for β_0. To state the assumption, we define the concentrated moment equation for β as

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

where $\boldsymbol{\alpha}(\beta)$ is the unique solution to $\mathbb{E}[\mathbf{m}_1(\boldsymbol{\alpha}, \beta)|\mathbf{x}, \boldsymbol{\alpha}_0] = \mathbf{0}_n$, a result from the proof of Lemma 1.

Assumption 4 (Identification of β_0)**.** *Suppose for all $\delta > 0$ and for n large enough*

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

Assumption 4 identifies 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. Assumption 4 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 van der Vaart (2000, Page 45).

Remark 3. 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 Assumption 4 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 the concentrated Jacobian matrix for β) is strictly large than 0, which is quite weak.

In the next theorem, we prove that $\hat{\beta}$ is consistent for β_0 and that $\widehat{\boldsymbol{\alpha}}$ is uniformly consistent for $\boldsymbol{\alpha}_0$ in the sup norm. Furthermore, we establish asymptotic normality for $\hat{\beta}$.

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

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

Furthermore, we have

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

where J_0, B_0, *and* Ω_0 *are defined in (A22), (A1), and (A2), respectively.*

Theorem 1 shows that the JMM estimator $\hat{\beta}$ is asymptotically normal, but 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 nonlinear panel fixed effects regression with large N and T. Moreover, $\hat{\beta}$ does not achieve the CRLB of I_0^{-1}. We refine $\hat{\beta}$ by the following modules.

3.2 One-Step Estimator

The first refinement concerns achieving the CRLB. We follow Le Cam (1969)'s one-step adjustment as in Module 2. Algebra shows

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

Therefore, a Taylor expansion on the right hand side of (6) yields

$$\hat{\beta}_{\mathrm{OS}} - \beta_0 \approx \mathrm{I}_n(\boldsymbol{\alpha}_0, \beta_0)^{-1} s_n(\boldsymbol{\alpha}_0, \beta_0) \tag{8}$$

in large samples. To establish (8) rigorously and hence the asymptotic normality of $\hat{\beta}_{\mathrm{OS}}$, we impose an additional assumption on the behavior of the information matrix (5). Let $w_{ki}(\boldsymbol{\alpha}, \beta) = \left[\mathbf{I}_{12}(\boldsymbol{\alpha}, \beta)^\top \mathbf{I}_{11}(\boldsymbol{\alpha}, \beta)^{-1}\right]_{ki}$, the $(k,i)^{\text{th}}$ 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* $1 \le i \ne j \le n$, *suppose that* $\sup_{1 \le k \le K, 1 \le i \le n} |w_{ki}(\boldsymbol{\alpha}, \beta)|$ *is* $O(1)$ *and continuously differentiable in both arguments. Furthermore, the following conditions on* $w_{ki}(\boldsymbol{\alpha}, \beta)$ *are satisfied:*

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

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

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

Assumption 5 is mild under $\sup_{1\le k\le K, 1\le i\le n}|w_{ki}(\boldsymbol{\alpha},\beta)| = O(1)$. 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, then $w_{ki}(\boldsymbol{\alpha},\beta)$ depends only on α_i and β. Hence, $|\partial w_{ki}(\boldsymbol{\alpha},\beta)/\partial\alpha_j| = 0$, satisfying Assumption 5(c). Therefore, Assumption 5(c) controls the extent to which $w_{ki}(\boldsymbol{\alpha},\beta)$ depends on α_j for $j \neq i$.

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

Theorem 2. *If Assumptions 1–5 hold, then*

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

where I_0 *and* b_0 *are defined in (A39) and (A34), respectively.*

Theorem 2 shows $\hat{\beta}_{\text{OS}}$ achieves the CRLB asymptotically. In the proof of Theorem 2, 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 (6) uses information from $\mathbf{s}_1$ to concentrate out $\boldsymbol{\alpha}$. As a result, the covariance between $\mathbf{m}_1$ and $\mathbf{s}_1$ determines the magnitude of the term b_0 in the asymptotic bias of Theorem 2.

3.3 Bagging

While reaching the CRLB, Theorem 2 reveals that $\hat{\beta}_{\text{OS}}$ incurs an asymptotic bias. As discussed in Module 3, one way to debias $\hat{\beta}_{\text{OS}}$ is to use split-network jackknife to self-cancel the leading bias. However, it doubles the asymptotic variance (see (A42)–(A43)) and suffers from computational instability. The solution we propose is to use bagging on a split-network jackknife estimator.

To motivate the bagging method, in theory there are a total of $T_n := \binom{n}{n/2}$ possible ways to divide the network. However, T_n can be very large for a moderate sample size n. For example, $n = 100$ produces $T_n = \binom{100}{50} \simeq 1.009 \times 10^{29}$, which is an astronomical number. We solve this problem by choosing $\tilde{T}_n \ll T_n$ in the BG module. In the simulations, we set $\tilde{T}_n = 2n$ and find that the results are robust to this choice.

The next theorem shows that when n and $\tilde{T}_n$ go to infinity, $\hat{\beta}_{\text{BG}}$ is asymptotically normal, unbiased, and efficient.

Theorem 3. *If Assumptions 1–5 hold, then*

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

as $n \to \infty$ *and* $\tilde{T}_n \to \infty$.

Theorem 3 is the main theoretical result of this paper. A few remarks are in place to discuss its implications and connections with the literature. First, $\hat{\beta}_{\text{BG}}$ involves three modules—JMM, OS, BG—which play different roles. Module JMM provides an initial consistent yet biased estimator, which is fed into Module OS to achieve the CRLB. Module BG corrects for the bias in the OS estimator via split-network jackknife while maintaining the efficiency through bagging.

Second, a similar idea to $\hat{\beta}_{\text{OS-SJ}}$ in a panel setting with fixed effects is presented in Dhaene and Jochmans (2015). Although related, $\hat{\beta}_{\text{BG}}$ is preferred over $\hat{\beta}_{\text{OS-SJ}}$ because $\hat{\beta}_{\text{OS-SJ}}$ has an asymptotic variance of 2I_0^{-1} while $\hat{\beta}_{\text{BG}}$'s is I_0^{-1}.

Third, one may be inclined to apply BG to the initial JMM estimator directly and bypass the one-step approximation. Indeed, BG can correct for the asymptotic bias of the JMM estimator. However, the JMM-BG estimator is not efficient because the asymptotic variance of the initial JMM estimator is preserved through the bagging procedure.

Finally, sample splitting across individuals introduces a degree of extra randomness, which motivates Fernández-Val and Weidner (2016, Footnote 8) to suggest averaging of all possible T_n partitions and point out that the average over $\tilde{T}_n \ll T_n$ splits is sufficient. The BG estimator in our context not only eliminates randomness from sample splitting but also simultaneously achieves efficiency and bias correction. Furthermore, our Theorem 3 provides formal asymptotic results to justify the use of the BG estimator.

4 Extensions

We have established the asymptotic properties of $\hat{\beta}_{\text{BG}}$ for β_0. The homophily coefficient is interpretable and useful. For instance, it enables the comparison of the relative importance of covariates underlying the formation of informal risk-sharing networks in rural villages. Beyond the model parameters themselves, additional policy-relevant quantities—such as the APEs—can be derived. This section establishes their theoretical properties. Furthermore, the preceding results are derived under the assumption that the link function is correctly specified. We examine the consequences of misspecification, and thereby assessing the robustness of the proposed method.

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

$$\mathbb{E}\left[Y_{ij}\middle|\, X_{ij}=x_{ij},\boldsymbol{\alpha}\right]=F(x_{ij}^{\top}\beta_0+\alpha_i)F(x_{ij}^{\top}\beta_0+\alpha_j). \tag{9}$$

Here, the partial effects are defined as the differences or derivatives of (9) 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 the 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 APEs are

$$\delta_0=\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{n}\sum_{j>i}\Delta_{ij}(\alpha_i,\alpha_j,\beta_0)\right]. \tag{10}$$

Plugging the JMM estimator $(\widehat{\boldsymbol{\alpha}},\widehat{\beta})$ into (10) yields an estimator for the APEs

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

Define an (infeasible) $\bar{\Delta}_n=\frac{1}{N}\sum_{i=1}^{n}\sum_{j>i}\Delta_{ij}(\alpha_{i0},\alpha_{j0},\beta_0)$. Let the split-jackknife estimator and the bagging estimator of the APE be

$$\hat{\delta}_{\text{SJ}}:=2\hat{\delta}-\frac{1}{2}(\hat{\delta}_1+\hat{\delta}_2)\quad\text{and}\quad\hat{\delta}_{\text{BG}}:=\frac{1}{\tilde{T}_n}\sum_{t=1}^{\tilde{T}_n}\hat{\delta}_{\text{SJ}}^{(t)},$$

respectively. Here, $(\hat{\delta}_1,\hat{\delta}_2)$ are the plug-in estimators based on two sub-networks after a random split of the nodes and $\{\hat{\delta}_{\text{SJ}}^{(t)}\}_{t=1}^{\tilde{T}_n}$ are split-network jackknife estimators based on $\tilde{T}_n$ random splits. The next theorem shows that $\hat{\delta}$ is asymptotically unbiased. We use a central limit theorem for U-statistics (van der Vaart, 2000, Theorem 12.3) to prove it. To state the result precisely, we incorporate the asymptotically vanishing bias terms, as in Fernández-Val and Weidner (2016, Theorem 4.2), and establish that $\hat{\delta}_{\text{SJ}}$ and $\hat{\delta}_{\text{BG}}$ offer no meaningful improvement over $\hat{\delta}$. Sections 5 and 6 present numerical evidence that supports this claim.

To state the next theorem, we define $\sigma_{\delta,n}:=\frac{\Sigma_\Delta}{N}+\frac{4\Sigma_\delta}{n}$, where Σ_Δ is defined in (A54) and $\Sigma_\delta=\mathbb{E}\left[\Delta_{ij}(\alpha_i,\alpha_j,\beta_0)\Delta_{ik}(\alpha_i,\alpha_k,\beta_0)\right]$. Furthermore, let

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

where $\mathbf{R}_k^{\mu}$ for $k=1,\ldots,K$ and $(\Delta_{\boldsymbol{\alpha}},\Delta_\beta)$ are characterized in (A50) and (A47), respectively.

Theorem 4. *If Assumptions 1–4 hold, the sequence $\{\alpha_i\}_{i\in\mathcal{I}_n}$ is i.i.d., and $\bar{\Delta}_n$ is a non-*

degenerate U-statistic, then

$$\sigma_{\delta,n}^{-1/2}\left(\hat{\delta}-\delta_0-\frac{1}{\sqrt{N}}B_\beta-\frac{1}{\sqrt{N}}B_\alpha\right)\xrightarrow{d}\mathcal{N}(0,I_K),\ \textit{and} \tag{12}$$
$$\sigma_{\delta,n}^{-1/2}\left(\hat{\delta}_{\mathrm{BG}}-\delta_0\right)\xrightarrow{d}\mathcal{N}(0,I_K).$$

In Theorem 4, the rate of convergence of $\hat{\delta}$ (and $\hat{\delta}_{\mathrm{BG}}$) is $\sqrt{n}$ instead of $\sqrt{N}$. The slower convergence rate in (12) makes the bias terms introduced by estimating the individual fixed effects asymptotically negligible. Note that B_β originates from the bias of the plug-in estimator $\hat{\beta}$ whereas B_α arises from the incidental parameter bias of the 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 $\sqrt{n}(\bar{\Delta}_n-\delta_0)$. 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 (12), we find in simulations that including it improves the coverage probabilities.

Remark 4. If one is interested in $\bar{\Delta}_n$, the asymptotic result becomes

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

which generalizes Theorem 2 of Chen et al. (2021) to the NTU setting.

4.2 Link Function Misspecification

Our analysis so far relies on the correct specification of the link function $F(\cdot)$. A natural question is what if $F(\cdot)$ is misspecified? Graham (2024) provides an insightful analysis for sparse bipartite network models. However, this question has not been studied yet in the literature of dyadic network formation models with NTU. In this subsection, we present theoretical properties of our JMM, OS, and BG estimators under such misspecification.

