Statistica Sinica Preprint No: SS-2022-0341	
Title	Perfect Spectral Clustering with Discrete Covariates
Manuscript ID	SS-2022-0341
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0341
Complete List of Authors	Jonathan Hehir, Xiaoyue Niu and Aleksandra Slavkovic
Corresponding Authors	Xiaoyue Niu
E-mails	xiaoyue@psu.edu
Notice: Accepted version subject to English editing.	

PERFECT SPECTRAL CLUSTERING WITH DISCRETE COVARIATES

Jonathan Hehir, Xiaoyue Niu, and Aleksandra Slavković

Penn State University

Abstract: Among community detection methods, spectral clustering enjoys two desirable properties: computational efficiency and theoretical guarantees of consistency. Where most studies of spectral clustering consider only the edges of a network as input to the algorithm, we consider the problem of performing community detection in the presence of discrete node covariates, with network structure determined by a combination of a latent block model structure and homophily on the observed covariates. We propose a spectral algorithm that we prove achieves perfect clustering with high probability on a class of large, sparse networks with discrete covariates, effectively separating latent network structure from homophily on observed covariates. We apply this method to a network of online friendships among university students to uncover community structure not explained by covariates. To our knowledge, our method is the first to offer a guarantee of consistent latent structure recovery using spectral clustering in the setting where edge formation is dependent on both latent and observed factors.

Key words and phrases: community detection, homophily, spectral clustering, stochastic block model

1. Introduction

A structural pattern commonly observed in social networks is *homophily*, the tendency for two nodes sharing a certain trait to be more (or sometimes less) likely to form a connection (McPherson et al., 2001). Homophily may occur on any number of traits, observed or latent, and is known to confound problems of causal inference in the social sciences (Smith and Christakis, 2008; Shalizi and Thomas, 2011; Goldsmith-Pinkham and Imbens, 2013; Lee and Ogburn, 2021). Homophily, meanwhile, lies at the heart of such issues as segregation (Shrum et al., 1988; Henry et al., 2011), job access (Ibarra, 1992), and political partisanship (Huber and Malhotra, 2017), where homophily on observed traits may be the subject of estimation in its own right. In order to fully understand the effects of network patterns like observed homophily, we first need to separate them from further latent network structure.

In the literature on community detection, latent structure is frequently recovered through a clustering process involving only the network edges, reserving node covariates to validate the clustering results in an approach that conflates latent structure with observed structure (Peel et al., 2017). What we wish to do instead is to separate the latent from the observed structural patterns. To this end, we consider an extension of the stochastic block model (SBM) (Holland et al., 1983) that incorporates homophily on observed, discrete node covariates into a generalized linear model (GLM). We define this model, which we call the *additive-covariate SBM (ACSBM)*,

in Section 2. The model was previously studied by Mele et al. (2019) and allows for flexible modeling choices in which latent communities take a block model structure, covariates may or may not depend on community membership, and the effects of homophily may be modeled through a range of link functions. We give an explicit representation of this model as an SBM (Proposition 1), which motivates the use of spectral clustering to estimate the latent structure.

In the context of SBMs, spectral clustering is known as a fast method that achieves consistency in community detection down to established recovery thresholds (McSherry, 2001; Von Luxburg, 2007; Rohe et al., 2011; Lei and Rinaldo, 2015; Su et al., 2019; Abbe et al., 2020). In Section 3 of this work, we propose a computationally efficient spectral algorithm for recovering the latent structure of the ACSBM. Building on techniques from the field of random dot product graphs (Young and Scheinerman, 2007; Rubin-Delanchy et al., 2017), we develop new algebraic tools to synthesize latent structure over an ACSBM network partitioned by its covariate data. We are able to prove that our method recovers the latent communities of the ACSBM perfectly for sufficiently large networks with node degree at least polylogarithmic in n. Our theoretical analysis is outlined in Section 4, with proofs and derivations deferred to Supplementary Materials, Sections S1 and S2. We provide simulation-based evidence in Section 5 and apply our method to network of Facebook friendships among Harvard students in Section 6. In the Harvard example, we see both strong and subtle homophily over observed covariates (class year and gender, respectively), and

we uncover additional latent structure not explained by these covariates using our method. We conclude with a discussion of the results, their implications, and future generalizations in Section 7.

Related Work. Community detection with covariates is a very active area of research, with a wide variety of methods for modeling community structure, estimating effects of covariates in edge formation, and recovering community memberships. Studies that demonstrate consistency in community recovery assume a generating process with ground-truth communities. Quite commonly, these generating processes feature conditional independence between covariates and edges, given community memberships (e.g., Binkiewicz et al., 2017; Deshpande et al., 2018; Yang et al., 2013; Tallberg, 2004; Newman and Clauset, 2016; Weng and Feng, 2021). In these models, any two nodes belonging to the same latent community have the same connectivity patterns, regardless of their observed covariates.

Explicit separation of latent from observed effects in edge formation is possible in models lacking this conditional independence structure. Such models include (e.g., Hoff, 2007; Handcock et al., 2007; Choi et al., 2012; Vu et al., 2013; Sweet, 2015; Huang and Feng, 2018; Mele et al., 2019; Zhang et al., 2019; Roy et al., 2019; Ma et al., 2020), many of which could be considered broader cases of the model we consider. For example, Hoff (2007); Handcock et al. (2007); Ma et al. (2020) model latent network structure via more general latent position models, which include SBM as a special case. The remainder focus more explicitly on extending SBM but usually

allow greater flexibility in the role of covariates, up to and including allowing arbitrary edge covariates. Since working with SBM likelihood is computationally expensive (Snijders and Nowicki, 1997), many of these studies rely on approximate methods; only a small handful offer methods that scale to large networks and carry a theoretical guarantee of consistent classification. In particular, Huang and Feng (2018) provides a consistency guarantee for spectral clustering only when covariates are independent of community membership, and Ma et al. (2020) provides guarantees only under the assumption of a positive semi-definite latent structure. Our results do not require these assumptions.

By far the most similar paper to ours is Mele et al. (2019), which considers the same model, ACSBM, but under a different spectral estimation method. The main results concern estimation of covariate effects, while we focus on consistency of latent community recovery. Moreover, the results of Mele et al. (2019) implicitly rely on strong assumptions about the community structure that we wish to avoid (see Section 3) and require node degrees of larger order than $\sqrt{n}$. A follow-up paper (Mu et al., 2020) proposes a modification to the algorithm to improve robustness, but results are limited to the specific case of a single covariate under the identity link, with linear node degree.

Contribution. We propose a novel spectral algorithm that is computationally efficient and yields perfect clustering for sufficiently large ACSBM networks with high probability. We prove this result for networks with node degree at least polyloga-

rithmic in n in which homophily effects are multiplicative on the probabilty of edge formation; empirical results suggest greater generality. To our knowledge, our method is the first to offer a guarantee of consistent latent structure recovery using spectral clustering in the important setting where edge formation is dependent on both latent and observed factors.

