Journal of Machine Learning Research 23 (2022) 1-61 Submitted 10/21; Revised 6/22; Published 11/22

Detecting Latent Communities in Network Formation Models

Shujie Ma SHUJIE.MA@UCR.EDU
Department of Statistics
University of California
Riverside, CA 92521, USA

Liangjun Su SULJ@SEM.TSINGHUA.EDU.CN
School of Economics and Mangement
Tsinghua University
Beijing, 100084, China

Yichong Zhang YCZHANG@SMU.EDU.SG
School of Economics
Singapore Management Univeristy
178903, Singapore

Editor: Ji Zhu

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 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. To address the second issue, we note that the 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[1] 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 contributes to three strands of literature in statistics and econometrics. First, our paper contributes to the large literature on the application of spectral clustering to detect communities in stochastic block models (SBMs) by studying the estimation and inference of a network formation model with both covariates and latent community structures in the regression coefficients. Since the pioneering work of Holland et al. (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), Paul and Chen (2020), Qin and Rohe (2013), Rohe et al. (2011), Sarkar and Bickel (2015), Sengupta and Chen (2015), Vu (2018), Wang and Wong (1987), Yun and Proutiere (2014), Yun and Proutiere (2016), and Zhao et al. (2012) among others.[2] From an information theory perspective, Abbe and Sandon (2015), Abbe et al. (2016), Mossel et al. (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 $\log n$. See Abbe (2018) for an excellent survey on the recent development of estimation of SBMs and degree-corrected SBMs. Nevertheless, most existing SBMs do not include covariates. A few exceptions include (Binkiewicz et al., 2017), Weng and Feng (2016), Yan and Sarkar (2020) and Zhang et al. (2016), who consider covariates-assisted community detection but not inferences on the underlying parameters.

1. Let Y_{ij} be a binary variable which equals one when there is an edge between nodes i and j and zero otherwise. Then, the expected degree for node i is defined as $\mathbb{E}\sum_{j\neq i} Y_{ij}$.
2. Other methods to detect communities include but are not limited to modularity maximization (Newman and Girvan, 2004), likelihood-based methods (Amini et al., 2013; Bickel and Chen, 2009; Choi et al., 2012; Zhao et al., 2012), the method of moments (Bickel et al., 2011), and spectral embedding (Lyzinski et al., 2014; Sussman et al., 2012).

In addition, our multi-step procedure combines likelihood maximization with low-rank estimation and a spectral clustering step and provides an effective and reliable tool for the estimation of the complex network model considered in the paper. The variational EM algorithm is an alternative method widely used to estimate complicated models that can incorporate both covariates and community structures. However, its performance highly hinges on the proper choice of initial values and its solution is not globally optimal. Our method, on the other hand, enjoys both the optimal fitting of the data based on the likelihood and the easy computation of the spectral clustering. Furthermore, despite the fact that the regression coefficient matrices have to be 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 rate on the expected node degree and to conduct inferences on the underlying parameters in the model.

Second, our paper contributes to the burgeoning literature on network formation models and panel structure models. For the former, see Chatterjee et al. (2011), Graham (2017), Graham (2019), Graham (2020), Graham and de Paula (2019), Holland and Leinhardt (1981), Hoff et al. (2002), Jochmans (2019), Leung (2015), Mele (2017a), Rinaldo et al. (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 inference. 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 into the network formation model is mainly inspired by the recent works of Bonhomme and Manresa (2015) and Su et al. (2016), who introduce latent group structures into panel data analyses. When the community structure is unobserved, it is analogous to the latent group structure in panel data models. For recent analyses of panel data models with latent group structures, see Ando and Bai (2016), Chen (2019), Cheng et al. (2019), Dzemski and Okui (2018), Huang et al. (2020), Huang et al. (2021), Liu et al. (2020), Lu and Su (2017), Su and Ju (2018), Su et al. (2019), Vogt and Linton (2020), Wang and Su (2021), and Xu et al. (2020), among others. In particular, Wang and Su (2021) establish the connection between SBMs and panel data models with latent group structures and propose to adopt the spectral clustering techniques to recover the latent group structures in panel data models.

Last, our paper also contributes to the literature on the use of nuclear norm regularization in various contexts; see Alidaee et al. (2020), Belloni et al. (2019), Bai and Ng (2019), Chernozhukov et al. (2020), Fan et al. (2019), Feng (2019), Koltchinskii et al. (2011), Moon and Weidner (2018), Negahban and Wainwright (2011), Negahban et al. (2012), and Rohde and Tsybakov (2011), among others. All these previous works focus on the error bounds (in Frobenius norm) for the nuclear norm regularized estimates, except Moon and Weidner (2018) and Chernozhukov et al. (2020) who study the inference problem in linear panel data models with a low-rank structure. Like Moon and Weidner (2018) and Chernozhukov et al. (2020), we simply use the nuclear norm regularization to obtain consistent initial estimates. Unlike Moon and Weidner (2018) and Chernozhukov et al. (2020), we study a nonlinear logistic network formation model with a latent community structure and propose the iterative row- and column-wise logistic regressions to improve the error bounds (in row-wise Euclidean norm) for the singular vectors of the nuclear norm regularized estimates.

Relying on such an improvement, we can fully recover the community memberships. Then, we can estimate the community specific parameters and make statistical inferences.

The rest of the paper is organized as follows. In Section 2, we introduce the model and basic assumptions. In Section 3, we provide our multi-step estimation procedure. Section 4 establishes the statistical properties of the proposed estimators of the singular vectors. Section 5 studies the K-means estimation of the community memberships when the regression coefficient matrix is assumed to exhibit some community structure. Section 6 studies the asymptotic properties of the regression coefficient estimates in the presence of latent community structures. Section 7 discusses the determination of the ranks of the regression coefficient matrices. Section 8 reports simulation results. In Section 9, 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 10 concludes. The Appendix provides the proofs of all theoretical results and the associated technical lemmas, and some additional technical details.

Notation. Throughout the paper, we write "w.p.a.1" for "with probability approaching one," $M = \{M_{ij}\}$ 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

$$\text{vech}(B) = (B_{11}, ..., B_{1K}, B_{22}, ..., B_{2K}, \cdots, B_{K-1,K-1}, B_{K-1,K}, B_{KK})^\top.$$

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. Let $\odot$ denote Hadamard product.

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

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

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 $i \in [n]$. Our estimation and inference methods can be extended to directed networks by considering the upper- and lower-triangular submatrices separately.

Apparently, without making any assumptions on $\Theta_l^* = \{\Theta_{l,ij}^*\}$ for $l \in [p] \cup \{0\}$, one cannot estimate all the parameters in (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 regressions and assume that each Θ_l^* exhibits a certain low rank structure. Even so, it is easy to see that our model in (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

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq a_i + a_j\}. \tag{2}$$

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 et al. (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

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq a_i + a_j + W_{ij}^\top \beta\}, \tag{3}$$

where $\beta = (\beta_1, ..., \beta_p)^\top$. Apparently, (3) is the undirected dyadic link formation model with degree heterogeneity studied in Graham (2017). See also Yan et al. (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 a 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

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \Theta_{0,ij}\}. \tag{4}$$

Corresponding to the simple SBM with K communities, 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

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \Theta_{0,ij} + W_{ij}^\top \beta\}. \tag{5}$$

Then (5) defines a stochastic block model with covariates considered in Sweet (2015), Leger (2016), and Roy et al. (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.[3] Of course, if ζ_n is fixed or convergent to a positive constant as $n \to \infty$, the network becomes dense.

Admittedly, our model is somewhat restrictive in the sense that it requires that different communities should share the same order of expected degree. But this is also the common case in the community detection literature. Ideally, we can allow different communities to have different orders of expected degree in which case one should replace ζ_n by $\zeta_{n,ij}$ in (1), where

$$\zeta_{n,ij} = \zeta^o_{n,kl} \text{ for nodes in } i \text{ and } j \text{ in communities } k \text{ and } l, \text{ respectively,}$$

and $k, l = 1, ..., K_0$. Since we now allow $\zeta^o_{n,kl}$ to be n-dependent and they may decay to zero at different rates for different pairs (k, l) , we allow different communities to have different expected degrees of connection. For instance, we can assume that $\zeta^o_{n,kl} = n^{-\alpha_{kl}}$ for some group specific rate $\alpha_{kl} \geq 0$. Then $\log(\zeta_{n,ij}) = -\alpha_{kl} \log n$. With such a change, we need to combine $\log(\zeta_{n,ij})$ with $\Theta^*_{0,ij}$ in the estimation procedure. Notice that $\log(\zeta_{n,ij})+\Theta^*_{0,ij}$ may diverge to negative infinity at rate $\log n$ and $\{(\log(\zeta_{n,ij}) + \Theta^*_{0,ij})/(\log n)\}$ converges to a matrix $\{A_{ij}\}$ which contains the group structure, i.e., $A_{ij} = -\alpha_{kl}$ if $i \in k$ and $j \in l$. We can apply K-means to the estimator of $\{(\log(\zeta_{n,ij}) + \Theta^*_{0,ij})/(\log n)\}$ to estimate α_{kl} and then estimate $\Theta^*_{0,ij}$. This will complicate the subsequent analyses in the paper. We leave the theoretical study of this extension for future research and focus on the case with a universal rate ζ_n for the rest of the paper.

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 (1) as

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq W_{ij}^\top \Gamma^*_{ij}\}. \tag{6}$$

Below, we will let $\Gamma^*_l = \Theta^*_l$ for $l \in [p]$ and 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 (1). The first assumption is about the data generating process (DGP).

Assumption 1

1. *For $l \in [p]$, 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 $\{e_{ij}\}_{1 \leq i < j \leq n}$ are two independent and identically distributed (i.i.d.) sequences of random variables, and $e_{ij} = e_{ji}$ for $i \neq j$.*
2. *$\{\varepsilon_{ij}\}_{1 \leq i < j \leq n}$ is an i.i.d. sequence of logistic random variables. Moreover, $\{\varepsilon_{ij}\}_{1 \leq i < j \leq n} \perp\!\!\!\perp (\{X_i\}_{i=1}^n \cup \{e_{ij}\}_{1 \leq i < j \leq n})$. Let $\varepsilon_{ij} = \varepsilon_{ji}$ for $i > j$.*
3. *$\max_{l \in [p]} \max_{i \neq j \in [n]} |W_{l,ij}| \leq M_W$ for some constant $M_W < \infty$.*

Assumption 1 specifies how the covariates and error terms are generated. As the labels of the nodes are exchangeable, Assumption 1.1 is innocuous due to the Aldous and Hoover representation

3. A network is dense if the expected degree grows at rate-n and semi-dense if it diverges to infinity at a rate slower than n.

theorem. 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. We further assume that it is uniformly bounded to simplify the analysis. Assumption 1.2 is standard.

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

Assumption 2 1. *Suppose* $\sum_{i,j\in[n]} \Theta^*_{0,ij} = 0$.

2. *Suppose* Θ_l^* *is symmetric and of low rank* K_l *for* $l \in [p]\cup\{0\}$*. 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* $\Sigma_l = diag(\sigma_{1,l}, \cdots, \sigma_{K_l,l})$ *with singular values* $\sigma_{1,l} \ge \cdots \ge \sigma_{K_l,l}$*. We further denote* $U_l = \sqrt{n}\mathcal{U}_l\Sigma_l$ *and* $V_l = \sqrt{n}\mathcal{V}_l$*. Then,*

$$\Theta_l^* = n\mathcal{U}_l\Sigma_l\mathcal{V}_l^\top = U_l V_l^\top \text{ for } l = 0, ..., p. \tag{7}$$

Let $u_{i,l}^\top$ *and* $v_{i,l}^\top$ *denote the* i*-th row of* U_l *and* V_l*, respectively for* $l \in [p]\cup\{0\}$*. Then,* $\max_{i\in[n],l\in[p]}(||u_{i,l}|| \vee ||v_{i,l}||) \le M$ *for some constant* $M < \infty$ *and there are constants* C_σ *and* c_σ *such that*

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

We note (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 (1). The low-rank structure of Θ_l^* incorporates two special cases: (1) additive structure and (2) latent community structure, as illustrated in detail in Examples 1 and 2 below, respectively. When there are no covariates in regression and Θ_0 belongs to the two cases in Examples 1 and 2, the model becomes the so-called Beta model and stochastic block model, respectively. We extend these models to the scenario with edge-wise characteristics and latent community structure for the slope coefficients. The following two examples further show that all four models, namely (2)–(5), considered in Section 2.1 satisfy Assumption 2.

Example 1 *Let* $\Theta^*_{l,ij} = \alpha_{l,i} + \alpha_{l,j}$*. In this case,* $K_l = 2$ *and* $n^{-1}\Theta_l^* = \mathcal{U}_l\Sigma_l\mathcal{V}_l^T$*, where*

$$\mathcal{U}_l = \begin{pmatrix} \frac{1}{\sqrt{2n}}(1+\frac{\alpha_{l,1}}{s_{l,n}}) & \frac{-1}{\sqrt{2n}}(1-\frac{\alpha_{l,1}}{s_{l,n}}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2n}}(1+\frac{\alpha_{l,n}}{s_{l,n}}) & \frac{-1}{\sqrt{2n}}(1-\frac{\alpha_{l,n}}{s_{l,n}}) \end{pmatrix},\ \Sigma_l = \begin{pmatrix} s_{l,n} & 0 \\ 0 & s_{l,n} \end{pmatrix},$$
$$\mathcal{V}_l = \begin{pmatrix} \frac{1}{\sqrt{2n}}(1+\frac{\alpha_{l,1}}{s_{l,n}}) & \frac{1}{\sqrt{2n}}(1-\frac{\alpha_{l,1}}{s_{l,n}}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2n}}(1+\frac{\alpha_{l,n}}{s_{l,n}}) & \frac{1}{\sqrt{2n}}(1-\frac{\alpha_{l,n}}{s_{l,n}}) \end{pmatrix},$$

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

$$U_l = \begin{pmatrix} \frac{1}{\sqrt{2}}(s_{l,n}+\alpha_{l,n}) & \frac{-1}{\sqrt{2}}(s_{l,n}-\alpha_{l,n}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2}}(s_{l,n}+\alpha_{l,n}) & \frac{-1}{\sqrt{2}}(s_{l,n}-\alpha_{l,n}) \end{pmatrix} \text{ and } V_l = \sqrt{n}\mathcal{V}_l.$$

When $l = 0$, we further impose $\sum_{i,j\in[n]} \Theta_0^* = 0$, which implies $\sum_{i=1}^n \alpha_{0,i} = 0$. We also allow $\{\alpha_{l,i}\}_{i=1}^n$ to depend on $\{W_{ij}\}_{1\leq i<j\leq n}$ so that $\{\alpha_{l,i}\}_{i=1}^n$ are usually referred to as individual fixed effects in the literature.