First, we analyze the pseudo values that our estimators $\hat{\beta}$ and $\hat{\beta}_{\mathrm{OS}}$ converge to under misspecification of the link function. Suppose researchers misspecify the distribution function of ϵ_{ij} to be $G(\cdot)$ which differs from $F(\cdot)$ at points with strictly positive probability measure. For a fixed n, we impose the following identification assumption to facilitate the analysis. 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 Link Function Misspecification)**.** *For a fixed n, the nonlinear function*

$$\tilde{S}_n(\beta) := \sum_{i=1}^{n} \sum_{j>i} \left[p_{ij} - q_{ij}(\boldsymbol{\alpha}(\beta), \beta) \right] x_{ij} \tag{13}$$

has a unique root β_{n*}, and satisfies

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

for all $\delta > 0$ and sufficiently large n, where $\boldsymbol{\alpha}(\beta)$ is the unique solution to the following system of equations

$$\left(\textstyle\sum_{j \neq 1} p_{1j} - \sum_{j \neq 1} q_{1j}(\boldsymbol{\alpha}, \beta), \cdots, \sum_{j \neq n} p_{nj} - \sum_{j \neq n} q_{nj}(\boldsymbol{\alpha}, \beta) \right)^{\top} = 0. \tag{14}$$

Assumption 6 is the counterpart of Assumption 4 under a misspecified link function. Similarly to Lemma 1, (14) has a unique solution with high probability under mild conditions on $(\boldsymbol{\alpha}_0, \beta_0)$ and β. Thus, Assumption 6 identifies the homophily parameter under link function misspecification. Notice that β_{n*} depends on the true link function $F(\cdot)$, misspecified link function $G(\cdot)$, and the true parameter values. As a result, β_{n*} may vary with n. The following theorem shows that the JMM estimator based on the misspecified link function $G(\cdot)$ is centered at β_{n*} up to a bias, which the split-network jackknife procedure removes asymptotically. Let $\boldsymbol{\alpha}_* := \boldsymbol{\alpha}(\beta_{n*})$ with $\boldsymbol{\alpha}(\cdot)$ satisfying (14).

In the next Theorem, J_*, B_*, and $\boldsymbol{\Omega}_*$ are defined analogously to J_0, B_0, and Ω_0 in Theorem 1, but with the pseudo values $(\boldsymbol{\alpha}_*, \beta_{n*})$ and the misspecified link function $G(\cdot)$ replacing $(\boldsymbol{\alpha}_0, \beta_0)$ and $F(\cdot)$, respectively. Furthermore, the sandwich-form variance $\boldsymbol{\Omega}_*$ can be consistently estimated as in (A3).

Theorem 5 (JMM Estimation under Link Function Misspecification)**.** *If Assumptions 1–3 and 6 hold, then* $\sqrt{N}(\hat{\beta} - \beta_{n*}) - \mathrm{J}_*^{-1} B_* \xrightarrow{d} \mathcal{N}(0, \Omega_*)$.

Theorem 5 establishes that, if the researcher assumes the moment equations hold in population, the JMM estimator $\widehat{\beta}$ remains robust to link function misspecification. In particular, $\hat{\beta}$ is consistent for β_{n*}, the unique solution to the pseudo-population moment equations (13).

Under the link function misspecification, the one-step estimator becomes

$$\hat{\beta}_{\mathrm{OS}} := \hat{\beta} - \mathrm{H}(\widehat{\boldsymbol{\alpha}}, \hat{\beta})^{-1} s_n(\widehat{\boldsymbol{\alpha}}, \hat{\beta}), \tag{15}$$

with the JMM estimator $(\widehat{\boldsymbol{\alpha}}, \hat{\beta})$ substituted in. Note that

$$\mathrm{H}(\boldsymbol{\alpha}, \beta) := \mathrm{H}_{22}(\boldsymbol{\alpha}, \beta) - \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta)^{\top} \mathbf{H}_{22}(\boldsymbol{\alpha}, \beta)^{-1} \mathbf{H}_{12}(\boldsymbol{\alpha}, \beta), \tag{16}$$

is the concentrated Hessian matrix with the full expression in Appendix A.1, 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 link function misspecification, $\hat{\beta}_{\text{OS}}$ in (15) centers around

$$\beta_{n\sharp} := \beta_{n*} - \text{H}(\boldsymbol{\alpha}_*, \beta_{n*})^{-1} \mathbb{E} s_n(\boldsymbol{\alpha}_*, \beta_{n*}), \tag{17}$$

which can be seen as a projection of β_{n*} by concentrating out the fixed effects. When the link function is correctly specified, $(\boldsymbol{\alpha}_*, \beta_{n*}) \equiv (\boldsymbol{\alpha}_0, \beta_0)$, thus $\beta_{n\sharp} \equiv \beta_{n*} \equiv \beta_0$ because $\mathbb{E} s_n(\boldsymbol{\alpha}_*, \beta_{n*}) = \mathbb{E} s_n(\boldsymbol{\alpha}_0, \beta_0) \equiv 0$. Furthermore, our OS and BG estimators in the misspecified case share similar asymptotic properties from their counterparts when the link function is correctly specified, except that they now center around the projected pseudo value $\beta_{n\sharp}$ instead of β_0.

For the next theorem, let $\text{H}_* = \text{plim}_{n\to\infty} \text{H}(\boldsymbol{\alpha}_*, \beta_{n*})$, and define b_* similarly to b_0, but with $(\boldsymbol{\alpha}_*, \beta_{n*})$ and the misspecified link function $G(\cdot)$. The asymptotic covariance matrix Γ_* is

$$\Gamma_* := \lim_{n\to\infty} N^{-1} \text{H}_*^{-1} \begin{bmatrix} \text{I}_{22*} + \mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{11*} (\mathbf{H}_{11*}^{-1} \mathbf{H}_{12*}^\top)^\top \\ -\mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{12*} - (\mathbf{H}_{12*}^\top \mathbf{H}_{11*}^{-1} \mathbf{I}_{12*})^\top \end{bmatrix} (\text{H}_*^{-1})^\top. \tag{18}$$

Theorem 6 (OS and BG Estimation under Misspecified Link Function)**.** *Suppose all the bounds in Assumption 5 hold for each element of* $\mathbf{H}_{12}^\top \mathbf{H}_{11}^{-1}$*. If Assumptions 1–3 and 6 are satisfied, then*

$$\begin{aligned} \sqrt{N}(\hat{\beta}_{\text{OS}} - \beta_{n\sharp}) - \text{H}_*^{-1} b_* &\xrightarrow{d} \mathcal{N}(0, \Gamma_*) \text{ and} \\ \sqrt{N}(\hat{\beta}_{\text{BG}} - \beta_{n\sharp}) &\xrightarrow{d} \mathcal{N}(0, \Gamma_*). \end{aligned}$$

We point out that the limits of the variance term and Hessian term are functions of $(\boldsymbol{\alpha}_*, \beta_{n*})$ because $\beta_{n\sharp}$ is a function of $(\boldsymbol{\alpha}_*, \beta_{n*})$ by (17). Theorem 6 demonstrates that $\hat{\beta}_{\text{BG}}$ serves as a robust estimator for common parameters in the following sense. If the link function is correctly specified, $\hat{\beta}_{\text{BG}}$ centers around β_0 without bias and achieves the CRLB asymptotically. Otherwise, $\hat{\beta}_{\text{BG}}$ centers around a projected pseudo value with no asymptotic bias. Finally, we can estimate $\mathbf{I}_*$ by $\hat{\mathbf{I}} = \mathbf{s}(\hat{\boldsymbol{\alpha}}, \hat{\beta}) \mathbf{s}(\hat{\boldsymbol{\alpha}}, \hat{\beta})^\top$ and $\mathbf{H}_*$ by plugging $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ into (16), which together give a consistent estimator for Γ_* by (18).

To summarize, we extend our analysis beyond the core parameters of interest. We first establish that APEs can be consistently estimated, with the plug-in, SJ, and bagging procedures yielding asymptotically unbiased estimators whose slower $\sqrt{n}$-convergence rate renders the incidental parameter bias negligible. We then show that, under link function misspecifi-

cation, the JMM, OS, and BG estimators remain well-behaved by converging to pseudo-true values, with the bagging estimator in particular providing robustness and efficiency, thereby ensuring the practical relevance of the methodology in both well-specified and misspecified settings.

5 Monte Carlo Simulations

Building on the theoretical properties established above, we now turn to Monte Carlo simulations to evaluate the finite-sample performance of our estimators. We first assess the accuracy of the JMM, OS, and BG estimators for β_0, and then examine how well the method recovers the individual fixed effects $\boldsymbol{\alpha}_0$ and the APEs. Finally, we study the robustness of our estimators under link function misspecification and in sparser network settings.

The data generating process (DGP) is as follows. We set $\beta_0 = (1, -1)^\top$, and draw the first covariate of X_{ij} as $X_{1,ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.3)$, $X_{1,ij} = X_{1,ji}$. 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 all other variables so that α_i and X_{ij} are correlated via X_i. We independently draw the idiosyncratic shock to each dyad, ϵ_{ij}, from the standard logistic distribution, and obtain the outcome of each ij pair 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 data used in our empirical applications. To further investigate the performance of the estimator of the high-dimensional fixed effects, we also let $n = 500$ and $1,000$. We report the mean and median bias, standard deviation, mean and median absolute bias, and the root mean squared error (RMSE) across replications.

5.1 Main Estimation Results: β_0 and $\boldsymbol{\alpha}_0$

Table 1 reports the results of estimating the common parameter β_0 for $n = 100$ and 200 when the network has a density of 25%. Here are the main observations when $n = 100$. First, in terms of the bias, the BG estimator performs significantly better than JMM and OS, which is consistent with Theorem 3. Second, BG works very well in simultaneously achieving bias-correction and low standard deviation, leading to the lowest RMSE.[3] Third,

[3]Though not reported in the tables, we find that SJ (without BG) doubles the variance of JMM and OS estimators in the simulations, which is in line with the theory.

Table 1: Baseline estimation results for β_0

$n = 100$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	3.04	-2.88	2.91	-2.82	-0.26	0.28
Median Bias	3.02	-3.45	2.87	-3.61	-0.22	-0.34
Standard Deviation	5.94	13.49	5.91	13.52	5.73	13.18
Mean Standard Error	5.68	12.96	5.68	12.93	5.68	12.93
Mean Absolute Bias	5.38	11.15	5.32	11.16	4.59	10.58
Median Absolute Bias	4.53	9.31	4.37	9.38	4.01	9.25
RMSE	6.67	13.8	6.59	13.81	5.74	13.18
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
$n = 200$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	1.42	-1.71	1.37	-1.66	-0.17	-0.14
Median Bias	1.47	-2.03	1.37	-1.81	-0.19	-0.29
Standard Deviation	2.89	6.52	2.88	6.50	2.84	6.40
Mean Standard Error	2.78	6.36	2.78	6.34	2.78	6.34
Mean Absolute Bias	2.59	5.40	2.56	5.36	2.29	5.08
Median Absolute Bias	2.26	4.60	2.18	4.50	1.93	4.21
RMSE	3.22	6.74	3.19	6.70	2.85	6.40
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

Note: All values have been multiplied by 100.

the coverage probabilities of the confidence intervals constructed using the asymptotic distribution of each estimator are close to their nominal levels. Finally, the mean standard errors implied by the asymptotic theory are close to the standard deviations computed from the Monte Carlo simulations across all estimators. We also find that the quantiles of the empirical distributions for all estimators are well approximated by the same quantiles of the corresponding asymptotic normal distributions. These results further support our theoretical findings. When $n = 200$, the performance of all the estimators improve. The 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 also improve.

Given the large number of individual fixed effects, we plot in Figure 2 the histogram of $\widehat{\alpha}_i - \alpha_{i0}$ for $i \in \mathcal{I}_n$. All the histograms are well centered around 0. As n increases, the performance of $\widehat{\alpha}_i$ improves. Moreover, the range of estimation errors shrinks toward zero as the sample size increases, consistent with our theoretical predictions.

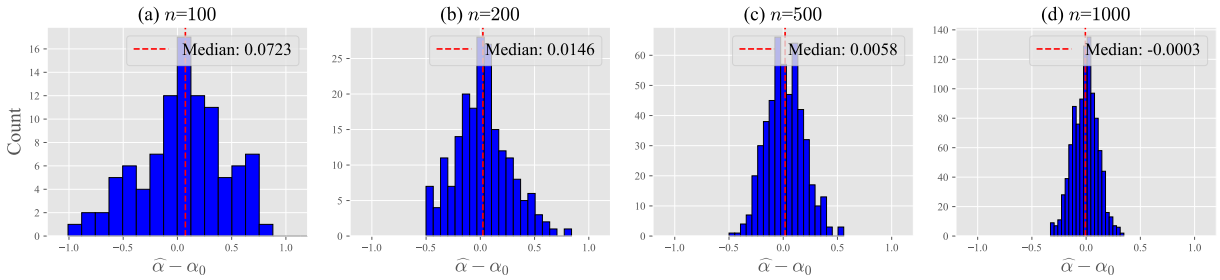


Figure 2: Histograms of $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0$ for different n

5.2 Extended Estimation Results: APEs, Link Function Misspecification, and Sparser Network

Table 2 summarizes the APEs for each coordinate of X_{ij} defined in (10). Our plug-in estimator performs well with respect to RMSE and coverage probabilities. When applied to the estimation of APEs, the split-network jackknife bagging method does not yield meaningful improvement. As predicted by Theorem 4, the asymptotic bias in estimating APEs is asymptotically negligible, of a smaller order than the bias in estimating β_0.