Notation. Let $[n] = \{1, \ldots, n\}$, with $S_{[n]}$ denoting the set of all permutations $[n] \to [n]$. The function $\mathbb{I}(\cdot)$ is the indicator function. We represent networks as adjacency matrices, e.g., $Y \in \{0,1\}^{n \times n}$. The i-th row of the matrix Y is denoted Y_{i*}, and the i-th column Y_{*i}. $\mathbf{1}_n$ denotes a column vector of n ones. We use $\|x\|_2$ to denote the ℓ_2 norm of a vector x, $\|A\|_F$ to denote the Frobenius norm of a matrix, and $\|A\|_2$ to denote the spectral norm of the matrix A, i.e., $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$. We write $A \succeq 0$ for positive semi-definite matrices and $A \succ 0$ for positive-definite matrices. All functions of matrices are taken element-wise, with the exception of the matrix absolute value, $|A| = \sqrt{A^T A}$. When $n \to \infty$, we write $a_n = o(b_n)$ if $|a_n/b_n| \to 0$; $a_n = \omega(b_n)$ if $|a_n/b_n| \to \infty$; $a_n = O(b_n)$ if $|a_n/b_n| \leq C$ for some $C > 0$ and all n; and $a_n = \Theta(b_n)$ if $|a_n/b_n| \in (C_1, C_2)$ for some $C_2 > C_1 > 0$ and all n. Finally, we write $X_n = O_P(b_n)$ if for any $\alpha > 0$ there exists a constant C such that $\mathbf{P}(|X_n/b_n| > C) < \alpha$ for all large n; and $X_n = o_P(a_n)$ if $\mathbf{P}(|X_n/a_n| > \varepsilon) \to 0$ for all $\varepsilon > 0$. Further notation is defined in text as needed.

Code. A Python implementation of our proposed method, including simulation code and additional examples, is available at `https://github.com/jonhehir/acsbm`.

2. Network Model and Representation

The network model we consider is an extension of the popular stochastic block model (SBM) (Holland et al., 1983), which we recall in Definition 1.

Definition 1. Conditioned on community membership $\theta \in [K]^n$, the undirected network $Y \sim \text{SBM}(\theta, B)$ is an SBM with edge probabilities $B \in [0,1]^{K\times K}$ if:

$$Y_{ij} \overset{ind}{\sim} \text{Bernoulli}(B_{\theta_i\theta_j}), \quad i < j.$$

The extension we study is what we call the *additive-covariate stochastic block model* (ACSBM), which is also the model studied in Mele et al. (2019). In this setting, we observe a network with n nodes and K communities, along with a set of M discrete covariates. Links are formed independently, depending on community assignments, as in SBM, as well as on covariate similarity, allowing for explicit modeling of homophily based on the observed covariates. Homophily is therefore modeled in a manner similar to exponential random graph models (Goodreau et al., 2009), with latent structure modeled like SBM. The specific nature of the covariate influence is captured by a known link function g. We state a formal definition of this model in Definition 2.

Definition 2. For nodes $i \in [n]$, let $\theta_i \in [K]$ denote latent community membership, and let $Z_i \in [L_1] \times \cdots \times [L_M]$ be a vector of M discrete, observed covariates. Let $Z = [Z_1 \mid \cdots \mid Z_n]^T$. Conditioned on θ and Z, the undirected network $Y \sim \text{ACSBM}(\theta, Z, B, \beta, g)$ is an additive-covariate SBM with covariate effects $\beta \in \mathbb{R}^M$

and known link function g if:

$$Y_{ij} \overset{ind}{\sim} \text{Bernoulli}\left(g^{-1}\left(B_{\theta_i\theta_j} + \sum_{m=1}^{M} \beta_m \mathbb{I}(Z_{im} = Z_{jm})\right)\right), \quad i < j.$$

While the link function g could in principle be any strictly increasing function whose range includes $[0, 1]$, typical choices inspired by similar models include the logit link (e.g., Handcock et al., 2007; Choi et al., 2012; Vu et al., 2013; Roy et al., 2019; Ma et al., 2020), log link (e.g., Huang and Feng, 2018), probit link (e.g., Hoff, 2007), or identity link (Mu et al., 2020). Choice of link function should be informed by the nature in which covariates are believed to affect edge formation. Our theoretical analysis in Section 4 focuses primarily on the log link, in which the effects of observed homophily are multiplicative on the probability of edge formation. Such effects are particularly reasonable to assume in sparse networks, easily interpreted (if estimated), and mimic the form of other popular models like the degree-corrected block model (Karrer and Newman, 2011). We offer generalizations to other link functions as well as more flexible models of homophily under additional theoretical assumptions.

The ACSBM's combination of independent edges and discrete attributes leads to an important representation result: the ACSBM, viewed one way as an extension of the SBM, may also be represented by a special case of the SBM. Specifically, Proposition 1 represents the ACSBM as an SBM by subdividing each latent community in ACSBM by the observed covariates, yielding an SBM over the resulting set of "subcommunities." This generalizes a similar result stated by Mele et al. (2019).

Proposition 1. *If* $Y \sim \text{ACSBM}(\theta, Z, B, \beta, g)$*, then* Y *is equal in distribution to a* $(K\tilde{L})$*-block SBM, namely* $Y \overset{D}{=} \text{SBM}(\tilde{\theta}, \tilde{B})$ *for:*

$$\tilde{L} = \prod_{m=1}^{M} L_m$$

$$\tilde{\theta} = \tilde{L}(\theta - \mathbf{1}_n) + \sum_{m=1}^{M-1} \left[\prod_{m'=m+1}^{M} L_{m'} \right] (Z_{*m} - \mathbf{1}_n) + Z_{*M},$$

$$\tilde{B} = g^{-1}(B \boxplus \beta_1 I_{L_1} \boxplus \cdots \boxplus \beta_P I_{L_M}),$$

where g^{-1} *is taken element-wise, and* $A_1 \boxplus A_2 = (A_1 \otimes \mathbf{1}_{d_2}\mathbf{1}_{d_2}^T) + (\mathbf{1}_{d_1}\mathbf{1}_{d_1}^T \otimes A_2)$ *for matrices* $A_1 \in \mathbb{R}^{d_1 \times d_1}, A_2 \in \mathbb{R}^{d_2 \times d_2}$.

Remark 1. $\tilde{\theta}$ is formed from a bijection from $[K] \times [L_1] \times \cdots \times [L_M]$ to $[K\tilde{L}]$. In an abuse of notation, we will refer to this mapping later in the paper as $\tilde{\theta}(\cdot, \cdot)$ where for $k \in [K], z \in [L_1] \times \cdots \times [L_M]$, $\tilde{\theta}(k, z) = \tilde{L}(k-1) + \sum_{m=1}^{M-1} \left[\prod_{m'=m+1}^{M} L_{m'} \right] (z_m - 1) + z_M$.

The proof of Proposition 1 is constructive and is given in Supplementary Materials, Section S2. This representation result leads to a natural idea: since any ACSBM network is equivalently represented as an SBM, perhaps familiar SBM-fitting methods can be adapted to fit the ACSBM.

2.1 Random Dot Product Graphs

Spectral clustering of SBMs has been studied extensively in the context of (generalized) random dot product graphs (RDPGs) (Athreya et al., 2017; Rubin-Delanchy et al., 2017). The class of (g)RDPGs lends itself well to spectral estimation methods,

and any binary, undirected, independent-edge network can be formulated as a generalized random dot product graph. In particular, it is well established that SBMs may be represented as gRDPGs (Rubin-Delanchy et al., 2017). Below we state the definition of a gRDPG and follow it with a representation result for ACSBM analogous to Proposition 1.

Definition 3. The matrix $I_{pq} = \text{diag}(I_p, -I_q)$ is the diagonal matrix whose first p diagonal entries are equal to $+1$ and whose remaining q diagonal entries are equal to -1. For $x, y \in \mathbb{R}^d$ and some nonnegative integers $p + q = d$, the indefinite inner product of x and y with signature (p, q) is given by $\langle x, y \rangle_{pq} = \langle x, I_{pq} y \rangle = x^T I_{pq} y$. The indefinite orthogonal group with signature (p, q) is given by the set of matrices $\mathbb{O}(p, q) = \{Q \in \mathbb{R}^{d \times d} : Q^T I_{pq} Q = I_{pq}\}$.

Definition 4. Let F_X be a distribution on $\mathbb{R}^d$. We say the undirected network $Y \sim \text{gRDPG}(n, F_X)$ is a generalized random dot product graph with signature (p, q) if $X_1, \ldots, X_n \overset{iid}{\sim} F_X$, and $Y_{ij} \mid X_1, \ldots, X_n \overset{ind}{\sim} \text{Bernoulli}(\langle X_i, X_j \rangle_{pq})$ for $i < j$. The variable X_i is referred to as the latent position of the i-th node.

