Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

PCABM: Pairwise Covariates-Adjusted Block Model for Community Detection

Sihan Huang, Jiajin Sun & Yang Feng

To cite this article: Sihan Huang, Jiajin Sun & Yang Feng (2024) PCABM: Pairwise Covariates-Adjusted Block Model for Community Detection, Journal of the American Statistical Association, 119:547, 2092-2104, DOI: 10.1080/01621459.2023.2244731

To link to this article: https://doi.org/10.1080/01621459.2023.2244731

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 547, 2092–2104: Theory and Methods
https://doi.org/10.1080/01621459.2023.2244731

PCABM: Pairwise Covariates-Adjusted Block Model for Community Detection

Sihan Huang[a], Jiajin Sun[a], and Yang Feng[b]

[a]Department of Statistics, Columbia University, New York, NY; [b]Department of Biostatistics, New York University, New York, NY

ABSTRACT

One of the most fundamental problems in network study is community detection. The stochastic block model (SBM) is a widely used model, and various estimation methods have been developed with their community detection consistency results unveiled. However, the SBM is restricted by the strong assumption that all nodes in the same community are stochastically equivalent, which may not be suitable for practical applications. We introduce a pairwise covariates-adjusted stochastic block model (PCABM), a generalization of SBM that incorporates pairwise covariate information. We study the maximum likelihood estimators of the coefficients for the covariates as well as the community assignments, and show they are consistent under suitable sparsity conditions. Spectral clustering with adjustment (SCWA) is introduced to efficiently solve PCABM. Under certain conditions, we derive the error bound of community detection for SCWA and show that it is community detection consistent. In addition, we investigate model selection in terms of the number of communities and feature selection for the pairwise covariates, and propose two corresponding algorithms. PCABM compares favorably with the SBM or degree-corrected stochastic block model (DCBM) under a wide range of simulated and real networks when covariate information is accessible. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2020
Accepted July 2023

KEYWORDS
Community detection; Consistency; Covariates-adjusted; Model selection; Network; Spectral clustering with adjustment

1. Introduction

Networks are used to represent connections among subjects within a population of interest, and their wide range of applications has drawn researchers from various fields. In social media, network analysis can reveal people's behaviors and interests through their connections, such as Facebook friends and Twitter followers. In ecology, a food web depicting predator–prey interactions offers valuable insights into individual habits and the structure of biocoenosis. Network analysis also has extensive applications in computer science, biology, physics, and economics (Newman 1963; Getoor and Diehl 2005; Goldenberg et al. 2010; Graham 2014).

Community detection, one of the most studied problems for network data, is concerned with identifying groups of nodes that are densely connected within groups and sparsely connected between groups. Detecting network communities not only aids in understanding the structural features of networks, but also has practical applications. For instance, communities in social networks often share similar interests, which can help the development of recommendation systems. Community detection methods primarily fall into two categories: algorithm-based and model-based. Algorithm-based methods (Newman 2006; Bickel and Chen 2009; Zhao, Levina, and Zhu 2011; Wilson et al. 2014, 2017) involve devising an objective function (e.g., modularity) and optimizing it for community detection, while model-based methods assume that edges are generated from a probabilistic model. Popular models include the stochastic block model (Holland, Laskey, and Leinhardt 1983), mixture model (Newman and Leicht 2007), degree-corrected stochastic block model (Karrer and Newman 2011), latent space model (Hoff, Raftery, and Handcock 2002; Handcock, Raftery, and Tantrum 2007; Hoff 2008), and hypergraph block model (Ghoshdastidar and Dukkipati 2014; Yuan et al. 2022). For a comprehensive review of statistical network models, refer to Goldenberg et al. (2010) and Fortunato (2010).

The classical stochastic block model (SBM) posits that the connection between each pair of nodes depends solely on their community labels. For SBM, community detection consistency has been established for various methods, such as modularity maximization (Newman 2006), profile likelihood (Bickel and Chen 2009; Choi, Wolfe, and Airoldi 2012), spectral clustering (Rohe, Chatterjee, and Yu 2011; Lei and Rinaldo 2015), variational inference (Bickel et al. 2013), and penalized local maximum likelihood estimation (Gao et al. 2017), among others. However, in real-world scenarios, node connections may depend not only on community structure but also on nodal or pairwise covariates. For example, in an ecological network, predator–prey links between species could be influenced by factors such as prey types, habits, body sizes, and living environments. By incorporating nodal and pairwise information into network models, a more accurate community structure can be obtained.

Depending on the relationship between communities and covariates, there are generally two classes of models, as depicted in Figure 1: *covariates-adjusted* and *covariates-confounding*.

CONTACT Yang Feng *yang.feng@nyu.edu* Department of Biostatistics, New York University, New York, NY
Sihan Huang and Jiajin Sun have contributed equally to this work.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

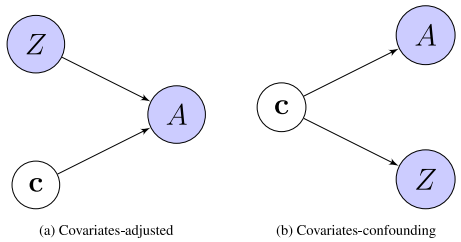


Figure 1. Two different network models including covariates.

The symbols $\mathbf{c}$, Z, and A represent latent community labels, pairwise covariates, and the adjacency matrix, respectively. In Figure 1(a), the latent community and covariates jointly determine the network structure. One example of this model is the friendship network among students. Students may become friends for various reasons, such as being in the same class, sharing hobbies, or belonging to the same ethnic group. Without adjusting for these covariates, it is difficult to infer a single community membership from A. We will analyze one such example in detail in Section 8. Conversely, covariates may carry the same community information as the adjacency matrix, as shown in Figure 1(b). The term "confounding" originates from graph models (Greenland, Robins, and Pearl 1999). The citation network serves as an excellent example of this model (Tan, Chan, and Zheng 2016). When research topics are treated as community labels for articles, citation links largely depend on the research topics of the article pair. Simultaneously, the distribution of keywords is likely driven by the specific topic an article addresses.

Researchers often modify the SBM in the above two ways to incorporate covariate information. For the covariates-confounding model, Newman and Clauset (2016) uses covariates to construct the prior for community labels and then generates edges using a degree-corrected model. Zhang, Levina, and Zhu (2016) proposes a joint community detection criterion, an analog of modularity, to incorporate nodal features. Deshpande et al. (2018) establishes information-theoretic bounds for combining a block model and a spike covariance model that are conditionally independent given class assignments. Yan and Sarkar (2021) suggests a semidefinite programming framework to aggregate network and covariate information, while Xu, Zhen, and Wang (2022) considers an augmented adjacency tensor approach under an analogous setting in multilayer SBM. Weng and Feng (2022) employs a logistic model as the prior for community labels. For the covariates-adjusted model, Yan et al. (2019) proposes a directed network model with a logistic function, but it does not consider potential community structures. Wu, Levina, and Zhu (2017) introduces a generalized linear model with low-rank effects to model network edges, which could imply a community structure or a latent space structure, although not explicitly mentioned; Ma, Ma, and Yuan (2020) presents algorithms for a latent space model that incorporates edge covariates; both of these works consider penalized MLE with convex relaxation and gradient-based algorithms.