Table 2: Estimation results for the APEs

	$n = 100$				$n = 200$			
	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	-0.28	0.14	-0.06	-0.16	-0.14	0.07	-0.02	-0.06
Median Bias	-0.32	0.07	-0.06	-0.15	-0.14	0.14	-0.02	-0.01
Standard Deviation	1.41	2.73	1.44	2.79	0.75	1.42	0.77	1.45
Mean Standard Error	1.48	2.85	1.48	2.85	0.75	1.42	0.75	1.42
Mean Absolute Bias	1.14	2.21	1.14	2.26	0.61	1.12	0.62	1.14
Median Absolute Bias	0.97	1.87	0.91	1.92	0.52	0.93	0.52	0.97
RMSE	1.44	2.74	1.44	2.80	0.77	1.42	0.77	1.45
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: All values have been multiplied by 100; true values of APEs are calibrated from a simulation with $n = 10,000$ agents.

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 logistic link function in the estimation. We compare $\widehat{\beta}$ to the pseudo true value $\beta_{n\sharp}$ defined in (17) 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.

Table 3: Estimation results for β_0 under link function misspecification

$n = 100$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	6.11	-5.70	6.47	-5.50	-0.25	0.87
Median Bias	6.08	-5.68	6.48	-5.56	-0.25	0.83
Standard Deviation	6.34	15.10	6.36	15.00	6.09	14.41
Mean Standard Error	6.21	14.57	6.36	14.82	6.36	14.82
Mean Absolute Bias	7.22	12.79	7.46	12.66	4.84	11.49
Median Absolute Bias	6.43	11.02	6.62	10.60	4.11	9.80
RMSE	8.81	16.14	9.08	15.97	6.09	14.43
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
$n = 200$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	2.81	-2.94	2.96	-2.73	-0.17	0.14
Median Bias	2.90	-3.11	3.10	-2.86	-0.01	0.06
Standard Deviation	3.05	7.66	3.04	7.61	2.98	7.48
Mean Standard Error	3.04	7.48	3.06	7.51	3.06	7.51
Mean Absolute Bias	3.45	6.62	3.54	6.50	2.36	5.94
Median Absolute Bias	3.11	5.89	3.22	5.63	1.97	4.98
RMSE	4.15	8.20	4.24	8.09	2.98	7.48
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

Note: All values have been multiplied by 100.

Finally, we examine the performance of the method in networks with fewer links on average. To this end, we lower all α_i's by one, resulting in a network density of 8.6%. As reported in Table 4, network sparsity worsens the performance of all estimators. Nevertheless, the BG estimator continues to outperform the others across nearly all metrics.

6 Empirical Applications

This section presents two empirical applications. First, we apply our method to the Nyakatoke risk-sharing network dataset (De Weerdt, 2004). Our empirical findings complement Gao et al. (2023) by showing that wealth difference has no statistically significant

Table 4: Estimation results for β_0 under a sparser network

$n = 100$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	4.47	-8.25	5.17	-5.15	-0.23	0.18
Median Bias	4.46	-8.02	5.06	-4.94	-0.31	0.50
Standard Deviation	7.47	17.80	7.49	17.95	7.08	17.00
Mean Standard Error	7.40	17.91	7.42	17.93	7.42	17.93
Mean Absolute Bias	7.01	15.82	7.35	15.03	5.64	13.70
Median Absolute Bias	6.02	13.46	6.22	13.05	4.72	11.79
RMSE	8.70	19.62	9.10	18.67	7.09	17.00
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
$n = 200$	JMM		OS		BG	
	β_1	β_2	β_1	β_2	β_1	β_2
Mean Bias	1.89	-3.60	2.15	-2.21	-0.35	0.29
Median Bias	1.97	-3.64	2.21	-2.42	-0.25	0.07
Standard Deviation	3.73	8.69	3.71	8.71	3.63	8.50
Mean Standard Error	3.59	8.73	3.60	8.72	3.60	8.72
Mean Absolute Bias	3.35	7.57	3.46	7.22	2.90	6.79
Median Absolute Bias	2.88	6.51	3.01	6.28	2.48	5.82
RMSE	4.18	9.40	4.29	8.99	3.65	8.51
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

Note: All values have been multiplied by 100.

impact on the link formation. Second, we use the India microfinance network dataset (Banerjee et al., 2013) to study the influence of caste and various measures of wealth difference on forming an information and favor link between households. We find that while belonging to the same caste has a significantly positive effect, the relationship becomes more nuanced when considering wealth differences among households.

6.1 Nyakatoke Risk-Sharing Network

6.1.1 Data

The network data of Nyakatoke, located in the Kagera Region of Tanzania, covers a small Haya community of all 119 households. We investigate how important are wealth difference, distance, and blood or religious ties in deciding the formation of risk-sharing links among local residents. The dataset includes the following variables: (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. To define the dependent variable *link*, each household was asked:

"*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. In the context of the village economies, the risk-sharing links are unlikely to be driven by efficient arrangements of side-payment transfers, thereby satisfying NTU.

We estimate the coefficients for three regressors: *wealth difference*, *distance* and *tie* between households. *Wealth* is defined as the total assets in USD owned by each household, including livestocks, durables and land. *Distance* measures how far away two households are located in kilometers. *Tie* is a discrete variable, with the value "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. Following the literature we take natural logarithm on *wealth* and *distance*, and we construct the *wealth difference* variable as the absolute difference in *wealth*, i.e.,

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

Five households in the data have no information on *wealth* and/or *distance*. We drop these observations, resulting in a sample size $n = 114$, which creates a network data of $N = 12,882$ observations. Table 5 reports the summary statistics.

Table 5: Summary statistics for the Nyakatoke network

Variables	Mean	Std. Dev.	Min	Max
link	0.0732	0.2606	0	1
(ln) wealth difference	1.0365	0.8228	0.0004	5.8898
(ln) distance	6.0553	0.7092	2.6672	7.4603
tie	0.4260	0.6123	0	3

6.1.2 Results and Discussion

Table 6 presents the estimation results for the homophily coefficients and the APEs. The estimated coefficient for wealth difference is negative using all three methods. However, we cannot reject the null that it is zero based on the test using the asymptotic distribution of the BG estimator. To interpret this result, consider two scenarios. First, when two households have similar wealth levels, everything else being equal, they may still be reluctant to form a risk-sharing link since neither household would likely have sufficient capacity to insure the other against unpredictable shocks, such as natural disasters or severe illnesses. Second, consider households with substantial wealth differences. In this case, according to the link formation rule (7) under NTU, the linking decision is primarily driven by the wealthier household. Here, link formation is again unlikely because the richer household's expected surplus from the risk-sharing arrangement would typically be negative. Thus, the net effect of absolute wealth difference on link formation is expected to be close to zero. Clearly, the requirement for bilateral agreement to form a link under NTU plays a crucial role in both scenarios. Our estimates of the homophily coefficient for wealth difference align well with these explanations. While Gao et al. (2023) also obtain a negative coefficient for wealth difference, they do not provide inference results, given their emphasis on semiparametric identification. In contrast, our framework leverages the link function to conduct comprehensive inference, thereby quantifying the statistical uncertainty of the estimates. This contribution makes our paper complementary to theirs

In addition to the wealth difference, under BG the coefficient for *distance* is significantly negative at -0.8187, and that of *tie* is significantly positive at 0.5817. The results are intuitive. We further report the APEs in the last two columns of Table 6. We find that the APE of wealth difference is not significant based on either the plug-in or the bagging estimator. Distances between households and social ties, on the other hand, matter more significantly in terms of the APE.

Finally, we estimate the individual fixed effects α_i and plot their distribution in Figure 3. We find that most estimated fixed effects are in the range of $[2, 4]$, although some exceed this range, reflecting heterogeneity in unobserved household characteristics.

6.2 India Microfinance Network

6.2.1 Data

The Indian microfinance network dataset of Banerjee et al. (2013) is based on a detailed survey of villagers in India, which records their daily interactions and demographic

Table 6: Estimation results for the Nyakatoke network

Variables	Coefficients			APEs	
	JMM	OS	BG	Plug-in	Bagging
(ln) wealth difference	-0.0882	-0.0974	-0.0777	-0.0065	-0.0083
	(0.0676)	(0.0641)	(0.0641)	(0.0052)	(0.0052)
(ln) distance	-0.7824	-0.8636	-0.8187	-0.0576	-0.0641
	(0.0530)	(0.0536)	(0.0536)	(0.0065)	(0.0065)
Tie	0.6714	0.6287	0.5817	0.0514	0.0501
	(0.0546)	(0.0556)	(0.0556)	(0.0059)	(0.0059)

Note: Standard errors are reported in the parentheses.

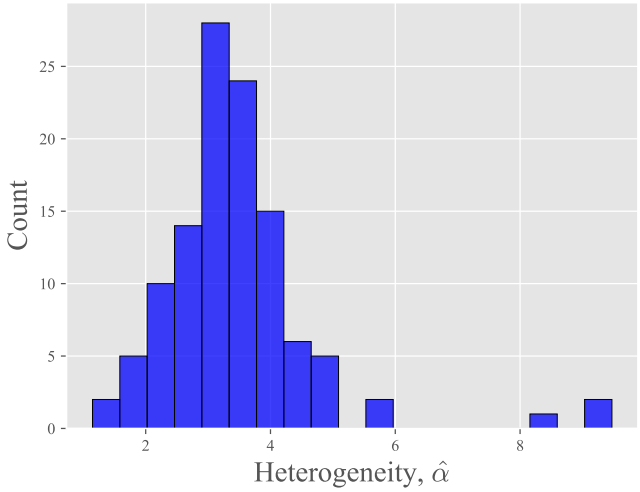


Figure 3: Histogram of $\widehat{\boldsymbol{\alpha}}$ for the Nyakatoke network

information such as caste, family size, and wealth. The survey covers 89.14% of the 16,476 households across 75 villages. Village sizes vary, with an average of 220 households. This yields a total of $N = \sum_{r=1}^{75} \binom{n_r}{2} = 1,238,970$ links in the full sample.

For the dependent variable, we follow Chandrasekhar and Jackson (2025) and consider two types of links: an "information link," defined when two households exchange advice, and a "favor link," defined when they borrow or lend material goods. Both are binary variables corresponding to Y_{ij} in (2). As for the covariates, we use six dyadic variables that are constructed based on the demographics of each household. The first set of covariates is binary, taking the value 1 if two households share the same characteristics and 0 otherwise. These binary characteristics include (i) what caste group the household belongs to, (ii) whether the household has access to electricity, (iii) what type of latrine the household uses,

and (iv) whether the household owns or rents a house. The second set of covariates includes the absolute differences in the number of beds and rooms between pairs of households. Summary statistics for these variables are reported in Table 7.

Table 7: Summary statistics for the Indian networks

Variable	Mean	Std. Dev.	Min	Max
Information link (Dependent Variable 1)	0.0330	0.1787	0	1
Favor link (Dependent Variable 2)	0.0388	0.1932	0	1
Same caste	0.4828	0.4997	0	1
Same electricity	0.5244	0.4994	0	1
Same latrine	0.6201	0.4854	0	1
Same ownrent	0.8488	0.3583	0	1
Bed number difference	1.0371	1.4439	0	50
Room number difference	1.2789	1.2898	0	18

6.2.2 Estimation Results

We estimate $\boldsymbol{\alpha}_0$ and β_0 for each of the 75 villages, for both the information and favor networks, using our BG estimator. Figures 4 and 5 present histograms of the village-level t-statistics, computed from the BG estimator $\widehat{\beta}_{\text{BG}}$ and compared with their asymptotic distribution, for the information and favor networks, respectively. Figures 4 and 5 show the distributions of the estimated fixed effects $\widehat{\alpha}_i$ for each network.

We highlight several findings. For the first four binary covariates labeled with "same," which capture whether two households share a characteristic, the BG estimates $\widehat{\beta}_{\text{BG}}$are generally significantly positive in both networks. Most corresponding t-statistics exceed 1.645, the 95^{th} percentile of the standard normal distribution. Intuitively, households with the same caste, access to electricity, latrine, or housing tenure are more likely to be linked. By contrast, the BG estimates for the last two discrete covariates—differences in the number of beds and rooms—are negative in both networks, indicating that greater differences in these characteristics reduce the likelihood of connection. Moreover, the t-statistics for "same caste" are much larger in absolute value than those for the other covariates, showing that caste plays a crucial role in link formation in Indian villages. Finally, Figures 6a and 6b reveal substantial heterogeneity in the distributions of the estimated fixed effects $\widehat{\alpha}_i$ across individuals, underscoring the importance of incorporating fixed effects in the model.

