



Spectral Estimation of Large Stochastic Blockmodels with Discrete Nodal Covariates

Angelo Mele, Lingxin Hao, Joshua Cape & Carey E. Priebe

To cite this article: Angelo Mele, Lingxin Hao, Joshua Cape & Carey E. Priebe (2022): Spectral Estimation of Large Stochastic Blockmodels with Discrete Nodal Covariates, Journal of Business & Economic Statistics, DOI: [10.1080/07350015.2022.2139709](https://doi.org/10.1080/07350015.2022.2139709)

To link to this article: <https://doi.org/10.1080/07350015.2022.2139709>



View supplementary material [↗](#)



Published online: 15 Nov 2022.



Submit your article to this journal [↗](#)



Article views: 73



View related articles [↗](#)





CrossMark

View Crossmark data [↗](#)



Spectral Estimation of Large Stochastic Blockmodels with Discrete Nodal Covariates

Angelo Mele^a , Lingxin Hao^b, Joshua Cape^c , and Carey E. Priebe^d

^aCarey Business School, Johns Hopkins University, Baltimore, MD; ^bDepartment of Sociology, Johns Hopkins University, Baltimore, MD; ^cDepartment of Statistics, University of Wisconsin–Madison, Madison, WI; ^dDepartment of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD

ABSTRACT

In many applications of network analysis, it is important to distinguish between observed and unobserved factors affecting network structure. We show that a network model with discrete unobserved link heterogeneity and binary (or discrete) covariates corresponds to a stochastic blockmodel (SBM). We develop a spectral estimator for the effect of covariates on link probabilities, exploiting the correspondence of SBMs and generalized random dot product graphs (GRDPG). We show that computing our estimator is much faster than standard variational expectation–maximization algorithms and scales well for large networks. Monte Carlo experiments suggest that the estimator performs well under different data generating processes. Our application to Facebook data shows evidence of homophily in gender, role and campus-residence, while allowing us to discover unobserved communities. Finally, we establish asymptotic normality of our estimators.

ARTICLE HISTORY

Received June 2021
Accepted October 2022

KEYWORDS

Asymptotic normality; Conditionally independent links; Generalized random dot product graphs; Large networks; Spectral estimation; Stochastic blockmodel

1. Introduction

The analysis and modeling of network data has far-reaching applications in economics, sociology, public health, computer science, neuroscience, and marketing, among other areas. Social networks have been shown to affect socioeconomic performance, such as education (Calvo-Armengol, Patacchini and Zenou 2009; DeGiorgi, Pellizzari, and Redaelli 2009; Carrell, Sacerdote, and West 2013), health and risky behaviors (Nakajima 2007; Badev *forthcoming*), risk sharing arrangements (Fafchamps and Gubert 2007), and employment opportunities (Beaman 2012). Often, both observed and unobserved factors contribute to the global structure of networks and the processes that generate them (Graham 2017; Graham and dePaula 2020; Mele 2022). For example, in social networks, factors including gender, race, and personality affect the likelihood that two people interact. In many applications, race and gender are typically observed, while personality is usually unobserved. It is therefore crucial to develop ways to disentangle the effect of observed and unobserved variables on link formation in networks.

In this article, we provide a simple estimation method for network models with conditionally independent links, where the probability of a link depends on observable covariates and unobservable link effects (Graham 2017; Dzemski 2017; Jochmans 2017). Our approach is mostly tailored to the estimation of random effects models with binary or discrete covariates, where the unobservable heterogeneity of nodes also has discrete support. We provide a computationally efficient estimator for the effect of discrete covariates and unobserved heterogeneity on link probabilities, by leveraging the fact that our model corresponds to popular models in the network science literature,

namely the Stochastic Blockmodel (SBM) and the Generalized Random Dot Product Graph (GRDPG) (Athreya et al. 2018; Rubin-Delanchy et al. *forthcoming*). Our algorithm is based on spectral methods recently developed in the applied probability and statistics literature (Athreya et al. 2018; Tang, Cape, and Priebe 2022; Rubin-Delanchy et al. *forthcoming*) and works well when the network is large and the number of support points for the unobservable heterogeneity is not too large.

The analysis of network models poses several econometric challenges, arising from the structure of correlation among links (Chandrasekhar 2016; DePaula 2017; Dzemski 2017; Graham 2020; Graham and dePaula 2020). In this article, we focus on models with conditionally independent links, where the unobserved heterogeneity is modeled by latent positions in low-dimensional Euclidean space (Graham 2017; Auerbach 2019). While this formulation rules out externalities and strategic considerations in link formation, which are dominant focal points in much of the econometric literature on strategic network formation games (DePaula 2017; Mele 2017; Menzel 2017; Graham 2020; Graham and dePaula 2020), there is a growing literature showing how models with conditionally independent links provide useful approximations of strategic models, in special instances of interest (Diaconis and Chatterjee 2013; Mele 2017; Graham 2020; Mele and Zhu *forthcoming*).

Our model assumes that the marginal surplus generated by a link can be separated in a component that depends on binary or discrete observable characteristics of the individuals, and an unobservable part that includes an iid shock and a link-specific unobservable parameter, which we assume can take on a finite number of possible values. While this setup is

similar to other models in the literature with additive unobservable heterogeneity (Graham 2017; Dzemski 2017; Jochmans 2017), our approach considers nonadditive unobservables. Previous models aim at capturing the degree heterogeneity in the network; for example, in Graham (2017) the observable part models homophily, while the additive unobservable part models degree heterogeneity. In our model both the observable and unobservable part capture homophily (and heterophily) as well as degree heterogeneity. Our estimator moves away from the additive modeling in Graham (2017) and provides a simpler, computationally faster estimator than the nonparametric approach in Zeleneev (2020), instead focusing on a more restrictive specification.

A crucial step of our approach is showing that our model with discrete unobserved heterogeneity and binary or discrete covariates corresponds to a Stochastic Block Model (SBM), a very popular model in the network science literature (Nowicki and Snijders 2001; Airoldi et al. 2008; Abbe 2018). In standard K -block SBMs, each node belongs to one of K unobserved blocks (communities); conditional on the block assignments, links form independently as Bernoulli variables with probabilities that depend 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 (Choi, Wolfe, and Airoldi 2011; Sweet 2015; Roy, Atchade, and Michailidis 2019). A possible reason is that estimation for stochastic blockmodels is computationally burdensome, and including covariates in the specification imposes significant additional challenges to modeling, estimation, and inference. Exact maximum likelihood estimation is infeasible, causing most estimation strategies to rely on approximations based on expectation–maximization algorithms and variational methods (Airoldi et al. 2008; Daudin, Picard, and Robin 2008; Latouche, Birmele, and Ambroise 2012; Bickel et al. 2013). However, these algorithms may converge slowly to the (approximate) solution and become impractical for networks with thousands of nodes.¹

The development of our estimator depends crucially on several observations. First, in a random effect setting, a K -block stochastic blockmodel with one binary covariate can be reformulated as a (different yet related) $2K$ -block SBM. Analogously, a K -block SBM with two binary covariates can be reformulated as a $4K$ -block SBM; and so on. Second, a stochastic blockmodel graph is a generalized random dot product graph (GRDPG) whose latent positions are fixed within blocks (Athreya et al. 2018; Tang and Priebe 2018; Tang, Cape, and Priebe 2022; Mu et al. 2022; Rubin-Delanchy et al. forthcoming). In a GRDPG, each node is characterized by an unobserved latent position (vector), and each pair of nodes links with probability determined via a (possibly indefinite) inner product of the pair's latent positions. Our estimation method is tied to the asymptotic behavior of spectral estimators for SBM block probability matrix entries recently studied in Tang, Cape, and Priebe (2022) and

have been shown to be successful both in terms of feasibility and scalability in related settings (Zheng et al. 2017).

Our theoretical machinery used to perform estimation and inference both applies and extends methods developed for the analysis of latent positions network models (Rohe, Chatterjee, and Yu 2011; Athreya et al. 2018; Tang, Cape, and Priebe 2022). 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, Cape, and Priebe 2022). Furthermore, we prove asymptotic normality for the parameter measuring the effect of the covariates on the link probabilities, extending previous results in (Tang, Cape, and Priebe 2022). These results provide a first step in the direction of practical inference. We show that our estimator is asymptotically normal as long as the parameter(s) for covariate effects can be written as sufficiently well-behaved functions of the SBM block-specific probabilities, and the estimates are asymptotically unbiased. However, a plug-in estimator using our formula for the variance is impractically conservative (see simulations in Appendix, supplementary materials).

In our simulations (see Section 4) we compare our approach to the variational EM (VEM) algorithm (Daudin, Picard, and Robin 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. 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 min to estimate the model using the (parallelized) VEM algorithm. When we add a binary covariate, our estimator converges in under 30 sec, while in contrast it takes almost 10 hr when using the VEM algorithm. Our methods are implemented in the R package `grdpg`, which we made available in the supplementary materials. All replication files can be found in the supplementary materials.

We perform a Monte Carlo study to examine the performance of our spectral estimator. We simulate networks in which the block structure is unbalanced (blocks have different sizes); we include scenarios in which the covariates are binary and independent, as well as cases in which the covariates are correlated. Our results confirm the existence of bias in small samples, but the bias decreases in magnitude when the size of the network increases. As expected, the bias is larger when the blocks are unbalanced and the covariates are correlated. These insights confirm that our estimator works best in very large networks, where the bias problem is attenuated, thus, adding to the advantage of computational speed.

Finally, we apply our method to the study of Facebook friendship data using the Facebook 100 dataset, initially collected and analyzed in Traud, Mucha, and Porter (2012). These data contain networks of friendships and node (person) information

¹Recent advances use further approximations and parallelization to improve computational efficiency (Roy, Atchade, and Michailidis 2019; Vu, Hunter, and Schweinberger 2013). We do not pursue such extensions in this article.

for 100 universities in the United States in the year 2005. We estimate a stochastic blockmodel for the Harvard University network, consisting of more than 13,000 nodes, using information on gender, off-campus residence, and university role (i.e., student, faculty, staff, etc.) of the users, with results similar to Roy, Atchade, and Michailidis (2019). We find evidence of homophily, as suggested by the positive effect of gender, role, and off-campus residence on the probability of linking. Qualitatively similar conclusion are obtained when using the joint fixed effect (JFE) estimator of Graham (2017) (see Appendix B, supplementary materials). This suggests that including information about observable covariates in the estimation may allow researchers to better recover unobservable community structure.