Example 2 *Let $\Theta_l^* = Z_l B_l^* Z_l^\top$, 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 for Θ_l^*, and $B_l^* \in \mathbb{R}^{K_l\times K_l}$ is symmetric with rank K_l. Let $p_l^\top = (\frac{n_{1,l}}{n}, \cdots, \frac{n_{K_l,l}}{n})$ and $n_{k,l}$ denotes the size of Θ_l^*'s k-th community for $k \in [K_l]$. Then, as Lemma 2.1 below shows,*

$$U_l = Z_l^\top (\Pi_{l,n})^{-1/2} S_l' \Sigma_l \quad \textit{and} \quad V_l = Z_l^\top (\Pi_{l,n})^{-1/2} S_l,$$

where S_l and S_l' are two $K_l \times K_l$ matrices such that $S_l^\top S_l = I_{K_l} = (S_l')^\top S_l'$, $\Pi_{l,n} = diag(p_l)$, and Σ_l is the singular value matrix of $\Pi_{l,n}^{1/2} B_l^* \Pi_{l,n}^{1/2}$. Let ι_n denote an $n \times 1$ vector of ones. If $l = 0$, we further impose that $\iota_n^\top Z_0 B_0^* Z_0^\top \iota_n = p_0^\top B_0^* p_0 = 0$.

For classification and inference, we need to impose the latent community structure for Θ_l^*, $l \in [p]$ as in Example 2. This is summarized in the following assumption.

Assumption 3

1. *$\Theta_l^* = Z_l B_l^* Z_l^\top$, where $Z_l \in \mathbb{R}^{n\times K_l}$ is as defined in Example 2.*
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 if Θ_l^* has a latent community structure, the size of each community must be 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 shows Assumption 3 is a special case of Assumption 2 and lays down the foundation for our classification procedure Section 3.2.

Lemma 2.1 *Suppose Assumption 3 holds. 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 $\max_{i\in[n]} ||u_{i,l}|| \leq c_1^{-1/2} C_\sigma < \infty$.*
3. *If $z_{i,l} \neq z_{j,l}$, then $\left\| \frac{v_{i,l}}{||v_{i,l}||} - \frac{v_{j,l}}{||v_{j,l}||} \right\| = ||(z_{i,l} - z_{j,l}) S_l|| = \sqrt{2}$.*

Lemma 2.1 implies that, if Θ_l^* has the community structure, its singular vectors $\{v_{i,l}\}_{i\in[n]}$ 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).

In Section 4, we only require Θ_l^*, $l \in [p] \cup \{0\}$ to be of low-rank and derive the uniform convergence rate of the estimators of $(u_{i,l}, v_{i,l})$ across $i \in [n]$. In Section 5, we further suppose that *some* coefficient Θ_l^* has a special community structure as in Assumption 3 and apply the K-means algorithm to exactly recover their group identities. Last, for inference in Section 6, we impose that *all* coefficients Θ_l^*, $l \in [p]$ have (potentially different) community structures while Θ_0^* follows the structure in either Example 1 or 2.

3. The Estimation Algorithm

For notational simplicity, we will focus on the case of $p = 1$. 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, if we further impose Θ_1^* has a community structure, then we can 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 7.

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 3.1.1 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 n 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 3.1.2 and obtain $\{\widehat{V}_l^{(1)}\}_{l=0,1}$, a preliminary estimate of $\{V_l\}_{l=0,1}$, where the superscript (1) means we use the first subsample to conduct the nuclear norm estimation. 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 $\{v_{j,l}\}_{l=0,1}$, where 0 in the superscript $(0, 1)$ means it is the 0-th step estimator for the full sample iteration below and 1 in the superscript means it is computes when the first subsample is used for the nuclear norm estimation. See Section 3.1.3 for details.

4. Based on $\{\dot{v}_{j,l}^{(0,1)}\}_{j\in[n], l=0,1}$, obtain the iterative estimates

$$(\dot{u}_{i,0}^{(h,1)}, \dot{u}_{i,1}^{(h,1)})_{i\in[n]} \quad \text{and} \quad (\dot{v}_{j,0}^{(h,1)}, \dot{v}_{j,1}^{(h,1)})_{j\in[n]}$$

of the singular vectors as in Step 3 for $h = 1, 2, \cdots, H$. See Section 3.1.4 for details.

5. Switch the roles of I_1 and I_2 and repeat Steps 2–4 to obtain

$$(\dot{u}_{i,0}^{(h,2)}, \dot{u}_{i,1}^{(h,2)})_{i\in[n]} \quad \text{and} \quad (\dot{v}_{j,0}^{(h,2)}, \dot{v}_{j,1}^{(h,2)})_{j\in[n]} \quad \forall h \in [H],$$

where h in the superscript $(h, 2)$ means it is the h-th step iteration of the full sample estimator and 2 in the superscript means the second subsample is used for the nuclear norm estimation. 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 3.2.

In the following, we provide explanations for our proposed multi-step estimation procedure. In Step 1, we obtain $\widehat{\tau}_n$ that is needed in the iterative logistic regression in Steps 3-5. We also obtain $\{\widehat{\Sigma}_l\}_{l=0,1}$ which is used to estimate the ranks $\{K_l\}_{l=0,1}$ of the matrices $\{\Theta_l^*\}_{l=0,1}$ in Section 7 below.

In Step 2, we obtain $\widehat{V}_l^{(1)}$ via nuclear norm regularized estimation. It serves as an initial estimate for the iterative logistic regression in Steps 3-4. However, we cannot directly classify nodes using $\widehat{V}_l^{(1)}$, as we can only control its estimation error in Frobenius norm, as shown in Theorem 4.1. In order to show the exact recovery of latent communities, we need to control the estimation error in the row-wise L_2 norm (denoted as $||\cdot||_{2\to\infty}$).

In Step 3, we run row-wise and column-wise logistic regression to obtain refined estimates of $\{u_{i,l}\}_{l=0,1}$ and $\{v_{j,l}\}_{l=0,1}$, respectively. It is worth noting that in Steps 2 and 3, we employ a sample-splitting technique to create independence (conditional on covariates $\{X_i\}_{i=1}^n \cup \{e_{i,j}\}_{1\le i<j\le n}$) between the edges, so that the estimation error of the resulting row-wise logistic regression estimators can be well controlled in $||\cdot||_{2\to\infty}$ norm. To see the effect of sample splitting, we note that in the row-wise logistic regression, the estimation error of $\widehat{u}_{i,l}^{(1)}$ (the i-th row of $\widehat{U}_l^{(1)}$) for $i \in I_2$ in L_2 norm is determined by the score function

$$\frac{1}{n} \sum_{j\in I_2, j\neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))W_{l,ij}(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)},$$

where $\Lambda(\cdot)$ is the logistic CDF, $\widehat{O}_l^{(1)}$ is a $K_l \times K_l$ orthogonal matrix defined in Theorem 4.1 below, and $\widehat{v}_{j,l}$ is the j-th row of $\widehat{V}_l$. We see that

$$\begin{aligned}
&\frac{1}{n} \sum_{j\in I_2, j\neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))W_{l,ij}(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} \\
&= \frac{1}{n} \sum_{j\in I_2, j\neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))W_{l,ij}v_{j,l} \\
&+ \frac{1}{n} \sum_{j\in I_2, j\neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))W_{l,ij}((\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} - v_{j,l}). \qquad (8)
\end{aligned}$$

The first term on the right hand side (RHS) of (8) is $O_p\left(\sqrt{\frac{(\log n)\zeta_n}{n}}\right)$ uniformly in $i \in I_2$, where the rate ζ_n comes from the fact that the network is possibly semi-dense. However, without sample splitting, we do not have independence between Y_{ij} and $(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)}$ and we can apply the Cauchy-Schwartz inequality to obtain the crude bound for the second term on the RHS of (8) as follows

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{j \in I_2, j \neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} ((\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} - v_{j,l}) \right| \\
&\leq M_W \left(\frac{1}{n} \sum_{j \in I_2, j \neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))^2 \right)^{1/2} \frac{||V_l - \widehat{V}_l^{(1)} \widehat{O}_l^{(1)}||_F}{\sqrt{n}} \\
&= O_p\left(\sqrt{\frac{\log n}{n}} \right) \text{ uniformly in } i \in I_2,
\end{aligned}$$

where we use the fact that

$$\max_{i \in [n]} \left(\frac{1}{n} \sum_{j \in [n]} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij}))^2 \right)^{1/2} = O_p(\zeta_n^{1/2}) \quad \text{and} \quad ||V - \widehat{V}_l^{(1)} \widehat{O}_l^{(1)}||_F / \sqrt{n} = O_p\left(\eta_n\right)$$

by Theorem 4.1 below, and $\eta_n = \sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}$. The same conclusion holds with $\sum_{j \in I_2, j \neq i}$ replaced by $\sum_{j \in [n]}$ in the absence of sample splitting. When $\zeta_n = o(1)$, compared with the first term, the second term has a rate loss by a magnitude of $\sqrt{\zeta_n}$, which can be close to $n^{-1/2}$ because we allow $\zeta_n = C \log(n)/n$ for some constant C. That is, without sample splitting, we can only show that $\frac{1}{n} \sum_{j \in I_2, j \neq i} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)}$ is $O_p\left(\sqrt{\frac{\log n}{n}}\right)$ uniformly in $i \in I_2$. In contrast, with sample splitting, we can employ the independence between Y_{ij} and $(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)}$ and show that it is $O_p\left(\sqrt{\frac{(\log n)\zeta_n}{n}}\right)$ uniformly in $i \in I_2$, which improves over the rate $O_p\left(\sqrt{\frac{\log n}{n}}\right)$ when $\zeta_n = o(1)$.

Note that we cannot use $\{\widehat{u}_{i,l}^{(1)}\}_{i \in I_2}$ to cluster all nodes $i \in [n]$. Therefore, also in Step 3, given $(\widehat{u}_{i,l}^{(1)})_{i \in I_2}$, we need to compute $\widehat{v}_{j,l}^{(1)}$ from a column-wise logistic regression using nodes in I_2 for each $j \in [n]$. Similarly, the estimation error of $\widehat{v}_{j,l}^{(1)}$ in L_2 norm is determined by the score function

$$\begin{aligned}
&\frac{1}{n} \sum_{i \in I_2, i \neq j} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} \\
&= \frac{1}{n} \sum_{i \in I_2, i \neq j} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} u_{i,l} \\
&+ \frac{1}{n} \sum_{i \in I_2, i \neq j} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} ((\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}),
\end{aligned} \tag{9}$$

where we do not have independence between $(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)}$ and Y_{ij}. However, we have already obtained the $|| \cdot ||_{2 \to \infty}$ rate of $((\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l})_{i \in I_2}$. To bound the second term on the RHS of

(9), instead of the Cauchy-Schwartz inequality, we can apply the Hölder inequality to obtain

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i \in I_2} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) W_{l,ij} ((\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}) \right\|_2 \\
&\leq \frac{M_W}{n} \sum_{i \in I_2} |Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})| \max_{i \in I_2} \|(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}\|_2 \\
&= O_p\left(\sqrt{\frac{(\log n)\zeta_n}{n}}\right) \text{ uniformly in } j \in [n], \qquad (10)
\end{aligned}$$

where we use the fact that

$$\frac{1}{n} \sum_{i \in I_2} |Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})| \leq \frac{1}{n} \sum_{i \in I_2} (Y_{ij} - \Lambda(W_{ij}^\top \Gamma_{ij})) + \frac{2}{n} \sum_{i \in I_2} \Lambda(W_{ij}^\top \Gamma_{ij}) = O_p\left(\sqrt{\frac{\zeta_n}{n}} + \zeta_n\right),$$

the result that $\max_{i \in I_2} \|(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}\|_2 = O_p(\eta_n)$ as established in Theorem 4.2, and the condition that $\zeta_n \geq C(\log n)/n$. The order in (10) is also the uniform probability order of the first term on the RHS of (9) and thus does not cause any efficiency loss to bound the left hand side of (9) in terms of uniform convergence rate.

In Step 4, we iterate the row- and column-wise logistic regressions on the full sample multiple times. Following the above argument, the $|| \cdot ||_{2 \to \infty}$ of $(\dot{u}_{i,l}^{(h,1)})_{i \in [n]}$ and $(\dot{v}_{j,l}^{(h,1)})_{j \in [n]}$ will be the same as shown in Theorem 4.3. We use this step to reduce the instability of our estimator due to the random sample splitting.

In Step 5, we apply the K-means algorithm to the normalized $\{(\dot{u}_{i,l}^{(h,1)})_{i \in [n]}, (\dot{v}_{j,l}^{(h,1)})_{j \in [n]}\}$. We show in Theorem 5.1 that if $n\zeta_n / \log n \geq C$ for some sufficiently large and positive constant C, we can exactly recover the latent communities. We provide the implementation detail of each step below.

3.1 The Estimation of $(u_{i,l}, v_{i,l})$

In the estimation of $(u_{i,l}, v_{i,l})$ (see Steps 1–5 in the above procedure), we only require that Θ_0^* and Θ_1^* be of low-rank.

3.1.1 FULL-SAMPLE LOW-RANK ESTIMATION

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

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

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

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

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:

$$\widetilde{\Gamma} = \underset{\Gamma \in \mathbb{T}(0, \log n)}{\arg\min}\ Q_n(\Gamma) + \lambda_n \sum_{l=0}^{1} ||\Gamma_l||_*, \qquad (11)$$

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.[4] 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.[5] 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 $\widehat{\tau}_n = \frac{1}{n(n-1)}\sum_{i\neq j}\widehat{\Gamma}_{0,ij}$. We are then able to show $\widehat{\tau}_n - \tau_n = O_p\left(\sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}\right)$ in Theorem 4.1 below.

Since $\Theta_l^* = \{\Theta_{l,ij}^*\}$ are symmetric, we define their preliminary low-rank estimators as $\widehat{\Theta}_l = \{\widehat{\Theta}_{l,ij}\}$, where

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

$\delta_{l0} = \mathbf{1}\{l = 0\}$, $f_M(u) = u\cdot\mathbf{1}\{|u|\leq M\} + M\cdot\mathbf{1}\{u > M\} - M\cdot\mathbf{1}\{u < -M\}$ is the round function, 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 $\widehat{V}_l = \sqrt{n}\widehat{\mathcal{V}}_l$. We will establish in Theorem 4.1 below that $||V_l - \widehat{V}_l\widehat{O}_l||_F/\sqrt{n} = O_p(\eta_n)$.