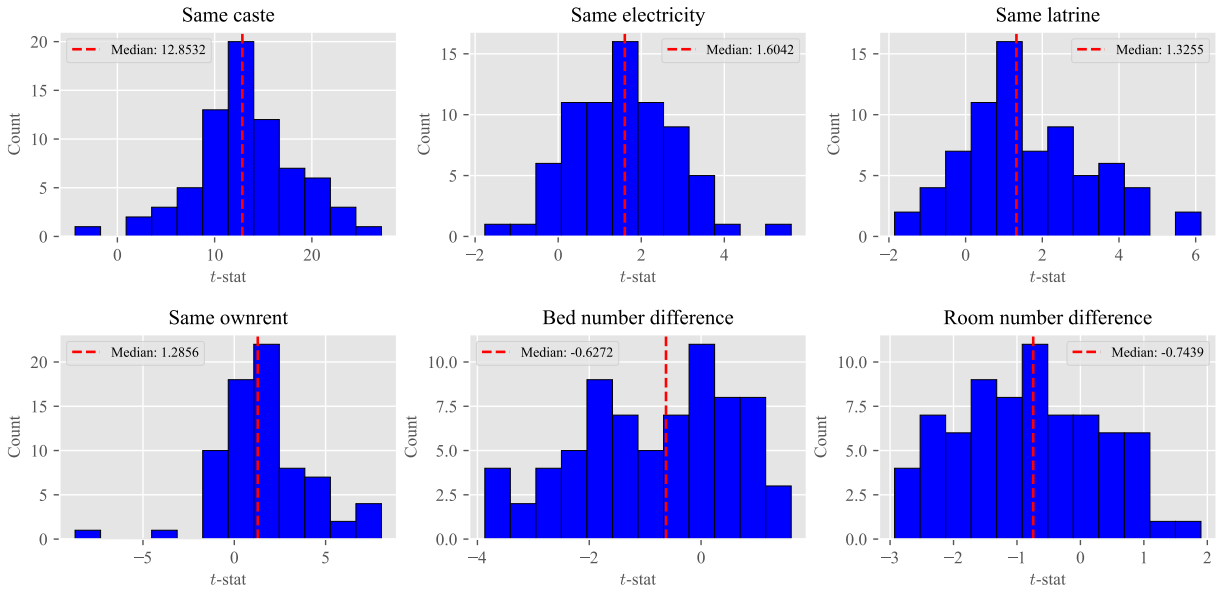


Figure 4: Histograms of t-statistics for the information network

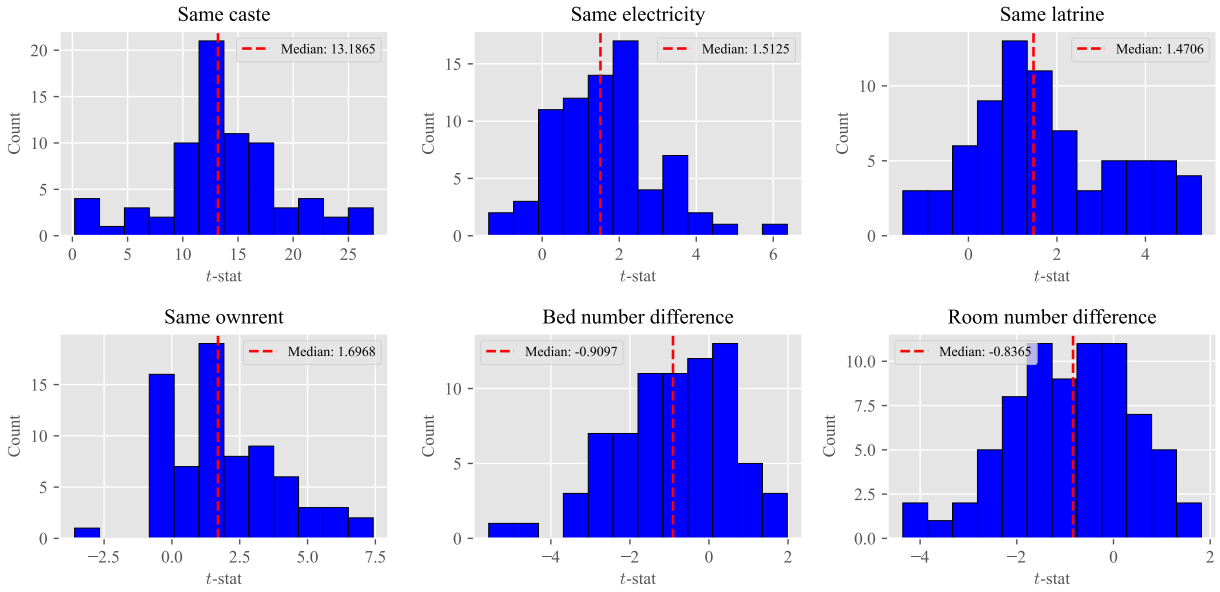
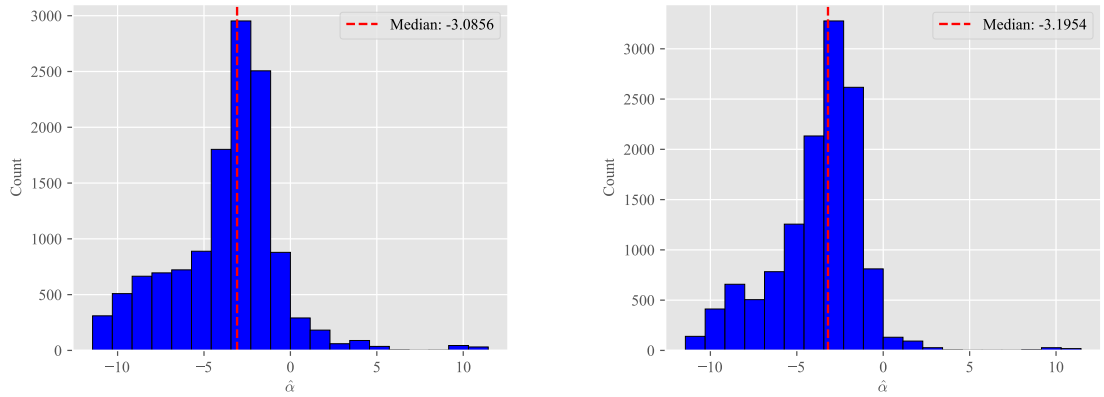


Figure 5: Histograms of t-statistics for the favor network



(a) Information network (b) Favor network

Figure 6: Histograms of $\widehat{\alpha}_i$

7 Conclusion

In this paper, we propose a bagging estimator for the homophily coefficients in a dyadic network formation model. The estimator is asymptotically efficient, a property implied by Le Cam (1969)'s result on the one-step approximation to the ML estimator, and it is unbiased owing to the use of bagging in the split-network jackknife procedure. We also estimate the high-dimensional individual fixed effects and establish their consistency. In addition, we extend the framework to study the average partial effects (APEs) and link function misspecification. Extensive simulations show that the estimators perform well under various settings. Finally, two empirical applications—the Nyakatoke risk-sharing network and the Indian microfinance network—demonstrate the practical relevance of our approach.

This paper serves as a stepping stone toward more flexible models. Our theory currently relies on an additive specification of the utility surplus for each individual (e.g., $\alpha_{i0} + x_{ij}^{\top}\beta_0$) and correct specification of the link function distribution $F(\cdot)$. Relaxing these assumptions would enhance robustness, and the insights from the sieve MLE literature (Shen, 1997; Chen, 2007) could be useful. Moreover, as noted in the introduction, our focus on dyadic link formation under NTU excludes interdependence in link preferences. An important direction for future work is to develop tests for this assumption in dyadic network formation models.

References

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): "Diffusion of Microfinance," *Science*, 341, 1236498.

Bonhomme, S. (2012): "Functional Differencing," *Econometrica*, 80, 1337–1385.

Bonhomme, S. and K. Dano (2024): "Functional Differencing in Networks," *Revue économique*, 75, 147–175.

Bonhomme, S., K. Jochmans, and M. Weidner (2024): "A Neyman-Orthogonalization Approach to the Incidental Parameter Problem," arXiv preprint arXiv:2412.10304.

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.

Candelaria, L. E. and Y. Zhang (2024): "Robust Inference in Locally Misspecified Bipartite Networks," arXiv preprint arXiv:2403.13725.

Chandrasekhar, A. G. and M. O. Jackson (2025): "A Network Formation Model Based on Subgraphs," *Review of Economic Studies*, rdaf013.

Chatterjee, S., P. Diaconis, and A. Sly (2011): "Random Graphs with Given Degree Sequence," *Annals of Applied Probability*, 21, 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, vol. 6, 5549–5632.

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. (2020): "Econometric Models of Network Formation," *Annual Review of Economics*, 12, 775–799.

de Paula, Á., S. Richards-Shubik, and E. Tamer (2018): "Identifying Preferences in Networks with Bounded Degree," *Econometrica*, 86, 263–288.

De Weerdt, J. (2004): "Risk-Sharing and Endogenous Network Formation," in *Insurance Against Poverty*, Oxford University Press.

Dhaene, G. and K. Jochmans (2015): "Split-Panel Jackknife Estimation of Fixed-Effect Models," *Review of Economic Studies*, 82, 991–1030.

Dzemski, A. (2019): "An Empirical Model of Dyadic Link Formation in 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 and Economic Statistics*, 31, 253–264.

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.

Hahn, J. and G. Kuersteiner (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27, 1152–1191.

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.

Hsieh, C.-S. and L. F. Lee (2016): "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. O., Z. Lin, and N. N. Yu (2024): "Adjusting for Peer-Influence in Propensity Scoring when Estimating Treatment Effects," *Available at SSRN 3522256*.

Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71, 44–74.

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.

Le Cam, L. M. (1969): *Théorie Asymptotique de la Décision Statistique*, Presses de l'Université de Montréal.

Liao, C., Z. Mei, and Z. Shi (2024): "Nickell Meets Stambaugh: A Tale of Two Biases in Panel Predictive Regressions," *arXiv preprint arXiv:2410.09825*.

Mei, Z., L. Sheng, and Z. Shi (2023): "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.

Pelican, A. and B. S. Graham (2024): "An Optimal Test for Strategic Interaction in Network Formation Games," Working Paper.

Qu, L., L. Chen, T. Yan, and Y. Chen (2025): "Inference in Semiparametric Formation Models for Directed Networks," *Journal of Business & Economic Statistics*, 0, 1–15.

Rao, C. R. (1992): "Information and the Accuracy Attainable in the Estimation of Statistical Parameters," in *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer, 235–247.

Shen, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics*, 25, 2555–2591.

Sheng, S. (2020): "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," Available at 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.

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., H. Qin, and H. Wang (2016): "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," Working Paper.

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. We write "$a_n \asymp b_n$" to denote $a_n = O(b_n)$ and $b_n = O(a_n)$ simultaneously. We use $C_1, C_2, \ldots$ to represent strictly positive and finite constants.

A.1 Definitions of Matrices in the Main Text

Let the Jacobian matrix of $\mathbf{m}(\boldsymbol{\alpha}, \beta)$ be

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

where we separate it into four blocks according to the variables of differentiation. In Appendix A, we provide explicit expressions for these blocks. It is worth emphasizing that $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta) \neq \mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)^\top$ and $\mathbf{J}_{12}(\boldsymbol{\alpha}, \beta) \neq \mathbf{J}_{21}(\boldsymbol{\alpha}, \beta)^\top$, and thus $\mathbf{J}(\boldsymbol{\alpha}, \beta)$ is asymmetric. The consequence of the asymmetry is 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 the unknown parameter $(\boldsymbol{\alpha}, \beta)$ because, as a covariance matrix, the demeaning operation cancels them out. Define

$$B_{k0} = \lim_{n \to \infty} \frac{1}{2\sqrt{N}} \mathrm{Tr}\left[\mathbf{J}_{11}^{-1} \mathbf{V}_{11} \left(\mathbf{J}_{11}^{-1} \right)^\top \mathbf{R}_k \right] \tag{A1}$$

where $\mathbf{R}_k$ is defined by (A25) in the Appendix B. Let $B_0 = (B_{10}, \ldots, B_{K0})^\top$ and the limiting variance matrix be

$$\Omega_0 := \lim_{n \to \infty} N^{-1} \mathrm{J}_0^{-1} \left[\mathrm{V}_{22} + \mathbf{J}_{21} \mathbf{J}_{11}^{-1} \mathbf{V}_{11} (\mathbf{J}_{21} \mathbf{J}_{11}^{-1})^\top - \mathbf{J}_{21} \mathbf{J}_{11}^{-1} \mathbf{V}_{12} - (\mathbf{J}_{21} \mathbf{J}_{11}^{-1} \mathbf{V}_{12})^\top \right] (\mathrm{J}_0^{-1})^\top, \tag{A2}$$

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 $\hat{\beta} - \beta_0$ reduces to $\mathrm{J}_{22}^{-1}\mathrm{V}_{22}\mathrm{J}_{22}^{-1}$, and the additional terms in (A2) arise from estimating $\boldsymbol{\alpha}_0$.