We acknowledge several limitations. First, the estimator works well in a random effects setup but is not necessarily suitable for fixed-effects settings. Consider a model with K unobserved blocks and two binary covariates. In a random effect approach, this model would correspond to a GRDPG-SBM model with $4K$ blocks. However, in a fixed effect setting, there is the possibility that one or more blocks are homogeneous by one observable covariate. This implies that the GRDPG model to estimate would have between K and $4K$ blocks, thus, increasing the complexity of estimation. Second, the algorithm is applicable in large networks. In smaller networks, our procedure may prove to be impractical, especially when K is relatively large and when we have many discrete covariates with large support. Finally, the algorithm requires knowledge of the number of blocks to perform estimation. While there are many ways to estimate the number of blocks, we suggest to try different estimated values in increasing order and estimate a Gaussian Mixture Model, choosing the best fit according to BIC (Fraley and Raftery 1999). Such a choice is motivated by asymptotic results for GRDPGs and SBMs in previous work (Athreya et al. 2018; Cape, Tang, and Priebe 2018; Tang, Cape, and Priebe 2022).

2. Background and Methodology

We consider the situation where the researcher has access to data on a single network and possibly to observable covariates for each node. Network formation is modeled as a process where links form independently, after conditioning on unobserved and observed heterogeneity. The existence of a link A_{ij} between two individuals i and j in the network is modeled as follows

$$A_{ij} = \mathbf{1} \left\{ \mathbf{B}_{ij} + \beta \mathbf{1}_{\{Z_i=Z_j\}} - \varepsilon_{ij} \geq 0 \right\}, \quad (1)$$

where A_{ij} is the entry of adjacency matrix, describing the existence of a link between i and j ; the parameter \mathbf{B}_{ij} is a random or fixed effect that relates to unobserved heterogeneity; Z_i and Z_j are observed covariates for i and j , which we assume to be binary or discrete; β is a parameter that measures the effect of observed covariates on the probability of a link; and ε_{ij} is an error term. Throughout the article we assume that ε_{ij} are iid and have full support. In most applications of our approach we use a logistic model, but other distributions are possible as well.

In economic terms, the quantity $\mathbf{B}_{ij} + \beta \mathbf{1}_{\{Z_i=Z_j\}} - \varepsilon_{ij}$ can be thought of as the (marginal) surplus generated by a link between i and j , as in simple random utility models of network formation

(DePaula 2017; Mele 2017; Graham 2020; Graham and dePaula 2020). The parameter β is interpretable as the marginal effect of homophily in observables: if $\beta > 0$ there is homophily in observables, making links between similar people more likely; vice versa, if $\beta < 0$ then we have heterophily and links are more likely among people who are different in their Z_i 's.

In previous work, researchers have used a specification with $\mathbf{B}_{ij} = g(\mathbf{X}_i, \mathbf{X}_j)$ where the unobserved heterogeneity is modeled as a function g of the latent vectors \mathbf{X}_i and \mathbf{X}_j in low-dimensional Euclidean space. A possible specification is $\mathbf{B}_{ij} = \mathbf{X}_i + \mathbf{X}_j$, where the unobserved heterogeneity is additive and the latent positions are scalars (Graham 2017; Dzemski 2017). This specification corresponds to the so-called β -model with covariates analyzed in Graham (2017); the model generates heterogeneous degrees for the nodes.

Another possible specification allows modeling of homophily driven by unobservables, with $\mathbf{B}_{ij} = \text{dist}(\mathbf{X}_i, \mathbf{X}_j)$, where the \mathbf{X}_i 's are vectors of coordinates in an Euclidean space and dist is a distance metric. For example, one can use an Euclidean distance, with $\mathbf{B}_{ij} = \|\mathbf{X}_i - \mathbf{X}_j\|$ (Fraley and Raftery 1999; Hoff, Raftery, and Handcock 2002). Alternatively, we can consider a specification based on angular distance, with $\mathbf{B}_{ij} = \mathbf{X}_i^\top \mathbf{X}_j$.

The model in (1) with $\beta = 0$ (i.e., no covariate effects) and $\mathbf{B}_{ij} = \mathbf{X}_i^\top \mathbf{X}_j$ is known in the network science literature as a *Random Dot Product Graph (RDPG)* (Athreya et al. 2018). In such a framework, links are more likely to occur when the angular distance of latent positions for nodes i and j is smaller, and the probability of a link is given by the dot product of the latent positions, that is, $P_{ij} := P(A_{ij} = 1 | \mathbf{X}_i, \mathbf{X}_j) = h(\mathbf{X}_i^\top \mathbf{X}_j)$, for a known function h . The RDPG generates assortative networks, but there are generalizations that make this framework applicable to more general networks, as we now discuss below.

2.1. Generalized Random Dot Product Graphs and Stochastic Blockmodels

The *Generalized Random Dot Product Graph (GRPDG)* model is a latent space model of network formation with conditionally independent links. It extends the random dot product graph model to allow for disassortative networks. In a GRPDG, each node i is characterized by a d -dimensional vector (i.e., an *unobserved* latent position) $\mathbf{X}_i = (X_{i1}, \dots, X_{id}) \in \mathcal{X}_d \subseteq \mathbb{R}^d$. The latent positions are iid draws from a distribution F with support \mathcal{X}_d , that is $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} F$. Let $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]^\top$ denote the matrix formed by row-wise stacking all unobserved vectors \mathbf{X}_i .

Let $d_1 \geq 1$ and $d_2 \geq 0$ be integers, and define $d = d_1 + d_2$. Let \mathbf{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 GRPDG with signature (d_1, d_2) , the entries of the adjacency matrix A_{ij} are specified to be independent, after conditioning on the latent positions \mathbf{X}_i and \mathbf{X}_j , namely

$$A_{ij} | \mathbf{X}_i, \mathbf{X}_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}(P_{ij}), \quad (2)$$

with link probability given by the (indefinite) dot product

$$P_{ij} = \mathbf{X}_i^\top \mathbf{I}_{d_1, d_2} \mathbf{X}_j. \quad (3)$$

For this setting, we write $(X, A) \sim \text{GRDPG}_{d_1, d_2}(F)$.²

A K -block *Stochastic Blockmodel* (SBM) is a special case of a GRDPG, where each latent position vector X_i can only take values from $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K]$. Two nodes i and j belong to the same block k if $X_i = X_j = \mathbf{v}_k$. We use the variable τ_i to indicate the block of node i , so that node i belongs to block k if $\tau_i = k$. These random variables τ are such that $\tau_1, \dots, \tau_n \stackrel{\text{iid}}{\sim} \text{Multinomial}(1; \pi_1, \dots, \pi_K)$, with $\pi \in (0, 1)^K$ and $\sum_{k=1}^K \pi_k = 1$ ³ and the probability of a link between i and j is $P_{ij} = \mathbf{v}_{\tau_i}^\top \mathbf{I}_{d_1, d_2} \mathbf{v}_{\tau_j}$. This corresponds to a block structure where a link between a node i in block k and another node j in block ℓ forms independently with probability $\theta_{k\ell} = \mathbf{v}_k^\top \mathbf{I}_{d_1, d_2} \mathbf{v}_\ell$. After collecting all these probabilities in the matrix θ , we denote the stochastic blockmodel as $(A, \tau) \sim \text{SBM}(\theta, \pi)$.

Vice versa, the GRDPG corresponding to the SBM model $(A, \tau) \sim \text{SBM}(\theta, \pi)$ can be obtained by an eigendecomposition of the matrix $\theta = \mathbf{U}\Sigma\mathbf{U}^\top$ and by defining $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K$ as the rows of $\mathbf{U}|\Sigma|^{1/2}$. The distribution F is $F = \sum_{k=1}^K \pi_k \delta_{\mathbf{v}_k}$, where δ is the Dirac-delta; importantly, d is the rank of the block-probabilities matrix θ , and d_1, d_2 are the number of positive and negative eigenvalues of matrix θ , respectively.

This article uses spectral methods for GRDPGs to estimate SBMs (Athreya et al. 2018; Tang, Cape, and Priebe 2022). The correspondence noted above between SBMs and GRDPGs also holds for *known* link functions and $\theta_{\tau_i \tau_j} = h(\mathbf{B}_{\tau_i \tau_j})$, where h is a known function that maps to $[0, 1]$ and \mathbf{B} is a $K \times K$ matrix of real numbers. For example, h could be the logistic function or the cumulative density function of the Gaussian distribution. Our stochastic blockmodel would have adjacency matrix A with elements

$$A_{ij} | \tau_i, \tau_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}\left(h(\mathbf{B}_{\tau_i \tau_j})\right). \quad (5)$$

It follows that the SBM is a special case of model (1), where $\beta = 0$ and $\mathbf{B}_{ij} = \mathbf{B}_{\tau_i \tau_j}$, where the function h is determined by the density of the error terms ε_{ij} .

The stochastic blockmodel can be extended to include the effect of observed covariates (Choi, Wolfe, and Airolidi 2011; Sweet 2015; Roy, Atchade, and Michailidis 2019). Let node i be characterized by a binary covariate $\mathbf{Z}_i \in \{0, 1\}$ and let the stochastic blockmodel be⁴

$$A_{ij} | \tau_i, \tau_j, \mathbf{Z}_i, \mathbf{Z}_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \quad (6)$$

$$\mathbf{P}_{ij} = h(\mathbf{B}_{\tau_i \tau_j} + \beta \mathbf{1}_{\{\mathbf{Z}_i = \mathbf{Z}_j\}}). \quad (7)$$

Therefore, the SBM with covariates, can be thought of as a special case of (1) with

$$A_{ij} = \mathbf{1}\left\{\mathbf{B}_{\tau_i \tau_j} + \beta \mathbf{1}_{\{\mathbf{Z}_i = \mathbf{Z}_j\}} - \varepsilon_{ij} \geq 0\right\}, \quad (8)$$

² It must be noted that the support X_d of F , is a subset of \mathbb{R}^d such that $\mathbf{x}^\top \mathbf{I}_{d_1, d_2} \mathbf{y} \in [0, 1]$ for all $\mathbf{x}, \mathbf{y} \in X_d$.

