SPECTRAL INFERENCE FOR LARGE STOCHASTIC BLOCKMODELS WITH NODAL COVARIATES

ANGELO MELE, LINGXIN HAO, JOSHUA CAPE, AND CAREY E. PRIEBE

ABSTRACT. In many applications of network analysis, it is important to distinguish between observed and unobserved factors affecting network structure. To this end, we develop spectral estimators for both unobserved blocks and the effect of covariates in stochastic blockmodels. On the theoretical side, we establish asymptotic normality of our estimators for the subsequent purpose of performing inference. On the applied side, we show that computing our estimator is much faster than standard variational expectation–maximization algorithms and scales well for large networks. The results in this paper provide a foundation to estimate the effect of observed covariates as well as unobserved latent community structure on the probability of link formation in networks.

1. INTRODUCTION

1.1. **Motivation.** The analysis and modeling of network data has important applications in sociology, economics, public health, computer science, neuroscience, and marketing, among other areas. Often, both observed and unobserved factors contribute to the global structure of networks and the processes that generate them. For example, in social networks, factors including gender, race, and personality affect the likelihood that two people interact. In such data, race and gender are typically observed, while personality is typically unobserved. It is therefore crucial to develop ways to disentangle the effect of observed and unobserved variables on link formation.

The stochastic blockmodel (SBM) is a popular network model and workhorse in the literature on community detection (Abbe, 2018; Nowicki and Snijders, 2001; Airoldi et al., 2008). In a K-block SBM, each node belongs to one of K unobserved blocks (communities); conditional on the block assignments, links form independently as Bernoulli trials with probabilities that depends on the community memberships. Many existing works use SBMs to estimate or approximate unobserved block structure in networks. In contrast, applications involving SBMs that incorporate observed nodal attributes as covariates are comparatively few (Sweet, 2015; Choi et al., 2011; Roy et al., 2019). A possible reason is that estimation for stochastic blockmodels is computational burdensome, and including covariates in the specification imposes significant additional challenges to modelling, estimation, and inference. Exact maximum likelihood estimation is infeasible, therefore most estimation strategies rely

This version: August 18, 2019. First version: February 25, 2019. Contact: angelo.mele@jhu.edu; hao@jhu.edu; jrcape@umich.edu; cep@jhu.edu. We are grateful to Cong Mu and Jipeng Zhang for excellent research assistance. We thank Eric Auerbach, Stephane Bonhomme, and Eleonora Patacchini for comments and suggestions. Funding from the Institute of Data Intensive Engineering and Science (IDIES) at Johns Hopkins University is gratefully acknowledged. Joshua Cape also gratefully acknowledges support from NSF grant DMS-1902755.

on approximations based on expectation–maximization algorithms and variational methods (Airoldi et al., 2008; Daudin et al., 2008; Bickel et al., 2013; Latouche et al., 2012; Vu et al., 2013). However, these algorithms may converge slowly to the (approximate) solution and become impractical for networks with thousands of nodes.[1]

1.2. **Overview.** The goal of our paper is to develop an inference procedure for SBMs with observed nodal covariates that is computationally feasible and scales well to large networks. At a high level, our main strategy consists of formulating this goal as an inference problem in the context of the generalized random dot product graph (GRDPG) model (Athreya et al., 2018; Tang and Priebe, 2018; Tang et al., 2017; Rubin-Delanchy et al., 2018). In a GRDPG, each node is characterized by an unobserved latent position (vector), and each pair of nodes link with probability determined via a (possibly indefinite) inner product of the pair's latent positions; crucially, any SBM can be reformulated as a GRDPG where the latent positions are fixed within blocks. To address our computational goals, we turn to the use of spectral methods, which, in addition to enjoying theoretical guarantees (Tang et al., 2017), have been shown to be successful both in terms of feasibility and scalability in related settings (Zheng et al., 2015, 2016, 2017).

We provide several contributions to the literature on statistical network analysis. First, we present a GRDPG model framework that incorporates the effect of observed covariates on linkage probabilities. Second, we develop a spectral estimator for inference in stochastic blockmodels with covariates, adapting the spectral estimators developed for our new class of GRDPG models. Crucially, we obtain a new central limit theorem for the spectral estimator of the covariates' effect. Our estimator is asymptotically normal as long as the parameter(s) for covariate effect can be written as sufficiently well-behaved functions of the SBM block-specific probabilities. We provide explicit formulas for bias and variance properties of the estimator, and we show that the estimator is computationally fast, scaling well for large networks. Our method provides a statistical and algorithmic foundation for inference in a broad class of models for large network data, including networks that are relatively sparse in the sense that their average degree scales sub-linearly with network size.

Our exposition focuses on SBMs with a single binary (or discrete) observed covariate, though we emphasize that the theoretical and computational properties set forth in this work extend to settings involving multiple discrete covariates. Asymptotic normality continues to hold as long as the estimator for the effect of the covariate(s) can be expressed as suitably well-behaved a function of the SBM probabilities. The case involving continuous covariates is more complicated and an active area of contemporaneous research. Current progress on this front is being facilitated by recently investigated Latent Structure Models (LSM) (Athreya et al., forthcoming) and related ideas.

The development of our estimator depends crucially on several observations. First, a K-block stochastic blockmodel with one binary covariate can be equivalently reformulated as a (different) $2K$-block stochastic blockmodel. Second, as discussed previously, a stochastic blockmodel can be viewed as a generalized random dot product graph whose latent positions are fixed within blocks (Athreya et al., 2018; Tang and Priebe, 2018; Tang et al., 2017). The

[1]Recent advances use further approximations and parallelization to improve computational efficiency (Roy et al., 2019; Vu et al., 2013). We do not pursue such extensions in this paper.

behavior of our spectral estimation method is tied to the asymptotic behavior of spectral estimators for SBM block probability matrix entries recently studied in Tang et al. (2017). Our asymptotic analysis provides explicit formulas for standard errors and establishes the existence of a bias term; however, this bias vanishes at rate proportional to the size of the network.[2]

The theoretical machinery used to perform inference extends methods developed for the analysis of latent positions network models (Athreya et al., 2018; Tang et al., 2017). In particular, we use Adjacency Spectral Embedding (ASE) for random graphs to embed the network in a low-dimensional space and to recover the latent positions of the nodes. Our method is motivated by the (verifiable) intuition that the adjacency matrix can be viewed as a (mild) perturbation of the probability matrix that generates the network data, and thus, that the eigenstructure of the adjacency matrix resembles that of the edge probability matrix (Tang and Priebe, 2018; Athreya et al., 2018). In particular, spectrally decomposing the adjacency matrix provides accurate information about the structure of sufficiently large networks (Tang et al., 2017).

In addition to providing statistical guarantees, one of the many advantages of our method is the speed of computation, obtained without sacrificing estimation accuracy. In our simulations (see Section 4) we compare our approach to the variational EM (VEM) algorithm (Daudin et al., 2008; Bickel et al., 2013), as implemented in the `blockmodels` package in `R`. Even for the simplest case of a stochastic blockmodel without covariates, our spectral method is faster by several orders of magnitude. For example, in a network with $n = 5000$ nodes and $K = 2$ blocks, we can estimate the model in few seconds using our spectral method, while it takes almost 10 minutes to estimate the model using the variational EM algorithm. When we add a binary covariate, our estimator converges in less than 30 seconds, while in contrast it takes almost 10 hours when using a parallelized version of the VEM algorithm in `blockmodels`. Our methods are implemented in the R package `grdpg` available at `https://github.com/meleangelo/grdpg`.

We also apply our method to the study of Facebook friendship data using the Facebook 100 dataset initially collected and analyzed in Traud et al. (2012). These data contain the network of friendships and node information about 100 universities in the United States in the year 2005. We estimate a stochastic blockmodel for the network of Rice University, consisting of approximately 4000 nodes, using information on gender, dorm, and year of graduation of the users (see also Roy et al. (2019)). We find evidence of homophily by gender, as suggested by the positive effect of gender on the probability of linking.

Another way to use our models and methods is to correct for the endogeneity of the network in empirical models of network effects (Shalizi and McFowland, 2018; Goldsmith-Pinkham and Imbens, 2013; Boucher and Fortin, 2016; **?**; **?**). Currently most of these studies rely on an auxiliary model of network formation to capture unobserved heterogeneity that affects the outcome. Our model and computational method allow the researcher to perform this type of correction for large network data.

[2]Since we have a closed-form expression for the bias term, in principle we can naively correct for it in estimation, using a plug-in estimate. In our simulations we find that the bias term is usually so small that the correction is not necessary, at least for networks with a few thousand nodes. On the other hand, the bias is demonstrably substantial in the empirical application to Facebook data in Section 4.

2. Background and methodology

2.1. **Stochastic blockmodels and generalized random dot product graphs.** In a K-block *Stochastic Blockmodel (SBM)* nodes are randomly assigned to one of K blocks; conditional on the blocks, nodes form links independently. A K-block SBM is characterized by the $K \times K$ matrix of probabilities $\boldsymbol{\theta} \in [0,1]^{K\times K}$, where the entry $\boldsymbol{\theta}_{k\ell}$ is the probability of a link occurring between nodes in blocks k and ℓ. The random variables comprising $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ describe the assignments of each node to a block, and they are i.i.d., such that the probability that node i belongs to block k is $P(\tau_i = k) = \pi_k$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Conditional on the assignment to blocks $\boldsymbol{\tau}$, the probability that nodes i and j have a link is $\boldsymbol{P}_{ij} = \boldsymbol{\theta}_{\tau_i\tau_j}$. We use the $n \times n$ adjacency matrix $\boldsymbol{A}$ to describe the network, conditioning on the unobserved blocks. According to the SBM the entries of the adjacency matrix are generated as

$$\boldsymbol{A}_{ij}|\tau_i, \tau_j, \overset{ind}{\sim} Bernoulli(\boldsymbol{P}_{ij}), \tag{1}$$

and we write $(\boldsymbol{A}, \boldsymbol{\tau}) \sim SBM(\boldsymbol{\theta}, \boldsymbol{\pi})$ to denote the adjacency matrix drawn from a K-block SBM with probability matrix $\boldsymbol{\theta}$ and block assignment probabilities $\boldsymbol{\pi}$.

Conditioning on $\boldsymbol{\tau}$, the likelihood of the SBM with K blocks is

$$P(\boldsymbol{A}|\boldsymbol{\tau}) \;=\; \prod_{i\le j} \boldsymbol{P}_{ij}^{\boldsymbol{A}_{ij}}(1-\boldsymbol{P}_{ij})^{1-\boldsymbol{A}_{ij}} = \prod_{i\le j} \boldsymbol{\theta}_{\tau_i\tau_j}^{\boldsymbol{A}_{ij}}(1-\boldsymbol{\theta}_{\tau_i\tau_j})^{1-\boldsymbol{A}_{ij}}. \tag{2}$$

The *Generalized Random Dot Product Graph* (GRPDG) model is an alternative model for network formation with conditionally independent links. In a GRDPG, each node i is characterized by a d-dimensional vector (i.e., an *unobserved* latent position) $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X}_d \subseteq \mathbb{R}^d$. The latent positions are i.i.d. draws from a distribution F with support $\mathcal{X}_d$, that is $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} F$. Let $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^T$ denote the matrix stacking all unobserved $\boldsymbol{X}_i$ vectors by row.

Let $d_1 \ge 1$ and $d_2 \ge 0$ be integers, and define $d = d_1 + d_2$. Let $\boldsymbol{I}_{d_1,d_2}$ be a $d \times d$ diagonal matrix containing 1's in d_1 diagonal entries and -1 in the remaining d_2 diagonal entries. For a *GRDPG with signature* (d_1, d_2), the entries of the adjacency matrix $\boldsymbol{A}_{ij}$ are specified to be independent, after conditioning on the latent positions $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, namely

$$\boldsymbol{A}_{ij}|\boldsymbol{X}_i, \boldsymbol{X}_j, \overset{ind}{\sim} Bernoulli(\boldsymbol{P}_{ij}) \tag{3}$$

with link probability given by

$$\boldsymbol{P}_{ij} = \boldsymbol{X}_i^T \boldsymbol{I}_{d_1,d_2} \boldsymbol{X}_j. \tag{4}$$

For this setting, we write $(\boldsymbol{X}, \boldsymbol{A}) \sim GRDPG_{d_1,d_2}(F)$.[3]

2.2. **The relationship between SBMs and GRDPGs.** A remarkable property of GRDPGs is that they encompass or approximate any conditionally independent network model. In particular, *any SBM can be represented as a GRDPG* with latent positions fixed within blocks. That is, the K blocks are represented by a fixed location, so that each $\boldsymbol{X}_i$ can only take values $\boldsymbol{\nu} = [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_K]$. Two nodes i and j belong to the same block k if $\boldsymbol{X}_i = \boldsymbol{X}_j = \boldsymbol{\nu}_k$. The

[3] It must be noted that the support $\mathcal{X}_d$ of F, is a subset of $\mathbb{R}^d$ such that $\boldsymbol{x}^T \boldsymbol{I}_{d_1,d_2} \boldsymbol{y} \in [0,1]$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}_d$.

random variables $\boldsymbol{\tau}$ are such that $\tau_1, \ldots, \tau_n \overset{iid}{\sim} Multinomial(1; \pi_1, \ldots, \pi_K)$, with $\boldsymbol{\pi} \in (0,1)^K$ and $\sum_{k=1}^K \pi_k = 1$.[4] The GRDPG corresponding to model $(\boldsymbol{A}, \boldsymbol{\tau}) \sim SBM(\boldsymbol{\theta}, \boldsymbol{\pi})$ can be obtained by an eigendecomposition of the matrix $\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$ and by defining $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_K$ as the rows of $\boldsymbol{U}|\boldsymbol{\Sigma}|^{1/2}$. The distribution F is $F = \sum_{k=1}^K \pi_k \delta_{\boldsymbol{\nu}_k}$, where δ is the Dirac-delta; importantly, d is the rank of the block-probabilities matrix $\boldsymbol{\theta}$, and d_1, d_2 are the number of positive and negative eigenvalues of matrix $\boldsymbol{\theta}$, respectively.

This paper utilizes the spectral methods for inference developed for GRDPGs to estimate SBMs (Athreya et al., 2018; Tang et al., 2017). The same relationship between SBMs and GRDPG holds for *known* link functions and $\boldsymbol{\theta}_{\tau_i\tau_j} = h\left(\boldsymbol{B}_{\tau_i\tau_j}\right)$, where h is a known function that maps to $[0,1]$ and $\boldsymbol{B}$ is a $K \times K$ matrix of real numbers.[5] For example, h could be the logistic function or the cumulative density function of the Gaussian distribution. Our stochastic blockmodel would have adjacency matrix $\boldsymbol{A}$ with elements

$$\boldsymbol{A}_{ij}|\tau_i, \tau_j \overset{ind}{\sim} Bernoulli\left(h(\boldsymbol{B}_{\tau_i\tau_j})\right). \tag{6}$$

The stochastic blockmodel can be extended to include the effect of observed covariates (Choi et al., 2011; Sweet, 2015; Roy et al., 2019). Such models allow researchers to disentangle the effect of observed and unobserved nodal heterogeneity on the probability of linking. In particular, in social science, such models are used to estimate to what extent the network exhibits homophily or heterophily. Let node i be characterized by an r-dimensional vector of *observed* covariates $\boldsymbol{Z}_i = (Z_i^{(1)}, \ldots, Z_i^{(r)}) \in \mathcal{Z} \subseteq \mathbb{R}^r$ and let the stochastic blockmodel be

$$\boldsymbol{A}_{ij}|\tau_i, \tau_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j \overset{ind}{\sim} Bernoulli\left(h(\boldsymbol{B}_{\tau_i\tau_j} + f(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta}))\right), \tag{7}$$

where f is a known function, $\boldsymbol{\beta}$ is a vector of parameters, and where we allow $\boldsymbol{Z}_i$ to (possibly) depend on the latent blocks.

In this paper *we will focus on the case of a single binary (or discrete) covariate $\boldsymbol{Z}_i$'s, and we will assume that the function f is an indicator variables that indicates whether i and j's covariates have the same value*, i.e.,

$$f(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta}) = \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}. \tag{8}$$

Here β can be interpreted in terms of homophily. Namely, if $\beta > 0$, the probability of a link between i and j is higher when their observables $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ are the same, and the network displays homophily in the observable variable. Viceversa, when $\beta < 0$, we have heterophily. The extension to multiple discrete covariates has similar properties and will be discussed further below.

