5-2020

Detecting latent communities in network formation models

Shujie MA

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Yichong ZHANG
Singapore Management University, yczhang@smu.edu.sg

SMU ECONOMICS & STATISTICS

Detecting Latent Communities in Network Formation Models

Shujie Ma, Liangjun Su, Yichong Zhang

May 2020

Paper No. 12-2020

Detecting Latent Communities in Network Formation Models*

Shujie Ma† Liangjun Su‡ Yichong Zhang§

May 7, 2020

Abstract

This paper proposes a logistic undirected network formation model which allows for assortative matching on observed individual characteristics and the presence of edge-wise fixed effects. We model the coefficients of observed characteristics to have a latent community structure and the edge-wise fixed effects to be of low rank. We propose a multi-step estimation procedure involving nuclear norm regularization, sample splitting, iterative logistic regression and spectral clustering to detect the latent communities. We show that the latent communities can be exactly recovered when the expected degree of the network is of order $\log n$ or higher, where n is the number of nodes in the network. The finite sample performance of the new estimation and inference methods is illustrated through both simulated and real datasets.

Keywords: Community detection, homophily, spectral clustering, strong consistency, unobserved heterogeneity

1 Introduction

In real world social and economic networks, individuals tend to form links with someones who are alike to themselves, resulting in assortative matching on observed individual characteristics (homophily). In addition, network data often exhibit natural communities such that individuals in the same community may share similar preferences for a certain type of homophily while those in different communities tend to have quite distinctive preferences. In many cases, such a community structure is latent and has to be identified from the data. The detection of such community structures is challenging yet crucial for network analyses. It prompts a couple of important questions that need to be addressed: How do we formulate a network formation model with individual

*Ma's research was partially supported by NSF grant DMS 1712558 and a UCR Academic Senate CoR grant. Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian fellowship. Any and all errors are our own.

†Department of Statistics, University of California, Riverside. E-mail address: shujie.ma@ucr.edu.

‡Tsinghua University and Singapore Management University. E-mail address: ljsu@smu.edu.sg.

§Singapore Management University. E-mail address: yczhang@smu.edu.sg.

characteristics, unobserved edge-wise fixed effects, and latent communities? When the model is formulated, how do we recover the community structure and estimate the community-specific parameters effectively in the model?

To address the first issue above, we propose a logistic undirected network formation model with observed measurements of homophily as regressors. We allow the regression coefficients to have a latent community structure such that the regression coefficient for covariate l in the network formation model is B_{l,k_1k_2} for any nodes i and j in communities k_1 and k_2, respectively. The edge-wise fixed effects are assumed to have a low-rank structure. This includes the commonly used discretized fixed effects and additive fixed effects as special cases. Estimation of this latent model is challenging, and it has to involve a multi-step procedure. In the first step, we estimate the coefficient matrices by a nuclear norm regularized logistic regression given their low-rank structures; we then obtain the estimators of their singular vectors which contain information about the community memberships via the singular value decomposition (SVD). Such singular vector estimates are only consistent in Frobenius norms but not in uniform row-wise Euclidean norm. A refined estimation is needed for accurate community detection. In the second step, we use the singular vector estimates from the first step as the initial values and iteratively run row-wise and column-wise logistic regressions to reestimate the singular vectors. The efficiency of the resulting estimator can be improved through this iterative procedure. In the third step, we apply the standard K-means algorithm to the singular vector estimates obtained in the second step. For technical reasons, we have to resort to sample-splitting techniques to estimate the singular vectors, and for numerical stability, both iterative procedures and multiple-splits are called upon. We establish the exact recovery of the latent community (strong consistency) under the condition that the expected degree of the network diverges to infinity at the rate $\log n$ or higher order, where n is the number of nodes. Under the exact recovery property, we can treat the estimated community memberships as the truth and further estimate the community-specific regression coefficients.

Our paper is closely related to three strands of literature in statistics and econometrics. First, our paper is closely tied to the large literature on the application of spectral clustering to detect communities in stochastic block models (SBMs). Since the pioneering work of Holland, Laskey, and Leinhardt (1983), SBM has become the most popular model for community detection. The statistical properties of spectral clustering in such models have been studied by Jin (2015), Joseph and Yu (2016), Lei and Rinaldo (2015), Qin and Rohe (2013), Rohe, Chatterjee, and Yu (2011), Sarkar and Bickel (2015), Vu (2018), Yun and Proutiere (2014), and Yun and Proutiere (2016), among others. From an information theory perspective, Abbe and Sandon (2015), Abbe, Bandeira, and Hall (2016), Mossel, Neeman, and Sly (2014), and Vu (2018) establish the phase transition threshold for the exact recovery of communities in SBMs, which requires the expected degree to diverge to infinity at a rate no slower than $\log n$. Su, Wang, and Zhang (2020) show that spectral clustering can achieve this information-theoretical minimum rate for the exact recovery. Nevertheless, existing SBMs either do not include covariates or include covariates in a non-regression fashion (see, e.g., Binkiewicz, Vogelstein, and Rohe, 2017), which makes them too simple for practical uses. For more complicated models that can incorporate both covariates and community structures, people often resort to variational EM algorithm, the performance of which highly hinges on the proper choice of initial values. In contrast, the network formation model proposed in this paper extends the SBM to a complex logistic regression model with both latent community structures and covariates, and our multi-step procedure provides an effective and reliable tool for the estimation of such a complex network model. Despite the fact that the regression coefficient matrices have to estimated from the data in order to obtain the associated singular vectors for spectral clustering, we are able to obtain the exact recovery of the community structures at the minimal conditions on the expected node degree.

Second, our paper is closely tied to the literature on network formation models; see, for example, Chatterjee, Diaconis, and Sly (2011), Graham (2017), Jochmans (2019), Leung (2015), Mele (2017a), Rinaldo, Petrović, and Fienberg (2013), and Yan and Xu (2013). We complement these works by allowing for community structures on the regression coefficients, which can capture a rich set of unobserved heterogeneity in the network data. In a working paper, Mele (2017b) also considers a network formation model with heterogeneous players and latent community structure. He assumes that the community structure follows an i.i.d. multinomial distribution and imposes a prior distribution over communities and parameters before conducting Bayesian estimation and inferences. In contrast, we treat the community memberships as fixed parameters and aim to recover them from a single observation of a large network. Our idea of introducing the community structure in network formation model is also inspired by the recent works of Bonhomme and Manresa (2015) and Su, Shi, and Phillips (2016), who introduce latent group structures into panel data analyses.

Last, our paper is related to the literature on the use of nuclear norm regularization in various contexts; see Alidaee, Auerbach, and Leung (2020), Belloni, Chen, and Padilla (2019), Chernozhukov, Hansen, Liao, and Zhu (2018), Fan, Gong, and Zhu (2019), Feng (2019), Koltchinskii, Lounici, and Tsybakov (2011), Moon and Weidner (2018), Negahban and Wainwright (2011), Negahban, Ravikumar, Wainwright, and Yu (2012), and Rohde and Tsybakov (2011), among others. Except Moon and Weidner (2018) and Chernozhukov et al. (2018), all these previous works focus on the error bounds for the nuclear norm regularized estimates but not the asymptotic distributions or statistical inferences. Like Moon and Weidner (2018) and Chernozhukov et al. (2018), we simply use the nuclear norm regularization to obtain consistent initial estimates and our interest is not the error bounds but some asymptotic distribution results. Unlike Moon and Weidner (2018) and Chernozhukov et al. (2018) who study linear panel data models with a low-rank structure, we study a logistic network formation model with a latent community structure, and we need to fully recover the community membership before estimating the community-specific parameters and making statistical inferences.

The rest of the paper is organized as follows. In Section 2, we introduce the model and several basic assumptions. In Section 3, we provide our multi-step estimation procedure. In Section 4, we establish the statistical properties of our proposed estimators of the community memberships and regression coefficients. Section 5 reports simulation results. In Section 6, we apply the new methods to study the community structure of a Facebook friendship networks at one hundred American colleges and universities at a single time point. Section 7 concludes. We provide the proofs of all theoretical results in the appendix. Some additional technical results are contained in the online supplement.

Notation. Throughout the paper, we write $M = \{M_{ij}\}$ as a matrix with its (i, j)-th entry denoted as M_{ij}. We use $||\cdot||_{op}$, $||\cdot||_F$, and $||\cdot||_*$ to denote matrix spectral, Frobenius, and nuclear norms, respectively. We use $[n]$ to denote $\{1, \cdots, n\}$ for some positive integer n. For a vector u, $||u||$ and $u^\top$ denote its L_2 norm and transpose, respectively. For a vector $a = (a_1, \cdots, a_n)$, let $\text{diag}(a)$ be the diagonal matrix whose diagonal is a. For a symmetric matrix $B \in \mathbb{R}^{K\times K}$, we define

$$\text{vech}(B) = (B_{11}, B_{12}, B_{22}, \cdots, B_{1K_1}, \cdots, B_{K-1K}, B_{KK})^\top.$$

We define $\max(u, v) = u \vee v$ and $\min(u, v) = u \wedge v$ for two real numbers u and v. We write $\mathbf{1}\{A\}$ to denote the usual indicator function that takes value 1 if event A happens and 0 otherwise.

2 The Model and Basic Assumptions

In this section, we introduce the model and basic assumptions.

2.1 The Model

For $i \neq j \in [n]$, let Y_{ij} denote the dummy variable for a link between nodes i and j. It takes value 1 if nodes i and j are linked and 0 otherwise. Let $W_{ij} = (W_{1,ij}, ..., W_{p,ij})^\top$ denote a p-vector of measurements of homophily between nodes i and j. Researchers observe the network adjacency matrix $\{Y_{ij}\}$ and covariates $\{W_{ij}\}$. We model the link formation between i and j is as

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \log \zeta_n + \sum_{l=0}^{p} W_{l,ij}\Theta^*_{l,ij}\},\ i < j, \tag{2.1}$$

where $\{\zeta_n\}_{n\geq 1}$ is a deterministic sequence that may decay to zero and is used to control the expected degree in the network, $W_{0,ij} = 1$, and $W_{l,ij} = W_{l,ji}$ for $j \neq i$ and $l \in [p]$. For clarity, we consider undirected network so that $Y_{ij} = Y_{ji}$ and $\Theta^*_{l,ij} = \Theta^*_{l,ji}$ $\forall l$ if $i \neq j$, ε_{ij} follows the standard logistic distribution for $i > j$, and $\varepsilon_{ij} = \varepsilon_{ji}$. Let $Y_{ii} = 0$ for all $i \in [n]$.

Apparently, without making any assumptions on $\Theta^*_l = \{\Theta^*_{l,ij}\}$ for $l \in [p] \cup \{0\}$, one cannot estimate all the parameters in (2.1) as the number of parameters can easily exceed the number of observations in the model. Specifically, we will follow the literature on reduced rank regression and assume that each Θ_l^* exhibits a certain low rank structure. Even so, it is easy to see that our model in (2.1) is fairly general, and it includes a variety of network formation models as special cases.

1. If $\log(\zeta_n) = 2\bar{a} = \frac{2}{n}\sum_{i=1}^n a_i$, $\alpha_i = a_i - \bar{a}$, $\Theta_{0,ij}^* = \alpha_i + \alpha_j$, and $p = 0$, then

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq a_i + a_j\}. \tag{2.2}$$

   Under the standard logistic distribution assumption on ε_{ij}, $\mathbb{P}(Y_{ij} = 1) = \frac{\exp(a_i+a_j)}{1+\exp(a_i+a_j)}$ for all $i \neq j$, and we have the simplest exponential graph model (Beta model) considered in the literature; see, e.g., Lusher, Koskinen, and Robins (2013).

2. If $\log(\zeta_n)$ and $\Theta_{0,ij}^*$ are defined as above and $\Theta_{l,ij}^* = \beta_l$ for $l \in [p]$, then

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq a_i + a_j + W_{ij}^\top \beta\}, \tag{2.3}$$

   where $\beta = (\beta_1, ..., \beta_p)^\top$. Apparently, (2.3) is the undirected dyadic link formation model with degree heterogeneity studied in Graham (2017). See also Yan, Jiang, Fienberg, and Leng (2019) for the case of a directed network.

3. Let $\Theta_{0,ij} = \Theta_{0,ij}^* + \log\zeta_n$. If $p = 0$, and $\Theta_0 = \{\Theta_{0,ij}\}$ is assumed to exhibit the stochastic block structure such that $\Theta_{0,ij} = b_{kl}$ if nodes i and j belong to communities k and l, respectively, then we have

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \Theta_{0,ij}\}. \tag{2.4}$$

   Corresponding to the simple SBM with K communities, the the probability matrix $P = \{P_{ij}\}$ with $P_{ij} = \mathbb{P}(Y_{ij} = 1)$ can be written as $P = ZBZ^\top$ where $Z = \{Z_{ik}\}$ denotes an $n \times K$ binary matrix providing the cluster membership of each node, i.e., $Z_{ik} = 1$ if node i is in community k and $Z_{ik} = 0$ otherwise, and $B = \{B_{kl}\}$ denotes the block probability matrix that depends on b_{kl}. See Holland et al. (1983) and the references cited in the introduction section.

4. Let $\Theta_{0,ij} = \Theta_{0,ij}^* + \log\zeta_n$. If $\Theta_0 = \{\Theta_{0,ij}\}$ is assumed to exhibit the stochastic block structure such that $\Theta_{0,ij} = b_{kl}$ if nodes i and j belong to communities k and l, respectively, and $\Theta_{l,ij}^* = \beta_l$ for $l \in [p]$, then

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \Theta_{0,ij} + W_{ij}^\top \beta\}. \tag{2.5}$$

   Then (2.5) defines a stochastic block model with covariates considered in Sweet (2015), Leger (2016), and Roy, Atchade, and Michailidis (2019).

Under the assumptions specified in the next subsection, it is easy to see that the expected degree of the network is of order $n\zeta_n$. In the theory to be developed below, we allow ζ_n to shrink to zero at a rate as slow as $n^{-1}\log n$, so that the expected degree can be as small as $C\log n$ for some sufficiently large constant C and the network is semi-dense.[1] Of course, if ζ_n is fixed or convergent to a positive constant as $n\to\infty$, the network becomes dense.

To proceed, let $\tau_n=\log(\zeta_n)$, $\Gamma^*_{0,ij}=\tau_n+\Theta^*_{0,ij}$, $\Gamma^*_{ij}=(\Gamma^*_{0,ij},\Theta^*_{1,ij},...,\Theta^*_{p,ij})^\top$, and $W_{ij}=(W_{0,ij},W_{1,ij},...,W_{p,ij})^\top$, where $W_{0,ij}=1$. Let $\Gamma^*=(\Gamma^*_0,\Theta^*_1,...,\Theta^*_p)$, where $\Gamma^*_0=\{\Gamma^*_{0,ij}\}$ and $\Theta^*_l=\{\Theta^*_{l,ij}\}$ for $l\in[p]$. Then, we can rewrite the model in (2.1) as

$$Y_{ij}=\mathbf{1}\{\varepsilon_{ij}\le W_{ij}^\top\Gamma^*_{ij}\}. \tag{2.6}$$

Below, we impose some basic assumptions on the model in order to propose a multiple-step procedure to estimate the parameters of interest in the model.

2.2 Basic Assumptions

Now, we state a set of basic assumptions to characterize the model in (2.1). The first assumption is about the data generating process (DGP).

Assumption 1.

1. *For $l\in[p]$, $\{W_{l,ij}\}_{1\le i<j\le n}$ is exchangeable under node permutations, and thus, there exists a function $g_l(\cdot)$ such that $W_{l,ij}=g_l(X_i,X_j,e_{ij})$, where $g_l(\cdot,\cdot,e)$ is symmetric in its first two arguments, $\{X_i\}_{i=1}^n$ and $\{e_{ij}\}_{1\le i<j\le n}$ are two independent and identically distributed (i.i.d.) sequences of random variables, and $e_{ij}=e_{ji}$ for $i\ne j$.*
2. *$\{\varepsilon_{ij}\}_{1\le i<j\le n}$ is an i.i.d. sequence of logistic random variables. Moreover, $\{\varepsilon_{ij}\}_{1\le i<j\le n}\perp\!\!\!\perp(\{X_i\}_{i=1}^n\cup\{e_{ij}\}_{1\le i<j\le n})$. Let $\varepsilon_{ij}=\varepsilon_{ji}$ for $i>j$.*
3. *$\max_{l\in[p]}\max_{i\ne j\in[n]}|W_{l,ij}|\le M_W$ a.s.*

Assumption 1 specifies how the covariates and error terms are generated. In some applications, e_{ij} is absent and $W_{l,ij}$ depend on (X_i,X_j) only. For example, $W_{l,ij}=\|X_i-X_j\|$ for some l where $\|\cdot\|$ denotes the Euclidean norm. We further assume that it is uniformly bounded to simplify the analysis. Part 2 of Assumption 1 is standard.

The next assumption imposes some structures on $\{\Theta^*_l\}_{0\le l\le p}$.

Assumption 2.

1. *Θ^*_0 is symmetric with fixed rank K_0 such that $\max_{i,j\in[n]}|\Theta^*_{0,ij}|\le M$, and $\sum_{i,j\in[n]}\Theta^*_{0,ij}=0$.*
2. *$\Theta^*_l=Z_lB^*_lZ_l^\top$ for $l\in[p]$, where $Z_l\in\mathbb{R}^{n\times K_l}$ is the membership matrix with one entry in each row taking value one and the rest taking value zero, K_l denotes the number of distinctive communities, and $B^*_l\in\mathbb{R}^{K_l\times K_l}$ is symmetric with fixed rank K_l. In addition, $\max_{l,k_1,k_2}|B^*_{l,k_1k_2}|\le M$.*

[1] A network is dense if the expect degree grows at rate-n and semi-dense if it diverges to infinity at a rate slower than n.

Assumption 2 is a key assumption on the model parameters. It assumes that Θ_0^* is of low rank and each matrix in $\{\Theta_l^*\}_{l\in[p]}$ has a community structure. For $l = 0, ..., p$, since Θ_l^* is of rank K_l, the singular value decomposition of $n^{-1}\Theta_l^*$ is $\mathcal{U}_l\Sigma_l\mathcal{V}_l^\top$, where $\mathcal{U}_l$ and $\mathcal{V}_l$ are $n \times K_l$ matrices such that $\mathcal{U}_l^\top\mathcal{U}_l = I_{K_l} = \mathcal{V}_l^\top\mathcal{V}_l$ and $\Sigma_l = \text{diag}(\sigma_{1,l}, \cdots, \sigma_{K_l,l})$. We further denote $U_l = \sqrt{n}\mathcal{U}_l\Sigma_l$ and $V_l = \sqrt{n}\mathcal{V}_l$. Then,

$$\Theta_l^* = n\mathcal{U}_l\Sigma_l\mathcal{V}_l^\top = U_lV_l^\top \text{ for } l = 0, ..., p. \tag{2.7}$$

Let $z_{i,l}^\top$ denote the i-th row of Z_l for $l \in [p]$. Similarly, let $u_{i,l}^\top$ and $v_{i,l}^\top$ denote the i-th row of U_l and V_l, respectively for $l \in [p] \cup \{0\}$. (2.7) implies that $\Theta_{l,ij}^* = u_{i,l}^\top v_{j,l}$.

We view $\Theta_{0,ij}^*$ as the edge-wise fixed effects for the network formation model. We impose the normalization that $\sum_{i,j\in[n]}\Theta_{0,ij}^* = 0$ in the first part of Assumption 2 because we have included the grand intercept term $\tau_n(\equiv \log(\zeta_n))$ in (2.1). The low-rank structure of Θ_0^* incorporates two special cases: (1) additive fixed effects and (2) discretized fixed effects, as illustrated in detail in Examples 1 and 2 below, respectively. The models in Examples 1 and 2 extend, respectively, the so-called Beta model and stochastic block model to the scenario with edge-wise characteristics and latent community structure for the slope coefficients.

Example 1. *Let $\Theta_{0,ij}^* = \alpha_i + \alpha_j$ where $\sum_{i=1}^n \alpha_i = 0$. In this case, $K_0 = 2$ and $n^{-1}\Theta_0^* = \mathcal{U}_0\Sigma_0\mathcal{V}_0^T$, where*

$$\mathcal{U}_0 = \begin{pmatrix} \frac{1}{\sqrt{2n}}(1+\frac{\alpha_1}{s_n}) & \frac{-1}{\sqrt{2n}}(1-\frac{\alpha_1}{s_n}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2n}}(1+\frac{\alpha_n}{s_n}) & \frac{-1}{\sqrt{2n}}(1-\frac{\alpha_n}{s_n}) \end{pmatrix}, \ \mathcal{V}_0 = \begin{pmatrix} \frac{1}{\sqrt{2n}}(1+\frac{\alpha_1}{s_n}) & \frac{1}{\sqrt{2n}}(1-\frac{\alpha_1}{s_n}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2n}}(1+\frac{\alpha_n}{s_n}) & \frac{1}{\sqrt{2n}}(1-\frac{\alpha_n}{s_n}) \end{pmatrix}, \ \Sigma_0 = \begin{pmatrix} s_n & 0 \\ 0 & s_n \end{pmatrix}$$

and $s_n^2 = \frac{1}{n}\sum_{i=1}^n \alpha_i^2$. Similarly, it is easy to verify that

$$U_0 = \begin{pmatrix} \frac{1}{\sqrt{2}}(s_n+\alpha_1) & \frac{-1}{\sqrt{2}}(s_n-\alpha_1) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2}}(s_n+\alpha_n) & \frac{-1}{\sqrt{2}}(s_n-\alpha_n) \end{pmatrix} \quad \text{and } V_0 = \sqrt{n}\mathcal{V}_0.$$

Note that we allow $\{\alpha_i\}_{i=1}^n$ to depend on $\{W_{ij}\}_{1\le i<j\le n}$ so that $\{\alpha_i\}_{i=1}^n$ are usually referred to as individual fixed effects in the literature.

Example 2. *Let $\Theta_0^* = Z_0B_0^*Z_0^\top$, where $Z_0 \in \mathbb{R}^{n\times K_0}$ is the membership matrix, K_0 denotes the number of distinctive communities for Θ_0^*, and $B_0^* \in \mathbb{R}^{K_0\times K_0}$ is symmetric with rank K_0. Let ι_n denote an $n \times 1$ vector of ones. For normalization, we assume $\iota_n^\top Z_0B_0^*Z_0^\top\iota_n = p_0^\top B_0^*p_0 = 0$, where $p_0^\top = (\frac{n_{1,0}}{n}, \cdots, \frac{n_{K_0,0}}{n})$ and $n_{k,0}$ denotes the size of Θ_0^*'s k-th community for $k \in [K_0]$. Then, as Lemma 2.1 below shows,*

$$U_0 = Z_0^\top(\Pi_{0,n})^{-1/2}S_0'\Sigma_0 \quad \text{and} \quad V_0 = Z_0^\top(\Pi_{0,n})^{-1/2}S_0,$$

where S_0 *and* S_0' *are two* $K_0 \times K_0$ *matrices such that* $S_0^\top S_0 = I_{K_0} = (S_0')^\top S_0'$, $\Pi_{0,n} = diag(p_0)$, *and* Σ_0 *is the singular value matrix of* $\Pi_{0,n}^{1/2} B_0^* \Pi_{0,n}^{1/2}$. *Note that in this example, we allow the group structures* Z_0 *and* $Z_l, l \in [p]$ *to be different.*

Let $n_{l,k}$ denote the number of nodes in Θ_l^*'s k-th group for $k \in [K_l]$ and $l \in [p]$. Let $\pi_{l,kn} = n_{l,k}/n$ and $\Pi_{l,n} = \text{diag}(\pi_{l,1n}, \cdots, \pi_{l,K_l n})$ for $l \in [p]$. The next assumption imposes some conditions on the community size.

Assumption 3. *1. There exist some constants* C_σ *and* c_σ *such that*

$$\infty > C_\sigma \geq \limsup_n \max_{l \in [p] \cup \{0\}} \sigma_{1,l} \geq \liminf_n \min_{l \in [p] \cup \{0\}} \sigma_{K_l,l} \geq c_\sigma > 0.$$

2. There exist some constants C_1 *and* c_1 *such that*

$$\infty > C_1 \geq \limsup_n \max_{k \in [K_l],\, l \in [p]} \pi_{l,kn} \geq \liminf_n \min_{k \in [K_l],\, l \in [p]} \pi_{l,kn} \geq c_1 > 0.$$

Two remarks are in order. First, Assumption 3 implies that the size of each community of Θ_l^* is proportional to the number of nodes n. Such an assumption is common in the literature on network community detection and panel data latent structure detection. Second, it is possible to allow for $\pi_{l,kn}$ and/or $\sigma_{k,l}$ to vary with n. In this case, one just needs to keep track of all these terms in the proofs.