³ Alternatively, we can think of a network where the vectors X_i are drawn from a discrete mixture with mass centered at \mathbf{v} , that is,

$$X_i \sim \pi_1 \delta_{\mathbf{v}_1} + \pi_2 \delta_{\mathbf{v}_2} + \dots + \pi_K \delta_{\mathbf{v}_K}. \quad (4)$$

⁴ The extensions to multiple binary or discrete covariates is shown later in the article.

where $\mathbf{B}_{\tau_i \tau_j}$ has a block structure, and the density of the error term ε_{ij} determines the functional form of h .

Our goal is to develop a *spectral method for estimation* of β and $\mathbf{B}_{\tau_i \tau_j}$. To achieve this, we need to extend results from previous work on GRDPGs and SBMs (Athreya et al. 2018; Tang and Priebe 2018; Tang, Cape, and Priebe 2022). In the following sections we review some of the spectral methods we use in the article, and we provide an example that highlights the core aspects of our method.

2.2. 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, Picard, and Robin 2008; Airolidi et al. 2008); however, even these methods are computationally prohibitive for large networks.

We make use of spectral methods, which have been shown in the literature to scale well with network size. Our spectral approach embeds the network into a low(er) dimensional space, thus, reducing the dimensionality of the problem, while maintaining the geometric properties of the data. 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 ambient data space. The intuition about the spectral method is that if \mathbf{P} is a low-rank matrix, then we can view the adjacency matrix A as a perturbation of \mathbf{P} , that is $A_{ij} = \mathbf{P}_{ij} + E_{ij}$, where E_{ij} is a matrix of independent stochastic perturbations.⁵ If A and \mathbf{P} are *close enough*, namely if E is *small enough*, then the leading eigenvalues and eigenvectors of A and \mathbf{P} will be similar (Tang, Cape, and Priebe 2022). As a consequence, the spectral decomposition of A will provide an estimate of the latent structure of the network, that is, the latent positions X .

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

⁵ In the Bernoulli case, E_{ij} is a shifted Bernoulli variable, with values $E_{ij} = 1 - \mathbf{P}_{ij}$ with probability \mathbf{P}_{ij} and $E_{ij} = \mathbf{P}_{ij}$ with probability $1 - \mathbf{P}_{ij}$.

in absolute value and 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 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, Cape, and Priebe 2022). This implies that asymptotically the blocks are recovered exactly (up to relabeling, namely a permutation of the block labels). The same logic and results hold for nonpositive definite matrices P , allowing us to study more general stochastic blockmodels (Rubin-Delanchy et al. forthcoming).

2.3. 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 $K = 2$ stochastic blockmodel with a single binary covariate. The corresponding GRDPG has latent positions in 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.). In this example we can illustrate the geometry of the method in a low-dimensional space. The matrix B is given by

$$B = \begin{matrix} & \text{block}_1 & \text{block}_2 \\ \begin{matrix} \text{block}_1 \\ \text{block}_2 \end{matrix} & \begin{pmatrix} p^2 & pq \\ pq & q^2 \end{pmatrix} \end{matrix}, \quad (9)$$

where $p, q \in [0, 1]$. We can conveniently rewrite the matrix B as a dot-product of vector $\mathbf{v} = [p \ q]^\top$, with $p, q \in [0, 1]$, that is $B = \mathbf{v}\mathbf{v}^\top$, so that the SBM can be rewritten as a random dot-product graph model with $X_i = p$ if i is in block 1, $X_i = q$ if i is in block 2. The probability of linking is

$$P_{ij} = h(X_i^\top X_j + \beta \mathbf{1}_{\{Z_i=Z_j\}}). \quad (10)$$

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 ($Z_i = 0$); we assume random effects, so Z_i and X_i are independent.

The *crucial observation* is that the random effects model specified via (10) 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 covariate, we will have a $\tilde{K} = 2K$ -block SBM. The possible values of $X_i^\top X_j$ are $\{p^2, pq, q^2\}$. Therefore, the 4-block model can be completely characterized by the 4×4 matrix B_Z

$$B_Z = \begin{matrix} & \text{male}_1 & \text{female}_1 & \text{male}_2 & \text{female}_2 \\ \begin{matrix} \text{male}_1 \\ \text{female}_1 \\ \text{male}_2 \\ \text{female}_2 \end{matrix} & \begin{pmatrix} p^2 + \beta & p^2 & pq + \beta & pq \\ p^2 & p^2 + \beta & pq & pq + \beta \\ pq + \beta & pq & q^2 + \beta & q^2 \\ pq & pq + \beta & q^2 & q^2 + \beta \end{pmatrix} \end{matrix}. \quad (11)$$

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

$h(B_{Z,31}) = h(pq + \beta)$ 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 $P = YI_{d_1, d_2}Y^\top$, for some $n \times d$ matrix of latent positions 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 A , letting S_A denote the matrix whose diagonal contains the largest \hat{d} eigenvalues of A in absolute value and U_A denote the matrix whose columns are corresponding unit norm eigenvectors. We chose \hat{d} according to the profile likelihood method in Zhu and Ghodsi (2006). The Adjacency Spectral Embedding (ASE) of A gives an estimate of the latent positions of the 4-block model as

$$\hat{Y} = U_A |S_A|^{1/2}, \quad (12)$$

where $|\cdot|$ indicates the (entrywise) absolute value operation applied to a matrix.

2. We use a clustering procedure to assign each row of \hat{Y} to one of $\tilde{K} = 4$ blocks. We use a Gaussian Mixture Modeling approach (GMM) and estimate the centers of the clusters $\hat{\mu} = [\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4]$, that is, the means of the Gaussian components from the GMM. We choose \tilde{K} by running the algorithm for $K = 1, 2, \dots, K_{\max}$ and comparing the Bayesian Information Criterion (BIC) for each model.
3. We compute an estimate of θ_Z as

$$\hat{\theta}_Z = \hat{\mu}^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu} \quad (13)$$

$$= \begin{pmatrix} \text{male}_1 & \text{female}_1 & \text{male}_2 & \text{female}_2 \\ \text{male}_1 & \hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1 & \hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2 & \hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_3 & \hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_4 \\ \text{female}_1 & \hat{\mu}_2^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1 & \hat{\mu}_2^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2 & \hat{\mu}_2^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_3 & \hat{\mu}_2^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_4 \\ \text{male}_2 & \hat{\mu}_3^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1 & \hat{\mu}_3^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2 & \hat{\mu}_3^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_3 & \hat{\mu}_3^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_4 \\ \text{female}_2 & \hat{\mu}_4^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1 & \hat{\mu}_4^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2 & \hat{\mu}_4^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_3 & \hat{\mu}_4^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_4 \end{pmatrix}$$

By comparing the matrix $\hat{\theta}_Z$ and the population-level matrix θ_Z , we can assign each of the four blocks to the original two blocks. In fact, we know that the diagonal terms of $\hat{\theta}_Z$ 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 practice we use a GMM clustering algorithm to cluster the diagonal entries. In our example, the results of the diagonal clustering will group $\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1$ and $\hat{\mu}_2^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2$ in one block, and $\hat{\mu}_3^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_3$ and $\hat{\mu}_4^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_4$ in another block. Therefore, blocks 1 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.

4. We finally estimate β from the entries of the matrix $\hat{B}_Z = h^{-1}(\hat{\theta}_Z)$, where the inverse function h^{-1} is applied element-wise. For example, we know that $h^{-1}(\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1)$ 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 $h^{-1}(\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1)$ is an estimate of $p^2 + \beta$; therefore, the entry $h^{-1}(\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2)$ is an estimate of p^2 . A point estimate of β is then given by

$$\hat{\beta} = h^{-1}(\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_1) - h^{-1}(\hat{\mu}_1^\top I_{\hat{d}_1, \hat{d}_2} \hat{\mu}_2). \quad (14)$$

We can estimate β from many such pairs in the matrix $\hat{\mathbf{B}}_Z$ and our practical estimator is a weighted mean of such values.

5. The latent positions p and q can be estimated from the matrix $\hat{\mathbf{B}}_Z$ by using the submatrix

$$\begin{array}{cc} & \text{female}_1 & \text{female}_2 \\ \text{male}_1 & \left(h^{-1}(\hat{\boldsymbol{\mu}}_1^\top \mathbf{I}_{\hat{d}_1, \hat{d}_2} \hat{\boldsymbol{\mu}}_2) & h^{-1}(\hat{\boldsymbol{\mu}}_1^\top \mathbf{I}_{\hat{d}_1, \hat{d}_2} \hat{\boldsymbol{\mu}}_4) \right) \\ \text{male}_2 & \left(h^{-1}(\hat{\boldsymbol{\mu}}_3^\top \mathbf{I}_{\hat{d}_1, \hat{d}_2} \hat{\boldsymbol{\mu}}_2) & h^{-1}(\hat{\boldsymbol{\mu}}_3^\top \mathbf{I}_{\hat{d}_1, \hat{d}_2} \hat{\boldsymbol{\mu}}_4) \right) \end{array}$$

$$= \begin{pmatrix} \hat{p}^2 & \hat{p}q \\ \hat{p}q & \hat{q}^2 \end{pmatrix}. \quad (15)$$

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 estimate β from multiple entries of the matrix \mathbf{B}_Z , for example $\beta = \mathbf{B}_{Z,11} - \mathbf{B}_{Z,12} = \mathbf{B}_{Z,33} - \mathbf{B}_{Z,34}$, and weight each estimate by the size of the blocks, as shown in (45). This could improve the estimate, since some blocks are larger than others, consequently delivering more precise estimates. Our code implements this idea, which is helpful for empirical applications.⁶

2.4. Comparison with Variational EM

We compare our spectral methods to a standard algorithm used in the literature, the VEM algorithm, as implemented in the R package `blockmodels`. Our methods are implemented in the package `grdp` and are available as supplementary material, as well as the replication files for simulations and empirical application. We also note that the VEM algorithm uses parallelization to increase computational efficiency, while our method is implemented without any parallelization.