3.1.2 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 $\ell = 1, 2$.

4. We provide the detailed estimation algorithm in Section E in the Appendix.

5. 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.

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

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

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

$$\widetilde{\Gamma}^{(1)} = \underset{\Gamma \in \mathbb{T}^{(1)}(0, \log n)}{\arg\min} \; Q_n^{(1)}(\Gamma) + \lambda_n^{(1)} \sum_{l=0}^{1} ||\Gamma_l||_*, \tag{12}$$

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

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

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

where $\widehat{\widehat{\Sigma}}_l^{(1)}$ 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, $\widehat{\sigma}_{1,l}^{(1)} \geq \cdots \geq \widehat{\sigma}_{n_1,l}^{(1)} \geq 0$, and $\widehat{\widehat{\mathcal{U}}}_l^{(1)}$ and $\widehat{\widehat{\mathcal{V}}}_l^{(1)}$ 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 $\widehat{\widehat{\mathcal{V}}}_l^{(1)}$ 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 $j \in [n]$. We will establish in Theorem 4.1 below that $||V_l - \widehat{V}_l^{(1)} \widehat{O}_l^{(1)}||_F / \sqrt{n} = O_p(\eta_n)$.

3.1.3 SPLIT-SAMPLE ROW- AND COLUMN-WISE LOGISTIC REGRESSIONS

We note that $\Theta_{l,ij}^* = u_{i,l}^\top v_{j,l}$ for $i \in I_2$ and $j \in [n]$. For the i-th row when $i \in I_2$, we can view $\{v_{j,l}\}_{j \in [n]}$ and $u_{i,l}$ as regressors and the parameter, respectively, and estimate $u_{i,l}$ by the row-wise logistic regression. Although $\{v_{j,l}\}_{j \in [n]}$ are unobservable, we can replace them by their estimators obtained from the previous step.

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 $\ell_{ij}^{\text{left}}(\mu) = Y_{ij} \log(\Lambda_{ij}^{\text{left}}(\mu)) + (1-Y_{ij}) \log(1-\Lambda_{ij}^{\text{left}}(\mu))$, where the superscript "left" means these functions are used to estimate the left singular vector $u_{i,l}$. 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 $Q_{in,U}^{(0)}(\mu) = \frac{-1}{n_2} \sum_{j\in I_2, j\neq i} \ell_{ij}^{\text{left}}(\mu)$ and the superscript (0) means it is the initial step for the full sample iteration below. To keep the independence between $\{\widehat{v}_{j,l}^{(1)}\}_{j\in[n]}$ and the data in this regression, we only use $j \in I_2$ to run the regression.

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 $\ell_{ij}^{\text{right}}(\nu) = Y_{ij} \log(\Lambda_{ij}^{\text{right}}(\nu)) + (1-Y_{ij}) \log(1-\Lambda_{ij}^{\text{right}}(\nu))$, where the superscript "right" means the functions are used to estimate the right singular vector $v_{j,l}$. 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 $Q_{jn,V}^{(0)}(\nu) = \frac{-1}{n_2} \sum_{i\in I_2, i\neq j} \ell_{ij}^{\text{right}}(\nu)$.

We will establish in Theorem 4.2 below that

$$\max_{i\in I_2} ||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| = O_p(\eta_n) \quad \text{and} \quad \max_{j\in[n]} ||(\widehat{O}_l^{(1)})^\top \dot{v}_{j,l}^{(0,1)} - v_{j,l}|| = O_p(\eta_n).$$

Our final objective is to obtain accurate estimates of $\{v_{j,l}\}_{j\in[n],l=0,1}$ in finite samples. 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.

3.1.4 Full-Sample Iteration

Given the initial estimates, we use the full sample and iteratively run row- and column-wise logistic regressions to estimate $\{u_{i,l}, v_{i,l}\}_{i\in[n]}$. For $h = 1, 2, ..., H$, let

$$\Lambda_{ij}^{\text{left},h}(\mu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \mu_l^\top \dot{v}_{j,l}^{(h-1,1)} W_{l,ij}) \quad \text{and}$$
$$\ell_{ij}^{\text{left},h}(\mu) = Y_{ij} \log(\Lambda_{ij}^{\text{left},h}(\mu)) + (1 - Y_{ij}) \log(1 - \Lambda_{ij}^{\text{left},h}(\mu)).$$

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 the row-wise logistic regression

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

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 the column-wise logistic regression

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

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 $h = 1, 2, \cdots, H$. We will establish in Theorem 4.3 below that

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

3.2 K-means Classification

In this step, we further assume Θ_1^* has the latent community structure and Θ_0^* remains to be of low-rank. 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

$$\hat{g}_j = \underset{1\le k\le K_1}{\arg\min} \left\| \overline{v}_{j,1} - \widehat{\beta}_k \right\|, \tag{13}$$

where if there are multiple k's that achieve the minimum, $\hat{g}_j$ takes value of the smallest one. We establish in Theorem 5.1 below that $\hat{g}_j$ estimates the true group identity g_j^0 for node j uniformly well $w.p.a.1$.

As mentioned previously, we can repeat Steps 2–6 R times to obtain R membership estimates, denoted as $\{\hat{g}_{j,r}\}_{j\in[n], r\in[R]}$. Recall that

$$\text{vech}(B_1^*) = (B_{1,11}^*, ..., B_{1,1K_1}^*, B_{1,22}^*, \cdots, B_{1,2K_1}^*, \cdots, B_{1,K_1-1,K_1-1}^*, B_{1,K_1-1,K_1}^*, B_{1,K_1K_1}^*)^\top,$$

which is a $K_1(K_1+1)/2$-vector. In addition, let $\chi_{1,ij}$ be the vectorization of the upper triangular part of the $K_1 \times K_1$ matrix whose (g_i^0, g_j^0) and (g_j^0, g_i^0) entries are one and the rest entries are zero, i.e., $\chi_{1,ij}$ is 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 5.[6] 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

$$r^* = \arg\max_{r\in[R]} \widehat{\mathcal{L}}(r). \tag{14}$$

4. Statistical Properties of the Estimators of $(u_{i,l}, v_{j,l})$

In this section, we study the asymptotic properties of the estimators of $(u_{i,l}, v_{j,l})$ proposed in the last section.

4.1 Full- and Split-Sample Low-Rank Estimations

Suppose the singular value decomposition of Γ_l^* is $\Gamma_l^* = \overline{U}_l\Sigma_l\overline{V}_l^\top$ for $l = 0, 1$ and $\overline{U}_{l,c}$ and $\overline{V}_{l,c}$ are the left and right singular matrices corresponding to the zero singular values. Let $\mathcal{P}_l(\Delta) = \overline{U}_{l,c}\overline{U}_{l,c}^\top\Delta\overline{V}_{l,c}\overline{V}_{l,c}^\top$ for some $n \times n$ matrix Δ and $\mathcal{M}_j(\Delta) = \Delta - \mathcal{P}_j(\Delta)$. Define the restricted low-rank set as, for some $c_1 > 0$

$$\mathcal{C}(c_1) = \left\{(\Delta_0, \Delta_1): \quad ||\mathcal{P}_0(\Delta_0)||_* + ||\mathcal{P}_1(\Delta_1)||_* \le c_1||\mathcal{M}_0(\Delta_0)||_* + c_1||\mathcal{M}_1(\Delta_1)||_*\right\}. \tag{15}$$

Assumption 4 *For any $c_1 > 0$, there exist constants $\kappa, c_2, c_3 > 0$,*

$$\begin{aligned}
\mathcal{C}_1(c_2) &= \{(\Delta_0, \Delta_1): ||\Delta_0||_F^2 + ||\Delta_1||_F^2 \le c_2\log(n)n/\zeta_n\}, \text{ and} \\
\mathcal{C}_2(c_3) &= \{(\Delta_0, \Delta_1): ||\Delta_0 + \Delta_1 \odot W_1||_F^2 \ge \kappa(||\Delta_0||_F^2 + ||\Delta_1||_F^2) - c_3\log(n)n/\zeta_n\},
\end{aligned}$$

such that

$$\mathcal{C}(c_1) \subset \mathcal{C}_1(c_2) \cup \mathcal{C}_2(c_3) \ w.p.a.1.$$

The same condition holds when (Γ_0^*, Γ_1^*) are replaced by $(\Gamma_0^*(I_1), \Gamma_1^*(I_1))$ and $(\Gamma_0^*(I_2), \Gamma_1^*(I_2))$.

Several remarks are in order. First, Assumption 4 is a slight generalization of Chernozhukov et al. (2020, Assumption 3.1) where, in terms of our notation, $\mathcal{C}_1(c_2)$ and $\mathcal{C}_2(c_3)$ take the forms:

$$\mathcal{C}_1(c_2) = \{(\Delta_0, \Delta_1): ||\Delta_0||_F^2 + ||\Delta_1||_F^2 \le c_2 n\} \text{ and}$$

6. If we have multiple covariates W_l, $l \in [p]$, to compute $\mathcal{L}_{n,r}(b)$, we let $\widehat{\Theta}_{l,ij} = [(\dot{u}_{i,l}^{(H,1)})^\top \dot{v}_{j,l}^{(H,1)} + (\dot{u}_{i,l}^{(H,2)})^\top \dot{v}_{j,l}^{(H,2)}]/2$ when Θ_l^* is only assumed to be of low-rank. For those Θ_l^*'s that have latent communities, for the r-th split, we can estimate their memberships by $\hat{g}_{i,l,r}$ and construct $\hat{\chi}'_{lr,ij}$ similarly. Then, we can define $\mathcal{L}_{n,r}(b)$ and $\widehat{\mathcal{L}}(r)$ in the same manner.

$$\mathcal{C}_2(c_3) \quad = \quad \{(\Delta_0, \Delta_1) : ||\Delta_0 + \Delta_1 \odot W_1||_F^2 \geq \kappa(||\Delta_0||_F^2 + ||\Delta_1||_F^2) - nc_3\}.$$

Such a generalization is due to the fact that the network can be semi-dense, and thus, the convergence rates of our estimators of the singular vectors are slower than those of Chernozhukov et al. (2020)'s estimators.

Second, even if there are two sets of parameters (Γ_0^*, Γ_1^*) and $(\Gamma_0^\dagger, \Gamma_1^\dagger)$ with $\Gamma_l^* \neq \Gamma_l^\dagger$ for some $l \in \{0, 1\}$ such that both satisfy Assumption 4 and

$$\Gamma_0^* + \Gamma_1^* \odot W_1 = \Gamma_0^\dagger + \Gamma_1^\dagger \odot W_1,$$

such an ambiguity will not affect the the rate of convergence. To see this, note the singular value decomposition of $\Gamma_l^\dagger$ is $\Gamma_l^\dagger = \tilde{U}_l \tilde{\Sigma}_l \tilde{V}_l^\top$ for $l = 0, 1$ and $\tilde{U}_{l,c}$ and $\tilde{V}_{l,c}$ are the left and right singular matrices corresponding to the zero singular values. Denote $\widetilde{\mathcal{P}}_l(\Delta) = \tilde{U}_{l,c}\tilde{U}_{l,c}^T \Delta \tilde{V}_{l_c} \tilde{V}_{l,c}^T$ and $\widetilde{\mathcal{M}}_l(\Delta) = \Delta - \widetilde{\mathcal{P}}_l(\Delta)$. Suppose that

$$\left\{\tilde{\mathcal{C}}(c_1) = (\Delta_0, \Delta_1) : \quad ||\widetilde{\mathcal{P}}_0(\Delta_0)||_* + ||\widetilde{\mathcal{P}}_1(\Delta_1)||_* \leq c_1||\widetilde{\mathcal{M}}_0(\Delta_0)||_* + c_1||\widetilde{\mathcal{M}}_1(\Delta_1)||_*\right\},$$

Assumption 4 holds for both $\mathcal{C}(c_1)$ and $\tilde{\mathcal{C}}(c_1)$. Denote $\Delta_l = \Gamma_l^\dagger - \Gamma_l^*, l = 0, 1$. Then $\Delta_0 + \Delta_1 \odot W_1 = 0$ and it is possible to show that (Δ_0, Δ_1) belongs to either $\mathcal{C}(1)$ or $\tilde{\mathcal{C}}(1)$.[7] If $(\Delta_0, \Delta_1) \notin \mathcal{C}_1(c_2)$, then Assumption 4 implies

$$0 = ||\Delta_0 + \Delta_1 \odot W_1||_F^2 \geq \kappa(||\Delta_0||_F^2 + ||\Delta_1||_F^2) - c_3 \log(n)n/\zeta_n,$$

and thus,

$$c_3 \log(n)n/(\kappa\zeta_n) \geq ||\Delta_0||_F^2 + ||\Delta_1||_F^2 > c_2 \log(n)n/\zeta_n.$$

Therefore,

$$||\Delta_0||_F^2 + ||\Delta_1||_F^2 \leq (c_2 \vee \kappa^{-1}c_3)\log(n)n/\zeta_n.$$

For any estimator $\hat{\Gamma}_l$ of Γ_l^*, $l = 0, 1$, we have, w.p.a.1,

$$\left|\frac{1}{n}(\sum_{l=0}^{1} ||\hat{\Gamma}_l - \Gamma_l^*||_F) - \frac{1}{n}(\sum_{l=0}^{1} ||\hat{\Gamma}_l - \Gamma_l^\dagger||_F)\right| \leq \frac{1}{n}(||\Delta_0||_F + ||\Delta_1||_F) \leq \sqrt{\frac{2(c_2 \vee \kappa^{-1}c_3)\log(n)}{n\zeta_n}}$$

7. Without loss of generality, we assume that $||\Gamma_0^\dagger||_* + ||\Gamma_1^\dagger||_* \leq ||\Gamma_0^*||_* + ||\Gamma_1^*||_*$. Noting that

$$\begin{aligned} ||\Gamma_l^\dagger||_* =& ||\Gamma_l^* + \mathcal{M}_l(\Delta_l) + \mathcal{P}_l(\Delta_l)||_* \\ \geq& ||\Gamma_l^* + \mathcal{P}_l(\Delta_l)||_* - ||\mathcal{M}_l(\Delta_l)||_* \\ =& ||\Gamma_l^*||_* + ||\mathcal{P}_l(\Delta_l)||_* - ||\mathcal{M}_l(\Delta_l)||_* \text{ for } l = 0, 1, \end{aligned}$$

where the last equality holds due to Chernozhukov et al. (2020, Lemma D.2(i)), we have

$$\begin{aligned} ||\Gamma_0^*||_* + ||\Gamma_1^*||_* \geq& ||\Gamma_0^\dagger||_* + ||\Gamma_1^\dagger||_* \\ \geq& ||\Gamma_0^*||_* + ||\mathcal{P}_0(\Delta_0)||_* - ||\mathcal{M}_0(\Delta_0)||_* + ||\Theta_1^*||_* + ||\mathcal{P}_1(\Delta_1)||_* - ||\mathcal{M}_1(\Delta_1)||_*, \end{aligned}$$

which implies

$$||\mathcal{P}_0(\Delta_0)||_* + ||\mathcal{P}_1(\Delta_1)||_* \leq ||\mathcal{M}_0(\Delta_0)||_* + ||\mathcal{M}_1(\Delta_1)||_*,$$

i.e., $(\Delta_0, \Delta_1) \in \mathcal{C}(1)$.