Next, we derive the explicit expressions of the matrices introduced above.

Jacobian matrix. First, $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ is an $n \times n$ matrix whose off-diagonal and diagonal elements are given by

$$\begin{aligned}
[\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)]_{ij} &= -F_{ij}(\boldsymbol{\alpha}, \beta) f_{ji}(\boldsymbol{\alpha}, \beta), \qquad 1 \leq i \neq j \leq n \text{ and} \\
[\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} &= -\sum_{j \neq i} f_{ij}(\boldsymbol{\alpha}, \beta) F_{ji}(\boldsymbol{\alpha}, \beta), \qquad i \in \mathcal{I}_n,
\end{aligned}$$

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

$$[\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} = \sum_{j \neq i} [\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)]_{ji}, \qquad i \in \mathcal{I}_n.$$

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 in Lemma A2. Next, $\mathbf{J}_{12}(\boldsymbol{\alpha}, \beta)$ is an $n \times K$ matrix whose ith is

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

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

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

is a $K \times K$ matrix.

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

$$\begin{aligned}
[\mathbf{V}_{11}(\boldsymbol{\alpha}, \beta)]_{ij} &= p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)), \qquad 1 \leq i \neq j \leq n \text{ and} \\
[\mathbf{V}_{11}(\boldsymbol{\alpha}, \beta)]_{ii} &= \sum_{j \neq i} p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ji}(\boldsymbol{\alpha}, \beta)), \qquad i \in \mathcal{I}_n,
\end{aligned}$$

respectively. $\mathbf{V}_{12}$ is an $n \times K$ matrix whose ith row is $\sum_{j \neq i} p_{ij}(\boldsymbol{\alpha}, \beta)(1 - p_{ij}(\boldsymbol{\alpha}, \beta)) x_{ij}^\top$. 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:

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

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

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

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

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

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

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

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

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

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

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

In what follows, we apply the mean value theorem for vector-valued functions in its integral form, as in Chatterjee et al. (2011). For example,

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

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, and 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 property of being diagonally dominant from $\mathbf{J}_{11}(\boldsymbol{\alpha} + t(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \beta)$.

Sandwich form covariance matrix estimator under link function misspecification. Statistical inference for the JMM estimator under link function misspecification requires an estimator of the limiting 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 $\hat{\mathbf{J}}_*$. Recall that J_* is the concentrated Jacobian matrix for β_{n*}. Then, we propose to estimate Ω_* by

$$\hat{\Omega}_* := N^{-1} \hat{\mathrm{J}}_*^{-1} \left[\hat{\mathrm{V}}_{22*} + \hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{11*} (\hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{J}}_{21*})^{\top} - \hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{12*} - (\hat{\mathbf{J}}_{21*} \hat{\mathbf{J}}_{11*}^{-1} \hat{\mathbf{V}}_{12*})^{\top} \right] (\hat{\mathrm{J}}_*^{-1})^{\top}, \tag{A3}$$

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

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

We adapt Theorem 1 of Yan (2019) to the NTU framework to analytically approximate the inverse of the Jacobian matrix $\mathbf{J}_{11}(\boldsymbol{\alpha}, \beta)$ and bound the approximation errors. Similar

techniques have been used to prove asymptotic normality in network estimation problems; see, for example, Yan and Xu (2013), Graham (2017), and Yan et al. (2019). 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 A1 (Yan, 2019)**.** *Suppose an $n \times n$ matrix $\mathbf{A} = (a_{ij})_{n\times n}$ is invertible with all its entries positive and $a_{ii} \geq \sum_{j\neq i} a_{ji}$. Let $\mathbf{B} = [\text{diag}(a_{11}, a_{22}, \ldots, a_{nn})]^{-1}$, $\Delta_i = a_{ii} - \sum_{j\neq i} a_{ji}$, $M \equiv \max\{\max_{1\leq i\neq j\leq n} a_{ij}, \max_{i=1,\ldots,n} \Delta_i\}$, and $m \equiv \min_{1\leq i\neq j\leq n} a_{ij}$. If $M \asymp 1$ and $m \asymp 1$, then we have*

$$\|\mathbf{A}^{-1} - \mathbf{B}\|_{\max} = O(n^{-2}). \tag{A4}$$

Proof. The proof proceeds along the lines of Yan (2019). 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

$$\mathbf{F} = \mathbf{A}^{-1} - \mathbf{B} = (\mathbf{A}^{-1} - \mathbf{B})(I_n - \mathbf{AB}) + \mathbf{B}(I_n - \mathbf{AB}) = \mathbf{FU} + \mathbf{W}. \tag{A5}$$

Some algebra leads to

$$u_{ij} = \delta_{ij} - \sum_{k=1}^{n} a_{ik}b_{kj} = \delta_{ij} - \sum_{k=1}^{n} a_{ik}\frac{\delta_{kj}}{a_{jj}} = \delta_{ij} - \frac{a_{ij}}{a_{jj}} = (\delta_{ij} - 1)\frac{a_{ij}}{a_{jj}}, \tag{A6}$$

and

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

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

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

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

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

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

It follows that

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

We use (A5) to obtain a bound for the approximate error $\|\mathbf{F}\|_{\max}$. By (A5) and (A6), for any $i \leq n$, we have

$$f_{ij} = \sum_{k=1}^{n} f_{ik} u_{kj} + w_{ij} = \sum_{k=1}^{n} f_{ik}(\delta_{kj} - 1)\frac{a_{kj}}{a_{jj}} + w_{ij}. \tag{A7}$$

Define $f_{i\theta} = \max_{1\leq k\leq n} f_{ik}$ and $f_{i\xi} = \min_{1\leq k\leq n} f_{ik}$. 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, $f_{i\xi} \sum_{k=1}^{n} a_{ki} \leq \sum_{k=1}^{n} f_{ik} a_{ki} = 0$. So, we have $f_{i\xi} < 0$. Similarly, we have that $f_{i\theta} > 0$. Recall that

$$a_{\theta\theta} = \sum_{k\neq\theta} a_{k\theta} + \Delta_\theta = \sum_{k=1}^{n} (1 - \delta_{k\theta}) a_{k\theta} + \Delta_\theta, \text{ hence, } 1 \equiv \sum_{k=1}^{n} (1 - \delta_{k\theta}) \frac{a_{k\theta}}{a_{\theta\theta}} + \frac{\Delta_\theta}{a_{\theta\theta}} \tag{A8}$$

for any θ, which yields the following identities

$$\begin{aligned} f_{i\xi} =& f_{i\xi} \left[\sum_{k=1}^{n} (1 - \delta_{k\theta}) \frac{a_{k\theta}}{a_{\theta\theta}} + \frac{\Delta_\theta}{a_{\theta\theta}} \right] = \sum_{k=1}^{n} f_{i\xi} (1 - \delta_{k\theta}) \frac{a_{k\theta}}{a_{\theta\theta}} + \frac{f_{i\xi}\Delta_\theta}{a_{\theta\theta}}, \\ f_{i\xi} =& f_{i\xi} \left[\sum_{k=1}^{n} (1 - \delta_{k\xi}) \frac{a_{k\xi}}{a_{\xi\xi}} + \frac{\Delta_\xi}{a_{\xi\xi}} \right] = \sum_{k=1}^{n} f_{i\xi} (1 - \delta_{k\xi}) \frac{a_{k\xi}}{a_{\xi\xi}} + \frac{f_{i\xi}\Delta_\xi}{a_{\xi\xi}}, \end{aligned} \tag{A9}$$

where the first and second part of this equation use (A8) for $a_{\theta\theta}$ and $a_{\xi\xi}$, respectively.

By combining (A7) with the first part of (A9) where we set $j = \theta$ in (A7), we have

$$f_{i\theta} + f_{i\xi} = \sum_{k=1}^{n} (f_{i\xi} - f_{ik})(1 - \delta_{k\theta}) \frac{a_{k\theta}}{a_{\theta\theta}} + w_{i\theta} + \frac{f_{i\xi}\Delta_\theta}{a_{\theta\theta}}. \tag{A10}$$

Similarly, we have

$$2 f_{i\xi} = \sum_{k=1}^{n} (f_{i\xi} - f_{ik})(1 - \delta_{k\xi}) \frac{a_{k\xi}}{a_{\xi\xi}} + w_{i\xi} + \frac{f_{i\xi}\Delta_\xi}{a_{\xi\xi}}. \tag{A11}$$

Subtracting (A11) from (A10), we have

$$f_{i\theta} - f_{i\xi} = \sum_{k=1}^{n} (f_{ik} - f_{i\xi}) \left[(1 - \delta_{k\xi}) \frac{a_{k\xi}}{a_{\xi\xi}} - (1 - \delta_{k\theta}) \frac{a_{k\theta}}{a_{\theta\theta}} \right] + w_{i\theta} - w_{i\xi} + f_{i\xi} \left(\frac{\Delta_\theta}{a_{\theta\theta}} - \frac{\Delta_\xi}{a_{\xi\xi}} \right). \tag{A12}$$

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. By the fact 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 $1 \le \lambda \le n-1$, and

$$
\begin{aligned}
&\sum_{k=1}^{n}(f_{ik}-f_{i\xi})\left[(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}}-(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}}\right] \\
\le &\sum_{k\in\Omega}(f_{ik}-f_{i\xi})\left[(1-\delta_{k\xi})\frac{a_{k\xi}}{a_{\xi\xi}}-(1-\delta_{k\theta})\frac{a_{k\theta}}{a_{\theta\theta}}\right] \\
\le &\,(f_{i\theta}-f_{i\xi})\left[\frac{\sum_{k\in\Omega}a_{k\xi}}{a_{\xi\xi}}-\frac{\sum_{k\in\Omega}(1-\delta_{k\theta})a_{k\theta}}{a_{\theta\theta}}\right] \\
\le &\,(f_{i\theta}-f_{i\xi})\left[\frac{\lambda M}{\lambda M+(n-1-\lambda)m}-\frac{(\lambda-1)m}{(\lambda-1)m+(n-\lambda+1)M}\right] \\
\le &\,(f_{i\theta}-f_{i\xi})\left\{\frac{nM-(n-2)m}{nM+(n-2)m}+\frac{(n-2)Mm}{[(n-2)m+M][(n-2)m+2M]}\right\},
\end{aligned}
\tag{A13}
$$

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

$$
f_{i\xi}\left(\frac{\Delta_\theta}{a_{\theta\theta}}-\frac{\Delta_\xi}{a_{\xi\xi}}\right)\le(f_{i\theta}-f_{i\xi})\frac{2M}{m(n-1)}. \tag{A14}
$$

Combining (A12), (A13), and (A14), 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 $\max_{k=1,\ldots,n}|f_{ik}| \le f_{i\theta}-f_{i\xi} = O(n^{-2})$ as $m, M \asymp 1$. Hence, we have shown $\|\mathbf{A}^{-1}-\mathbf{B}\|_{\max}=\|\mathbf{F}\|_{\max}=O(n^{-2})$. □

Based on Lemma A1, 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 A2. *If Assumptions 2 and 3 hold, 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 A2. 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 A1 demonstrates that its inverse can be approximated by $\operatorname{diag}\left([\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{11}^{-1},\ldots,[\mathbf{J}_{11}(\boldsymbol{\alpha},\beta)]_{n-1n-1}^{-1}\right)$ 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

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

Thus, we have

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

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

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

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

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 Lemma A1. Let $\mathbf{T}(\boldsymbol{\alpha},\beta)=[\operatorname{diag}(\mathbf{J}_{11}(\boldsymbol{\alpha},\beta))]^{-1}$. Applying Lemma A1 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 be applied to $\mathbf{J}_{11}^{\circ}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})$, where $\mathbf{T}^{\circ}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})$ denotes the diagonal approximation for $[\mathbf{J}_{11}^{\circ}(\hat{\boldsymbol{\alpha}},\boldsymbol{\alpha})]^{-1}$.