[4] Alternatively, we can think of a K-block stochastic blockmodel as a network where the $\boldsymbol{X}_i$'s are drawn from a mixture of degenerate distributions with mass centered at $\boldsymbol{\nu}$, i.e.,

$$\boldsymbol{X}_i \sim \pi_1 \delta_{\boldsymbol{\nu}_1} + \pi_2 \delta_{\boldsymbol{\nu}_2} + \cdots + \pi_K \delta_{\boldsymbol{\nu}_K}. \tag{5}$$

[5] If h is unknown we cannot in general expect to be able to accurately estimate the latent positions. See Tang et al. (2013).

Our goal is to develop a *general spectral method of inference* for the parameter β and for $\boldsymbol{B}_{\tau_i\tau_j}$ in the following stochastic blockmodel with a discrete nodal covariate:

$$\boldsymbol{A}_{ij}|\tau_i,\tau_j,\boldsymbol{Z}_i,\boldsymbol{Z}_j \overset{ind}{\sim} Bernoulli\left(\boldsymbol{P}_{ij}\right), \tag{9}$$

$$\boldsymbol{P}_{ij} = h\left(\boldsymbol{B}_{\tau_i\tau_j} + \beta \mathbf{1}_{\{\boldsymbol{Z}_i=\boldsymbol{Z}_j\}}\right). \tag{10}$$

We further wish to disentangle the effect of observed and unobserved heterogeneity on link probabilities. To achieve this, we need to extend results from previous work on GRDPGs and SBMs (Athreya et al., 2018; Tang and Priebe, 2018; Tang et al., 2017). In the following subsections we review some of the spectral methods we use in the paper, and we provide an example that highlights the core of our method.

2.3. **Spectral methods and spectral embeddings.** Estimation of SBMs for large networks, with or without observed covariates, is computationally challenging. The exact MLE problem is intractable because of the high-dimensional combinatorial problem of considering all possible partitions of the nodes in blocks (Bickel et al., 2013). Approximate methods are available, based on variational approximations (Daudin et al., 2008; Airoldi et al., 2008; Wainwright and Jordan, 2008); however, even these methods are too computationally burdensome for large networks.

We make use of spectral methods, which have been shown in the literature to scale well with network size. Our spectral method embeds the network in a low(er) dimensional space, thus reducing the dimensionality of the problem, while maintaining the geometric properties of the model. In particular we use the *Adjacency Spectral Embedding* (ASE) to estimate the latent positions of the GRDPG (Athreya et al., 2018). In this sense, our method can be considered a dimension-reduction tool that decreases the complexity of the data by reducing the dimensionality of the space. The intuition about the spectral method is that if $\boldsymbol{P}$ is a low-rank matrix, then we can think of the adjacency matrix $\boldsymbol{A}$ as a perturbation of $\boldsymbol{P}$, that is $\boldsymbol{A}_{ij} = \boldsymbol{P}_{ij} + \boldsymbol{E}_{ij}$, where $\boldsymbol{E}_{ij}$ is a matrix of independent stochastic perturbations.[6] If $\boldsymbol{A}$ and $\boldsymbol{P}$ are *close enough*, namely if $\boldsymbol{E}$ is *small enough*, then the leading eigenvalues and eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{P}$ will be similar (Tang et al., 2017). As a consequence, the spectral decomposition of $\boldsymbol{A}$ will provide an estimate of the latent structure of the network, that is, the latent positions $\boldsymbol{X}$.

Consider first the case without observed covariates. Let $\boldsymbol{P}$ be positive semidefinite and let h be the identity function $h(u) = u$. In this setting, we only have latent positions $\boldsymbol{X}$, that are unobserved. If we were able to observe $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{X}^T$, estimation of $\boldsymbol{X}$ would be straightforward. Furthermore, we could use spectral embeddings for $\boldsymbol{P}$ by exploiting the fact that $\boldsymbol{P}$ is positive semidefinite of rank d and has spectral decomposition $\boldsymbol{P} = \boldsymbol{U}_{\boldsymbol{P}}\boldsymbol{S}_{\boldsymbol{P}}\boldsymbol{U}_{\boldsymbol{P}}^T$, where $\boldsymbol{S}_{\boldsymbol{P}}$ is a diagonal matrix containing the largest d eigenvalues (in absolute value) of $\boldsymbol{P}$ and $\boldsymbol{U}_{\boldsymbol{P}}$ is the matrix with the corresponding eigenvectors. This implies that a good estimate for $\boldsymbol{X}$ is $\widehat{\boldsymbol{X}} = \boldsymbol{U}_{\boldsymbol{P}}|\boldsymbol{S}_{\boldsymbol{P}}|^{1/2}$, where $|\cdot|$ denotes entrywise absolute values. The estimation problem arises because *we only observe* $\boldsymbol{A}$*, a perturbed version of* $\boldsymbol{P}$. The Adjacency Spectral Embedding of $\boldsymbol{A}$ into $\mathbb{R}^d$ is then $\widehat{\boldsymbol{X}} = \boldsymbol{U}_{\boldsymbol{A}}|\boldsymbol{S}_{\boldsymbol{A}}|^{1/2}$ where $\boldsymbol{S}_{\boldsymbol{A}}$ is a diagonal matrix containing

[6] In the Bernoulli case, $\boldsymbol{E}_{ij}$ is a shifted Bernoulli variable, with values $\boldsymbol{E}_{ij} = 1 - \boldsymbol{P}_{ij}$ with probability $\boldsymbol{P}_{ij}$ and $\boldsymbol{E}_{ij} = \boldsymbol{P}_{ij}$ with probability $1 - \boldsymbol{P}_{ij}$.

the largest d eigenvalues of $\boldsymbol{A}$ in absolute value and $\boldsymbol{U_A}$ is the matrix with the corresponding eigenvectors.

In our asymptotic results for the above setup, we use the fact that ASE estimates of latent positions $\boldsymbol{X}$ asymptotically achieve perfect clustering (moreover, are asymptotically normal) and can be identified up to multiplication by an orthogonal matrix (Athreya et al., 2018; Tang et al., 2017). This implies that asymptotically the blocks are recovered exactly (up to relabeling). The same logic and results hold for non-positive definite matrices $\boldsymbol{P}$, allowing us to study more general stochastic blockmodels (Rubin-Delanchy et al., 2018).

2.4. **Overview of the method in a 2-block SBM with one binary covariate.** To illustrate the methodology and to develop intuition, we focus on the special case of a $K = 2$ block stochastic blockmodel with a single discrete covariate and with latent positions on the unit interval $[0, 1]$, yielding $d_1 = 1, d_2 = 0$, and $r = 1$, where $Z_i \in \{0, 1\}$ is a binary variable (e.g., male/female, white/nonwhite, rich/poor, etc.) and the function $f(Z_i, Z_j; \beta) = \beta \mathbf{1}_{\{Z_i = Z_j\}}$ is an indicator for the equality of the covariates for i and j, weighted by parameter β. The main advantage of this approach is that we can illustrate the geometry of the method in a low-dimensional space. In our simple example, the matrix $\boldsymbol{B}$ is given by

$$\boldsymbol{B} = \begin{array}{c} \\ block_1 \\ block_2 \end{array} \begin{array}{c} \begin{array}{cc} block_1 & block_2 \end{array} \\ \begin{pmatrix} p^2 & pq \\ pq & q^2 \end{pmatrix} \end{array} \tag{11}$$

where $p, q \in [0, 1]$. We can conveniently re-write the matrix $\boldsymbol{B}$ as a dot-product of vector $\boldsymbol{\nu} = [p \quad q]^T$, with $p, q \in [0, 1]$, that is $\boldsymbol{B} = \boldsymbol{\nu}\boldsymbol{\nu}^T$, so that the SBM can be re-written as a random dot-product graph model with $\boldsymbol{X}_i = p$ if i is in block 1, $\boldsymbol{X}_i = q$ if i is in block 2. The probability of linking can then be written as

$$\boldsymbol{P}_{ij} = h\left(\boldsymbol{X}_i^T \boldsymbol{X}_j + \beta \mathbf{1}_{\{Z_i = Z_j\}}\right). \tag{12}$$

For ease of exposition the network blocks have the same probability, so $(\pi_1, \pi_2) = (0.5, 0.5)$ and each community contains half males ($Z_i = 1$) and half females ($Z_i = 0$). However, we note that our algorithm and the theoretical results are valid when we allow the blocks to be of different size, and the observed covariates to be correlated with the unobserved blocks.

The model specified via (12) corresponds to a 4-block stochastic blockmodel. Indeed, we have 2 unobserved blocks, that are split in two additional blocks by the observed binary variable. Therefore, the final result is a 4-block SBM. More generally, if there are K latent blocks and one binary covariates, we will have a $\tilde{K} = 2K$-block SBM.

The possible values of $\boldsymbol{X}_i^T \boldsymbol{X}_j$ are $\{p^2, pq, q^2\}$. Therefore the 4-block model can be completely characterized by the 4×4 matrix

$$\boldsymbol{B}_Z = \begin{array}{c} \\ male_1 \\ female_1 \\ male_2 \\ female_2 \end{array} \begin{array}{c} \begin{array}{cccc} male_1 & female_1 & male_2 & female_2 \end{array} \\ \begin{pmatrix} p^2+\beta & p^2 & pq+\beta & pq \\ p^2 & p^2+\beta & pq & pq+\beta \\ pq+\beta & pq & q^2+\beta & q^2 \\ pq & pq+\beta & q^2 & q^2+\beta \end{pmatrix} \end{array}. \tag{13}$$

The value $h(\boldsymbol{B}_{Z,11}) = h(p^2 + \beta)$ is the probability that two males in block 1 form a link; on the other hand, $h(\boldsymbol{B}_{Z,12}) = h(p^2)$ is the probability that a male and a female in block 1 form

a link; $h(\boldsymbol{B}_{Z,31}) = h(pq + \beta)$ is the probability that two males, one in block 1 and one in block 2, form a link; and so on.

The above observations imply that, for this four block SBM, there exists a corresponding GRDPG with link probability matrix

$$\boldsymbol{P} = \boldsymbol{Y}\boldsymbol{I}_{d_1,d_2}\boldsymbol{Y}^T, \tag{14}$$

for some $n \times d$ matrix of latent positions $\boldsymbol{Y}$ with $d_1 \geq 1$, $d_2 \geq 0$, and $d = d_1 + d_2$.

To estimate the parameter β and the latent positions p and q we use the following algorithmic approach.

1. We compute an eigendecomposition of the adjacency matrix $\boldsymbol{A}$, letting $\boldsymbol{S}_A$ denote the matrix whose diagonal contains the largest $\hat{d}$ eigenvalues of $\boldsymbol{A}$ in absolute value and $\boldsymbol{U}_A$ denote the matrix whose columns are corresponding unit norm eigenvectors.[7] The Adjacency Spectral Embedding (ASE) of $\boldsymbol{A}$ gives an estimate of the latent positions of the 4-block model as
$$\widehat{\boldsymbol{Y}} = \boldsymbol{U}_A|\boldsymbol{S}_A|^{1/2}, \tag{15}$$
where $|\cdot|$ indicates the absolute value (entrywise).
2. We use $\widehat{\boldsymbol{Y}}$ to estimate $\boldsymbol{P}$ as
$$\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{Y}}\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{Y}}^T, \tag{16}$$
where $\hat{d} := \hat{d}_1 + \hat{d}_2$ is the number of largest eigenvalues (in magnitude) of $\boldsymbol{A}$ beyond a prescribed threshold, and $\hat{d}_1$ and $\hat{d}_2$ are the number of these eigenvalues that are positive and negative, respectively.
3. We use a clustering procedure to assign each row of $\widehat{\boldsymbol{Y}}$ to one of $\tilde{K} = 4$ blocks. We use a Gaussian Mixture Modeling approach (GMM) and estimate the center of the clusters $\widehat{\boldsymbol{\mu}} = [\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\mu}}_3, \widehat{\boldsymbol{\mu}}_4]$, that is the means of the Gaussians from the GMM.
4. We compute an estimate for $\boldsymbol{\theta}_Z$ as
$$\begin{aligned}\widehat{\boldsymbol{\theta}}_Z &= \widehat{\boldsymbol{\mu}}^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}} \\ &= \begin{array}{c} \\ male_1 \\ female_1 \\ male_2 \\ female_2 \end{array} \begin{array}{c} \begin{array}{cccc} male_1 & female_1 & male_2 & female_2 \end{array} \\ \begin{pmatrix} \widehat{\boldsymbol{\mu}}_1^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_1 & \widehat{\boldsymbol{\mu}}_1^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_2 & \widehat{\boldsymbol{\mu}}_1^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_3 & \widehat{\boldsymbol{\mu}}_1^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_4 \\ \widehat{\boldsymbol{\mu}}_2^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_1 & \widehat{\boldsymbol{\mu}}_2^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_2 & \widehat{\boldsymbol{\mu}}_2^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_3 & \widehat{\boldsymbol{\mu}}_2^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_4 \\ \widehat{\boldsymbol{\mu}}_3^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_1 & \widehat{\boldsymbol{\mu}}_3^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_2 & \widehat{\boldsymbol{\mu}}_3^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_3 & \widehat{\boldsymbol{\mu}}_3^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_4 \\ \widehat{\boldsymbol{\mu}}_4^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_1 & \widehat{\boldsymbol{\mu}}_4^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_2 & \widehat{\boldsymbol{\mu}}_4^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_3 & \widehat{\boldsymbol{\mu}}_4^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_4 \end{pmatrix} \end{array}. \end{aligned} \tag{17}$$
By comparing the matrix $\widehat{\boldsymbol{\theta}}_Z$ and the population matrix $\boldsymbol{\theta}_Z$, we can assign each of the 4 blocks to the original 2 blocks. In fact, we know that the diagonal terms of $\widehat{\boldsymbol{\theta}}_Z$ are estimates of $h(p^2 + \beta)$ if the nodes belong to block 1, or $h(q^2 + \beta)$ if nodes belong to block 2. This observation shows that we can group the 4 entries of the diagonal, by checking which values are close. In our case, we will group $\widehat{\boldsymbol{\mu}}_1^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_2$ in one block, and $\widehat{\boldsymbol{\mu}}_3^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_3$ and $\widehat{\boldsymbol{\mu}}_4^T\boldsymbol{I}_{\hat{d}_1,\hat{d}_2}\widehat{\boldsymbol{\mu}}_4$ in another block. Therefore, blocks 1

[7] In principle, the optimal value for $\hat{d} = rank(\boldsymbol{B}_Z)$; however we do not observe $\boldsymbol{B}_Z$, therefore we estimate $\hat{d}$ by profile likelihood methods (Zhu and Ghodsi, 2006).

and 2 in the 4-block model are assigned to the original latent block 1, while blocks 3 and 4 are assigned to original latent block 2.

5. We estimate β from the entries of the matrix $\widehat{\boldsymbol{B}}_Z = h^{-1}(\widehat{\boldsymbol{\theta}_Z})$, where the inverse function h^{-1} is applied element-wise. For example we know that $h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_1)$ is an estimate of $p^2+\beta$ or $q^2+\beta$ (because of the invariance of the model to relabeling of the blocks). Without loss of generality, assume that $h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_1)$ is an estimate of $p^2+\beta$; therefore, the entry $h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_2)$ is an estimate of p^2. Therefore, the point estimate of β is then
$$\widehat{\beta} = h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_1) - h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_2). \tag{18}$$

6. The latent positions p and q can be estimated from the matrix $\widehat{\boldsymbol{B}}_Z$, by using the submatrix
$$\begin{array}{c} \\ male_1 \\ male_2 \end{array} \begin{array}{c} \begin{array}{cc} female_1 & female_2 \end{array} \\ \begin{pmatrix} h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_2) & h^{-1}(\widehat{\boldsymbol{\mu}}_1^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_4) \\ h^{-1}(\widehat{\boldsymbol{\mu}}_3^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_2) & h^{-1}(\widehat{\boldsymbol{\mu}}_3^T \boldsymbol{I}_{\hat{d}_1,\hat{d}_2} \widehat{\boldsymbol{\mu}}_4) \end{pmatrix} \end{array} = \begin{pmatrix} \widehat{p^2} & \widehat{pq} \\ \widehat{pq} & \widehat{q^2} \end{pmatrix}. \tag{19}$$
The spectral embedding of this matrix provides estimates for the latent positions $\hat{p}$ and $\hat{q}$, that are identified up to an orthogonal transformation.

In practice, we can estimate β from multiple entries of the matrix $\boldsymbol{B}_Z$, for example $\beta = \boldsymbol{B}_{Z,11} - \boldsymbol{B}_{Z,12} = \boldsymbol{B}_{Z,33} - \boldsymbol{B}_{Z,34}$, and weight each estimate by the size of the blocks. This could improve the estimate, since some blocks are larger than others, so delivering more precise estimates. Our code implements this idea, which is more practical for empirical applications.