Based on Assumption 4 and other conditions in the paper, we can show that (see Theorem 4.1 below)

$$\frac{1}{n}(\sum_{l=0}^{1}||\hat{\Gamma}_l - \Gamma_l^*||_F) \leq 48C_{F,1}\left(\sqrt{\frac{\log(n)}{n\zeta_n}} + \frac{\log(n)}{n\zeta_n}\right) \ w.p.a.1,$$

where $C_{F,1}$ is some constant. This implies

$$\frac{1}{n}(\sum_{l=0}^{1}||\hat{\Gamma}_l - \Gamma_l^\dagger||_F) \leq \left(48C_{F,1} + \sqrt{2(c_2 \vee \kappa^{-1}c_3)}\right)\left(\sqrt{\frac{\log(n)}{n\zeta_n}} + \frac{\log(n)}{n\zeta_n}\right) \ w.p.a.1$$

and vise versa. The same conclusion holds if $(\Delta_0, \Delta_1) \in \mathcal{C}_1(c_2)$. As a result, the ambiguity between Θ_l^* and $\tilde{\Theta}_l$ is asymptotically negligible and will not affect the convergence rates of their estimators.

Third, Chernozhukov et al. (2020, Appendix D.3) provide a sufficient condition for Assumption 4. Recall $W_{1,ij} = g_1(X_i, X_j, e_{ij})$. Following the same arguments in Chernozhukov et al. (2020, Appendix D.3), it is possible to show that Assumption 4 holds if $W_{1,ij}$ is bounded and $Var(W_{1,ij}|X_i, X_j) > 0$.[8] The sufficient condition basically requires the existence of e_{ij} in $g_l(\cdot)$ which is a sequence of i.i.d. random variables across i, j.[9] Note that the presence of e_{ij} is sufficient, but may not be necessary. In our simulation, we generate $W_{1,ij} = |X_i - X_j|$ with $\{X_i\}_{i\in[n]}$ being a sequence of i.i.d. standard normal random variables, and find that our method works well.

Fourth, Assumption 4 rules out the case $W_{1,ij} = g_1(X_i, X_j)$ when X_i is discrete, which is equivalent to a community structure of $W_{1,ij}$. Suppose $W_{1,ij} = w_{k_1k_2} > 0 \ \forall k_1, k_2$ where i, j are in groups k_1 and k_2. Then, we can let Δ_1 share the same community structure as W_1 and $\Delta_{1,ij} = w_{k_1k_2}^{-1}$. Let $\Delta_0 = -\iota_n\iota_n^\top$. Then we have

$$\Delta_0 + \Delta_1 \odot W_1 = 0 \quad \text{and} \quad ||\Delta_0||_F^2 + ||\Delta_1||_F^2 \geq ||\Delta_0||_F^2 = n^2.$$

Because both Δ_1 and Δ_0 are of low-rank, we have

$$||\mathcal{P}_0(\Delta_0)||_* + ||\mathcal{P}_1(\Delta_1)||_* \leq ||\Delta_0||_* + ||\Delta_1||_* \leq Cn,$$

for some constant $C > 0$. In addition, the singular value decomposition of Δ_0 is $\Delta_0 = (-\iota_n/\sqrt{n}) \times n \times (\iota_n/\sqrt{n})^\top$. It is possible to find some parameter Θ_0 such that $||\mathcal{M}_0(\Delta_0)||_* \geq cn$ for some $c > 0$.[10] Then we can take $c_1 = C/c$ so that

$$||\mathcal{P}_0(\Delta_0)||_* + ||\mathcal{P}_1(\Delta_1)||_* \leq c_1||\mathcal{M}_0(\Delta_0)||_* \leq c_1||\mathcal{M}_0(\Delta_0)||_* + c_1||\mathcal{M}_1(\Delta_1)||_*.$$

In this case, Assumption 4 does not hold because $||\Delta_0||_F^2 + ||\Delta_1||_F^2 \geq n^2 > c_2\log(n)n/\zeta_n$[11] and

$$0 = ||\Delta_0 + \Delta_1 \odot W_1||_F^2 < \kappa n^2 - c_3\log(n)n/\zeta_n \leq \kappa(||\Delta_0||_F^2 + ||\Delta_1||_F^2) - c_3\log(n)n/\zeta_n.$$

8. In the general case with multiple covariates, they require $\min_{i,j}\lambda_{\min}(\mathbb{E}W_{ij}W_{ij}^\top|X_i, X_j) \geq c > 0$ where $\lambda_{\min}(A)$ is the minimum eigenvalue of matrix A and $W_{ij} = (1, W_{1,ij}, \cdots, W_{p,ij})^\top$.
9. Note there are two key differences between the setups in our paper and Chernozhukov et al. (2020). First, Chernozhukov et al. (2020) consider the panel data with indexes $i \in [N]$ and $t \in [T]$ while we consider the network data with indexes $(i,j) \in \{1 \leq i < j \leq n\}$. Second, Chernozhukov et al. (2020) consider $X_{it} = \mu_{it} + e_{it}$ such that given $\{\mu_{it}\}_{i\in[N],t\in[T]}$, X_{it} is independent across both t and t. Instead, we consider $W_{1,ij} = g_1(X_i, X_j, e_{ij})$ such that given $\{X_i\}_{i\in[n]}$, $W_{1,ij}$ is independent across $1 \leq i < j \leq n$. By examining the proofs of Chernozhukov et al. (2020, Lemmas D.3 and D.4), we note that their argument does not rely on the special structure of $X_{it} = \mu_{it} + e_{it}$ and works if $X_{it} = f(\mu_{it}, e_{it})$ for some non-additive function f.
10. This occurs, say, when $\iota_n/\sqrt{n}$ is in the spaces spanned by the left and right singular vectors of Θ_0 that correspond to its nonzero singular values.
11. When $\zeta_n = C_\varsigma n^{-1}\log(n)$, we require that C_ς is sufficiently large.

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

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

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 5.1 below. Assumption 5.4 usually holds as the sample is split randomly and Θ_0^* satisfies the normalization condition in Assumption 2.1. If Θ_0^* satisfies the additive structure as in Example 1, then Assumption 5.4 holds provided that $\frac{1}{n_1}\sum_{i \in I_1} \alpha_i = o(\sqrt{\frac{\log(n)}{n\zeta_n}})$. Such a requirement holds almost surely 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. If Θ_0^* has the community structure as in Example 2, then Assumption 5.4 holds provided that $p_0^\top(I_1) B_0^* p_0 = o(\sqrt{\frac{\log(n)}{n\zeta_n}})$, where $p_0^\top(I_1) = (\frac{n_{1,0}(I_1)}{n_1}, \cdots, \frac{n_{K_0,0}(I_1)}{n_1})$ 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 almost surely 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, 2, 4, and 5 hold and* $\eta_n = \sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}$*. Then for* $l = 0, 1$ *and w.p.a.1, we have*

1. $|\widehat{\tau}_n - \tau_n| \leq 30 C_{F,1}\eta_n$, $|\widehat{\tau}_n^{(1)} - \tau_n| \leq 30 C_{F,1}\eta_n$,

2. $\frac{1}{n}||\widehat{\Theta}_l - \Theta_l^*||_F \leq 48 C_{F,1}\eta_n$, $\frac{1}{n}||\widehat{\Theta}_l^{(1)} - \Theta_l^*(I_1)||_F \leq 48 C_{F,1}\eta_n$,

3. $\max_{k \in [K_l]} |\widehat{\sigma}_{k,l} - \sigma_{k,l}| \leq 48 C_{F,1}\eta_n$, $\max_{k \in [K_l]} |\widehat{\sigma}_{k,l}^{(1)} - \sigma_{k,l}| \leq 48 C_{F,1}\eta_n$,

4. $||V_l - \widehat{V}_l \widehat{O}_l||_F \leq 136 C_{F,2}\sqrt{n}\eta_n$*, and* $||V_l - \widehat{V}_l^{(1)}\widehat{O}_l^{(1)}||_F \leq 136 C_{F,2}\sqrt{n}\eta_n$,

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 (31) and (32) 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 $\widehat{V}_l^{(1)}/\sqrt{n}$. In part 4 of Theorem 4.1, the orthogonal matrices $\widehat{O}_l$ and $\widehat{O}_l^{(1)}$ are present because the singular values of Θ_l^* can be the same and its singular vectors can only be identified up to some rotation.

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 w.p.a.1,*

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

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 Γ_1 follows the community structure as in Example 2 with K_1 equal-sized communities and $B_1^* = I_{K_1}$, then $\Pi_{1,n} = \text{diag}(1/K_1, \cdots, 1/K_1)$. By Lemma B.4 in the Appendix, 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. In addition, suppose Θ_0 follows the specification in Example 1. Then,

$$\Phi_i(I_2) = \frac{1}{n_2} \sum_{j \in I_2} \begin{pmatrix} \frac{1}{\sqrt{2}}(1 + \frac{\alpha_{0,j}}{s_{0,n}}) \\ \frac{1}{\sqrt{2}}(1 - \frac{\alpha_{0,j}}{s_{0,n}}) \\ v_{j,1} W_{1,ij} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}}(1 + \frac{\alpha_{0,j}}{s_{0,n}}) \\ \frac{1}{\sqrt{2}}(1 - \frac{\alpha_{0,j}}{s_{0,n}}) \\ v_{j,1} W_{1,ij} \end{pmatrix}^\top.$$

Suppose that $\alpha_{0,i} = a_i - \bar{a}$ for some i.i.d. sequence $\{a_i\}_{i=1}^n$ with $\bar{a} = \frac{1}{n}\sum_{i=1}^n a_i$, and the group identities of Θ_1^* ($\{z_i\}_{i \in [n]}$) are independent of Θ_0^* and $\{X_i\}_{i \in [n]}$ and $\{e_{ij}\}_{i,j \in [n]}$. Further suppose $\mathbb{E}(W_{1,ij} a_j | X_i) = 0$, $\mathbb{E}(W_{1,ij} | X_i) = 0$, and $\mathbb{E}(W_{1,ij}^2 | X_i) \geq c > 0$ for some constant c. Then, 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.

If Θ_0^* has the community structure as in Example 2. Further suppose Θ_0^* and Θ_1^* share the same community structure Z_1, which is independent of W_1, $\mathbb{E}(W_{1,ij} | X_i) = 0$ 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 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, 2, 4–6 hold. Then,*

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

where C_1^* and $C_{0,v}$ are some constants defined respectively in (44) and (48) 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. 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.

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 $w.p.a.1$*

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

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

Theorem 4.3 *Suppose that Assumptions 1, 2, 4–7 hold. Then, for $h=1,\cdots,H$ and $l=0,1$,*

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

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

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

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 $h=1,\cdots,H$,

$$\max_{i\in[n]}||(\widehat{O}_l^{(2)})^\top \dot{u}_{i,l}^{(h,2)} - u_{i,l}|| \le C_{h,u}\eta_n \quad \text{and} \quad \max_{i\in[n]}||(\widehat{O}_l^{(2)})^\top \dot{v}_{i,l}^{(h,2)} - v_{i,l}|| \le C_{h,v}\eta_n \;\; w.p.a.1$$

5. K-means Classification

If we further assume Θ_1^* has the community structure and satisfies Assumption 3, then 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 $\overline{v}_{j,l}$ defined in Section 3.2 to recover the memberships as long as the estimation error is sufficiently small.

Let $g_i^0 \in [K_1]$ denote the true group identity for the i-th node in Θ_1^*. To establish the strong consistency of the membership estimator $\hat{g}_i$ defined in (13), we add the following condition.

Assumption 8 *Suppose* $145K_1^{3/2}C_{H,v}C_1\eta_n \leq 1$*, 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 $\eta_n = o(1)$. The constant in the statement is not optimal.

Theorem 5.1 *If Assumptions 1, 2, 4 –8 hold and* Θ_1^* *further satisfies Assumption 3, then up to some label permutation,*

$$\max_{1\leq i\leq n} \mathbf{1}\{\hat{g}_i \neq g_i^0\} = 0 \; w.p.a.1.$$

Several remarks are in order. First, Theorem 5.1 implies the K-means algorithm can exactly recover the latent community structure of Θ_1^* $w.p.a.1$. Second, if we repeat the sample split R times, we need to maintain Assumption 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 (14) also enjoys the property that

$$\max_{1\leq i\leq n} \mathbf{1}\{\hat{g}_{i,r^*} \neq g_i^0\} = 0 \; w.p.a.1.$$

Third, if Θ_0^* also has the latent community 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.

6. Inference for B_1^*

In this section, we maintain the assumption that Θ_1^* has a latent community structure. In the general model with multiple covariates, we allow $\{\Theta_l^*\}_{l\in[p]}$ to have potentially different community structures. Note this includes the case that some of the Θ_l^*'s are homogeneous. We can recover the community structures by applying the K-means algorithm in the previous section to each Θ_l^*.

For the rest of the section, for notation simplicity, we continue to consider the case that there is only one covariate W_1 and Θ_1^* has a latent community structure, which is estimated by $\{\hat{g}_i\}_{i\in[n]}$ defined in the previous section. Given the exact recovery of the community memberships asymptotically, we can just treat $\hat{g}_i$ as g_i^0.

We discuss the inference for B_1^* for two specifications of Θ_0^*: (1) $\Theta_{0,ij}^*$ has an additive structure as in Example 1 and (2) $\Theta_{0,ij}^*$ has a latent community structure as in Example 2. In the first specification, once the group membership of Θ_1^* is recovered, the model boils down to the one studied by Graham (2017). For the second specification, when the memberships of both Θ_0^* and Θ_1^* are recovered, the model boils down to the standard logistic regression with finite-number of parameters.