We consider a model with a binary nodal covariate, $\mathbf{Z}_i \sim \text{Bernoulli}(0.5)$ and

$$\text{logit}(\mathbf{P}_{ij}) = \mathbf{X}_i^\top \mathbf{X}_j + \beta \mathbf{1}_{\{\mathbf{Z}_i = \mathbf{Z}_j\}}. \quad (16)$$

In this example we use two blocks with equal probabilities, latent positions $[p, q] = [-1.5, 1]$ and $\beta = 1.5$, thus, the matrices $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_Z$ are

$$\text{logit}(\boldsymbol{\theta}) = \begin{bmatrix} 2.25 & -1.5 \\ -1.5 & 1 \end{bmatrix}; \quad (17)$$

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

We choose the latent position dimension \hat{d} via the profile likelihood method in (Zhu and Ghodsi 2006). In Figure 1 we show the screeplots corresponding to this estimation exercise. In the upper-left, we display the screeplot for the adjacency matrix, which suggests using $\hat{d} = 4$ as an estimate of the dimension

for $\hat{\mathbf{Y}}$. We note that the fourth largest eigenvalue (in magnitude) is negative, and the GRDPG model takes this into account. In the center-left plot, we show the screeplot of the adjacency matrix after *netting out the effect of the covariates*, which suggests the estimate $\hat{d} = 1$ for determining the dimension of the unobserved latent positions \mathbf{X} . For a network with $n = 2000$ nodes, the point estimates for $\boldsymbol{\theta}_Z$ (up to a permutation of the block labels) are respectively

$$\text{logit}(\hat{\boldsymbol{\theta}}_{Z, \text{VEM}}) \quad (18)$$

$$= \begin{bmatrix} 3.7443 & 2.2410 & -0.00367 & -1.5069 \\ 2.2410 & 3.7443 & -1.5069 & -0.0036 \\ -0.00367 & -1.5069 & 2.5013 & 0.9980 \\ -1.5069 & -0.0036 & 0.9980 & 2.5013 \end{bmatrix},$$

$$\hat{\mathbf{B}}_Z = \text{logit}(\hat{\boldsymbol{\theta}}_{Z, \text{GRDPG}}) \quad (19)$$

$$= \begin{bmatrix} 3.7762 & 2.2336 & -0.0062 & -1.5095 \\ 2.2336 & 3.7821 & -1.5007 & -0.0042 \\ -0.0062 & -1.5007 & 2.4979 & 0.9985 \\ -1.5095 & -0.0042 & 0.9985 & 2.5045 \end{bmatrix}.$$

According to our procedure, there are several ways to obtain an estimate of β . From matrix (19), 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 $\hat{\mathbf{B}}_{Z,11} - \hat{\mathbf{B}}_{Z,12}$ or $\hat{\mathbf{B}}_{Z,22} - \hat{\mathbf{B}}_{Z,21}$ or $\hat{\mathbf{B}}_{Z,33} - \hat{\mathbf{B}}_{Z,34}$. 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 $\hat{\mathbf{B}}_{Z,11} - \hat{\mathbf{B}}_{Z,12}$ is weighted by the proportion of links in the network that are used to estimate it.

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) (Hubert and Arabie 1985), and the computational time is given by the CPU time in seconds.

The point estimates for β reported in Table 1 are $\hat{\beta}_{\text{GRDPG}} = 1.51201$ and $\hat{\beta}_{\text{VEM}} = 1.50335$, for a network with $n = 2000$ nodes. 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 hr to obtain the VEM results, while it only takes 7.5 sec with our estimator. The left plots in Figure 1 show the latent positions $\hat{\mathbf{Y}}$ 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 four blocks, as our theory predicts. In the bottom-left plot in Figure 1 we display the estimated latent positions $\hat{\mathbf{X}}$, estimated by netting out the effect of the covariates. The figure shows how the estimated latent positions $\hat{\mathbf{X}}$ cluster around the true values p and q (the black vertical lines).

When we increase the size of the network to $n = 5000$, the estimated parameters are essentially the same for VEM and GRDPG. However, the GRDPG estimator takes less than 30 sec to compute the final result; the VEM estimate takes almost 10 hr. We ran more simulations and additional examples are shown in the Appendix, supplementary materials, confirming that our estimator is faster than VEM, while providing similar estimates.

In summary, our simple examples and simulations show that our GRDPG-based estimator is quite fast and scales well to large

⁶A possible alternative is to exploit the overidentification of β in the matrix $\hat{\boldsymbol{\theta}}_Z$ to develop a minimum distance estimator. We do not pursue this direction here, as we focus on a *fully* spectral estimator.

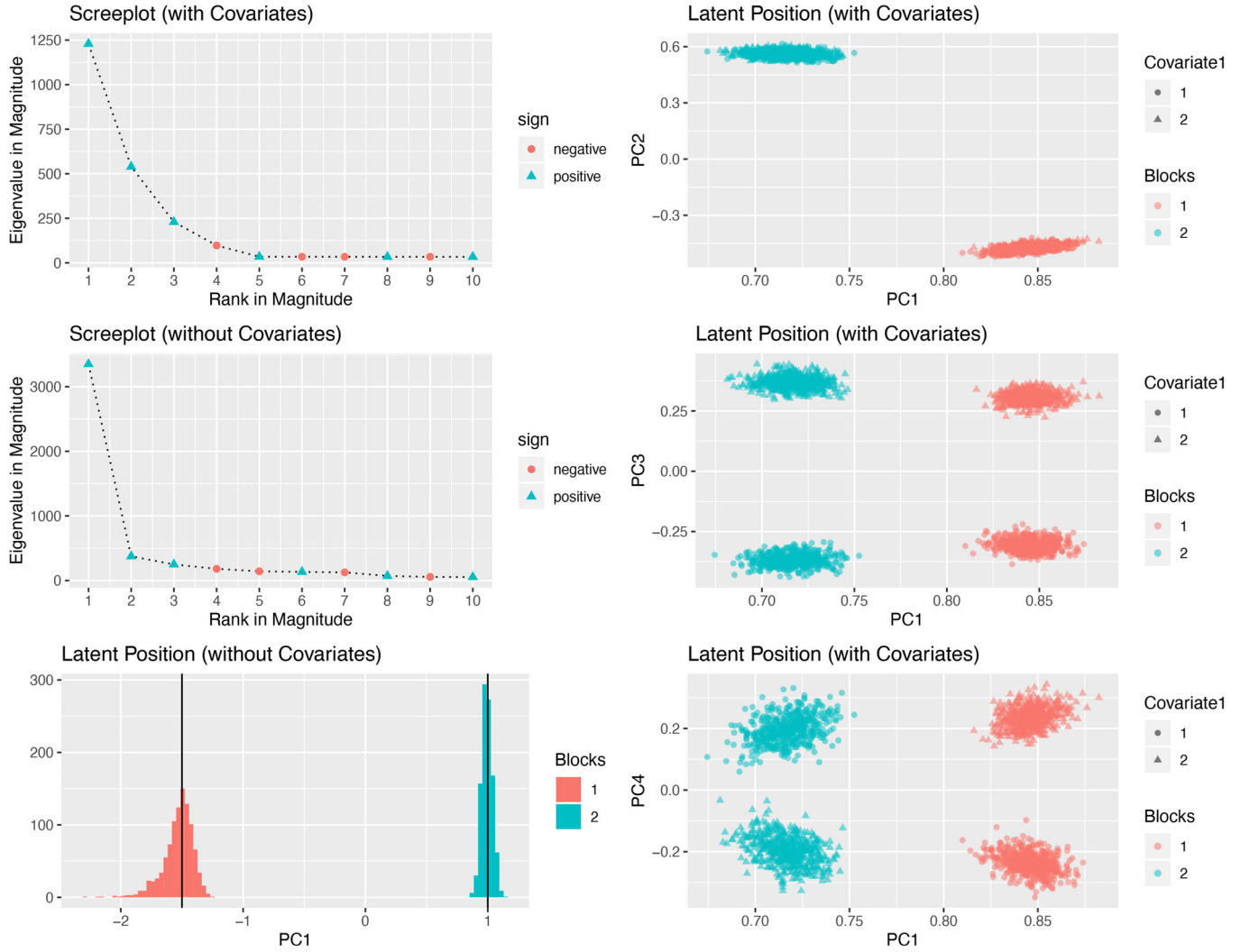


Figure 1. Estimation results for example.

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

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.

2.5. Multiple Observed Covariates

Consider the case with two covariates, $\mathbf{Z}_i = [\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}]$, assumed both binary for simplicity; let $\mathbf{Z}_i^{(1)} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(b_{\tau_i}^{(1)})$ and $\mathbf{Z}_i^{(2)} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(b_{\tau_i}^{(2)})$. The model is

$$\mathbf{A}_{ij} | \mathbf{X}_i, \mathbf{X}_j, \mathbf{Z}_i, \mathbf{Z}_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \quad (20)$$

$$\mathbf{P}_{ij} = h \left(\mathbf{X}_i^\top \mathbf{X}_j + \beta_1 \mathbf{1}_{\{\mathbf{Z}_i^{(1)} = \mathbf{Z}_j^{(1)}\}} + \beta_2 \mathbf{1}_{\{\mathbf{Z}_i^{(2)} = \mathbf{Z}_j^{(2)}\}} \right). \quad (21)$$

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

$$\boldsymbol{\theta}_Z = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \dots & \mathbf{W}_{1\tilde{K}} \\ \mathbf{W}_{21} & \mathbf{W}_{22} & \dots & \mathbf{W}_{2\tilde{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{\tilde{K}1} & \mathbf{W}_{\tilde{K}2} & \dots & \mathbf{W}_{\tilde{K}\tilde{K}} \end{bmatrix}, \quad (22)$$

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

	$\tau = \ell Z^{(1)} = 0 Z^{(2)} = 0$	$\tau = \ell Z^{(1)} = 1 Z^{(2)} = 0$	$\tau = \ell Z^{(1)} = 0 Z^{(2)} = 1$	$\tau = \ell Z^{(1)} = 1 Z^{(2)} = 1$
$\tau = k; Z^{(1)} = 0; Z^{(2)} = 0$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1 + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell)$
$\tau = k; Z^{(1)} = 1; Z^{(2)} = 0$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1 + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1)$
$\tau = k; Z^{(1)} = 0; Z^{(2)} = 1$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1 + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_2)$
$\tau = k; Z^{(1)} = 1; Z^{(2)} = 1$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_2)$	$h(\mathbf{v}_k^\top \mathbf{v}_\ell + \beta_1 + \beta_2)$

(23)

Table 1. Point estimates and CPU time for example 2 ($K = 2$).