To proceed, we state a lemma that lays down the foundation for our estimation procedure in the next section.

Lemma 2.1. *Suppose that Assumptions 2 and 3 hold. Then,*

1. $V_l = Z_l(\Pi_{l,n})^{-1/2} S_l$ *and* $U_l = Z_l(\Pi_{l,n})^{-1/2} S_l' \Sigma_l$ *for* $l \in [p]$, *where* S_l *and* S_l' *are two* $K_l \times K_l$ *matrices such that* $S_l^\top S_l = I_{K_l} = (S_l')^\top S_l'$.
2. $\max_{j \in [n]} ||v_{j,l}|| \leq c_1^{-1/2} < \infty$ *and* $\max_{i \in [n]} ||u_{i,l}|| \leq c_1^{-1/2} C_\sigma < \infty$ *for* $l \in [p]$.
3. *If* $z_{i,l} \neq z_{j,l}$, *then* $\left\| \frac{v_{i,l}}{||v_{i,l}||} - \frac{v_{j,l}}{||v_{j,l}||} \right\| = ||(z_{i,l} - z_{j,l}) S_l|| = \sqrt{2}$ *for* $l \in [p]$.

Lemma 2.1 implies the singular vectors $\{v_{i,l}\}_{i \in [n]}$ of Θ_l^* contain information about the community structure. A similar result has been established in the community detection literature; see, e.g., Rohe et al. (2011, Lemma 3.1) and Su et al. (2020, Theorem II.1). If we also assume Θ_0^* exhibits a community structure, similar results also hold for it.

3 The Estimation

For notational simplicity, we will focus on the case of $p = 1$ and study the recovery of latent community structure in Θ_1^* below. The general case with multiple covariates involves fundamentally no new ideas but more complicated notations.

First, we recognize that Γ_0^* and Γ_1^* are both low rank matrices with ranks bounded from above by $K_0 + 1$ and K_1, respectively. We can obtain their preliminary estimates via the nuclear norm penalized logistic regression. Second, based on the normalization imposed in Assumption 2.1, we can estimate τ_n and Θ_0 separately. We then apply the SVD to the preliminary estimates of Θ_0 and Θ_1 and obtain the estimates of U_l, Σ_l, and V_l, $l = 0, 1$. Third, we plug back the second step estimates of $\{V_l\}_{l=0,1}$ and re-estimate each row of U_l by a row-wise logistic regression. We can further iterate this procedure and estimate U_l and V_l alternatively. Last, we apply the K-means algorithm to the final estimate of V_1 to recover the community memberships. We rely on a sample splitting technique along with the estimation. Throughout, we assume the ranks K_0 and K_1 are known. We will propose an singular-value-ratio-based criterion to select them in Section 4.6.

Below is an overview of the multi-step estimation procedure that we propose.

1. Using the full sample, run the nuclear norm regularized estimation twice as detailed in Section 3.1 and obtain $\widehat{\tau}_n$ and $\{\widehat{\Sigma}_l\}_{l=0,1}$, the preliminary estimates of τ_n and $\{\Sigma_l\}_{l=0,1}$.

2. Randomly split the nodes into two subsets, denoted as I_1 and I_2. Using edges $(i, j) \in I_1 \times [n]$, run the nuclear norm estimation twice as detailed in Section 3.2 and obtain $\{\widehat{V}_l^{(1)}\}_{l=0,1}$, a preliminary estimate of $\{V_l\}_{l=0,1}$. For $j \in [n]$, denote the j-th row of $\widehat{V}_l^{(1)}$ as $(\widehat{v}_{j,l}^{(1)})^\top$, which is a preliminary estimate of $v_{j,l}^\top$.

3. For each $i \in I_2$, take $\{\widehat{v}_{j,l}^{(1)}\}_{j\in I_2, l=0,1}$ as regressors and run the row-wise logistic regression to obtain $\{\widehat{u}_{i,l}^{(1)}\}_{l=0,1}$, the estimates of $\{u_{i,l}\}_{l=0,1}$. For each $j \in [n]$, take $\{\widehat{u}_{i,l}^{(1)}\}_{i\in I_2, l=0,1}$ as regressors and run the column-wise logistic regression to obtain updated estimates, $\{\dot{v}_{j,l}^{(0,1)}\}_{l=0,1}$ of $\{v_{j,l}\}_{l=0,1}$. See Section 3.3 for details.

4. Based on $\{\dot{v}_{j,l}^{(0,1)}\}_{j\in[n], l=0,1}$, obtain the iterative estimates $(\dot{u}_{i,0}^{(h,1)}, \dot{u}_{i,1}^{(h,1)})_{i\in[n]}$ and $(\dot{v}_{j,0}^{(h,1)}, \dot{v}_{j,1}^{(h,1)})_{j\in[n]}$ of the singular vectors as in Step 3 for $h = 1, 2, \cdots, H$. See Section 3.4 for details.

5. Switch the roles of I_1 and I_2 and repeat Steps 2–4 to obtain $(\dot{u}_{i,0}^{(h,2)}, \dot{u}_{i,1}^{(h,2)})_{i\in[n]}$ and $(\dot{v}_{j,0}^{(h,2)}, \dot{v}_{j,1}^{(h,2)})_{j\in[n]}$ for $h \in [H]$. Let $\overline{v}_{j,1} = \left(\frac{(\dot{v}_{j,1}^{(H,1)})^\top}{||\dot{v}_{j,1}^{(H,1)}||}, \frac{(\dot{v}_{j,1}^{(H,2)})^\top}{||\dot{v}_{j,1}^{(H,2)}||} \right)^\top$. Then, apply the K-means algorithm on $\{\overline{v}_{j,1}\}_{j\in[n]}$ to recover the community memberships in Θ_1^* as detailed in Section 3.5.

Several remarks are in order. First, $\widehat{\tau}_n$ and $\{\widehat{\Sigma}_l\}_{l=0,1}$ obtained in Step 1 are used in Steps 3-5 and to determine $\{K_l\}_{l=0,1}$ in Section 4.6, respectively. Second, we employ the sample-splitting technique to create independence between the edges used for Steps 2 and 3. As $\widehat{V}_l^{(1)}$ in Step 2 is estimated by the nuclear-norm regularized logistic regression, we can only control the estimation error in Frobenius norm, as shown in Theorem 4.1. On the other hand, to analyze the row-wise estimator, we need to control for the estimation error of $\widehat{V}_l^{(1)}$ in row-wise L_2 norm (denoted as $||\cdot||_{2\to\infty}$). We overcome the discrepancy between $||\cdot||_F$ and $||\cdot||_{2\to\infty}$ by the independence structure. Third, one may propose to use each row of the full-sample lower-rank estimator $\widehat{V}_l$ as $\{\dot{v}_{j,l}^0\}_{j\in[n]}$, the initial estimates in Step 4. However, as $\widehat{V}_l$ is estimated using the full sample, it is not independent of, say, the i-th row of the edges if we want to estimate $(u_{i,0}^\top, u_{i,1}^\top)$. Fourth, in the literature, researchers overcome this difficulty by using the "leave-one-out" technique. See, for example, Abbe, Fan, Wang, and Zhong (2017), Bean, Bickel, El Karoui, and Yu (2013), Javanmard and Montanari (2018), Su et al. (2020), and Zhong and Boumal (2018), among others. Denote $\widehat{\Theta}_l^{(i)}$ as the low-rank estimator of Θ_l^* using all the edges except those on the i-th row and column and $\widehat{V}_l^{(i)}$ is obtained by applying the SVD on $\widehat{\Theta}_l^{(i)}$. The key step for the "leave-one-out" technique is to establish a perturbation theory to bound $\widehat{\Theta}_l^{(i)} - \widehat{\Theta}_l$, and thus, $\widehat{V}_l^{(i)} - \widehat{V}_l$. However, unlike the community detection literature, $\widehat{\Theta}_l$ and $\widehat{\Theta}_l^{(i)}$ are not directly observed but estimated by the nuclear-norm regularized logistic regression. It is interesting but very challenging, if possible, to establish such a perturbation theory. Fifth, although the sample-splitting can result in information loss, we compensate it in three aspects: (1) we just treat the sample-split estimator $\dot{v}_{j,l}^{(0,1)}$ as an initial value and in Step 4, we update it via an iterative algorithm which uses all the edges; (2) we can switch the roles of I_1 and I_2 and obtain $\dot{v}_{j,l}^{(H,1)}$ and $\dot{v}_{j,l}^{(H,2)}$ after H iterations; (3) to mitigate the concern of the randomness caused by a single sample split, in Section 3.5, we propose to repeat the sample-splitting R times to obtain R classifications, and select one of them based on the maximum-likelihood principle.

3.1 Full-Sample Low-Rank Estimation

Recall that $\Gamma_0^* = \tau_n + \Theta_0^*$ and $\Gamma_1^* = \Theta_1^*$. Let $\Gamma^* = (\Gamma_0^*, \Gamma_1^*)$. Let $\Lambda(u) = \frac{1}{1+\exp(-u)}$ denote the standard logistic probability density function. Let

$$\ell_{ij}(\Gamma_{ij}) = Y_{ij}\log(\Lambda(W_{ij}^\top \Gamma_{ij})) + (1-Y_{ij})\log(1-\Lambda(W_{ij}^\top \Gamma_{ij}))$$

denote the conditional logistic log-likelihood function associated with nodes i and j. Let

$$\mathbb{T}(\tau, c_n) = \{(\Gamma_0, \Gamma_1) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} : |\Gamma_{0,ij} - \tau| \le c_n, |\Gamma_{1,ij}| \le c_n\}.$$

We propose to estimate Γ^* by $\widetilde{\Gamma} = (\widetilde{\Gamma}_0, \widetilde{\Gamma}_1)$ via minimizing the negative logistic log-likelihood function with the nuclear norm regularization:

$$\widetilde{\Gamma} = \underset{\Gamma\in\mathbb{T}(0,\log n)}{\arg\min}\; Q_n(\Gamma) + \lambda_n \sum_{l=0}^{1} ||\Gamma_l||_*, \tag{3.1}$$

where $Q_n(\Gamma) = \frac{-1}{n(n-1)} \sum_{i,j\in[n],i\neq j} \ell_{ij}(\Gamma_{ij})$ and $\lambda_n > 0$ is a regularization parameter. As mentioned above, we allow ζ_n to shrink to zero at a rate as slow as $n^{-1}\log n$ so that $\tau_n = \log(\zeta_n)$ is slightly smaller than $\log n$ in magnitude. So it is sufficient to consider a parameter space $\mathbb{T}(0, \log n)$ that expands at rate-$\log n$. Later on, we specify $\lambda_n = \frac{C_\lambda(\sqrt{\zeta_n n}+\sqrt{\log n})}{n(n-1)}$ for some constant tuning parameter C_λ. Throughout the paper, we assume $W_{1,ij}$ has been rescaled so that its standard error is one. Therefore, we do not need to consider different penalty loads for $||\Gamma_0||_*$ and $||\Gamma_1||_*$. Many statistical softwares automatically normalize the regressors when estimating a generalized linear model. We recommend this normalization in practice before using our algorithm.

Let $\widetilde{\tau}_n = \frac{1}{n(n-1)} \sum_{i\neq j} \widetilde{\Gamma}_{0,ij}$. We will show that $\widetilde{\tau}_n$ lies within $c_\tau\sqrt{\log n}$-neighborhood of the true value τ_n, where c_τ can be made arbitrarily small provided that the expected degree is larger than $C\log n$ for some sufficiently large C.[2] This rate is insufficient and remains to be refined. Given $\widetilde{\tau}_n$, we propose to reestimate Γ^* by $\widehat{\Gamma} = (\widehat{\Gamma}_0, \widehat{\Gamma}_1)$, where

$$\widehat{\Gamma} = \underset{\Gamma\in\mathbb{T}(\widetilde{\tau}_n, C_M\sqrt{\log n})}{\arg\min} Q_n(\Gamma) + \lambda_n \sum_{l=0}^{1} ||\Gamma_l||_*,$$

and C_M is some constant to be specified later. Note that we now restrict the parameter space to expand at rate-$\sqrt{\log n}$ only.

Let $\widehat{\tau}_n = \frac{1}{n(n-1)} \sum_{i\neq j} \widehat{\Gamma}_{0,ij}$. Since $\Theta_l^* = \{\Theta_{l,ij}^*\}$ are symmetric, we define their preliminary low-rank estimators as $\widehat{\Theta}_l = \{\widehat{\Theta}_{l,ij}\}$, where

$$\widehat{\Theta}_{l,ij} = \begin{cases} f_M((\widehat{\Gamma}_{l,ij} + \widehat{\Gamma}_{l,ji})/2 - \widehat{\tau}_n\delta_{l0}) & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j \end{cases} \quad \text{for } l = 0, 1,$$

$\delta_{l0} = \mathbf{1}\{l = 0\}$, $f_M(u) = u \cdot \mathbf{1}\{|u| \leq M\} + M \cdot \mathbf{1}\{u > M\} - M \cdot \mathbf{1}\{u < -M\}$, and M is some positive constant. For $l = 0, 1$, we denote the SVD of $n^{-1}\widehat{\Theta}_l$ as

$$n^{-1}\widehat{\Theta}_l = \widehat{\widehat{\mathcal{U}}}_l \widehat{\widehat{\Sigma}}_l (\widehat{\widehat{\mathcal{V}}}_l)^\top,$$

where $\widehat{\widehat{\Sigma}}_l = \text{diag}(\widehat{\sigma}_{1,l}, ..., \widehat{\sigma}_{n,l})$, $\widehat{\sigma}_{1,l} \geq \cdots \geq \widehat{\sigma}_{n,l} \geq 0$, and both $\widehat{\widehat{\mathcal{U}}}_l$ and $\widehat{\widehat{\mathcal{V}}}_l$ are $n \times n$ unitary matrices. Let $\widehat{\mathcal{V}}_l$ consist of the first K_l columns of $\widehat{\widehat{\mathcal{V}}}_l$, such that $(\widehat{\mathcal{V}}_l)^\top \widehat{\mathcal{V}}_l = I_{K_l}$ and $\widehat{\Sigma}_l = \text{diag}(\widehat{\sigma}_{1,l}, \cdots, \widehat{\sigma}_{K_l,l})$. Then $\widehat{V}_l = \sqrt{n}\widehat{\mathcal{V}}_l$.

[2]Let $\eta_{0n} = \sqrt{\frac{\log n}{n\zeta_n}}$ and $\eta_n = \eta_{0n} + \eta_{0n}^2$. The proof of Theorem 4.1.1 suggests that $\widetilde{\tau}_n - \tau_n = O_p(\eta_n\sqrt{\log n})$, which is $o_p(\sqrt{\log n})$ (resp. $o_p(1)$) if one assumes that the magnitude $n\zeta_n$ of the expected degree is of order higher than $\log n$ (resp. $(\log n)^2$). But we will only assume that $\eta_{0n} \leq C_F \leq \frac{1}{4}$ for some sufficiently small constant C_F below.

3.2 Split-Sample Low-Rank Estimation

We divide the n nodes into two roughly equal-sized subsets (I_1, I_2). Let $n_\ell = \# I_\ell$ denote the cardinality of the set I_ℓ. If n is even, one can simply set $n_\ell = n/2$ for $\ell = 1, 2$.

Now, we only use the pair of observations $(i,j) \in I_1 \times [n]$ to conduct the low-rank estimation. Let $\Gamma_l^*(I_1)$ consist of the i-th row of Γ_l^* for $i \in I_1$, $l = 0, 1$. Let $\Gamma^*(I_1) = (\Gamma_0^*(I_1), \Gamma_1^*(I_1))$. Define

$$\mathbb{T}^{(1)}(\tau, c_n) = \{(\Gamma_0, \Gamma_1) \in \mathbb{R}^{n_1 \times n} \times \mathbb{R}^{n_1 \times n} : |\Gamma_{0,ij} - \tau| \le c_n, |\Gamma_{1,ij}| \le c_n\}.$$

We estimate $\Gamma^*(I_1)$ via the following nuclear-norm regularized estimation

$$\widetilde{\Gamma}^{(1)} = \underset{\Gamma \in \mathbb{T}^{(1)}(0, \log n)}{\arg\min} \; Q_n^{(1)}(\Gamma) + \lambda_n^{(1)} \sum_{l=0}^{1} ||\Gamma_l||_*, \tag{3.2}$$

where $Q_n^{(1)}(\Gamma) = \frac{-1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \ell_{ij}(\Gamma_{ij})$ and $\lambda_n^{(1)} = \frac{C_\lambda(\sqrt{\zeta_n n} + \sqrt{\log n})}{n_1(n-1)}$.

Let $\widetilde{\tau}_n^{(1)} = \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \widetilde{\Gamma}_{0,ij}^{(1)}$. As above, this estimate lies within $c_\tau \sqrt{\log n}$-neighborhood of the true value τ_n. To refine it, we can reestimate $\Gamma^*(I_1)$ by $\widehat{\Gamma}^{(1)} = (\widehat{\Gamma}_0^{(1)}, \widehat{\Gamma}_{1j}^{(1)})$:

$$\widehat{\Gamma}^{(1)} = \underset{\Gamma \in \mathbb{T}^{(1)}(\widetilde{\tau}_n^{(1)}, C_M\sqrt{\log n})}{\arg\min} \; Q_n^{(1)}(\Gamma) + \lambda_n^{(1)} \sum_{l=0}^{1} ||\Gamma_l||_*.$$

Let $\widehat{\tau}_n^{(1)} = \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \widehat{\Gamma}_{0,ij}^{(1)}$. Noting that $\{\Gamma_l^*\}_{l=0,1}$ are symmetric, we define the preliminary low-rank estimates for the $n_1 \times n$ matrices $\Theta_l^*(I_1)$ by $\widehat{\Theta}_l^{(1)}$ for $l = 0, 1$, where

$$\widehat{\Theta}_{l,ij}^{(1)} = \begin{cases} f_M((\widehat{\Gamma}_{l,ij}^{(1)} + \widehat{\Gamma}_{l,ji}^{(1)})/2 - \widehat{\tau}_n^{(1)} \delta_{l0}) & \text{if} \quad (i,j) \in I_1 \times I_1, \; i \neq j \\ 0 & \text{if} \quad (i,j) \in I_1 \times I_1, \; i = j \;, \\ f_M(\widehat{\Gamma}_{l,ij}^{(1)} - \widehat{\tau}_n^{(1)} \delta_{l0}) & \text{if} \quad i \in I_1, \; j \notin I_1 \end{cases}$$

and δ_{l0}, $f_M(u)$ and M are defined in Step 1. For $l = 0, 1$, we denote the SVD of $n^{-1}\widehat{\Theta}_l^{(1)}$ as

$$n^{-1}\widehat{\Theta}_l^{(1)} = \widehat{\widetilde{\mathcal{U}}}_l^{(1)} \widehat{\widetilde{\Sigma}}_l^{(1)} (\widehat{\widetilde{\mathcal{V}}}_l^{(1)})^\top,$$

where $\widehat{\widetilde{\Sigma}}_l^{(1)}$ is a rectangular $(n_1 \times n)$ diagonal matrix with $\widehat{\sigma}_{i,l}^{(1)}$ appearing in the (i,i)th position and zeros elsewhere, $\widehat{\sigma}_{1,l}^{(1)} \ge \cdots \ge \widehat{\sigma}_{n_1,l}^{(1)} \ge 0$, and $\widehat{\widetilde{\mathcal{U}}}_l^{(1)}$ and $\widehat{\widetilde{\mathcal{V}}}_l^{(1)}$ are $n_1 \times n_1$ and $n \times n$ unitary matrices, respectively. Let $\widehat{\mathcal{V}}_l^{(1)}$ consist of the first K_l columns of $\widehat{\widetilde{\mathcal{V}}}_l^{(1)}$ such that $(\widehat{\mathcal{V}}_l^{(1)})^\top \widehat{\mathcal{V}}_l^{(1)} = I_{K_l}$. Let $\widehat{\Sigma}_l^{(1)} = \text{diag}(\widehat{\sigma}_{1,l}^{(1)}, \cdots, \widehat{\sigma}_{K_l,l}^{(1)})$. Then $\widehat{V}_l^{(1)} = \sqrt{n}\widehat{\mathcal{V}}_l^{(1)}$, and $(\widehat{v}_{j,l}^{(1)})^\top$ is the j-th row of $\widehat{V}_l^{(1)}$ for $j \in [n]$.

3.3 Split-Sample Row- and Column-Wise Logistic Regressions

Let $\mu = (\mu_0^\top, \mu_1^\top)^\top$ and $\Lambda_{ij}^{\text{left}}(\mu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \mu_l^\top \widehat{v}_{j,l}^{(1)} W_{l,ij})$ and $\ell_{ij}^{\text{left}}(\mu) = Y_{ij} \log(\Lambda_{ij}^{\text{left}}(\mu)) + (1 - Y_{ij}) \log(1 - \Lambda_{ij}^{\text{left}}(\mu))$. Given the preliminary estimate $\{\widehat{v}_{j,l}^{(1)}\}$ obtained in Step 2, we can estimate the left singular vectors $\{u_{i,0}, u_{i,1}\}$ for each $i \in I_2$ by $\{\widehat{u}_{i,0}^{(1)}, \widehat{u}_{i,1}^{(1)}\}$ via the row-wise logistic regression:

$$((\widehat{u}_{i,0}^{(1)})^\top, (\widehat{u}_{i,10}^{(1)})^\top)^\top = \underset{\mu=(\mu_0^\top,\mu_1^\top)^\top \in \mathbb{R}^{K_0+K_1}}{\arg\min} Q_{in,U}^{(0)}(\mu),$$

where $Q_{in,U}^{(0)}(\mu) = \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} \ell_{ij}^{\text{left}}(\mu)$.

Let $\nu = (\nu_0^\top, \nu_1^\top)^\top$ and $\Lambda_{ij}^{\text{right}}(\nu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \nu_l^\top \widehat{u}_{i,l}^{(1)} W_{l,ij})$ and $\ell_{ij}^{\text{right}}(\nu) = Y_{ij} \log(\Lambda_{ij}^{\text{right}}(\nu_1)) + (1 - Y_{ij}) \log(1 - \Lambda_{ij}^{\text{right}}(\nu))$. Given $(\widehat{u}_{i,0}^{(1)}, \widehat{u}_{i,1}^{(1)})$, we update the estimate of the right singular vectors $\{v_{i,0}, v_{i,1}\}$ for each $j \in [n]$ by $\{\dot{v}_{j,0}^{(0,1)}, \dot{v}_{j,1}^{(0,1)}\}$ via the column-wise logistic regression:

$$((\dot{v}_{j,0}^{(0,1)})^\top, (\dot{v}_{j,1}^{(0,1)})^\top)^\top = \underset{\nu=(\nu_0^\top,\nu_1^\top)^\top \in \mathbb{R}^{K_0+K_1}}{\arg\min} Q_{jn,V}^{(0)}(\nu),$$

where $Q_{jn,V}^{(0)}(\nu) = \frac{-1}{n_2} \sum_{i \in I_2, i \neq j} \ell_{ij}^{\text{right}}(\nu)$.

Our final objective is to obtain accurate estimates of $\{v_{j,l}\}_{j \in [n], l=0,1}$. To this end, we treat $\{\dot{v}_{j,0}^{(0,1)}, \dot{v}_{j,1}^{(0,1)}\}_{j \in [n]}$ as the initial estimate in the following full-sample iteration procedure.

3.4 Full-Sample Iteration

For $h = 1, 2, ..., H$, let $\Lambda_{ij}^{\text{left},h}(\mu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \mu_l^\top \dot{v}_{j,l}^{(h-1,1)} W_{l,ij}))$ and $\ell_{ij}^{\text{left},h}(\mu) = Y_{ij} \log(\Lambda_{ij}^{\text{left},h}(\mu)) + (1 - Y_{ij}) \log(1 - \Lambda_{ij}^{\text{left},h}(\mu))$. Given $\{\dot{v}_{i,0}^{(h-1,1)}, \dot{v}_{i,1}^{(h-1,1)}\}$, we can compute $\{\dot{u}_{i,0}^{(h,1)}, \dot{u}_{i,1}^{(h,1)}\}$ via

$$((\dot{u}_{i,0}^{(h,1)})^\top, (\dot{u}_{i,01}^{(h,1)})^\top)^\top = \underset{\mu=(\mu_0^\top,\mu_1^\top)^\top \in \mathbb{R}^{K_0+K_1}}{\arg\min} Q_{in,U}^{(h)}(\mu),$$

where $Q_{in,U}^{(h)}(\mu) = \frac{-1}{n} \sum_{j \in [n], j \neq i} \ell_{ij}^{\text{left},h}(\mu)$.