3. Asymptotic theory

In this section, we derive a central limit theorem for the spectral estimator of β. For ease of exposition, we focus on the case of a single binary observed covariate and scalar β, though our method works for other specifications in which the effect of the observed covariates β can be written as a function of the stochastic blockmodel's probability matrix $\boldsymbol{\theta}_Z$. Extensions to multiple binary or discrete observed covariates are straightforward.

We desire to estimate a stochastic blockmodel with observed covariates, where
$$\tau_i \overset{iid}{\sim} Multinomial(1; \pi_1, \ldots, \pi_K), \tag{20}$$
$$\boldsymbol{Z}_i|\tau_i \overset{ind}{\sim} Bernoulli(b_{\tau_i}), \tag{21}$$
$$\boldsymbol{A}_{ij}|\tau_i, \tau_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j \overset{ind}{\sim} Bernoulli(\boldsymbol{P}_{ij}), \tag{22}$$
$$\boldsymbol{P}_{ij} = h\left(\boldsymbol{B}_{\tau_i\tau_j} + \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}\right). \tag{23}$$

We assume that the observed covariates are binary and can depend on the block assignment, that is $\boldsymbol{Z}_i|\tau_i \overset{ind}{\sim} Bernoulli(b_{\tau_i})$, where $b_{\tau_i} = P(\boldsymbol{Z}_i = 1|\tau_i)$. Our asymptotic results are easily extended to the case of discrete observed covariates with three or more possible outcomes.

As explained above in the simple example, our strategy consists of rewriting the SBM as a GRDPG. First, notice that the matrix $\boldsymbol{B}$ can be written as $\boldsymbol{B}_{\tau_i\tau_j} = \boldsymbol{X}_i^T \boldsymbol{X}_j$, where $\boldsymbol{X}_i$ is a $d \times 1$ vector of latent positions that has K possible values $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K$. In practice,

$\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K)$ are the centers of the K blocks $\boldsymbol{X}$, such that i and j belong to *unobserved* block k when $\boldsymbol{X}_i = \boldsymbol{X}_j = \boldsymbol{\nu}_k$. Let $\boldsymbol{\tau}$ be the function that assigns nodes to unobserved blocks; then $\tau_i = k$ if $\boldsymbol{X}_i = \boldsymbol{\nu}_k$. We can thus rewrite the stochastic blockmodel above as a random dot product graph with observed covariates as follows:

$$\begin{aligned}
\boldsymbol{X}_i &\overset{iid}{\sim} \pi_1\delta_{\nu_1} + \pi_2\delta_{\nu_2} + \cdots + \pi_K\delta_{\nu_K}, && (24)\\
\boldsymbol{Z}_i|\boldsymbol{X}_i &\overset{ind}{\sim} Bernoulli(b_{\tau_i}), && (25)\\
\boldsymbol{A}_{ij}|\boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j &\overset{ind}{\sim} Bernoulli(\boldsymbol{P}_{ij}), && (26)\\
\boldsymbol{P}_{ij} &= h\left(\boldsymbol{X}_i^T\boldsymbol{X}_j + \beta\mathbf{1}_{\{\boldsymbol{Z}_i=\boldsymbol{Z}_j\}}\right). && (27)
\end{aligned}$$

We first notice that both models are stochastic blockmodels with $\tilde{K} = 2K$ blocks, because the indicator variable $\mathbf{1}_{\{\boldsymbol{Z}_i=\boldsymbol{Z}_j\}}$ splits each unobserved block in two blocks. The probabilities of belonging to a block k for this $\widetilde{K}$-block SBM are denoted as $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{\widetilde{K}}) = (\pi_1 \cdot b_1, \pi_1 \cdot (1-b_1), \pi_2 \cdot b_2, \pi_2 \cdot (1-b_2), ..., \pi_K \cdot b_K, \pi_K \cdot (1-b_K))$; and the functions that assign nodes to blocks are $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, such that $\xi_i = 1$ if $\tau_i = 1$ and $\boldsymbol{Z}_i = 0$; $\xi_i = 2$ if $\tau_i = 1$ and $\boldsymbol{Z}_i = 1$; $\xi_i = 3$ if $\tau_i = 2$ and $\boldsymbol{Z}_i = 0$; $\xi_i = 4$ if $\tau_i = 2$ and $\boldsymbol{Z}_i = 1$; and so on.

So we have a stochastic blockmodel $(\boldsymbol{A}, \boldsymbol{\xi}, \boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z, \boldsymbol{\eta})$ with $\widetilde{K} \times \widetilde{K}$ matrix of probabilities $\boldsymbol{\theta}_Z$

$$\boldsymbol{\theta}_Z = \begin{array}{c} \\ \tau=1;Z=0 \\ \tau=1;Z=1 \\ \tau=2;Z=0 \\ \tau=2;Z=1 \\ \vdots \\ \tau=K;Z=0 \\ \tau=K;Z=1 \end{array} \begin{array}{c} \begin{array}{cccccccc} \tau=1;Z=0 & \tau=1;Z=1 & \tau=2;Z=0 & \tau=2;Z=1 & \cdots & \tau=K;Z=0 & \tau=K;Z=1 \end{array} \\ \begin{pmatrix} h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2+\beta\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2\right) & \cdots & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K+\beta\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K\right) \\ h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2+\beta\right) & \cdots & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K\right) & h\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K+\beta\right) \\ h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2+\beta\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2\right) & \cdots & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K+\beta\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K\right) \\ h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2+\beta\right) & \cdots & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K\right) & h\left(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K+\beta\right) \\ \vdots & \vdots & \vdots & & \ddots & & \\ h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2+\beta\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2\right) & \cdots & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K+\beta\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K\right) \\ h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1+\beta\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2+\beta\right) & \cdots & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K\right) & h\left(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K+\beta\right) \end{pmatrix} \end{array} \quad (28)$$

The stochastic blockmodel characterized by matrix $\boldsymbol{\theta}_Z$ can be re-formulated as a GRDPG. Indeed, consider the eigendecomposition of matrix $\boldsymbol{\theta}_Z = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$, and define $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{\widetilde{K}}]$ as the rows of $\boldsymbol{U}|\boldsymbol{\Sigma}|^{1/2}$; then let $F = \sum_{k=1}^{\widetilde{K}} \eta_k\delta_{\boldsymbol{\mu}_k}$, where δ is the Dirac-delta; and d_1 and d_2 are the number of positive and negative eigenvalues of $\boldsymbol{\theta}_Z$, respectively. So the Generalized Random Dot Product Graph model $(\boldsymbol{Y}, \boldsymbol{A}) \sim GRDPG_{d_1,d_2}(F)$ corresponding to our stochastic blockmodel $(\boldsymbol{A}, \boldsymbol{\xi}, \boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z, \boldsymbol{\eta})$ is given by

$$\begin{aligned}
\boldsymbol{Y}_i &\overset{iid}{\sim} \eta_1\delta_{\boldsymbol{\mu}_1} + \cdots + \eta_{\tilde{K}}\delta_{\boldsymbol{\mu}_{\tilde{K}}} && (29)\\
\boldsymbol{A}_{ij}|\boldsymbol{Y}_i, \boldsymbol{Y}_j &\overset{ind}{\sim} Bernoulli(\boldsymbol{Y}_i^T\boldsymbol{I}_{d_1,d_2}\boldsymbol{Y}_j) && (30)
\end{aligned}$$

where $d_1 + d_2 = \widetilde{d} = rank(\boldsymbol{\theta}_Z)$ and $\boldsymbol{Y}$ is the $n \times \widetilde{d}$ vector of latent positions with centers $\boldsymbol{\mu}$.

We can now extend asymptotic results for estimation of RDPGs in Athreya et al. (2018); Tang et al. (2017) to estimate block assignments and the effect of the covariates (see Rubin-Delanchy et al. (2018) for the corresponding generalization to GRDPGs).

3.1. **Main theoretical result.** Because the functions $\boldsymbol{\tau}$ that describe the assignments to blocks are unknown, the $\widetilde{K}$ SBM model assignment functions $\boldsymbol{\xi}$ are also unknown. Applying the Adjacency Spectral Embedding procedure, we recover an estimate $\widehat{\boldsymbol{\xi}}$.

We prove asymptotic normality for the parameter β, exploiting the fact that β can be written as a function of the SBM probabilities, that is

$$\beta = h^{-1}\left(\boldsymbol{\theta}_{Z,11}\right) - h^{-1}\left(\boldsymbol{\theta}_{Z,12}\right) = h^{-1}\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 + \beta\right) - h^{-1}\left(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1\right). \tag{31}$$

If the blocks were known at the onset, we could use the estimator $\widehat{\beta} = h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,11}) - h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,12})$. However, all that we have access to is the estimate $\widehat{\boldsymbol{\xi}}$, so it is crucial that this estimate be consistent. For RDPGs this is indeed the case, as one can prove that the latent blocks are recovered up to an orthogonal transformation matrix in the large n limit (Lemma 4 in Tang et al. (2017)). Therefore we can recover the parameter β up to relabeling of the blocks. This is summarized in the following theorem.

THEOREM 1. ***Central limit theorem for*** β
Let $\boldsymbol{\tau}$ be unknown and K known. Let $\widehat{\boldsymbol{\tau}} : [n] \to [K]$ be the function that assigns nodes to clusters, estimated using GMM or K-means clustering on the rows of $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{U}}|\widehat{\boldsymbol{S}}|^{1/2}$. Let function g be defined as the inverse of h, that is $g(\cdot) = h^{-1}(\cdot)$, with first derivative $g'(\cdot)$. Let $g'(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 + \beta) \neq 0$ and $g'(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2) \neq 0$. Then there exists a sequence of permutations $\phi \equiv \phi_n$ on $[K]$ such that the estimator $\widehat{\beta} = h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,\phi(1)\phi(1)}) - h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,\phi(1)\phi(2)})$ is asymptotically normal, that is

$$n\left(\widehat{\beta} - \beta - \frac{\widehat{\psi}_\beta}{n}\right) \xrightarrow{d} N(0, \widehat{\sigma}_\beta^2) \tag{32}$$

as $n \to \infty$. The values $\widehat{\psi}_\beta$ and $\widehat{\sigma}_\beta^2$ are computed in the proof.

Proof. See Appendix. □

This implies that we can recover the parameter β, and our estimator is asymptotically unbiased. In Appendix we provide the expression for the bias term.

3.2. **Sparsity.** The previous theoretical result implicitly assumes a dense network. However, many social and economic networks of interest in applications display some degree of sparsity. This is an empirical regularity that social scientists have observed in many settings, as most people do not form many links. Economists rationalize sparsity with the fact that people have constraints on time to spend with their friends.

To take this feature of the data into account, while still allowing estimation in massive networks, we multiply the probability $\boldsymbol{P}_{ij}$ by a scalar ρ_n that governs the sparsity of the network, that is, the probability of a link between nodes i and j becomes

$$\boldsymbol{P}_{ij} = \rho_n h\left(\boldsymbol{X}_i^T\boldsymbol{X}_j + \beta\mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}\right). \tag{33}$$

Our previous result in Theorem 1 applies to dense networks; that is when $\rho_n \to c$ where $c \in (0, 1]$ is a constant. For simplicity and without loss of generality, in Theorem 1 we have assumed $c = 1$.

In this section we consider formally the case of $\rho_n \to 0$ as $n \to \infty$. We have to limit the rate of convergence for ρ_n, because the network could become too sparse, not allowing estimation. We will describe this regime a *semi-sparse*, because we will allow $\rho_n \to 0$ but $n\rho_n = \omega(\sqrt{n})$, that is the average degree of the network grows sub-linearly in n.[8] The intuition for this restriction is that too much sparsity makes links "too rare" and therefore spectral estimation and inference are impeded by having too few observations.

THEOREM 2. ***Central limit theorem for sparse networks***
Let model (23) include a sparsity coefficient ρ_n

$$\boldsymbol{P}_{ij} = \rho_n h\left(\boldsymbol{X}_i^T \boldsymbol{X}_j + \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}\right) \tag{34}$$

such that $\rho_n \to 0$ *and* $n\rho_n = \omega(\sqrt{n})$ *as* $n \to \infty$. *Let* $\hat{\boldsymbol{\tau}}$ *be assignment of each node to a block, estimated using ASE and GMM (or K-means) clustering. Then there exists a sequence of permutations* $\phi \equiv \phi_n$ *on* $[K]$ *such that the estimator* $\widehat{\beta} = h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,\phi(1)\phi(1)}) - h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,\phi(1)\phi(2)})$ *is asymptotically normal, that is*

$$n\rho_n^{1/2}\left(\hat{\beta} - \beta - \frac{\ddot{\psi}_\beta}{n\rho_n}\right) \xrightarrow{d} N\left(0, \ddot{\sigma}_\beta^2\right) \tag{35}$$

where $\ddot{\psi}_\beta$ *and* $\ddot{\sigma}_\beta^2$ *are computed in the proof.*

Proof. See Appendix. □

Theorem 2 says that as long as the network is not too sparse, the estimator of β will be asymptotically normal. In practical estimation exercises, we can use a rule-of-thumb procedure and check whether the number of links of the network is proportional to $\sqrt{n}$, the square root of the network size.

3.3. **Multiple observed covariates.** The asymptotic results hold for discrete observed covariates and more general models, as long as the effect of the observed covariates on the probability of linking can be written as a function of the block probabilities. Let the observed variables $\boldsymbol{Z}_i = [\boldsymbol{Z}_i^{(1)}, \boldsymbol{Z}_i^{(2)}]$ include two covariates. For simplicity, we consider the case of binary variables, and we assume $\boldsymbol{Z}_i^{(1)} \overset{ind}{\sim} Bernoulli(b_{\tau_i}^{(1)})$ and $\boldsymbol{Z}_i^{(2)} \overset{ind}{\sim} Bernoulli(b_{\tau_i}^{(2)})$. The results still hold for discrete variables. The model is

$$\boldsymbol{A}_{ij} | \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j \overset{ind}{\sim} Bernoulli\left(\boldsymbol{P}_{ij}\right), \tag{36}$$

$$\boldsymbol{P}_{ij} = h\left(\boldsymbol{X}_i^T \boldsymbol{X}_j + \beta_1 \mathbf{1}_{\{\boldsymbol{Z}_i^{(1)} = \boldsymbol{Z}_j^{(1)}\}} + \beta_2 \mathbf{1}_{\{\boldsymbol{Z}_i^{(2)} = \boldsymbol{Z}_j^{(2)}\}}\right). \tag{37}$$

This stochastic blockmodel has $\tilde{K} = 4K$ blocks, $(\boldsymbol{A}, \boldsymbol{\xi}, \boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z, \boldsymbol{\eta})$ with $\tilde{K} \times \tilde{K}$ matrix of probabilities $\boldsymbol{\theta}_Z$ given by

[8] The notation $n\rho_n = \omega(\sqrt{n})$ means that for any real constant $a > 0$ there exists an $n_0 \geq 1$ such that $\rho_n > a/\sqrt{n} \geq 0$ for every integer $n \geq n_0$.

$$\boldsymbol{\theta}_Z = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} & \dots & \boldsymbol{W}_{1\tilde{K}} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} & \dots & \boldsymbol{W}_{2\tilde{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{W}_{\tilde{K}1} & \boldsymbol{W}_{\tilde{K}2} & \dots & \boldsymbol{W}_{\tilde{K}\tilde{K}} \end{bmatrix} \tag{38}$$

where each matrix $\boldsymbol{W}_{k\ell}$ is given by

	$\tau = \ell$ $Z^{(1)} = 0$ $Z^{(2)} = 0$	$\tau = \ell$ $Z^{(1)} = 1$ $Z^{(2)} = 0$	$\tau = \ell$ $Z^{(1)} = 0$ $Z^{(2)} = 1$	$\tau = \ell$ $Z^{(1)} = 1$ $Z^{(2)} = 1$
$\tau = k$ $Z^{(1)} = 0$ $Z^{(2)} = 0$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1 + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell)$
$\tau = k$ $Z^{(1)} = 1$ $Z^{(2)} = 0$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1 + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1)$
$\tau = k$ $Z^{(1)} = 0$ $Z^{(2)} = 1$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1 + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_2)$
$\tau = k$ $Z^{(1)} = 1$ $Z^{(2)} = 1$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_2)$	$h(\boldsymbol{\nu}_k^T\boldsymbol{\nu}_\ell + \beta_1 + \beta_2)$

(39)

The intuition is the same as the model with one covariate. The blocks can be inferred by clustering the diagonal elements of matrix $\boldsymbol{\theta}_Z$, and the parameters β_1 and β_2 are functions of the $\boldsymbol{\theta}_Z$ entries, namely

$$\beta_1 = h^{-1}\left(\boldsymbol{\theta}_{Z,11}\right) - h^{-1}\left(\boldsymbol{\theta}_{Z,12}\right); \quad \beta_2 = h^{-1}\left(\boldsymbol{\theta}_{Z,11}\right) - h^{-1}\left(\boldsymbol{\theta}_{Z,13}\right). \tag{40}$$

As such, the main characterization of the central limit theorem holds in this case with minimal modifications.