In this work, we propose a simple yet effective model called Pairwise Covariates-Adjusted Stochastic Block Model (PCABM), which extends the SBM by adjusting the probability of connections according to the contribution of pairwise covariates.[1] Through this model, we can learn how each covariate affects the connections by examining its corresponding regression coefficient, for which asymptotic normality is established. In addition, we investigate the likelihood-based community detection method and propose an efficient pseudo-likelihood expectation-maximization (PLEM) algorithm. Consistency results for both the MLE and the PLEM algorithm are provided. Apart from likelihood methods, we also propose a novel spectral clustering method for PCABM. We prove desirable theoretical properties for the spectral clustering method, and demonstrate that, as a fast algorithm, using it as an initial estimator for the likelihood method results in more accurate community detection than random initialization. Furthermore, we consider the model selection problems of estimating the number of communities and selecting the important confounding covariates, providing algorithms to address these two issues based on the edge cross-validation framework proposed by Li, Levina, and Zhu (2020).

The remainder of the article is organized as follows. In Section 2, we introduce the PCABM. We then present the asymptotic properties of the coefficient estimates in Section 3. After that, we introduce two methods for community detection: a likelihood approach in Section 4 and a spectral approach in Section 5. In addition, we present two algorithms for model selection in Section 6. Simulations and applications on real networks are discussed in Sections 7 and 8, respectively. We conclude the article with a brief discussion in Section 9. All proofs are relegated to the supplementary materials.

Here, we introduce some notations to facilitate the discussion. For a square matrix $M \in \mathbb{R}^{n\times n}$, let $\|M\|$ be the operator norm of M, $\|M\|_F = \sqrt{\text{trace}(M^TM)}$, $\|M\|_\infty = \max_i \sum_{j=1}^n |M_{ij}|$, $\|M\|_0 = \#\{(i,j)|M_{ij} \neq 0\}$, and $\|M\|_{\max} = \max_{ij} |M_{ij}|$. $\lambda_{\min}(M)$ is the minimum eigenvalue of M. For index sets $I, J \subset [n] := \{1, 2, \ldots, n\}$, $M_{I\cdot}$ and $M_{\cdot J}$ are the submatrices of M consisting the corresponding rows and columns, respectively. For a vector $\mathbf{x} \in \mathbb{R}^n$, let $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_\infty = \max_i |x_i|$. We define the Kronecker power by $\mathbf{x}^{\otimes(k+1)} = \mathbf{x}^{\otimes k} \otimes \mathbf{x}$, where $\otimes$ is the Kronecker product.

For any positive integer K, we define $I_K \in \mathbb{R}^{K\times K}$ to be the identity matrix and $\mathbf{1}_K$ to be the all-one vector. When there is no confusion, we will sometimes omit the subscript K. For a vector $\mathbf{x} \in \mathbb{R}^K$, $D(\mathbf{x}) \in \mathbb{R}^{K\times K}$ represents the diagonal matrix whose diagonal elements take the value of $\mathbf{x}$. For an event A, its indicator function is written as $\mathbb{1}(A)$. For two real number sequences x_n and y_n, we say $x_n = o(y_n)$ or $y_n = \omega(x_n)$ if $\lim_{n\to\infty} x_n/y_n = 0$, $x_n = O(y_n)$ or $y_n = \Omega(x_n)$ if $\limsup_{n\to\infty} |x_n/y_n| \leq \infty$.

2. Pairwise Covariates-Adjusted Stochastic Block Model

We consider a graph with n nodes and K communities, where K could be fixed or increase with n. In this article, we focus on

[1] Note that these are "edge-level" covariates instead of the nodal or vertex-level covariates that are often considered in other parts of the literature. Having said that, one can incorporate nodal information into our model by converting it into pairwise covariates, where an example will be presented in Section 8.1.

undirected weighted graphs without self-loops. All edge information is incorporated into a symmetric adjacency matrix $A = [A_{ij}] \in \mathbb{N}^{n\times n}$ with diagonal elements being zero, where $\mathbb{N}$ represents the set of nonnegative integers. The total number of possible edges is denoted by $N_n = n(n-1)/2$. The true node labels $\mathbf{c} = \{c_1, \ldots, c_n\} \in \{1, \ldots, K\}^n$ are drawn independently from a multinomial distribution with parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, where $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$ for all k. The community detection problem aims to find a disjoint partition of the nodes, or equivalently, estimated node labels $\mathbf{e} = \{e_1, \ldots, e_n\} \in \{1, \ldots, K\}^n$ that is close to $\mathbf{c}$, where $e_i \in \{1, \ldots, K\}$ is the label for node i.

In the classical SBM, we assume $\Pr(A_{ij} = 1|\boldsymbol{c}) = B_{c_i c_j}$, where $B = [B_{ab}] \in [0,1]^{K\times K}$ is a symmetric matrix with no identical rows. In practice, the connection between two nodes may depend not only on the communities they belong to, but also on the nodal information (e.g., gender, age, religion). To fix the idea, assume in addition to A, we have observed a pairwise p-dimensional vector $\mathbf{z}_{ij}$ between nodes i and j. Denote the collection of the pairwise covariates among nodes as $Z = [\mathbf{z}_{ij}^T] \in \mathbb{R}^{n^2\times p}$. Here, we assume $\mathbf{z}_{ij} = \mathbf{z}_{ji}$ and $\mathbf{z}_{ii} = \mathbf{0}$, for all i and j.

Now, we are ready to introduce the Pairwise Covariates-Adjusted Stochastic Block Model (PCABM). For $i < j$, conditional on the community label $\mathbf{c}$ and the pairwise covariate matrix Z, A_{ij}'s are independent and

$$A_{ij} \sim \text{Poisson}(\lambda_{ij}),\ \lambda_{ij} = B_{c_i c_j} e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0},$$

where $\boldsymbol{\gamma}^0$ is the true coefficient vector for the pairwise covariates. In addition to the goal of recovering the community membership vector $\mathbf{c}$, we would also like to get an accurate estimate for $\boldsymbol{\gamma}^0$.

The specific term $\exp(\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0)$ is introduced here to adjust the connectivity between nodes i and j. Here, as in the vanilla SBM, we assume a sparse setting for $B = \rho_n \bar{B}$, with $\bar{B}$ fixed and $\rho_n \to 0$ as $n \to \infty$. Note that due to the contribution of Z, $\varphi_n = n\rho_n$ is no longer the expected degree as in the vanilla SBM (Zhao, Levina, and Zhu 2012), but it is still useful as a measure of the network sparsity. It is easy to observe that when $\boldsymbol{\gamma}^0 = 0$, PCABM reduces into the vanilla Poisson SBM.