Remark 2. When $q = 0$, we say Y is a random dot product graph (without the "generalized" qualification) (Young and Scheinerman, 2007). In this case, $I_{pq} = I$, the indefinite inner product coincides with the usual dot product (i.e., $\langle x, y \rangle_{pq} = \langle x, y \rangle$), and $\mathbb{O}(p, q)$ coincides with the familiar group of $p \times p$ orthogonal matrices.

Both RDPGs and gRDPGs suffer from inherent identifiability issues. In the

case of RDPGs, for example, if any set of latent positions is altered by a common orthogonal transformation, the resulting RDPG has the same distribution, since $\langle x, y\rangle = \langle Qx, Qy\rangle$ for any orthogonal Q. In gRDPGs, latent positions can only be identified up a common indefinite orthogonal transformation (Rubin-Delanchy et al., 2017). (For a comprehensive approach to the non-identifiability of gRDPGs, see Agterberg et al. (2020).) Unlike orthogonal transformations, indefinite orthogonal transformations do not preserve distances or angles, rendering them more burdensome to work with. In the following proposition, we choose our canonical latent positions based on a spectral decomposition, but we clarify that this choice of latent positions is not unique. The proof of Proposition 2, given in the Supplementary Materials, follows as a corollary to Proposition 1, based on well known results in the gRDPG literature (e.g., Rubin-Delanchy et al., 2017, Section 2.1).

Proposition 2. *If* $(\theta_i, Z_i) \in [K]\times[L_1]\times\cdots\times[L_M]$ *are drawn i.i.d. from a distribution with p.m.f.* $\mathbf{P}_{\theta,Z}$*, and* $Y \mid \theta, Z \sim \text{ACSBM}(\theta, Z, B, \beta, g)$ *for* $Z = [Z_1 \mid \cdots \mid Z_n]^T$ *and some* $\beta \in \mathbb{R}^M$*, then* Y *is equal in distribution to a gRDPG,* Y_{grdpg}*, with latent positions sampled i.i.d. from a mixture of point masses. A canonical distribution for these latent positions is as follows. Let* $\tilde{B}$ *as in Proposition 1, and let* $U_{\tilde{B}}\Lambda_{\tilde{B}}U_{\tilde{B}}^T$ *be an eigendecomposition of* $\tilde{B}$*. Let* $X_{\tilde{B}} = U_{\tilde{B}}|\Lambda_{\tilde{B}}|^{1/2}$*, and let* $X_{\tilde{B}}(k,z)$ *denote the* $\tilde{\theta}(k,z)$*-th row of* $X_{\tilde{B}}$*. Let* $F_{X_{\tilde{B}}}$ *as follows:*

$$F_{X_{\tilde{B}}} = \sum_{\substack{k\in[K],\\ z\in[L_1]\times\cdots\times[L_M]}} \mathbf{P}_{\theta,Z}(\theta = k, Z = z)\delta_{X_{\tilde{B}}(k,z)}.$$

Letting q denote the number of negative entries in $\Lambda_{\tilde{B}}$, we have $Y_{grdpg} \sim \text{gRDPG}(n, F_{X_{\tilde{B}}})$ with signature $(p, q) = (K\tilde{L} - q, q)$.

3. Proposed Spectral Clustering Procedure

We propose a three-part algorithm (Algorithm 1) to estimate the latent community membership θ for an ACSBM network. Since an ACSBM with K latent communities is equivalently a $(K\tilde{L})$-block SBM per Proposition 1, we begin by trying to find the $K\tilde{L}$ "subcommunities" (i.e., $\tilde{\theta}$) of the SBM representation. Assuming we can recover the $K\tilde{L}$ subcommunities suitably, the primary remaining challenge is to merge these subcommunities into the original K desired communities (i.e., θ).

This fundamental idea is similar to that underlying Mele et al. (2019); Mu et al. (2020), but we propose a new method for delineating the subcommunities and matching each subcommunity back to its original latent community, allowing for provably consistent results under mild assumptions. In both Mele et al. (2019) and Mu et al. (2020), the process of finding the $K\tilde{L}$ subcommunities relies only on the expected separation of their spectral embeddings in Euclidean space—a condition not met if any β_m is sufficiently small (or zero). Moreover, subsequent estimation of β in Mele et al. (2019); Mu et al. (2020) relies implicitly on an assumption that the diagonal entries in B are unique, so that an estimate of $\text{diag}(\tilde{B})$ can be clustered into K sets of similar values corresponding to the K latent communities. In contrast, our method is robust to non-significant homophily effects and allows for any choice of B that

satisfies a full-rank assumption.

Remark 3. Algorithm 1 takes as input an embedding dimension d. This corresponds to the dimension of the latent positions in Proposition 2, which cannot exceed $K\tilde{L}$. In the absence of oracle knowledge, this maximum value appears to be a suitable choice for d.

Part 1 of the algorithm essentially seeks to recover $\tilde{\theta}$ of Proposition 1. To do so, we first find adjacency spectral embeddings for the full network. Then we consider each possible covariate configuration $z \in [L_1] \times \cdots \times [L_M]$ (of which there are $\tilde{L}$ total), and cluster the embeddings corresponding to nodes bearing this covariate configuration into K clusters. This yields a set of subcommunities that are each pure in their covariate distribution, since we know that $Z_i \neq Z_j \implies \tilde{\theta}_i \neq \tilde{\theta}_j$. A range of clustering methods (e.g., K-means) may be used here; existing theory suggests Gaussian mixture models may provide the best finite-sample performance (Athreya et al., 2016; Rubin-Delanchy et al., 2017). The computational complexity of Part 1 will depend on the specific clustering method employed.

Part 2 of the algorithm estimates $\tilde{B}$ so that we may estimate a latent position for each subcommunity. While the embeddings of Part 1 also serve as estimates of latent positions, these estimates are only consistent up to an indefinite orthogonal transformation, which would pose problems for the geometry of Part 3. In practical implementations, Part 2 can be performed in linear time, relative to the number of

Algorithm 1 Spectral Clustering of ACSBM

Input: adjacency matrix $Y \in \{0,1\}^{n\times n}$, discrete covariates $Z = [z_1 \mid \cdots \mid z_n]^T$, number of latent communities K, embedding dimension d

Output: estimated block membership $\hat{\theta} \in [K]^n$

 # Part 1: Recover the subcommunities $\tilde{\theta}$

Let $\hat{X}_Y := U|\Lambda|^{1/2}$, where $U\Lambda U^T$ is the truncated eigendecomposition of Y with dimension d

Let $L_1, \ldots, L_M := \max(Z_{*1}), \ldots, \max(Z_{*M})$

for z in $[L_1] \times \cdots \times [L_M]$ **do**

Let $\mathcal{I}_z := \{i : z_i = z\}$

Let $\hat{\theta}_z : \mathcal{I}_z \to [K]$ be a function returning cluster assignments over the rows of $\hat{X}_Y$ corresponding to the indices $\mathcal{I}_z$

end for

 # Part 2: Estimate $\tilde{B}$

for $1 \le k_1 \le k_2 \le K\tilde{L}$ **do**

Let $D_{k_1,k_2} := \{(i,j) \in [n] \times [n] : i \ne j, \tilde{\theta}(\hat{\theta}_{z_i}(i), z_i) = k_1, \tilde{\theta}(\hat{\theta}_{z_j}(j), z_j) = k_2\}$

Set $\hat{\tilde{B}}_{k_1,k_2} = \hat{\tilde{B}}_{k_2,k_1} := \sum_{(i,j)\in D_{k_1,k_2}} A_{ij} / \max\{1, |D_{k_1,k_2}|\}$

end for

 # Part 3: Reconcile θ using $z = \mathbf{1}_M$ as reference level

Let $\hat{X}_{\tilde{B}}(k,z)$ be the $\tilde{\theta}(k,z)$-th row of $V|\Psi|^{1/2}$, where $V\Psi V^T$ is an eigendecomposition of $\hat{\tilde{B}}$

for z in $[L_1] \times \cdots \times [L_M]$ **do**

Let $\hat{\sigma}_z := \arg\min_{\sigma\in S_{[K]}} \sum_{k=1}^{K} \|\hat{X}_{\tilde{B}}(\sigma(k), z) - \hat{X}_{\tilde{B}}(k, \mathbf{1}_M)\|_2^2$

end for

return $\hat{\theta} = [\hat{\sigma}_{z_i}(\hat{\theta}_{z_i}(i))]_{i=1}^n$

edges in the network.