4. Simulation and empirical results

4.1. **Comparison with Variational EM.** We compare our spectral methods to a standard algorithm used in the literature, the variational EM algorithm, as implemented in the R package `blockmodels`. Our methods are implemented in the package `grdpg`, available on Github at `https://github.com/meleangelo/grdpg`. We also note that the variational EM algorithm uses parallelization to increase computational efficiency, while our method is implemented without any parallelization, for networks with thousands of nodes.

4.1.1. *Example 1 (No covariates).* In our first example, we do not include any covariates and we assume h is the identity function, so that the probabilities are $\boldsymbol{P}_{ij} = \boldsymbol{X}_i^T \boldsymbol{X}_j$. We simulate networks with $n = 2000, 5000, 10000$, and 20000 nodes, with latent space dimension $d = 1$. In Table 1 we report the results for $K = 2$, with block centers $[p, q] = [0.1, 0.7]$, and matrix of probabilities

$$\boldsymbol{\theta} = \begin{bmatrix} 0.01 & 0.07 \\ 0.07 & 0.49 \end{bmatrix}. \tag{41}$$

For simplicity, we assume that blocks are equally likely, that is $(\pi_1, \pi_2) = (0.5, .0.5)$. To evaluate the performance of the algorithms, we compare clustering accuracy and computational time. The assignment of nodes to the correct block is summarized by the Adjusted Rand Index (ARI), and the computational time is given by the CPU time in seconds. Our point estimates are shown in Table 1, and below we report the estimated block probabilities for $n = 2000$.

$$\widehat{\boldsymbol{\theta}}_{VEM} = \begin{bmatrix} 0.01002 & 0.07012 \\ 0.07012 & 0.49086 \end{bmatrix}, \qquad \widehat{\boldsymbol{\theta}}_{GRDPG} = \begin{bmatrix} 0.00998 & 0.07001 \\ 0.07001 & 0.49091 \end{bmatrix}. \tag{42}$$

The values $\hat{p}, \hat{q}$ shown in in Table 1 are obtained by singular value decomposition of the estimated probability matrix (and rotation). We notice that the VEM and GRDPG estimators produce similar point estimates and very precise clustering of the nodes, as indicated by the ARI. However, our GRDPG estimator converges much faster than the VEM. For networks with $n = 10000$ nodes, our method provides estimates in approximately 30 seconds, while the VEM takes more than one hour to converge to the final approximation. When $n = 20000$ and $n = 30000$, our GRDPG approach converges in about 2 minutes and less than 7 minutes, respectively, while the VEM is impractical.

Here, a crucial choice is the number of dimensions for the spectral embedding. In our simulation we know that the rank of the matrix $\boldsymbol{\theta}$ is 1, therefore this is the optimal dimension (see Athreya et al. (2018)). We choose $\hat{d}$ by profile likelihood methods as in Zhu and Ghodsi (2006).[9] The clustering of the latent positions in blocks is performed using the MCLUST method implemented in the package `Mclust` in R (Fraley and Raftery, 1999).

In Table 2 we report results from the same model with $K = 5$ and latent positions $\boldsymbol{\nu} = (0.1, 0.3, 0.5, 0.7, 0.9)$. The results are comparable to the previous table, our estimator scales very well with the size of the network, while obtaining the same point estimates of the VEM algorithm. In this example, the difference in scaling for the two estimators is more pronounced. In particular, going from $K = 2$ to $K = 5$ blocks does not increase the computational burden too much for the GRDPG-based estimator.

4.1.2. *Example 2 (logit link and binary covariate).* We consider a model with a binary nodal covariate, $\boldsymbol{Z}_i \sim Bernoulli(0.5)$ and link probabilities

$$\log\left(\frac{\boldsymbol{P}_{ij}}{1 - \boldsymbol{P}_{ij}}\right) = \boldsymbol{X}_i^T \boldsymbol{X}_j + \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}. \tag{43}$$

[9]The screeplot, not shown, displays a huge step down in the (absolute) value of the eigenvalues of the adjacency matrix at the largest eigenvalue, which suggests that 1 dimension is sufficient to approximate the structure of the adjacency matrix.

In this example we use $[p, q] = [-1.5, 1]$ and $\beta = 1.5$, thus the matrix $\boldsymbol{\theta}$ is

$$logit(\boldsymbol{\theta}) = \begin{bmatrix} 2.25 & -1.5 \\ -1.5 & 1 \end{bmatrix} \tag{44}$$

while the full matrix $\boldsymbol{\theta}_Z$ that includes the effect of covariates is

$$logit(\boldsymbol{\theta}_Z) = \begin{bmatrix} 3.75 & 2.25 & 0.00 & -1.50 \\ 2.25 & 3.75 & -1.50 & 0.00 \\ 0.00 & -1.50 & 2.50 & 1.00 \\ -1.50 & 0.00 & 1.00 & 2.50 \end{bmatrix}. \tag{45}$$

We choose $\hat{d}$ by profile likelihood (Zhu and Ghodsi, 2006). In Figure 1 we show the screeplots. In the upper-left, we display the screeplot for the adjacency matrix, which suggest to use $\hat{d} = 4$. We note that the fourth largest eigenvalue is negative, and the GRDPG model takes this into account. In the center-left plot, we show the screeplot of the adjacency matrix *net of the effect of the covariates*,which we use to estimate the dimension of the unobserved latent positions $\boldsymbol{X}$, which suggests a dimension $\hat{d} = 1$.

The point estimates for $\boldsymbol{\theta}_Z$ (up to a permutation of the block labels) when the network has $n = 2000$ nodes are respectively

$$logit(\widehat{\boldsymbol{\theta}}_{Z,VEM}) = \begin{bmatrix} 3.7443 & 2.2410 & -0.00367 & -1.5069 \\ 2.2410 & 3.7443 & -1.5069 & -0.0036 \\ -0.00367 & -1.5069 & 2.5013 & 0.9980 \\ -1.5069 & -0.0036 & 0.9980 & 2.5013 \end{bmatrix}, \tag{46}$$

$$\widehat{\boldsymbol{B}}_Z = logit(\widehat{\boldsymbol{\theta}}_{Z,GRDPG}) = \begin{bmatrix} 3.7762 & 2.2336 & -0.0062 & -1.5095 \\ 2.2336 & 3.7821 & -1.5007 & -0.0042 \\ -0.0062 & -1.5007 & 2.4979 & 0.9985 \\ -1.5095 & -0.0042 & 0.9985 & 2.5045 \end{bmatrix}. \tag{47}$$

According to our procedure, there are several ways to obtain an estimate of β. From matrix (47), we group rows 1 and 2 in one block, and rows 3 and 4 in another block by clustering the diagonal entries. We can get an estimate of β as $\widehat{\boldsymbol{B}}_{Z,11} - \widehat{\boldsymbol{B}}_{Z,12}$ or $\widehat{\boldsymbol{B}}_{Z,22} - \widehat{\boldsymbol{B}}_{Z,21}$ or $\widehat{\boldsymbol{B}}_{Z,33} - \widehat{\boldsymbol{B}}_{Z,34}$, etc. We know by our theorem that each of these estimators is asymptotically normal. Instead of choosing which entries to use to estimate β, we pool all possible estimates, weighting them by the proportion of observations that are assigned to each block. For example, the estimate $\widehat{\boldsymbol{B}}_{Z,11} - \widehat{\boldsymbol{B}}_{Z,12}$ is weighted by the proportion of links in the network that are used to estimate it.

The point estimates for β reported in Table 3 are $\widehat{\beta}_{GRDPG} = 1.51201$ and $\widehat{\beta}_{VEM} = 1.50335$. The estimated latent positions are $\hat{p} = -1.49712$ and $\hat{q} = 1.00067$ for the VEM; and $\hat{p} = -1.49454$ and $\hat{q} = 0.99926$ for the GRDPG estimator. However, it takes almost 2 hours to obtain the VEM results, while it only takes 7 seconds with our estimator. The left plots in Figure 1 show the latent positions of the GRDPG (including the effect of covariates) estimated by ASE. We plot the first coordinate against each of the other three. In the second and third plot from the top, we can notice that the latent positions nicely cluster into 4 blocks, as our theory predicts. In the bottom-left plot in Figure 1 we display the estimated

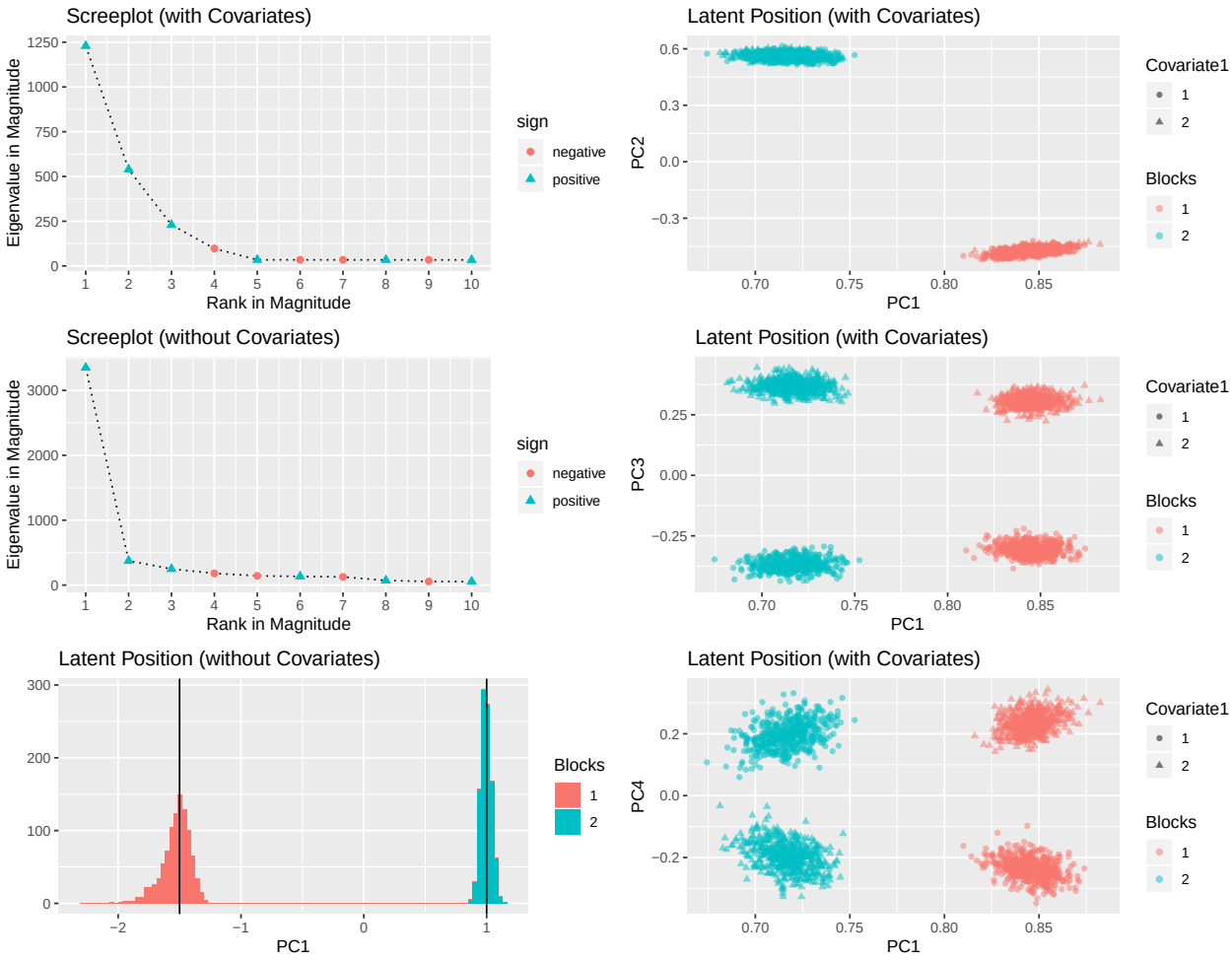

FIGURE 1. Screeplots (upper left and center left), Estimated latent positions $\widehat{\boldsymbol{Y}}$ (right, only 2 dimensions out of 4 per plot) and estimated latent positions $\hat{\boldsymbol{X}}$, that is $\hat{p}$ and $\hat{q}$ in Example 2 (bottom left, up to orthogonal transformation) for $n = 2000$.

latent positions *net of the covariate effect*, showing how the estimated latent positions $\widehat{\boldsymbol{X}}$ cluster around the true values p and q (the black vertical lines).

As explained above, a central advantage of our approach is computational speed. Indeed, in Table 3 we show that when we increase the size of the network to $n = 5000$, the estimated parameters are essentially the same for VEM and GRPDG. However, the GRDPG estimator take less than 30 seconds to converge; the VEM estimate takes almost 10 hours.

4.1.3. *Example 3 (logit link, binary covariate and $d = 2$).* In Table 4 we show estimates for models with latent positions $\boldsymbol{\nu}_1 = (-1.5, -1.0)$ and $\boldsymbol{\nu}_2 = (1.0, 0.5)$. For the simulations in the first 3 rows we set $\beta = 1.5$. It is quite remarkable that the computational time does not increase much, with respect to the case of $d = 1$.

TABLE 1. Point Estimates and CPU time for example 1 ($K = 2$)

Estimator	n	K	p	$\hat{p}$	q	$\hat{q}$	CPU Time (s)	ARI
GRDPG	2000	2	0.1	0.09993	0.7	0.70065	1.513	1
VEM	2000	2	0.1	0.10008	0.7	0.70061	39.679	1
GRDPG	5000	2	0.1	0.10004	0.7	0.69977	8.548	1
VEM	5000	2	0.1	0.10008	0.7	0.69975	593.203	1
GRDPG	10000	2	0.1	0.09994	0.7	0.69988	32.169	1
VEM	10000	2	0.1	0.09996	0.7	0.69987	4171.218	1
GRDPG	20000	2	0.1	0.09998	0.7	0.70005	128.633	1
VEM	20000	2	NA					
GRDPG	30000	2	0.1	0.09998	0.7	0.69995	386.210	1

TABLE 2. Point Estimates and CPU time for example 1 ($K = 5$)

			latent positions/blocks						
Estimator	n	K	0.1	0.3	0.5	0.7	0.9	CPU Time (s)	ARI
GRDPG	2000	5	0.09976	0.30122	0.49819	0.70027	0.89987	1.543	1
VEM	2000	5	0.09994	0.30133	0.49825	0.70018	0.89973	257.713	1
GRDPG	5000	5	0.10015	0.29952	0.49994	0.69962	0.90003	7.982	1
VEM	5000	5	0.10021	0.29958	0.49996	0.69958	0.89999	926.330	1
GRDPG	10000	5	0.09982	0.29975	0.49990	0.70006	0.90006	44.659	1
VEM	10000	5	0.09985	0.29977	0.49990	0.70004	0.90004	8128.253	1
GRDPG	20000	5	0.10000	0.30001	0.50005	0.70019	0.89999	186.073	1
VEM	20000	5	NA						

‘

TABLE 3. Point Estimates and CPU time for example 2 ($K = 2$)

Estimator	n	K	p	$\hat{p}$	q	$\hat{q}$	β	$\hat{\beta}$	CPU Time (s)	ARI
GRDPG	2000	2	-1.5	-1.49744	1	1.00077	no covariates		4.672	1
VEM	2000	2	-1.5	-1.49712	1	1.00067	no covariates		48.619	1
GRDPG	2000	2	-1.5	-1.49454	1	0.99926	1.5	1.51201	7.557	1
VEM	2000	2	-1.5	-1.49712	1	1.00067	1.5	1.50335	6903.673	1
GRDPG	5000	2	-1.5	-1.50029	1	1.00030	no covariates		17.539	1
VEM	5000	2	-1.5	-1.50019	1	1.00024	no covariates		537.831	1
GRDPG	5000	2	-1.5	-1.49995	1	1.00064	1.5	1.49981	27.312	1
VEM	5000	2	-1.5	-1.50019	1	1.00024	1.5	1.49955	35331.012	1
GRDPG	10000	2	-1.5	-1.49989	1	1.00029	no covariates		55.428	1
GRDPG	10000	2	-1.5	1.49992	1	0.99992	1.5	1.50190	91.067	1

TABLE 4. Point estimates, standard errors and time for Example 3.