Under PCABM, the likelihood function is

$$\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, B, \boldsymbol{\pi}|A, Z) \propto \prod_{i=1}^n \pi_{e_i} \prod_{i<j} B_{e_i e_j}^{A_{ij}} e^{A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma}} \exp\left(-B_{e_i e_j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}}\right).$$

Define

$$n_k(\mathbf{e}) = \sum_{i=1}^n \mathbb{1}(e_i = k),\ O_{kl}(\mathbf{e}) = \sum_{ij} A_{ij}\mathbb{1}(e_i = k, e_j = l),$$
$$E_{kl}(\mathbf{e}, \boldsymbol{\gamma}) = \sum_{i\neq j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}}\mathbb{1}(e_i = k, e_j = l) = \sum_{(i,j)\in s_{\mathbf{e}}(k,l)} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}},$$

where $s_{\mathbf{e}}(k,l) = \{(i,j)|e_i = k, e_j = l, i \neq j\}$. Under the assignment $\mathbf{e}$, $n_k(\mathbf{e})$ represents the number of nodes estimated to be in the community k. For $k \neq l$, O_{kl} is the total number of edges between estimated communities k and l; for $k = l$, O_{kk} is twice the number of edges within estimated community k. E_{kl} is the summation of all pair-level factors between estimated communities k and l. Up to a constant term, we can write the log-likelihood function as

$$\log\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, B, \boldsymbol{\pi}|A, Z) = \sum_k n_k(\mathbf{e})\log\pi_k + \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log B_{kl} - \frac{1}{2}\sum_{kl} B_{kl}E_{kl}(\mathbf{e}, \boldsymbol{\gamma}) + \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma}.$$

Given $\mathbf{e}$ and $\boldsymbol{\gamma}$, we derive the MLE $\hat{\pi}_k(\mathbf{e}) = \frac{n_k(\mathbf{e})}{n}$ and $\hat{B}_{kl}(\mathbf{e}, \boldsymbol{\gamma}) = \frac{O_{kl}(\mathbf{e})}{E_{kl}(\mathbf{e},\boldsymbol{\gamma})}$. Plugging $\hat{B}(\mathbf{e}, \boldsymbol{\gamma})$ and $\hat{\boldsymbol{\pi}}(\mathbf{e})$ into the original log-likelihood and discarding the constant terms, we have

$$\log\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, \hat{B}, \hat{\boldsymbol{\pi}}|A, Z) \propto \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log\frac{O_{kl}(\mathbf{e})}{E_{kl}(\mathbf{e}, \boldsymbol{\gamma})} + \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma} + \sum_k n_k(\mathbf{e})\log\frac{n_k(\mathbf{e})}{n}. \quad (1)$$

Out target is to maximize (1) with respect to $\mathbf{e}$ and $\boldsymbol{\gamma}$. We consider a two-step sequential estimation procedure by first studying the estimation of $\boldsymbol{\gamma}^0$ in Section 3 and then the estimation of $\mathbf{c}$ in Section 4 (likelihood method) and Section 5 (spectral method).

It is worth mentioning that the proposed model includes DCBM in the following sense: by choosing $p = 1, \mathbf{z}_{ij} = \log(d_i d_j)$ and $\boldsymbol{\gamma} = 1$ where d_i is the degree of node i, (1) becomes

$$\log\mathcal{L}\left(\mathbf{e}, \boldsymbol{\gamma} = 1, \hat{B}, \hat{\boldsymbol{\pi}}|A, Z = (\log(d_i d_j))_{n^2\times 1}\right) \propto \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log\frac{O_{kl}(\mathbf{e})}{n_k(\mathbf{e})n_l(\mathbf{e})},$$

which is exactly the profile log-likelihood under DCBM derived by maximizing over "θ and P" (degree parameter and block connection probability) in DCBM. From this perspective, one can view PCABM as a generalization of DCBM.

3. Estimation of Coefficients for Pairwise Covariates

As the first step to maximize the log-likelihood, we consider the estimation of coefficients $\boldsymbol{\gamma}^0$ for pairwise covariates. To this end, we impose the following conditions on Z.

Condition 1. $\{\mathbf{z}_{ij}, i < j\}$ are iid and uniformly bounded, that is, for $\forall i < j$, $\|\mathbf{z}_{ij}\|_\infty \le \zeta$, where $\zeta > 0$ is some constant. $\|\boldsymbol{\gamma}^0\|_1$ is also bounded by a constant. Denote $\xi = \exp(\zeta\|\boldsymbol{\gamma}^0\|_1)$.

Remark 1. The bounded support condition for $\mathbf{z}_{ij}$ is introduced to simplify the proof. It could be relaxed to $\mathbf{z}_{ij}$ to have a light tail or to allow the upper bound to grow slowly with network size n. For example, our proofs could still go through if $\exp(\mathbf{z}_{ij}^\top\boldsymbol{\gamma}^0)$ follows a sub-Gaussian distribution (with $\|\boldsymbol{\gamma}^0\|_1$ bounded), under slightly stronger conditions on the sparsity of the network.

Under Condition 1, the following expectations exist: $\theta(\boldsymbol{\gamma}^0) \equiv \mathbb{E}e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0} \in \mathbb{R}^+$, $\boldsymbol{\mu}(\boldsymbol{\gamma}^0) \equiv \mathbb{E}\mathbf{z}_{ij}e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0} \in \mathbb{R}^p$, and $\Sigma(\boldsymbol{\gamma}^0) \equiv \mathbb{E}\mathbf{z}_{ij}\mathbf{z}_{ij}^T e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0} \in \mathbb{R}^{p\times p}$. To ensure that $\boldsymbol{\gamma}^0$ is the unique solution to maximize the likelihood in the population version, we impose the following regularity condition at the true $\boldsymbol{\gamma}^0$.

Condition 2. $\Sigma(\boldsymbol{\gamma}^0) - \theta(\boldsymbol{\gamma}^0)^{-1}\boldsymbol{\mu}(\boldsymbol{\gamma}^0)^{\otimes 2}$ is positive definite.

Remark 2. To understand the implication of Condition 2, consider the function $g(\boldsymbol{\gamma}) = \theta(\boldsymbol{\gamma})\Sigma(\boldsymbol{\gamma}) - \boldsymbol{\mu}(\boldsymbol{\gamma})^{\otimes 2}$. In the special case of SBM where $\boldsymbol{\gamma}^0 = \mathbf{0}$, we have $g(\mathbf{0}) = \mathbb{E}[\mathbf{z}^{\otimes 2}] - \mathbb{E}[\mathbf{z}]^{\otimes 2} = \mathrm{cov}(\mathbf{z})$. To avoid multicollinearity, it's natural for us to require $\mathrm{cov}(\mathbf{z})$ to be positive definite. For a general PCABM, we require $g(\boldsymbol{\gamma})$ to be positive definite at the true value $\boldsymbol{\gamma}^0$.

For a given initial community assignment $\boldsymbol{e}_0$, denote by $\ell_{\mathbf{e}_0}$ the log-likelihood terms in (1) containing $\boldsymbol{\gamma}$, which is

$$\ell_{\mathbf{e}_0}(\boldsymbol{\gamma}) \equiv \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma} - \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e}_0)\log E_{kl}(\mathbf{e}_0, \boldsymbol{\gamma}).$$

We consider the following estimate:

$$\hat{\boldsymbol{\gamma}}(\boldsymbol{e}_0) = \arg\max_{\boldsymbol{\gamma}} \ell_{\mathbf{e}_0}(\boldsymbol{\gamma}). \tag{2}$$

We point out that $\ell_{\mathbf{e}_0}(\boldsymbol{\gamma})$ is concave in $\boldsymbol{\gamma}$, so the global optimizer in (2) can be efficiently solved by a BFGS algorithm. When there is no ambiguity, we will just write it as $\hat{\boldsymbol{\gamma}}$. As we will see in the theory that under some mild conditions, the asymptotic property does not depend on the choice of $\mathbf{e}_0$. In fact, one could simply choose $\mathbf{e}_0 = \mathbf{1}_n$, the all-one vector, when estimating $\boldsymbol{\gamma}^0$.