Estimator	n	K	p	\hat{p}	q	\hat{q}	β	$\hat{\beta}$	CPU Time (sec)	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	10,000	2	-1.5	-1.49989	1	1.00029	No covariates		55.428	1
GRDPG	10,000	2	-1.5	1.49992	1	0.99992	1.5	1.50190	91.067	1

The intuition is the same as the model with one covariate. The blocks can be inferred by clustering the diagonal elements of matrix θ_Z , and the parameters β_1 and β_2 are functions of the θ_Z entries, namely

$$\beta_1 = h^{-1}(\theta_{Z,11}) - h^{-1}(\theta_{Z,12}); \quad (24)$$

$$\beta_2 = h^{-1}(\theta_{Z,11}) - h^{-1}(\theta_{Z,13}). \quad (25)$$

Other extensions are possible; for example, in Appendix, supplementary materials we show how to estimate differential homophily.

3. Asymptotic Theory

In this section, we derive a central limit theorem for the spectral estimator of β . We want to estimate the following model

$$\tau_i \stackrel{\text{iid}}{\sim} \text{Multinomial}(1; \pi_1, \dots, \pi_K), \quad (26)$$

$$\mathbf{Z}_i | \tau_i \stackrel{\text{ind}}{\sim} \text{Bernoulli}(b_{\tau_i}), \quad (27)$$

$$\mathbf{A}_{ij} | \tau_i, \tau_j, \mathbf{Z}_i, \mathbf{Z}_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}(P_{ij}), \quad (28)$$

$$P_{ij} = h(\mathbf{B}_{\tau_i \tau_j} + \beta \mathbf{1}_{\{Z_i = Z_j\}}). \quad (29)$$

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 θ_Z . Extensions to multiple binary or discrete observed covariates are straightforward albeit tedious. We assume that the observed covariates are binary and can depend on the block assignment,

that is $\mathbf{Z}_i | \tau_i \stackrel{\text{ind}}{\sim} \text{Bernoulli}(b_{\tau_i})$, where $b_{\tau_i} = P(\mathbf{Z}_i = 1 | \tau_i) > 0$. Our asymptotic results are easily extended to the case of discrete observed covariates with three or more possible outcomes.

We rewrite this model as a GRDPG. The matrix \mathbf{B} can be written as $\mathbf{B}_{\tau_i \tau_j} = \mathbf{X}_i^\top \mathbf{X}_j$, where \mathbf{X}_i is a $d \times 1$ vector of latent positions that has K possible values $\mathbf{v}_1, \dots, \mathbf{v}_K$. In practice, $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_K)$ are the centers of the K blocks \mathbf{X} , such that i and j belong to *unobserved* block k when $\mathbf{X}_i = \mathbf{X}_j = \mathbf{v}_k$. Let τ be the function that assigns nodes to unobserved blocks; then $\tau_i = k$ if $\mathbf{X}_i = \mathbf{v}_k$. We can thus rewrite the stochastic blockmodel above as a generalized random dot product graph with observed covariates:

$$\mathbf{X}_i \stackrel{\text{iid}}{\sim} \pi_1 \delta_{\mathbf{v}_1} + \pi_2 \delta_{\mathbf{v}_2} + \dots + \pi_K \delta_{\mathbf{v}_K}, \quad (30)$$

$$\mathbf{Z}_i | \mathbf{X}_i \stackrel{\text{ind}}{\sim} \text{Bernoulli}(b_{\tau_i}), \quad (31)$$

$$\mathbf{A}_{ij} | \mathbf{X}_i, \mathbf{X}_j, \mathbf{Z}_i, \mathbf{Z}_j \stackrel{\text{ind}}{\sim} \text{Bernoulli}(P_{ij}), \quad (32)$$

$$P_{ij} = h(\mathbf{X}_i^\top \mathbf{X}_j + \beta \mathbf{1}_{\{Z_i = Z_j\}}). \quad (33)$$

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

So we have a stochastic blockmodel $(\mathbf{A}, \xi, \mathbf{Z}) \sim \text{SBM}(\theta_Z, \eta)$ with the $\tilde{K} \times \tilde{K}$ matrix of probabilities θ_Z

$$\theta_Z = \begin{matrix} & \begin{matrix} \tau = 1; Z = 0 & \tau = 1; Z = 1 & \tau = 2; Z = 0 & \tau = 2; Z = 1 & \dots & \tau = K; Z = 0 & \tau = K; Z = 1 \end{matrix} \\ \begin{matrix} \tau = 1; Z = 0 \\ \tau = 1; Z = 1 \\ \tau = 2; Z = 0 \\ \tau = 2; Z = 1 \\ \vdots \\ \tau = K; Z = 0 \\ \tau = K; Z = 1 \end{matrix} & \begin{pmatrix} h(\mathbf{v}_1^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_1^\top \mathbf{v}_1) & h(\mathbf{v}_1^\top \mathbf{v}_2 + \beta) & h(\mathbf{v}_1^\top \mathbf{v}_2) & \dots & h(\mathbf{v}_1^\top \mathbf{v}_K + \beta) & h(\mathbf{v}_1^\top \mathbf{v}_K) \\ h(\mathbf{v}_1^\top \mathbf{v}_1) & h(\mathbf{v}_1^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_1^\top \mathbf{v}_2) & h(\mathbf{v}_1^\top \mathbf{v}_2 + \beta) & \dots & h(\mathbf{v}_1^\top \mathbf{v}_K) & h(\mathbf{v}_1^\top \mathbf{v}_K + \beta) \\ h(\mathbf{v}_2^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_2^\top \mathbf{v}_1) & h(\mathbf{v}_2^\top \mathbf{v}_2 + \beta) & h(\mathbf{v}_2^\top \mathbf{v}_2) & \dots & h(\mathbf{v}_2^\top \mathbf{v}_K + \beta) & h(\mathbf{v}_2^\top \mathbf{v}_K) \\ h(\mathbf{v}_2^\top \mathbf{v}_1) & h(\mathbf{v}_2^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_2^\top \mathbf{v}_2) & h(\mathbf{v}_2^\top \mathbf{v}_2 + \beta) & \dots & h(\mathbf{v}_2^\top \mathbf{v}_K) & h(\mathbf{v}_2^\top \mathbf{v}_K + \beta) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h(\mathbf{v}_K^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_K^\top \mathbf{v}_1) & h(\mathbf{v}_K^\top \mathbf{v}_2 + \beta) & h(\mathbf{v}_K^\top \mathbf{v}_2) & \dots & h(\mathbf{v}_K^\top \mathbf{v}_K + \beta) & h(\mathbf{v}_K^\top \mathbf{v}_K) \\ h(\mathbf{v}_K^\top \mathbf{v}_1) & h(\mathbf{v}_K^\top \mathbf{v}_1 + \beta) & h(\mathbf{v}_K^\top \mathbf{v}_2) & h(\mathbf{v}_K^\top \mathbf{v}_2 + \beta) & \dots & h(\mathbf{v}_K^\top \mathbf{v}_K) & h(\mathbf{v}_K^\top \mathbf{v}_K + \beta) \end{pmatrix} \end{matrix} \quad (34)$$

The stochastic blockmodel characterized by the matrix θ_Z can be reformulated as a GRDPG. Indeed, consider the eigen-decomposition $\theta_Z \equiv U \Sigma U^\top$, and define $\mu = [\mu_1, \mu_2, \dots, \mu_{\tilde{K}}]$ as the rows of $U|\Sigma|^{1/2}$; then let $F = \sum_{k=1}^{\tilde{K}} \eta_k \delta_{\mu_k}$, where δ is the Dirac-delta; and d_1 and d_2 are the number of positive and negative eigenvalues of θ_Z , respectively. Then, the Generalized Random Dot Product Graph model $(Y, A) \sim \text{GRDPG}_{d_1, d_2}(F)$ corresponding to our stochastic blockmodel $(A, \xi, Z) \sim \text{SBM}(\theta_Z, \eta)$ is given by

$$Y_i \stackrel{\text{iid}}{\sim} \eta_1 \delta_{\mu_1} + \dots + \eta_{\tilde{K}} \delta_{\mu_{\tilde{K}}}, \quad (35)$$

$$A_{ij} | Y_i, Y_j \stackrel{\text{iid}}{\sim} \text{Bernoulli}(Y_i^\top I_{d_1, d_2} Y_j), \quad (36)$$

where $d_1 + d_2 = \tilde{d} = \text{rank}(\theta_Z)$ and Y is the $n \times \tilde{d}$ vector of latent positions with centroids μ .

We can now extend asymptotic results for estimation of RDPGs in Athreya et al. (2018) and Tang, Cape, and Priebe (2022) to estimate block assignments and the effect of the covariates (see Rubin-Delanchy et al. (forthcoming) for the corresponding generalization to GRDPGs).

3.1. Main Theoretical Result

Because the functions τ that describe the assignments to blocks are unknown, the \tilde{K} SBM model assignment functions ξ are also unknown. Applying the Adjacency Spectral Embedding procedure, we recover an estimate $\hat{\xi}$.