Successful clustering in Part 1 of the algorithm implies that we are able to recover θ up to a separate permutation for any set of nodes with the same covariates. Part 3 of the algorithm seeks a common permutation for all nodes by attempting to reconcile each covariate configuration with a given reference level (canonically $z = \mathbf{1}_M$). This is achieved by finding the matching that minimizes the sum of squared distances between estimates of latent positions for each cluster. This optimization is a case of the assignment problem, which can be completed efficiently using the Hungarian algorithm (Edmonds and Karp, 1972). The computational complexity of Part 3 depends only on K and $\tilde{L}$. The analysis in Section 4 assumes these quantities are constant in n. If allowed to grow, however, we would only expect consistency of subcommunity recovery (i.e., Part 1) if $K\tilde{L} = O(\sqrt{n})$, based on existing results in SBM recovery (e.g., Choi et al., 2012). Under this assumption, the overall complexity of Part 3 of the algorithm is $O(n^{1.5})$ in time and $O(n)$ in space.

4. Consistency Results

Breaking Algorithm 1 into its three main parts, we first show that Part 1 consistently recovers $\tilde{\theta}$ from Proposition 1. Next, Part 2 yields a consistent estimate of $\tilde{B}$, given $\tilde{\theta}$ from Part 1. Finally, Part 3 yields a consistent estimate of θ, given $\tilde{\theta}$ from Part 1 and a suitable approximation of $\tilde{B}$ from Part 2. While detailed proofs of these results are left to the Supplementary Materials, we state the major theorems and give an outline

of the proof ideas here. To make things concrete, we consider the following setting.

Setting. Let M be a positive integer, and let $K, L_1, \ldots, L_M$ be integers greater than 1. Let $\mathbf{P}_{\theta Z}$ be a probability mass function on $[K] \times [L_1] \times \cdots \times [L_M]$. Let $\beta \in \mathbb{R}^M$ be a vector of covariate coefficients and $B_0 \in \mathbb{R}^{K \times K}$ be a symmetric matrix of latent block coefficients. To allow for sparsity, let $\alpha_n \in (0, 1]$ be a sequence controlling the expected degree of our networks. For each $n \geq 1$, we draw $\{(\theta_i, Z_i)\}_{i=1}^n \in ([K] \times [L_1] \times \cdots \times [L_M])^n$ from $(\mathbf{P}_{\theta Z})^n$. Letting $B = B_0 + \log(\alpha_n)\mathbf{1}_K\mathbf{1}_K^T$, we then draw $Y \mid \theta, Z \sim \text{ACSBM}(\theta, Z, B, \beta, \log)$.

As discussed in Section 2, under the log link, the effects of observed homophily are multiplicative on the probability of edge formation. When $\alpha_n \to 0$, this is essentially equivalent to the canonical logit link in the limit, since $\lim_{n\to\infty} \log^{-1}(b + \log(\alpha_n))/\text{logit}^{-1}(b + \log(\alpha_n)) = 1$ for any constant b. We note that in this setting, all edge probabilities scale by α_n, so the expected degree of each node is $\Theta(n\alpha_n)$. Although we drop the subscripts, the quantities $\tilde{B}$ and $X_{\tilde{B}}$ depend on n. When we desire constant quantities, we will use $\alpha_n^{-1}\tilde{B}$ and $\alpha_n^{-1/2}X_{\tilde{B}}$.

Assumptions. Our full set of results will require the following assumptions. Assumption ***(A1)*** limits the sparsity of the network, as is standard in the SBM recovery literature. We note that the polylogarithmic (rather than logarithmic) term used here matches the assumptions of Rubin-Delanchy et al. (2017), on which our theory is built. Assumption ***(A2)*** is equivalent to saying the latent SBM structure is full-rank, which is also a common assumption. Assumption ***(A3)*** requires that each

latent community contains a node of each type with nonzero probability.

(A1) $\alpha_n = \omega(\log^{4c} n/n)$ for the universal constant c in Lemma 1.

(A2) $\exp(B_0)$ is full-rank.

(A3) $\mathbf{P}_{\theta Z}(\theta = k, Z = z) > 0$ for all $(k, z) \in [K] \times [L_1] \times \cdots \times [L_M]$.

Our consistency analysis assumes that $M, K, L_1, \ldots, L_M$ are constant in n. We expect that these could be allowed to grow in an appropriate asymptotic regime, but we note that based on prior results in the SBM literature, we expect any growth would need to be reasonably slow. For example, the subcommunity recovery of Algorithm 1, Part 1 likely requires $K\tilde{L} = KL_1 \cdots L_m = O(\sqrt{n})$ (e.g., Choi et al., 2012).

We begin by recasting the ACSBM as a gRDPG with signature (p, q), as prescribed by Proposition 2. Let $\hat{X}_Y = U|\Lambda|^{1/2}$ (where $Y \approx U\Lambda U^T$) as in Algorithm 1, and let $\hat{X}_i$ denote the i-th row of $\hat{X}_Y$ (i.e., the spectral embedding for node i). Results from the gRDPG literature tell us that these spectral embeddings will be consistent estimates of the latent positions of the gRDPG, up to an unknown transformation from the indefinite orthogonal group $\mathbb{O}(p, q)$. This is stated in Lemma 1, which follows from Rubin-Delanchy et al. (2017, Theorem 3).

Lemma 1 (Rubin-Delanchy et al. (2017))**.** *Under assumptions* ***(A1)*** *and* ***(A3)****, there exists a universal constant $c > 1$ and a sequence of matrices $Q \in \mathbb{O}(p, q)$ such*

that:

$$\max_{i\in[n]} \|Q\hat{X}_i - X_{\tilde{B}}(\theta_i, Z_i)\|_2 = O_P\left(\frac{\log^c n}{\sqrt{n}}\right).$$

The uniform consistency of Lemma 1 is the key to Part 1 of the algorithm. In particular, when we look at the spectral embeddings for nodes of a given covariate configuration $z \in [L_1] \times \cdots \times [L_M]$, this result yields perfect separation of the embeddings with high probability (Theorem 1).

Theorem 1. *Fix $z \in [L_1] \times \cdots \times [L_M]$. Let $\mathcal{I}_z = \{i : Z_i = z\}$. Assuming **(A1)** and **(A3)**, there exist K sequences of balls $\mathcal{B}_{1,z}, \ldots, \mathcal{B}_{K,z}$ such that $\hat{X}_i \in \mathcal{B}_{\theta_i,z}$ for all $i \in \mathcal{I}_z$ and $\mathcal{B}_{1,z}, \ldots, \mathcal{B}_{K,z}$ are disjoint with probability approaching 1.*

Theorem 1 is proven in Supplementary Materials, Section S2, and is sufficient to support exact recovery of $\tilde{\theta}$ with high probability under a variety of clustering algorithms, such as K-means (Lyzinski et al., 2014). However, while Lemma 1 states spherical concentration bounds, the clusters of embeddings generally are not spherical but are asymptotically normal, per the discussion in Rubin-Delanchy et al. (2017). For this reason, Gaussian mixture modeling is often preferred over K-means for finite-sample performance (Athreya et al., 2016; Rubin-Delanchy et al., 2017).