To accommodate the "K growing with n" case, we also need the following stability condition.

Condition 3. $\bar{B}_{\lim} = \lim_{n\to\infty}\sum_{a=1}^{K}\sum_{b=1}^{K}\pi_a\pi_b\bar{B}_{ab}$ exists.

Remark 3. Note that when K is fixed, Condition 3 is automatically satisfied. When K grows with n, we need the $\boldsymbol{\pi}$-weighted average of matrix $\bar{B}$ to have a limit. This is a mild condition since, otherwise, the sequence of observed graphs indexed by n does not come from a consistent data generating process.

Now we are ready to present the consistency and asymptotic normality of $\hat{\boldsymbol{\gamma}}$.

Theorem 1 (Consistency and asymptotic normality of MLE of $\boldsymbol{\gamma}$). Under PCABM, assume Conditions 1, 2, and 3 hold, where the number of communities K could either be fixed or grow to ∞ at an arbitrary rate. Then fixing $\mathbf{e}_0 = \mathbf{1}_n$, as $n \to \infty$, if $N_n\rho_n \to \infty$ and $\rho_n \to 0$, we have $\hat{\boldsymbol{\gamma}}(\mathbf{e}_0) \xrightarrow{p} \boldsymbol{\gamma}^0$ and

$$\sqrt{N_n\rho_n}\left[\hat{\boldsymbol{\gamma}}(\mathbf{e}_0) - \boldsymbol{\gamma}^0\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_\infty^{-1}(\boldsymbol{\gamma}^0)), \tag{3}$$

where $\Sigma_\infty(\boldsymbol{\gamma}^0) = \bar{B}_{\lim}[\Sigma(\boldsymbol{\gamma}^0) - \theta(\boldsymbol{\gamma}^0)^{-1}\boldsymbol{\mu}(\boldsymbol{\gamma}^0)^{\otimes 2}]$.

Different from Yan et al. (2019), in which the network is dense, the convergence rate is $\sqrt{N_n\rho_n}$ rather than $\sqrt{N_n}$ since the effective number of edges is reduced from N_n to $N_n\rho_n$. The asymptotic covariance matrix $\Sigma_\infty^{-1}(\boldsymbol{\gamma}^0)$ depends on $\theta(\boldsymbol{\gamma}^0)$, $\boldsymbol{\mu}(\boldsymbol{\gamma}^0)$, and $\Sigma(\boldsymbol{\gamma}^0)$, which can be estimated empirically by the plug-in method.

Now, with a consistent estimate of $\boldsymbol{\gamma}^0$, we are ready to study the estimation of $\boldsymbol{c}$. In the next two sections, we will present two different methods for estimating $\boldsymbol{c}$, namely the likelihood-based estimate in Section 4 and the spectral method in Section 5.

4. Likelihood Based Estimate for Community Labels

This section presents a likelihood-based estimate for community labels by maximizing $\log\mathcal{L}$ regarding $\mathbf{e}_0$ with $\hat{\boldsymbol{\gamma}}$ from Section 3. We only present the fixed K setting here, and the results for the growing K scenario are relegated to the supplementary materials, partly because in the class label MLE for growing K, we consider a slightly different regime from $B = \rho_n\bar{B}$: we need the signal-noise-ratio, or approximately in-class probability over between-class probability, to also grow with n and K; and the conditions are imposed on $\sum_{i<j}B_{c_ic_j}$ rather than ρ_n. See Section A.4 in the supplementary materials for details.

We will show that with $\hat{\boldsymbol{\gamma}}(\mathbf{e}_0)$ satisfying the conclusions in Theorem 1, the consistency of $\hat{\mathbf{c}}(\hat{\boldsymbol{\gamma}})$ is guaranteed. Plugging $\hat{\boldsymbol{\gamma}}$ into (1), the log-likelihood function can be rewritten as

$$\ell_{\hat{\boldsymbol{\gamma}}}(\mathbf{e}) = \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log\frac{O_{kl}(\mathbf{e})}{E_{kl}(\mathbf{e}, \hat{\boldsymbol{\gamma}})} + \sum_k n_k(\mathbf{e})\log\frac{n_k(\mathbf{e})}{n}.$$

Then, our maximum likelihood estimate for the community label is

$$\hat{\mathbf{c}} = \hat{\mathbf{c}}(\hat{\boldsymbol{\gamma}}) := \arg\max_{\mathbf{e}} \ell_{\hat{\boldsymbol{\gamma}}}(\mathbf{e}). \tag{4}$$

Note that here we omit $\mathbf{e}_0$ to avoid confusion. Following Zhao, Levina, and Zhu (2012), we consider two versions of community detection consistency. Note that the consistency in community detection is understood under any permutation of the labels. To be more precise, let $\mathcal{P}_K$ be the collection of all permutation functions of $[K]$. (a) We say the label estimate $\hat{\mathbf{c}}$ is *weakly consistent* if $\Pr[n^{-1}\min_{\sigma\in\mathcal{P}_K}\sum_{i=1}^{n}\mathbb{1}(\sigma(\hat{c}_i) \neq c_i) < \varepsilon] \to 1$ for any $\varepsilon > 0$ as $n \to \infty$. (b) We say $\hat{\mathbf{c}}$ is *strongly consistent* if $\Pr[\min_{\sigma\in\mathcal{P}_K}\sum_{i=1}^{n}\mathbb{1}(\sigma(\hat{c}_i) \neq c_i) = 0] \to 1$, as $n \to \infty$. We establish both versions of consistency for MLE $\hat{\mathbf{c}}$ in the following theorem.

Theorem 2. Under PCABM that satisfies the Conditions 1 and 2, when K is fixed, the community label estimate $\hat{\boldsymbol{c}}$ defined in (4) is weakly consistent if $\varphi_n \to \infty$ and strongly consistent if $\varphi_n/\log n \to \infty$, where $\varphi_n = n\rho_n$.

In addition to the fixed K case considered in Theorem 2, we have also shown that $\hat{\mathbf{c}}$ is consistent when K grows as fast as $K = O(\sqrt{n})$, where we require a slightly stronger condition on the sparsity, $\varphi_n/(\log n)^{3+\delta} \to \infty$. Details are presented in the supplementary materials, Section A.4.

Finding the MLE involves optimizing over all possible label assignments, which is, in principle, NP-hard. General discrete optimization methods such as the tabu search (Beasley 1998; Zhao, Levina, and Zhu 2012) could be time-consuming and unstable. Taking advantage of the specific structure in our problem, we propose a pseudo-likelihood EM algorithm (PLEM) that computes an approximate solution to (4) efficiently.