Given $\{\dot{u}_{i,0}^{(h,1)}, \dot{u}_{i,1}^{(h,1)}\}$, by letting $\Lambda_{ij}^{\text{right},h}(\nu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \nu_l^\top \dot{u}_{i,l}^{(h,1)} W_{l,ij}))$ and $\ell_{ij}^{\text{right},h}(\nu) = Y_{ij} \log(\Lambda_{ij}^{\text{right},h}(\nu)) + (1 - Y_{ij}) \log(1 - \Lambda_{ij}^{\text{right},h}(\nu))$, we compute $\{\dot{v}_{j,0}^{(h,1)}, \dot{v}_{j,1}^{(h,1)}\}$ via

$$((\dot{v}_{j,0}^{(h,1)})^\top, (\dot{v}_{j,1}^{(h,1)})^\top)^\top = \underset{\nu=(\nu_0^\top,\nu_1^\top)^\top \in \mathbb{R}^{K_0+K_1}}{\arg\min} Q_{jn,V}^{(h)}(\nu),$$

where $Q_{jn,V}^{(h)}(\nu) = \frac{-1}{n} \sum_{i \in [n], i \neq j} \ell_{ij}^{\text{right},h}(\nu)$.

We can stop iteration when certain convergence criterion is met for sufficiently large H.

Switching the roles of I_1 and I_2 and repeating the procedure in the last three steps, we can obtain the iterative estimates $\{\dot{u}_{i,0}^{(h,2)}, \dot{u}_{i,1}^{(h,2)}\}_{i \in [n]}$ and $\{\dot{v}_{j,0}^{(h,2)}, \dot{v}_{j,1}^{(h,2)}\}_{j \in [n]}$ for $h = 1, 2, \cdots, H$.

3.5 K-means Classification

Recall that $\overline{v}_{j,1} = \left(\frac{(\dot{v}_{j,1}^{(H,1)})^\top}{||\dot{v}_{j,1}^{(H,1)}||}, \frac{(\dot{v}_{j,1}^{(H,2)})^\top}{||\dot{v}_{j,1}^{(H,2)}||} \right)^\top$, a $2K_1 \times 1$ vector. We now apply the K-means algorithm to $\{\overline{v}_{j,1}\}_{j\in[n]}$. Let $\mathcal{B} = \{\beta_1, \ldots, \beta_{K_1}\}$ be a set of K_1 arbitrary $2K_1 \times 1$ vectors: $\beta_1, \ldots, \beta_{K_1}$. Define

$$\widehat{Q}_n(\mathcal{B}) = \frac{1}{n} \sum_{j=1}^{n} \min_{1\le k\le K_1} \|\overline{v}_{j,1} - \beta_k\|^2$$

and $\widehat{\mathcal{B}}_n = \{\widehat{\beta}_1, \ldots, \widehat{\beta}_{K_1}\}$, where $\widehat{\mathcal{B}}_n = \arg\min_{\mathcal{B}} \widehat{Q}_n(\mathcal{B})$. For each $j \in [n]$, we estimate the group identity by

$$\hat{g}_j = \arg\min_{1\le k\le K_1} \left\| \overline{v}_{j,1} - \widehat{\beta}_k \right\|, \tag{3.3}$$

where if there are multiple k's that achieve the minimum, $\hat{g}_j$ takes value of the smallest one.

As mentioned previously, we can repeat Steps 2–5 R times to obtain R membership estimates, denoted as $\{\hat{g}_{j,r}\}_{j\in[n],r\in[R]}$. Recall that

$$\text{vech}(B_1^*) = (B_{1,11}^*, B_{1,12}^*, B_{1,22}^*, \cdots, B_{1,1K_1}^*, \cdots, B_{1,K_1-1K_1}^*, B_{1,K_1K_1}^*)^\top,$$

which is a $K_1(K_1+1)/2$-vector. In addition, let $\chi_{1,ij}$ be a $K_1(K_1+1)/2$ vector such that the $((g_i^0 \vee g_j^0 - 1)(g_i^0 \vee g_j^0)/2 + g_i^0 \wedge g_j^0)$-th element is one and the rest are zeros, where $g_i^0 \in [K_1]$ denotes the true group membership of the i-th node in Θ_1^*. By construction,

$$\chi_{1,ij}^\top \text{vech}(B_1^*) = B_{1,g_i^0 g_j^0}^*.$$

Analogously, for the r-th split, denote $\hat{\chi}_{1r,ij}$ as a $K_1(K_1+1)/2$ vector such that the $((\hat{g}_{i,r} \vee \hat{g}_{j,r} - 1)(\hat{g}_{i,r} \vee \hat{g}_{j,r})/2 + \hat{g}_{i,r} \wedge \hat{g}_{j,r})$-th element is one and the rest are zeros. We then estimate B_1^* by $\widehat{B}_{1,r}$, a symmetric matrix constructed from $\widehat{b}_r$ by reversing the vech operator:

$$\widehat{b}_r = \arg\max_{b} \mathcal{L}_{n,r}(b),$$

where $\mathcal{L}_{n,r}(b) = \sum_{i<j}[Y_{ij}\log(\hat{\Lambda}_{ij}(b)) + (1 - Y_{ij})\log(1 - \hat{\Lambda}_{ij}(b)))]$ with $\hat{\Lambda}_{ij}(b) = \Lambda(\widehat{\tau}_n + \widehat{\Theta}_{0,ij} + W_{1,ij}\hat{\chi}_{1r,ij}' b)$, $\widehat{\tau}_n$ is obtained in Step 1, $\widehat{\Theta}_{0,ij} = [(\dot{u}_{i,0}^{(H,1)})^\top \dot{v}_{j,0}^{(H,1)} + (\dot{u}_{i,0}^{(H,2)})^\top \dot{v}_{j,0}^{(H,2)}]/2$, and $(\dot{u}_{i,0}^{(H,1)}, \dot{v}_{j,0}^{(H,1)}, \dot{u}_{i,0}^{(H,2)}, \dot{v}_{j,0}^{(H,2)})$ are obtained in Step 4. Then, the likelihood of the r-th split is defined as $\widehat{\mathcal{L}}(r) = \mathcal{L}_{n,r}(\widehat{b}_r)$. Our final estimator $\{\hat{g}_{i,r^*}\}_{i\in[n]}$ of the membership corresponds to the r^*-th split, where

$$r^* = \arg\max_{r\in[R]} \widehat{\mathcal{L}}(r). \tag{3.4}$$

4 Statistical Properties

In this section, we study the asymptotic properties of the estimators proposed in the last section.

4.1 Full- and Split-Sample Low-Rank Estimations

To study the asymptotic properties of the first two-step estimators, we add two assumptions.

Assumption 4. *For some positive constant $\tilde{c}$, let*

$$\mathcal{C}(\tilde{c}) = \{(\Delta_0, \Delta_1) : \Delta_l = \Delta_l' + \Delta_l^{''} \text{ for } l = 0, 1, \ \sum_{l=0}^{1} ||\Delta_l^{''}||_* \leq \tilde{c} \sum_{l=0}^{1} ||\Delta_l^{'}||_*, \\ rank(\Delta_0^{'}) \leq 2K_0 + 2, \ rank(\Delta_1^{'}) \leq 2K_1\}$$

If $(\Delta_0, \Delta_1) \in \mathcal{C}(\tilde{c})$, then there is a constant $\kappa > 0$ that potentially depends on $\tilde{c}$ such that

$$\sum_{1 \leq i,j \leq n} (\Delta_{0,ij} + W_{1,ij}\Delta_{1,ij})^2 \geq \kappa \sum_{1 \leq i,j \leq n} \left(\Delta_{0,ij}^2 + \Delta_{1,ij}^2\right) \ a.s.$$

Assumption 5. *1. $C_\lambda > C_\Upsilon M_W$, where C_Υ is a constant defined in Lemma S1.1 in the online supplement.*

2. There exist constants $0 < \underline{c} \leq \overline{c} < \infty$ such that $\zeta_n \underline{c} \leq \Lambda_{n,ij} \leq \zeta_n \overline{c}$, where $\Lambda_{n,ij} \equiv \Lambda(W_{ij}^\top \Gamma_{ij}^*)$.

3. $\sqrt{\frac{\log n}{n\zeta_n}} \leq c_F \leq \frac{1}{4}$ for some sufficiently small constant c_F.

4. There exists a constant $C_{0,u}$ such that $\max_{i\in[n]} |u_{i,0}| \leq C_{0,u}$.

5. $\sum_{i\in I_1, j\in[n]} \Theta_{0,ij}^* = o(\sqrt{\frac{\log(n)}{n\zeta_n}})$.

Assumption 4 is the restricted strong convexity condition commonly assumed in the literature. See, e.g., Negahban and Wainwright (2011), Negahban et al. (2012), Chernozhukov et al. (2018), and Moon and Weidner (2018), among others. Assumption 5 is a regularity condition. In particular, Assumptions 5.2 implies the order of the average degree in the network is $n\zeta_n$. Assumption 5.3 means that the average degree diverges to infinity at a rate that is not slower than $\log n$. Such a rate is the slowest for exact recovery in the SBM, as established by Abbe et al. (2016), Abbe and Sandon (2015), Mossel et al. (2014), and Vu (2018). As our model incorporates the SBM as a special case, the rate is also the minimal requirement for the exact recovery of Z_1, which is established in Theorem 4.4 below. Assumption 5.5 usually holds as the sample is split randomly and Θ_0^* satisfies the normalization condition in Assumption 2.1. For the specification in Example 1, this assumption is satisfied if $\frac{1}{n_1}\sum_{i\in I_1} \alpha_i = o(\sqrt{\frac{\log(n)}{n\zeta_n}})$. Such a requirement holds almost surely (*a.s.*)

if $\alpha_i = a_i - \frac{1}{n}\sum_{i\in[n]} a_i$ and $\{a_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with finite second moments. For the specification in Example 2, Assumption 5.5 is satisfied if $p_0^\top(I_1)B_0^*p_0 = o(\sqrt{\frac{\log(n)}{n\zeta_n}})$, where $p_0^\top(I_1) = (\frac{n_{1,0}(I_1)}{n_1}, \cdots, \frac{n_{k,0}(I_1)}{n_1})$ and $n_{k,0}(I_1)$ denotes the size of Θ_0^*'s k-th community for the subsample of nodes with index $i \in I_1$. As $p_0^\top B_0^* p_0 = 0$, the requirement holds *a.s.* if community memberships are generated from a multinomial distribution so that $||p_0 - p_0(I_1)|| = o_{a.s.}(\sqrt{\frac{\log(n)}{n\zeta_n}})$.

Theorem 4.1. *Let Assumptions 1–5 hold and* $\eta_n = \sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}$. *Then for* $l = 0, 1$, *we have that a.s.,*

1. $|\widehat{\tau}_n - \tau_n| \leq 30C_{F,1}\eta_n$, $|\widehat{\tau}_n^{(1)} - \tau_n| \leq 30C_{F,1}\eta_n$,
2. $\frac{1}{n}||\widehat{\Theta}_l - \Theta_l^*||_F \leq 48C_{F,1}\eta_n$, $\frac{1}{n}||\widehat{\Theta}_l^{(1)} - \Theta_l^*(I_1)||_F \leq 48C_{F,1}\eta_n$,
3. $\max_{k\in[K_l]} |\widehat{\sigma}_{k,l} - \sigma_{k,l}| \leq 48C_{F,1}\eta_n$, $\max_{k\in[K_l]} |\widehat{\sigma}_{k,l}^{(1)} - \sigma_{k,l}| \leq 48C_{F,1}\eta_n$,
4. $||V_l - \widehat{V}_l\widehat{O}_l||_F \leq 136C_{F,2}\sqrt{n}\eta_n$, *and* $||V_l - \widehat{V}_l^{(1)}\widehat{O}_l^{(1)}||_F \leq 136C_{F,2}\sqrt{n}\eta_n$,
   where $\widehat{O}_l$ *and* $\widehat{O}_l^{(1)}$ *are two* $K_l \times K_l$ *orthogonal matrices that depend on* $(V_l, \widehat{V}_l)$ *and* $(V_l, \widehat{V}_l^{(1)})$, *respectively, and* $C_{F,1}$ *and* $C_{F,2}$ *are two constants defined respectively after (A.12) and (A.13) in the Appendix.*

Part 1 of Theorem 4.1 indicates that despite the possible divergence of the grand intercept τ_n, we can estimate it consistently up to rate η_n. In the dense network, $\zeta_n \asymp 1$ where $a \asymp b$ denotes both a/b and b/a are stochastically bounded. In this case, $\tau_n \asymp 1$ and it can be estimated consistently at rate-$\sqrt{(\log n)/n}$. Note that the convergence rate of $\widehat{\Theta}_l$ and $\widehat{\Theta}_l^{(1)}$ in terms of the Frobenius norm is also driven by η_n. Similarly for $\widehat{\sigma}_{k,l}$ $\widehat{\sigma}_{k,l}^{(1)}$, $\widehat{V}_l/\sqrt{n}$ and $\widehat{V}_l^{(1)}/\sqrt{n}$.

4.2 Split-Sample Row- and Column-Wise Logistic Regressions

Define two $(K_0 + K_1) \times (K_0 + K_1)$ matrices:

$$\Psi_j(I_2) = \frac{1}{n_2}\sum_{i\in I_2, i\neq j} \begin{bmatrix} u_{i,0} \\ u_{i,1}W_{1,ij} \end{bmatrix} \begin{bmatrix} u_{i,0} \\ u_{i,1}W_{1,ij} \end{bmatrix}^\top \text{ and } \Phi_i(I_2) = \frac{1}{n_2}\sum_{j\in I_2, j\neq i} \begin{bmatrix} v_{j,0} \\ v_{j,1}W_{1,ij} \end{bmatrix} \begin{bmatrix} v_{j,0} \\ v_{j,1}W_{1,ij} \end{bmatrix}^\top.$$

To study the asymptotic properties of the third step estimator, we assume that both matrices are well behaved uniformly in i and j in the following assumption.

Assumption 6. *There exist constants* C_ϕ *and* c_ϕ *such that a.s.*

$$\begin{aligned} \infty &> C_\phi \geq \limsup_n \max_{j\in[n]} \lambda_{\max}(\Psi_j(I_2)) \geq \liminf_n \min_{j\in[n]} \lambda_{\min}(\Psi_j(I_2)) \geq c_\phi > 0 \text{ and} \\ \infty &> C_\phi \geq \limsup_n \max_{i\in I_2} \lambda_{\max}(\Phi_i(I_2)) \geq \liminf_n \min_{i\in I_2} \lambda_{\min}(\Phi_i(I_2)) \geq c_\phi > 0, \end{aligned}$$

where $\lambda_{\max}(\cdot)$ *and* $\lambda_{\min}(\cdot)$ *denote the maximum and minimum eigenvalues, respectively.*

Assumption 6 assumes that $\Phi_i(I_2)$ and $\Psi_j(I_2)$ are positive definite (p.d.) uniformly in i and j asymptotically. Suppose there are K_1 equal-sized communities and $B_1^* = I_{K_1}$ in Assumption 2, then $\Pi_{1,n} = \text{diag}(1/K_1, \cdots, 1/K_1)$. By Lemma S1.4 in the online supplement, if node j is in community k, then $v_{j,1} = \sqrt{n}\sqrt{\frac{K_1}{n}} z_{j,1} = \sqrt{K_1} e_{K_1,k}$, where $e_{K_1,k}$ denotes a $K_1 \times 1$ vector with the k-th unit being 1 and all other units being 0. For the specification in Example 1,

$$\Phi_i(I_2) = \frac{1}{n_2} \sum_{j \in I_2} \begin{pmatrix} \frac{1}{\sqrt{2}}(1 + \frac{\alpha_j}{s_n}) \\ \frac{1}{\sqrt{2}}(1 - \frac{\alpha_j}{s_n}) \\ z_{j,1} W_{1,ij} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}}(1 + \frac{\alpha_j}{s_n}) \\ \frac{1}{\sqrt{2}}(1 - \frac{\alpha_j}{s_n}) \\ z_{j,1} W_{1,ij} \end{pmatrix}^\top .$$

Suppose that $\alpha_i = a_i - \bar{a}$ for some i.i.d. sequence $\{a_i\}_{i=1}^n$ with $\bar{a} = \frac{1}{n}\sum_{i=1}^n a_i$, $\mathbb{E}(W_{1,ij} a_j | X_i) = 0$, $\mathbb{E}(W_{1,ij}^2 | X_i) \geq c > 0$ for some constant c, then $\frac{1}{n_2}\sum_{j \in I_2} a_j^2 / s_n \to 1$ *a.s.* Therefore, we can expect that, uniformly over $i \in I_2$,

$$\Phi_i(I_2) \to \text{diag}(1, 1, \mathbb{E}(W_{1,ij}^2 | X_i), \cdots, \mathbb{E}(W_{1,ij}^2 | X_i)) \ a.s.,$$

which implies Assumption 6 holds. For the specification in Example 2, if we further assume Θ_0^* and Θ_1^* share the same group structure Z_1, then

$$\Phi_i(I_2) = \frac{1}{n_2} \sum_{j \in I_2} \begin{pmatrix} z_{j,0} \\ z_{j,1} W_{1,ij} \end{pmatrix} \begin{pmatrix} z_{j,0} \\ z_{j,1} W_{1,ij} \end{pmatrix}^\top .$$

Suppose $\mathbb{E}(W_{1,ij} | X_i) = 0$ and $\mathbb{E}(W_{1,ij}^2 | X_i) \geq c > 0$ for some constant c, then one can expect that $\Phi_i(I_2)$ has the same *a.s.* limit as above uniformly over $i \in I_2$.

The following theorem studies the asymptotic properties of $\widehat{u}_{i,l}^{(1)}$ and $\dot{v}_{j,l}^{(0,1)}$ defined in Step 3.

Theorem 4.2. *Suppose that Assumptions 1–6 hold. Then,*

$$\max_{i \in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \leq C_1^* \eta_n \quad \textit{and} \quad \max_{j \in [n]} ||(\widehat{O}_l^{(1)})^\top \dot{v}_{j,l}^{(0,1)} - v_{j,l}|| \leq C_{0,v} \eta_n \ a.s.,$$

where C_1^* and $C_{0,v}$ are some constants defined respectively in (A.25) and (A.29) in the Appendix.

Theorem 4.2 establishes the uniform bound for the estimation error of $\dot{v}_{j,l}^{(0,1)}$ up to some rotation. Since Lemma 2.1 shows $\{v_{j,1}\}_{j \in [n]}$ contains information about the community memberships, it is intuitive to expect that we can use $\dot{v}_{j,l}^{(0,1)}$ to recover the memberships as long as the estimation error is sufficiently small. However, we only use half of the edges to estimate $\dot{v}_{j,l}^{(0,1)}$, which may result in information loss. In the next section, we treat $\dot{v}_{j,l}^{(0,1)}$ as an initial value and iteratively re-estimate $\{u_{i,l}\}_{i \in [n]}$ and $\{v_{j,l}\}_{i \in [n]}$ using all the edges in the network. We will show that the iteration can preserve the error bound established in Theorem 4.2 and improve the efficiency of the estimates.

4.3 Full-Sample Iteration

Define two $(K_0 + K_1) \times (K_0 + K_1)$ matrices:

$$\Psi_j = \frac{1}{n} \sum_{i \in [n], i \neq j} \begin{bmatrix} u_{i,0} \\ u_{i,1} W_{1,ij} \end{bmatrix} \begin{bmatrix} u_{i,0} \\ u_{i,1} W_{1,ij} \end{bmatrix}^\top \quad \text{and} \quad \Phi_i = \frac{1}{n} \sum_{j \in [n], j \neq i} \begin{bmatrix} v_{j,0} \\ v_{j,1} W_{1,ij} \end{bmatrix} \begin{bmatrix} v_{j,0} \\ v_{j,1} W_{1,ij} \end{bmatrix}^\top .$$

To study the asymptotic properties of the fourth step estimators, we add an assumption.

Assumption 7. *There exist constants C_ϕ and c_ϕ such that a.s.*

$$\begin{aligned} \infty &> C_\phi \geq \limsup_n \max_{j \in [n]} \lambda_{\max}(\Psi_j) \geq \liminf_n \min_{j \in [n]} \lambda_{\min}(\Psi_j) \geq c_\phi > 0 \text{ and} \\ \infty &> C_\phi \geq \limsup_n \max_{i \in [n]} \lambda_{\max}(\Phi_i) \geq \liminf_n \min_{i \in [n]} \lambda_{\min}(\Phi_i) \geq c_\phi > 0. \end{aligned}$$

The above assumption parallels Assumption 6 and is now imposed for the full sample.

Theorem 4.3. *Suppose that Assumptions 1–7 hold. Then, for $h = 1, \cdots, H$ and $l = 0, 1$,*

$$\max_{i \in [n]} ||(\widehat{O}_l^{(1)})^\top \dot{u}_{i,l}^{(h,1)} - u_{i,l}|| \leq C_{h,u} \eta_n \quad \text{and} \quad \max_{i \in [n]} ||(\widehat{O}_l^{(1)})^\top \dot{v}_{i,l}^{(h,1)} - v_{i,l}|| \leq C_{h,v} \eta_n \text{ a.s.},$$

where $\{C_{h,u}\}_{h=1}^H$ and $\{C_{h,v}\}_{h=1}^H$ are two sequences of constants defined in the proof of this theorem.

Theorem 4.3 establishes the uniform bound for the estimation error in the iterated estimators $\{\dot{u}_{i,l}^{(h,1)}\}$ and $\{\dot{v}_{i,l}^{(h,1)}\}$.

By switching the roles of I_1 and I_2, we have, similar to Theorem 4.1, that

$$||V_l - \widehat{V}_l^{(2)} \widehat{O}_l^{(2)}||_F \leq 136 C_{F,2} \sqrt{n} \eta_n,$$

where $\widehat{O}_l^{(2)}$ is a $K_l \times K_l$ rotation matrix that depends on V_l and $\widehat{V}_l^{(2)}$. Then, following the same derivations of Theorems 4.2 and 4.3, we have, for $h = 1, \cdots, H$,

$$\max_{i \in [n]} ||(\widehat{O}_l^{(2)})^\top \dot{u}_{i,l}^{(h,2)} - u_{i,l}|| \leq C_{h,u} \eta_n \quad \text{and} \quad \max_{i \in [n]} ||(\widehat{O}_l^{(2)})^\top \dot{v}_{i,l}^{(h,2)} - v_{i,l}|| \leq C_{h,v} \eta_n \text{ a.s.}$$

4.4 K-means classification

Let $g_i^0 \in [K_1]$ denote the true group identity for the i-th node. To establish the strong consistency of the membership estimator $\hat{g}_i$ defined in (3.3), we add the following side condition.

Assumption 8. *Suppose $145 K_1^{3/2} C_{H,v} C_1 \eta_n \leq 1$, where $C_{H,v}$ is the constant defined in the proof of Theorem 4.3.*

Apparently, Assumption 8 is automatically satisfied in large samples if $\eta_n = o(1)$.

Theorem 4.4. *If Assumptions 1–8 hold, then*

$$\max_{1\le i\le n} \mathbf{1}\{\hat{g}_i \neq g_i^0\} = 0 \ a.s.$$

Several remarks are in order. First, Theorem 4.4 implies the K-means algorithm can exactly recover the latent group structure of Θ_1^* *a.s.* Second, if we repeat the sample split R times, we need to maintain Assumptions 6 for each split. Then, we can show the exact recovery of $\hat{g}_{i,r}$ for $r \in [R]$ in the exact same manner, as long as R is fixed. This implies $\hat{g}_{i,r*}$ for r^* selected in (3.4) also enjoys the property that

$$\max_{1\le i\le n} \mathbf{1}\{\hat{g}_{i,r^*} \neq g_i^0\} = 0 \ a.s.$$

Third, if Θ_0^* also has the latent group structure as in Example 2, we can apply the same K-means algorithm to $\{\overline{v}_{j,0}\}_{j\in[n]}$ with $\overline{v}_{j,0} \equiv (\dot{v}_{j,0}^{(H,1)\top}/||\dot{v}_{j,0}^{(H,1)}||, \dot{v}_{j,0}^{(H,2)\top}/||\dot{v}_{j,0}^{(H,2)}||)^\top$ to recover the group identities of Θ_0^*. Last, if we further assume $Z_0 = Z_1 = Z$ (which implies $K_0 = K_1$), then we can catenate $\overline{v}_{j,0}$ and $\overline{v}_{j,1}$ as a $4K_1 \times 1$ vector and apply the same K-means algorithm to this vector to recover the group membership for each node.