6.1 Additive Fixed Effects

Suppose $\Gamma_{0,ij}^* = \tau_n + \alpha_i + \alpha_j$ 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 3.2 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 6.1 *Suppose Assumptions 1, 2, 4–8 hold and Θ_1^* further satisfies Assumption 3. Then $\hat{\chi}_{1,ij} = \chi_{1,ij}\ \forall i < j$ w.p.a.1.*

Corollary 6.1 directly follows from Theorem 5.1 and implies that we can treat $\chi_{1,ij}$ as observed. Then, (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 $\text{vec}(B_1^*)$.

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 $\text{vech}(B^*)$ 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 $t_{q,n} = \mathbb{P}(\top_{i_1 i_2 i_3 i_4} = 1, \top_{j j_2 j_3 j_4} = 1)$ 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 $\Upsilon_0 \equiv \lim_{n\to\infty} t_{4,n}^{-1} \sum_{i<i'<j<j'} \nabla_{BB} \bar{\ell}_{ij,i'j'}(B)$ is a finite nonsingular matrix.*

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

Theorem 6.1 *Suppose that Assumptions 1, 2, 4–9 hold. Suppose that $\Gamma_0^* = \tau_n + \alpha_i + \alpha_j$ and Θ_1^* satisfies Assumption 3. Then $\widehat{B} \xrightarrow{p}$ vec(B^*) and*

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

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.

Theorem 6.1 imposes two additional structures in order to make the inferences on B^* by borrowing the asymptotic results from Graham (2017). One is that Γ_0^* exhibits the usual additive fixed effects structure (with $K_0 = 2$) and the other is Γ_1^* has a latent community structure. The model reduces to that of Graham (2017) in the special case of $K_1 = 1$.

6.2 Latent Community Structure in the Fixed Effects

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 denotes a $b\times 1$ vector of ones. Therefore, $\Gamma_0^* = \tau_n\iota_n\iota_n^\top + Z_0B_0^*Z_0^\top = Z_0(B_0^* + \tau_n\iota_{K_0}\iota_{K_0}^\top)Z_0^\top \equiv Z_0B_0^{**}Z_0^\top$, i.e., Γ_0^* shares the same community structure as Θ_0^*. We then define $\chi_{0,ij}$ as 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 $\hat{\chi}_{0,ij}$ as 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. Similar to Corollary 6.1, we have the following corollary.

Corollary 6.2 *Suppose that Assumptions 1, 2, 4–8 hold. Suppose that Θ_l^*, $l = 0, 1,$ further satisfy Assumption 3. Then, $\hat{\chi}_{l,ij} = \chi_{l,ij}$ $\forall i < j$ for $l = 0, 1$ $w.p.a.1$.*

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

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

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

$$\hat{\Lambda}_{ij}(b) = \Lambda(\hat{\chi}_{0,ij}^\top b_0 + \hat{\chi}_{1,ij}^\top W_{1,ij}b_1).$$

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 $\Lambda_{n,ij} = \Lambda(W_{ij}^\top\Gamma_{ij}^*)$.

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 6.2 *Suppose that Assumptions 1, 2, 4–8, 10 hold and* Θ_l^*, $l = 0, 1$, *further satisfy Assumption 3. 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*

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

Although in theory, the inference for B_1^* in the above two cases is straightforward, there are two finite-sample issues. First, the Tetrad logit regression does not scale with the number of nodes n as it needs to scan over all four-node figurations, which contains a total of $\Omega(n^4)$ operations in a brutal force implementation. Such a step is inevitable even when we know the true memberships. 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 a small bias may not be ignored in making inferences. 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.

7. Determination of 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, for any given k satisfying $1 \leq k \leq K_{\max}$, where $K_{\max}$ is a large but fixed integer, we first obtain the singular value estimates $\{\hat{\sigma}_{k,l}\}_{k\in[K_{\max}],l=0,1}$ from Step 1 of the estimation algorithm given in Section 3. We then 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

$$\widehat{K}_l = \arg\max_{1\leq k\leq K_{\max}-1} \frac{\widehat{\sigma}_{k,l}}{\widehat{\sigma}_{k+1,l}} \mathbf{1}\left\{\widehat{\sigma}_{k,l} \geq c_l \left(\sqrt{\frac{\log n}{n\bar{Y}}} + \frac{\log n}{n\bar{Y}}\right)\right\}, \tag{16}$$

where $\bar{Y} = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} Y_{ij}$, and c_l is a 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 $\hat{K}_l$.

8. Monte Carlo Simulations

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

8.1 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 5.

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.

8.2 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 3.1.1. We estimate K_1 by $\widehat{K}_1$ defined in (16) 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 $\lambda^{(1)}_n$. To require that the estimator of $\widehat{\Theta}_{l,ij}$ is 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.

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

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

Rand Index (RI) and the proportion (PROP) of nodes whose memberships are correctly identified. 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 6 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 (in the Appendix) 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.

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	0.945	0.945	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	0.945	0.955	0.945	0.945	0.945	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

9. Empirical applications

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

9.1 Pokec social network

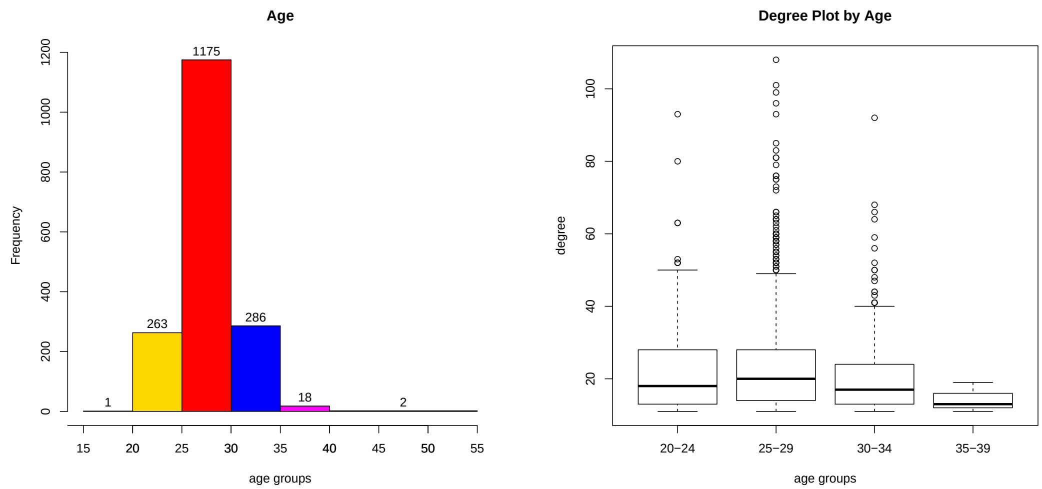
9.1.1 The dataset and model

Pokec is a popular on-line social network in Slovakia. The whole dataset has more than 1.6 million users, and it can be downloaded from https://snap.stanford.edu/data/soc-Pokec.html. In this social network, nodes are anonymized users of Pokec and edges represent friendships. Moreover, demo-

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 2 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

Figure 1: left panel depicts the number of nodes in different age groups; right panel shows the boxplots of degrees by age groups.



graphical features of the users are provided, including gender, age, hobbies, interest, education, etc. To illustrate our method, we select the first 10000 users. Each user is a node in the graph. After deleting the nodes with missing values in age and with degree less than 10, we have 1745 nodes in our dataset. We use the continuous variable, age, as the covariate in our model, and use the friendship network to create an undirected adjacency matrix which has 1745 nodes and 39650 edges. The average degree in this dataset is 22.72. The left panel of Figure 1 shows the number of nodes in different age groups. We see that the age group of 25-29 is the largest group with 1175 users and the age groups of 20-24 and 30-34 have similar number of users. Around 98.8% of users are between the ages of 20 and 35 years old. Moreover, in the right panel of Figure 1, we depict the boxplots of degrees (the number of users connected to each user) for the four age groups 20-24, 25-29, 30-34 and 35-39 that include most users. The plots of degrees vary across different age groups, indicating that age may play a role in the prediction of connections between users.

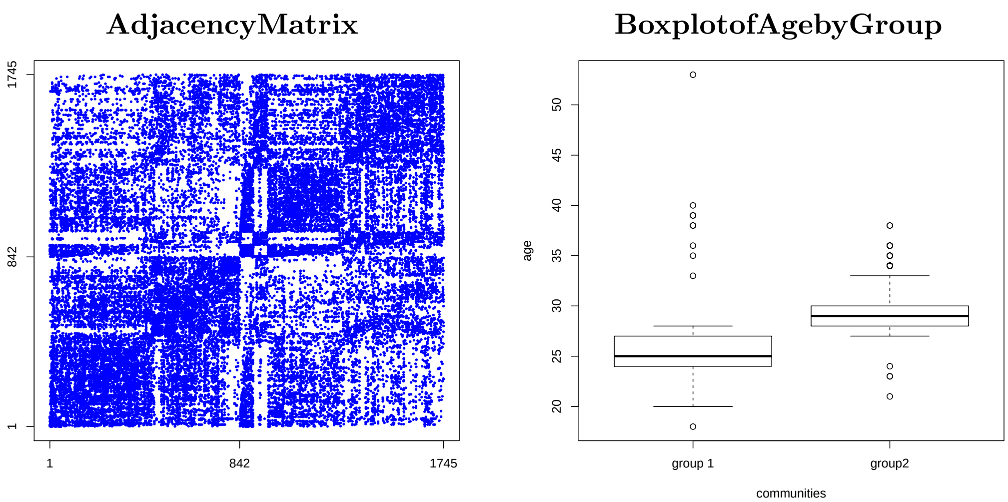
We consider fitting the model:

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \leq \tau_n + \Theta^*_{0,ij} + W_{1,ij}\Theta^*_{1,ij}\}, \quad i > j, \tag{17}$$

for $i = 1, ..., 1745$, where Y_{ij} is the observed value (0 or 1) of the adjacency matrix in our dataset, and $W_{1,ij} = |X_i - X_j|/(\sqrt{X_i^2 + X_j^2})$, in which X_i is the normalized age of the i^{th} customer.[12] 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. Model (17) considered for this real application is similar to Model 2 in the simulation, and it allows for not only the main effect but also possible interaction effects of age and the latent community structure.

12. The variable $W_{1,ij}$ takes 1444 distinctive values. Given there are only 1745 nodes in our dataset, we can view W_{ij} as continuous.

Figure 2: left panel depicts the friendship network with two communities; right panel shows the adjacency matrix reorganized according to the node's memberships.



9.1.2 ESTIMATION RESULTS

We first use the singular-value ratio method to obtain the estimated number of groups for Θ_0^* and Θ_1^*: $\widehat{K}_0 = 2$ and $\widehat{K}_1 = 2$, i.e., we identify two subgroups in the friendship network.

Next, we use our proposed method to obtain the estimated membership for each node. As a result, we have identified 842 nodes in one community and 903 nodes in the other community. We reorganize the observed adjacency matrix according to the estimated memberships of the nodes, i.e., the nodes in the same estimated community are put together in the adjacency matrix. We use blue dots to represent the edges between nodes. The left panel of Figure 2 displays the reorganized adjacency. We see that nodes within each community are generally more densely connected than nodes between communities. In the right panel of Figure 2, we show the boxplots of age for the two identified subgroups. We can observe that in general, the values of age in group 1 are smaller than those in group 2.

Last, Table 5 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 covariate age has a significant effect on the prediction of the friendships between users.

Table 5: The estimates of B_0^* and B_1^* and their standard errors (s.e.).

	$B_{0,11}^*$	$B_{0,12}^*$	$B_{0,22}^*$	$B_{1,11}^*$	$B_{1,12}^*$	$B_{1,22}^*$
estimate	-3.922	-4.119	-3.425	-0.444	-0.518	-0.477
s.e.	0.017	0.025	0.017	0.027	0.019	0.016

9.2 Facebook friendship network

9.2.1 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 et al. (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:

$$Y_{ij} = \mathbf{1}\{\varepsilon_{ij} \le \tau_n + \Theta^*_{0,ij} + W_{1,ij}\Theta^*_{1,ij}\}, \quad i > j, \tag{18}$$

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 $k = 1, 2, 3$.[13] 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. (2019) fit a similar regression model as (18) 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.

9.2.2 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 6 presents the number of students 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

13. We note that $W_{1,ij}$ takes 1512 distinctive values. Given there are just 3073 nodes in the dataset, we can view $W_{1,ij}$ as continuous.

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 7 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.

Table 6: 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 7: 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

10. 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.

11. Acknowledgments and Disclosure of Funding

Ma's research was partially supported by NSF grants DMS 1712558 and DMS 2014221 and a UCR Academic Senate CoR grant. Su gratefully acknowledges the support from NSFC under the grant No. 72133002. 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.

Appendix

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 4.1–5.1 in Appendices A.2–A.6, respectively.

A.1 List of constants

Before we prove the main results, we first list the frequently used constants in Table 8. 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 8: Table of Constants

Name	Description
M_W	$\lvert W_{1,ij}\rvert \le M_W$.
M	$\max_{i\in[n],l=0,1} \lvert\Theta^*_{l,ij}\rvert \le M$, used in the definition of $f_M(\cdot)$ and Assumption 2.
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.
$\overline{c}, \underline{c}, c_F$	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 B.1.