The algorithm is outlined in Algorithm 1. In the outer loop, we update the label estimate and related quantities in each iteration. The inner loop employs a latent class EM algorithm to derive a new label estimate based on an initial one. For each edge A_{ij}, the pseudo-likelihood treats node i as belonging to the true community c_i and node j as belonging to an estimated community e_j. With this approximation, the latent class variables c_i's are separated in the pseudo log-likelihood function, enabling

Algorithm 1: PCABM.PLEM

Input : Adjacency matrix A; pairwise covariates Z; initial community assignment $\mathbf{e}$; number of communities K; iteration number T.

Output: Coefficient estimate $\hat{\boldsymbol{\gamma}}$; community estimate $\hat{\mathbf{c}}$.

Maximize $\ell(\boldsymbol{\gamma})$ in (2) by some optimization algorithm (e.g., BFGS) to get $\hat{\boldsymbol{\gamma}}$.

Initialization:

$$\hat{\pi}_l = \frac{n_l(\mathbf{e})}{n}, \hat{E}_{lk}(\mathbf{e}) = \sum_{(i,j)\in s_{\mathbf{e}}(l,k)} e^{\mathbf{z}_{ij}^\top \hat{\boldsymbol{\gamma}}}, \hat{B}_{lk} = \frac{O_{lk}(\mathbf{e})}{\hat{E}_{lk}(\mathbf{e})}.$$

for $t = 1$ **to** T **do**

  Calculate $b_{ik}(\mathbf{e}) = \sum_{j=1}^n A_{ij}\mathbb{1}(e_j = k)$ and $\hat{\Xi}_{ik}(\mathbf{e}) = \sum_{j=1}^n e^{\mathbf{z}_{ij}^\top \hat{\boldsymbol{\gamma}}}\mathbb{1}(e_j = k)$;

  while *pseudo-likelihood has not converged* **do**

    E-step: $\hat{\pi}_{il} = \frac{\hat{\pi}_l \prod_{k=1}^K \exp(b_{ik}\log \hat{B}_{lk} - \hat{\Xi}_{ik}\hat{B}_{lk})}{\sum_{m=1}^K \hat{\pi}_m \prod_{k=1}^K \exp(b_{ik}\log \hat{B}_{mk} - \hat{\Xi}_{ik}\hat{B}_{mk})}$;

    M-step: $\hat{\pi}_l = \frac{1}{n}\sum_{i=1}^n \hat{\pi}_{il}, \hat{B}_{lk} = \frac{\sum_{i=1}^n \hat{\pi}_{il} b_{ik}}{\sum_{i=1}^n \hat{\pi}_{il}\hat{\Xi}_{ik}}$;

  Update label estimates: $e_i = \arg\max_l \hat{\pi}_{il}$.

Output $\hat{\mathbf{c}}$ with $\hat{c}_i = e_i$.

an analytic expression for the EM updates. A similar idea was proposed in Amini et al. (2013) for SBM and DCBM. In the case of PCABM, we adjust the algorithm to account for the covariates. A detailed derivation of the PLEM algorithm under PCABM, as well as its theoretical guarantees, are provided in Section A.5 of the supplementary materials.

5. Spectral Clustering with Adjustment

Though the likelihood-based method has appealing theoretical properties, it can sometimes be computationally slow when the network size is large. In addition, the community detection results can be sensitive to the initial label assignments $\mathbf{e}$. With that in mind, we aim to propose a computationally efficient algorithm in the flavor of spectral clustering (Rohe, Chatterjee, and Yu 2011), which can also be used as the initial community label assignments for the likelihood-based methods.

5.1. A Brief Review on Spectral Clustering

First, we introduce some notations and briefly review the classical spectral clustering with K-means for SBM. Let $\mathbb{M}_{n,K}$ be the space of all $n \times K$ matrices where each row has exactly one 1 and $(K-1)$ 0's. We usually call $M \in \mathbb{M}_{n,K}$ a *membership matrix* with $M_{ic_i} = 1$ for node i with community label c_i. Note that M contains the same information as $\mathbf{c}$, and is only introduced to facilitate the discussion.

From now on, we use PCABM$(M, B, Z, \boldsymbol{\gamma}^0)$ to represent PCABM generated with parameters in parentheses. Let $G_k = G_k(M) = \{1 \le i \le n : c_i = k\}$ and $n_k = |G_k|$ for $k = 1, \ldots, K$. Let $n_{\min} = \min_{1\le k\le K} n_k$, $n_{\max} = \max_{1\le k\le K} n_k$ and $n'_{\max}$ is the second largest community size.

For convenience, we define matrix $P = [P_{ij}] \in [0,\infty)^{n\times n}$, where $P_{ij} = B_{c_i c_j}$. Then it is easy to observe $P = MBM^T$. When A is generated from a SBM with (M, B), the K-dimensional eigen-decomposition of $P = UDU^T$ and $A = \hat{U}\hat{D}\hat{U}^T$ are expected to be close, where $\hat{U}^T\hat{U} = I_K$ and $D, \hat{D} \in \mathbb{R}^{K\times K}$. Since U has only K unique rows, which represent the community labels, the K-means clustering on the rows of $\hat{U}$ usually leads to a good estimate of M. While finding a global minimizer for the K-means problem is NP-hard (Aloise et al. 2009), for any positive constant ϵ, we have efficient algorithms to find an $(1+\epsilon)$-approximate solution (Kumar, Sabharwal, and Sen 2004; Lu and Zhou 2016):

$$(\hat{M}, \hat{X}) \in \mathbb{M}_{n,K} \times \mathbb{R}^{K\times K}$$
$$s.t. \quad \|\hat{M}\hat{X} - \hat{U}\|_F^2 \le (1+\epsilon) \min_{M\in\mathbb{M}_{n,K}, X\in\mathbb{R}^{K\times K}} \|MX - \hat{U}\|_F^2.$$

The goal of community detection is to find $\hat{M}$ that is close to M. To define a loss function, we need to take permutation into account. Let $\mathcal{S}_K$ be the space of all $K \times K$ permutation matrices. Following Lei and Rinaldo (2015), we define two measures of estimation error: the overall error and the worst-case relative error:

$$L_1(\hat{M}, M) = n^{-1} \min_{S\in\mathcal{S}_K} \|\hat{M}S - M\|_0,$$
$$L_2(\hat{M}, M) = \min_{S\in\mathcal{S}_K} \max_{1\le k\le K} n_k^{-1} \|(\hat{M}S)_{G_k\cdot} - M_{G_k\cdot}\|_0.$$

It can be seen that $0 \le L_1(\hat{M}, M) \le L_2(\hat{M}, M) \le 2$. While L_1 measures the overall proportion of mis-clustered nodes, L_2 measures the worst-case performance across all communities.