4.5 Inference for B_1^*

Given the exact recovery of the community memberships asymptotically, we can just treat $\hat{g}_i$ as g_i^0. In this case, the inference for B_1^* for the model in Example 1 has been studied by Graham (2017). The model in Example 2 boils down to the standard logistic regression with finite-number of parameters, whose inference theory is established in Appendix S2. In the following, we discuss the two specifications in Examples 1 and 2.

Example 1 (cont.). Suppose the model is specified as in (2.6) with $\Gamma_{0,ij}^* = \tau_n + \alpha_i + \alpha_j$ and $\Gamma_1^* = \Theta_1^* = Z_1 B_1^* Z_1^\top$. Recall the definitions of $\chi_{1,ij}$, $\hat{\chi}_{1r,ij}$, and $\text{vech}(B_1^*)$ in Section 3.5 such that $\chi_{1,ij}^\top \text{vech}(B_1^*) = B^*_{1,g_i^0 g_j^0}$. We further denote $\hat{\chi}_{1,ij}$ as either $\hat{\chi}_{1,ij}$ if one single split is used or $\hat{\chi}_{1r^*,ij}$ if R splits are used and the r^*-th split is selected.

Corollary 4.1. *Suppose that Assumptions 1–8 hold. Then* $\hat{\chi}_{1,ij} = \chi_{1,ij}, \ \forall i < j \ a.s.$

Corollary 4.1 directly follows from Theorem 4.4 and implies that we can treat $\chi_{1,ij}$ as observed. Then, (2.6) can be written as

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \le \tau_n + \alpha_i + \alpha_j + \omega_{1,ij}^\top \text{vech}(B_1^*)\},$$

where $\omega_{1,ij} = W_{1,ij}\chi_{1,ij}$. This model has already been studied by Graham (2017). We can directly apply his Tetrad logit regression to estimate $\text{vec}(B_1^*)$. We provide more details on the estimation and inference in Section S2 in the online supplement.

Example 2 (cont.). Let $g_{i,0}^0$ be the true memberships of node i for Θ_0^* and $\hat{g}_{i,0}$ be its estimator which can be computed by applying the K-means algorithm to $\{\overline{v}_{j,0}\}_{j\in[n]}$. Further note $Z_0\iota_{K_0} = \iota_n$ where recall that ι_b denote a $b\times 1$ vector of ones. Therefore, $\Gamma_0^* = \tau_n\iota_n\iota_n^\top + Z_0B_0^*Z_0^\top = Z_0(B_0^* + \tau_n\iota_{K_0}\iota_{K_0}^\top)Z_0^\top \equiv Z_0B_0^{**}Z_0^\top$. As above, we define $\chi_{0,ij}$ be a $K_0(K_0+1)/2\times 1$ vector whose $((g_{i,0}^0 \vee g_{j,0}^0 - 1)(g_{i,0}^0 \vee g_{j,0}^0)/2 + g_{i,0}^0 \wedge g_{j,0}^0)$-th element is one and the rest are zeros and $\hat{\chi}_{0,ij}$ be a $K_0(K_0+1)/2\times 1$ vector whose $((\hat{g}_{i,0} \vee \hat{g}_{j,0} - 1)(\hat{g}_{i,0} \vee \hat{g}_{j,0})/2 + \hat{g}_{i,0} \wedge \hat{g}_{j,0})$-th element is one and the rest are zeros. Similarly, we have the following corollary.

Corollary 4.2. *If Assumptions 1–8 hold and the model is specified in Example 2, then* $\hat{\chi}_{l,ij} = \chi_{l,ij}$, $\forall i<j, l=0,1$ *a.s.*

We propose to estimate $\text{vech}(B^*) \equiv (\text{vech}(B_0^{**})^\top, \text{vech}(B_1^*)^\top)^\top$ by

$$\widehat{b} \equiv (\widehat{b}_0^\top, \widehat{b}_1^\top)^\top = \underset{b=(b_0^\top,b_1^\top)^\top\in\mathbb{R}^{K_0(K_0+1)/2}\times\mathbb{R}^{K_1(K_1+1)/2}}{\arg\min} Q_n(b),$$

where $Q_n(b) = \frac{-1}{n(n-1)}\sum_{1\le i<j\le n}[Y_{ij}\log(\hat{\Lambda}_{ij}(b)) + (1-Y_{ij})\log(1-\hat{\Lambda}_{ij}(b))]$, and $\hat{\Lambda}_{ij}(b) = \Lambda(\hat{\chi}_{0,ij}^\top b_0 + \hat{\chi}_{1,ij}^\top W_{1,ij}b_1)$. As we can view the estimated memberships as the truth, the asymptotic distribution of $\widehat{b}$ can be established by standard arguments. See Section S2 in the online supplement for details.

Although in theory, the inference for B_1^* in the above two examples is straightforward, there are two finite-sample issues. First, the tetrad logistic regression does not scale with the number of nodes n because the algorithm scans over all four-nodes figurations, which contains a total of $O(n^4)$ operations in a brutal force implementation. Although the Python code by Graham (2017) incorporates a number of computational speed-ups by keeping careful track of non-contributing configurations as the estimation proceeds, we still find in our simulations that the implementation turns extremely hard for networks with over 1000 nodes. One can, instead, use subsampling or divide-and-conquer algorithm for estimation. To establish the theoretical properties of such an estimator is an important and interesting topic for future research.

Second, for the specification in the second example, based on unreported simulation results, we find that $\widehat{b}_1$ has a small bias if there are some misclassified nodes. However, as the standard error of our estimator is even smaller, such small bias cannot be ignored in making inference. If we further increase the sample size, then the classification indeed achieves exact recovery and such a bias vanishes quickly. However, in practice, researchers cannot know whether their sample size is sufficiently large. It is interesting to further investigate such a bias issue and make proper bias-corrections. This is, again, left as a topic for future research.

4.6 Determining K_0 and K_1

In practice, K_0 and K_1 are unknown and need to be estimated from the data. In this case, we propose to replace them by a large but fixed integer $K_{\max}$ in the first step estimation to obtain the singular value estimates $\{\hat{\sigma}_{k,l}\}_{k\in[K_{\max}],l=0,1}$. We propose a version of singular-value ratio (SVR) statistic in the spirit of the eigenvalue-ratio statistics of Ahn and Horenstein (2013) and Lam and Yao (2012). That is, for $l = 0, 1$, we estimate K_l by

$$\widehat{K}_l = \arg\max_{1\le k\le K_{\max}-1} \frac{\widehat{\sigma}_{k,l}}{\widehat{\sigma}_{k+1,l}} \mathbf{1}\left\{\widehat{\sigma}_{k,l} \ge c_l \left(\sqrt{\frac{\log n}{n\overline{Y}}} + \frac{\log n}{n\overline{Y}}\right)\right\} \tag{4.1}$$

where $\bar{Y} = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} Y_{ij}$, and c_l is the tuning parameter to be specified. Without the indicator function in the above definition, $\widehat{K}_l$ is nothing but the SVR statistic. The use of the indicator function helps to avoid the overestimation of the ranks. Apparently, $n\bar{Y}$ consistently estimate the expected degree that is of order $n\zeta_n$. By using Assumption 3 and the results in Theorem 4.1, we can readily establish the consistency of $\hat{K}_l$.

5 Monte Carlo Simulations

In this section, we conduct some simulations to evaluate the performance of our procedure.

5.1 Data generation mechanisms

We generate data from the following two models.

Model 1. We simulate the responses Y_{ij} from the Bernoulli distribution with mean $\Lambda(\log(\zeta_n) + \Theta^*_{0,ij} + W_{1,ij}\Theta^*_{1,ij})$ for $i > j$, where $\Theta^*_{0,ij} = \alpha_i + \alpha_j$ and $\Theta^*_1 = ZB^*_1 Z^\top$. We generate $\alpha_i \overset{i.i.d}{\sim} \mathcal{U}(-1/2, 1/2)$ for $i = 1, ..., n$, and $W_{1,ij} = |X_i - X_j|$ for $i \neq j$, where $X_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$. For the i^{th} row of the membership matrix $Z \in \mathbb{R}^{n\times K_1}$, the C_i^{th} component is 1 and other entries are 0, where $C = (C_1, ..., C_n)^\top \in \mathbb{R}^n$ is the membership vector with $C_i \in [K_1]$.

Case 1. Let $K_1 = 2$ and $B^*_1 = ((0.6, 0.2)^\top, (0.2, 0.7)^\top)^\top$. The membership vector $C = (C_1, ..., C_n)^\top$ is generated by sampling each entry independently from $\{1, 2\}$ with probabilities $\{0.4, 0.6\}$. Let $\zeta_n = 0.7n^{-1/2}\log n$.

Case 2. Let $K_1 = 3$ and $B^*_1 = ((0.8, 0.4, 0.3)^\top, (0.4, 0.7, 0.4)^\top, (0.3, 0.4, 0.8)^\top)^\top$. The membership vector $C = (C_1, ..., C_n)^\top$ is generated by sampling each entry independently from $\{1, 2, 3\}$ with probabilities $\{0.3, 0.3, 0.4\}$. Let $\zeta_n = 1.5n^{-1/2}\log n$.

Model 2. We simulate the responses Y_{ij} from the Bernoulli distribution with mean $\Lambda(\log(\zeta_n) + \Theta^*_{0,ij} + W_{1,ij}\Theta^*_{1,ij})$ for $i > j$, where $\Theta^*_0 = ZB^*_0 Z^\top$, $\Theta^*_1 = ZB^*_1 Z^\top$, and $W_{1,ij}$ is simulated in the same way as in Model 1. Note here we impose that the latent community structures for Θ^*_0 and Θ^*_1 are the same. We then apply the K-means algorithm to the $4K_1 \times 1$ vector $\{\overline{v}^\top_{j,0}, \overline{v}^\top_{j,1}\}_{j\in[n]}$ to recover the community membership, as described in Section 4.4.

Case 1. Let $K_0 = K_1 = 2$ and $B^*_0 = ((0.6, 0.2)^\top, (0.2, 0.7)^\top)^\top$, $B^*_1 = ((0.6, 0.2)^\top, (0.2, 0.5)^\top)^\top$. The membership vector $C = (C_1, ..., C_n)^\top$ is generated by sampling each entry independently from

$\{1, 2\}$ with probabilities $\{0.3, 0.7\}$. Let $\zeta_n = 0.5n^{-1/2} \log n$.

Case 2. Let $K_0 = K_1 = 3$ and $B_0^* = ((0.7, 0.2, 0.2)^\top, (0.2, 0.6, 0.2)^\top, (0.2, 0.2, 0.7)^\top)^\top$, $B_1^* = ((0.7, 0.3, 0.2)^\top, (0.3, 0.7, 0.2)^\top, (0.2, 0.2, 0.6)^\top)^\top$. The membership vector is generated in the same way as given in Case 2 of Model 1. Let $\zeta_n = 1.5n^{-1/2} \log n$.

We consider $n = 500$, 1000, and 1500. All simulation results are based on 200 realizations.

5.2 Simulation Results

We select the number of communities K_1 by an eigenvalue ratio method given as follows. Let $\widehat{\sigma}_{1,1} \geq \cdots \geq \widehat{\sigma}_{K_{\max},1}$ be the first $K_{\max}$ singular values of the SVD decomposition of $\widehat{\Theta}_1$ from the nuclear norm penalization method given in Section 3.1. We estimate K_1 by $\widehat{K}_1$ defined in (4.1) by setting $c_1 = 0.1$ and $K_{\max} = 10$. We set the tuning parameter $\lambda_n = C_\lambda\{\sqrt{n\overline{Y}} + \sqrt{\log n}\}/\{n(n-1)\}$ with $C_\lambda = 2$ and similarly for $\lambda_n^{(1)}$. To require that the estimator of $\widehat{\Theta}_{l,ij}$ be bounded by finite constants, we let $M = 2$ and $C_M = 2$. The performance of the method is not sensitive to the choice of these finite constants. Define the mean squared error (MSE) of the nuclear norm estimator $\widehat{\Theta}_l$ for Θ_l as $\sum_{i \neq j}(\widehat{\Theta}_{l,ij} - \Theta^*_{l,ij})^2/\{n(n-1)\}$ for $l = 0, 1$.

Table 1 reports the MSEs for $\widehat{\Theta}_l$, the mean of $\widehat{K}_1$ and the percentage of correctly estimating K_1 based on the 200 realizations. We observe that the mean value of $\widehat{K}_1$ gets closer to the true number of communities K_1 and, the percentage of correctly estimating K_1 approaches to 1, as the samples size n increases. When n is large enough ($n = 1500$), the mean value of $\widehat{K}_1$ is the same as K_1 and the percentage of correctly estimating K is exactly equal to 1.

Table 1: The MSEs for $\widehat{\Theta}_l$, the mean of $\widehat{K}_1$ and the percentage of correctly estimating K_1 based on the 200 realizations for Models 1 and 2.

	$K_1 = 2$			$K_1 = 3$		
n	500	1000	1500	500	1000	1500
Model 1						
MSE for $\widehat{\Theta}_0$	0.083	0.079	0.092	0.112	0.091	0.088
MSE for $\widehat{\Theta}_1$	0.226	0.215	0.211	0.256	0.263	0.265
mean of $\widehat{K}_1$	1.990	2.000	2.000	2.990	3.000	3.000
percentage	0.990	1.000	1.000	0.990	1.000	1.000
Model 2						
MSE for $\widehat{\Theta}_0$	0.304	0.318	0.328	0.173	0.184	0.196
MSE for $\widehat{\Theta}_1$	0.150	0.157	0.170	0.153	0.155	0.151
mean of $\widehat{K}_1$	1.980	2.005	2.000	2.725	3.000	3.000
percentage	0.980	0.995	1.000	0.705	1.000	1.000

Next, we use three commonly used criteria for evaluating the accuracy of membership estimation for our proposed method. These criteria include the Normalized Mutual Information (NMI), the Rand Index (RI) and the proportion (PROP) of correctly identifying the membership for each individual node. They all give a value between 0 and 1, where 1 means a perfect membership estimation. Table 2 presents the mean of the NMI, RI and PROP values based on the 200 realizations for Models 1 and 2. The values of NMI, RI and PROP increase to 1 as the sample size increases for all cases. These results demonstrate that our method is quite effective for membership estimation in both models, and corroborate our large-sample theory.

Table 2: The means of the NMI, RI and PROP values based on the 200 realizations for Models 1 and 2.

	$K_1 = 2$			$K_1 = 3$		
n	500	1000	1500	500	1000	1500
			Model 1			
NMI	0.9247	0.9976	0.9978	0.5494	0.7867	0.8973
RI	0.9807	0.9995	0.9996	0.7998	0.9062	0.9593
PROP	0.9903	0.9999	0.9999	0.8063	0.9089	0.9670
			Model 2			
NMI	0.9488	0.9977	0.9984	0.9664	0.9843	0.9977
RI	0.9881	0.9966	0.9998	0.9790	0.9909	0.9987
PROP	0.9940	0.9978	0.9999	0.9838	0.9928	0.9988

Last, we estimate the parameters B_0^* and B_1^* by our proposed method given in Section 4.5 for Model 2. Tables 3 and 4 show the empirical coverage rate (coverage) of the 95% confidence intervals, the absolute value of bias (bias), the empirical standard deviation (emp_sd), and the average value of the estimated asymptotic standard deviation (asym_sd) of the estimates for B_0^* and B_1^* in cases 1 and 2 of model 2, respectively, based on 200 realizations. We observe that the emp_sd and asym_sd decrease and the empirical coverage rate gets close to the nominal level 0.95, as the sample size increases. Moreover, the value of emp_sd is similar to that of asym_sd for each parameter. This result confirms our established formula for the asymptotic variances of the estimators for the parameters. When the sample size is large enough ($n = 1500$), the value of bias is very small compared to asym_sd, so that it can be negligible for constructing confidence intervals of the parameters.

6 Empirical application

In this section, we apply the proposed method to study the community structure of a social network dataset.

Table 3: The empirical coverage rate (coverage), the absolute bias (bias), empirical standard deviation (emp_sd) and asymptotic standard deviation (asym_sd) of the estimators for B_0^* and B_1^* in case 1 of Model 2 based on 200 realizations.

n		$B^*_{0,11}$	$B^*_{0,12}$	$B^*_{0,22}$	$B^*_{1,11}$	$B^*_{1,12}$	$B^*_{1,22}$
	coverage	0.880	0.860	0.975	0.960	0.915	0.955
500	bias	0.023	0.020	0.003	0.002	0.007	0.001
	emp_sd	0.042	0.036	0.014	0.021	0.018	0.009
	asym_sd	0.035	0.029	0.015	0.020	0.017	0.009
	coverage	0.960	0.940	0.945	0.944	0.944	0.940
1000	bias	0.004	0.001	< 0.001	0.002	0.002	< 0.001
	emp_sd	0.017	0.016	0.008	0.010	0.009	0.005
	asym_sd	0.018	0.015	0.008	0.011	0.008	0.005
	coverage	0.940	0.955	0.940	0.940	0.940	0.940
1500	bias	< 0.001	0.001	0.001	0.001	0.001	< 0.001
	emp_sd	0.014	0.011	0.006	0.008	0.006	0.003
	asym_sd	0.013	0.011	0.005	0.007	0.006	0.003

6.1 The dataset and model

The dataset contains Facebook friendship networks at one hundred American colleges and universities at a single point in time. It was provided and analyzed by Traud, Mucha, and Porter (2012), and can be downloaded from https://archive.org/details/oxford-2005-facebook-matrix. Traud et al. (2012) used the dataset to illustrate the relative importance of different characteristics of individuals across different institutions, and showed that gender, dormitory residence and class year may play a role in network partitions by using assortativity coefficients. We, therefore, use these three user attributes as the covariates $X_i = (X_{i1}, X_{i2}, X_{i3})^\top$, where X_{i1} =binary indicator for gender, X_{i2} =multi-category variable for dorm number (e.g. “202”, “203”, etc.), and X_{i3} =integer valued variable for class year (e.g. “2004”, “2005”, etc.). We use the dataset of Rice University to identify the latent community structure interacted with the covariates by our proposed method.

We use the dataset to fit the model:

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \le \tau_n + \Theta^*_{0,ij} + W_{1,ij}\Theta^*_{1,ij}\}, \quad i > j, \tag{6.1}$$

where Y_{ij} is the observed value (0 or 1) of the adjacency matrix in the dataset, and $W_{1,ij} = \{\sum_{k=1}^3 (2D_{ij,k}/\Delta_k)^2\}^{1/2}$, where $\Delta_k = \max(D_{ij,k}) - \min(D_{ij,k})$ and $D_{ij,k} = X_{ik} - X_{jk}$ for $k = 1, 2, 3$. In this model, $(\tau_n, \Theta^*_{0,ij}, \Theta_{1,ij})$ are unknown parameters, and $\Theta^*_{0,ij}$ and $\Theta^*_{1,ij}$ have the latent group structures $\Theta^*_0 = ZB^*_0Z^\top$ and $\Theta^*_1 = ZB^*_1Z^\top$, respectively. Following model 2 in the simulation, we impose that Θ^*_0 and Θ^*_1 share the same community structure. It is worth noting that Roy et al.

Table 4: The empirical coverage rate (coverage), the absolute bias (bias), empirical standard deviation (emp_sd) and asymptotic standard deviation (asym_sd) of the estimators for B_0^* and B_1^* in case of Model 2 based on 200 realizations.

n		$B^*_{0,11}$	$B^*_{0,12}$	$B^*_{0,13}$	$B^*_{0,22}$	$B^*_{0,23}$	$B^*_{0,33}$
	coverage	0.910	0.920	0.900	0.875	0.925	0.960
500	bias	0.018	0.025	< 0.001	0.008	0.002	0.009
	emp_sd	0.033	0.029	0.035	0.030	0.028	0.032
	asym_sd	0.033	0.031	0.032	0.028	0.027	0.032
	coverage	0.915	0.935	0.955	0.930	0.950	0.925
1000	bias	0.005	0.005	0.001	0.004	0.006	0.006
	emp_sd	0.018	0.016	0.015	0.014	0.014	0.017
	asym_sd	0.017	0.015	0.017	0.013	0.014	0.016
	coverage	0.940	0.945	0.940	0.960	0.940	0.955
1500	bias	0.001	0.001	< 0.001	0.001	0.002	< 0.001
	emp_sd	0.012	0.010	0.012	0.008	0.009	0.011
	asym_sd	0.011	0.010	0.011	0.009	0.010	0.011
n		$B^*_{1,11}$	$B^*_{1,12}$	$B^*_{1,13}$	$B^*_{1,22}$	$B^*_{1,23}$	$B^*_{1,33}$
	coverage	0.885	0.900	0.915	0.900	0.960	0.925
500	bias	0.020	0.005	0.001	0.016	< 0.001	0.005
	emp_sd	0.023	0.019	0.020	0.021	0.017	0.022
	asym_sd	0.025	0.019	0.019	0.020	0.016	0.022
	coverage	0.930	0.905	0.945	0.925	0.940	0.930
1000	bias	0.003	0.001	0.006	0.007	0.002	0.002
	emp_sd	0.011	0.011	0.011	0.009	0.008	0.011
	asym_sd	0.012	0.009	0.010	0.009	0.008	0.011
	coverage	0.940	0.955	0.940	0.960	0.960	0.950
1500	bias	< 0.001	< 0.001	< 0.001	0.001	< 0.001	0.001
	emp_sd	0.009	0.006	0.007	0.005	0.005	0.007
	asym_sd	0.008	0.006	0.007	0.006	0.006	0.007

(2019) fit a similar regression model as (6.1) but let the coefficient of the pairwise covariate be an unknown constant with respect to (i, j) such that $\Theta^*_{1,ij} = \Theta^*_1$. Although Roy et al.'s 2019 model can take into account the covariate effect for community detection, it does not consider possible interaction effects of the observed covariates and the latent community structure. As a result, it may cause the number of estimated groups to be inflated. In the dataset of Rice University, we delete the nodes with missing values and with degree less than 10, and consider the class year from 2004 to 2009. After the cleanup, there are $n = 3073$ nodes and 279916 edges in the dataset for our analysis.

6.2 Estimation results

We first use the eigenvalue ratio method to obtain the estimated number of groups for Θ^*_0 and Θ^*_1: $\widehat{K}_0 = 4$ and $\widehat{K}_1 = 4$.

Next, we use our proposed method to obtain the estimated membership for each node. Table 5 presents the number of persons in each estimated group for female and male, for different class years, and for different dorm numbers. It is interesting to observe that most female students belong to either group 2 or group 4, and most male students belong to either group 1 or group 3. There is a clear community division between female and male; within each gender category, the students are further separated into two large groups. Moreover, most students in the class years of 2004 and 2005 are in either group 1 or group 2, while most students in the class years of 2008 and 2009 are in either group 3 or group 4. Students in the class years of 2006 and 2007 are almost evenly distributed across the four groups, with a tendency that more students will join groups 3 and group 4 when they are in later class years. This result indicates that students tend to be in different groups as the gap between their class years becomes larger. Last, Table 6 shows the estimates of B^*_0 and B^*_1 and their standard errors (s.e.). We obtain the p-value< 0.01 for testing each coefficient in B^*_1 equal to zero, indicating that the three covariates are useful for identifying the community structure.

7 Conclusion

In this paper, we proposed a network formation model which can capture heterogeneous effects of homophily via a latent community structure. When the expected degree diverges at a rate no slower than rate-$\log n$, we established that the proposed method can exactly recover the latent community memberships almost surely. By treating the estimated community memberships as the truth, we can then estimate the regression coefficients in the model by existing methods in the literature.