Estimator	n	K	β	$\hat{\beta}$	Time (s)	ARI
GRDPG	2000	2	1.5	1.51760	7.335	1
GRDPG	5000	2	1.5	1.49946	28.153	1
GRDPG	10000	2	1.5	1.50257	99.128	1
GRDPG	2000	2	0.5	0.64145	7.815	0.998
GRDPG	5000	2	0.5	0.56222	26.763	1
GRDPG	10000	2	0.5	0.51617	88.562	1

The second group of three rows shows the results of simulations with smaller $\beta = 0.5$. This makes the estimation of the covariate effect more challenging. Indeed when $n = 2000$ the classification in blocks and the point estimate are imprecise, as indicated by the ARI. When we increase the network size to $n = 5000$ and $n = 10000$, the accuracy of the point estimates improve significantly. This example shows that our approach is extremely useful in very large networks, where VEM may become computationally impractical.

In summary, our simple examples and simulations show that our GRDPG-based estimator is quite fast and scales well to large networks. These good computational properties are obtained without sacrificing the accuracy of the estimates, as we prove that the algorithm produces the same point estimates as the variational EM in all the examples.

4.2. **Application to Facebook 100 dataset.** We apply our method to study the network of Facebook friendships, using the Facebook 100 dataset from Traud et al. (2012).[10] We follow the analysis in Roy et al. (2019), using the Rice University network data. This network consists of 4087 nodes and 7 nodal covariates: role, gender, major, minor, dorm, year, and high school. These are all discrete variables. We focus on dorm, gender, and graduation year, and we exclude the nodes that have a missing value in any of these variables as in Roy et al. (2019). In addition, we keep only nodes whose graduation year is between 2004 and 2010 and that have at least two links.

We estimate two models with one binary covariate, focusing on gender and dorm. For the Adjacency Spectral Embedding we estimate the dimension of the latent space $\hat{d} = 12$, using the profile likelihood method in Zhu and Ghodsi (2006). The Gaussian mixture model clustering is performed using the `MCLUST` implementation of Fraley and Raftery (1999) in `R`. We obtain $\widehat{\widehat{K}} = 43$ blocks. Our procedure nicely adapts to this case of odd number of clusters. Indeed once we obtain the estimated matrix $\widehat{\boldsymbol{B}}_Z$, we cluster its diagonal to recover the (unobserved) blocks.

When we estimate the model with one binary covariate (gender), we obtain an estimate of $\hat{\beta} = 1.78096$. This means that in this network there is homophily by gender, that is higher probability of linking when two students share the same gender.

We also estimate the effect of living in the same dorm on the probability of forming a Facebook link. We find a $\hat{\beta} = 1.884547$. Not surprisingly, living in the same dorm increase the probability of becoming friends on Facebook. These results are consistent with both

[10]The entire dataset is availale at `https://archive.org/details/oxford-2005-facebook-matrix`.

Traud et al. (2012) and Roy et al. (2019), that found dorm to be an important determinant of linking.

5. Conclusion

We have developed a spectral estimator for large stochastic blockmodels with nodal covariates. The main theoretical contribution is an asymptotic normality result for the estimator of the covariates' effect on the probability of linking. Our work leverages the relationship between generalized random dot product graphs and stochastic blockmodels, extending existing frameworks to include observed covariates and constructing an estimator that is fast and scalable for large networks. Our theoretical results also apply to relatively sparse graphs, which is important in a host of applications in social sciences, public health, and computer science, where network data are usually sparse.

We have shown that our method delivers the same accuracy as the variational EM algorithm, while converging much faster. Our simulations and the empirical application show that this method works best in very large networks, when the variational EM becomes impractical.

We consider the present work a first step in the study of this class of models and the foundation for inference for SBMs and other latent position models for large networks with nodal covariates. While we have focused on binary and discrete covariates in this work, extensions to continuous covariates are currently being pursued via recently developed Latent Structure Models (Athreya et al., forthcoming). In future work, similar ideas can also be applied to directed networks and bipartite networks, significantly expanding the realm of GRDPG applications.

References

Abbe, Emmanuel (2018), 'Community detection and stochastic block models: Recent developments', *Journal of Machine Learning Research* **18**(177), 1–86.
URL: *http://jmlr.org/papers/v18/16-480.html*

Airoldi, Edoardo, David M. Blei, Stephen E. Fienberg and Eric P. Xing (2008), 'Mixed memberships stochastic blockmodels', *Journal of Machine Learning Research* **9**, 1981–2014.

Athreya, Avanti, Donniell E. Fishkind, Keith Levin, Vince Lyzinski, Youngser Park, Yichen Qin, Daniel L. Sussman, Mihn Tang, Joshua T. Vogelstein and Carey E. Priebe (2018), 'Statistical inference on random dot product graphs: A survey', *Journal of Machine Learning Research* **18**(226), 1–92.

Athreya, Avanti, Minh Tang, Youngser Park and Carey E. Priebe (forthcoming), 'On estimation and inference in latent structure random graphs', *Statistical Science* .

Bickel, Peter, David Choi, Xiangyu Chang and Hai Zhang (2013), 'Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels', *Ann. Statist.* **41**(4), 1922–1943.

Boucher, Vintent and Bernard Fortin (2016), *Oxford Handbook on the Economics of Networks*, Oxford University Press, chapter Some Challenges in the Empirics of the Effects of Networks.

Choi, David, Patrick J Wolfe and Edoardo M Airoldi (2011), 'Stochastic blockmodels with a growing number of classes', *Biometrika* **99**(2), 273–284.

Daudin, J.-J., F. Picard and S. Robin (2008), 'A mixture model for random graphs', *Statistics and Computing* **18**(2), 173–183.

Fraley, C. and A. E. Raftery (1999), 'Mclust: Software for model-based cluster analysis.', *Journal of Classification* (16), 297–306.

Goldsmith-Pinkham, Paul and Guido W. Imbens (2013), 'Social networks and the identification of peer effects', *Journal of Business and Economic Statistics* **31**(3), 253–264.
URL: *https://doi.org/10.1080/07350015.2013.801251*

Latouche, P, E Birmele and C Ambroise (2012), 'Variational bayesian inference and complexity control for stochastic block models', *Statistical Modelling* **12**(1), 93–115.

Nimczik, Jan Sebastian (2018), Job mobility networks and endogenous labor markets. working paper.

Nowicki, Krzysztof and Tom A. B Snijders (2001), 'Estimation and prediction for stochastic blockstructures', *Journal of the American Statistical Association* **96**(455), 1077–1087.

Roy, Sandipan, Yves Atchade and George Michailidis (2019), 'Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication', *Journal of Computational and Graphical Statistics* **0**(0), 1–22.

Rubin-Delanchy, Patrick, Carey E Priebe, Minh Tang and Joshua Cape (2018), A statistical interpretation of spectral embedding: the generalised random dot product graph. working paper, arXiv:1709.05506.

Shalizi, Cosma Rohilla and Edward III McFowland (2018), Estimating causal peer influence in homophilous social networks by inferring latent locations.
URL: *https://arxiv.org/abs/1607.06565*

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

Tang, Minh and Carey E. Priebe (2018), 'Limit theorems for eigenvectors of the normalized laplacian for random graphs.', *Annals of Statistics* **46**, 2360–2415.

Tang, Minh, Daniel L. Sussman and Carey E. Priebe (2013), 'Universally consistent vertex classification for latent positions graphs', *The Annals of Statistics* **41**(3), 1406–1430.
URL: *https://doi.org/10.1214/13-AOS1112*

Tang, Minh, Joshua Cape and Carey E Priebe (2017), Asymptotically efficient estimators for stochastic blockmodels: The naive mle, the rank-constrained mle, and the spectral. working paper, https://arxiv.org/abs/1710.10936.

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

Vu, Duy Q, David R Hunter and Michael Schweinberger (2013), 'Model-based clustering of large networks', *The annals of applied statistics* **7**(2), 1010–1039.

Wainwright, M.J. and M.l. Jordan (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends@ in Machine Learning* **1**(1-2), 1–305.

Zheng, D., R. Burns, J. Vogelstein, C. E. Priebe and A. S. Szalay (2016), An ssd-based eigensolver for spectral analysis on billion-node graph. working paper, https://arxiv.org/abs/1602.01421.

Zheng, Da, Disa Mhembere, Randal Burns, Joshua Vogelstein, Carey E. Priebe and Alexander S. Szalay (2015), Flashgraph: Processing billion-node graphs on an array of commodity ssds, *in* '13th USENIX Conference on File and Storage Technologies'.

Zheng, Da, Disa Mhembere, Vince Lyzinski, Randal Burns, Joshua Vogelstein and Carey E. Priebe (2017), 'Semi-external memory sparse matrix multiplication for billion-node graphs', *IEEE Transactions on Parallel and Distributed Systems* **28**(5), 1470–1483.

Zhu, M. and A. Ghodsi (2006), 'Automatic dimensionality selection from the scree plot via the use of profile likelihood.', *Computational Statistics and Data Analysis* **51**, 918–930.

Appendix A. Proofs

We first provide the general proof strategy for the simple case in which the block assignment function $\boldsymbol{\tau}$ is known. The next two theorems provide a foundation and roadmap for the proof for the more general case.

A.1. **Blocks known.** The simplest case is when the latent block assignments are known, the value of d and K are known, and h is the identity function. Let $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K)$ be the centers of the K blocks $\boldsymbol{X}$, such that i and j belong to *unobserved* block k when $\boldsymbol{X}_i = \boldsymbol{X}_j = \boldsymbol{\nu}_k$. Let $\boldsymbol{\tau}$ be the function to assigns nodes to unobserved blocks, that is $\tau_i = k$ if $\boldsymbol{X}_i = \boldsymbol{\nu}_k$. *For this subsection we will assume that $\boldsymbol{\tau}$ is known.* Our model is

$$\boldsymbol{A}_{ij} | \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j, \beta \overset{ind}{\sim} Bernoulli\left(\boldsymbol{X}_i^T \boldsymbol{X}_j + \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}\right) \tag{A.1}$$

with $\boldsymbol{Z}_i \overset{ind}{\sim} Bernoulli(b_{\tau_i})$. Therefore, model (A.1) is a $\widetilde{K} = 2K$ SBM, because we have K unobserved blocks, each split in 2 by the observed covariates. The probabilities of belonging to a block k for this $\widetilde{K}$-block SBM are $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{\widetilde{K}}) = (\pi_1 \cdot b_1, \pi_1 \cdot (1 - b_1), \pi_2 \cdot b_2, \pi_2 \cdot (1 - b_2), \ldots, \pi_K \cdot b_K, \pi_K \cdot (1 - b_K))$. Additionally, the assignment functions are $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, such that $\xi_i = 1$ if $\tau_i = 1$ and $\boldsymbol{Z}_i = 0$; $\xi_i = 2$ if $\tau_i = 1$ and $\boldsymbol{Z}_i = 1$; $\xi_i = 3$ if $\tau_i = 2$ and $\boldsymbol{Z}_i = 0$; $\xi_i = 4$ if $\tau_i = 2$ and $\boldsymbol{Z}_i = 1$; and so on.

Our SBM is $(\boldsymbol{A}, \boldsymbol{\xi}, \boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z, \boldsymbol{\eta})$ with $\widetilde{K} \times \widetilde{K}$ matrix of probabilities $\boldsymbol{\theta}_Z$

$$\boldsymbol{\theta}_Z = \begin{array}{c} \\ \tau=1;Z=0 \\ \tau=1;Z=1 \\ \tau=2;Z=0 \\ \tau=2;Z=1 \\ \vdots \\ \tau=K;Z=0 \\ \tau=K;Z=1 \end{array} \begin{array}{c} \begin{array}{ccccccc} \tau=1;Z=0 & \tau=1;Z=1 & \tau=2;Z=0 & \tau=2;Z=1 & \cdots & \tau=K;Z=0 & \tau=K;Z=1 \end{array} \\ \begin{pmatrix} \boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_2 + \beta & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_2 & \cdots & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_K + \beta & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_K \\ \boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_2 & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_2 + \beta & \cdots & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_K & \boldsymbol{\nu}_1^T\boldsymbol{\nu}_K + \beta \\ \boldsymbol{\nu}_2^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_2 + \beta & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_2 & \cdots & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_K + \beta & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_K \\ \boldsymbol{\nu}_2^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_2 & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_2 + \beta & \cdots & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_K & \boldsymbol{\nu}_2^T\boldsymbol{\nu}_K + \beta \\ \vdots & \vdots & & & \ddots & & \\ \boldsymbol{\nu}_K^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_2 + \beta & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_2 & \cdots & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_K + \beta & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_K \\ \boldsymbol{\nu}_K^T\boldsymbol{\nu}_1 & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_1 + \beta & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_2 & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_2 + \beta & \cdots & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_K & \boldsymbol{\nu}_K^T\boldsymbol{\nu}_K + \beta \end{pmatrix} \end{array}. \tag{A.2}$$

The following theorem establishes asymptotic normality for the estimator $\hat{\beta} = \hat{\boldsymbol{\theta}}_{Z,11} - \hat{\boldsymbol{\theta}}_{Z,12}$.

THEOREM A.1. *Let $\boldsymbol{A}$ be an adjacency matrix from model (A.1) with h equal to the identity function $h(u) = u$. Let $\boldsymbol{\tau}$ be known. Then $\widehat{\beta} = \widehat{\boldsymbol{\theta}}_{Z,11} - \widehat{\boldsymbol{\theta}}_{Z,12}$ is asymptotically normal, that is*

$$n\left(\hat{\beta} - \beta - \frac{\psi_\beta}{n}\right) \xrightarrow{d} N(0, \sigma_\beta^2) \tag{A.3}$$

where both ψ_β and σ^2_β are derived in the appendix.

Proof. The results in the theorem exploit the fact that β is a linear function of the entries of $\boldsymbol{\theta}_Z$, whose spectral estimators also exhibit asymptotic normality (Tang et al., 2017). There is a small bias term that goes to zero with n, and we can use the result for inference.

To prove the central limit theorem using the machinery of spectral estimation of generalized random dot product graphs models, we proceed in several steps.

STEP 1: reformulate SBM as GRDPG
Our SBM can be thought of as a GRDPG. Indeed, consider the eigendecomposition of matrix $\boldsymbol{\theta}_Z = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$, and define $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{\widetilde{K}}]$ as the rows of $\boldsymbol{U}|\boldsymbol{\Sigma}|^{1/2}$, then $F = \sum_{k=1}^{\widetilde{K}} \eta_k \delta_{\boldsymbol{\mu}_k}$, where δ is the Dirac-delta; and d_1 and d_2 are the number of positive and negative eigenvalues of $\boldsymbol{\theta}_Z$, respectively. So the GRDPG corresponding to our SBM $(\boldsymbol{A}, \boldsymbol{\xi}, \boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z, \boldsymbol{\eta})$ is given by

$$\boldsymbol{A}_{ij}|\boldsymbol{Y}_i, \boldsymbol{Y}_j \overset{ind}{\sim} Bernoulli(\boldsymbol{Y}_i^T \boldsymbol{I}_{d_1,d_2} \boldsymbol{Y}_j) \tag{A.4}$$

where $d_1 + d_2 = \widetilde{d} = rank(\boldsymbol{\theta}_Z)$ and $\boldsymbol{Y}$ is the $n \times \widetilde{d}$ vector of latent positions with centers $\boldsymbol{\mu}$.

STEP 2: spectral estimation for GRDPG
Letting $\widetilde{d} = rank(\boldsymbol{\theta}_Z)$, we then perform Adjacency Spectral Embedding (ASE) for $\boldsymbol{A}$ and obtain $\boldsymbol{A} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{U}}^T + \widehat{\boldsymbol{U}}_\perp \widehat{\boldsymbol{S}}_\perp \widehat{\boldsymbol{U}}_\perp^T$, where $\widehat{\boldsymbol{S}}$ is the diagonal matrix containing the $\widetilde{d}$ largest eigenvalues of $\boldsymbol{A}$ in absolute value, and $\widehat{\boldsymbol{U}}$ is the $n \times \widetilde{d}$ matrix whose columns are the corresponding eigenvectors of $\boldsymbol{A}$. Using a clustering procedure we can cluster the estimated latent positions $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{U}}|\widehat{\boldsymbol{S}}|^{1/2}$ (using K-means or GMM), obtaining $\widetilde{K}$ clusters and estimates of the clusters centers $\hat{\boldsymbol{\mu}}$ and cluster assignments $\hat{\boldsymbol{\xi}}$. Notice that these estimates are consistent (Athreya et al., 2018; Tang and Priebe, 2018).