Vanilla spectral clustering on SBM requires the average degree of the network to be of the order $\Omega(\log n)$ (Lei and Rinaldo 2015), mainly because sparser networks do not have desired concentration properties like $||A - \mathbb{E}A|| = O(\sqrt{\varphi_n})$. In particular, because the true $\mathbb{E}A$ has elements of the same scale, one can imagine a node with a very large degree will harm the closeness between A and $\mathbb{E}A$, which is the basis that spectral clustering lies on. Recent works (Joseph and Yu 2016; Le, Levina, and Vershynin 2017; Gao et al. 2017) have shown that regularized versions of spectral clustering (Amini et al. 2013; Qin and Rohe 2013), which basically means performing spectral clustering on a regularized adjacency matrix, could enable the concentration of the adjacency matrix under sparser settings and thus relax the average degree assumption required in vanilla spectral clustering. In our algorithms, we adopt the "reduce weight of edges proportionally to the excess of degrees" version of regularization (Le, Levina, and Vershynin 2017), that is, assigning weight $\sqrt{\lambda_i\lambda_j}$ to A_{ij}, where $\lambda_i := \min\{2d/d_i, 1\}$, $d = \max_{ij} nP_{ij}$, and d_i is the degree of node i. As d is unknown, in practice we can take $\lambda_i = \min\{\lambda^R\bar{d}/d_i, 1\}$, where $\bar{d} = \sum_{i=1}^n d_i/n$ is the average degree, and λ^R is a constant. For theoretical guarantee we need λ^R to be large enough, but in practice $\lambda^R = 2$ is sufficient to give satisfactory results from our simulation experience.

5.2. Regularized Spectral Clustering with Adjustment

The existence of covariates in PCABM prevents us from applying (regularized) spectral clustering directly on A. Unlike SBM

where A is generated from a low-rank matrix P, A in PCABM consists of both community and covariate information. Since $P_{ij} = \mathbb{E}[A_{ij}/e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}]$, an intuitive idea to take advantage of the low-rank structure is to remove the covariate effects, that is, using the adjusted adjacency matrix $[A_{ij}/e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}]$ for spectral clustering.

In practice, we don't know the true value of the parameter $\boldsymbol{\gamma}^0$. Naturally, we replace $\boldsymbol{\gamma}^0$ with the empirical estimate $\hat{\boldsymbol{\gamma}}$ from (2), and define the *adjusted adjacency matrix* as $A' = [A'_{ij}]$ where $A'_{ij} = A_{ij}\exp(-\mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}})$. Furthermore, for regularized spectral clustering, define the weighted version of A' to be A'^R, called *weighted adjusted adjacency matrix*. By the asymptotic properties of $\hat{\boldsymbol{\gamma}}$ proved in Theorem 1, we show that $\|A'^R - P\|$ achieves the desirable spectral bound of order $O_p(\sqrt{\varphi_n})$; the proof is given in Section A.6 of the supplementary materials.

Based on this bound, we could then apply the regularized spectral clustering algorithm on matrix A' to detect the communities. We call this adjustment scheme the Spectral Clustering with Adjustment (SCWA) algorithm, which is elaborated in Algorithm 2.

To show the consistency of Algorithm 2, one natural requirement is that A'^R and P are close enough, which is stated rigorously in the following theorem.

Theorem 3 (Spectral bound of adjusted, regularized Poisson random matrices). Let A be the adjacency matrix generated by the undirected PCABM $(M, B, Z, \boldsymbol{\gamma}^0)$. Assume Conditions 1, 2, 3 hold. Further assume each element of $\bar{B}$ is bounded from above by a constant $C_{\bar{B}}$ and below by a constant $c_{\bar{B}}$. For any $r > 1$, the following holds with probability at least $1 - 5n^{-r} - C_\eta \exp(-\nu_\eta n)$ (where $\eta = (p\zeta)^{-1}$, C_η and ν_η are constants in Lemma A.6.11 of the supplementary materials): the regularized adjusted adjacency matrix A'^R in Algorithm 2 satisfies

$$\|A'^R - P\| \le C\sqrt{\varphi_n} \tag{5}$$

where C is a constant that depends on $p, r, \xi, \zeta, C_{\bar{B}}$ and $c_{\bar{B}}$.

Algorithm 2: PCABM.SCWA

Input : Adjacency matrix A; pairwise covariates Z; initial community assignment $\mathbf{c}$; number of communities K; approximation parameter ϵ; constant λ^R.

Output: Coefficient estimate $\hat{\boldsymbol{\gamma}}$; community estimate $\hat{\mathbf{c}}$.

1 Maximize $\ell(\boldsymbol{\gamma})$ as in (2) by some optimization algorithm (e.g., BFGS) to derive $\hat{\boldsymbol{\gamma}}$.

2 Compute the adjusted adjacency matrix $A' = [A'_{ij}]$ where $A'_{ij} = A_{ij}\exp(-\mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}})$.

3 Compute the regularized adjusted adjacency matrix $A'^R = [A'^R_{ij}]$, where $A'^R_{ij} = A'_{ij}\sqrt{\lambda_i\lambda_j}$, $\lambda_i = \min\{\lambda^R d'/d'_i, 1\}$, d'_i is the degree of node i in A' and $d' = \sum_i d'_i/n$.

4 Calculate $\hat{U} \in \mathbb{R}^{n\times K}$ consisting of the leading K eigenvectors (ordered in absolute eigenvalue) of A'^R.

5 Calculate the $(1+\epsilon)$-approximate solution $\hat{M}$ to the K-means problem with K clusters and input matrix $\hat{U}$.

6 Output $\hat{\mathbf{c}}$ according to $\hat{M}$.

Similarly to the proof of Theorem 3.1 in Lei and Rinaldo (2015), we can prove the following Theorem 4 by combining Lemmas 5.1 and 5.3 in Lei and Rinaldo (2015), and Theorem 3.

Theorem 4. In addition to the conditions of Theorem 3, assume that $P = MBM^T$ is of rank K with the smallest absolute nonzero eigenvalue at least ϱ_n, and that $\max_{kl}\bar{B}(k,l) = 1$. Let $\hat{M}$ be the output of spectral clustering using $(1+\epsilon)$ approximate K-means on A'^R (defined in Algorithm 2, step 3). For any constant $r > 0$, there exists an absolute constant $C > 0$, such that, if

$$(2+\epsilon)\frac{Kn\rho_n}{\varrho_n^2} < C, \tag{6}$$

then, with probability at least $1 - 5n^{-r} - C_\eta\exp(-\nu_\eta n)$, there exist subsets $H_k \subset G_k$ for $k = 1, \ldots, K$, and a $K\times K$ permutation matrix J such that $\hat{M}_{G\cdot}J = M_{G\cdot}$, where $G = \cup_{k=1}^K (G_k \setminus H_k)$, and

$$\sum_{k=1}^K \frac{|H_k|}{n_k} \le C^{-1}(2+\epsilon)\frac{Kn\rho_n}{\varrho_n^2}. \tag{7}$$

Inequality (7) provides an error bound for the overall relative error. Theorem 4 doesn't provide us with an error bound in a straightforward form since ϱ_n contains ρ_n. The following corollary gives us a clearer view of the error bound in terms of model parameters. The condition that the maximum normalized probability equals 1 can be replaced by any constant, but we just use 1 here for simplicity, since any constant can always be absorbed into the sparsity parameter ρ_n.