Table 5: The number of persons in each estimated group for female and male, for different class years, and for different dorm numbers.

	gender			class year					
	female	male		2004	2005	2006	2007	2008	2009
group 1	1	515		112	139	147	110	37	1
group 2	540	4		103	135	116	165	50	2
group 3	4	1050		38	79	152	178	277	300
group 4	958	1		30	62	125	156	288	271
	dorm number								
	202	203	204	205	206	207	208	209	210
group 1	71	67	36	42	41	50	57	59	93
group 2	65	98	53	46	20	63	56	56	84
group 3	94	116	142	138	129	130	121	101	83
group 4	92	72	124	125	139	95	122	110	83

Table 6: The estimates of B_0^* and B_1^* and their standard errors (s.e.).

	$B_{0,11}^*$	$B_{0,12}^*$	$B_{0,13}^*$	$B_{0,14}^*$	$B_{0,22}^*$	$B_{0,23}^*$	$B_{0,24}^*$	$B_{0,33}^*$	$B_{0,34}^*$	$B_{0,44}^*$
estimate	-0.730	4.912	-1.543	6.197	-0.751	4.123	-1.624	-1.702	5.933	-1.419
s.e.	0.018	0.112	0.024	0.171	0.017	0.195	0.024	0.017	0.207	0.016
	$B_{1,11}^*$	$B_{1,12}^*$	$B_{1,13}^*$	$B_{1,14}^*$	$B_{1,22}^*$	$B_{1,23}^*$	$B_{1,24}^*$	$B_{1,33}^*$	$B_{1,34}^*$	$B_{1,44}^*$
estimate	-3.397	-6.381	-4.398	-5.656	-3.600	-5.628	-4.387	-6.384	-6.704	-7.567
s.e.	0.042	0.102	0.057	0.155	0.042	0.180	0.059	0.059	0.196	0.060

Appendix

A Proofs of the Main Results

In this appendix, we prove the main results in the paper. Given the fact that our proofs involve a lot of constants defined in the assumptions and proofs, we first provide a list of these constants in Appendix A.1. Then we prove Lemma 2.1 and Theorems 4.1–4.4 in Appendices A.2–A.6, respectively.

A.1 List of constants

Before we prove the main results, we first list all the constants in Table 7. We specify each constant to illustrate that all our results hold as long as $\sqrt{\log n/(n\zeta_n)} \le c_F \le \frac{1}{4}$ for some sufficiently small constant c_F. Apparently, if $\log n/(n\zeta_n) \to 0$, c_F can be arbitrarily small as long as n is sufficiently large. Then all the rate requirements in the proof hold automatically. However, $\log n/(n\zeta_n) \to 0$ is sufficient but not necessary.

Table 7: Table of constants

Name	Description
M_W	$\|W_{1,ij}\| \le M_W$.
M	$\max_{i\in[n],l=0,1} \|\Theta^*_{l,ij}\| \le M$, used in the definition of $f_M(\cdot)$.
C_λ	Used in the definition of $\lambda_n^{(1)}$.
C_M	Used in the definition of $\mathbb{T}^{(1)}$.
$C_\sigma, c_\sigma, C_1, c_1$	Defined in Assumption 3.
κ	Defined in Assumption 4.
$\overline{c}, \underline{c}, C_{0,u}, c_F$	Defined in Assumption 5.
C_ϕ, c_ϕ	Defined in Assumption 6.
$C_F, C_{F,1}, C_{F,2}$	Defined in Theorem 4.1.
C_1^*	Defined in Theorem 4.2.
$C_{h,u}, C_{h,v}$	Defined in Theorem 4.3.
C_Υ	Defined in Lemma S1.1.

A.2 Proof of Lemma 2.1

We prove the results for U_1 first. Let $\Pi_{1,n} = Z_1^\top Z_1/n = \mathrm{diag}(\pi_{1,1n}, \cdots, \pi_{1,K_1 n})$. Then,

$$(n^{-1}\Theta_1^*)(n^{-1}\Theta_1^*)^\top = n^{-1}Z_1 B_1^* \Pi_n B_1^* Z_1^\top.$$

Consider the spectral decomposition of $\chi \equiv \Pi_{1,n}^{1/2} B_1^* \Pi_{1,n} B_1^* \Pi_{1,n}^{1/2} : \chi = S_1' \tilde{\Omega}_l^2 (S_1')^\top$. Let $\mathcal{U}_1 = Z_1(Z_1^\top Z_1)^{-1/2} S_1'$, where S_1 is a $K_1 \times K_1$ matrix such that $(S_1')^\top S_1' = I_{K_1}$. Then, we have

$$\mathcal{U}_1 \tilde{\Omega}_1^2 \mathcal{U}_1^\top = n^{-1} Z_1 \Pi_n^{-1/2} S_1 \tilde{\Omega}_1^2 S_1^\top \Pi_n^{-1/2} Z_1^\top = n^{-1} Z_1 B_1^* \Pi_n B_1^* Z_1^\top = (n^{-1}\Theta_1^*)^2.$$

In addition, note that $\mathcal{U}_1^\top \mathcal{U}_1 = I_{K_1}$ and $\tilde{\Omega}_1^2$ is a diagonal matrix. This implies $\tilde{\Omega}_1^2 = \Sigma_1^2$ (after reordering the eigenvalues) and $\mathcal{U}_l$ is the corresponding singular vector matrix. Then, by definition,

$$U_1 = \sqrt{n}\mathcal{U}_1 \Sigma_1 = Z_1(\Pi_{1,n})^{-1/2} S_1' \Sigma_1.$$

Similarly, by considering the spectral decomposition of $(n^{-1}\Theta_1^*)^\top (n^{-1}\Theta_1^*)$, we can show that $V_1 = Z_1(\Pi_{1,n})^{-1/2} S_1$ for some rotation matrix S_1. Parts (2) and (3) can be verified directly by noting that S_1 and S_1' are orthonormal, $\Pi_{1,n}$ is diagonal, and Assumption 3 holds.

A.3 Proof of Theorem 4.1

We focus on the split-sample low-rank estimators. The full-sample results can be derived in the same manner. Denote $Q_{n,ij}(\Gamma_{ij}) = -[Y_{ij} \log(\Lambda(W_{ij}^\top \Gamma_{ij})) + (1 - Y_{ij}) \log(1 - \Lambda(W_{ij}^\top \Gamma_{ij}))]$, which is a convex function for each element in $\Gamma_{ij} = (\Gamma_{0,ij}, \Gamma_{1,ij})^\top$. In addition, we note that the true parameter $\Gamma^*(I_1) \in \mathbb{T}^{(1)}(0, \log n)$. Denote $\widetilde{\Gamma}^{(1)} = \{\widetilde{\Gamma}_{ij}^{(1)}\}_{i \in I_1, j \in [n]}$, $\widetilde{\Gamma}_{ij}^{(1)} = (\widetilde{\Gamma}_{0,ij}^{(1)}, \widetilde{\Gamma}_{1,ij}^{(1)})^\top$ and $\Delta_{ij} = \widetilde{\Gamma}_{ij}^{(1)} - \Gamma_{ij}^* \equiv (\Delta_{0,ij}, \Delta_{1,ij})^\top$, for $i \in I_1$, $j \in [n]$. Then, we have

$$\begin{aligned}
\lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* \right) \geq & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \left(Q_{n,ij}(\widetilde{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) \right) \\
\geq & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \left(\partial_{\Gamma_{ij}} Q_{n,ij}^\top(\Gamma_{ij}^*) \right)^\top \Delta_{ij} \\
= & \frac{-1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], i \neq j} \left(Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}^*) \right) W_{ij}^\top \Delta_{ij} \\
\equiv & \frac{-1}{n_1(n-1)} \sum_{l=0}^{1} \text{trace}(\Upsilon_l^\top \Delta_l), \qquad \text{(A.1)}
\end{aligned}$$

where $\partial_{\Gamma_{ij}} Q_{n,ij}^\top(\Gamma_{ij}^*) = \partial Q_{n,ij}(\Gamma_{ij}^*)/\partial \Gamma_{ij}$, Υ_l is an $n_1 \times n$ matrix with (i,j)-th entry

$$\Upsilon_{l,ij} = \begin{cases} \left(Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}^*) \right) W_{l,ij} & \text{if} \quad i \in I_1, j \in [n], \ j \neq i \\ 0 & \text{if} \quad i = j \in I_1 \end{cases},$$

and trace$(\cdot)$ is the trace operator. By (A.1), we have

$$0 \leq \lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* \right) + \frac{1}{n_1(n-1)} \left| \sum_{l=0}^{1} \text{trace}(\Upsilon_l^\top \Delta_l) \right|$$

$$\leq \lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* \right) + \frac{1}{n_1(n-1)} \sum_{l=0}^{1} ||\Upsilon_l||_{op} ||\Delta_l||_*. \tag{A.2}$$

By Chernozhukov et al. (2018, Lemma C.2) and the fact that Γ_0^* and Γ_1^* are exact low-rank matrices with ranks upper bounded by K_0+1 and K_1, respectively, there exist $\{\Delta_l', \Delta_l''\}_{l=0}^1$ such that $\Delta_l = \Delta_l' + \Delta_l''$, $\text{rank}(\Delta_0') \leq 2K_0+2$, $\text{rank}(\Delta_1') \leq 2K_1$, and for $l = 0, 1$,

$$||\Delta_l||_F^2 = ||\Delta_l'||_F^2 + ||\Delta_l''||_F^2 \text{ and } ||\Gamma_l^*(I_1) + \Delta_l''||_* = ||\Gamma_l^*(I_1)||_* + ||\Delta_l''||_*. \tag{A.3}$$

This implies that

$$||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* = ||\Gamma_l^*(I_1)||_* - ||\Gamma_l^*(I_1) + \Delta_l' + \Delta_l''||_* \leq ||\Delta_l'||_* - ||\Delta_l''||_*, \quad l = 0, 1. \tag{A.4}$$

Therefore, combining (A.2), Lemma S1.1, and (A.4), we have

$$0 \leq \lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Delta_l'||_* - ||\Delta_l''||_* \right) + \frac{C_\Upsilon M_W(\sqrt{\zeta_n n} + \sqrt{\log n})}{n_1(n-1)} \sum_{l=0}^{1} \left(||\Delta_l'||_* + ||\Delta_l''||_* \right).$$

Noting that $\lambda_n^{(1)} = \frac{C_\lambda(\sqrt{\zeta_n n} + \sqrt{\log n})}{n_1(n-1)}$ and $C_\lambda > C_\Upsilon M_W$, the last inequality implies that

$$(C_\lambda - C_\Upsilon M_W) \sum_{l=0}^{1} ||\Delta_l''||_* \leq (C_\lambda + C_\Upsilon M_W) \sum_{l=0}^{1} ||\Delta_l'||_*, \tag{A.5}$$

and that $(\Delta_0, \Delta_1) \in \mathcal{C}(\tilde{c})$ for $\tilde{c} = \frac{C_\lambda + C_\Upsilon M_W}{C_\lambda - C_\Upsilon M_W} > 0$, with a slight abuse of notation. Note although Δ_l' and Δ_l'' are $n_1 \times n$ matrices, we can make them square matrices by adding rows of zeros. This will not affect the matrices' nuclear norm, and thus, (A.5) still holds for the associated square matrices.

Next, we consider the second-order Taylor expansion of $Q_{n,ij}(\Gamma_{ij})$, following the argument in Belloni, Chernozhukov, Fernández-Val, and Hansen (2017). Let $f_{ij}(t) = \log\{1 + \exp(W_{ij}^\top(\Gamma_{ij}^* + t\Delta_{ij}))\}$, where $\Delta_{ij} = (\Delta_{0,ij}, \cdots, \Delta_{p,ij})^\top$. Then,

$$Q_{n,ij}(\widetilde{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) - \partial_{\Gamma_{ij}} Q_{n,ij}^\top(\Gamma_{ij}^*)\Delta_{ij} = f_{ij}(1) - f_{ij}(0) - f_{ij}'(0).$$

Note that $f_{ij}(\cdot)$ is a three times differentiable convex function such that for all $t \in \mathbb{R}$,

$$\begin{aligned} |f_{ij}'''(t)| =& |W_{ij}^\top \Delta_{ij}|^3 \Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij}))(1 - \Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij}))) |1 - 2\Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij}))| \\ \leq& |W_{ij}^\top \Delta_{ij}| f_{ij}''(t). \end{aligned}$$

Then, by Bach (2010, Lemma 1) we have

$$\begin{aligned}
f_{ij}(1) - f_{ij}(0) - f'_{ij}(0) \geq & \frac{f''_{ij}(0)}{(W_{ij}^\top \Delta_{ij})^2} \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
= & \Lambda(W_{ij}^\top \Gamma_{ij}^*)(1 - \Lambda(W_{ij}^\top \Gamma_{ij}^*)) \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
\geq & \underline{c}\zeta_n \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
\geq & \underline{c}\zeta_n \left(\frac{(W_{ij}^\top \Delta_{ij})^2}{4(\max_{i,j} |W_{ij}^\top \Delta_{ij}| \vee \log(2))} \right) \\
\geq & \frac{\zeta_n \underline{c} (W_{ij}^\top \Delta_{ij})^2}{8(M_W + 1) \log n}, \qquad (A.6)
\end{aligned}$$

where the third inequality holds by Lemma S1.2 and the last inequality holds because of Assumption 5 and the fact that $|W_{ij}^\top \Delta_{ij}| \leq |\widetilde{\Gamma}_{0,ij} - \Gamma_{0,ij}| + M_W |\widetilde{\Gamma}_{1,ij} - \Gamma_{1,ij}| \leq 2(M_W + 1) \log n$. Therefore,

$$\begin{aligned}
F_n(\Delta_0, \Delta_1) \equiv & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], j \neq i} \left[Q_{n,ij}(\widetilde{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) - \partial_{\Gamma_{ij}} Q_{n,ij}^\top (\Gamma_{ij}^*) \Delta_{ij} \right] \\
\geq & \frac{\zeta_n \underline{c}}{8 n_1 (n-1)(M_W + 1) \log n} \sum_{i \in I_1, j \in [n], j \neq i} (W_{ij}^\top \Delta_{ij})^2 \\
\geq & \frac{\zeta_n \underline{c}}{8 n_1 (n-1)(M_W + 1) \log n} \left[\kappa \sum_{l=0}^{1} ||\Delta_l||_F^2 - 4(M_W + 1)^2 (\log n)^2 n_1 \right], \qquad (A.7)
\end{aligned}$$

where, by an abuse of notation, we still view Δ_l as a square matrix with some rows filled in by zeros, and the last inequality holds by Assumption 4 and the fact that $|\Delta_{l,ii}| \leq 2 \log n$, $i \in I_1$.

On the other hand, by (A.1),

$$\begin{aligned}
F_n(\Delta_0, \Delta_1) \leq & \lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* \right) + \left| \frac{1}{n_1(n-1)} \sum_{l=0}^{1} \operatorname{trace}(\Upsilon_l^\top \Delta_l) \right| \\
\leq & \lambda_n^{(1)} \sum_{l=0}^{1} \left(||\Delta'_l||_* - ||\Delta''_l||_* \right) + \frac{1}{n_1(n-1)} \sum_{l=0}^{1} ||\Upsilon_l||_{op} ||\Delta_l||_* \\
\leq & \frac{\sqrt{\zeta_n n} + \sqrt{\log n}}{n_1(n-1)} \left[\sum_{l=0}^{1} (C_\lambda + C_\Upsilon M_W) ||\Delta'_l||_* - \sum_{l=0}^{1} (C_\lambda - C_\Upsilon M_W) ||\Delta''_l||_* \right] \\
\leq & \frac{\sqrt{\zeta_n n} + \sqrt{\log n}}{n_1(n-1)} (C_\lambda + C_\Upsilon M_W) (\sum_{l=0}^{1} ||\Delta'_l||_*) \\
\leq & \frac{\sqrt{\zeta_n n} + \sqrt{\log n}}{n_1(n-1)} (C_\lambda + C_\Upsilon M_W) \sqrt{2\bar{K}} (\sum_{l=0}^{1} ||\Delta'_l||_F)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\sqrt{\zeta_n n}+\sqrt{\log n}}{n_1(n-1)}(C_\lambda + C_\Upsilon M_W)\sqrt{2\bar{K}}(\sum_{l=0}^{1}||\Delta_l||_F) \\
&\leq \frac{\sqrt{\zeta_n n}+\sqrt{\log n}}{n_1(n-1)}(C_\lambda + C_\Upsilon M_W)2\sqrt{\bar{K}}(\sum_{l=0}^{1}||\Delta_l||_F^2)^{1/2},
\end{aligned}
\tag{A.8}
$$

where $\bar{K} = \max(K_0+1, K_1)$, the first inequality is due to (A.1), the second inequality is due to (A.4) and the trace inequality, the third inequality holds by the definition of $\lambda_n^{(1)}$ and Lemma S1.1, the fourth inequality is due to the fact that $C_\lambda - C_\Upsilon M_W > 0$, the fifth inequality is due to the fact that $\text{rank}(\Delta_l') \leq 2\bar{K}$, the second last inequality is due to (A.3), and the last inequality is due to the Cauchy's inequality.

Combining (A.7) and (A.8), we have

$$
\begin{aligned}
&\left[(\sum_{l=0}^{1}||\Delta_l||_F^2)^{1/2} - \frac{8\sqrt{\bar{K}}(M_W+1)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa}\frac{\log n[\sqrt{n\zeta_n}+\sqrt{\log n}]}{\zeta_n}\right]^2 \\
\leq& \bar{K}\left[\frac{8(M_W+1)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa}\right]^2\left(\frac{\log n[\sqrt{n\zeta_n}+\sqrt{\log n}]}{\zeta_n}\right)^2 + \frac{4n_1(M_W+1)^2(\log n)^2}{\kappa},
\end{aligned}
$$

and thus

$$
\frac{1}{n}(\sum_{l=0}^{1}||\Delta_l||_F^2)^{1/2} \leq 17C_F\left(\frac{\log n}{\sqrt{n\zeta_n}} + \frac{(\log n)^{3/2}}{n\zeta_n}\right)\ a.s., \tag{A.9}
$$

where $C_F = \frac{\sqrt{\bar{K}}(M_W+1)(C_\lambda+C_\Upsilon M_W)}{\underline{c}\kappa}$. Then,

$$
\begin{aligned}
|\widetilde{\tau}_n^{(1)} - \tau_n| =& \left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}(\widetilde{\Gamma}_{0,ij} - \tau_n)\right| \leq \left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}(\widetilde{\Gamma}_{0,ij} - \Gamma_{0,ij}^*)\right| + \left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}\Theta_{0,ij}^*\right| \\
\leq& \frac{1}{\sqrt{n_1 n}}||\Delta_0||_F + M \leq 30C_F\left(\frac{\log n}{\sqrt{n\zeta_n}} + \frac{(\log n)^{3/2}}{n\zeta_n}\right) \\
\leq& 30C_F(c_F + c_F^2)\sqrt{\log n},
\end{aligned}
\tag{A.10}
$$

where the last inequality follows Assumption 5.3.

Next, we rerun the nuclear norm regularized logistic regression with the parameter space restriction $\mathbb{T}^{(1)}(0, \log n)$ replaced by $\mathbb{T}^{(1)}(\widetilde{\tau}_n^{(1)}, C_M\sqrt{\log n})$. First, we note that the true parameter $\Gamma^*(I_1) \in \mathbb{T}^{(1)}(\widetilde{\tau}_n^{(1)}, C_M\sqrt{\log n})$ because

$$
|\Gamma_{0,ij}^* - \widetilde{\tau}_n^{(1)}| \leq |\Theta_{0,ij}^*| + |\widetilde{\tau}_n^{(1)} - \tau_n| \leq |\Theta_{0,ij}^*| + 30C_F(c_F + c_F^2)\sqrt{\log n} \leq C_M\sqrt{\log n}, \tag{A.11}
$$

where we use the fact that $|\Theta_{1,ij}^*| \leq \sqrt{\log n}$ and c_F, and thus, $30(c_F + c_F^2)C_F$ is sufficiently small.

Therefore, following the same arguments used to obtain (A.5), we can show that $\widehat{\Delta} \equiv (\widehat{\Delta}_0, \widehat{\Delta}_1) \in \mathcal{C}(\tilde{c})$, where $\widehat{\Delta}_l = \widehat{\Gamma}_l^{(1)} - \Gamma_l^*(I_1)$. Let $\widehat{\Delta}_{ij} = (\widehat{\Delta}_{0,ij}, \widehat{\Delta}_{1,ij})^\top$. Now let $f_{ij}(t) = \log(1 + \exp(W_{ij}^\top(\Gamma_{ij}^* +$

$t\widehat{\Delta}_{ij})))$. Then, following (A.6),

$$f_{ij}(1) - f_{ij}(0) - f'_{ij}(0) \geq \underline{c}\zeta_n \left(\frac{(W_{ij}^\top \widehat{\Delta}_{ij})^2}{4(\max_{i,j} |W_{ij}^\top \widehat{\Delta}_{ij}| \vee \log(2))} \right) \geq \frac{\zeta_n \underline{c} (W_{ij}^\top \widehat{\Delta}_{ij})^2}{8(C_M + M_W)\sqrt{\log n}},$$

where the last inequality holds because of (A.11) and uniformly in (i, j)

$$\begin{aligned} |W_{ij}^\top \widehat{\Delta}_{ij}| \leq & |\hat{\Gamma}_{0,ij}^{(1)} - \Gamma_{0,ij}^*| + M_W |\hat{\Gamma}_{1,ij}^{(1)} - \Theta_{1,ij}^*| \\ \leq & |\hat{\Gamma}_{0,ij}^{(1)} - \widetilde{\tau}_n^{(1)}| + |\widetilde{\tau}_n^{(1)} - \Gamma_{0,ij}^*| + M_W(\sqrt{\log n} + M) \leq 2(C_M + M_W)\sqrt{\log n}. \end{aligned}$$

Then, similar to (A.7) and (A.8),

$$\begin{aligned} F_n(\widehat{\Delta}_0, \widehat{\Delta}_1) \equiv & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], j \neq i} \left(Q_{n,ij}(\widehat{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) - \partial_{\Gamma_{ij}} Q_{n,ij}^\top (\Gamma_{ij}^*) \widehat{\Delta}_{ij} \right) \\ \geq & \frac{\zeta_n \underline{c}}{8n_1(n-1)(M_W + C_M)\sqrt{\log n}} \left[\kappa \left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2 \right) - 4(M_W + C_M)^2 \log n n_1 \right] \end{aligned}$$

and

$$F_n(\widehat{\Delta}_0, \widehat{\Delta}_1) \leq \frac{\sqrt{\zeta_n n} + \sqrt{\log n}}{n_1(n-1)} (C_\lambda + C_\Upsilon M_W) 2\sqrt{\bar{K}} (\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2)^{1/2}.$$

Therefore, we have

$$\begin{aligned} & \left[\left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2 \right)^{1/2} - \frac{8\sqrt{\bar{K}}(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} \left(\frac{\sqrt{\log n}(\sqrt{n\zeta_n} + \sqrt{\log n})}{\zeta_n} \right) \right]^2 \\ \leq & \bar{K} \left[\frac{8(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} \right]^2 \left(\frac{\sqrt{\log n}(\sqrt{n\zeta_n} + \sqrt{\log n})}{\zeta_n} \right)^2 + 4(M_W + C_M)^2 \log n n_1, \end{aligned}$$

and thus,

$$\frac{1}{n} \left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2 \right)^{1/2} \leq 17 C_{F,1} \eta_n, \tag{A.12}$$

where $C_{F,1} = \frac{\sqrt{\bar{K}}(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa}$ and $\eta_n = \sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}$. Then, similar to (A.10) and by Assumption 5.5, we have $|\widehat{\tau}_n^{(1)} - \tau_n| \leq \frac{1}{\sqrt{n_1 n}} ||\widehat{\Delta}_0||_F + o(\eta_n) \leq 30 C_{F,1} \eta_n$. This establishes the first result in Theorem 4.1.