In view of Theorem 1, from here we assume knowledge of $\tilde{\theta}$ in order to demonstrate consistency in Parts 2 and 3 of the algorithm. Recall that Part 2 of the algorithm estimates $\tilde{B}$ from Proposition 1. While this estimate is not our end goal, we will use this reconstruction of $\tilde{B}$ to estimate the canonical latent positions $X_{\tilde{B}}$ from Proposition 2.

Theorem 2. *Let $\hat{\theta}_z : \mathcal{I}_z \to [K]$. Suppose for each $z \in [L_1] \times \cdots \times [L_M]$, there exists $\tau_z \in S_{[K]}$ such that $\hat{\theta}_z(i) = \tau_z(\theta_i)$ for all $i \in \mathcal{I}_z$. Assuming **(A1)**–**(A3)**, if $\hat{\tilde{B}}$ is constructed as in Algorithm 1, then there exists a sequence of $K\tilde{L} \times K\tilde{L}$ permutation matrices T such that:*

$$\alpha_n^{-1} \|\hat{\tilde{B}} - T\tilde{B}T^{-1}\|_F = o_P\left(\frac{1}{\sqrt{n \log^c n}}\right).$$

Theorem 2 follows from the fact that, conditioned on $\tilde{\theta}$, $\hat{\tilde{B}}$ is the maximum likelihood estimate for a matrix of SBM probabilities corresponding to the subcommunities of $\tilde{\theta}$ (up to relabeling). The bounds thus follow from a bit of algebraic manipulation of well-known results (Bickel et al., 2013; Tang et al., 2022), as outlined in Supplementary Materials, Section S2. Finally, we move on to the main act: reconciling the $\tilde{L}$ per-covariate clusterings into a single clustering for all nodes.

Theorem 3. *Let $\hat{\theta}_z : \mathcal{I}_z \to [K]$ and $\hat{X}_{\tilde{B}}(k, z)$ as in Algorithm 1. Suppose for each $z \in [L_1] \times \cdots \times [L_M]$, there exists $\tau_z \in S_{[K]}$ such that $\hat{\theta}_z(i) = \tau_z(\theta_i)$ for all $i \in \mathcal{I}_z$. Let:*

$$\hat{\sigma}_z = \arg\min_{\sigma \in S_{[K]}} \sum_{k=1}^{K} \|\hat{X}_{\tilde{B}}(\sigma(k), z) - \hat{X}_{\tilde{B}}(k, \mathbf{1}_M)\|_2^2. \tag{4.1}$$

*Then, assuming **(A1)**–**(A3)**, $\hat{\sigma}_z(\hat{\theta}_z(i)) = \tau_{\mathbf{1}_M}(\theta_i)$ for all $i \in [n]$ with probability approaching 1.*

Theorem 3 involves an abundance of permutations. We assume that for each covariate configuration z, we have a function $\hat{\theta}_z(\cdot)$ that recovers the values of θ_i up to a permutation τ_z. We can find such functions with high probability from Part 1 of our

algorithm. Then, for each z, we estimate a permutation $\hat{\sigma}_z$ in an attempt to "reverse" these permutations. Since the true permutations τ_z are unknowable, we cannot hope to invert τ_z exactly. Instead, we seek a permutation that satisfies $\hat{\sigma}_z \circ \tau_z = \tau_0$ for some common unidentifiable permutation $\tau_0 \in S_{[K]}$. By using $z = \mathbf{1}_M$ as our reference level, we end up recovering $\tau_0 = \tau_{\mathbf{1}_M}$.

The proof of Theorem 3 is broken into a number of intermediate results in the supplementary materials, of which we give an overview here. We first consider the task of solving an analog to the matching problem (4.1) using the true latent positions $X_{\tilde{B}}$ (Section S2, Theorem S15). An intuitive explanation for minimizing the sum of squared distances is clearest in the case when the latent communities follow an assortative homophily structure and (correspondingly) $\exp(B_0)$ is positive-definite. In this case, nodes in the network are more likely to connect with other nodes within the same latent community rather than across communities. Correspondingly, the embeddings of nodes that are most likely to connect with each other tend to be closer together in latent space. Minimizing the expected sum of squared distances yields the permutation of rows of $\exp(B_0)$ that maximizes the matrix trace. Since $\exp(B_0)$ is positive-definite, this is the "correct" arrangement of rows (Section S2, Fact S14). A similar concept applies even when $\exp(B_0)$ is not positive-definite, as the sum of embedding distances being minimized reduces to a calculation involving entries in $|\exp(B_0)|$, which is positive-definite.

Having shown that the matching problem yields the desired result in the absence of

estimation error, it remains to show that the estimation error vanishes asymptotically (Section S2, Lemma S17). The estimation error is bounded by a multiple of $\| \, |\hat{\tilde{B}}| - |T\tilde{B}T^{-1}| \, \|_F$, a bound for which follows from Theorem 2. This, indeed, shrinks to zero faster than the gap between the optimal and second-best matching. A formal proof of Theorem 3 tying these results together is given in Section S2 of the Supplementary Materials.

In sum, Theorems 1–3 demonstrate that Algorithm 1 perfectly recovers ACSBM's latent community assignment variable, θ, in the limit. From this, asymptotically unbiased estimation of the remaining ACSBM parameters—including the marginal homophily effects, β—follows in a straightforward manner, using standard GLM-fitting approaches with $\hat{\theta}$ as a plug-in estimator for θ.

Generalizations. The preceding consistency results assume the use of a log link function, which yields a simple decomposition of the matrix $\tilde{B}$ (of Proposition 1) and its matrix absolute value $|\tilde{B}|$. It is natural to ask whether the consistency of the proposed algorithm indeed depends on the choice of log link. While the simulations of Section 5 provide empirical evidence that the algorithm is robust to the choice of link function, we can also point to signs that this may provably be so.

The challenges involved with analyzing $|\tilde{B}|$ are largely avoided in the case when $\tilde{B}$ is positive-definite, in which case $|\tilde{B}|$ is simply $\tilde{B}$ itself. Such positive-definite assumptions are common (e.g., Ma et al., 2020) and capture typical cases of *assortative homophily*, i.e., the tendency for similar (rather than dissimilar) nodes to connect. In

the presence of the appropriate positive-definite assumptions, some generalizations follow as immediate corollaries.

In addition to allowing any link function g, we may also generalize the ACSBM definition (Definition 2) to allow for *differential homophily*, i.e., covariate effects that vary across different levels of a covariate. In other words, keeping M, L, θ, and Z as defined for the ACSBM, let $\beta_{m\ell} \in \mathbb{R}$ for $m \in [M], \ell \in [L_m]$ denote differential homophily coefficients for the model

$$Y_{ij} \overset{ind}{\sim} \text{Bernoulli}\left(\alpha_n g^{-1}\left(B_{\theta_i\theta_j} + \sum_{m=1}^{M}\sum_{\ell=1}^{L_m} \beta_{m\ell}\mathbb{I}(Z_{im} = Z_{jm} = \ell)\right)\right),$$

where g is any link function. If $\tilde{B} \succeq 0, g^{-1}(B) \succ 0$ and ***(A1)***, ***(A3)*** hold, then we will also achieve consistent recovery of θ (Section S2, Corollary S16).