A.3 Deviation Bound

We derive several useful non-asymptotic deviation bounds in this subsection. The following probabilities are defined conditional on $\boldsymbol{\alpha}$ and $\mathbf{x}$; for brevity, we suppress its conditioning whenever it is clear from the context. Lemma A3 provides a bound on the deviation of the weighted sum of centered Bernoulli random variables, $\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})$. This result is used extensively in the proof.

Let $\{\lambda_{ij}\}_{i,j=1}^{n}$ denote a sequence of bounded constants that satisfy $\max_{i,j}|\lambda_{ij}| < C_1$.

Lemma A3. *If Assumptions 2 and 3 hold, , then we have*

$$\Pr\left\{\max_{1\le i\le n}\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right| > C_1\sqrt{\frac{6\log n}{n-1}}\right\} \le 2n^{-2}. \tag{A15}$$

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 et al., 2013) for the sum of bounded and independent random variables, we have

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

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

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

By Boole's inequality,

$$\Pr\left(\max_{1\le i\le n}\frac{1}{n-1}\left|\sum_{j\neq i}\lambda_{ij}(y_{ij}-p_{ij})\right| > C_1\sqrt{\frac{6\log n}{n-1}}\right) \le n\cdot 2n^{-3} = 2n^{-2}. \tag{A16}$$

We complete the proof. □

Using Lemma A3, we can bound the estimation error of $\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0$, which guarantees that our moment estimator is consistent for $\boldsymbol{\alpha}_0$ when β_0 is known. This result can be strengthened to prove the second part of Theorem 1, which we do in Appendix B.

Lemma A4. *If Assumptions 2 and 3 hold, then we have*

$$\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{n}\sum_{j>i}\lambda_{ij}(y_{ij}-p_{ij})\right| > C_1\sqrt{\frac{2\log N}{N}}\right\} \le (n(n-1))^{-1}. \tag{A17}$$

Proof. Similar to the proof of Lemma A3, by Hoeffding's inequality, we have

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

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

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

Lemma A5. *If Assumptions 2 and 3 hold, 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

$$\left\|\sqrt{n}[\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0] + \left(\frac{\mathbf{J}_{11}}{n}\right)^{-1} \frac{\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)}{\sqrt{n}}\right\|_\infty = O\left(\frac{\log n}{\sqrt{n}}\right). \tag{A18}$$

Proof. The rest of proof is conditional on the following event, which happens with probability at least $1 - 2n^{-2}$ by Lemma (A3):

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

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

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

which implies that

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

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 $[\mathbf{J}_{11}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)]^{-1}$ is $\mathbf{T}^\circ(\widehat{\boldsymbol{\alpha}}(\beta_0), \boldsymbol{\alpha}_0)$. By Lemma A1, we decompose $\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0$ into two parts and apply the triangle inequality:

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

We 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 A3, we obtain

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

For the second part, by Lemma A1, we have

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

Hence, we have

$$\begin{aligned}&\|\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta_0),\boldsymbol{\alpha}_0)-\mathbf{T}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta_0),\boldsymbol{\alpha}_0)\|_{\infty}\|\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\|_{\infty}\\ &=O(n^{-1})\cdot\max_{1\le i\le n}\left|\sum_{j\neq i}(y_{ij}-p_{ij})\right|=O\left(\sqrt{\frac{\log n}{n}}\right).\end{aligned}$$

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 (A18). By the second-order Taylor expansion, which is also used in the proof of Lemma 6 of Graham (2017), we have

$$\begin{aligned}&\mathbf{m}_1(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)-\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\\ &=\mathbf{J}_{11}(\boldsymbol{\alpha}_0,\beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]+\frac{1}{2}\left[\sum_{k=1}^{n}(\hat{\alpha}_k(\beta_0)-\alpha_{k0})\frac{\partial\mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}}^k,\beta_0)}{\partial\alpha_k}\right][\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\end{aligned}\tag{A19}$$

with the mean value $\tilde{\boldsymbol{\alpha}}^k$ lying between $\widehat{\boldsymbol{\alpha}}(\beta_0)$ and $\boldsymbol{\alpha}_0$ and possibly varying 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 write 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 $\Lambda=\sum_{k=1}^{n}(\hat{\alpha}_k(\beta_0)-\alpha_{k0})\frac{\partial\mathbf{J}_{11}(\tilde{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_k}$ whose 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

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

Then, 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, by (A19), we have

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

B Proofs of Main Results

In this section, we prove Lemma 1 and Theorems 1–6.

B.1 Proof of Lemma 1

Before proving Lemma 1, we state a different version of Lemma 2.1 of Chatterjee et al. (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 A6. *If $\mathbf{A}, \mathbf{B} \in \mathcal{G}_n(\delta)$, we have*

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

Proof. This is equivalent to proving if $A, B \in \mathcal{G}_n(\delta)$, then $\|\mathbf{B}^\top\mathbf{A}^\top\|_\infty \leq 1 - \frac{2(n-2)}{n-1}\delta^2$, which is a direct application of Lemma 2.1 of Chatterjee et al. (2011). $\square$

We prove Lemma 1 based on this lemma.

Proof of Lemma 1. First, suppose a solution to $\mathbf{m}_1(\boldsymbol{\alpha}, \beta) = 0$ exists. 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 $i \neq j$, $\partial r_j/\partial \alpha_i = -(n-1)f_{ij}(\boldsymbol{\alpha}, \beta)F_{ji}(\boldsymbol{\alpha}, \beta) < 0$; while for each i, $\partial r_i/\partial \alpha_i = 1 - (n-1)\sum_{j\neq i} f_{ij}(\boldsymbol{\alpha}, \beta)F_{ji}(\boldsymbol{\alpha}, \beta) > 0$. Moreover, for each i,

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

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 $f_{ij}(\boldsymbol{\alpha}, \beta)F_{ji}(\boldsymbol{\alpha}, \beta) \in [C_1C_2, (1-C_1)(1-C_2)]$, which implies

$$\frac{\partial r_j}{\partial \alpha_i} \leq -\frac{C_1C_2}{n-1}, \quad \text{and} \quad \frac{\partial r_i}{\partial \alpha_i} \geq C_1 + C_2 - C_1C_2 \geq C_1C_2,$$

where the last inequality holds because $C_1 + C_2 \geq 2\sqrt{C_1C_2} \geq 2C_2C_2$ provided $C_1, C_2 \leq 1/2$. So, if we choose $\delta_1 = C_1C_2$ ($\leq 1/4$), it follows 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 (4), after every two updates, we have

$$\begin{aligned}
\|\boldsymbol{\alpha}^{k+2}(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1 &= \|\mathbf{r}(\mathbf{r}(\boldsymbol{\alpha}^k(\beta))) - \mathbf{r}(\mathbf{r}(\widehat{\boldsymbol{\alpha}}(\beta)))\|_1 \\
&= \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)) - \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \\
&\leq \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))\mathbf{G}^\circ(\boldsymbol{\alpha}^k(\beta), \widehat{\boldsymbol{\alpha}}(\beta))(\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \\
&\leq \|\mathbf{G}^\circ(\mathbf{r}(\boldsymbol{\alpha}^k(\beta)), \widehat{\boldsymbol{\alpha}}(\beta))\mathbf{G}^\circ(\boldsymbol{\alpha}^k(\beta), \widehat{\boldsymbol{\alpha}}(\beta))\|_1 \|\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1 \\
&\leq \left(1 - \frac{2(n-2)}{n-1}\delta_1^2\right)\|\boldsymbol{\alpha}^k(\beta) - \widehat{\boldsymbol{\alpha}}(\beta)\|_1,
\end{aligned}$$

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 A6. We write $\delta := 1 - \frac{2(n-2)}{n-1}\delta_1^2$, and the second inequality of Lemma 1 follows.

The proof of the first inequality is identical to the argument above and is 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 $\widehat{\boldsymbol{\alpha}}(\beta) \in \mathbb{A}$, the solution is unique. Now we show the existence of the solution,

where the main technique is adapted from Yan et al. (2016) and Yan et al. (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 et al. (2016), in a convex subset $\mathbb{D} \subset \mathbb{A}$ that contains $\boldsymbol{\alpha}_0$, to obtain the existence of the solution 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 A1. 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 A3 and the Lipschitz continuity of $F(\cdot)$ under Assumption 3. Then, by an application of Proposition A.1 of Yan et al. (2016), we have $\lim_{k\to\infty} \boldsymbol{\alpha}^{(k)}$ exists and the limit equals $\widehat{\boldsymbol{\alpha}}(\beta)$ if $\|\beta - \beta_0\|_2 < c$ for some constant $c > 0$. □

B.2 Proof of Theorem 1

We separate the proof into two parts: the first establishes consistency, and the second establishes asymptotic normality.

B.2.1 Consistency

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 to $S_n(\beta)=0$ 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 A7. *If Assumptions 1–4 hold, 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,

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

By an integral type mean-value theorem, we have

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

Recall that $\mathbf{J}_{21}(\boldsymbol{\alpha},\beta):=\frac{\partial m_2(\boldsymbol{\alpha},\beta)}{\partial\boldsymbol{\alpha}^\top}$. Therefore, we have

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

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 to $\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}(\beta),\boldsymbol{\alpha}(\beta))$ by Lemma A1.

R_1 is of order $O_p(\sqrt{(\log N)/N})$ by Lemma A4 and the fact that x_{ij} is bounded. 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 A3, 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 A1 to bound it as

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

Moreover, all these bounds hold uniformly in β, thereby completing the proof. □

Proof of the consistency part of Theorem 1. By the definitions of $\hat{\beta}$ and β_0, we have $S_n(\hat{\beta}) = 0$ and $\bar{S}_n(\beta_0) = 0$. Combining this fact with Lemma A7, we have

$$\left\|\bar{S}_n(\hat{\beta})\right\|_2 = \left\|\bar{S}_n(\hat{\beta}) - S_n(\hat{\beta})\right\|_2 \leq \sup_{\beta \in \mathbb{B}} \left\|S_n(\beta) - \bar{S}_n(\beta)\right\|_2 \xrightarrow{p} 0. \tag{A20}$$

Fix $\delta > 0$. By Assumption 4, 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 (A20).

We turn to the proof of the convergence of $\widehat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}_0$ in the ℓ_{∞} norm. By the integral type mean-value theorem, we have

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = & -\left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) \\
= & -\left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0) - \left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\right]^{-1} \left(\mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\right)
\end{aligned} \tag{A21}$$

Following the proof of Lemma A5, we have $\left\|\left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\right]^{-1}\right\|_{\infty} = O\left(n^{-1}\right)$ and $\|\mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_{\infty} = O_p\left(\sqrt{n \log n}\right)$, hence $\left\|\left[\mathbf{J}_{11}^{\circ}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\right]^{-1} \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\right\|_{\infty} \xrightarrow{p} 0$. 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

$$\begin{aligned}
& \|\mathbf{m}_1(\boldsymbol{\alpha}_0, \hat{\beta}) - \mathbf{m}_1(\boldsymbol{\alpha}_0, \beta_0)\|_{\infty} \\
= & \max_{1 \leq i \leq n} \left|\sum_{j \neq i} \left[p_{ij}(\boldsymbol{\alpha}_0, \hat{\beta}) - p_{ij}(\boldsymbol{\alpha}_0, \beta_0)\right]\right|
\end{aligned}$$

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

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. □

B.2.2 Asymptotic Normality

Before we prove the asymptotic normality part of Theorem 1, we characterize the limit of the concentrated Jacobian matrices. By Lemma A1,

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

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

$$N^{-1}\mathrm{J}_n(\bar{\beta}) \xrightarrow{p} \mathrm{J}_0 := \mathrm{plim}_{n\to\infty} N^{-1}\mathrm{J}_n(\beta_0) \tag{A22}$$

for any $\bar{\beta}$ lies between $\hat{\beta}$ and β_0. The existence of J_0 is guaranteed by the identification Assumption 1.