In addition,

$$\begin{aligned} & \frac{1}{n} ||\widehat{\Theta}_1^{(1)} - \Theta_1^*(I_1)||_F \\ \leq & \frac{1}{n} \left[\sum_{(i,j) \in I_1 \times I_1 i \neq j} \left(\frac{1}{2} (\widehat{\Gamma}_{1,ij}^{(1)} + \widehat{\Gamma}_{1,ji}^{(1)}) - \Theta_{1,ij}^* \right)^2 + \sum_{(i,j): i \in I_1, j \notin I_1} (\widehat{\Gamma}_{1,ij}^{(1)} - \Theta_{1,ij}^*)^2 \right]^{1/2} + \frac{1}{n} \left(\sum_{i \in I_1} \Theta_{1,ii}^{*2} \right)^{1/2} \end{aligned}$$

$$\leq \frac{1}{n}\left[\sum_{i\in I_1, j\in[n], i\neq j}(\widehat{\Gamma}^{(1)}_{1,ij}-\Theta^*_{1,ij})^2\right]^{1/2}+\frac{1}{n}\left(\sum_{i\in I_1}\Theta^{*2}_{1,ii}\right)^{1/2}$$

$$\leq \frac{1}{n}\left(\sum_{l=0}^{1}||\widehat{\Delta}_l||_F^2\right)^{1/2}+\sqrt{\frac{M^2}{3n}}\leq 18C_{F,1}\eta_n \ a.s.,$$

where the first inequality holds due to the fact that $f_M(\cdot)$ is 1-Lipschitz continuous, $\Theta^*_1=(\Theta^*_1)^\top$, and $|\Theta^*_{1,ij}|\leq M$, and the last inequality holds due to the fact that $\frac{1}{n}\left(\sum_{i\in I_1}(\Theta^*_{1,ii})^2\right)^{1/2}=o(\eta_n)$. Similarly,

$$\frac{1}{n}||\widehat{\Theta}^{(1)}_0-\Theta^*_0(I_1)||_F$$

$$\leq\frac{1}{n}\left[\sum_{(i,j)\in I_1\times I_1, i\neq j}\left(\frac{1}{2}(\widehat{\Gamma}^{(1)}_{0,ij}+\widehat{\Gamma}^{(1)}_{0,ji})-\Theta^*_{0,ij}-\widehat{\tau}^{(1)}_n\right)^2+\sum_{(i,j):i\in I_1, j\notin I_1}(\Gamma^*_{0,ij}-\widehat{\tau}^{(1)}_n-\Theta^*_{0,ij})^2\right]^{1/2}$$

$$+\frac{1}{n}\left(\sum_{i\in I_1}\Theta^{*2}_{0,ii}\right)^{1/2}$$

$$\leq\frac{1}{n}\left[\sum_{i\in I_1, j\in[n], i\neq j}(\widetilde{\Gamma}^{(1)}_{0,ij}-\Gamma^*_{0,ij})^2\right]^{1/2}+|\widehat{\tau}^{(1)}_n-\tau_n|+\sqrt{\frac{M^2}{3n}}\leq 48C_{F,1}\eta_n \ a.s.$$

Then, by the Weyl's inequality, $\max_{k=1,\cdots,K_l}|\widehat{\sigma}^{(1)}_{k,l}-\sigma_{k,l}|\leq 48C_{F,1}\eta_n$ $a.s.$ for $l=0,1$.

Last, noting that $\widehat{V}^{(1)}_l$ consists of the first K_l eigenvectors of $(\frac{1}{n}\widehat{\Theta}^{(1)}_l)^\top(\frac{1}{n}\widehat{\Theta}^{(1)}_l)$, we have

$$\left\|\frac{1}{n}\widehat{\Theta}^{(1)\top}_l\left(\frac{1}{n}\widehat{\Theta}^{(1)}_l\right)-\frac{1}{n}\Theta^{*\top}_l(I_1)\left(\frac{1}{n}\Theta^*_l(I_1)\right)\right\|_{op}\leq\frac{2C_\sigma}{n}||\widehat{\Theta}^{(1)}_l-\Theta^*_l(I_1)||_F\leq 96C_{F,1}C_\sigma\eta_n.$$

Then by the Davis-Kahan $\sin\Theta$ Theorem (Su et al. (2020, Lemma C.1)), we have

$$\begin{aligned}
||\mathcal{V}_l-\widehat{\mathcal{V}}^{(1)}_l\widehat{O}^{(1)}_l||_F &\leq \sqrt{K_l}||\mathcal{V}_l-\widehat{\mathcal{V}}^{(1)}_l\widehat{O}^{(1)}_l||_{op}\leq\frac{96\sqrt{2K_l}C_{F,1}C_\sigma s\eta_n}{\sigma^2_{K_l,l}-96C_{F,1}C_\sigma\eta_n}\\
&\leq \frac{96\sqrt{2K_l}C_{F,1}C_\sigma s\eta_n}{c^2_\sigma-96C_{F,1}C_\sigma\eta_n}\leq\frac{136\sqrt{K_l}C_{F,1}C_\sigma\eta_n}{c^2_\sigma}\\
&\leq 136C_{F,2}\eta_n, \qquad\qquad (A.13)
\end{aligned}$$

where $C_{F,2}=\max_{l=0,1}\sqrt{K_l}C_{F,1}C_\sigma c_\sigma^{-2}$, and the third inequality holds due to Assumption 5 and the second last inequality is due to the fact that we can set c_F to be sufficiently small to ensure that $1-96\sqrt{2}C_{F,1}C_\sigma(c_F+c_F^2)c_\sigma^{-2}\geq\frac{96\sqrt{2}}{136}$.

Recall that $\widehat{V}^{(1)}_l=\sqrt{n}\widehat{\mathcal{V}}^{(1)}_l$ and $V_l=\sqrt{n}\mathcal{V}_l$, we have the desired result that $||V_l-\widehat{V}^{(1)}_l\widehat{O}^{(1)}_l||_F\leq 136C_{F,2}\sqrt{n}\eta_n$. ■

A.4 Proof of Theorem 4.2

First, we prove the first result in the theorem. Let $\Delta_{i,l} = (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}$ for $l = 0, 1$, and $\Delta_{iu} = (\Delta_{i,0}^\top, \Delta_{i,1}^\top)^\top$. Denote

$$\widehat{\Lambda}_{n,ij} = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} u_{i,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij}). \tag{A.14}$$

Recall that $\Lambda_{n,ij} = \Lambda(\tau_n + \sum_{l=0}^1 u_{i,l}^\top v_{j,l} W_{l,ij}) = \Lambda(\tau_n + \Theta_{0,ij}^* + \Theta_{1,ij}^* W_{1,ij})$. Let

$$\tilde{\Lambda}_{n,ij} = \Lambda(\dot{a}_{n,ij}), \tag{A.15}$$

where $\dot{a}_{n,ij}$ is an intermediate value that is between $\tau_n + \Theta_{0,ij}^* + \Theta_{1,ij}^* W_{1,ij}$ and $\widehat{\tau}_n + \sum_{l=0}^1 u_{i,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij}$. Define

$$\widehat{\phi}_{ij}^{(1)} = \begin{bmatrix} (\widehat{O}_0^{(1)})^\top \widehat{v}_{j,0}^{(1)} \\ (\widehat{O}_1^{(1)})^\top \widehat{v}_{j,1}^{(1)} W_{1,ij} \end{bmatrix} \text{ and } \widehat{\Phi}_i^{(1)} = \frac{1}{n_2} \sum_{j \in I_2, j \neq i} \widehat{\phi}_{ij}^{(1)} (\widehat{\phi}_{ij}^{(1)})^\top.$$

Let $\widetilde{\Lambda}_{ij}^{(1)}(\mu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^1 \mu_l^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij})$ and $\ell_{ij}^{(1)}(\mu) = Y_{ij} \log(\widetilde{\Lambda}_{ij}^{(1)}(\mu)) + (1 - Y_{ij}) \log(1 - \widetilde{\Lambda}_{ij}^{(1)}(\mu))$. Define $\widetilde{Q}_{in}^{(1)}(\mu) = \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} \ell_{ij}^{(1)}(\mu)$. Then,

$$\begin{aligned}
0 \geq & Q_{in,U}^{(0)}(\widehat{u}_{i,0}^{(1)}, \widehat{u}_{i,1}^{(1)}) - Q_{in}^{(1)}((\widehat{O}_0^{(1)}) u_{i,0}, (\widehat{O}_1^{(1)}) u_{i,1}) \\
= & \widetilde{Q}_{in}^{(1)}(u_{i,0} + \Delta_{i,0}, u_{i,1} + \Delta_{i,1}) - \widetilde{Q}_{in}^{(1)}(u_{i,0}, u_{i,1}) \\
\geq & \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} \\
& + \frac{1}{n_2} \sum_{j \in I_2, j \neq i} \widehat{\Lambda}_{n,ij} (1 - \widehat{\Lambda}_{n,ij}) \left[\exp(-|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|) + |(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}| - 1 \right] \\
\geq & \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} + \frac{\underline{c}' \zeta_n}{n_2} \sum_{j \in I_2, j \neq i} \left[\exp(-|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|) + |(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}| - 1 \right] \\
\geq & \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} + \frac{\underline{c}' \zeta_n}{n_2} \sum_{j \in I_2, j \neq i} \left[\frac{((\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu})^2}{2} - \frac{|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|^3}{6} \right]
\end{aligned} \tag{A.16}$$

where the second inequality is due to Bach (2010, Lemma 1), the third inequality is due to the fact that $\exp(-t) + t - 1 \geq 0$ and Lemma S1.3(2), the constant $\underline{c}'$ is defined in Lemma S1.3, and the last inequality is due to the fact that $\exp(-t) + t - 1 \geq \frac{t^2}{2} - \frac{t^3}{6}$. The following argument follows Belloni et al. (2017). Let

$$F(\Delta_{iu}) = \widetilde{Q}_{in}^{(1)}(u_{i,0} + \Delta_{i,0}, u_{i,1} + \Delta_{i,1}) - \widetilde{Q}_{in}^{(1)}(u_{i,0}, u_{i,1}) + \frac{1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu},$$

which is convex in Δ_{iu}. Let

$$q_{in} = \inf_{\Delta} \frac{\left[\frac{1}{n_2}\sum_{j\in I_2, j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \Delta)^2\right]^{3/2}}{\frac{1}{n_2}\sum_{j\in I_2, j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \Delta)^3} \quad \text{and} \quad \delta_{in} = \left[\frac{1}{n_2}\sum_{j\in I_2, j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu})^2\right]^{1/2}. \tag{A.17}$$

If $\delta_{in} \leq q_{in}$, then $\frac{1}{n_2}\sum_{j\in I_2, j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu})^3 \leq \delta_{in}^2$, and thus $F(\Delta_{iu}) \geq \frac{\underline{c}'\zeta_n}{3}\delta_{in}^2$. On the other hand, if $\delta_{in} > q_{in}$, let $\tilde{\Delta}_{iu} = \frac{\Delta_{iu} q_{in}}{\delta_{in}}$, then $\left[\frac{1}{n_2}\sum_{j\in I_2, j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \tilde{\Delta}_{iu})^2\right]^{1/2} \leq q_{in}$. Then, we have

$$F(\Delta_{iu}) = F(\frac{\delta_{in}\tilde{\Delta}_{iu}}{q_{in}}) \geq \frac{\delta_{in}}{q_{in}}F(\tilde{\Delta}_{iu}) \geq \frac{\underline{c}'\zeta_n\delta_{in}}{3n_2q_{in}}\sum_{j\in[n], j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top \tilde{\Delta}_{iu})^2 = \frac{\underline{c}'\zeta_n q_{in}\delta_{in}}{3}.$$

Therefore, by Lemma S1.4,

$$F(\Delta_{iu}) \geq \min\left(\frac{\underline{c}'\zeta_n\delta_{in}^2}{3}, \frac{\underline{c}'\zeta_n q_{in}\delta_{in}}{3}\right) \geq \min\left(\frac{\underline{c}'c_\phi\zeta_n\underline{c}||\Delta_{iu}||^2}{6}, \frac{\underline{c}'\zeta_n q_{in}\sqrt{c_\phi}||\Delta_{iu}||}{3\sqrt{2}}\right). \tag{A.18}$$

On the other hand, we have $|F(\Delta_{iu})| \leq \left|\frac{1}{n}\sum_{j\in I_2, j\neq i}(Y_{ij} - \widehat{\Lambda}_{n,ij})(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}\right| \leq I_i + II_i$, where

$$I_i = \left|\frac{1}{n}\sum_{j\in I_2, j\neq i}(Y_{ij} - \Lambda_{n,ij})(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}\right| \text{ and } II_i = \left|\frac{1}{n}\sum_{j\in I_2, j\neq i}(\widehat{\Lambda}_{n,ij} - \Lambda_{n,ij})(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}\right|.$$

We aim to upper bound I_i and II_i uniformly in i below.

We first bound II_i. Note that

$$\begin{aligned}
II_i \leq& \frac{1}{n_2}\sum_{j\in I_2, j\neq i}\tilde{\Lambda}_{n,ij}(1-\tilde{\Lambda}_{n,ij})\left(|\widehat{\tau}_n - \tau_n| + \sum_{l=0}^{1}\left|u_{i,l}^\top((\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)} - v_{j,l})W_{l,ij}\right|\right)|(\widehat{\phi}_{ij}^{(1)})^\top\Delta_{iu}| \\
\leq& \frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{n_2 c_\sigma}\sum_{j\in I_2, j\neq i}\left(|\widehat{\tau}_n - \tau_n| + \sum_{l=0}^{1}\left|u_{i,l}^\top((\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)} - v_{j,l})W_{l,ij}\right|\right) \\
\leq& \frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{c_\sigma}\left[48C_{F,1}\eta_n + c_{II}\sum_{l=0}^{1}\frac{1}{n_2}\sum_{j\in I_2, j\neq i}\left\|(\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)} - v_{j,l}\right\|\right] \\
\leq& \frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{c_\sigma}\left[48C_{F,1}\eta_n + c_{II}\sum_{l=0}^{1}\frac{1}{\sqrt{n_2}}\left\|\widehat{V}_l^{(1)}\widehat{O}_l^{(1)} - V_l\right\|_F\right] \\
\leq& C_{II}||\Delta_{iu}||\zeta_n\eta_n,
\end{aligned} \tag{A.19}$$

where $c_{II} = \max(C_{0,u}, c_1^{-1/2}M_W)C_\sigma$, $C_{II} = 2\overline{c}'M(1+M_W)\zeta_n\left(48C_{F,1} + 136c_{II}C_{F,2}\right)c_\sigma^{-1}$, the first inequality holds by the Taylor expansion, the second inequality holds by Lemma S1.3

$$\max_{i,j\in I_2, i\neq j}||\widehat{\phi}_{ij}^{(1)}|| \leq \max_{i,j\in I_2, i\neq j}\left(||\widehat{O}_0^{(1)}\widehat{v}_{j,0}^{(1)}|| + M_W||\widehat{O}_1^{(1)}\widehat{v}_{j,1}^{(1)}||\right)$$

$$\leq 2M\sigma_{K_0,0}^{-1} + 2M_W M\sigma_{K_1,1}^{-1} \leq 2M(1+M_W)c_\sigma^{-1}, \tag{A.20}$$

the third inequality is due to Theorem 4.1 and the fact that $||u_{i,l}^\top W_{l,ij}|| \leq c_{II}$, the fourth inequality is due to Cauchy's inequality, and the last inequality is due to Theorem 4.1. Note that the constant C_{II} does not depend on i, the above upper bound for II_i holds uniformly over i.

Next, we turn to the upper bound for I_i. Let $\mathcal{F}_n$ be the σ-field generated by $\{X_i\}_{i=1}^n \cup \{\varepsilon_{ij}\}_{i\in I_1, j\in[n], j\neq i} \cup \{e_{ij}\}_{1\leq i,j\leq n}$ and $H_{ij} = (Y_{ij} - \Lambda_{n,ij})\widehat{\phi}_{ij}^{(1)}$. Then, conditional on $\mathcal{F}_n$, $\{H_{ij}\}_{j\in I_2, j\neq i}$ only depends on $\{\varepsilon_{ij}\}_{j\in I_2, j\neq i}$, and thus, is a sequence of independent random vectors. Note that $I_i \leq ||\frac{1}{n_2}\sum_{j\in I_2, j\neq i} H_{ij}||||\Delta_{iu}||$. Let $H_{k,ij}$ be the k-th coordinate of H_{ij} where $k \in [K_0 + K_1]$. By Lemma S1.3, (A.20) and Assumption 5,

$$\max_{1\leq i,j\leq n} |H_{k,ij}| \leq \left[2M(1+M_W)c_\sigma^{-1} + 1\right]^2 (1+\bar{c}) \equiv C_H \tag{A.21}$$

and $\sum_{j\in I_2, j\neq i} \mathbb{E}(H_{k,ij}^2 | \mathcal{F}_n) \leq C_H \zeta_n n_2$. Therefore, by the Bernstein inequality, for any $t > 0$,

$$\mathbb{P}\left(\max_{i\in I_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq n_2 t \middle| \mathcal{F}_n\right) \leq \sum_{i\in I_2} 2\exp\left(-\frac{\frac{n_2^2 t^2}{2}}{C_H\zeta_n n_2 + \frac{C_H t n_2}{3}}\right).$$

Taking $t = 4C_H\sqrt{\frac{\zeta_n \log n}{n}}$, we have

$$\begin{aligned}\mathbb{P}\left(\max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq t \middle| \mathcal{F}_n\right) \leq & 2n_2\exp\left(-\frac{\frac{16C_H^2\zeta_n \log n n_2^2}{2n}}{C_H\zeta_n n_2 + \frac{4C_H^2\sqrt{\frac{\zeta_n \log n}{n}} n_2}{3}}\right)\\ &\leq 2n_2\exp\left(-\frac{8\log n}{7}\right) \leq n^{-1.1},\end{aligned}$$

where the second inequality holds because $\log n/(n\zeta_n) \leq c_F < 1$ and $C_H > 1$. Applying Expectation on both sides, we have

$$\mathbb{P}\left(\max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq t\right) \leq n^{-1.1}.$$

Then by Borel-Cantelli Lemma, we have

$$\max_{i\in I_2} I_i \leq \max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \leq 4C_H\sqrt{\frac{\log n\zeta_n}{n}}\ a.s. \tag{A.22}$$

Combining (A.19) and (A.22), we have

$$|F(\Delta_{iu})| \leq (4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}||. \tag{A.23}$$

Then, (A.18) and (A.23) imply

$$(4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}|| \geq \min\left(\frac{\underline{c}c_\phi\zeta_n|||\Delta_{iu}||^2}{6}, \frac{\underline{c}\sqrt{c_\phi}\zeta_n q_{in}||\Delta_{iu}||}{3\sqrt{2}}\right). \tag{A.24}$$

On the other hand, by Lemma S1.5, we have

$$\liminf_n \min_{i\in I_2} \frac{\underline{c}'\sqrt{c_\phi}\zeta_n q_{in}||\Delta_{iu}||}{3\sqrt{2}} \geq \frac{c_\phi c_\sigma}{2M(1+M_W)}\frac{\underline{c}'\sqrt{c_\phi}\zeta_n||\Delta_{iu}||}{3\sqrt{2}} > (4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}||,$$

where the first inequality holds by Lemma S1.5 and the second inequality holds due to the fact that c_F is sufficiently small so that

$$(4C_H + C_{II})(c_F + c_F^2) < \frac{\underline{c}'(c_\phi)^{3/2}c_\sigma}{6\sqrt{2}M(1+M_W)}.$$

Therefore, (A.24) implies

$$||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \leq ||\Delta_u|| \leq \frac{6(4C_H + C_{II})}{\underline{c}c_\phi}\eta_n \leq C_1^*\eta_n \ a.s., \tag{A.25}$$

where $C_1^* = \max\left(\frac{6(4C_H+C_{II})}{\underline{c}c_\phi}, \frac{12(4C_H+C_{II})}{\underline{c}c_\phi c_\sigma} + \frac{4C_\sigma C_u C_\Sigma}{c_\sigma^2} + 48C_{F,1}C_u\right)$. Because the constant C_1^* does not depend on index i, the above inequality holds uniformly over $i \in I_2$.

Now, we prove the second result in the theorem. The proof follows that of the first result with a notable difference: $\{\widehat{u}_{i,l}^{(1)}\}_{i\in I_2, l=0,1}$ are not independent of the observations $\{Y_{ij}\}$ given the covariates, thus the conditional Bernstein inequality argument cannot be directly used. Recall that

$$(\dot{v}_{j,0}^{(0,1)}, \dot{v}_{j,1}^{(0,1)}) = \arg\min Q_{jn,V}^{(0)}(\nu_0, \nu_1),$$

where $Q_{jn,V}^{(0)}(\nu)$ with $\nu = (\nu_0^\top, \nu_1^\top)^\top$ is defined in Section 3.3. Let $\tilde{\Lambda}_{ij}^{(0)}(\nu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^1 \nu_l^\top (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij})$, $\ell_{ij}^{(0)}(\nu) = Y_{ij}\log(\tilde{\Lambda}_{ij}^{(0)}(\nu)) + (1-Y_{ij})\log(1-\tilde{\Lambda}_{ij}^{(0)}(\nu))$. Define $\widetilde{Q}_{jn,V}^{(0)}(\nu) = \frac{-1}{n_2}\sum_{j\in I_2, j\neq i}\ell_{ij}^{(1)}(\nu)$. Then

$$Q_{jn,V}^{(0)}(\nu_0, \nu_1) = \widetilde{Q}_{jn,V}^{(0)}((\widehat{O}_0^{(1)})^\top\nu_0, (\widehat{O}_1^{(1)})^\top\nu_1).$$

Recall that $\Lambda_{n,ij} = \Lambda(\tau_n + \sum_{l=0}^1 u_{i,l}^\top v_{j,l} W_{l,ij}) = \Lambda(\tau_n + \Theta_{0,ij}^* + \Theta_{1,ij}^* W_{1,ij})$. Let $\dot{\Lambda}_{n,ij} = \Lambda(\widehat{\tau}_n + \sum_{l=0}^1 v_{j,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij})$ and $\tilde{\Lambda}_{n,ij} = \Lambda(\dot{a}_{n,ij})$, where $\dot{a}_{n,ij}$ is an intermediate value that is between $\tau_n + \Theta_{0,ij}^* + \Theta_{1,ij}^* W_{1,ij}$ and $\widehat{\tau}_n + \sum_{l=0}^1 v_{j,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij}$. Define

$$\dot{\psi}_{ij} = \begin{bmatrix} (\widehat{O}_0^{(1)})^\top \widehat{u}_{i,0}^{(1)} \\ (\widehat{O}_1^{(1)})^\top \widehat{u}_{i,1}^{(1)} W_{1,ij} \end{bmatrix} \quad \text{and} \quad \dot{\Psi}_j = \frac{1}{n_2}\sum_{i\in I_2, i\neq j}\dot{\psi}_{ij}(\dot{\psi}_{ij})^\top.$$

Let $\Delta_{jv} \equiv (\Delta_{j,0}^\top, \Delta_{j,1}^\top)^\top$, where $\Delta_{j,l} = (\widehat{O}_l^{(1)})^\top \dot{v}_{j,l}^{(0,1)} - v_{j,l}$ for $l = 0, 1$. Then we have

$$
\begin{aligned}
0 \geq & Q_{jn,V}^{(0)}(\dot{v}_{j,0}^{(0,1)}, \dot{v}_{j,1}^{(0,1)}) - Q_{jn,V}^{(0)}((\widehat{O}_0^{(1)})^\top v_{j,0}, (\widehat{O}_1^{(1)})^\top v_{j,1}) \\
= & \widetilde{Q}_{jn,V}^{(0)}((\widehat{O}_0^{(1)})^\top \dot{v}_{j,0}^{(0,1)}, (\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(0,1)}) - \widetilde{Q}_{jn,V}^{(0)}(v_{j,0}, v_{j,1}) \\
\geq & \frac{-1}{n} \sum_{i \in I_2, i \neq j} (Y_{ij} - \dot{\Lambda}_{n,ij})(\dot{\psi}_{ij})^\top \Delta_v + \frac{\underline{c}'\zeta_n}{n} \sum_{i \in I_2, i \neq j} \left[\frac{((\dot{\psi}_{ij})^\top \Delta_v)^2}{2} - \frac{|(\dot{\psi}_{ij})^\top \Delta_v|^3}{6} \right].
\end{aligned}
$$

By the first result that $\max_{i \in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \leq C_1^* \eta_n$, we have

$$
\max_{i \in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij}|| \leq M_W \max_{i \in I_2} \left[||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| + ||u_{i,l}|| \right] \leq M_W(C_1^* \eta_n + C_\sigma C_u) < \infty.
$$

Therefore, similar to (S1.1), we have

$$
\begin{aligned}
||\dot{\Psi}_j - \Psi_j(I_2)|| \leq & \frac{2M_W(C_1^* \eta_n + C_\sigma C_u)}{n} \sum_{l=0}^{1} \sum_{i \in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \\
\leq & 4M_W(C_1^* \eta_n + C_\sigma C_u) C_1^* \eta_n \ a.s.
\end{aligned}
$$

As c_F is sufficiently small so that $4M_W(C_1^* \eta_n + C_\sigma C_u) C_1^*(c_F + c_F^2) \leq c_\phi/2$ can be ensured and Assumption 7 holds, we have $\min_{j \in [n]} \lambda_{\min}(\dot{\Psi}_j) \geq c_\phi/2 \ a.s.$

Let

$$
F(\Delta_{jv}) = \widetilde{Q}_{jn}^{(0)}(v_{j,0} + \Delta_{j,0}, v_{j,1} + \Delta_{j,1}) - \widetilde{Q}_{jn}^{(0)}(v_{j,0}, v_{j,1}) + \frac{1}{n} \sum_{i \in I_2, i \neq j} (Y_{ij} - \dot{\Lambda}_{n,ij})(\dot{\psi}_{ij})^\top \Delta_{jv}.
$$

Following the same argument in the proof of Theorem 4.2, we have

$$
F(\Delta_{jv}) \geq \min \left(\frac{\underline{c}' c_\phi \zeta_n \underline{c} ||\Delta_{jv}||^2}{6}, \frac{\underline{c}' \zeta_n q_{jn} \sqrt{c_\phi} ||\Delta_{jv}||}{3\sqrt{2}} \right),
$$

where $q_{jn} = \inf_\Delta \frac{\left[\frac{1}{n_2} \sum_{i \in I_2, i \neq j} ((\dot{\psi}_{ij})^\top \Delta)^2\right]^{3/2}}{\frac{1}{n_2} \sum_{i \in I_2, i \neq j} ((\dot{\psi}_{ij})^\top \Delta)^3}$. For the upper bound of $F(\Delta_{jv})$, we can show that

$$
F(\Delta_{jv}) \leq \left| \frac{1}{n_2} \sum_{i \in I_2, i \neq j} (Y_{ij} - \Lambda_{n,ij}) (\dot{\psi}_{ij})^\top \Delta_{jv} \right| + \left| \frac{1}{n_2} \sum_{i \in I_2, i \neq j} (\dot{\Lambda}_{n,ij} - \Lambda_{n,ij})(\dot{\psi}_{ij})^\top \Delta_{jv} \right| \equiv \tilde{I}_j + \widetilde{II}_j.
$$

We first bound $\widetilde{II}_j$. Following Lemma S1.3(1), we have

$$
||v_{j,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij}|| \lesssim ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| + ||u_{i,l}|| \leq C < \infty.
$$

Then, by the same argument in the proof of Lemma S1.3(2), we have

$$\overline{c}'\zeta_n \geq \dot{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n \quad \text{and} \quad \overline{c}'\zeta_n \geq \tilde{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n,$$

for some constants $\infty > \overline{c}' > \underline{c}' > 0$. Following (A.19) and by noticing that $\frac{1}{n_2}\sum_{i \in I_2, i \neq j} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \leq C_1^* \eta_n$, we have

$$\widetilde{II}_j \leq C'_{II}\zeta_n\eta_n||\Delta_{jv}||, \tag{A.26}$$

for some constant $C'_{II} > 0$.