A.2 Proof of Lemma 2.1

We prove the results for U_l first. Let $\Pi_{l,n} = Z_l^\top Z_l/n = \text{diag}(\pi_{l,1n}, \cdots, \pi_{l,K_l n})$. Then,

$$(n^{-1}\Theta_l^*)(n^{-1}\Theta_l^*)^\top = n^{-1} Z_l B_l^* \Pi_{l,n} B_l^* Z_l^\top.$$

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

$$\mathcal{U}_1 \tilde{\Omega}_l^2 \mathcal{U}_l^\top = n^{-1} Z_l \Pi_n^{-1/2} S_l \tilde{\Omega}_l^2 S_l^\top \Pi_n^{-1/2} Z_l^\top = n^{-1} Z_l B_l^* \Pi_n B_l^* Z_l^\top = (n^{-1}\Theta_l^*)^2.$$

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

$$U_l = \sqrt{n}\mathcal{U}_l \Sigma_l = Z_l(\Pi_{l,n})^{-1/2} S_l' \Sigma_l.$$

Similarly, by considering the spectral decomposition of $(n^{-1}\Theta_l^*)^\top(n^{-1}\Theta_l^*)$, we can show that $V_l = Z_l(\Pi_{l,n})^{-1/2}S_l$ for some rotation matrix S_l. Parts (2) and (3) can be verified directly by noting that S_l and S_l' are orthonormal, $\Pi_{l,n}$ 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

$$\begin{aligned}
\lambda_n^{(1)}\sum_{l=0}^{1}\left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_*\right) \geq & \frac{1}{n_1(n-1)}\sum_{i\in I_1, j\in[n], i\neq j}\left(Q_{n,ij}(\widetilde{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*)\right) \\
\geq & \frac{1}{n_1(n-1)}\sum_{i\in I_1, j\in[n], i\neq j}\left(\partial_{\Gamma_{ij}}Q_{n,ij}^\top(\Gamma_{ij}^*)\right)^\top\Delta_{ij} \\
= & \frac{-1}{n_1(n-1)}\sum_{i\in I_1, j\in[n], i\neq j}\left(Y_{ij} - \Lambda(W_{ij}^\top\Gamma_{ij}^*)\right)W_{ij}^\top\Delta_{ij} \\
\equiv & \frac{-1}{n_1(n-1)}\sum_{l=0}^{1}\text{trace}(\Upsilon_l^\top\Delta_l),
\end{aligned} \tag{19}$$

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 (19), we have

$$\begin{aligned}
0 \leq & \lambda_n^{(1)}\sum_{l=0}^{1}\left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_*\right) + \frac{1}{n_1(n-1)}\left|\sum_{l=0}^{1}\text{trace}(\Upsilon_l^\top\Delta_l)\right| \\
\leq & \lambda_n^{(1)}\sum_{l=0}^{1}\left(||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_*\right) + \frac{1}{n_1(n-1)}\sum_{l=0}^{1}||\Upsilon_l||_{op}||\Delta_l||_*.
\end{aligned} \tag{20}$$

For some generic $n_1 \times n$ matrix Δ, let $\mathcal{M}_l^{(1)}(\Delta)$ and $\mathcal{P}_l^{(1)}(\Delta)$ be the residual and projection matrices of Δ with respect to $\Gamma_l^*(I_1)$, as defined in Assumption 4. By Chernozhukov et al. (2020, Lemma D.2) and the fact that $\Gamma_0^*(I_1)$ and $\Gamma_1^*(I_1)$ are exact low-rank matrices with ranks upper bounded by $K_0 + 1$ and K_1, respectively, we have $\Delta_l = \mathcal{M}_l^{(1)}(\Delta_l) + \mathcal{P}_l^{(1)}(\Delta_l)$, rank$(\mathcal{M}_0^{(1)}(\Delta_0)) \leq 2K_0 + 2$, rank$(\mathcal{M}_1^{(1)}(\Delta_1)) \leq 2K_1$, and for $l = 0, 1$,

$$||\Delta_l||_F^2 = ||\mathcal{M}_l^{(1)}(\Delta_l)||_F^2 + ||\mathcal{P}_l^{(1)}(\Delta_l)||_F^2 \text{ and } ||\Gamma_l^*(I_1) + \mathcal{P}_l^{(1)}(\Delta_l)||_* = ||\Gamma_l^*(I_1)||_* + ||\mathcal{P}_l^{(1)}(\Delta_l)||_*. \tag{21}$$

This implies that

$$||\Gamma_l^*(I_1)||_* - ||\widetilde{\Gamma}_l^{(1)}||_* = ||\Gamma_l^*(I_1)||_* - ||\Gamma_l^*(I_1) + \mathcal{M}_l^{(1)}(\Delta_l) + \mathcal{P}_l^{(1)}(\Delta_l)||_*$$
$$\leq ||\mathcal{M}_l^{(1)}(\Delta_l)||_* - ||\mathcal{P}_l^{(1)}(\Delta_l)||_*, \quad l = 0, 1. \tag{22}$$

Therefore, combining (20), Lemma B.1, and (22), we have

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

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

$$(C_\lambda - C_\Upsilon M_W) \sum_{l=0}^{1} ||\mathcal{P}_l^{(1)}(\Delta_l)||_* \leq (C_\lambda + C_\Upsilon M_W) \sum_{l=0}^{1} ||\mathcal{M}_l^{(1)}(\Delta_l)||_*, \tag{23}$$

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.
Next, we first aim to show

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

where $C_F = \frac{\sqrt{K}(M_W+1)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} + \sqrt{\frac{c_3}{\kappa}} + \sqrt{c_2}$. We suppose $(\Delta_0, \Delta_1) \notin \mathcal{C}_1(c_2)$, i.e.,

$$\sum_{l=0}^{1} ||\Delta_l||_F^2 > c_2 n \log(n)/\zeta_n, \tag{24}$$

otherwise,

$$\frac{1}{n}(\sum_{l=0}^{1} ||\Delta_l||_F^2)^{1/2} \leq \sqrt{\frac{c_2 \log(n)}{n\zeta_n}} < 17C_F \left(\frac{\log(n)}{\sqrt{n\zeta_n}} + \frac{(\log n)^{3/2}}{n\zeta_n} \right),$$

and we are done.
Now we consider the second-order Taylor expansion of $Q_{n,ij}(\Gamma_{ij})$, following the argument in Belloni et al. (2017). Let $f_{ij}(t) = \log\{1 + \exp(W_{ij}^\top(\Gamma_{ij}^* + t\Delta_{ij}))\}$, where $\Delta_{ij} = (\Delta_{0,ij}, \cdots, \Delta_{p,ij})^\top$. Note

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

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

$$|f_{ij}'''(t)| = |W_{ij}^\top \Delta_{ij}|^3 \Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij}))(1 - \Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij})))|1 - 2\Lambda(W_{ij}^\top(\Delta_{ij} + t\Delta_{ij}))|$$

$$\leq |W_{ij}^\top \Delta_{ij}| f_{ij}^{''}(t).$$

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

$$\begin{aligned}
f_{ij}(1) - f_{ij}(0) - f_{ij}'(0) \geq & \frac{f_{ij}^{''}(0)}{(W_{ij}^\top \Delta_{ij})^2} \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
= & \Lambda(W_{ij}^\top \Gamma_{ij}^*)(1 - \Lambda(W_{ij}^\top \Gamma_{ij}^*)) \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
\geq & \underline{c}\zeta_n \left[\exp(-|W_{ij}^\top \Delta_{ij}|) + |W_{ij}^\top \Delta_{ij}| - 1\right] \\
\geq & \underline{c}\zeta_n \left(\frac{(W_{ij}^\top \Delta_{ij})^2}{4(\max_{i,j} |W_{ij}^\top \Delta_{ij}| \vee \log(2))} \right) \\
\geq & \frac{\zeta_n \underline{c} (W_{ij}^\top \Delta_{ij})^2}{8(M_W + 1) \log n},
\end{aligned} \tag{25}$$

where the third inequality holds by Lemma B.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, w.p.a.1,

$$\begin{aligned}
& F_n(\Delta_0, \Delta_1) \\
\equiv & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], j \neq i} \left[Q_{n,ij}(\widetilde{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) - \partial_{\Gamma_{ij}} Q_{n,ij}^\top (\Gamma_{ij}^*) \Delta_{ij} \right] \\
\geq & \frac{\zeta_n \underline{c}}{8 n_1 (n-1)(M_W + 1) \log n} \sum_{i \in I_1, j \in [n], j \neq i} (W_{ij}^\top \Delta_{ij})^2 \\
\geq & \frac{\zeta_n \underline{c}}{8 n_1 (n-1)(M_W + 1) \log n} \left[\kappa \sum_{l=0}^{1} ||\Delta_l||_F^2 - 4(M_W + 1)^2 (\log n)^2 n_1 - c_3 n \log(n)/\zeta_n \right],
\end{aligned} \tag{26}$$

where the last inequality holds by Assumption 4, (24), and the fact that $|\Delta_{l,ii}| \leq 2 \log n$, $i \in I_1$. On the other hand, by (19),

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

$$
\begin{aligned}
&\leq \frac{\sqrt{\zeta_n n}+\sqrt{\log n}}{n_1(n-1)}(C_\lambda + C_\Upsilon M_W)\sqrt{2\bar{K}}(\sum_{l=0}^{1}||\mathcal{M}_l^{(1)}(\Delta_l)||_F) \\
&\leq \frac{\sqrt{\zeta_n n}+\sqrt{\log n}}{n_1(n-1)}(C_\lambda + C_\Upsilon M_W)\sqrt{2\bar{K}}(\sum_{l=0}^{1}||\Delta_l||_F) \\
&\leq \frac{\sqrt{\zeta_n n}+\sqrt{\log n}}{n_1(n-1)}(C_\lambda + C_\Upsilon M_W)2\sqrt{\bar{K}}(\sum_{l=0}^{1}||\Delta_l||_F^2)^{1/2},
\end{aligned} \tag{27}
$$

where $\bar{K} = \max(K_0+1, K_1)$, the first inequality is due to (19), the second inequality is due to (22) and the trace inequality, the third inequality holds by the definition of $\lambda_n^{(1)}$ and Lemma B.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 rank$(\mathcal{M}_l^{(1)}(\Delta_l)) \leq 2\bar{K}$, the second last inequality is due to (21), and the last inequality is due to the Cauchy's inequality.

Combining (26) and (27), we have

$$
\begin{aligned}
&\left[\left(\sum_{l=0}^{1}||\Delta_l||_F^2\right)^{1/2} - \frac{8\sqrt{\bar{K}}(M_W+1)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa}\frac{\log n[\sqrt{n\zeta_n}+\sqrt{\log n}]}{\zeta_n}\right]^2 \\
&\leq \bar{K}\left[\frac{8(M_W+1)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa}\right]^2\left(\frac{\log n[\sqrt{n\zeta_n}+\sqrt{\log n}]}{\zeta_n}\right)^2 \\
&+ \frac{4n_1(M_W+1)^2(\log n)^2}{\kappa} + \frac{c_3 n\log(n)}{\kappa\zeta_n},
\end{aligned}
$$

and thus,

$$
\frac{1}{n}(\sum_{l=0}^{1}||\Delta_l||_F^2)^{1/2} \leq 17C_F\left(\frac{\log n}{\sqrt{n\zeta_n}} + \frac{(\log n)^{3/2}}{n\zeta_n}\right) \ w.p.a.1. \tag{28}
$$

Then,

$$
\begin{aligned}
|\widetilde{\tau}_n^{(1)} - \tau_n| = &\left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}(\widetilde{\Gamma}_{0,ij}-\tau_n)\right| \leq \left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}(\widetilde{\Gamma}_{0,ij}-\Gamma^*_{0,ij})\right| + \left|\frac{1}{n_1 n}\sum_{i\in I_1, j\in[n]}\Theta^*_{0,ij}\right| \\
&\leq \frac{1}{\sqrt{n_1 n}}||\Delta_0||_F + M \leq 30C_F\left(\frac{\log n}{\sqrt{n\zeta_n}} + \frac{(\log n)^{3/2}}{n\zeta_n}\right) \\
&\leq 30C_F(c_F + c_F^2)\sqrt{\log n} \ w.p.a.1,
\end{aligned} \tag{29}
$$

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 $|\Gamma^*_{1,ij}| \leq C_M\sqrt{\log n}$ and

$$
|\Gamma^*_{0,ij} - \widetilde{\tau}_n^{(1)}| \leq |\Theta^*_{0,ij}| + |\widetilde{\tau}_n^{(1)} - \tau_n| \leq |\Theta^*_{0,ij}| + 30C_F(c_F + c_F^2)\sqrt{\log n} \leq C_M\sqrt{\log n}, \tag{30}
$$

where we use the fact that c_F, and thus, $30(c_F + c_F^2)C_F$ is sufficiently small.

Therefore, following the same arguments used to obtain (23), 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 $f_{ij}(t) = \log(1 + \exp(W_{ij}^\top(\Gamma_{ij}^* + t\widehat{\Delta}_{ij})))$. We aim to show that

$$\frac{1}{n}\left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2\right)^{1/2} \leq 17 C_{F,1}\eta_n \; w.p.a.1, \tag{31}$$

where with $C_{F,1} = \frac{\sqrt{\bar{K}}(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} + \sqrt{\frac{c_3}{\kappa}} + \sqrt{c_2}$ and $\eta_n = \sqrt{\frac{\log n}{n\zeta_n}} + \frac{\log n}{n\zeta_n}$. Following the same argument as before, we can suppose that $(\widehat{\Delta}_0, \widehat{\Delta}_1) \notin \mathcal{C}_1(c_2)$. Then, following (25),

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

where the last inequality holds because of (30) and uniformly in (i, j)

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

Then, similar to (26) and (27),

$$\begin{aligned} & F_n(\widehat{\Delta}_0, \widehat{\Delta}_1) \\ \equiv & \frac{1}{n_1(n-1)} \sum_{i \in I_1, j \in [n], j \neq i} \left(Q_{n,ij}(\widehat{\Gamma}_{ij}^{(1)}) - Q_{n,ij}(\Gamma_{ij}^*) - \partial_{\Gamma_{ij}} Q_{n,ij}^\top(\Gamma_{ij}^*)\widehat{\Delta}_{ij} \right) \\ \geq & \frac{\zeta_n \underline{c}}{8 n_1(n-1)(M_W + C_M)\sqrt{\log n}} \left[\kappa \left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2 \right) - 4(M_W + C_M)^2 \log(n) n_1 - c_3 \log(n) n/\zeta_n \right] \end{aligned}$$