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

Proof of asymptotic normality part of Theorem 1. 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 the mean-value between $\hat{\beta}$ and β_0. By $m_n(\hat{\beta}) = 0$, we obtain

$$\begin{aligned} \sqrt{N}(\hat{\beta} - \beta_0) &= -\left[N^{-1}\mathrm{J}_n(\bar{\beta})\right]^{-1}\frac{1}{\sqrt{N}} m_2(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0) \\ &= -\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\left[y_{ij} - p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta_0), \beta_0)\right] x_{ij}\right\} + o_p(1) \end{aligned} \tag{A23}$$

in view of (A22). Note that we cannot directly apply standard central limit theorem (CLT) to the term in the curly bracket of (A23) because of the existence of $\widehat{\boldsymbol{\alpha}}(\beta_0)$. By a third-order

Taylor expansion of $\widehat{\boldsymbol{\alpha}}(\beta_0)$ around $\boldsymbol{\alpha}_0$, we have

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\left[y_{ij}-p_{ij}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right]x_{ij}\\
&=\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0)+\frac{1}{\sqrt{N}}\mathbf{J}_{21}(\boldsymbol{\alpha}_0,\beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\\
&\quad+\frac{1}{2}\left\{-\frac{1}{\sqrt{N}}\sum_{k=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_k\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}_n(\beta_0)-\boldsymbol{\alpha}_0]x_{ij}\right\}\\
&\quad+\frac{1}{6}\left\{-\frac{1}{\sqrt{N}}\sum_{k=1}^{n}\sum_{l=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}][\hat{\alpha}_l(\beta_0)-\alpha_{l0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^3 p_{ij}(\bar{\boldsymbol{\alpha}}_n,\beta_0)}{\partial\alpha_k\partial\alpha_l\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}_n(\beta_0)-\boldsymbol{\alpha}_{n0}]x_{ij}\right\}\\
&=:\ (I)+(II)+(III)+(IV)\,.
\end{aligned}
\tag{A24}
$$

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

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

By Lemma A5, $\sup_i|\hat{\alpha}_i(\beta_0)-\alpha_{i0}|=O_p(\sqrt{(\log n)/n})$. 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 algebra, its kth entry, which involves several third-order derivatives, equals

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

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

$$
\begin{aligned}
(\mathbf{R}_k)_{ij}&=\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_i\partial\alpha_j}x_{ij,k},\ 1<i\neq j<n,\\
(\mathbf{R}_k)_{ii}&=\sum_{j\neq i}\frac{\partial^2 p_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial^2\alpha_i}x_{ij,k},\ i=1,\ldots,n.
\end{aligned}
\tag{A25}
$$

Recalling the definition of B_{k0} in (A1), it follows that $(III)=-B_0+o_p(1)$.

We directly substitute the rest of terms, (I) and (II), into (A23) and obtain

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}-\beta_0)-\mathrm{J}_0^{-1}B_0 \\
=&-\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0)+\frac{1}{\sqrt{N}}\mathbf{J}_{21}(\boldsymbol{\alpha}_0,\beta_0)[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\right\}+o_p(1) \\
=&-\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0)-\frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\}+o_p(1).
\end{aligned} \tag{A26}$$

To apply the CLT to the first two terms of (A26), we verify the Lindeberg condition. Define

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}\xi_{ij}:=&-\mathrm{J}_0^{-1}\frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{n}\sum_{j>i}(y_{ij}-p_{ij})x_{ij}-\frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\} \\
=&-\mathrm{J}_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\sum_{j>i}(y_{ij}-p_{ij})\tilde{x}_{ij},
\end{aligned} \tag{A27}$$

where $\tilde{x}_{ij}$ represents the sum of the two multipliers to $(y_{ij}-p_{ij})$. Hence, (A27) 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 (A2). □

B.3 Proof of Theorem 2

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

$$s_n(\boldsymbol{\alpha},\beta)=s_2(\boldsymbol{\alpha},\beta)-\sum_{i=1}^{n}s_{1i}(\boldsymbol{\alpha},\beta)w_i(\boldsymbol{\alpha},\beta), \tag{A28}$$

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 A8. *If Assumptions 1–5 hold, for any $\tilde{\beta}$ such that $\|\tilde{\beta}-\beta_0\|_2 = O_p(N^{-1/2})$, we have*

$$\frac{1}{\sqrt{N}}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\,[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0] \xrightarrow{p} b_0, \tag{A29}$$

and

$$\frac{1}{N}\nabla_{\beta^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial\widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0 \xrightarrow{p} 0, \tag{A30}$$

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

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

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

where $z_{ij} = \frac{f_{ij}^{(1)}F_{ij}(1-p_{ij})-f_{ij}^2(1-2p_{ij})+F_{ij}^2 f_{ij}f_{ji}}{F_{ij}^2(1-p_{ij})^2}$. 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 A5. Hence, by the continuous mapping theorem (CMT) and the fact that $\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\|_\infty = o_p(1)$, we have

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

Similarly, we 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

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\beta_0)^{-1}\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right\|_{\max} \\
\leq\ &\frac{1}{\sqrt{N}}\left\|\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)+\mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\right\|_{\max} \\
&+\frac{1}{\sqrt{N}}\left\|\mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}\left[\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)+\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\right]\right\|_{\max} = o_p(1).
\end{aligned} \tag{A31}$$

By Lemma A5, 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 $\|\mathbf{c}\|_2 = 1$,

$$\sqrt{n}\mathbf{c}^\top(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0) = O_p(1). \tag{A32}$$

Combining (A31) and (A32) yields

$$\frac{1}{\sqrt{N}}[\mathbf{H}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^\top - \mathbf{I}_{12}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)\mathbf{I}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)^{-1}\mathbf{H}_{11}(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0)](\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0) = o_p(1). \tag{A33}$$

Next, similarly to the process of finding the bias term in the proof of Theorem 1, 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 by $\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0\|_\infty = o_p(1)$ and the CMT.

The asymptotic bias $b_0 := (b_{10},\ldots,b_{K0})^\top$ for the one-step estimator is defined as

$$b_{k0} = \lim_{n\to\infty}\frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k], \quad k=1,\ldots,K, \tag{A34}$$

where the entries of the $n\times n$ covariance matrix $\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)$ are

$$\begin{aligned}
[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ij} =&\mathbb{E}\left[\left(\sum_{k\neq i}(y_{ik}-p_{ik})\right)\left(\sum_{k\neq j}\frac{f_{jk}(y_{jk}-p_{jk})}{F_{jk}(1-p_{jk})}\right)\right]\\
=&\frac{f_{ji}\mathrm{Var}(y_{ij})}{F_{ji}(1-p_{ij})} = f_{ji}F_{ij}, \quad 1\le i\neq j\le n, \qquad \text{(A35)}\\
[\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)]_{ii} =&\mathbb{E}\left[\left(\sum_{k\neq i}(y_{ik}-p_{ik})\right)\left(\sum_{k\neq i}\frac{f_{ik}(y_{ik}-p_{ik})}{F_{ik}(1-p_{ik})}\right)\right]\\
=&\sum_{k\neq i}\frac{f_{ik}\mathrm{Var}(y_{ik})}{F_{ik}(1-p_{ik})} = \sum_{k\neq i}f_{ik}F_{ki}, \quad 1\le i\le n.
\end{aligned}$$

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 (A35). 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 A1, $\mathbf{J}_{11}^{-1}$ 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, uniformly for all i, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii}\\
=&\frac{1}{\sqrt{N}}[\mathbf{T}\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii}+\frac{1}{\sqrt{N}}[(\mathbf{J}_{11}^{-1}-\mathbf{T})\mathrm{Cov}(\mathbf{m}_1,\mathbf{s}_1)\mathbf{W}_k]_{ii}
\end{aligned}$$

$$= O(n^{-1} \cdot n^{-1}) \cdot [\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k]_{ii} + O(n^{-1} \cdot n^{-2}) \cdot [\mathbf{1}_n \mathbf{1}_n^\top \mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k]_{ii}$$
$$= O(n^{-2}) \cdot O(n) + O(n^{-3}) \cdot O(n^2) = O(n^{-1}).$$

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

$$\begin{aligned} &-\frac{1}{\sqrt{N}}\mathbf{s}_1^\top \mathbf{W}_k \mathbf{J}_{11}^{-1}\mathbf{m}_1 = \frac{1}{\sqrt{N}}\mathrm{Tr}(\mathbf{J}_{11}^{-1}\mathbf{m}_1\mathbf{s}_1^\top \mathbf{W}_k) \\ &= \frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k] + \left\{\frac{1}{\sqrt{N}}\mathrm{Tr}(\mathbf{J}_{11}^{-1}\mathbf{m}_1\mathbf{s}_1^\top \mathbf{W}_k) - \frac{1}{\sqrt{N}}\mathrm{Tr}[\mathbf{J}_{11}^{-1}\mathrm{Cov}(\mathbf{m}_1, \mathbf{s}_1)\mathbf{W}_k]\right\} \\ &= R_1 + R_2. \end{aligned} \tag{A36}$$

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 (A33) and (A36), 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 (A29).

We now turn to prove (A30). 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 (A21) 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 A4 and we have

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

Hence,

$$\frac{1}{N}\nabla_{\beta^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}), \bar{\beta}) + \mathrm{I}_0 = o_p(1). \tag{A37}$$

Finally, by (A31) and an argument identical to the proof of (A29), we have

$$\frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial\widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} = \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^{-1}\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) = o_p(1). \quad \text{(A38)}$$

Combining (A37) and (A38) completes the proof of (A30). □

By directly applying this lemma, we establish the asymptotic normality of $\hat{\beta}_{\mathrm{OS}}$.

Proof of Theorem 2. 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 JMM estimator $\widehat{\boldsymbol{\alpha}} := \widehat{\boldsymbol{\alpha}}(\hat{\beta})$ and $\widehat{\beta}$. Similar to the proof of Theorem 1, we apply 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 $\boldsymbol{\alpha}_0$. Together with the fact that

$$N^{-1}\mathrm{I}_n(\widehat{\boldsymbol{\alpha}},\hat{\beta}) \xrightarrow{p} \mathrm{I}_0, \quad \text{(A39)}$$

we obtain

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}_{\mathrm{OS}} - \beta_0) \\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1} s_n(\widehat{\boldsymbol{\alpha}}(\beta_0),\beta_0) \\
&\quad + \mathrm{I}_0^{-1}\left[\frac{1}{N}\nabla_{\beta^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial\widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0\right]\sqrt{N}(\hat{\beta} - \beta_0) + o_p(1) \quad \text{(A40)} \\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1} s_n(\boldsymbol{\alpha}_0,\beta_0) + \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}\nabla_{\boldsymbol{\alpha}^\top} s_n(\bar{\boldsymbol{\alpha}},\beta_0)(\widehat{\boldsymbol{\alpha}}(\beta_0) - \boldsymbol{\alpha}_0) \\
&\quad + \mathrm{I}_0^{-1}\left[\frac{1}{N}\nabla_{\beta^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta}) + \frac{1}{N}\nabla_{\boldsymbol{\alpha}^\top} s_n(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\frac{\partial\widehat{\boldsymbol{\alpha}}(\bar{\beta})}{\partial\beta^\top} + \mathrm{I}_0\right]\sqrt{N}(\hat{\beta} - \beta_0) + o_p(1).
\end{aligned}$$

By (A29) of Lemma A8, 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 (A30) of Lemma A8,

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

Hence, using the result that $\sqrt{N}(\hat{\beta} - \beta_0) = O_p(1)$ by Theorem 1, we simplify (A40) as

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{\mathrm{OS}} - \beta_0) &= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1} s_n(\boldsymbol{\alpha}_0,\beta_0) + \mathrm{I}_0^{-1} b_0 + o_p(1) \\
&= \frac{1}{\sqrt{N}}\mathrm{I}_0^{-1}\sum_{i=1}^{n}\sum_{j>i} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \mathrm{I}_0^{-1} b_0 + o_p(1), \quad \text{(A41)}
\end{aligned}$$

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 Theorem 1, we have the stated asymptotic normality. □

B.4 Proof of Theorem 3

Proof. We define $\hat{\beta}_{T_n} = \frac{1}{T_n}\sum_{s=1}^{T_n}\hat{\beta}^{(s)}_{\text{OS}-\text{SJ}}$ as the average of all possible OS-SJ estimators. Let $\mathcal{F}_n$ be the σ-algebra generated by all observed information. It is clear that $\hat{\beta}_{T_n}$ is $\mathcal{F}_n$–measurable. Let $\mathbb{E}^*$ represent the expectation over randomness from the random splits conditional on $\mathcal{F}_n$. Our proof has contains two immediate results: (i) $\sqrt{N}(\hat{\beta}_{T_n} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_0^{-1})$ as $n \to \infty$; (ii) $\sqrt{N}(\hat{\beta}_{\text{BG}} - \hat{\beta}_{T_n}) \xrightarrow{p} 0$ as $n \to \infty$ and $\tilde{T}_n \to \infty$. Theorem 3 follows by a combination of these two results.