The analysis of $\widetilde{I}_j$ is different from that of I_i as we no longer have the independence between $\dot{\psi}_{ij}$ and $Y_{ij} - \Lambda_{n,ij}$ given $\{W_{1,ij}\}_{1 \leq i < j \leq n}$. Instead, we let $\psi_{ij} = \begin{bmatrix} u_{i,0} \\ u_{i,1}W_{1,ij} \end{bmatrix}$. Note that ψ_{ij} is deterministic given $\{W_{1,ij}\}_{1 \leq i < j \leq n}$. In addition, $\max_{i,j \in [n], i \neq j} ||\dot{\psi}_{ij} - \psi_{ij}|| \leq (1 + M_W)C_1^*\eta_n$. Therefore,

$$\tilde{I}_j \leq \left[\left\|\frac{1}{n_2}\sum_{i \in I_2, i \neq j}(Y_{ij} - \Lambda_{n,ij})\psi_{ij}\right\| + \frac{1}{n_2}\sum_{i \in I_2, i \neq j}|Y_{ij} - \Lambda_{n,ij}|\,||\dot{\psi}_{ij} - \psi_{ij}||\right]||\Delta_{jv}||.$$

For the first term in the square brackets, by the conditional Bernstein inequality given $\{W_{1,ij}\}_{1 \leq i < j \leq n}$, we have

$$\max_{j \in [n]}\left\|\frac{1}{n_2}\sum_{i \in I_2, i \neq j}(Y_{ij} - \Lambda_{n,ij})\psi_{ij}\right\| \leq C'_H\sqrt{\frac{\log n\zeta_n}{n}} \ a.s., \tag{A.27}$$

where $C'_H = 4(1 + \overline{c})^2[C_uC_\sigma(M_W + 1) + 1]^4$. For the second term in the square brackets, we have

$$\begin{aligned}
\frac{1}{n_2}\sum_{i \in I_2, i \neq j}|Y_{ij} - \Lambda_{n,ij}| \cdot ||\dot{\psi}_{ij} - \psi_{ij}|| \leq & \frac{(1 + M_W)C_1^*\eta_n}{n_2}\sum_{i \in I_2, i \neq j}|Y_{ij} - \Lambda_{n,ij}| \\
\leq & (1 + M_W)C_1^*\eta_n\left[\frac{1}{n_2}\sum_{i \in I_2, i \neq j}(Y_{ij} - \Lambda_{n,ij}) + \frac{2}{n_2}\sum_{i \in I_2, i \neq j}\Lambda_{n,ij}\right] \\
\leq & (1 + M_W)C_1^*\eta_n\left(4\overline{c}\sqrt{\frac{\zeta_n\log n}{n}} + 2\overline{c}\zeta_n\right) \\
\leq & 3(1 + M_W)\overline{c}C_1^*\eta_n\zeta_n,
\end{aligned}$$

where the second last inequality is due to the Bernstein inequality and Assumption 5, and the last inequality holds because $4\sqrt{\frac{\log n}{n\zeta_n}} \leq 4c_F \leq 1$.

Combining the two estimates, we have uniformly in j

$$\tilde{I}_j \leq \left(C'_H\sqrt{\frac{\log n}{n\zeta_n}} + 3(1 + M_W)\overline{c}C_1^*\right)\eta_n\zeta_n||\Delta_v|| \leq 4(1 + M_W)\overline{c}C_1^*\eta_n\zeta_n||\Delta_{jv}||,$$

where the last inequality holds because c_F is sufficiently small so that $C'_H(c_F + c_F^2) \leq (1 + M_W)\overline{c}C_1^*$.

Combining the upper and lower bounds for $F(\Delta_{jv})$, we have

$$[4(1+M_W)\bar{c}C_1^* + C_{II}']\eta_n\zeta_n||\Delta_{jv}|| \geq \min\left(\frac{\underline{c}'c_\phi\zeta_n\underline{c}||\Delta_{jv}||^2}{6}, \frac{\underline{c}'\zeta_n q_{jn}\sqrt{c_\phi}||\Delta_{jv}||}{3\sqrt{2}}\right). \tag{A.28}$$

By the same argument in Lemma S1.5, we have

$$q_{jn} \geq \inf_{\Delta}\sqrt{\frac{\frac{1}{n}\sum_{i\in I_2, i\neq j}((\dot{\psi}_{ij})^\top\Delta)^2}{4(1+M_W)^2C_u^2C_\sigma^2||\Delta||^2}} \geq \frac{c_\sigma}{4(1+M_W)C_uC_\sigma} > 0.$$

In addition, because c_F can be made sufficiently small to ensure $(4(1+M_W)\bar{c}C_1^* + C_{II}')(c_F + c_F^2) < \frac{c_\sigma\underline{c}'\sqrt{c_\phi}}{12\sqrt{2}(1+M_W)C_uC_\sigma}$, we have

$$\begin{aligned}(4(1+M_W)\bar{c}C_1^* + C_{II}')\eta_n\zeta_n||\Delta_{jv}|| &\leq (4(1+M_W)\bar{c}C_1^* + C_{II}')(c_F + c_F^2)\zeta_n||\Delta_{jv}|| \\ &< \frac{c_\sigma\underline{c}'\sqrt{c_\phi}\zeta_n||\Delta_{jv}||}{12\sqrt{2}(1+M_W)C_uC_\sigma} \leq \frac{\underline{c}'\sqrt{c_\phi}\zeta_n q_{jn}||\Delta_{jv}||}{3\sqrt{2}}.\end{aligned}$$

Then, (A.28) implies

$$||\Delta_{jv}|| \leq \frac{6(4(1+M_W)\bar{c}C_1^* + C_{II}')}{\underline{c}'c_\phi\underline{c}}\eta_n \equiv C_{0,v}\eta_n. \tag{A.29}$$

Note the constant $C_{0,v}$ on the right hand side does not depend on j so that the desired result holds uniformly over $j \in [n]$. ■

A.5 Proof of Theorem 4.3

We can establish the desired results by induction. Given $\max_{j\in[n]}||(\widehat{O}_l^{(1)})^\top\dot{v}_{j,l}^{(h-1,1)} - v_{j,l}|| \leq C_{h-1,v}\eta_n$ *a.s.*, we can readily show that

$$\max_{i\in[n]}||(\widehat{O}_l^{(1)})^\top\dot{u}_{i,l}^{(h,1)} - u_{i,l}|| \leq C_{h,u}\eta_n \ a.s.$$

Then, given $\max_{i\in[n]}||(\widehat{O}_l^{(1)})^\top\dot{u}_{i,l}^{(h,1)} - u_{i,l}|| \leq C_{h,u}\eta_n$ *a.s.*, we can show that

$$\max_{j\in[n]}||(\widehat{O}_l^{(1)})^\top\dot{v}_{j,l}^{(h,1)} - v_{j,l}|| \leq C_{h,v}\eta_n \ a.s.$$

As the regressors in both iteration steps have the uniform bound, the proof of Theorem 4.3 is similar to that of the second result in Theorem 4.2, and is thus omitted for brevity. ■

A.6 Proof of Theorem 4.4

Let $v_j^* = \left(\frac{(\widehat{O}_1^{(1)} v_{j,1})^\top}{||\widehat{O}_1^{(1)} v_{j,1}||}, \frac{(\widehat{O}_1^{(2)} v_{j,1})^\top}{||\widehat{O}_1^{(2)} v_{j,1}||} \right)^\top$. Then we have

$$\begin{aligned}
||\overline{v}_j - v_j^*|| &\leq \left\| \frac{\dot{v}_{j,1}^{(H,1)}}{||\dot{v}_{j,1}^{(H,1)}||} - \frac{\widehat{O}_1^{(1)} v_{j,1}}{||\widehat{O}_1^{(1)} v_{j,1}||} \right\| + \left\| \frac{\dot{v}_{j,1}^{(H,2)}}{||\dot{v}_{j,1}^{(H,2)}||} - \frac{\widehat{O}_1^{(2)} v_{j,1}}{||\widehat{O}_1^{(2)} v_{j,1}||} \right\| \\
&= \left\| \frac{(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}}{||(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\| + \left\| \frac{(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}}{||(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\| \\
&\leq \frac{2 \left\| (\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)} - v_{j,1} \right\|}{||(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}||} + \frac{2 \left\| (\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)} - v_{j,1} \right\|}{||(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}||} \\
&\leq \frac{4C_{H,v}\eta_n}{||v_{j,1}|| - C_{H,v}\eta_n} \leq 5C_1^{-1/2} C_{H,v}\eta_n, \qquad \text{(A.30)}
\end{aligned}$$

where the last inequality is due to the fact that $||v_{j,1}|| \geq C_1^{-1/2}$ and $C_{H,v}\eta_n \leq C_{H,v}(c_F + c_F^2) \leq C_1^{-1/2}/5$ as c_F can be made sufficiently small. In addition, by Lemma 2.1, for $z_i \neq z_j$,

$$||v_j^* - v_i^*|| = \left[\left\| \frac{\widehat{O}_1^{(1)} v_{i,1}}{||\widehat{O}_1^{(1)} v_{i,1}||} - \frac{\widehat{O}_1^{(1)} v_{j,1}}{||\widehat{O}_1^{(1)} v_{j,1}||} \right\|^2 + \left\| \frac{\widehat{O}_1^{(2)} v_{i,1}}{||\widehat{O}_1^{(2)} v_{i,1}||} - \frac{\widehat{O}_1^{(2)} v_{j,1}}{||\widehat{O}_1^{(2)} v_{j,1}||} \right\|^2 \right]^{1/2} \qquad \text{(A.31)}$$

$$= \left[\left\| \frac{v_{i,1}}{||v_{i,1}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\|^2 + \left\| \frac{v_{i,1}}{||v_{i,1}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\|^2 \right]^{1/2} = 2. \qquad \text{(A.32)}$$

Given (A.30) and (A.31), the result of Theorem 4.4 is a direct consequence of Su et al. (2020, Theorem II.3). In particular, we only need to verify their Assumption 4 holds with $c_{1n} = 2$, $c_{2n} = 5C_1^{-1/2} C_{H,v}\eta_n$, and $M = 2$. Note when c_F is sufficiently small,

$$2(5C_1^{-1/2} c_1^{1/2} C_{H,v}\eta_n)^{1/2} \leq 2 \left[5C_1^{-1/2} c_1^{1/2} C_{H,v}(c_F + c_F^2) \right]^{1/2} \leq K_1^{3/4}\sqrt{2}.$$

Then their Assumption 4 holds as

$$\begin{aligned}
(2c_{2n}c_1^{1/2} + 16K_1^{3/4} M^{1/2} c_{2n}^{1/2})^2 &\leq (17K_1^{3/4} M^{1/2} c_{2n}^{1/2})^2 = 1734K_1^{3/2} C_1^{-1/2} C_{H,v}\eta_n \\
&\leq 1734K_1^{3/2} C_1^{-1/2} C_{H,v}(c_F + c_F^2) \leq 2c_1
\end{aligned}$$

when c_F is sufficiently small. ■

References

Abbe, E., A. S. Bandeira, and G. Hall (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory 62*(1), 471–487.

Abbe, E., J. Fan, K. Wang, and Y. Zhong (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

Abbe, E. and C. Sandon (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 670–688. IEEE.

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Alidaee, H., E. Auerbach, and M. P. Leung (2020). Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics 4*, 384–414.

Bandeira, A. S. and R. van Handel (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability 44*(4), 2479–2506.

Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences 110*(36), 14563–14568.

Belloni, A., M. Chen, and O. H. M. Padilla (2019). High dimensional latent panel quantile regression with an application to asset pricing. *arXiv preprint arXiv:1912.02151*.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation with high-dimensional data. *Econometrica 85*(1), 233–298.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2017). Covariate-assisted spectral clustering. *Biometrika 104*(2), 361–377.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Cabral, R., F. De la Torre, P. J. Costeira, and B. Alexandre (2013). Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. *IEEE International Conference on Computer Vision*, 2488–2495.

Chatterjee, S., P. Diaconis, and A. Sly (2011). Random graphs with a given degree sequence. *The Annals of Applied Probability 21*(4), 1400–1435.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2018). Inference for heterogeneous effects using low-rank estimations. *arXiv preprint arXiv:1812.08089*.

Fan, J., W. Gong, and Z. Zhu (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics 212*(1), 177–202.

Feng, J. (2019). Regularized quantile regression with interactive fixed effects. *arXiv preprint arXiv:1911.00166*.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica 85*(4), 1033–1063.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Javanmard, A. and A. Montanari (2018). De-biasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics 43*(1), 57–89.

Jochmans, K. (2019). Modified-likelihood estimation of fixed-effect models for dyadic data.

Joseph, A. and B. Yu (2016). Impact of regularization on spectral clustering. *The Annals of Statistics 44*(4), 1765–1791.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Leger, J.-B. (2016). Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates. *arXiv preprint arXiv:1602.07587*.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics 43*(1), 215–237.

Leung, M. P. (2015). Two-step estimation of network-formation models with incomplete information. *Journal of Econometrics 188*(1), 182–195.

Lusher, D., J. Koskinen, and G. Robins (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Mele, A. (2017a). A structural model of dense network formation. *Econometrica 85*(3), 825–850.

Mele, A. (2017b). A structural model of homophily and clustering in social networks. *Available at SSRN 3031489*.

Moon, H. R. and M. Weidner (2018). Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987*.

Mossel, E., J. Neeman, and A. Sly (2014). Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591 In proc. of STOC15*.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics 39*(2), 1069–1097.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*(2), 186–199.

Qin, T. and K. Rohe (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 26, pp. 3120–3128. Curran Associates, Inc.

Rinaldo, A., S. Petrović, and S. E. Fienberg (2013). Maximum lilkelihood estimation in the β-model. *The Annals of Statistics 41*(3), 1085–1110.

Rohde, A. and A. B. Tsybakov (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics 39*(2), 887–930.

Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*(4), 1878–1915.

Roy, S., Y. Atchade, and G. Michailidis (2019). Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. *Journal of Computational and Graphical Statistics*.

Sarkar, P. and P. J. Bickel (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics 43*(3), 962–990.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

Su, L., W. Wang, and Y. Zhang (2020). Strong consistency of spectral clustering for stochastic block models. *IEEE Transactions on Information Theory 66*(1), 324–338.

Sweet, T. M. (2015). Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics 40*(6), 635–664.

Traud, A., P. Mucha, and M. Porter (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications 391*(16), 4165–4180.

Vu, V. (2018). A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing 27*(1), 124–140.

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*(526), 857–868.

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*(2), 519–524.

Yun, S.-Y. and A. Proutiere (2014). Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*.

Yun, S.-Y. and A. Proutiere (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pp. 965–973.

Zhong, Y. and N. Boumal (2018). Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization 28*(2), 989–1016.

Online Supplement to "Detecting Latent Communities in Network Formation Models"

Shujie Ma[a], Liangjun Su[b,c], and Yichong Zhang[c]

[a]*Department of Statistics, University of California, Riverside*
[b] *School of Economics and Management, Tsinghua University,*
[c]*School of Economics, Singapore Management University*

This supplement is composed of three parts. Section S1 contains some technical lemmas used in the proofs of the main results in the paper. Section S2 provides more details on the inference of B_1^*. Section S3 describes the algorithm to implement the nuclear norm regularized estimation.

S1 Some Technical Lemmas

Lemma S1.1. *Let C_Υ be an sufficiently large and fixed constant. Suppose that the assumptions in Theorem 4.1 hold. Then*

$$\max_{l=0,1} ||\Upsilon_l||_{op} \leq C_\Upsilon M_W(\sqrt{\zeta_n n} + \sqrt{\log n}) \ a.s.$$

Proof. Let $\mathcal{C} = \{X_i\}_{i=1}^n \cup \{e_{ij}\}_{1\leq i<j\leq n}$ and $r_n = C_\Upsilon M_W\sqrt{\log(n)\zeta_n n}$ for some sufficiently large constant C_Υ whose value will be determined later. In addition, we augment the $n_1 \times n$ matrix Υ_l to a symmetric $n \times n$ matrix $\overline{\Upsilon}_l$ with (i,j)-th entry

$$\overline{\Upsilon}_{l,ij} = \begin{cases} \Upsilon_{l,ij} & \text{if} \quad i \in I_1, j = 1, \cdots, n \\ \Upsilon_{l,ji} & \text{if} \quad j \in I_1, i \in [n]/I_1 \\ 0 & \text{if} \quad i \notin I_1, j \notin I_1. \end{cases}$$

Then, by construction, $||\Upsilon_l||_{op} \leq ||\overline{\Upsilon}_l||_{op}$. Therefore,

$$\begin{aligned}\mathbb{P}(\max_{l=0,1} ||\Upsilon_l||_{op} \geq r_n) \leq & 2\max_{l=0,1} \mathbb{P}(||\Upsilon_l||_{op} \geq r_n) \leq 2\max_{l=0,1} \mathbb{E}\left[\mathbb{P}(||\Upsilon_l||_{op} \geq r_n|\mathcal{C})\right] \\ \leq & 2\max_{l=0,1} \mathbb{E}\left[\mathbb{P}(||\overline{\Upsilon}_l||_{op} \geq r_n|\mathcal{C})\right].\end{aligned}$$

Next, we bound $\mathbb{P}(||\overline{\Upsilon}_l||_{op} \geq r_n|\mathcal{C})$. Recall $\mathcal{I}_1 = \{(i,j) \in I_1 \times I_1, j > i\} \cup \{(i,j) : i \in I_1, j \notin I_1\}$. Given $\mathcal{C}$, the only randomness of $\overline{\Upsilon}_l$ comes from $\{\varepsilon_{ij}\}_{(i,j)\in\mathcal{I}_1}$, which is an i.i.d. sequence of logistic random variables. In addition, we have

$$\tilde{\sigma}^2 \equiv \max_{i\in[n]} \mathbb{E}\left(\sum_{l=1}^n \overline{\Upsilon}_{l,ij}^2 |\mathcal{C}\right) \leq \max_{i\in[n]} \sum_{j=1}^n \Lambda_{n,ij} M_W^2 \leq \bar{c} M_W^2 n\zeta_n$$

and $|\overline{\Upsilon}_{l,ij}| \leq M_W$. Then, by Bandeira and van Handel (2016, Corollary 3.12), there exists a universal constant $\tilde{c}$ such that

$$\mathbb{P}\left(||\overline{\Upsilon}_l||_{op} \geq 3\tilde{\sigma} + t\right) \leq n \exp\left(-\frac{t^2}{\tilde{c}M_W^2}\right).$$

Choosing $t = 3\sqrt{\tilde{c}}M_W$, we have

$$2\mathbb{P}\left(||\overline{\Upsilon}_l||_{op} \geq 3M_W\sqrt{\bar{c}n\zeta_n} + 3\sqrt{\tilde{c}\log(n)}M_W\right) \leq n^{-1.1},$$

and by the Borel-Cantelli Lemma,

$$||\overline{\Upsilon}_l||_{op} \leq 3M_W(\sqrt{\bar{c}n\zeta_n} + \sqrt{\tilde{c}\log(n)}) \leq C_\Upsilon M_W(\sqrt{n\zeta_n} + \sqrt{\log(n)}) \ a.s. \ \blacksquare$$

Lemma S1.2. *Suppose $M \geq t \geq 0$, for some $M \geq \log(2)$. Then* $\exp(-t) + t - 1 \geq \frac{t^2}{4M}$.

Proof. Let $f(t) = \exp(-t) + t - 1 - \frac{t^2}{4M}$. Then, $f'(t) = 1 - \exp(-t) - \frac{t}{2M}$. We want to show $f'(t) \geq 0$ for $t \in [0, M]$. This implies that $\min_{t\in[0,M]} f(t) = f(0) = 0$. Note that

$$f'(M) = 0.5 - \exp(-M) \geq 0.$$

In addition, we note that $f'(t)$ is concave so that for any $t \in [0, M]$,

$$f'(t) \geq \frac{f'(M)t}{M} \geq 0.$$

This concludes the proof. $\blacksquare$

Lemma S1.3. *Suppose that the Assumptions in Theorem 4.1 hold. Then,*

1. $\max_{j\in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)}|| \leq M\sigma_{K_l,l}^{-1}$ *a.s.*;
2. *There exist some constants $\infty > \overline{c}' > \underline{c}' > 0$ such that*

$$\overline{c}'\zeta_n \geq \widehat{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n \quad \text{and} \quad \overline{c}'\zeta_n \geq \tilde{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n \ a.s.,$$

where $\widehat{\Lambda}_{n,ij}$ and $\tilde{\Lambda}_{n,ij}$ are defined in (A.14) *and* (A.15)*, respectively.*

Proof. 1. Note that

$$\begin{aligned}||(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)}|| =&||\widehat{v}_{j,l}^{(1)}|| \leq \hat{\sigma}_{K_l,l}^{-1}||\widehat{\Sigma}_l^{(1)}\widehat{v}_{j,l}^{(1)}|| \\ =&n^{-1/2}\hat{\sigma}_{K_l,l}^{-1}\left\|[(\widehat{\mathcal{U}}_l^{(1)})^\top\widehat{\Theta}_l^{(1)}]_{\cdot j}\right\| \leq n^{-1/2}\hat{\sigma}_{K_l,l}^{-1}\left\|[\widehat{\Theta}_l^{(1)}]_{\cdot j}\right\| \leq 2M\sigma_{K_l,l}^{-1},\end{aligned}$$

where the first equality holds because $\widehat{O}_l^{(1)}$ is unitary, the second equality holds because

$$n^{-1/2}(\widehat{\mathcal{U}}_l^{(1)})^\top\widehat{\Theta}_l^{(1)} = \widehat{\Sigma}_l^{(1)}\sqrt{n}(\widehat{\mathcal{V}}_l^{(1)})^\top \equiv \widehat{\Sigma}_l^{(1)}(\widehat{V}_l^{(1)})^\top,$$

the first inequality holds because $(\widehat{\mathcal{U}}_l^{(1)})^\top\widehat{\mathcal{U}}_l^{(1)} = I_{K_l}$, and the last inequality holds because $|\widehat{\Theta}_{l,ij}| \leq M$ by construction and that by Theorem 4.1 and the fact that c_F is sufficiently small so that

$48C_{F,1}\eta_n \leq \sigma_{K_l,l}/2$, and thus,

$$|\hat{\sigma}^{-1}_{K_l,l} - \sigma^{-1}_{K_l,l}| \leq \frac{|\hat{\sigma}_{K_l,l} - \sigma_{K_l,l}|}{\sigma_{K_l,l}(\sigma_{K_l,l} - |\hat{\sigma}_{K_l,l} - \sigma_{K_l,l}|)} \leq \sigma^{-1}_{K_l,l} \ a.s.$$

As the constant M does not depend on j, the result holds uniformly over $j = 1, \cdots, n$.