5. Simulations

We evaluate the empirical performance of our method on a variety of sequences of ACSBM networks. First, we consider two sequences of sparse networks ($\alpha_n = n^{-0.8}$) with $K = 2$ latent communities and $M = 2$ covariates drawn i.i.d. as Bernoulli(0.5). The link function is chosen to be $g = \log$. In the first setting, we use a "regular" structure for the latent SBM, $B_0 = 1.5\,\mathbf{1}_2\mathbf{1}_2^T - I_2$. In the second, we consider something more "irregular," with $B_0 = \mathbf{1}_2\mathbf{1}_2^T + \text{diag}(1, -0.2)$. In both cases, covariate effects are $\beta_1 = 1, \beta_2 = -0.5$. The choice to include negative homophily effects (both latent and observed) is intentional, yielding matrices B_0 and corresponding $\tilde{B}$ that

are indefinite. For each of ten values of n ranging from $n = 125$ to $n = 128000$, we generate 100 networks, then apply Algorithm 1, using Gaussian mixture modeling as our clustering method for Part 1. We calculate a misclassification rate (up to relabeling) as $\min_{\sigma \in S_{[K]}} n^{-1} \sum_{i=1}^{n} \mathbb{I}(\sigma(\hat{\theta}_i) \neq \theta_i)$. The median misclassication rate is plotted in the left panel of Figure 1, with error bands denoting the interquartile range (IQR). The dashed line represents the worst possible misclassification rate of one half. As we might hope, as n increases, misclassification falls toward zero.

The second set of simulations evaluates the performance of the algorithm on dense networks ($\alpha_n = 1$), with four settings corresponding to different choices of link function: identity, log, logit, and probit. In each case, we model the underlying latent structure as an SBM with $K = 3$ communities and model $M = 2$ binary covariates, drawn i.i.d. as Bernoulli(0.5). For the identity link, we choose $B = 0.2\,\mathbf{1}_3\mathbf{1}_3^T - 0.1I_3, \beta_1 = 0.05, \beta_2 = -0.05$. For the remaining links, we use $B = -\mathbf{1}_3\mathbf{1}_3^T - 0.5I_3, \beta_1 = -0.7, \beta_2 = 0.1$. For seven values of n ranging from $n = 125$ to $n = 8000$, we simulate 100 networks and apply the same clustering methodology as in the previous set of simulations. The results are plotted in the right panel of Figure 1. Here we see consistency for a variety of link functions, even in the absence of the positive-definiteness assumption employed in the generalization to arbitrary link functions (Section S2, Corollary S16), suggesting even greater generality for our proposed method. In our dense simulations, we achieve perfect clustering in the overwhelming majority of cases when $n \geq 2000$.

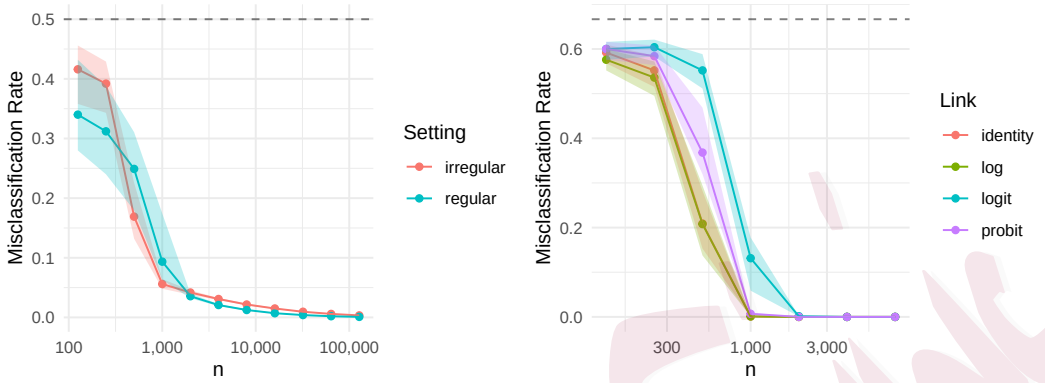

Figure 1: Median proportion (and IQR) of misclassified nodes on repeated simulations of ACSBM models. Left: Sparse settings with $K = 2, M = 2, g = \log, \alpha_n = n^{-0.8}$. Right: Dense settings with $K = 3, M = 2$, various g, $\alpha_n = 1$. Dashed line represents worst possible misclassification $(1 - 1/K)$. Specific parameters given in text.

We caution against direct comparisons of the simulation settings presented here. For example, in the dense network simulations, one may notice that convergence appears fastest for the log link and slowest for the logit link, but each setting is different in ways that complicate comparisons. While these two settings share the same parameters, the difference in link function subtly affects the relations between entries in $\tilde{B}$ and leads to a network of lower density for the logit link, since $\text{logit}^{-1}(x) < \log^{-1}(x)$ for any $x \in \mathbb{R}$.

These simulations were conducted on a high performance cluster, but each individual network was simulated and fit using a single CPU core (2.2 GHz Intel Xeon). The most demanding simulation setting was the sparse, regular setting at $n = 128000$

nodes, where each network had about 6.2 million edges on average. The average running time for this setting using our Python-based algorithm was 4.35 minutes per network, of which 4.25 minutes were spent in Part 1 of Algorithm 1.

Comparisons with other spectral algorithms. Comparing our algorithm against other spectral clustering–based competitors reveals a distinct advantage—robustness to model parameters. We specifically compare against three other algorithms: Mele et al. (2019), which implicitly requires the matrix B appearing in Definition 2 to have unique diagonal entries; Mu et al. (2020), which post-processes the results of Mele et al. (2019) by estimating covariate effects and re-clustering an adjacency matrix from which the estimated covariate effects have been removed; and a vanilla spectral clustering that simply ignores the known covariate information. (For the "Vanilla" method, we employ a modified version of the algorithm from Lei and Rinaldo (2015), replacing the use of K-medians clustering with the more typical K-means.)

We simulate a model in which each node corresponds to two random variables: an observed covariate $Z_i \in [3]$ and a latent community membership $\theta_i \in [2]$. We set Z_i and θ_i to be correlated, with relative frequencies as given in Table 1.

We draw 100 random undirected networks of size $n = 400$ nodes with the simple edge distribution,

$$Y_{ij} \overset{ind}{\sim} \text{Bernoulli}\left(0.05 + 0.3\mathbb{I}(\theta_i = \theta_j) + \beta\mathbb{I}(Z_i = Z_j)\right), \tag{5.2}$$

	$Z_i = 1$	$Z_i = 2$	$Z_i = 3$
$\theta_i = 1$	0.3	0.1	0.1
$\theta_i = 2$	0.1	0.1	0.3

Table 1: Relative joint frequencies of Z_i, θ_i used in algorithm comparison

where $0 \leq \beta \leq 0.5$ is varied in increments of 0.05. It is clear to see that these networks follow the ACSBM form of Definition 2, using an identity link. (The identity link is the link function supported in Mu et al., 2020.) Given the size and density of these networks, the spectral embeddings are well concentrated. As a result, misclassification of θ by a given algorithm can largely be attributed to a failure of the algorithm.