Remember that in the present setting we assume $\boldsymbol{\tau}$ is known, so our estimates for the probabilities are

$$\widehat{\boldsymbol{\theta}}_{Z,k\ell} = \hat{\boldsymbol{\mu}}_k^T \boldsymbol{I}_{d_1,d_2} \hat{\boldsymbol{\mu}}_\ell \tag{A.5}$$

for any pair $k, \ell = 1, \ldots, \widetilde{K}$.

STEP 3: estimate β from matrix $\widehat{\boldsymbol{\theta}}_Z$
From the matrix $\boldsymbol{\theta}_Z$ we notice that $\beta = \boldsymbol{\theta}_{Z,11} - \boldsymbol{\theta}_{Z,12}$, thus we can use the estimator $\hat{\beta} = \widehat{\boldsymbol{\theta}}_{Z,11} - \widehat{\boldsymbol{\theta}}_{Z,12}$. With some algebra we obtain

$$\begin{aligned}
\hat{\beta} &= \widehat{\boldsymbol{\theta}}_{Z,11} - \widehat{\boldsymbol{\theta}}_{Z,12} && \text{(A.6)}\\
&= \widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11} + \boldsymbol{\theta}_{Z,11} - \boldsymbol{\theta}_{Z,12} + \boldsymbol{\theta}_{Z,12} - \widehat{\boldsymbol{\theta}}_{Z,12} && \text{(A.7)}\\
&= (\widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11}) + (\boldsymbol{\theta}_{Z,11} - \boldsymbol{\theta}_{Z,12}) - (\widehat{\boldsymbol{\theta}}_{Z,12} - \boldsymbol{\theta}_{Z,12}) && \text{(A.8)}\\
&= (\widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11}) + \beta - (\widehat{\boldsymbol{\theta}}_{Z,12} - \boldsymbol{\theta}_{Z,12}) && \text{(A.9)}\\
&= \beta + (\widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11}) - (\widehat{\boldsymbol{\theta}}_{Z,12} - \boldsymbol{\theta}_{Z,12}). && \text{(A.10)}
\end{aligned}$$

Multiplying by n and rearranging terms we finally get

$$n(\hat{\beta} - \beta) = n(\widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11}) - n(\widehat{\boldsymbol{\theta}}_{Z,12} - \boldsymbol{\theta}_{Z,12}). \tag{A.11}$$

Hence, understanding the asymptotic behavior of $n(\hat{\beta}-\beta)$ is equivalent to understanding the asymptotic behavior of the difference between $n(\widehat{\boldsymbol{\theta}}_{Z,11}-\boldsymbol{\theta}_{Z,11})$ and $n(\widehat{\boldsymbol{\theta}}_{Z,12}-\boldsymbol{\theta}_{Z,12})$. It turns out that analogously to what is available for MLE and variational approximations (Bickel et al., 2013), we can prove normality of these two terms for the GRDPG (Tang et al., 2017).

The following Lemma A.1 (corresponding to Theorem 2 in Tang et al. (2017)), showing asymptotic normality of the spectral estimator for the SBM probabilities, will be used in the proof.

LEMMA A.1. *(Theorem 2 in Tang et al. (2017)) Let $\boldsymbol{A} \sim SBM(\boldsymbol{\theta}, \boldsymbol{\eta})$ be a K-block stochastic blockmodel graph on n vertices. Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ be point masses in $\mathbb{R}^d$ such that $\boldsymbol{\theta}_{k\ell} = \boldsymbol{\mu}_k^T \boldsymbol{I}_{d_1,d_2} \boldsymbol{\mu}_\ell$ and let $\Delta = \sum_k \eta_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$. For $k \in [K]$ and $\ell \in [K]$, let $\psi_{k\ell}$ be*

$$\psi_{k\ell} = \sum_{r=1}^{K} \eta_r \left(\boldsymbol{\theta}_{kr}(1-\boldsymbol{\theta}_{kr}) + \boldsymbol{\theta}_{\ell r}(1-\boldsymbol{\theta}_{\ell r})\right) \boldsymbol{\mu}_k^T \Delta^{-1} \boldsymbol{I}_{d_1,d_2} \Delta^{-1} \boldsymbol{\mu}_\ell \tag{A.12}$$

$$- \sum_{r=1}^{K}\sum_{s=1}^{K} \eta_r \eta_s \boldsymbol{\theta}_{sr}(1-\boldsymbol{\theta}_{sr}) \boldsymbol{\mu}_s^T \Delta^{-1} \boldsymbol{I}_{d_1,d_2} \Delta^{-1} (\boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k \boldsymbol{\mu}_\ell^T) \Delta^{-1} \boldsymbol{\mu}_s. \tag{A.13}$$

Let $\zeta_{k\ell} = \boldsymbol{\mu}_k^T \Delta^{-1} \boldsymbol{\mu}_\ell$ and define σ_{kk}^2 for $k \in [K]$ to be

$$\sigma_{kk}^2 = 4\boldsymbol{\theta}_{kk}(1-\boldsymbol{\theta}_{kk})\zeta_{kk}^2 + 4\sum_{r=1}^{K} \eta_r \boldsymbol{\theta}_{kr}(1-\boldsymbol{\theta}_{kr})\zeta_{kr}^2 \left(\frac{1}{\eta_k} - 2\zeta_{kk}\right) \tag{A.14}$$

$$+ 2\sum_{r=1}^{K}\sum_{s=1}^{K} \eta_r \eta_s \boldsymbol{\theta}_{rs}(1-\boldsymbol{\theta}_{rs})\zeta_{kr}^2 \zeta_{ks}^2 \tag{A.15}$$

and define $\sigma_{k\ell}^2$ for $k \in [K]$ and $\ell \in [K]$, $k \neq \ell$ to be

$$\sigma_{k\ell}^2 = \left(\boldsymbol{\theta}_{kk}(1-\boldsymbol{\theta}_{kk}) + \boldsymbol{\theta}_{\ell\ell}(1-\boldsymbol{\theta}_{\ell\ell})\right)\zeta_{kl}^2 + 2\boldsymbol{\theta}_{k\ell}(1-\boldsymbol{\theta}_{k\ell})\zeta_{kk}\zeta_{\ell\ell} \tag{A.16}$$

$$+ \sum_{r=1}^{K} \eta_r \boldsymbol{\theta}_{kr}(1-\boldsymbol{\theta}_{kr})\zeta_{\ell r}^2 \left(\frac{1}{\eta_k} - 2\zeta_{kk}\right) \tag{A.17}$$

$$+ \sum_{r=1}^{K} \eta_r \boldsymbol{\theta}_{\ell r}(1-\boldsymbol{\theta}_{\ell r})\zeta_{kr}^2 \left(\frac{1}{\eta_\ell} - 2\zeta_{\ell\ell}\right) \tag{A.18}$$

$$- 2\sum_{r=1}^{K} \eta_r \left(\boldsymbol{\theta}_{kr}(1-\boldsymbol{\theta}_{kr}) + \boldsymbol{\theta}_{\ell r}(1-\boldsymbol{\theta}_{\ell r})\right) \zeta_{kr}\zeta_{r\ell}\zeta_{k\ell} \tag{A.19}$$

$$+ \frac{1}{2}\sum_{r=1}^{K}\sum_{s=1}^{K} \eta_r \eta_s \boldsymbol{\theta}_{rs}(1-\boldsymbol{\theta}_{rs}) \left(\zeta_{kr}\zeta_{\ell s} + \zeta_{\ell r}\zeta_{ks}\right)^2. \tag{A.20}$$

Then for any $k \in [K]$ *and* $\ell \in [K]$,

$$n\left(\widehat{\boldsymbol{\theta}}_{k\ell} - \boldsymbol{\theta}_{k\ell} - \frac{\psi_{k\ell}}{n}\right) \xrightarrow{d} N(0, \sigma^2_{k\ell}) \tag{A.21}$$

as $n \to \infty$.

Proof. See Tang et al. (2017) for a detailed proof. □

Using the result in Lemma A.1, we can see that

$$n(\widehat{\boldsymbol{\theta}}_{Z,11} - \boldsymbol{\theta}_{Z,11}) \xrightarrow{d} N(\psi_{11}, \sigma^2_{11}) \tag{A.22}$$

$$n(\widehat{\boldsymbol{\theta}}_{Z,12} - \boldsymbol{\theta}_{Z,12}) \xrightarrow{d} N(\psi_{12}, \sigma^2_{12}) \tag{A.23}$$

and by consequence

$$n(\hat{\beta} - \beta) \xrightarrow{d} N(\psi_\beta, \sigma^2_\beta) \tag{A.24}$$

where

$$\psi_\beta = \psi_{11} - \psi_{12} \tag{A.25}$$

$$\sigma^2_\beta = \sigma^2_{11} + \sigma^2_{12} - 2\sigma_{11,12} \tag{A.26}$$

and we have used notation $\sigma_{k\ell,k'\ell'}$ to indicated the covariance terms, that is

$$\sigma_{k\ell,k'\ell'} = \mathbb{COV}(\widehat{\boldsymbol{\theta}}_{Z,k\ell}, \widehat{\boldsymbol{\theta}}_{Z,k'\ell'}) \tag{A.27}$$

for any $k, \ell, k', \ell' \in \{1, \ldots, K\}$. The first two terms of the variance are given above in Lemma A.1. The multivariate version of the CLT can be obtained by applying the Cramer-Wold device. For the covariance term $\sigma_{11,12}$, we provide the calculation below.

Computation of the covariance terms

We compute the covariance for a slightly more general model, that includes a sparsity coefficient ρ_n, that is

$$\boldsymbol{P}_{ij} = \rho_n \left(\boldsymbol{X}_i^T \boldsymbol{X}_j + \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}}\right). \tag{A.28}$$

In the main text we separately discuss the cases in which $\rho_n \to c$, where $c > 0$ is a constant, or $\rho_n \to 0$, but $n\rho_n = \omega(\sqrt{n})$ when $n \to \infty$. Let $\boldsymbol{s}_k$ be the vector in $\mathbb{R}^n$ whose i-th entry is 1 if $\xi_i = k$ and 0 otherwise, and let n_k be the number of nodes in block k, that is $n_k = |\{i : \xi_i = k\}|$.

We want to compute the correlation between $\hat{\boldsymbol{\theta}}_{Z,k\ell}$ and $\hat{\boldsymbol{\theta}}_{Z,k'\ell'}$, for $k \neq k'$ and $\ell \neq \ell'$ in general.

To simplify notation, we will omit the Z from the subscript, so we will refer to $\hat{\boldsymbol{\theta}}_{k\ell}$ instead of $\hat{\boldsymbol{\theta}}_{Z,k\ell}$ for any k, ℓ. Let $\boldsymbol{S}$ be the $d \times d$ diagonal matrix containing the largest d eigenvalues of $\boldsymbol{P}$ in absolute value, and let $\boldsymbol{U}$ be the $n \times d$ matrix whose rows are the corresponding eigenvectors of $\boldsymbol{P}$.

We start from equation (A.5) in the appendix of Tang et al. (2017).

$$\frac{n\rho_n^{1/2}}{n_k n_\ell}\left(\hat{\boldsymbol{\theta}}_{k\ell}-\boldsymbol{\theta}_{k\ell}\right) = \frac{n\rho_n^{-1/2}}{n_k n_\ell}\left(\boldsymbol{s}_k^T\boldsymbol{E}\boldsymbol{\Pi}_U\boldsymbol{s}_\ell+\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U^\perp\boldsymbol{E}\boldsymbol{\Pi}_U\boldsymbol{s}_k\right) \tag{A.29}$$

$$+ \frac{n\rho_n^{-1/2}}{n_k n_\ell}\left(\boldsymbol{s}_k^T\boldsymbol{\Pi}_U^\perp\boldsymbol{E}^2\boldsymbol{P}^\dagger\boldsymbol{s}_\ell+\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U^\perp\boldsymbol{E}^2\boldsymbol{P}^\dagger\boldsymbol{s}_k\right) \tag{A.30}$$

$$+ O_p\left(n^{-1/2}\rho_n^{-1}\right) \tag{A.31}$$

where $\boldsymbol{E}=\boldsymbol{A}-\boldsymbol{P}$, $\boldsymbol{\Pi}_U=\boldsymbol{U}\boldsymbol{U}^T$, $\boldsymbol{\Pi}_U^\perp=\boldsymbol{I}-\boldsymbol{\Pi}_U$ and $\boldsymbol{P}^\dagger=\boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^T$.

Term (A.31) goes to zero when $n\rho_n=\omega(\sqrt{n})$ as $n\to\infty$.

Term (A.30) is the bias term, corresponding to $\psi_{k\ell}$ or $\rho_n^{-1/2}\tilde{\psi}_{k\ell}$ depending on whether $\rho_n\equiv 1$ or $\rho_n\to 0$, respectively.

Term (A.29) is the leading order term which converges in distribution to a normal random variable. In particular, this is the term that must be considered when deriving asymptotic covariances.

Towards this end, define $\boldsymbol{\Upsilon}_{k\ell}$ below as in equation (A.6) found in Tang et al. (2017), namely

$$\boldsymbol{\Upsilon}_{k\ell} := \frac{n\rho_n^{-1/2}}{n_k n_\ell}\left(\boldsymbol{s}_k^T\boldsymbol{E}\boldsymbol{\Pi}_U\boldsymbol{s}_\ell+\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U^\perp\boldsymbol{E}\boldsymbol{\Pi}_U\boldsymbol{s}_k\right) \tag{A.32}$$

$$= \frac{n\rho_n^{-1/2}}{n_k n_\ell} tr\boldsymbol{E}\left(\boldsymbol{\Pi}_U\boldsymbol{s}_\ell\boldsymbol{s}_k^T+\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U^\perp\right) \tag{A.33}$$

$$= \frac{n\rho_n^{-1/2}}{n_k n_\ell} tr\boldsymbol{E}\left(\boldsymbol{\Pi}_U\boldsymbol{s}_\ell\boldsymbol{s}_k^T+\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T-\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U\right) \tag{A.34}$$

$$= \frac{n\rho_n^{-1/2}}{n_k n_\ell} tr\left(\boldsymbol{A}-\boldsymbol{P}\right)\left(\boldsymbol{\Pi}_U\boldsymbol{s}_\ell\boldsymbol{s}_k^T+\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T-\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U\right) \tag{A.35}$$

$$= \frac{n\rho_n^{-1/2}}{n_k n_\ell} tr\left(\boldsymbol{A}-\boldsymbol{P}\right)\boldsymbol{M} \tag{A.36}$$

where $\boldsymbol{M}:=\boldsymbol{\Pi}_U\boldsymbol{s}_\ell\boldsymbol{s}_k^T+\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T-\boldsymbol{\Pi}_U\boldsymbol{s}_k\boldsymbol{s}_\ell^T\boldsymbol{\Pi}_U$.

We therefore have:

$$\boldsymbol{\Upsilon}_{k\ell}=\frac{n\rho_n^{-1/2}}{n_k n_\ell}\sum_i\sum_j\left(\boldsymbol{A}_{ij}-\boldsymbol{P}_{ij}\right)\boldsymbol{M}_{ij}. \tag{A.37}$$

First, note that the variable $\boldsymbol{\Upsilon}_{k\ell}$ has expected value equal to zero, i.e., $\mathbb{E}[\boldsymbol{\Upsilon}_{k\ell}]=0$, since $\mathbb{E}[\boldsymbol{A}_{ij}]=\boldsymbol{P}_{ij}$ for each pair $\{i,j\}$. This implies that we can focus on computing covariances as

$$\mathbb{COV}[\boldsymbol{\Upsilon}_{k\ell},\boldsymbol{\Upsilon}_{k'\ell'}]=\mathbb{E}[\boldsymbol{\Upsilon}_{k\ell}\boldsymbol{\Upsilon}_{k'\ell'}]. \tag{A.38}$$

Towards this end, let the matrix $\boldsymbol{Q}$ be defined as

$$\boldsymbol{Q} = \boldsymbol{\Pi}_U \boldsymbol{s}_{\ell'} \boldsymbol{s}_{k'}^T + \boldsymbol{\Pi}_U \boldsymbol{s}_{k'} \boldsymbol{s}_{\ell'}^T - \boldsymbol{\Pi}_U \boldsymbol{s}_{k'} \boldsymbol{s}_{\ell'}^T \boldsymbol{\Pi}_U, \tag{A.39}$$

and so

$$\boldsymbol{\Upsilon}_{k'\ell'} = \frac{n\rho_n^{-1/2}}{n_{k'} n_{\ell'}} \sum_i \sum_j \left(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij}\right) \boldsymbol{Q}_{ij} \tag{A.40}$$

The product $\boldsymbol{\Upsilon}_{k\ell} \boldsymbol{\Upsilon}_{k'\ell'}$ is then given by

$$\begin{aligned} \boldsymbol{\Upsilon}_{k\ell} \boldsymbol{\Upsilon}_{k'\ell'} &= \left(\frac{n\rho_n^{-1/2}}{n_k n_\ell} \sum_i \sum_j \left(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij}\right) \boldsymbol{M}_{ij} \right) \left(\frac{n\rho_n^{-1/2}}{n_{k'} n_{\ell'}} \sum_{i'} \sum_{j'} \left(\boldsymbol{A}_{i'j'} - \boldsymbol{P}_{i'j'}\right) \boldsymbol{Q}_{i'j'} \right) && \text{(A.41)} \\ &= \frac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_i \sum_j \sum_{i'} \sum_{j'} \left(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij}\right) \left(\boldsymbol{A}_{i'j'} - \boldsymbol{P}_{i'j'}\right) \boldsymbol{M}_{ij} \boldsymbol{Q}_{i'j'} && \text{(A.42)} \end{aligned}$$

First note that when $\{i, j\} \neq \{i', j'\}$, then the expected value of the corresponding term in the above summation is zero, since then $(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij})$ and $(\boldsymbol{A}_{i'j'} - \boldsymbol{P}_{i'j'})$ are independent, centered random variables. Therefore we can focus on the case when $\{i, j\} = \{i', j'\}$.