2. By Theorem 4.1 and the previous result,

$$\left| \widehat{\tau}_n + \sum_{l=0}^{1} u_{i,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij} - \tau_n \right| \leq |\widehat{\tau}_n - \tau_n| + \left| \sum_{l=0}^{1} u_{i,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij} \right| \leq 30C_{F,1}\eta_n + C,$$

and thus, there exist some constants $\infty > \overline{c}' > \underline{c}' > 0$ such that

$$\overline{c}'\zeta_n \geq \widehat{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n.$$

For the same reason, we have $\overline{c}'\zeta_n \geq \tilde{\Lambda}_{n,ij} \geq \underline{c}'\zeta_n$. ■

Lemma S1.4. *Suppose Assumptions 1–6 hold. Recall that*

$$\widehat{\Phi}_i^{(1)} = \frac{1}{n_2} \sum_{j \in I_2, j \neq i} \begin{bmatrix} (\widehat{O}_0^{(1)})^\top \widehat{v}_{j,0}^{(1)} \\ (\widehat{O}_1^{(1)})^\top \widehat{v}_{j,1}^{(1)} W_{1,ij} \end{bmatrix} \begin{bmatrix} (\widehat{O}_0^{(1)})^\top \widehat{v}_{j,0}^{(1)} \\ (\widehat{O}_1^{(1)})^\top \widehat{v}_{j,1}^{(1)} W_{1,ij} \end{bmatrix}^\top.$$

Then, for the constant c_ϕ defined in Assumption 6,

$$\min_{i \in I_2} \lambda_{\min}(\widehat{\Phi}_i^{(1)}) \geq c_\phi/2 \ a.s.$$

Proof. By Lemma S1.3(1), $||(\widehat{O}_{l,U}^{(1)})^\top \widehat{v}_{j,l}^{(1)}|| \leq 2M\sigma^{-1}_{K_l,l}$ for $l = 0, 1$. Then, we have

$$\begin{aligned} ||\widehat{\Phi}_i^{(1)} - \Phi_i(I_2)|| \leq & \frac{4M}{n_2} \sum_{l=0}^{1} \sum_{j \in I_2} \sigma^{-1}_{K_l,l} ||(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} - v_{j,l}|| \\ \leq & 4M \sum_{l=0}^{1} \sigma^{-1}_{K_l,l} n_2^{-1/2} ||\widehat{V}_l \widehat{O}_l^{(1)} - V_l||_F \\ \leq & \left(544(\sqrt{K_0} + \sqrt{K_1}) C_\sigma M C_{F,1} c_\sigma^{-3}\right) \eta_n \ a.s. \end{aligned} \tag{S1.1}$$

where the second inequality holds due to Cauchy's inequality, and the last inequality holds due to Theorem 4.1. As c_F is sufficiently small so that $1088\sqrt{K_l} C_\sigma M C_{F,1} c_\sigma^{-3}(c_F + c_F^2) \leq c_\phi/2$, we have

$$\min_{i \in I_2} \lambda_{\min}(\widehat{\Phi}_i^{(1)}) \geq \min_{i \in I_2} \lambda_{\min}(\Phi_i(I_2)) - \left(544(\sqrt{K_0} + \sqrt{K_1}) C_\sigma M C_{F,1} c_\sigma^{-3}\right) \eta_n \geq c_\phi/2 \ a.s. \ ■$$

Lemma S1.5. *Let q_{in} be defined in (A.17). Suppose that Assumptions 1–6 hold. Then*

$$\liminf_n \min_{i \in I_2} q_{in} \geq \frac{c_\phi c_\sigma}{2M(1 + M_W)} > 0, \ a.s.,$$

where $\underline{c}$ and M are two constants in Assumption 6 and Lemma S1.3, respectively.

Proof. Note

$$q_{in} \geq \inf_{\Delta} \sqrt{\frac{c_\sigma^2 \frac{1}{n_2} \sum_{j \in I_2, j \neq i} ((\widehat{\phi}_{ij}^{(1)})^\top \Delta)^2}{M^2(1+M_W)^2 ||\Delta||^2}} \geq \frac{c_\sigma \liminf_n \min_{i \in I_2} \lambda_{\min}(\widehat{\Phi}_i^{(1)})}{M(1+M_W)} \geq \frac{c_\phi c_\sigma}{2M(1+M_W)} > 0,$$

where the first inequality is due to Lemma S1.3(1) and the second inequality is due to Lemma S1.4. ■

S2 More Details on the Inference for B_1^*

In this appendix we provide more details on the inference for B_1^* discussed in Section 4.5 via two examples.

S2.1 Example 1

In this example, we consider the tetrad logit regression of Graham (2017). Let $S_{ij,i'j'} = Y_{ij}Y_{i'j'}(1-Y_{ii'})(1-Y_{jj'}) - (1-Y_{ij})(1-Y_{i'j'})Y_{ii'}Y_{jj'}$. Then, for an arbitrary $K_1(K_1+1)/2$-vector B, the conditional likelihood of $S_{ij,i'j'}$ given $S_{ij,i'j'} \in \{-1,1\}$ is

$$\ell_{ij,i'j'}(B) = |S_{ij,i'j'}| \left[S_{ij,i'j'} \widetilde{\omega}_{1,ij,i'j'}^\top B - \log\left(1+\exp(S_{ij,i'j'}\widetilde{\omega}_{1,ij,i'j'}^\top B)\right)\right],$$

where $\widetilde{\omega}_{1,ij,i'j'} = \omega_{1,ij} + \omega_{1,i'j'} - (\omega_{1,ii'} + \omega_{1,jj'})$. Further denote

$$\bar{\ell}_{ij,i'j'}(B) = \frac{1}{3}\left(\ell_{ij,i'j'}(B) + \ell_{ij,j'i'}(B) + \ell_{ii',j'j}(B)\right).$$

Following Graham (2017), we define the tetrad regression estimator $\widehat{B}$ for vech(B^*) as

$$\widehat{B} = \arg\max_B \sum_{i<i'<j<j'} \bar{\ell}_{ij,i'j'}(B).$$

Let

$$\top_{iji'j'} = \begin{cases} 1 & \text{if } S_{ij,i'j'} \in \{-1,1\} \cup S_{ij,j'i'} \in \{-1,1\} \cup S_{ii',jj'} \in \{-1,1\} \\ 0 & \text{otherwise} \end{cases}$$

be the indicator that the tetrad $\{i,j,i',j'\}$ take an identifying configuration, and thus, contributes to the tetrad logit regression. Further denote $t_{q,n} = \mathbb{P}(\top_{i_1i_2i_3i_4} = 1, \top_{jj_2j_3j_4} = 1)$ as the probability that tetrads $\{i_1,i_2,i_3,i_4\}$ and $\{j,j_2,j_3,j_4\}$ both take an identifying configuration when sharing $q = 0,1,2,3,$ or 4 nodes in common. Then, we make the following assumption on the Hessian matrix.

Assumption 9. *Suppose that $\Gamma_0 \equiv \lim_{n\to\infty} t_{4,n}^{-1} \sum_{i<i'<j<j'} \nabla_{BB}\bar{\ell}_{ij,i'j'}(B)$ is a finite nonsingular matrix.*

The following theorem reports the asymptotic normality of $\widehat{B}$.

Theorem S2.1. *If Assumptions 1–9 hold, then* $\widehat{B} \xrightarrow{p} vec(B^*)$ *and*

$$\left[\frac{72}{(n-1)n}\hat{H}^{-1}\widehat{\Delta}_{2,n}\hat{H}^{-1}\right]^{-1/2}(\widehat{B} - vech(B^*)) \rightsquigarrow \mathcal{N}(0, I_{K_1(K_1+1)/2}),$$

where

$$\hat{H} = \binom{n}{4}^{-1}\sum_{i<j<i'<j'}\frac{\partial^2\bar{\ell}_{ij,i'j'}(\widehat{B})}{\partial B\partial B^\top}, \quad \widehat{\Delta}_{2,n} = \frac{2}{n(n-1)}\sum_{i<j}\hat{\bar{s}}_{ij}(\widehat{B})\hat{\bar{s}}_{ij}(\widehat{B})^\top,$$

$\hat{\bar{s}}_{ij}(B) = \frac{1}{n(n-1)/2-2(n-1)+1}\sum_{i'<j',\{i,j\}\cap\{i',j'\}=\emptyset} s_{ij,i'j'}(B)$, $s_{ij,i'j'}(B) = \nabla_B\bar{\ell}_{ij,i'j'}(B)$, *and* I_a *denotes an* $a \times a$ *identity matrix.*

Proof. Theorem S2.1 is the direct consequence of Graham (2017, Theorem 1). Note that Assumptions 1–3 in Graham (2017) hold in our setup. Although Graham (2017) requires that $W_{l,ij} = g_l(X_i, X_j)$, his proof remains valid if we have $W_{l,ij} = g_l(X_i, X_j, e_{ij})$ for some i.i.d. random variable e_{ij} such that $e_{ij} = e_{ji}$ and $e_{ij} \perp\!\!\!\perp (X_i, X_j, \varepsilon_{ij})$. In addition, Assumption 4(i)-(ii) in Graham (2017) hold as we have $n\zeta_n = \Omega(\log n)$. His Assumption 4(iii) is the same as our Assumption 9. ∎

S2.2 Example 2

In this example, we consider the logistic maximum likelihood estimation. Let $\Lambda_{n,ij}(u) = \Lambda(\omega_{ij}^\top[\text{vech}(B^*) + u(n^2\zeta_n)^{-1/2}])$ and $\Lambda_{n,ij} \equiv \Lambda_{n,ij}(0)$, where $\omega_{ij} = (\chi_{0,ij}^\top, \chi_{1,ij}^\top W_{1,ij})^\top$ is an $\mathcal{K}$-vector with $\mathcal{K} = \sum_{l=0}^1 K_l(K_l+1)/2$. Note that $\Lambda_{n,ij} = \Lambda(W_{ij}^\top\Gamma_{ij}^*)$. Now, we consider an alternative assumption.

Assumption 10. $\sup_{\|u\|\leq C}\frac{1}{n^2\zeta_n}\sum_{1\leq i<j\leq n}\Lambda_{n,ij}(u)(1-\Lambda_{n,ij}(u))\omega_{ij}\omega_{ij}^\top \xrightarrow{p} \mathcal{H}$ *for some positive-definite matrix* $\mathcal{H}$ *and large but fixed constant* C.

Theorem S2.2. *Suppose that Assumptions 1–8 and 10 hold. Let* $\widehat{\mathcal{H}}_n = \sum_{1\leq i<j\leq n}\Lambda(\omega_{ij}^\top\hat{B})(1 - \Lambda(\omega_{ij}^\top\hat{B}))\omega_{ij}\omega_{ij}^\top$. *Then*

$$\widehat{\mathcal{H}}_n^{-1/2}(\hat{B} - vech(B^*)) \rightsquigarrow \mathcal{N}(0, I_{\mathcal{K}}).$$

Proof. Let $B = \text{vech}(B^*) + u(n^2\zeta_n)^{-1/2}$ for some $\mathcal{K}\times 1$ vector u. Then, by the change of variables, we have $\hat{u} = \sqrt{n^2\zeta_n}(\widehat{B} - \text{vech}(B^*))$ and

$$\hat{u} = \arg\max_u\left[Q_n\left(\text{vech}(B^*) + u(n^2\zeta_n)^{-1/2}\right) - Q_n(\text{vech}(B^*))\right].$$

We divide the proof into two steps. In the first step, we show that for each u,

$$Q_n\left(\text{vec}(B^*) + u(n^2\zeta_n)^{-1/2}\right) - Q_n(\text{vec}(B^*)) + \upsilon_n^\top u - \frac{u^\top\mathcal{H}u}{2} = o_p(1), \tag{S2.1}$$

where $\upsilon_n = O_p(1)$ and $\mathcal{H}$ is positive definite. Then, by noticing that $Q_n\left(\text{vec}(B^*) + u(n^2\zeta_n)^{-1/2}\right)$ is convex in u, we can apply the convexity lemma of Pollard (1991) and conclude that

$$\hat{u} - \mathcal{H}^{-1}\upsilon_n = o_p(1). \tag{S2.2}$$

In the step second, we derive the asymptotic distribution of $\mathcal{H}^{-1}\upsilon_n$.

Step 1. By Taylor expansion,

$$
\begin{aligned}
&Q_n\left(\mathrm{vec}(B^*)+u(n^2\zeta_n)^{-1/2}\right)-Q_n(\mathrm{vec}(B^*))\\
=&-\frac{1}{\sqrt{n^2\zeta_n}}\sum_{1\le i<j\le n}(Y_{ij}-\Lambda_{n,ij})\omega_{ij}^\top u+\frac{1}{2}u^\top\frac{1}{n^2\zeta_n}\sum_{1\le i<j\le n}\Lambda_{n,ij}(\tilde{u})(1-\Lambda_{n,ij}(\tilde{u}))\omega_{ij}\omega_{ij}^\top u\\
\equiv&-\upsilon_n^\top u+\frac{1}{2}u^\top\mathcal{H}_n u,
\end{aligned}
$$

where $\Lambda_{n,ij}=\Lambda_{n,ij}(0)$, $\tilde{u}$ is between 0 and u, and the definitions of υ_n and $\mathcal{H}_n$ are evident. By Assumption 10, $\mathcal{H}_n\xrightarrow{p}\mathcal{H}$. In addition, $\mathbb{E}\upsilon_n=\mathbb{E}(\mathbb{E}(\upsilon_n|\omega_{ij}))=0$ and $\mathrm{Var}(\upsilon_n)<\infty$, implying that $\upsilon_n=O_p(1)$. Therefore, we have established (S2.1), and thus (S2.2).

Step 2. $\mathcal{H}$ is positive definite by Assumption 10. Noting that, $\{\varepsilon_{ij}\}_{1\le i<j\le n}\perp\!\!\!\perp\{W_{1,ij}\}_{1\le i<j\le n}$, and $\{\varepsilon_{ij}\}_{1\le i<j\le n}$ is independent across (i,j), we have

$$
\frac{1}{n^2\zeta_n}\mathbb{E}\left[(Y_{ij}-\Lambda_{n,ij})^2\omega_{ij}\omega_{ij}^\top|\{W_{1,ij}\}_{1\le i<j\le n}\right]=\frac{1}{n^2\zeta_n}\sum_{1\le i<j\le n}\Lambda_{n,ij}(1-\Lambda_{n,ij})\omega_{ij}\omega_{ij}^\top\xrightarrow{p}\mathcal{H},
$$

and for any $\varepsilon>0$, there exists n_0 sufficiently large so that for all $n\ge n_0$ and $k\in[\mathcal{K}]$,

$$
\frac{1}{n^2\zeta_n}\sum_{1\le i<j\le n}\mathbb{E}\left[(Y_{ij}-\Lambda_{n,ij})^2\omega_{k,ij}^2\mathbf{1}\{|(Y_{ij}-\Lambda_{n,ij})^2\omega_{k,ij}^2|\ge\sqrt{n^2\zeta_n}\varepsilon\}\right]\le M_W^2\mathbf{1}\{M_W^2\ge\sqrt{n^2\zeta_n}\varepsilon\}=0,
$$

where $\omega_{k,ij}$ denotes the k-th element of ω_{ij}. Therefore, by the Lindeberg-Feller central limit theorem, $\upsilon_n\rightsquigarrow\mathcal{N}(0,\mathcal{H})$ conditionally on $\{W_{1,ij}\}_{1\le i<j\le n}$. As $\mathcal{H}$ is deterministic, the above weak convergence holds unconditionally too. Therefore, $\hat{u}\rightsquigarrow\mathcal{N}(0,\mathcal{H}^{-1})=O_p(1)$. In addition, by Assumption 10,

$$
\frac{1}{n^2\zeta_n}\widehat{\mathcal{H}}_n=\frac{1}{n^2\zeta_n}\sum_{1\le i<j\le n}\Lambda_{n,ij}(\hat{u})(1-\Lambda_{n,ij}(\hat{u}))\omega_{ij}\omega_{ij}^\top\xrightarrow{p}\mathcal{H}.
$$

It follows that $\widehat{\mathcal{H}}_n^{-1/2}(\hat{B}-\mathrm{vec}(B^*))\rightsquigarrow\mathcal{N}(0,I_{\mathcal{K}})$. ■

S3 Algorithm for the Nuclear Norm Regularization

We apply the optimization algorithm proposed in Cabral, De la Torre, Costeira, and Alexandre (2013) to obtain the nuclear norm penalized estimator given in (3.2). For any given $r_l\ge K_l$ and $r_l\le n$, Γ_l can be written as $\Gamma_l=U_lV_l^\top$, where $U_l\in\mathbb{R}^{n\times r_l}$ and $V_l\in\mathbb{R}^{r_l\times n}$, for $l=0,...,p$. We consider the optimization problem:

$$
Q_n^{(1)}(\Gamma)+\frac{\lambda_n^{(1)}}{2}\sum_{l=0}^{p}\gamma_l(||U_l||_F^2+||V_l||_F^2), \tag{S3.1}
$$

where $\Gamma = (\Gamma_l, l = 0, ..., p)$, and

$$Q_n^{(1)}(\Gamma) = \sum_{i \in I_1, j \in [n], i \neq j} \left[-Y_{ij}(W_{ij}^\top \Gamma_{ij}) + \log\{1 + \exp(W_{ij}^\top \Gamma_{ij})\} \right],$$

subject to $\Gamma_l = U_l V_l^\top$ for $l = 0, ..., p$. Let $\lambda_n^{(1)} = C_\lambda(\sqrt{\zeta_n n} + \sqrt{\log n})$.

Let Γ_l^* for $l = 0, ..., p$ be an optimal solution of (3.2) with $\mathrm{rank}(\Gamma_l^*) = K_l^*$. Cabral et al. (2013) shows that any solution $\Gamma_l = U_l V_l^\top$ for $l = 0, ..., p$ of (S3.1) with $r_l \geq K_l^*$ is a solution of (3.1). Next we apply the Augmented Lagrange Multiplier (ALM) method given in Cabral et al. (2013) to solve (S3.1). The augmented Lagrangian function of (S3.1) is

$$Q_n^{(1)}(\Gamma) + \frac{\lambda_n^{(1)}}{2} \sum_{l=0}^p \gamma_l(||U_l||_F^2 + ||V_l||_F^2) + \sum_{l=0}^p \left\langle \Delta_l, \Gamma_l - U_l V_l^\top \right\rangle + \frac{\rho}{2} \sum_{l=0}^p ||\Gamma_l - U_l V_l^\top||_F^2,$$

where Δ_l are Lagrange multipliers and ρ is a penalty parameter to improve convergence.

1. At step $m + 1$, for given $(U_l^m, V_l^m, \Delta_l^m, \Theta^m, l = 0, ..., p)$, (Γ^{m+1}) minimizes

$$L_n(\Gamma) = Q_n^{(1)}(\Gamma) + \sum_{l=0}^p \left\langle \Delta_l^m, \Gamma_l - U_l^m V_l^{m\top} \right\rangle + \frac{\rho}{2} \sum_{l=0}^p ||\Gamma_l - U_l^m V_l^{m\top}||_F^2 + C.$$

Moreover, for $i \in I_1, j \in [n], i \neq j$,

$$\frac{\partial L_n(\Gamma)}{\partial \Gamma_{l,ij}} = (\mu_{ij} - Y_{ij}) W_{l,ij} + \Delta_{l,ij}^m + \rho(\Theta_{l,ij} - V_{l,ij}^{m\top} U_{l,ij}^m),$$

where $\mu_{ij} = \exp(\sum_{l=0}^1 W_{l,ij}\Gamma_{l,ij})\{1 + \exp(\sum_{l=0}^1 W_{l,ij}\Gamma_{l,ij})\}^{-1}$, and

$$\frac{\partial^2 L_n(\Gamma)}{\partial \Gamma_{l,ij}^2} = \mu_{ij}(1 - \mu_{ij}) W_{l,ij}^2 + \rho,$$

$$\frac{\partial^2 L_n(\Gamma)}{\partial \Gamma_{l,ij} \Gamma_{l',ij}} = \mu_{ij}(1 - \mu_{ij}) W_{l,ij} W_{l',ij}, \text{ for } l \neq l'$$

For $i = j \in I_1$,

$$\frac{\partial L_n(\Gamma)}{\partial \Gamma_{l,ij}} = \Delta_{l,ij}^m + \rho(\Gamma_{l,ij} - V_{l,ij}^{m\top} U_{l,ij}^m),$$

$\frac{\partial^2 L_n(\Gamma_0,\Gamma_1)}{\partial \Gamma_{l,ij}^2} = \rho$ and $\frac{\partial^2 L_n(\Gamma_0,\Gamma_1)}{\partial \Gamma_{l,ij}\Gamma_{l',ij}} = 0$. Then,

$$\Gamma^{m+1} = -\left(\frac{\partial^2 L_n(\Gamma^m)}{\partial \Gamma_{ij} \Gamma_{ij}^\top}\right)^{-1} \left(\frac{\partial L_n(\Gamma^m)}{\partial \Gamma_{ij}}\right) + \Gamma^m,$$

where $\Gamma_{ij} = (\Gamma_{0,ij}, ..., \Gamma_{p,ij})^\top$. Update $\Gamma_{l,ij}^{m+1} = \Gamma_{l,ij}^{m+1} I\{|\Gamma_{l,ij}^{m+1}| \leq \log n\} + \log n I\{|\Gamma_{l,ij}^{m+1}| > \log n\}$.

2. For given $(U_l^m, V_l^m, \Delta_l^m, \Gamma^{m+1}, l = 1, 2)$, U_l^{m+1} minimizes

$$\frac{\lambda_n^{(1)}}{2}\sum_{l=0}^{1}\gamma_l(||U_l||_F^2 + ||V_l^m||_F^2) + \sum_{l=0}^{1}\left\langle \Delta_l^m, \Gamma_l^{m+1} - U_l V_l^{m\top}\right\rangle + \frac{\rho}{2}||\Gamma_l^{m+1} - U_l V_l^{m\top}||_F^2 + C.$$

Then

$$U_l^{m+1} = (\Delta_l^m + \rho\Gamma_l^{m+1})V_l^m(\lambda_n^{(1)}\gamma_l I_{r_l} + \rho V_l^{m\top}V_l^m)^{-1}.$$

Similarly, $V_l^{m+1} = (\Delta_l^m + \rho\Gamma_l^{m+1})^\top U_l^{m+1}(\lambda_n^{(1)}\gamma_l I_{r_l} + \rho U_l^{m+1\top}U_l^{m+1})^{-1}$.

3. Let $\Delta_l^{m+1} = \Delta_l^m + \rho(\Theta_l^{m+1} - U_l^{m+1}V_l^{m+1\top})$.

4. Let $\rho = \min(\rho\mu, 10^{20})$.