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

$$\begin{aligned} & \left[\left(\sum_{l=0}^{1} ||\widehat{\Delta}_l||_F^2 \right)^{1/2} - \frac{8\sqrt{\bar{K}}(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} \left(\frac{\sqrt{\log n}(\sqrt{n\zeta_n} + \sqrt{\log n})}{\zeta_n} \right) \right]^2 \\ & \leq \bar{K} \left[\frac{8(M_W + C_M)(C_\lambda + C_\Upsilon M_W)}{\underline{c}\kappa} \right]^2 \left(\frac{\sqrt{\log n}(\sqrt{n\zeta_n} + \sqrt{\log n})}{\zeta_n} \right)^2 \\ & + \frac{4(M_W + C_M)^2 \log(n) n_1}{\kappa} + \frac{c_3 \log(n) n}{\kappa \zeta_n}, \end{aligned}$$

and thus, (31) holds. Then, similar to (29) and by Assumption 5.4, 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,

$$
\begin{aligned}
&\frac{1}{n}||\widehat{\Theta}_1^{(1)} - \Theta_1^*(I_1)||_F \\
&\leq \frac{1}{n}\left[\sum_{(i,j)\in I_1\times I_1, i\neq j}\left(\frac{1}{2}(\widehat{\Gamma}_{1,ij}^{(1)} + \widehat{\Gamma}_{1,ji}^{(1)}) - \Theta_{1,ij}^*\right)^2 + \sum_{(i,j):i\in I_1, j\notin I_1}(\widehat{\Gamma}_{1,ij}^{(1)} - \Theta_{1,ij}^*)^2\right]^{1/2} \\
&+ \frac{1}{n}\left(\sum_{i\in I_1}\Theta_{1,ii}^{*2}\right)^{1/2} \\
&\leq \frac{1}{n}\left[\sum_{i\in I_1, j\in[n], i\neq j}(\widehat{\Gamma}_{1,ij}^{(1)} - \Theta_{1,ij}^*)^2\right]^{1/2} + \frac{1}{n}\left(\sum_{i\in I_1}\Theta_{1,ii}^{*2}\right)^{1/2} \\
&\leq \frac{1}{n}\left(\sum_{l=0}^{1}||\widehat{\Delta}_l||_F^2\right)^{1/2} + \sqrt{\frac{M^2}{3n}} \leq 18C_{F,1}\eta_n \ w.p.a.1,
\end{aligned}
$$

where the first inequality holds due to the facts that $f_M(\cdot)$ is 1-Lipschitz continuous, $\Theta_1^* = (\Theta_1^*)^\top$, and $|\Theta_{1,ij}^*| \leq M$. Similarly,

$$
\begin{aligned}
&\frac{1}{n}||\widehat{\Theta}_0^{(1)} - \Theta_0^*(I_1)||_F \\
\leq&\frac{1}{n}\left[\sum_{(i,j)\in I_1\times I_1, i\neq j}\left(\frac{1}{2}(\widehat{\Gamma}_{0,ij}^{(1)}+\widehat{\Gamma}_{0,ji}^{(1)}) - \Theta_{0,ij}^* - \widehat{\tau}_n^{(1)}\right)^2 + \sum_{(i,j):i\in I_1, j\notin I_1}(\Gamma_{0,ij}^* - \widehat{\tau}_n^{(1)} - \Theta_{0,ij}^*)^2\right]^{1/2} \\
&+ \frac{1}{n}\left(\sum_{i\in I_1}\Theta_{0,ii}^{*2}\right)^{1/2} \\
\leq&\frac{1}{n}\left[\sum_{i\in I_1, j\in[n], i\neq j}(\widetilde{\Gamma}_{0,ij}^{(1)} - \Gamma_{0,ij}^*)^2\right]^{1/2} + |\widehat{\tau}_n^{(1)} - \tau_n| + \sqrt{\frac{M^2}{3n}} \leq 48C_{F,1}\eta_n \ w.p.a.1
\end{aligned}
$$

Then, by the Weyl's inequality, $\max_{k=1,\cdots,K_l}|\widehat{\sigma}_{k,l}^{(1)} - \sigma_{k,l}| \leq 48C_{F,1}\eta_n \ w.p.a.1$ for $l = 0, 1$. Last, noting that $\widehat{V}_l^{(1)}$ consists of the first K_l eigenvectors of $(\frac{1}{n}\widehat{\Theta}_l^{(1)})^\top(\frac{1}{n}\widehat{\Theta}_l^{(1)})$, we have

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

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

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

$$\leq \quad 136 C_{F,2} \eta_n, \tag{32}$$

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

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

$$\widehat{\Lambda}_{n,ij} = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} u_{i,l}^\top (\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} W_{l,ij}). \tag{33}$$

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

$$\tilde{\Lambda}_{n,ij} = \Lambda(\dot{a}_{n,ij}), \tag{34}$$

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,

$$\begin{aligned}
0 &\geq Q_{in,U}^{(0)}(\widehat{u}_{i,0}^{(1)}, \widehat{u}_{i,1}^{(1)}) - Q_{in,U}^{(0)}((\widehat{O}_0^{(1)}) u_{i,0}, (\widehat{O}_1^{(1)}) u_{i,1}) \\
&= \widetilde{Q}_{in}^{(1)}(u_{i,0} + \Delta_{i,0}, u_{i,1} + \Delta_{i,1}) - \widetilde{Q}_{in}^{(1)}(u_{i,0}, u_{i,1}) \\
&\geq \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} \\
&\quad + \frac{1}{n_2} \sum_{j \in I_2, j \neq i} \widehat{\Lambda}_{n,ij} (1 - \widehat{\Lambda}_{n,ij}) \left[\exp(-|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|) + |(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}| - 1 \right] \\
&\geq \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} \\
&\quad + \frac{\underline{c}' \zeta_n}{n_2} \sum_{j \in I_2, j \neq i} \left[\exp(-|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|) + |(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}| - 1 \right] \\
&\geq \frac{-1}{n_2} \sum_{j \in I_2, j \neq i} (Y_{ij} - \widehat{\Lambda}_{n,ij}) (\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu} + \frac{\underline{c}' \zeta_n}{n_2} \sum_{j \in I_2, j \neq i} \left[\frac{((\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu})^2}{2} - \frac{|(\widehat{\phi}_{ij}^{(1)})^\top \Delta_{iu}|^3}{6} \right]
\end{aligned} \tag{35}$$

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 B.3(2), the constant $\underline{c}'$ is defined in Lemma B.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

$$q_{in}=\inf_{\Delta}\frac{\left[\frac{1}{n_2}\sum_{j\in I_2,j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top\Delta)^2\right]^{3/2}}{\frac{1}{n_2}\sum_{j\in I_2,j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top\Delta)^3}\quad\text{and}\quad\delta_{in}=\left[\frac{1}{n_2}\sum_{j\in I_2,j\neq i}((\widehat{\phi}_{ij}^{(1)})^\top\Delta_{iu})^2\right]^{1/2}.\tag{36}$$

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

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

Therefore, by Lemma B.4,

$$F(\Delta_{iu})\geq\min\left(\frac{\underline{c}'\zeta_n\delta_{in}^2}{3},\frac{\underline{c}'\zeta_nq_{in}\delta_{in}}{3}\right)\geq\min\left(\frac{\underline{c}'c_\phi\zeta_n\underline{c}||\Delta_{iu}||^2}{6},\frac{\underline{c}'\zeta_nq_{in}\sqrt{c_\phi}||\Delta_{iu}||}{3\sqrt{2}}\right).\tag{37}$$

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

$$\begin{aligned}II_i\leq&\frac{1}{n_2}\sum_{j\in I_2,j\neq i}\tilde{\Lambda}_{n,ij}(1-\tilde{\Lambda}_{n,ij})\left(|\widehat{\tau}_n-\tau_n|+\sum_{l=0}^{1}\left|u_{i,l}^\top((\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)}-v_{j,l})W_{l,ij}\right|\right)|(\widehat{\phi}_{ij}^{(1)})^\top\Delta_{iu}|\\
\leq&\frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{n_2c_\sigma}\sum_{j\in I_2,j\neq i}\left(|\widehat{\tau}_n-\tau_n|+\sum_{l=0}^{1}\left|u_{i,l}^\top((\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)}-v_{j,l})W_{l,ij}\right|\right)\\
\leq&\frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{c_\sigma}\left[48C_{F,1}\eta_n+c_{II}\sum_{l=0}^{1}\frac{1}{n_2}\sum_{j\in I_2,j\neq i}\left\|(\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)}-v_{j,l}\right\|\right]\\
\leq&\frac{2\overline{c}'M(1+M_W)\zeta_n||\Delta_{iu}||}{c_\sigma}\left[48C_{F,1}\eta_n+c_{II}\sum_{l=0}^{1}\frac{1}{\sqrt{n_2}}\left\|\widehat{V}_l^{(1)}\widehat{O}_l^{(1)}-V_l\right\|_F\right]\end{aligned}$$

$$\leq C_{II}||\Delta_{iu}||\zeta_n\eta_n, \tag{38}$$

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

$$\begin{aligned}\max_{i,j\in I_2, i\neq j}||\widehat{\phi}_{ij}^{(1)}|| \leq &\max_{i,j\in I_2, i\neq j}\left(||\widehat{O}_0^{(1)}\widehat{v}_{j,0}^{(1)}|| + M_W||\widehat{O}_1^{(1)}\widehat{v}_{j,1}^{(1)}||\right)\\ \leq& 2M\sigma_{K_0,0}^{-1} + 2M_W M\sigma_{K_1,1}^{-1} \leq 2M(1+M_W)c_\sigma^{-1},\end{aligned} \tag{39}$$

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 $H_{ij} = (Y_{ij} - \Lambda_{n,ij})\widehat{\phi}_{ij}^{(1)}$. Further note that, for $i \in I_2$, $\{\varepsilon_{ij}\}_{j\in I_2, j\neq i}$ is independent of $\mathcal{F}_n$. Therefore, 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 $k \in [K_0 + K_1]$ and

$$\mathcal{A}_n = \{\max_{j\in I_2}||(\widehat{O}_l^{(1)})^\top\widehat{v}_{j,l}^{(1)}|| \leq 2M\sigma_{K_l,l}^{-1}\} \in \mathcal{F}_n.$$

By Lemma B.3, $\mathbb{P}(\mathcal{A}_n) \to 1$. Under $\mathcal{A}_n$ and Assumption 5, we have

$$\max_{1\leq i,j\leq n}|H_{k,ij}| \leq \left[2M(1+M_W)c_\sigma^{-1} + 1\right]^2(1+\overline{c}) \equiv C_H \tag{40}$$

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 $t > 0$,

$$\mathbb{P}\left(\max_{i\in I_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq n_2 t \middle| \mathcal{F}_n\right)1\{\mathcal{A}_n\} \leq \sum_{i\in I_2} 2\exp\left(-\frac{\frac{n_2^2 t^2}{2}}{C_H\zeta_n n_2 + \frac{C_H t n_2}{3}}\right).$$

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

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

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

$$\begin{aligned}\mathbb{P}\left(\max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq t\right) \leq& \mathbb{P}\left(\max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq t, \mathcal{A}_n\right) + \mathbb{P}(\mathcal{A}_n^c)\\ \leq& \mathbb{E}\left[\mathbb{P}\left(\max_{i\in I_2}\frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \geq t \middle| \mathcal{F}_n\right)1\{\mathcal{A}_n\}\right] + \mathbb{P}(\mathcal{A}_n^c)\end{aligned}$$

$$\leq n^{-1.1} + \mathbb{P}(\mathcal{A}_n^c) \to 0.$$

This means

$$\max_{i\in I_2} I_i \leq \max_{i\in I_2} \frac{1}{n_2}\left|\sum_{j\in I_2, j\neq i} H_{k,ij}\right| \leq 4C_H\sqrt{\frac{\log n\zeta_n}{n}}\ w.p.a.1. \tag{41}$$

Combining (38) and (41), we have

$$|F(\Delta_{iu})| \leq (4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}||. \tag{42}$$

Then, (37) and (42) imply

$$(4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}|| \geq \min\left(\frac{\underline{c}c_\phi\zeta_n|||\Delta_{iu}||^2}{6}, \frac{\underline{c}\sqrt{c_\phi}\zeta_n q_{in}||\Delta_{iu}||}{3\sqrt{2}}\right). \tag{43}$$

On the other hand, we have

$$\liminf_n \min_{i\in I_2} \frac{\underline{c}'\sqrt{c_\phi}\zeta_n q_{in}||\Delta_{iu}||}{3\sqrt{2}} \geq \frac{c_\sigma\underline{c}'c_\phi\zeta_n||\Delta_{iu}||}{24M(1+M_W)} > (4C_H + C_{II})\zeta_n\eta_n||\Delta_{iu}||,$$

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

$$(4C_H + C_{II})(c_F + c_F^2) < \frac{\underline{c}'c_\phi c_\sigma}{24M(1+M_W)}.$$

Therefore, (43) implies

$$||(\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} - u_{i,l}|| \leq ||\Delta_u|| \leq \frac{6(4C_H + C_{II})}{\underline{c}c_\phi}\eta_n \equiv C_1^*\eta_n\ w.p.a.1. \tag{44}$$

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: the regressors $\{\widehat{u}_{i,l}^{(1)}\}_{i\in I_2, l=0,1}$ obtained from the previous step are not independent of the observations $\{Y_{ij}\}$ given the covariates. Thus, the conditional Bernstein inequality argument above cannot be used again. 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 3.1.2. Let

$$\tilde{\Lambda}_{ij}^{(0)}(\nu) = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} \nu_l^\top (\widehat{O}_l^{(1)})^\top \widehat{u}_{i,l}^{(1)} W_{l,ij})$$

and

$$\ell_{ij}^{(0)}(\nu) = Y_{ij}\log(\tilde{\Lambda}_{ij}^{(0)}(\nu)) + (1 - Y_{ij})\log(1 - \tilde{\Lambda}_{ij}^{(0)}(\nu)).$$

Define $\widetilde{Q}^{(0)}_{jn,V}(\nu) = \frac{-1}{n_2}\sum_{j\in I_2, j\neq i}\ell^{(1)}_{ij}(\nu)$. Then

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

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 $\dot{\Lambda}_{n,ij} = \Lambda(\widehat{\tau}_n + \sum_{l=0}^{1} v_{j,l}^\top(\widehat{O}^{(1)}_l)^\top \widehat{u}^{(1)}_{i,l} W_{l,ij})$ 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 $\widehat{\tau}_n + \sum_{l=0}^{1} v_{j,l}^\top(\widehat{O}^{(1)}_l)^\top \widehat{u}^{(1)}_{i,l} W_{l,ij}$. Define

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

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

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

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

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

Therefore, similar to (52), we have

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

As c_F is sufficiently small so that $M_W(C_1^*\eta_n + M)C_1^*(c_F + c_F^2) \leq c_\phi/2$ can be ensured and Assumption 7 holds, we have $\min_{j\in[n]}\lambda_{\min}(\dot{\Psi}_j) \geq c_\phi/2 \ w.p.a.1.$
Let

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

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 B.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 B.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 (38) 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

$$\widetilde{II}_j \leq C_{II}'\zeta_n\eta_n ||\Delta_{jv}||, \tag{45}$$

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

$$\max_{j\in[n]} \left\|\frac{1}{n_2}\sum_{i\in I_2, i\neq j} (Y_{ij} - \Lambda_{n,ij})\,\psi_{ij}\right\| \leq C_H'\sqrt{\frac{\log n\zeta_n}{n}}\ w.p.a.1, \tag{46}$$

where $C_H' = 4(1+\overline{c})^2\left[C_u C_\sigma (M_W+1)+1\right]^4$. 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 and