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

$$\begin{aligned} \beta &= h^{-1}(\theta_{Z,11}) - h^{-1}(\theta_{Z,12}) \\ &= h^{-1}(\mathbf{v}_1^\top \mathbf{v}_1 + \beta) - h^{-1}(\mathbf{v}_1^\top \mathbf{v}_1). \end{aligned} \quad (37)$$

If the blocks were known at the onset, we could use the estimator $\hat{\beta} = h^{-1}(\hat{\theta}_{Z,11}) - h^{-1}(\hat{\theta}_{Z,12})$. However, all that we have access to is the estimate $\hat{\xi}$, 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, Cape, and Priebe 2022). 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 τ be unknown and K known. Let $\hat{\tau} : [n] \rightarrow [K]$ be the function that assigns nodes to clusters, estimated using GMM or K-means clustering on the rows of $\hat{Y} = \hat{U}|\hat{\Sigma}|^{1/2}$. Let function g be defined as the inverse of h , that is $g(\cdot) = h^{-1}(\cdot)$, with first derivative $g'(\cdot)$. Let $g'(\mathbf{v}_1^\top \mathbf{v}_1 + \beta) \neq 0$ and $g'(\mathbf{v}_1^\top \mathbf{v}_2) \neq 0$. Then there exists a sequence of permutations $\phi \equiv \phi_n$ on $[K]$ such that the estimator $\hat{\beta} = h^{-1}(\hat{\theta}_{Z, \phi(1)\phi(1)}) - h^{-1}(\hat{\theta}_{Z, \phi(1)\phi(2)})$ is asymptotically normal, that is

$$n \left(\hat{\beta} - \beta - \frac{\hat{\psi}_\beta}{n} \right) \xrightarrow{d} N(0, \tilde{\sigma}_\beta^2) \quad (38)$$

as $n \rightarrow \infty$. The bias term $\hat{\psi}_\beta$ and the asymptotic variance $\tilde{\sigma}_\beta^2$ are derived in the proof in equations (A.87) and (A.88), respectively.

Proof. See Appendix, supplementary materials. \square

3.2. Sparsity

Many social and economic networks of interest in applications display some degree of sparsity. Economists rationalize sparsity with the fact that people have constraints on time to spend with their friends (Jackson 2008). We follow the literature and assume that sparsity is an asymptotic feature of the data generating process. We multiply the probability 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

$$P_{ij} = \rho_n h \left(X_i^\top X_j + \beta \mathbf{1}_{\{Z_i=Z_j\}} \right). \quad (39)$$

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

For the central limit theorem we let $\rho_n \rightarrow 0$ as $n \rightarrow \infty$, with $n\rho_n = \omega(\sqrt{n})$, that is the average degree of the network grows sub-linearly in n .⁷ We will describe this regime a *semi-sparse*. The intuition is that too much sparsity, say below the order root- n regime, 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 (29) include a sparsity coefficient ρ_n where

$$P_{ij} = \rho_n h \left(X_i^\top X_j + \beta \mathbf{1}_{\{Z_i=Z_j\}} \right), \quad (40)$$

such that $\rho_n \rightarrow 0$ and $n\rho_n = \omega(\sqrt{n})$ as $n \rightarrow \infty$. Let $\hat{\tau}$ 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 $\hat{\beta} = h^{-1}(\hat{\theta}_{Z, \phi(1)\phi(1)}) - h^{-1}(\hat{\theta}_{Z, \phi(1)\phi(2)})$ is asymptotically normal, that is

$$n\rho_n^{1/2} \left(\hat{\beta} - \beta - \frac{\check{\psi}_\beta}{n\rho_n} \right) \xrightarrow{d} N(0, \check{\sigma}_\beta^2), \quad (41)$$

where the bias term $\check{\psi}_\beta$ and the asymptotic variance $\check{\sigma}_\beta^2$ are computed in (A.99) and (A.100), respectively.

Proof. See Appendix, supplementary materials. \square

Theorem 2 says that as long as the network is not too sparse, the estimator of β will be asymptotically normal. Notably, the bias term does not vanish asymptotically. Furthermore, we note that the formula for the standard errors is not a function of the sparsity term ρ_n .

3.3. Practical Details about the Estimators

Consider the model with one binary covariate in Theorem 1. Our central limit theorem focus on the differences of two entries of the matrix θ_Z . However, we can compute β in several

⁷The 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 $n \geq n_0$.

ways, using different entries of the matrix, for example, $\beta = h^{-1}(\theta_{Z,11}) - h^{-1}(\theta_{Z,12}) = h^{-1}(\theta_{Z,33}) - h^{-1}(\theta_{Z,34})$. Therefore, we rely on two ways to estimate the model. The first estimator consists of computing all the values of β from the relevant pairs of entries of $h^{-1}(\theta_Z)$ and then averaging out. The second estimator weights each estimated β by the size of the corresponding blocks. Formally, for the first estimator, after estimating the block assignments $\hat{\xi}$, we compute the number of nodes in the block with a particular value of the covariate $n_{1,k} = \sum_{i:\hat{\xi}_i=k} \mathbf{1}_{\{Z_i=1\}}$ and $n_{0,k} = \sum_{i:\hat{\xi}_i=k} \mathbf{1}_{\{Z_i=0\}}$, and we assign each block to a value of the covariate $Z_{\theta,k} = 1$ if $n_{1,k} > n_{0,k}$ and $Z_{\theta,k} = 0$ otherwise. Let $\psi = (\psi_1, \dots, \psi_{2K}) \in \{1, \dots, K\}^{2K}$ be the vector that assigns each element of the diagonal of θ_Z to the corresponding unobserved block; let $\hat{\psi}$ be the corresponding estimated assignment. We then consider the set of all pairs $M = \{(k\ell, k\ell'), k, \ell, \ell' \in \{1, 2, \dots, 2K\} | \hat{\psi}_\ell = \hat{\psi}_{\ell'}, Z_{\theta,k} = Z_{\theta,\ell}, Z_{\theta,k} \neq Z_{\theta,\ell'}\}$. For each element $m = (k\ell, k\ell') \in M$ we compute the estimate $\hat{\beta}_m$ as

$$\hat{\beta}_m = h^{-1}(\hat{\theta}_{Z,k\ell}) - h^{-1}(\hat{\theta}_{Z,k\ell'}). \quad (42)$$

We then average out the values of the $\hat{\beta}_m$ to obtain the final estimate

$$\hat{\beta}^{sa} = \frac{1}{|M|} \sum_{m \in M} \hat{\beta}_m, \quad (43)$$

where $|M|$ is the number of elements in set M , that is the number of paired entries in $h^{-1}(\hat{\theta}_Z)$ from which we can estimate β .

The second estimator weights each estimated parameter, $\hat{\beta}_{k\ell,k\ell'} = h^{-1}(\theta_{Z,k\ell}) - h^{-1}(\theta_{Z,k\ell'})$, by the size of the corresponding blocks used in its estimation. We compute the weight

$$\hat{\omega}_{k\ell,k\ell'} = \frac{n_{0,k}n_{0,\ell}n_{1,\ell'} + n_{1,k}n_{1,\ell}n_{0,\ell'}}{n_k n_\ell n_{\ell'}}, \quad (44)$$

where $n_k = \sum_{i=1}^n \mathbf{1}_{\{\hat{\xi}_i=k\}}$. We then consider the pairs in the set $\Omega = \{(\ell, \ell'), \ell, \ell' \in \{1, \dots, 2K\} | \hat{\psi}_\ell = \hat{\psi}_{\ell'}\}$ and estimate the weighted sum

$$\hat{\beta}^{wa} = \frac{1}{|\Omega|} \sum_{k=1}^{\hat{K}} \sum_{(\ell, \ell') \in \Omega} \hat{\omega}_{k\ell,k\ell'} \hat{\beta}_{k\ell,k\ell'}, \quad (45)$$

Both estimators work well in practice, as shown in the examples and simulations.

A note on the choice of K and d . The algorithm assumes that researchers know K and d . In practice we need to estimate both quantities. We suggest to choose d using the profile likelihood method of Zhu and Ghodsi (2006), as implemented in our replication package. We suggest to choose K using the BIC criterion of the `mclust` package in R (Fraley and Raftery 1999). In practice, once we estimate the latent positions, we try different values of K in increasing order; we then choose the one with the largest BIC.

4. Simulations and Empirical Application

4.1. Monte Carlo Experiments

Given the computational times shown in the previous section, we run a simple Monte Carlo to understand the potential bias

Table 2. Monte Carlo design.

Design	π_1	b_z	b_w	Correlation
1	0.5	0.5	0.5	Independent
2	0.5	0.5	0.5	0.3
3	0.3	0.5	0.5	Independent
4	0.3	0.4	0.6	Independent
5	0.3	0.4	0.6	0.3

In the Monte Carlo simulations we set the number of blocks $K = 2$ with centers $\mathbf{v} = (-1.5, 1)$; the true parameter vector for the covariates is $\beta = (0.5, 0.75)$ and we repeat the simulations for $n = 2000, 5000, 10,000$. We vary the probability of belonging to the first block π_1 , the probability of the Bernoulli covariates b_z and b_w , respectively for Z_i and W_i ; and we allow the covariates to have correlations in some of the simulations. Each Monte Carlo is the outcome of 1000 simulations.

of the spectral estimator in networks of moderate size.⁸ One of the potential advantages of our approach is that we have a formula for the bias term and could develop analytical correction methods, while it can be impractical to do the same for the VEM algorithm. We estimate a model with two binary observed covariates, generated as follows $Z_i \sim \text{Bernoulli}(b_z)$ and $W_i \sim \text{Bernoulli}(b_w)$ and vary the probabilities b_z and b_w , as well as the correlation among the two variables. We estimate the following model in each Monte Carlo design

$$\log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \mathbf{X}_i^\top \mathbf{X}_j + \beta_1 \mathbf{1}_{\{Z_i=Z_j\}} + \beta_2 \mathbf{1}_{\{W_i=W_j\}}. \quad (46)$$

The Monte Carlo design considers networks of size $n = 2000, 5000, 10,000$, and we set the number of blocks to be $K = 2$. For all the simulations, the parameter value that generates the data is $\beta = (0.5, 0.75)$ and the centers of the blocks are $\mathbf{v} = (-1.5, 1.0)$. We summarize the designs in Table 2. For each design and network size, we simulate 1000 networks and estimate the parameters of model (46) with the simple mean estimator and the weighted mean estimator.