Corollary 1. In addition to the conditions of Theorem 3, assume that $\bar{B}$'s minimum absolute eigenvalue is bounded below by $\tau > 0$ and $\max_{kl}\bar{B}(k,l) = 1$. Let $\hat{M}$ be the output of spectral clustering using $(1+\epsilon)$ approximate K-means on A'^R. For any constant $r > 0$, there exists an absolute constant C such that if

$$(2+\epsilon)\frac{Kn}{n_{\min}^2\tau^2\rho_n} < C,$$

then with probability at least $1 - 5n^{-r} - C_\eta\exp(-\nu_\eta n)$,

$$L_2(\hat{M}, M) \le C^{-1}(2+\epsilon)\frac{Kn}{n_{\min}^2\tau^2\rho_n},$$
$$L_1(\hat{M}, M) \le C^{-1}(2+\epsilon)\frac{Kn'_{\max}}{n_{\min}^2\tau^2\rho_n}.$$

It is worth mentioning that Theorems 3, 4, and Corollary 1 all allow K to go to infinity with n.

Compared to SCWA, the pseudo-likelihood EM algorithm can yield more accurate results, especially when provided with good initial labels. On the other hand, the SCWA algorithm is computationally more efficient. To combine the advantages of these two methods, we propose using the results of SCWA as the initial estimate for the pseudo-likelihood EM (PCABM.PL as described in Algorithm 3). We will conduct extensive simulation studies in Section 7 to evaluate the performance of both PCABM.SCWA and PCABM.PL.

Algorithm 3: PCABM.PL

Input : Adjacency matrix A; pairwise covariates Z; initial community assignment $\mathbf{e}$; number of communities K; approximation parameter ϵ; iteration number T; constant λ^R.

Output: Community estimate $\hat{\mathbf{c}}$.

1 Use Algorithm 2 to get an initial community estimate $\hat{\mathbf{e}}$.

2 Use Algorithm 1 with initial community estimate $\hat{\mathbf{e}}$ to get the community estimate $\hat{\mathbf{c}}$.

6. Model Selection

So far, we have been treating the number of communities K as given. In practice, the true value of K may be unknown to us. In that case, we would be interested in estimating K. Some existing work for estimating K include Wang and Bickel (2017), Saldana, Yu, and Feng (2017), Chen and Lei (2018), and Li, Levina, and Zhu (2020). To provide a systematic approach, we propose adapting the edge-sampling cross-validation (ECV) method (Li, Levina, and Zhu 2020) to the PCABM. The main idea of the ECV procedure can be summarized as follows: in each iteration, we randomly sample a certain proportion of node pairs in the network, and predict the remaining node pairs under specific models based on matrix completion on the adjacency matrix containing the true edge information of the selected node pairs. After all iterations, we compare the average prediction performance or hold-out losses under different models and choose the best model accordingly. Algorithm 4 presents a detailed process of applying this idea to estimate K in the PCABM. The notation $P_\Omega A$ represents the matrix that retains all elements of A in the index set Ω while setting other elements to 0.

In step 5 of Algorithm 4, $\hat{A}'_K$ denotes the rank-K matrix completion from $P_\Omega A'$. As suggested in Li, Levina, and Zhu (2020), we use the SVD truncation approach to obtain $\hat{A}'_K$. In the SVD of $P_\Omega A' = UDV^\top$, we keep the K largest elements of diagonal D and set $\hat{A}'_K = \frac{1}{p} U D_K V^\top$. This simple matrix completion method efficiently serves our model selection goal while remaining computationally inexpensive.

For the loss evaluated in step 7 of Algorithm 4, there are two options: the scaled negative log-likelihood (snll) $\sum_{(i,j)\in\Omega^c} \left[\hat{B}_{\hat{e}_i\hat{e}_j} - A_{ij}\exp(-\mathbf{z}_{ij}^\top\hat{\boldsymbol{\gamma}})\log \hat{B}_{\hat{e}_i\hat{e}_j}\right]$ and the scaled L_2 loss $\sum_{(i,j)\in\Omega^c} \left[\hat{B}_{\hat{e}_i\hat{e}_j} - A_{ij}\exp(-\mathbf{z}_{ij}^\top\hat{\boldsymbol{\gamma}})\right]^2$. We scale the loss functions by the covariate effect since the cross-validation is based on the block structure.

We present a theorem establishing the consistency of selecting K using Algorithm 4.

Theorem 5 (Consistency of Algorithm 4 under PCABM). Let A be the adjacency matrix generated by the undirected PCABM $(M, B, Z, \boldsymbol{\gamma}^0)$. Assume Conditions 1, 2 hold, and each element of $\bar{B}$ is bounded above by a constant $C_{\bar{B}}$, that is, $\|\bar{B}\|_{\max} \le C_{\bar{B}}$. The training proportion $p \in (0,1)$ is a constant. The number of communities K is fixed and to be estimated. Further assume $\varphi_n/\log n \to \infty$. Let $\hat{K}$ be the selected number of communities by using Algorithm 4 with the scaled L_2 loss. Then we have $\Pr(\hat{K} < K) \to 0$.

Algorithm 4: ECV for selecting K in PCABM

Input : Adjacency matrix A, covariates Z, the maximum number of communities to consider $K_{\max}$, training proportion p, number of replications (folds) N_{rep}.

Output: Estimated number of communities $\hat{K}$.

1 Calculate MLE $\hat{\boldsymbol{\gamma}}$ with A, Z with $\mathbf{e}$ being all 1 vector.

2 **for** $m = 1$ **to** N_{rep} **do**

3 Randomly choose a subset of node pairs Ω: selecting each pair $(i,j), i<j$ independently with probability p, and adding (j,i) if (i,j) is selected.

4 **for** $K = 1$ **to** $K_{\max}$ **do**

5 Apply matrix completion to $P_\Omega A'$ with rank constraint K to obtain $\hat{A}'_K$, where A' denotes the adjusted adjacency matrix $A'_{ij} = A_{ij}/\exp(\mathbf{z}_{ij}^\top\hat{\boldsymbol{\gamma}})$.

6 Run spectral clustering on $\hat{A}'_K$ to obtain the estimated membership vector $\hat{\mathbf{e}}_K^{(m)}$.

7 Estimate the probability matrix $\hat{B}_K^{(m)}$ with $\hat{B}_{kl}(\hat{\mathbf{e}},\hat{\boldsymbol{\gamma}}) = O_{kl}^{(\Omega)}(\hat{\mathbf{e}})/E_{kl}^{(\Omega)}(\hat{\mathbf{e}},\hat{\boldsymbol{\gamma}})$, and evaluate the corresponding losses $L_K^{(m)}$, by applying the loss function L with the estimated parameters to $A_{ij}, (i,j)\in\Omega^c$.

8 Let $L_K = \sum_{m=1}^{N_{rep}} L_K^{(m)}/N_{rep}$. Return $\hat{K} = \arg\min_{\{K=1,\ldots,K_{\max}\}} L_K$.

If we assume $\varphi_n/\sqrt{n} \to \infty$ and additionally assume all entries of $\bar{B}$ are bounded below by a constant $c_{\bar{B}}$, then the same result also holds for the scaled negative log-likelihood loss.

In addition to choosing the number of communities, another model selection problem of interest is distinguishing between covariate-adjusted and covariate-confounding models. In the covariate-adjusted model, covariates and class labels are independent, while in the covariate-confounding model, the distribution of covariates is governed by the community labels.