The mean misclassification rate for each algorithm and choice of β is plotted in Figure 2. Algorithm 1 ("ACSBM") performs well across all choices of β, showing error only in 1 of 100 simulated networks at $\beta = 0$. In sharp contrast, the algorithm of Mele et al. (2019) performs poorly across all choices of β, owing to the fact that $P(Y_{ij} = 1 \mid \theta_i = \theta_j = k, Z_i = Z_j = \ell)$ is constant in k. The performance of the remaining algorithms depends on the magnitude of the covariate effect β. The "Vanilla" algorithm performs well only when β is smaller than the latent homophily effect, i.e., when $\beta < 0.3$. On the other hand, Mu et al. (2020) performs well when β is sufficiently large. The difficulties for small β appear to be explained by biased covariate effect estimates, which are fed into the final clustering in Mu et al. (2020). Figure 3 depicts the bias in estimating β via Algorithm 1 vs. Mu et al. (2020). Mu

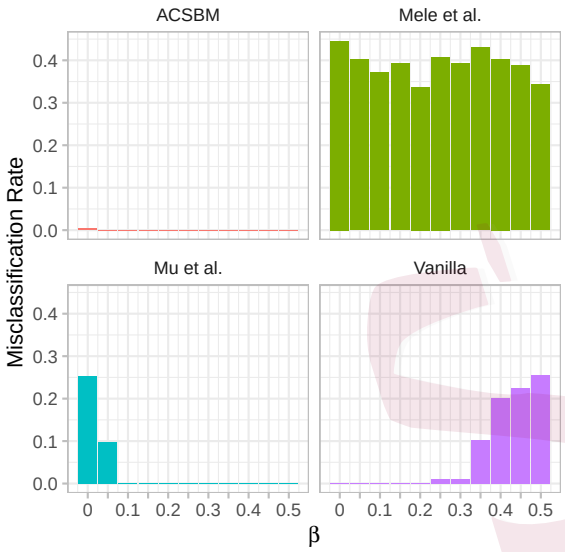

Figure 2: Mean misclassification rate vs. covariate effect β on simulations of the model from Eq. 5.2 for four different spectral methods. In contrast to the alternatives, Algorithm 1 ("ACSBM") performs well across the full range of covariate effects.

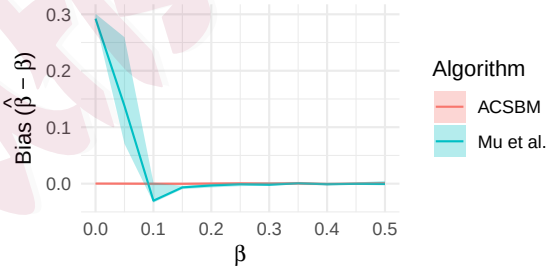

Figure 3: Median (and IQR) bias in estimates of covariate effect β from Algorithm 1 ("ACSBM") vs. Mu et al. (2020). Mu et al. (2020) exhibits substantial bias when β is small.

et al. (2020) exhibits a strong bias on small covariate effects.

6. Application to Harvard Facebook Data

We illustrate the real-world value of our method on a network of Facebook friendships between Harvard University students. The network we consider is a subgraph of a network originally published in Traud et al. (2012), consisting of Facebook friendships between 15,126 individuals from Harvard. Of the 5,970 profiles known to belong to students that declare their gender and a class year between 2006 and 2009, we restrict our attention to the largest connected component of $n = 5{,}917$ nodes and 629,864 edges. Letting $Z_{*1} \in [2]^n$ denote the genders of the students and $Z_{*2} \in [4]^n$ denote the four class years (suitably recoded), we consider a model with $K = 2$ latent communities. This choice of latent dimension was informed by the goodness of fit considered for several small values of K. (An array of methods for selecting K in the setting of stochastic block models can be used in this selection process for the ACSBM. See Lei, 2016; Hu et al., 2021; Li et al., 2020; Jin et al., 2023.)

We apply Algorithm 1, using as our clustering method a version of K-means that operates over the rows of $\hat{X}_Y$ normalized to have unit length. This is inspired by a popular method employed in fitting degree-corrected block models (Karrer and Newman, 2011; Lei and Rinaldo, 2015) and allows for varying node degrees within subcommunities. We denote the resulting latent communities $\hat{\theta}_{\text{covariate}}$. The first of these groups contains 3,772 nodes, and the second contains 2,145. The estimated $\tilde{B}$

matrix is depicted in Figure 4, where the latent structure is represented by the four main quadrants of the matrix, and the homophily patterns between gender and class year are captured by the repetitive structures within each of these quadrants.

We estimate the coefficients of the ACSBM model by fitting a logistic regression model using $\hat{\theta}_{\text{covariate}}$ as a plug-in estimator for θ and the empirical edge counts from $\hat{\hat{B}}$. The resulting estimated model is:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}}\right) = &\begin{bmatrix} -5.077 & -5.520 \\ -5.520 & -4.290 \end{bmatrix}_{\theta_i\theta_j} \\ &+ 0.102\,\mathbb{I}(\text{gender}_i = \text{gender}_j) \\ &+ 2.113\,\mathbb{I}(\text{year}_i = \text{year}_j),\end{aligned}$$

where π_{ij} denotes the probability of an edge occurring between nodes i and j. We interpret this as follows: after accounting for the latent structure we have discovered, sharing the same gender (or class year) is associated with odds of forming a friendship that are $e^{0.102} \approx 1.107$ times (or $e^{2.113} \approx 8.273$ times, respectively) higher, when holding the other variables constant for the pair. The remaining coefficients in the 2×2 matrix represent intercept terms that vary depending on the clusters to which a pair of nodes belongs. We note that the overall differences between these coefficients lie somewhere between the effect size of gender and that of class year. All of this matches our intuition from Figure 4, where differences between same-sex and opposite-sex friendship patterns are slight, differences by class year are stark, and the latent

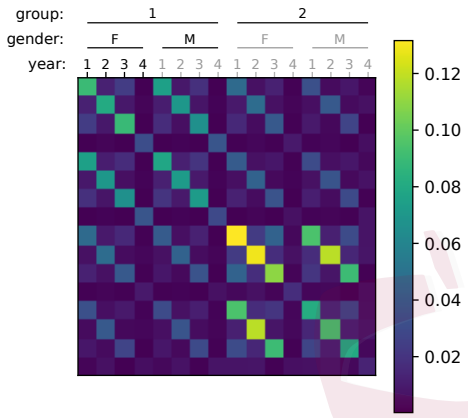

Figure 4: Estimated $\tilde{B}$ matrix applying Algorithm 1 to a network of Harvard students, accounting for observed covariates of class year and gender

structure contributes modestly.

The question of what the latent structure represents has no easy answer, since it corresponds to an unobserved feature of the network. We note, however, a curious correlation between this latent feature and class year. For students with class years of 2006, 2007, 2008, and 2009, the proportions assigned to latent group 1 are, respectively, 53.5%, 50.5%, 60.8%, and 90.1%. Noting that incoming freshmen at Harvard are traditionally assigned to designated housing and offered unique social programs, we conjecture that the latent feature captured by our method corresponds to residential and social patterns that correlate with the freshman experience, while going beyond what is captured by class year alone.

Comparison with vanilla spectral clustering. Comparing $\hat{\theta}_{\text{covariate}}$ to the

communities discovered through a spectral clustering process that ignores the covariates highlights the value of our proposed method. We compare our method to the same "Vanilla" method considered in Section 5, i.e., a K-means-based version of Lei and Rinaldo (2015). The resulting communities we denote $\hat{\theta}_{\text{vanilla}}$. As expected from the results of previous studies (e.g., Chen et al., 2018), the "latent" communities discovered in $\hat{\theta}_{\text{vanilla}}$ largely recover information already contained in our observed covariates: with $K = 2$, $\hat{\theta}_{\text{vanilla}}$ almost perfectly predicts whether the student is a freshman ($\text{year}_i = 4$), with agreement between the latent community and freshman status on 98.5% of nodes. Expanding the search to $K = 4$ recovers clusters that agree with class year on 95.3% of nodes (up to a permutation of labels). At $K = 8$, we begin to discover a latent structure: the recovered clusters $\hat{\theta}_{\text{vanilla}} \in [8]^n$ agree with a suitably recoded combination of class year and our $\hat{\theta}_{\text{covariate}}$ on 71.8% of nodes. However, recovering a hierarchical structure from these labels remains an elusive task. This is precisely the challenge addressed by our algorithm, which first identifies a flattened set of "subcommunities" in Algorithm 1, Part 1, before working backward to recover the full hierarchical structure in Part 3.