Step (i). By (A41), we have

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

where $\mathcal{R}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0)$ is a residual term of order $o_p(1)$, as shown in the proof of Lemma A8. Thus, the one-step estimators based on sub-networks are

$$\sqrt{N/4}(\hat{\beta}^{(t)}_{\text{OS},1} - \beta_0) = \mathrm{I}_0^{-1}b_0 + \frac{1}{\sqrt{N/4}}\mathrm{I}_0^{-1}\sum_{(i,j)\in\mathcal{I}^{(t)}_{1,n}\times\mathcal{I}^{(t)}_{1,n};j>i} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \mathcal{R}(\mathbf{y}^{(t)}_1,\mathbf{x}^{(t)}_1,\boldsymbol{\alpha}^{(t)}_{0,1}), \quad \text{(A42)}$$

$$\sqrt{N/4}(\hat{\beta}^{(t)}_{\text{OS},2} - \beta_0) = \mathrm{I}_0^{-1}b_0 + \frac{1}{\sqrt{N/4}}\mathrm{I}_0^{-1}\sum_{(i,j)\in\mathcal{I}^{(t)}_{2,n}\times\mathcal{I}^{(t)}_{2,n};j>i} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \mathcal{R}(\mathbf{y}^{(t)}_2,\mathbf{x}^{(t)}_2,\boldsymbol{\alpha}^{(t)}_{0,2}), \quad \text{(A43)}$$

where $\boldsymbol{\alpha}^{(t)}_{0,1}$ is the sub-vector of $\boldsymbol{\alpha}_0$ indexed by $\mathcal{I}^{(t)}_{1,n}$ and similarly for $\boldsymbol{\alpha}^{(t)}_{0,2}$. . Hence, we have

$$\sqrt{N}(\hat{\beta}^{(t)}_{\text{OS}-\text{SJ}} - \beta_0) = \frac{2}{\sqrt{N}}\mathrm{I}_0^{-1}\sum_{(i,j)\in\mathcal{I}^{(t)}_{1,n}\times\mathcal{I}^{(t)}_{2,n}} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \mathcal{R}^{(t)}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0), \quad \text{(A44)}$$

with $\mathcal{R}^{(t)}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0) := 2\mathcal{R}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0) - \left[\mathcal{R}(\mathbf{y}^{(t)}_1,\mathbf{x}^{(t)}_1,\boldsymbol{\alpha}^{(t)}_{0,1}) + \mathcal{R}(\mathbf{y}^{(t)}_2,\mathbf{x}^{(t)}_2,\boldsymbol{\alpha}^{(t)}_{0,2})\right]/2$, 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 implies $\max_{1\le t\le T_n}\left\|\mathcal{R}^{(t)}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0)\right\|_2 \le \mathcal{R}^*(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0)$ 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 (A44) over all $1 \le t \le T_n$ yields

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{T_n} - \beta_0) &= \frac{2}{\sqrt{N}}\mathrm{I}_0^{-1}\frac{1}{T_n}\sum_{t=1}^{T_n}\sum_{(i,j)\in\mathcal{I}^{(t)}_{1,n}\times\mathcal{I}^{(t)}_{2,n}} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + \frac{1}{T_n}\sum_{t=1}^{T_n}\mathcal{R}^{(t)}(\mathbf{y},\mathbf{x},\boldsymbol{\alpha}_0) \\
&= \frac{2}{\sqrt{N}}\mathrm{I}_0^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\binom{n-2}{n/2-1}}{\binom{n}{n/2}}s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + o_p(1) \\
&= \frac{1}{\sqrt{N}} \times \frac{n}{n-1}\mathrm{I}_0^{-1}\sum_{i=1}^{n}\sum_{j>i} s_{ij}(\boldsymbol{\alpha}_0,\beta_0) + o_p(1) \xrightarrow{d} \mathcal{N}(0,\mathrm{I}_0^{-1}), \quad \text{(A45)}
\end{aligned}$$

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

Step (ii). Conditional on $\mathcal{F}_n$, random draws $\hat{\beta}^{(t)}_{\text{OS-SJ}}$ for $t = 1, \ldots, \tilde{T}_n$ are independent and uniformly distributed over $\{\hat{\beta}^{(s)}_{\text{OS-SJ}}\}_{s=1}^{T_n}$. So that $\mathbb{E}^*[\hat{\beta}^{(t)}_{\text{OS-SJ}}] = \hat{\beta}_{T_n}$ and $\mathbb{E}^*[\|\hat{\beta}^{(t)}_{\text{OS-SJ}} - \hat{\beta}_{T_n}\|_2^2] = \frac{1}{T_n}\sum_{s=1}^{T_n} \|\hat{\beta}^{(s)}_{\text{OS-SJ}} - \hat{\beta}_{T_n}\|_2^2 := \sigma_n^2$. Note that σ_n^2 is $\mathcal{F}_n$-measurable and $\sigma_n^2 = O_p(N^{-1})$ by (A44) and (A45). By Markov's inequality, for any $\epsilon > 0$

$$\Pr(\|\sqrt{N}(\hat{\beta}_{\text{BG}} - \hat{\beta}_{T_n})\|_2 \geq \epsilon | \mathcal{F}_n) \leq \frac{N\mathbb{E}^*[\|\hat{\beta}_{\text{BG}} - \hat{\beta}_{T_n}\|_2^2]}{\epsilon^2} = \frac{N\sigma_n^2}{\tilde{T}_n \epsilon^2}.$$

Hence, $\Pr(\|\sqrt{N}(\hat{\beta}_{\text{BG}} - \hat{\beta}_{T_n})\|_2 \geq \epsilon) \leq \mathbb{E}[\tilde{T}_n^{-1}\epsilon^{-2}N\sigma_n^2] = O(\tilde{T}_n^{-1}) = o(1)$ as $n \to \infty$ and $\tilde{T}_n \to \infty$ for any $\epsilon > 0$. This proves the second result.

Combining Step (i) and Step (ii), we have established that

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

as $\tilde{T}_n \to \infty$ and $n \to \infty$. □

B.5 Proofs of Results in Section 4

Proof of Theorem 4. 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

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

with kernel $\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0) - \mathbb{E}\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0)$. So, if $\Sigma_\delta = \text{Cov}(\Delta_{ij}(\alpha_{i0}, \alpha_{j0}, \beta_0), \Delta_{ik}(\alpha_{i0}, \alpha_{k0}, \beta_0))$ exists, by Theorem 12.3 of van der Vaart (2000), we have

$$\sqrt{n}(\bar{\Delta}_n - \delta_0) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_\delta). \tag{A46}$$

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

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

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

Define

$$\Delta_\beta(\boldsymbol{\alpha},\beta) := \frac{1}{N}\sum_{i=1}^{n}\sum_{j>i}\frac{\partial\Delta_{ij}}{\partial\beta}(\alpha_i,\alpha_j,\beta), \qquad \Delta_{\boldsymbol{\alpha}}(\boldsymbol{\alpha},\beta) := \frac{1}{N}\begin{pmatrix}\sum_{j\neq 1}\frac{\partial\Delta_{1j}}{\partial\alpha_1}(\alpha_1,\alpha_j,\beta)\\ \vdots \\ \sum_{j\neq n}\frac{\partial\Delta_{nj}}{\partial\alpha_n}(\alpha_n,\alpha_j,\beta)\end{pmatrix}. \tag{A47}$$

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

$$\begin{aligned} U_1 &= \frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{n}\sum_{j>i}\frac{\partial\Delta_{ij}}{\partial\beta^\top}(\hat{\alpha}_i(\bar{\beta}),\hat{\alpha}_j(\bar{\beta}),\bar{\beta}) + \sum_{i=1}^{n}\sum_{j\neq i}\frac{\partial\Delta_{ij}}{\partial\alpha_i}(\hat{\alpha}_i(\bar{\beta}),\hat{\alpha}_j(\bar{\beta}),\bar{\beta})\frac{\partial\hat{\alpha}_i}{\partial\beta^\top}(\bar{\beta})\right\}(\hat{\beta}-\beta_0)\\ &= \left\{\Delta_\beta(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^\top - \Delta_{\boldsymbol{\alpha}}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^\top\mathbf{J}_{11}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})^{-1}\mathbf{J}_{12}(\widehat{\boldsymbol{\alpha}}(\bar{\beta}),\bar{\beta})\right\}\sqrt{N}(\hat{\beta}-\beta_0)\\ &= (\Delta_\beta^\top - \Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\sqrt{N}(\hat{\beta}-\beta_0)+o_p(1), \end{aligned} \tag{A48}$$

where $\bar{\beta}$ lies in the segment between $\hat{\beta}$ and β_0 and 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

$$\begin{aligned} U_2 = {}& \sqrt{N}\Delta_{\boldsymbol{\alpha}}^\top(\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0)\\ &+\frac{1}{2}\left\{\frac{1}{\sqrt{N}}\sum_{k=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^2\Delta_{ij}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_k\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0]\right\}\\ &+\frac{1}{6}\left\{\frac{1}{\sqrt{N}}\sum_{k=1}^{n}\sum_{l=1}^{n}[\hat{\alpha}_k(\beta_0)-\alpha_{k0}][\hat{\alpha}_l(\beta_0)-\alpha_{l0}]\sum_{i=1}^{n}\sum_{j>i}\frac{\partial^3\Delta_{ij}(\bar{\boldsymbol{\alpha}},\beta_0)}{\partial\alpha_k\partial\alpha_l\partial\boldsymbol{\alpha}^\top}[\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_{n0}]\right\}. \end{aligned} \tag{A49}$$

Similarly to the proof of Theorem 1, we can show that the second term of (A49) converges in probability to a bias term B_α defined by (11) and

$$\begin{aligned} (\mathbf{R}_k^\mu)_{ij} &= \frac{\partial^2\Delta_{ij,k}(\boldsymbol{\alpha}_0,\beta_0)}{\partial\alpha_i\partial\alpha_j},\ 1<i\neq j<n,\\ (\mathbf{R}_k^\mu)_{ii} &= \sum_{j\neq i}\frac{\partial^2\Delta_{ij,k}(\boldsymbol{\alpha}_0,\beta_0)}{\partial^2\alpha_i},\ i=1,\ldots,n. \end{aligned} \tag{A50}$$

The last term of (A49) is $o_p(1)$ (equivalent to the limit of part (IV) of (A24)). Additionally, from the proof of Theorem 1, we have

$$\sqrt{N}(\hat{\beta}-\beta_0)-\mathrm{J}_0^{-1}B_0 = -\mathrm{J}_0^{-1}\left\{\frac{1}{\sqrt{N}}m_2(\boldsymbol{\alpha}_0,\beta_0)-\frac{1}{\sqrt{N}}\mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\}+o_p(1) \tag{A51}$$

and

$$\left\|\widehat{\boldsymbol{\alpha}}(\beta_0)-\boldsymbol{\alpha}_0+\mathbf{J}_{11}^{-1}\mathbf{m}_1(\boldsymbol{\alpha}_0,\beta_0)\right\|_\infty = o_p(n^{-1/2}). \tag{A52}$$

Substituting (A51) and (A52) into (A48) and (A49) respectively, we have

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

where B_β is defined in (11).

Finally, by the Lindeberg-Feller CLT, we have

$$\sqrt{N}(\hat{\delta}-\bar{\Delta}_n)-B_\beta-B_\alpha\xrightarrow{d}\mathcal{N}(0,\Sigma_\Delta) \tag{A53}$$

with

$$\begin{aligned}
&\Sigma_\Delta \\
=&\lim_{n\to\infty}\frac{1}{N}\Big\{(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathrm{V}_{22}\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top \\
&+\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21}-N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{11}\left\{\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21}-N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\right\}^\top \\
&-\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21}-N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{12}\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top \\
&-\left\{\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\mathbf{J}_{21}-N\Delta_{\boldsymbol{\alpha}}^\top\right]\mathbf{J}_{11}^{-1}\mathbf{V}_{12}\left[(\Delta_\beta^\top-\Delta_{\boldsymbol{\alpha}}^\top\mathbf{J}_{11}^{-1}\mathbf{J}_{12})\mathrm{J}_0^{-1}\right]^\top\right\}^\top\Big\}.
\end{aligned}\tag{A54}$$

Combining (A46), (A53), and the fact that $\hat{\delta}-\bar{\Delta}_n$ is uncorrelated with $\bar{\Delta}_n-\delta_0$ asymptotically, we have

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

Since the asymptotic normality of the plug-in estimator $\hat{\delta}$ has already been established, the asymptotic normality of $\hat{\delta}_{\mathrm{BG}}$ follows by an argument analogous to the proof of Theorem 3, and is therefore omitted for brevity. □

Proof of Theorem 5. The argument proceeds as in the proof of Theorem 1, with $(\boldsymbol{\alpha}_0,\beta_0)$ replaced by $(\boldsymbol{\alpha}_*,\beta_{n*})$. To avoid redundancy, we omit the details. □

Proof of Theorem 6. The argument follows the proofs of Theorems 2 and 3, with $(\boldsymbol{\alpha}_0,\beta_0)$ replaced by $(\boldsymbol{\alpha}_*,\beta_{n\sharp})$. The details are omitted to avoid repetition. □