Further expanding $\mathbb{E}[\boldsymbol{\Upsilon}_{k\ell} \boldsymbol{\Upsilon}_{k'\ell'}]$ subsequently yields

$$\begin{aligned} \mathbb{E}[\boldsymbol{\Upsilon}_{k\ell} \boldsymbol{\Upsilon}_{k'\ell'}] &= \frac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_i \sum_j \sum_{i'} \sum_{j'} \mathbb{E}\left[\left(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij}\right) \left(\boldsymbol{A}_{i'j'} - \boldsymbol{P}_{i'j'}\right)\right] \boldsymbol{M}_{ij} \boldsymbol{Q}_{i'j} && \text{(A.43)} \\ &= \frac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_i \sum_j \mathbb{E}\left[\left(\boldsymbol{A}_{ij} - \boldsymbol{P}_{ij}\right)^2\right] \boldsymbol{M}_{ij} \boldsymbol{Q}_{ij} && \text{(A.44)} \\ &= \frac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_i \sum_j \boldsymbol{P}_{ij} \left(1 - \boldsymbol{P}_{ij}\right) \boldsymbol{M}_{ij} \boldsymbol{Q}_{ij}. && \text{(A.45)} \end{aligned}$$

We thus need to compute the entries of $\boldsymbol{M}$ and $\boldsymbol{Q}$ to obtain a computable formula.

We have for $\boldsymbol{M}$ that $\boldsymbol{M}_{ij} := \upsilon_{ij}^{(1)} + \upsilon_{ij}^{(2)} + \upsilon_{ij}^{(3)}$, where

$$\begin{aligned} \upsilon_{ij}^{(1)} := \left(\boldsymbol{\Pi}_U \boldsymbol{s}_\ell \boldsymbol{s}_k^T\right)_{ij} &= n_\ell \boldsymbol{Y}_i^T (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{\mu}_\ell \mathbb{1}\{\xi_j = k\} && \text{(A.46)} \\ \upsilon_{ij}^{(2)} := \left(\boldsymbol{\Pi}_U \boldsymbol{s}_k \boldsymbol{s}_\ell^T\right)_{ij} &= n_k \boldsymbol{Y}_i^T (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{\mu}_k \mathbb{1}\{\xi_j = \ell\} && \text{(A.47)} \\ \upsilon_{ij}^{(3)} := \left(\boldsymbol{\Pi}_U \boldsymbol{s}_k \boldsymbol{s}_\ell^T \boldsymbol{\Pi}_U\right)_{ij} &= n_k n_\ell \boldsymbol{Y}_i^T (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{\mu}_k \boldsymbol{\mu}_\ell^T (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{Y}_j && \text{(A.48)} \\ & && \text{(A.49)} \end{aligned}$$

and analogously for $\boldsymbol{Q}_{ij} := \varrho_{ij}^{(1)} + \varrho_{ij}^{(2)} + \varrho_{ij}^{(3)}$. Hence, $\boldsymbol{M}_{ij} \boldsymbol{Q}_{ij} = \sum_{1 \leq \alpha, \beta \leq 3} \upsilon_{ij}^{(\alpha)} \varrho_{ij}^{(\beta)}$. In what follows, it will sometimes be useful to write the scalars $\upsilon_{ij}^{(\alpha)}, \varrho_{ij}^{(\beta)}$ equivalently in terms of their transpose (i.e., see the right-hand sides of their definitions).

To begin, we make a preliminary observation that

$$\begin{aligned}
&\tfrac{\rho_n^{-1}}{n_k n_{k'}} \left(\sum_{i,j} \boldsymbol{P}_{ij}(1-\boldsymbol{P}_{ij}) \boldsymbol{Y}_i \boldsymbol{Y}_i^T \mathbb{1}\{\xi_j = k\} \mathbb{1}\{\xi_j = k'\} \right) \\
&\qquad \overset{a.s.}{\to} \begin{cases} \frac{1}{\eta_k} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} (1 - \boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] & \text{if } \rho_n \equiv 1 \text{ and } k = k', \\ \frac{1}{\eta_k} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] & \text{if } \rho_n \to 0 \text{ and } k = k', \\ 0 & \text{if } k \neq k'. \end{cases}
\end{aligned}$$

Consider the terms involving $\upsilon_{ij}^{(1)} \varrho_{ij}^{(1)}$, $\upsilon_{ij}^{(1)} \varrho_{ij}^{(2)}$, $\upsilon_{ij}^{(2)} \varrho_{ij}^{(1)}$, and $\upsilon_{ij}^{(2)} \varrho_{ij}^{(2)}$. Then

$$\begin{aligned}
&\tfrac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij} \left(1 - \boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(1)} \varrho_{ij}^{(1)} \\
&\quad = \tfrac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \left(n_\ell n_{\ell'} \boldsymbol{\mu}_\ell^T (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \left(\sum_{i,j} \boldsymbol{P}_{ij}(1-\boldsymbol{P}_{ij}) \boldsymbol{Y}_i \boldsymbol{Y}_i^T \mathbb{1}\{\xi_j = k\} \mathbb{1}\{\xi_j = k'\} \right) (\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{\mu}_{\ell'} \right) \\
&\quad \overset{a.s.}{\to} \begin{cases} \frac{1}{\eta_k} \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} (1 - \boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \equiv 1 \text{ and } k = k', \\ \frac{1}{\eta_k} \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \to 0 \text{ and } k = k', \\ 0 & \text{if } k \neq k', \end{cases}
\end{aligned}$$

and similarly

$$\begin{aligned}
&\tfrac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij} \left(1 - \boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(1)} \varrho_{ij}^{(2)} \\
&\qquad \overset{a.s.}{\to} \begin{cases} \frac{1}{\eta_k} \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} (1 - \boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \equiv 1 \text{ and } k = \ell', \\ \frac{1}{\eta_k} \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \to 0 \text{ and } k = \ell', \\ 0 & \text{if } k \neq \ell', \end{cases}
\end{aligned}$$

and similarly

$$\begin{aligned}
&\tfrac{n^2 \rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij} \left(1 - \boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(2)} \varrho_{ij}^{(1)} \\
&\qquad \overset{a.s.}{\to} \begin{cases} \frac{1}{\eta_\ell} \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)} (1 - \boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \equiv 1 \text{ and } \ell = k', \\ \frac{1}{\eta_\ell} \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T \right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \to 0 \text{ and } \ell = k', \\ 0 & \text{if } \ell \neq k'. \end{cases}
\end{aligned}$$

and similarly

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(2)} \varrho_{ij}^{(2)}$$

$$\overset{a.s.}{\to} \begin{cases} \frac{1}{\eta_\ell} \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \equiv 1 \text{ and } \ell = \ell', \\ \frac{1}{\eta_\ell} \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \to 0 \text{ and } \ell = \ell', \\ 0 & \text{if } \ell \neq \ell'. \end{cases}$$

Next, we consider the terms involving $\upsilon_{ij}^{(1)} \varrho_{ij}^{(3)}$, $\upsilon_{ij}^{(2)} \varrho_{ij}^{(3)}$, $\upsilon_{ij}^{(3)} \varrho_{ij}^{(1)}$, and $\upsilon_{ij}^{(3)} \varrho_{ij}^{(2)}$. In particular,

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(1)} \varrho_{ij}^{(3)}$$

$$= \frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \left(n_\ell n_{k'} n_{\ell'} \mathbb{1}\{\xi_j = k\} \boldsymbol{\mu}_\ell^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{Y}_j\right)$$

$$\overset{a.s.}{\to} \begin{cases} \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_k & \text{if } \rho_n \equiv 1, \\ \boldsymbol{\mu}_\ell^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_k & \text{if } \rho_n \to 0, \end{cases}$$

and similarly

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(2)} \varrho_{ij}^{(3)}$$

$$\overset{a.s.}{\to} \begin{cases} \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell & \text{if } \rho_n \equiv 1, \\ \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell & \text{if } \rho_n \to 0. \end{cases}$$

Along the same lines,

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(3)} \varrho_{ij}^{(1)}$$

$$= \frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \left(n_k n_\ell n_{\ell'} \mathbb{1}\{\xi_j = k'\} \boldsymbol{Y}_j^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{\mu}_{\ell'}\right)$$

$$\overset{a.s.}{\to} \begin{cases} \boldsymbol{\mu}_{k'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k'\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{k'\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \equiv 1, \\ \boldsymbol{\mu}_{k'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{k'\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{\ell'} & \text{if } \rho_n \to 0, \end{cases}$$

and similarly

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(3)} \varrho_{ij}^{(2)}$$

$$\overset{a.s.}{\to} \begin{cases} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell'\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{\ell'\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \equiv 1, \\ \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\ell'\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} & \text{if } \rho_n \to 0. \end{cases}$$

Finally, consider the term involving $\upsilon_{ij}^{(3)} \varrho_{ij}^{(3)}$. We see that

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(3)} \varrho_{ij}^{(3)}$$
$$= \frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \left(n_k n_\ell n_{k'} n_{\ell'} \boldsymbol{Y}_j^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T (\boldsymbol{Y}^T\boldsymbol{Y})^{-1} \boldsymbol{Y}_j\right).$$

In order to analyze this quantity, we now decompose the sum over the index j using the indicator variables $\mathbb{1}\{\xi_j = \alpha\}$ for all SBM blocks $\alpha \in \mathcal{A}$. For each such term we get

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(3)} \varrho_{ij}^{(3)} \mathbb{1}\{\xi_j = \alpha\}$$
$$\overset{a.s.}{\to} \begin{cases} \eta_\alpha \boldsymbol{\mu}_\alpha^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\alpha & \text{if } \rho_n \equiv 1, \\ \eta_\alpha \boldsymbol{\mu}_\alpha^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\alpha & \text{if } \rho_n \to 0. \end{cases}$$

Hence, by aggregating over all $\alpha \in \mathcal{A}$, we obtain

$$\frac{n^2\rho_n^{-1}}{n_k n_\ell n_{k'} n_{\ell'}} \sum_{i,j} \boldsymbol{P}_{ij}\left(1-\boldsymbol{P}_{ij}\right) \upsilon_{ij}^{(3)} \varrho_{ij}^{(3)}$$
$$\overset{a.s.}{\to} \begin{cases} \sum_\alpha \eta_\alpha \boldsymbol{\mu}_\alpha^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)}(1-\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)}) \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\alpha & \text{if } \rho_n \equiv 1, \\ \sum_\alpha \eta_\alpha \boldsymbol{\mu}_\alpha^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\ell \boldsymbol{\mu}_k^T \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\boldsymbol{\theta}_{\alpha\xi(\boldsymbol{Y}_1)} \boldsymbol{Y}_1 \boldsymbol{Y}_1^T\right] \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_{k'} \boldsymbol{\mu}_{\ell'}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\mu}_\alpha & \text{if } \rho_n \to 0. \end{cases}$$

Combining all of the above observations yields $\mathbb{COV}[\boldsymbol{\Upsilon}_{k\ell} \boldsymbol{\Upsilon}_{k'\ell'}]$ for all possible relationships between tuples $\{k, \ell\}, \{k', \ell'\}$ and for both regimes $\rho_n \equiv 1, \rho_n \to 0$.

□

A.2. **Blocks known and general link function.** If the link function is not the identity, then our general model becomes

$$\boldsymbol{A}_{ij} | \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j, \beta \overset{ind}{\sim} Bernoulli\left(h(\boldsymbol{X}_i^T \boldsymbol{X}_j + \beta \mathbf{1}_{\{\boldsymbol{Z}_i = \boldsymbol{Z}_j\}})\right). \tag{A.50}$$

A popular choice of h is the logistic specification, with $h(u) = e^u/(1+e^u)$ (Choi et al., 2011; Traud et al., 2012; Sweet, 2015; Nimczik, 2018). The generalization of the central limit theorem is as follows.

THEOREM A.2. *(General nonlinear h function) Let $\boldsymbol{A}$ be an adjacency matrix from model (A.50) and let h be a link function, $h : \mathcal{X} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{Z} \to [0,1]$. Let $\boldsymbol{\tau}$ be known and let function g be defined as the inverse of h, that is $g(\cdot) = h^{-1}(\cdot)$, with first derivative $g'(\cdot)$. Let $g'(\boldsymbol{\nu}_1^T \boldsymbol{\nu}_1 + \beta) \neq 0$ and $g'(\boldsymbol{\nu}_1^T \boldsymbol{\nu}_2) \neq 0$. Then $\widehat{\beta} = h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,11}) - h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,12})$ is asymptotically normal, in particular*

$$n\left(\hat{\beta} - \beta - \frac{\widetilde{\psi}_\beta}{n}\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}_\beta^2\right) \tag{A.51}$$

where $\widetilde{\psi}_\beta$ and $\widetilde{\sigma}_\beta^2$ are computed in appendix.

Proof. In this case model (23) is as follows:

$$\boldsymbol{A}_{ij}|\boldsymbol{X}_i,\boldsymbol{X}_j,\boldsymbol{Z}_i,\boldsymbol{Z}_j,\beta \overset{ind}{\sim} Bernoulli\left(h\left(\boldsymbol{X}_i^T\boldsymbol{X}_j+\beta\mathbf{1}_{\{\boldsymbol{Z}_i=\boldsymbol{Z}_j\}}\right)\right). \tag{A.52}$$

All the rest is the same, and we can write our SBM as $(\boldsymbol{A},\boldsymbol{\xi},\boldsymbol{Z}) \sim SBM(\boldsymbol{\theta}_Z,\boldsymbol{\eta})$ with $\widetilde{K}\times\widetilde{K}$ matrix of probabilities $\boldsymbol{\theta}_Z$ given by

$$\boldsymbol{\theta}_Z = \begin{array}{c} \\ \tau=1;Z=0 \\ \tau=1;Z=1 \\ \tau=2;Z=0 \\ \tau=2;Z=1 \\ \vdots \\ \tau=K;Z=0 \\ \tau=K;Z=1 \end{array} \begin{array}{c} \begin{array}{cccccccc} \tau=1;Z=0 & \tau=1;Z=1 & \tau=2;Z=0 & \tau=2;Z=1 & \cdots & \tau=K;Z=0 & \tau=K;Z=1 \end{array} \\ \begin{pmatrix} h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2+\beta) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2) & \cdots & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K+\beta) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K) \\ h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_2+\beta) & \cdots & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K) & h(\boldsymbol{\nu}_1^T\boldsymbol{\nu}_K+\beta) \\ h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2+\beta) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2) & \cdots & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K+\beta) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K) \\ h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_2+\beta) & \cdots & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K) & h(\boldsymbol{\nu}_2^T\boldsymbol{\nu}_K+\beta) \\ \vdots & \vdots & \vdots & & \ddots & & \\ h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2+\beta) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2) & \cdots & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K+\beta) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K) \\ h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_1+\beta) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_2+\beta) & \cdots & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K) & h(\boldsymbol{\nu}_K^T\boldsymbol{\nu}_K+\beta) \end{pmatrix} \end{array}. \tag{A.53}$$

So we know that $\beta = h^{-1}(\boldsymbol{\theta}_{Z,11}) - h^{-1}(\boldsymbol{\theta}_{Z,12})$, and we can use the estimator

$$\hat{\beta} = h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,11}) - h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,12}). \tag{A.54}$$

By Theorem A.1 we know that

$$n(\widehat{\boldsymbol{\theta}}_{Z,11}-\boldsymbol{\theta}_{Z,11}) \xrightarrow{d} N(\psi_{11},\sigma_{11}^2) \tag{A.55}$$

$$n(\widehat{\boldsymbol{\theta}}_{Z,12}-\boldsymbol{\theta}_{Z,12}) \xrightarrow{d} N(\psi_{12},\sigma_{12}^2) \tag{A.56}$$

Applying a Taylor expansion reveals

$$h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,11}) = h^{-1}(\boldsymbol{\theta}_{Z,11}) + \left(h^{-1}(\boldsymbol{\theta}_{Z,11})\right)'\left(\widehat{\boldsymbol{\theta}}_{Z,11}-\boldsymbol{\theta}_{Z,11}\right) + \text{smaller order terms} \tag{A.57}$$

$$h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,12}) = h^{-1}(\boldsymbol{\theta}_{Z,12}) + \left(h^{-1}(\boldsymbol{\theta}_{Z,12})\right)'\left(\widehat{\boldsymbol{\theta}}_{Z,12}-\boldsymbol{\theta}_{Z,12}\right) + \text{smaller order terms} \tag{A.58}$$

and this implies

$$n\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,11}) - h^{-1}(\boldsymbol{\theta}_{Z,11})\right) \xrightarrow{d} N(\widetilde{\psi}_{11},\widetilde{\sigma}_{11}^2), \tag{A.59}$$

$$n\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{Z,12}) - h^{-1}(\boldsymbol{\theta}_{Z,12})\right) \xrightarrow{d} N(\widetilde{\psi}_{12},\widetilde{\sigma}_{12}^2), \tag{A.60}$$

where $\widetilde{\psi}_{k\ell} = \psi_{k\ell}\left(h^{-1}(\boldsymbol{\theta}_{Z,k\ell})\right)'$ and $\widetilde{\sigma}_{k\ell}^2 = \sigma_{k\ell}^2\left[\left(h^{-1}(\boldsymbol{\theta}_{Z,k\ell})\right)'\right]^2$ for $k,\ell=1,2$.