The first design corresponds to the examples in the previous section. Blocks are assumed to have same membership probability and observables are independent Bernoulli variables with equal probability. The second design introduces correlation among the observables, as this is a realistic feature of many datasets. Designs 3 and 4 are intended to test the effect of unbalanced block size and unbalanced covariates, respectively, while maintaining the assumption of independence among observables. The final design assumes that we have unbalanced blocks, unbalanced covariates and correlated observables, allowing us to understand how the estimator behaves in a realistic setting. We expect that unbalancedness and correlation will increase bias, but this problem is less severe for larger networks, as our theory shows that the bias becomes vanishingly small as we increase the size of the network.

The results of our simulations are reported in Tables 3 (simple mean estimator) and 9 (weighted mean estimator). For each design, we report the absolute difference between the estimated parameter and the true value, the Monte Carlo standard error and the average time for estimation.⁹ When using the simple mean estimator, the estimates are precise, while displaying a

⁸We do not compare the spectral estimator to the variational EM estimator, because the latter is too slow for a Monte Carlo with 1000 repetitions, even after parallelizing the execution.

⁹The time of estimation reported in the table includes the following steps: (a) compute the ASE from the adjacency matrix; (b) compute the matrix

Table 3. Monte Carlo results, simple mean estimator (43).

	n	$ \hat{\beta}_1 - \beta_1 $	mcse	$ \hat{\beta}_2 - \beta_2 $	mcse	Time
Design 1	2000	0.0576	0.0001696	0.0350	0.0001366	12.35
	5000	0.0134	0.0000311	0.0082	0.0000281	31.68
	10,000	0.0016	0.0000093	0.0009	0.0000092	137.97
Design 2	2000	0.1118	0.0008083	0.0843	0.0007631	9.14
	5000	0.0240	0.0000475	0.0168	0.0000413	32.75
	10,000	0.0040	0.0000109	0.0026	0.0000104	140.20
Design 3	2000	0.1683	0.0008531	0.1039	0.0006886	12.71
	5000	0.0201	0.0000424	0.0089	0.0000379	52.91
	10,000	0.0015	0.0000140	0.0011	0.0000124	142.54
Design 4	2000	0.2144	0.0016978	0.1697	0.0014046	23.83
	5000	0.0399	0.0001094	0.0228	0.0000921	51.71
	10,000	0.0046	0.0000162	0.0016	0.0000143	138.92
Design 5	2000	0.1607	0.0012640	0.1714	0.0010800	14.98
	5000	0.0738	0.0007459	0.0949	0.0009128	54.29
	10,000	0.0498	0.0002895	0.0459	0.0003064	138.06

NOTE: Monte Carlo simulations for the simple mean estimator. Each Monte Carlo consists of 1000 simulated networks, with $K = 2$ unobserved blocks and $\beta = (0.5, 0.75)$. The centers of the blocks are $\nu = (-1.5, 1.0)$. Designs are summarized in Table 2.

small bias. The most challenging design for our estimator is Design 5, where we impose different unobserved block size, different Bernoulli probabilities for the observables and correlation among observed characteristics. As expected, these features increases the bias in our estimates; however, this problem becomes less severe with larger networks. Our weighted mean estimator has similar behavior and results are reported in the Appendix, supplementary materials, in Table 9.

4.2. Application to Facebook 100 Dataset

We apply our method to study a network of friendships, extracted from the Facebook 100 dataset of Traud, Mucha, and Porter (2012).¹⁰ This network was extracted from the Facebook platform in September 2005, providing a snapshot of the friendship among students, faculty, staff and alumni at 100 U.S. universities. We perform an analysis similar to Roy, Atchade, and Michailidis (2019), using the Harvard University network data. The dataset consists of 15,126 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 role in the analysis. We make each of these variables binary. So rather than specify specific dorm information, our control variable indicates whether the student lives on or off-campus. The role is binarized to indicate whether the node is a student or not.¹¹

The characteristics of the network are shown in Table 4. In the first column we report the descriptive statistics for the original data, containing $n = 15,126$ nodes, with an average degree of 109.03, an average clustering coefficient of 0.135, with

Table 4. Descriptive statistics of Harvard University's Facebook network.

	Original	Sample for estimation
n	15,126	13,003
avg. degree	109.03	105.02
clust. coeff.	0.135	0.137
prop. female	0.466	0.539
prop. students	0.508	0.520
prop. off-campus	0.229	0.153

NOTE: The first column is the original network data. The second column is the largest connected component of the network including all nodes with non-missing gender information.

46.6% females, 50.8% students and 22.9% of people living off-campus. The average degree and clustering coefficient suggest that this is a moderately sparse network.

The second column contains the descriptive statistics for our processed sample. We keep all nodes with non-missing gender information; once we delete all missing gender information, there are no missing values for the other covariates; we then determine the largest connected component of the resulting network.¹² The final network contains a higher proportion of females (53.9%) and students (52%); and a smaller fraction of people living off-campus (15.3%) than the original data. The average degree is slightly smaller (105.02), because we have deleted some nodes and edges; however, the clustering coefficients are of similar magnitude, 0.135 and 0.137, respectively.¹³

We estimate the following model with three covariates

$$\log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \mathbf{X}_i^\top \mathbf{X}_j + \beta_1 \mathbf{1}_{\{\text{female}_i = \text{female}_j\}} + \beta_2 \mathbf{1}_{\{\text{student}_i = \text{student}_j\}} + \beta_3 \mathbf{1}_{\{\text{off-campus}_i = \text{off-campus}_j\}}. \quad (47)$$

where $\text{female}_i = 1$ if the node is female, $\text{student}_i = 1$ if the node is a student and $\text{off-campus}_i = 1$ if the node lives off-campus.

of latent positions; (c) cluster latent positions to recover blocks; (d) compute matrix $\hat{\mathbf{B}}_Z$; (e) cluster diagonal entries of matrix $\hat{\mathbf{B}}_Z$ to recover the unobservable block structure; (f) estimate $\hat{\beta}$ using the information on the block structure and the entries of matrix $\hat{\mathbf{B}}_Z$; (g) compute simple mean and weighted mean of the estimated $\hat{\beta}$ according to (43) and (45). The simulation takes a little longer because we need to generate the data and the adjacency matrices for the Monte Carlo. Code is available in GitHub.

¹⁰The entire dataset is available at <https://archive.org/details/oxford-2005-facebook-matrix>.

¹¹Roles include students, faculty, staff, alumni. We focus on students because they are the ones mostly using the platform in 2005.

¹²This is a standard procedure in the literature on SBMs (Athreya et al. 2018; Abbe 2018; Roy, Atchade, and Michailidis 2019).

¹³Before we proceed to estimation, we regularize the adjacency matrix using the standard method proposed in (Le, Levina, and Vershynin 2017). This regularization step avoids numerical issues with the spectral decomposition arising from significant node degree heterogeneity.

Table 5. Estimation results for Harvard University network data (Facebook 100).

Variable	1	2	3	4
Female	0.6614 (0.0240)			0.5766 (0.0061)
Student		0.6883 (0.0360)		0.6463 (0.0144)
Off-campus			0.3269 (0.0113)	0.3315 (0.0020)
n	13,003	13,003	13,003	13,003
\hat{K}	16	16	16	4
\hat{d}	2	2	2	2

NOTE: Parameter estimates for the effect of observable covariates using Harvard University network data from the Facebook-100 dataset. The point estimate is obtained with the weighted mean estimator. The number in parenthesis is the naive standard error estimate, using a plug-in estimator and the formula for variance from [Theorem 1](#). All estimates are obtained using the first elbow of the screeplot (Zhu and Ghodsi 2006).

We first estimate models with one binary covariate, using each of our control variables individually. Next we estimate the full model (47) with three controls. For the Adjacency Spectral Embedding we choose the dimension of the latent space $\hat{d} = 2$, using the profile likelihood method in Zhu and Ghodsi (2006), as in our simulations.¹⁴

The clustering of the estimated latent positions is performed with a Gaussian mixture model, using the MCLUST implementation of Fraley and Raftery (1999) in R. We obtain latent positions estimates and $\hat{K} = 32$ blocks from the adjacency matrix. We then obtain the estimated matrix \hat{B}_Z , and cluster its diagonal entries to recover the (unobserved) blocks \hat{K} and estimate the vector of parameters β . The standard errors are computing using a plug-in estimator for the asymptotic variance. This leads to very conservative estimates as shown in Table 10 in Appendix, supplementary materials.

The estimated parameters are shown in Table 5. In the first three columns we report estimates for models with a single binary covariate. Each coefficient is precisely estimated, according to our naive plug-in standard error estimator; the estimated effects are all positive, which we interpret as evidence of homophily, a usual feature of many social networks. The point estimates are very similar when we estimate the full model (47) (Column 4). These results are also consistent with the analysis in Traud, Mucha, and Porter (2012) and Roy, Atchade, and Michailidis (2019) and qualitatively similar to the JFE approach of Graham (2017) (see Appendix B, supplementary materials). Our method estimates $\hat{K} = 4$ unobserved blocks. If we choose to include only one covariate, the number of blocks estimated is $\hat{K} = 16$ (Columns 1–3). The results indicate that there are additional unobserved characteristics that affect the network formation among Facebook users. In particular, the four blocks may capture shared interests, common preferences, similar schedules and additional information that is unobservable to the researcher.