7. Discussion

The task of separating latent from observed structure in networks is critical to a variety of network inference tasks. The method we have proposed is, to our knowledge, the first to offer a rigorous guarantee of consistency of latent structure recovery using

spectral clustering in the setting where edge formation is dependent on both observed and latent factors. Our proposed method is computationally efficient and theoretically appealing, using distance in latent space as a means of reconnecting a network partitioned by observed covariates.

We would like to note the limitations of our current work and highlight opportunities for future research. First and foremost, the combinatorial nature of the algorithm restricts its use to discrete covariates. Moreover, since Part 3 of the algorithm estimates permutations over network partitions, any error in permutation selection is likely to introduce considerable error in the final clustering of nodes. We believe that a post-processing step akin to spectral clustering with adjustment (SCWA) of Huang and Feng (2018) (or the correction of Mu et al. 2020 to the results of Mele et al. 2019) holds promise to mitigate finite-sample permutation errors, but we have not yet given this formal study. Finally, while we consider only a fixed number of latent communities and covariates, it would be useful to extend our analysis to the case where the dimension of the ACSBM network grows. Based on existing results for SBM recovery (e.g., Choi et al., 2012), we anticipate the total number of subcommunities of Proposition 1 is limited to $K\tilde{L} = O(\sqrt{n})$. It would be interesting, but well outside the scope of this paper, to extend these ideas to a continuous setting, which may alleviate these limitations.

We believe that our proposed method offers promise beyond what has been proven so far. As an example, the simulations of Section 5 suggest consistency for a wide

range of link functions that remains to be rigorously proven. Additionally, the analysis of Harvard students' Facebook friendships in Section 6 employed a degree-corrected post-processing step for the spectral embeddings, in a manner commonly employed with the degree-corrected block model (DCBM) of Karrer and Newman (2011). While the consistency of our method has not been rigorously proven in conjunction with degree correction, the empirical results look promising, as the method successfully identified a latent structure distinct from the covariates considered. Moreover, there is theoretical intuition behind this method. The matching problem of Algorithm 1, Part 3 may be recast as an optimization over the *angles* between subcommunities in latent space, while the latent positions in a degree-corrected analog of the ACSBM would be expected to fall along distinct rays corresponding to subcommunities. Such a theoretical extension for degree correction would greatly expand the practicality of the model we consider, allowing for nodes to exhibit greater variation in node degree, as commonly seen in observed networks, while retaining the simplicity and flexibility of the underlying latent block model structure.

Supplementary Materials

Proofs and derivations of all results are provided in the online supplementary material. A Python implementation of our proposed method and additional examples are available at `https://github.com/jonhehir/acsbm`.

Acknowledgements

Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. This work was supported in part by NIAID/NIH R01-AI136664 and NSF Award No. SES-1853209 to The Pennsylvania State University.

References

Abbe, E., J. Fan, K. Wang, and Y. Zhong (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics 48*(3), 1452–1474.

Agterberg, J., M. Tang, and C. E. Priebe (2020). On two distinct sources of nonidentifiability in latent position random graph models. *arXiv preprint arXiv:2003.14250*.

Athreya, A., D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research 18*(1), 8393–8484.

Athreya, A., C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman (2016). A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A 78*(1), 1–18.

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics 41*(4), 1922–1943.

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

Chen, Y., X. Li, and J. Xu (2018). Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics 46*(4), 1573–1602.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012). Stochastic blockmodels with a growing number of classes. *Biometrika 99*(2), 273–284.

Deshpande, Y., S. Sen, A. Montanari, and E. Mossel (2018). Contextual stochastic block models. *Advances in Neural Information Processing Systems 31*.

Edmonds, J. and R. M. Karp (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM) 19*(2), 248–264.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Goodreau, S. M., J. A. Kitts, and M. Morris (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography 46*(1), 103–125.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(2), 301–354.

Henry, A. D., P. Prałat, and C.-Q. Zhang (2011). Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences 108*(21), 8605–8610.

Hoff, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in neural information processing systems 20*.

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

Hu, J., J. Zhang, H. Qin, T. Yan, and J. Zhu (2021). Using maximum entry-wise deviation to test the goodness of fit for stochastic block models. *Journal of the American Statistical Association 116*(535), 1373–1382.

Huang, S. and Y. Feng (2018). Pairwise covariates-adjusted block model for community detection. *arXiv preprint arXiv:1807.03469*.

Huber, G. A. and N. Malhotra (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics 79*(1), 269–283.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, 422–447.

Jin, J., Z. T. Ke, S. Luo, and M. Wang (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association 118*(543), 2101–2116.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical review E 83*(1), 016107.

Lee, Y. and E. L. Ogburn (2021). Network dependence can lead to spurious associations and invalid inference. *Journal of the American Statistical Association 116*(535), 1060–1074.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics 44*(1), 401 – 424.

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

Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

Lyzinski, V., D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic journal of statistics 8*(2), 2905–2922.

Ma, Z., Z. Ma, and H. Yuan (2020). Universal latent space model fitting for large networks with edge covariates. *J. Mach. Learn. Res. 21*, 4–1.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology 27*(1), 415–444.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE.

Mele, A., L. Hao, J. Cape, and C. E. Priebe (2019). Spectral inference for large stochastic blockmodels with nodal covariates. *arXiv preprint arXiv:1908.06438*.

Mu, C., A. Mele, L. Hao, J. Cape, A. Athreya, and C. E. Priebe (2020). On spectral algorithms for community detection in stochastic blockmodel graphs with vertex covariates. *arXiv preprint arXiv:2007.02156*.

Newman, M. E. and A. Clauset (2016). Structure and inference in annotated networks. *Nature communications 7*(1), 1–11.

Peel, L., D. B. Larremore, and A. Clauset (2017). The ground truth about metadata and community detection in networks. *Science advances 3*(5), e1602548.

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

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

of Computational and Graphical Statistics 28(3), 609–619.

Rubin-Delanchy, P., J. Cape, M. Tang, and C. E. Priebe (2017). A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*.

Shalizi, C. R. and A. C. Thomas (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research 40*(2), 211–239.

Shrum, W., N. H. Cheek Jr, and S. MacD (1988). Friendship in school: Gender and racial homophily. *Sociology of Education*, 227–239.

Smith, K. P. and N. A. Christakis (2008). Social networks and health. *Annu. Rev. Sociol 34*, 405–429.

Snijders, T. A. and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification 14*(1), 75–100.

Su, L., W. Wang, and Y. Zhang (2019). Strong consistency of spectral clustering for stochastic block models. *IEEE Transactions on Information Theory 66*(1), 324–338.

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

Tallberg, C. (2004). A bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology 29*(1), 1–23.

Tang, M., J. Cape, and C. E. Priebe (2022). Asymptotically efficient estimators for stochastic blockmodels: The naive mle, the rank-constrained mle, and the spectral estimator. *Bernoulli 28*(2), 1049–1073.

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

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Vu, D. Q., D. R. Hunter, and M. Schweinberger (2013). Model-based clustering of large networks. *The annals of applied statistics 7*(2), 1010.

Weng, H. and Y. Feng (2021). Community detection with nodal information: likelihood and its variational approximation. *Stat*, e428.

Yang, J., J. McAuley, and J. Leskovec (2013). Community detection in networks with node attributes. In *2013 IEEE 13th international conference on data mining*, pp. 1151–1156. IEEE.

Young, S. J. and E. R. Scheinerman (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pp. 138–149. Springer.

Zhang, Y., K. Chen, A. Sampson, K. Hwang, and B. Luna (2019). Node features adjusted stochastic block model. *Journal of Computational and Graphical Statistics 28*(2), 362–373.

Department of Statistics, Penn State University, University Park, PA, USA.

E-mail: jh@psu.edu

Department of Statistics, Penn State University, University Park, PA, USA.

E-mail: xiaoyue@psu.edu

Department of Statistics, Penn State University, University Park, PA, USA.

E-mail: sesa@psu.edu