Finally this implies that our estimator behaves in the manner

$$n(\hat{\beta}-\beta) \xrightarrow{d} N(\widetilde{\psi}_{\beta},\widetilde{\sigma}_{\beta}^2), \tag{A.61}$$

where the bias term is

$$\widetilde{\psi}_{\beta} = \widetilde{\psi}_{11} - \widetilde{\psi}_{12}, \tag{A.62}$$

where the variance term $\widetilde{\sigma}_{\beta}^2$ satisfies

$$\widetilde{\sigma}_{\beta}^2 = \widetilde{\sigma}_{11}^2 + \widetilde{\sigma}_{12}^2 - 2\widetilde{\sigma}_{11,12}, \tag{A.63}$$

and where the covariance term is given by

$$\widetilde{\sigma}_{11,12} = cov(h^{-1}(\widehat{\boldsymbol{\theta}}_{11}), h^{-1}(\widehat{\boldsymbol{\theta}}_{12})) = \sigma_{11,12}\left[\left(h^{-1}(\boldsymbol{\theta}_{Z,11})\right)'\right]\left[\left(h^{-1}(\boldsymbol{\theta}_{Z,12})\right)'\right]. \tag{A.64}$$

□

A.3. **Proof of THEOREM 1.** Let K be known. When $\boldsymbol{\tau}$ is not known, we can estimate it using Adjacency Spectral Embedding (ASE) to get an estimate $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{U}}|\widehat{\boldsymbol{S}}|^{1/2}$; we then cluster the rows of $\widehat{\boldsymbol{Y}}$ using a Gaussian Mixture Model (GMM) or K-means clustering. This gives estimates $\widehat{\boldsymbol{\xi}}$ and therefore $\widehat{\boldsymbol{\tau}}$ as we can easily estimate the probability b_k of the Bernoulli covariates from the observed $\boldsymbol{Z}_i$'s, within each estimated block.

Because of the estimation the blocks are recovered up to a permutation of the blocks' labels (as it is in any mixture model). Nonetheless, we can still obtain an asymptotic result for $\widehat{\beta}$ using the following Lemma A.2, taken from Tang et al. (2017).

LEMMA A.2. *(Corollary 2 in Tang et al. (2017)) Let the setting and notation be as in Theorem A.1. Let K be known and let $\widehat{\boldsymbol{\xi}} : [n] \to [\tilde{K}]$ be the function that assigns nodes to clusters, estimated using GMM or K-means clustering on the rows of $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{U}}|\widehat{\boldsymbol{S}}|^{1/2}$ (as in the proof of Theorem A.1). Let $\widehat{\boldsymbol{\theta}}_{Z,k\ell} = \widehat{\boldsymbol{\mu}}_k^T \mathbf{I}_{d_1,d_2} \widehat{\boldsymbol{\mu}}_\ell$ and let $\widehat{\boldsymbol{\Delta}} = \sum_{k=1}^K \widehat{\eta}_k \widehat{\boldsymbol{\mu}}_k \widehat{\boldsymbol{\mu}}_\ell^T$. For $k \in [K]$ and $\ell \in [K]$ define $\widehat{\psi}_{k\ell}$ as*

$$\widehat{\psi}_{k\ell} = \sum_{r=1}^{K} \widehat{\xi}_r \left(\widehat{\boldsymbol{\theta}}_{kr}(1 - \widehat{\boldsymbol{\theta}}_{kr}) + \widehat{\boldsymbol{\theta}}_{\ell r}(1 - \widehat{\boldsymbol{\theta}}_{\ell r}) \right) \widehat{\boldsymbol{\mu}}_k^T \widehat{\Delta}^{-1} \boldsymbol{I}_{d_1,d_2} \widehat{\Delta}^{-1} \widehat{\boldsymbol{\mu}}_\ell \tag{A.65}$$

$$- \sum_{r=1}^{K} \sum_{s=1}^{K} \widehat{\eta}_r \widehat{\eta}_s \widehat{\boldsymbol{\theta}}_{sr}(1 - \widehat{\boldsymbol{\theta}}_{sr}) \widehat{\boldsymbol{\mu}}_s^T \widehat{\Delta}^{-1} \boldsymbol{I}_{d_1,d_2} \widehat{\Delta}^{-1} (\widehat{\boldsymbol{\mu}}_\ell \widehat{\boldsymbol{\mu}}_k^T + \widehat{\boldsymbol{\mu}}_k \widehat{\boldsymbol{\mu}}_\ell^T) \widehat{\Delta}^{-1} \widehat{\boldsymbol{\mu}}_s \tag{A.66}$$

Then there exists a sequence of permutations $\phi \equiv \phi_n$ on $[K]$ such that for any $k \in [K]$ and $\ell \in [K]$,

$$n \left(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)} - \boldsymbol{\theta}_{k\ell} - \frac{\widehat{\psi}_{k\ell}}{n} \right) \xrightarrow{d} N(0, \sigma_{k\ell}^2) \tag{A.67}$$

as $n \to \infty$.

Proof. See Tang et al. (2017) for the detailed proof. □

Let us first focus on the **linear** case, in which h is the identity function. If $h(u) = u$, then we can estimate β as

$$\widehat{\beta} = \widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)} - \widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)} \tag{A.68}$$

$$= \widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)} - \boldsymbol{\theta}_{11} + \boldsymbol{\theta}_{11} - \boldsymbol{\theta}_{12} + \boldsymbol{\theta}_{12} - \widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)} \tag{A.69}$$

$$= \left(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)} - \boldsymbol{\theta}_{11} \right) + \beta - \left(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)} - \boldsymbol{\theta}_{12} \right) \tag{A.70}$$

and rearranging we obtain

$$n \left(\widehat{\beta} - \beta \right) = n \left(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)} - \boldsymbol{\theta}_{11} \right) - n \left(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)} - \boldsymbol{\theta}_{12} \right) \tag{A.71}$$

which by Lemma A.2 implies that there exists a (sequence of) permutation(s) ϕ such that

$$n\left(\widehat{\beta}-\beta-\frac{\widehat{\psi}_\beta}{n}\right)\xrightarrow{d}N(0,\sigma_\beta^2) \tag{A.72}$$

where $\widehat{\psi}_\beta=(\widehat{\psi}_{11}-\widehat{\psi}_{12})$ and $\sigma_\beta^2=\sigma_{11}^2+\sigma_{12}^2-2cov\left(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)},\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)}\right)$.

For the **nonlinear** link function, we use a Taylor expansion

$$h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)}) \;=\; h^{-1}(\boldsymbol{\theta}_{k\ell})+[h^{-1}(\boldsymbol{\theta}_{k\ell})]'\left(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)}-\boldsymbol{\theta}_{k\ell}\right)+\text{smaller order terms} \tag{A.73}$$

and thus we have

$$h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)})-h^{-1}(\boldsymbol{\theta}_{k\ell}) \;=\; [h^{-1}(\boldsymbol{\theta}_{k\ell})]'\left(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)}-\boldsymbol{\theta}_{k\ell}\right)+\text{smaller order terms} \tag{A.74}$$

which implies that

$$n\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)})-h^{-1}(\boldsymbol{\theta}_{k\ell})\right)\xrightarrow{d}N\left(\tilde{\widehat{\psi}}_{k\ell},\tilde{\widehat{\sigma}}^2_{k\ell}\right), \tag{A.75}$$

where $\tilde{\widehat{\psi}}_{k\ell}=[h^{-1}(\boldsymbol{\theta}_{k\ell})]'\widehat{\psi}_{k\ell}$ and $\tilde{\widehat{\sigma}}^2_{k\ell}=([h^{-1}(\boldsymbol{\theta}_{k\ell})]')^2\sigma^2_{k\ell}$.

We therefore obtain the result

$$n(\widehat{\beta}-\beta)\xrightarrow{d}N(\widehat{\psi}_\beta,\widehat{\sigma}^2_\beta) \tag{A.76}$$

where $\widehat{\psi}_\beta=\tilde{\widehat{\psi}}_{11}-\tilde{\widehat{\psi}}_{12}$ and $\widehat{\sigma}^2_\beta=\tilde{\widehat{\sigma}}^2_{11}+\tilde{\widehat{\sigma}}^2_{12}-2\tilde{\widehat{\sigma}}_{11,12}$ and

$$\tilde{\widehat{\sigma}}_{11,12}:=cov\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(1)}),h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)})\right)=\sigma_{11,12}\left[\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(1)\phi(1)})\right)'\right]\left[\left(h^{-1}(\widehat{\boldsymbol{\theta}}_{\phi(1),\phi(2)})\right)'\right] \tag{A.77}$$

A.4. **Proof of THEOREM 2.** In the semi-sparse regime the proof follows the same steps as the proof of Theorem 1. The only difference is in the bias terms, variances, and covariance terms, that are computed according to the following lemma.

LEMMA A.3. *(Corollary 2 extension for semi-sparse case in Tang et al. (2017)). Let $\boldsymbol{A}\sim SBM(\boldsymbol{\xi},\boldsymbol{\theta},\rho_n)$ be a $\widetilde{K}$-block stochastic blockmodel adjacency matrix on n vertices with sparsity factor ρ_n. Let $\mu_1,\ldots,\mu_{\widetilde{K}}$ be the center of the blocks in $\mathbb{R}^d$ and let $\boldsymbol{\theta}_{k\ell}=\boldsymbol{\mu}_k^T\boldsymbol{I}_{d_1,d_2}\boldsymbol{\mu}_\ell$ be the probability of a link between nodes in blocks k and ℓ. Let K be known and let $\widehat{\boldsymbol{\xi}}:[n]\to[\tilde{K}]$ be the function that assigns nodes to clusters, estimated using GMM or K-means clustering on the rows of $\widehat{\boldsymbol{Y}}=\widehat{\boldsymbol{U}}|\widehat{\boldsymbol{S}}|^{1/2}$ (as in the proof of Theorem A.1). Let $\widehat{\boldsymbol{\theta}}_{k\ell}=\widehat{\boldsymbol{\mu}}_k^T\mathbf{I}_{d_1,d_2}\widehat{\boldsymbol{\mu}}_\ell$ and let $\widehat{\boldsymbol{\Delta}}=\sum_{k=1}^K\widehat{\eta}_k\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_\ell^T$. Define $\Delta=\sum_{k=1}^K\eta_k\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$ and let $\zeta_{k\ell}=\boldsymbol{\mu}_k^T\Delta^{-1}\boldsymbol{\mu}_\ell$. Define $\widetilde{\sigma}^2_{kk}$ for $k\in[K]$ to be*

$$\widetilde{\sigma}^2_{kk} \;=\; 4\boldsymbol{\theta}_{kk}\zeta^2_{kk}+4\sum_{r=1}^K\eta_r\boldsymbol{\theta}_{kr}\zeta^2_{kr}\left(\frac{1}{\eta_k}-2\zeta_{kk}\right)^2+2\sum_{r=1}^K\sum_{s=1}^K\eta_r\eta_s\boldsymbol{\theta}_{rs}\zeta^2_{kr}\zeta^2_{ks} \tag{A.78}$$

and define $\widetilde{\sigma}^2_{k\ell}$ *for* $k \in [\tilde{K}]$ *and* $\ell \in [\tilde{K}]$, $k \neq \ell$ *to be*

$$\widetilde{\sigma}^2_{k\ell} = (\boldsymbol{\theta}_{kk} + \boldsymbol{\theta}_{\ell\ell})\,\zeta^2_{kk} + 2\boldsymbol{\theta}_{k\ell}\zeta_{kk}\zeta_{\ell\ell} + \sum_{r=1}^{K} \eta_r \boldsymbol{\theta}_{kr}\zeta^2_{\ell r}\left(\frac{1}{\eta_k} - 2\zeta_{kk}\right) \tag{A.79}$$

$$+ \sum_{r=1}^{K} \eta_r \boldsymbol{\theta}_{\ell r}\zeta^2_{kr}\left(\frac{1}{\eta_\ell} - 2\zeta_{\ell}\right) - 2\sum_{r=1}^{K} \eta_r\,(\boldsymbol{\theta}_{kr} + \boldsymbol{\theta}_{\ell r})\,\zeta_{kr}\zeta_{r\ell}\zeta_{k\ell} \tag{A.80}$$

$$+ \frac{1}{2}\sum_{r=1}^{K}\sum_{s=1}^{K} \eta_r\eta_s\boldsymbol{\theta}_{rs}\left(\zeta_{kr}\zeta_{\ell s} + \zeta_{\ell r}\zeta_{ks}\right)^2. \tag{A.81}$$

Let $\ddot{\psi}_{k\ell}$ *be defined as*

$$\ddot{\psi}_{k\ell} = \sum_{r=1}^{K} \widehat{\eta}_r\left(\widehat{\boldsymbol{\theta}}_{kr} + \widehat{\boldsymbol{\theta}}_{\ell r}\right)\widehat{\boldsymbol{\mu}}_k^T\widehat{\Delta}^{-1}\widehat{\boldsymbol{\mu}}_\ell \tag{A.82}$$

$$- \sum_{r=1}^{K}\sum_{s=1}^{K} \widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_s\widehat{\boldsymbol{\theta}}_{sr}\widehat{\boldsymbol{\mu}}_s^T\widehat{\Delta}^{-1}\mathbf{I}_{d_1,d_2}\widehat{\Delta}^{-1}\left(\widehat{\boldsymbol{\mu}}_\ell\widehat{\boldsymbol{\mu}}_k^T + \widehat{\boldsymbol{\mu}}_l\widehat{\boldsymbol{\mu}}_\ell^T\right)\widehat{\Delta}^{-1}\widehat{\boldsymbol{\mu}}_s. \tag{A.83}$$

Then there exists a (sequence of) permutation(s) $\phi \equiv \phi_n$ *on* $[K]$ *such that for any* $k \in [K]$ *and* $\ell \in [K]$,

$$n\rho_n^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\phi(k),\phi(\ell)} - \boldsymbol{\theta}_{k\ell} - \frac{\ddot{\psi}_{k\ell}}{n\rho_n}\right) \xrightarrow{d} N(0, \widetilde{\sigma}_{k\ell}) \tag{A.84}$$

as $n \to \infty$, $\rho_n \to 0$, *and* $n\rho_n = \omega(\sqrt{n})$.

Carey Business School, Johns Hopkins University, 100 International Dr., Baltimore, MD 21202

Department of Sociology, Johns Hopkins University, 3400 North Charles St., Baltimore, MD 21218

Department of Statistics, University of Michigan, 1085 South University Ave., Ann Arbor, MI 48109

Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N Charles St., Baltimore, MD 21218