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

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, w.p.a.1,

$$[4(1+M_W)\overline{c}C_1^* + C'_{II}]\eta_n\zeta_n||\Delta_{jv}|| \geq \min\left(\frac{\underline{c}' c_\phi \zeta_n \underline{c} ||\Delta_{jv}||^2}{6}, \frac{\underline{c}' \zeta_n q_{jn} \sqrt{c_\phi} ||\Delta_{jv}||}{3\sqrt{2}} \right). \tag{47}$$

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

$$q_{jn} \geq \inf_{\Delta} \sqrt{\frac{\frac{c_\sigma^2}{n} \sum_{i \in I_2, i \neq j} ((\dot{\psi}_{ij})^\top \Delta)^2}{16(1+M_W)^2 M^2 ||\Delta||^2}} \geq \frac{c_\sigma \sqrt{c_\phi/2}}{4(1+M_W)M} > 0.$$

In addition, because c_F can be made sufficiently small to ensure $(4(1+M_W)\overline{c}C_1^* + C'_{II})(c_F + c_F^2) < \frac{c_\sigma \underline{c}' \sqrt{c_\phi}}{24(1+M_W)M}$, we have

$$\begin{aligned}(4(1+M_W)\overline{c}C_1^* + C'_{II})\eta_n\zeta_n||\Delta_{jv}|| &\leq (4(1+M_W)\overline{c}C_1^* + C'_{II})(c_F + c_F^2)\zeta_n||\Delta_{jv}|| \\ &< \frac{c_\sigma \underline{c}' c_\phi \zeta_n ||\Delta_{jv}||}{24(1+M_W)M} \leq \frac{\underline{c}' \sqrt{c_\phi} \zeta_n q_{jn} ||\Delta_{jv}||}{3\sqrt{2}}.\end{aligned}$$

Then, (47) implies

$$||\Delta_{jv}|| \leq \frac{6(4(1+M_W)\overline{c}C_1^* + C'_{II})}{\underline{c}' c_\phi \underline{c}} \eta_n \equiv C_{0,v}\eta_n \; w.p.a.1. \tag{48}$$

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

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

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

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

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 5.1

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

$$
\begin{aligned}
||\overline{v}_j - v_j^*|| &\leq \left\| \frac{\dot{v}_{j,1}^{(H,1)}}{||\dot{v}_{j,1}^{(H,1)}||} - \frac{\widehat{O}_1^{(1)} v_{j,1}}{||\widehat{O}_1^{(1)} v_{j,1}||} \right\| + \left\| \frac{\dot{v}_{j,1}^{(H,2)}}{||\dot{v}_{j,1}^{(H,2)}||} - \frac{\widehat{O}_1^{(2)} v_{j,1}}{||\widehat{O}_1^{(2)} v_{j,1}||} \right\| \\
&= \left\| \frac{(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}}{||(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\| + \left\| \frac{(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}}{||(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\| \\
&\leq \frac{2 \left\| (\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)} - v_{j,1} \right\|}{||(\widehat{O}_1^{(1)})^\top \dot{v}_{j,1}^{(H,1)}||} + \frac{2 \left\| (\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)} - v_{j,1} \right\|}{||(\widehat{O}_1^{(2)})^\top \dot{v}_{j,1}^{(H,2)}||} \\
&\leq \frac{4C_{H,v}\eta_n}{||v_{j,1}|| - C_{H,v}\eta_n} \leq 5C_1^{-1/2} C_{H,v} \eta_n, \qquad (49)
\end{aligned}
$$

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 $z_i \neq z_j$,

$$
\begin{aligned}
||v_j^* - v_i^*|| &= \left[\left\| \frac{\widehat{O}_1^{(1)} v_{i,1}}{||\widehat{O}_1^{(1)} v_{i,1}||} - \frac{\widehat{O}_1^{(1)} v_{j,1}}{||\widehat{O}_1^{(1)} v_{j,1}||} \right\|^2 + \left\| \frac{\widehat{O}_1^{(2)} v_{i,1}}{||\widehat{O}_1^{(2)} v_{i,1}||} - \frac{\widehat{O}_1^{(2)} v_{j,1}}{||\widehat{O}_1^{(2)} v_{j,1}||} \right\|^2 \right]^{1/2} \qquad (50) \\
&= \left[\left\| \frac{v_{i,1}}{||v_{i,1}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\|^2 + \left\| \frac{v_{i,1}}{||v_{i,1}||} - \frac{v_{j,1}}{||v_{j,1}||} \right\|^2 \right]^{1/2} = 2. \qquad (51)
\end{aligned}
$$

Given (49) and (50), the result of Theorem 5.1 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

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

when c_F is sufficiently small. ■

Appendix B. Some Technical Lemmas

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

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

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 $\{\varepsilon_{ij}\}_{(i,j) \in \mathcal{I}_1 \times [n]}$, which is an i.i.d. sequence of logistic random variables. In addition, $\{\varepsilon_{ij}\}_{(i,j) \in \mathcal{I}_1 \times [n]}$ is independent of $\mathcal{C}$,

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

and $|\overline{\Upsilon}_{l,ij}| \leq M_W$. Then, by Bandeira and van Handel (2016, Corollary 3.12 and Remark 3.13), there exists a universal constant $\tilde{c}$ such that

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

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

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

and thus,

$$||\overline{\Upsilon}_l||_{op} \leq 3M_W(\sqrt{2\bar{c}n\zeta_n} + \sqrt{\tilde{c}\log(n)}) \leq C_\Upsilon M_W(\sqrt{n\zeta_n} + \sqrt{\log(n)}) \; w.p.a.1. \blacksquare$$

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

Proof. First, suppose $M \geq \log(2)$. 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 leads to the desired result.
Next, suppose $M < \log(2)$. Then, we have

$$\exp(-t) + t - 1 \geq \frac{t^2}{2} - \frac{t^3}{6} \geq \frac{(3-\log(2))t^2}{6} \geq \frac{t^2}{4\log(2)}.$$

This concludes the proof. ■

Lemma B.3 *Suppose that the Assumptions in Theorem 4.1 hold. Then, w.p.a.1,*

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

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

where $\widehat{\Lambda}_{n,ij}$ *and* $\tilde{\Lambda}_{n,ij}$ *are defined in* (33) *and* (34)*, 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 second inequality holds because $||\widehat{\mathcal{U}}_l^{(1)}||_{op} \leq 1$, 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,

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

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 B.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,

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

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

$$\begin{aligned} ||\widehat{\Phi}_i^{(1)} - \Phi_i(I_2)|| \leq & \frac{4M}{n_2} \sum_{l=0}^{1} \sum_{j \in I_2} \sigma_{K_l,l}^{-1} ||(\widehat{O}_l^{(1)})^\top \widehat{v}_{j,l}^{(1)} - v_{j,l}|| \\ \leq & 4M \sum_{l=0}^{1} \sigma_{K_l,l}^{-1} n_2^{-1/2} ||\widehat{V}_l \widehat{O}_l^{(1)} - V_l||_F \\ \leq & 1088\sqrt{2} M C_{F,2} c_\sigma^{-1} \eta_n, \end{aligned} \tag{52}$$

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 $1088\sqrt{2}MC_{F,2}c_\sigma^{-1}(c_F + c_F^2) \leq c_\phi/2$, we have, w.p.a.1,

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

Lemma B.5 *Let q_{in} be defined in (36). Suppose that Assumptions 1–6 hold. Then,*

$$\liminf_n \min_{i \in I_2} q_{in} \geq \frac{\sqrt{c_\phi/2}\, c_\sigma}{4M(1 + M_W)} > 0 \; w.p.a.1,$$

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

Proof. Note

$$q_{in} \geq \inf_\Delta \sqrt{\frac{c_\sigma^2 \frac{1}{n_2} \sum_{j \in I_2, j \neq i} ((\widehat{\phi}_{ij}^{(1)})^\top \Delta)^2}{16M^2(1 + M_W)^2 ||\Delta||^2}} \geq \frac{c_\sigma \liminf_n \min_{i \in I_2} \lambda_{\min}(\widehat{\Phi}_i^{(1)})}{4M(1 + M_W)} \geq \frac{\sqrt{c_\phi/2}\, c_\sigma}{4M(1 + M_W)} > 0,$$

where the first inequality is due to Lemma B.3(1) and the second inequality is due to Lemma B.4. $\blacksquare$

Appendix C. Proof of Theorem 6.1

Theorem 6.1 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. $\blacksquare$

Appendix D. Proof of Theorem 6.2

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,

$$Q_n\left(\text{vec}(B^*) + u(n^2\zeta_n)^{-1/2}\right) - Q_n(\text{vec}(B^*)) + \upsilon_n^\top u - \frac{u^\top \mathcal{H} u}{2} = o_p(1), \tag{53}$$

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

$$\hat{u} - \mathcal{H}^{-1}\upsilon_n = o_p(1). \tag{54}$$

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

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

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 $\text{Var}(\upsilon_n) < \infty$, implying that $\upsilon_n = O_p(1)$. Therefore, we have established (53), and thus (54).
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 $\epsilon > 0$, there exists n_0 sufficiently large so that for all $n \ge n_0$ and $k \in [\mathcal{K}]$,

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

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

Appendix E. Algorithm for the Nuclear Norm Regularization

We apply the optimization algorithm proposed in Cabral et al. (2013) to obtain the nuclear norm penalized estimator given in (12). For any given $r_l \geq K_l$ and $r_l \leq n$, Γ_l can be written as $\Gamma_l = U_l V_l^\top$, 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:

$$Q_n^{(1)}(\Gamma) + \frac{\lambda_n^{(1)}}{2} \sum_{l=0}^{p} \gamma_l(||U_l||_F^2 + ||V_l||_F^2), \tag{55}$$

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 (12) with 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 (55) with $r_l \geq K_l^*$ is a solution of (11). Next we apply the Augmented Lagrange Multiplier (ALM) method given in Cabral et al. (2013) to solve (55). The augmented Lagrangian function of (55) 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,

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

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

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Emmanuel Abbe and Colin Sandon. 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*, pages 670–688. IEEE, 2015.

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

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

Hossein Alidaee, Eric Auerbach, and Michael P Leung. Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*, 2020.

Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

Tomohiro Ando and Jushan Bai. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191, 2016.

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

Jushan Bai and Serena Ng. Rank regularized estimation of approximate factor models. *Journal of Econometrics*, 212(1):78–96, 2019.

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

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Alexandre Belloni, Mingli Chen, and Oscar Hernan Madrid Padilla. High dimensional latent panel quantile regression with an application to asset pricing. *arXiv preprint arXiv:1912.02151*, 2019.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50): 21068–21073, 2009.

Peter J Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

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

Ricardo Cabral, Fernando De la Torre, P. Joao Costeira, and Bernardino Alexandre. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. *IEEE International Conference on Computer Vision*, pages 2488–2495, 2013.

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435, 2011.

Jia Chen. Estimating latent group structure in time-varying coefficient panel data models. *The Econometrics Journal*, 22(3):223–240, 2019.

Xu Cheng, Frank Schorfheide, and Peng Shao. Clustering for multi-dimensional heterogeneity. Technical report, University of Pennsylvania, 2019.

Victor Chernozhukov, Christian Hansen, Yuan Liao, and Yinchu Zhu. Inference for heterogeneous effects using low-rank estimations. *arXiv preprint arXiv:1812.08089*, 2020.

David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

Andreas Dzemski and Ryo Okui. Confidence set for group membership. *Available at SSRN 3133878*, 2018.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

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

Bryan Graham and Aureo de Paula. *The Econometric Analysis of Network Data*. Academic Press, 2019.

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

Bryan S Graham. Network data. Technical report, National Bureau of Economic Research, 2019.

Bryan S Graham. Sparse network asymptotics for logistic regression. Technical report, National Bureau of Economic Research, 2020.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Wenxin Huang, Sainan Jin, and Liangjun Su. Identifying latent grouped patterns in cointegrated panels. *Econometric Theory*, 36(3):410–456, 2020.

Wenxin Huang, Sainan Jin, Peter CB Phillips, and Liangjun Su. Nonstationary panel models with latent group structures and cross-section dependence. *Journal of Econometrics*, 2021.

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

Koen Jochmans. Modified-likelihood estimation of fixed-effect models for dyadic data. 2019.

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

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

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

Jean-Benoist Leger. 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*, 2016.

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

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

Ruiqi Liu, Zuofeng Shang, Yonghui Zhang, and Qiankun Zhou. Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, 215(2):574–590, 2020.

Xun Lu and Liangjun Su. Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8(3):729–760, 2017.

Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2013.

Vince Lyzinski, Daniel Sussman, Minh Tang, Avanti Athreya, and Carey Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.

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

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

Hyungsik Roger Moon and Martin Weidner. Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987*, 2018.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, In proc. of STOC15, 2014.

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

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250, 2020.

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

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3120–3128. Curran Associates, Inc., 2013.

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

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

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Sandipan Roy, Yves Atchade, and George Michailidis. 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*, 28(3):609–619, 2019.

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

Srijan Sengupta and Yuguo Chen. Spectral clustering in heterogeneous networks. *Statistica Sinica*, 25(3):1081–1106, 2015.

Liangjun Su and Gaosheng Ju. Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573, 2018.

Liangjun Su, Zhentao Shi, and Peter CB Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016.

Liangjun Su, Xia Wang, and Sainan Jin. Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 37(2):334–349, 2019.

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

Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

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

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

Michael Vogt and Oliver Linton. Multiscale clustering of nonparametric regression curves. *Journal of Econometrics*, 216(1):305–325, 2020.

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

Wuyi Wang and Liangjun Su. Identifying latent group structures in nonlinear panels. *Journal of Econometrics*, 220(2), 2021.

Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

Haolei Weng and Yang Feng. Community detection with nodal information. *arXiv preprint arXiv:1610.09735*, 2016.

Jinfeng Xu, Mu Yue, and Wenyang Zhang. A new multilevel modelling approach for clustered survival data. *Econometric Theory*, 36(2):707–750, 2020.

Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, pages 1–12, 2020.

Ting Yan and Jinfeng Xu. A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika*, 100(2):519–524, 2013.

Ting Yan, Binyan Jiang, Stephen E Fienberg, and Chenlei Leng. Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 114(526): 857–868, 2019.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014.

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.