¹⁴Multiple methods exist for selecting the embedding dimension in practice, and this remains a topic of current research. In the context of networks, choosing a dimension smaller than the true d will introduce bias in the estimated latent positions; on the other hand, using an embedding dimension larger than the true d will increase the variance of the estimated latent positions. In this tradeoff we prefer to err on the side of overestimating d . Specifically, we choose the value one plus the location of the first elbow in the screeplot.

5. Conclusion

We have developed a spectral estimator for stochastic blockmodels with nodal covariates in large networks. In a random effect setting, we can show that our model with discrete covariates and discrete unobserved heterogeneity corresponds to a stochastic blockmodel. 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. We prove a central limit theorem for the estimator, whose theoretical results also apply to moderately sparse graphs, which is important in a host of applications in economics and more generally in social sciences, public health, and computer science, where network data are often viewed as being sparse.

We have shown that our method delivers the same accuracy as the variational EM algorithm, while converging much faster. Our Monte Carlo simulations and the empirical application show that this method works best in very large networks, when the variational EM becomes impractical. Therefore, our estimator is a valid and practical alternative, as long as the number of covariates and the number of blocks is not too large.

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. 2021). In future work, similar ideas can also be applied to directed networks and bipartite networks (Abowd, Kramarz, and Margolis 1999; Bonhomme, Lamadon, and Manresa 2019), significantly expanding the realm of GRDPG applications in economics and the social sciences.

Supplementary Materials

Appendix: Contains all the proofs, additional simulations and extensions to multiple covariates and differential homophily. (.pdf file) grdp: R-package containing the functions and methods developed in this paper. (.zip file) grdp-supplement: R Code for the tables and figures in the paper. Also contains the network data for the empirical application. (.zip file).

Acknowledgments

We are grateful to Cong Mu and Jipeng Zhang for excellent research assistance. We thank the editor, associate editor and two referees, Avanti Athreya, Eric Auerbach, Federico Bandi, Stephane Bonhomme, Youngser Park and Eleonora Patacchini for comments and suggestions.

Funding

Funding from the Institute of Data Intensive Engineering and Science (IDIES) at Johns Hopkins University and NSF grant SES-1951005 is gratefully acknowledged. Joshua Cape also gratefully acknowledges support from NSF grant DMS-1902755.

ORCID

Angelo Mele  <http://orcid.org/0000-0003-2890-6042>
Joshua Cape  <http://orcid.org/0000-0002-1471-1650>

References

- Abbe, E. (2018), "Community Detection and Stochastic Block Models: Recent Developments," *Journal of Machine Learning Research*, 18, 1–86. [2,11]
- Abowd, J., Kramarz, F., and Margolis, D. (1999), "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333. [12]
- Airoldi, E., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Memberships Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [2,4]
- Athreya, A., Fishkind, D. E., Levin, K., Lyzinski, V., Park, Y., Qin, Y., Sussman, D. L., Tang, M., Vogelstein, J. T., and Priebe, C. E. (2018), "Statistical Inference on Random Dot Product Graphs: A Survey," *Journal of Machine Learning Research*, 18, 1–92. [1,2,3,4,5,9,11]
- Athreya, A., Tang, M., Park, Y., and Priebe, C. E. (2021), "On Estimation and Inference in Latent Structure Random Graphs," *Statistical Science*, 36, 68–88. [12]
- Auerbach, E. (2019), "Identification and Estimation of a Partially Linear Regression Model Using Network Data," working paper. [1]
- Badev, A. (forthcoming), "Nash Equilibria on (un)stable Networks," *Econometrica*. [1]
- Beaman, L. A. (2012), "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S.," *The Review of Economic Studies*, 79, 128–161. [1]
- Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic Normality of Maximum Likelihood and its Variational Approximation for Stochastic Blockmodels," *Annals of Statistics*, 41, 1922–1943. [2,4]
- Bonhomme, S., Lamadon, T., and Manresa, E. (2019), "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739. [12]
- Calvo-Armengol, A., Patacchini, E., and Zenou, Y. (2009), "Peer Effects and Social Networks in Education," *Review of Economic Studies*, 76, 1239–1267. [1]
- Cape, J., Tang, M., and Priebe, C. E. (2018), "On Spectral Embedding Performance and Elucidating Network Structure in Stochastic Block Model Graphs," working paper, arXiv:1808.04855. [3]
- Carrell, S. E., Sacerdote, B., and West, J. (2013), "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81, 855–882. [1]
- Chandrasekhar, A. G. (2016), "Econometrics of network formation," in *Oxford Handbook on the Economics of Networks*, eds. Y. Bramoullé, A. Galeotti, and B. W. Rogers, pp. 303–357, Oxford: Oxford University Press. [1]
- Choi, D., Wolfe, P. J., and Airoldi, E. M. (2011), "Stochastic Blockmodels with a Growing Number of Classes," *Biometrika*, 99, 273–284. [2,4]
- Daudin, J.-J., Picard, F., and Robin, S. (2008), "A Mixture Model for Random Graphs," *Statistics and Computing*, 18, 173–183. [2,4]
- DeGiorgi, G., Pellizzari, M., and Redaelli, S. (2009), "Be as Careful of the Books You Read as of the Company You Keep. Evidence on Peer Effects in Educational Choices," IZA discussion paper no. 2833. [1]
- DePaula, A. (2017), "Econometrics of Network Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, eds. B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge: Cambridge University Press. [1,3]
- Diaconis, P., and Chatterjee, S. (2013), "Estimating and Understanding Exponential Random Graph Models," *Annals of Statistics*, 41, 2428–2461. [1]
- Dzermiski, A. (2017), "An Empirical Model of Dyadic Link Formation in a Network with Unobserved Heterogeneity," working paper. [1,2,3]
- Fafchamps, M., and Gubert, F. (2007), "Risk Sharing and Network Formation," *American Economic Review Papers and Proceedings*, 97, 75–79. [1]
- Fraley, C., and Raftery, A. E. (1999), "Mclust: Software for Model-based Cluster Analysis," *Journal of Classification*, (16), 297–306. [3,10,12]
- Graham, B. (2017), "An Empirical Model of Network Formation: With Degree Heterogeneity," *Econometrica*, 85, 1033–1063. [1,2,3,12]
- (2020), "Network Data," in *Handbook of Econometrics 7A*, eds. S. Durlauf, L. Hansen, J. Heckman, and R. Matzkin, Amsterdam: North-Holland. [1,3]
- Graham, B., and dePaula, A., eds. (2020), *The Econometric Analysis of Network Data*, Amsterdam: Academic Press. [1,3]
- Hoff, P., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [3]
- Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [6]
- Jackson, M. ed. (2008), *Social and Economic Networks*, Princeton, NJ: Princeton. [9]
- Jochmans, K. (2017), "Two-Way Models for Gravity," *Review of Economics and Statistics*, 99, 478–485. [1,2]
- Latouche, P., Birmele, E., and Ambroise, C. (2012), "Variational Bayesian Inference and Complexity Control for Stochastic Block Models," *Statistical Modelling*, 12, 93–115. [2]
- Le, C. M., Levina, E., and Vershynin, R. (2017), "Concentration and Regularization of Random Graphs," *Random Structures & Algorithms*, 51, 538–561. [11]
- Mele, A. (2017), "A Structural Model of Dense Network Formation," *Econometrica*, 85, 825–850. [1,3]
- (2022), "A Structural Model of Homophily and Clustering in Social Networks," *Journal of Business and Economic Statistics*, 40, 1377–1389. [1]
- Mele, A., and Zhu, L. (forthcoming), "Approximate Variational Estimation for a Model of Network Formation," *Review of Economics and Statistics*. [1]
- Menzel, K. (2017), "Strategic Network Formation with Many Agents," working paper. [1]
- Mu, C., Mele, A., Hao, L., Cape, J., Athreya, A., and Priebe Carey, E. (2022), "On Spectral Algorithms for Community Detection in Stochastic Blockmodel Graphs with Vertex Covariates," *IEEE Transactions on Network Science and Engineering*, 9, 3373–3384. [2]
- Nakajima, R. (2007), "Measuring Peer Effects on Youth Smoking Behavior," *Review of Economic Studies*, 74, 897–935. [1]
- Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [2]
- Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [2]
- Roy, S., Atchade, Y., and Michailidis, G. (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*, 28, 609–619. [2,3,4,11,12]
- Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (forthcoming), "A Statistical Interpretation of Spectral Embedding: The Generalised Random Dot Product Graph," *Journal of the Royal Statistical Society, Series B*. [1,2,5,9]
- Sweet, T. M. (2015), "Incorporating Covariates into Stochastic Blockmodels," *Journal of Educational and Behavioral Statistics*, 40, 635–664. [2,4]
- Tang, M., and Priebe, C. E. (2018), "Limit Theorems for Eigenvectors of the Normalized Laplacian for Random Graphs," *Annals of Statistics*, 46, 2360–2415. [2,4]
- Tang, M., Cape, J., and Priebe, C. E. (2022), "Asymptotically Efficient Estimators for Stochastic Blockmodels: The Naive MLE, the Rank-Constrained MLE, and the Spectral Estimator," *Bernoulli*, 28, 1049–1073. [1,2,3,4,5,9]
- Traud, A. L., Mucha, P. J., and Porter, M. A. (2012), "Social Structure of Facebook Networks," *Physica A: Statistical Mechanics and its Applications*, 391, 4165–4180. [2,11,12]
- Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013), "Model-Based Clustering of Large Networks," *The Annals of Applied Statistics*, 7, 1010–1039. [2]
- Zelenyev, A. (2020), "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," working paper, https://www.princeton.edu/~zelenyev/azelenyev_jmp.pdf. [2]
- Zheng, D., Mhembere, D., Lyzinski, V., Burns, R., Vogelstein, J., and Priebe, C. E. (2017), "Semi-External Memory Sparse Matrix Multiplication for Billion-Node Graphs," *IEEE Transactions on Parallel and Distributed Systems*, 28, 1470–1483. [2]
- Zhu, M., and Ghodsi, A. (2006), "Automatic Dimensionality Selection from the Scree Plot via the Use of Profile Likelihood," *Computational Statistics and Data Analysis*, 51, 918–930. [5,6,10,12]