Given prior knowledge that a covariate is correlated with the block effect, one can extract cluster information from both the covariate and the network to improve the estimation accuracy of community labels. However, without that prior knowledge, fitting a confounding covariate in a covariate-adjusted model can undermine clustering performance. This phenomenon is illustrated in a simulation example provided in Section B of the supplementary material. A heuristic explanation is that the incorrect model mistakenly identifies the true underlying block effect as the covariate effect of the confounding covariate.

Motivated by the model-selection nature of the problem, we propose a covariate selection procedure based on the ECV framework. We present the detailed procedure in Algorithm 5 in Section B of the supplementary material and demonstrate that the proposed algorithm almost perfectly screens out false covariates and selects the correct model under various simulation study settings.

We note that while the proposed feature selection algorithm represents an interesting initial attempt to address confounding

covariates, it is still based on the PCABM, which only models covariate adjusting. It would be desirable to propose a comprehensive covariate block model and corresponding community detection methods that could integrate both covariate-adjusted and covariate-confounding models.

7. Simulations

For all simulations, we consider K communities with prior probabilities $\pi_i = 1/K, i = 1, \ldots, K$. In addition, we fix $\bar{B}$ to have all diagonal elements equaling 2 and off-diagonal elements 1; and we fix $K = 2$ except in Section 7.5 where K varies. We generate data by applying the following procedure:

S1. Determine parameters ρ_n and $\boldsymbol{\gamma}^0$. Generate $\mathbf{z}_{ij}$ from certain distributions.

S2. Generate adjacency matrix $A = [A_{ij}]$ using PCABM with parameters in S1.

7.1. γ Estimation

For PCABM, estimating $\boldsymbol{\gamma}$ would be the first step, so we check the consistency and asymptotic normality of $\hat{\boldsymbol{\gamma}}$ claimed in our theory section.

The pairwise covariate vector $\mathbf{z}_{ij}$ has five variables, generated independently from Bernoulli(0.1), Poisson(0.1), Uniform[0, 1], Exponential(0.3), and $N(0, 0.3)$, respectively. The parameters for each distribution are chosen to make the variances of covariates similar.

We ran 100 simulations respectively for $n = 100, 300, 500$. The parameters are set as $\rho_n = 5 \log n/n$, $\boldsymbol{\gamma}^0 = (0.4, 0.8, 1.2, 1.6, 2)^T$. We obtained $\hat{\boldsymbol{\gamma}}$ by using BFGS to optimize the likelihood function under the initial community assignment $\mathbf{e}_0 = \mathbf{1}_n$. We present the mean and standard deviation of $\hat{\boldsymbol{\gamma}}$ in Table 1. It is clear that $\hat{\boldsymbol{\gamma}}$ is very close to $\boldsymbol{\gamma}^0$ even for a small network. The shrinkage of standard deviation implies the consistency of $\hat{\boldsymbol{\gamma}}$. We also repeated the experiment by initializing with random community assignments, which leads to very similar results (Table C.3 of supplementary materials). This validates the observation that estimating $\boldsymbol{\gamma}$ and communities is decoupled.

By taking a closer look at the network of size $n = 500$, we compare the distribution of $\hat{\boldsymbol{\gamma}}$ with the theoretical asymptotic normal distribution derived in Theorem 1. We show the histogram for the first three coefficients in Figure 2. We can see that the empirical distribution matches well with the theoretical counterpart.

7.2. Community Detection

After obtaining $\hat{\boldsymbol{\gamma}}$, we now move on to the estimation of community labels. There are three parameters that we could tune to change the property of the network: $\boldsymbol{\gamma}^0$, ρ_n, and n. To illustrate the impact of these parameters on the performance of community detection, we vary one parameter while fixing the remaining two in each experiment. More specifically, we consider the form $\rho_n = c_\rho \log n/n$ and $\boldsymbol{\gamma}^0 = c_\gamma (0.4, 0.8, 1.2, 1.6, 2)$ in which we will vary the multipliers c_ρ and c_γ. The detailed parameter settings for the three experiments are as follows.

(a) $n \in \{200, 400, 600, 800, 1000\}$, with $c_\rho = 5$ and $c_\gamma = 1.2$.

(b) $c_\rho \in \{2, 3, 4, 5, 6\}$, with $n = 200$ and $c_\gamma = 1.2$.

(c) $c_\gamma \in \{0, 0.4, 0.8, 1.2, 1.6, 2.0\}$, with $n = 200$ and $c_\rho = 5$.

The results for the three experiments are presented in panels (a), (b), and (c) in Figure 3. Each setting is simulated 100 times. The error rate is reported in terms of the average Adjusted Rand Index (ARI) (Hubert and Arabie 1985), which is a measure of the similarity between two data clusterings. SBM.MLE and SBM.SC refer to the likelihood and spectral clustering methods under SBM, respectively; DCBM.MLE is the maximum likelihood method based on DCBM (Zhao, Levina, and Zhu 2012); PCABM.PL and PCABM.SCWA refer to Algorithms 3 and 2, respectively.

As the number of nodes increases, it is evident from the first panel (a) in Figure 3 that both PCABM-based algorithms perform exceptionally well, with PCABM.PL achieving nearly perfect community detection performance across all values of n. Spectral clustering under SBM results in nearly random guesses. DCBM and MLE under SBM perform better when n is large but still underperform PCABM-based algorithms. As the density of the network increases, the performance does not change significantly within this range. When the scale of $\boldsymbol{\gamma}^0$ is changed,

Table 1. Simulated results of $\hat{\boldsymbol{\gamma}}$ over 100 repetitions, displayed as mean (standard deviation).

n	$\gamma_1^0 = 0.4$	$\gamma_2^0 = 0.8$	$\gamma_3^0 = 1.2$	$\gamma_4^0 = 1.6$	$\gamma_5^0 = 2$
100	0.393(0.0471)	0.796(0.0345)	1.206(0.0560)	1.596(0.0410)	2.005(0.0454)
300	0.399(0.0198)	0.801(0.0160)	1.198(0.0256)	1.603(0.0180)	2.003(0.0213)
500	0.399(0.0147)	0.800(0.0117)	1.197(0.0162)	1.599(0.0148)	2.002(0.0155)

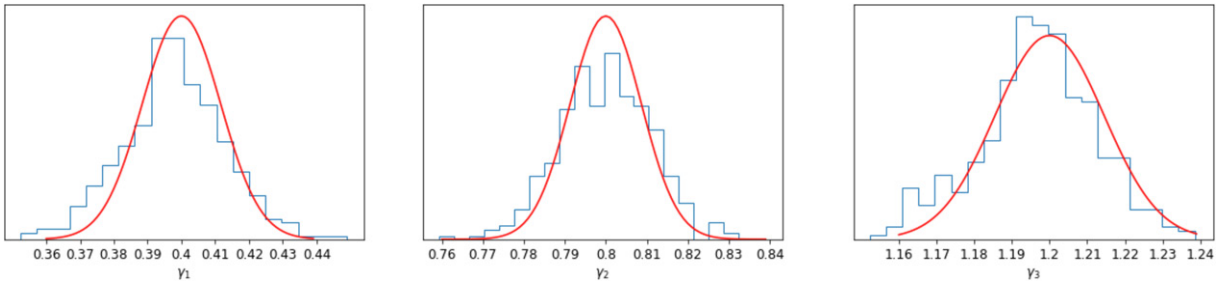
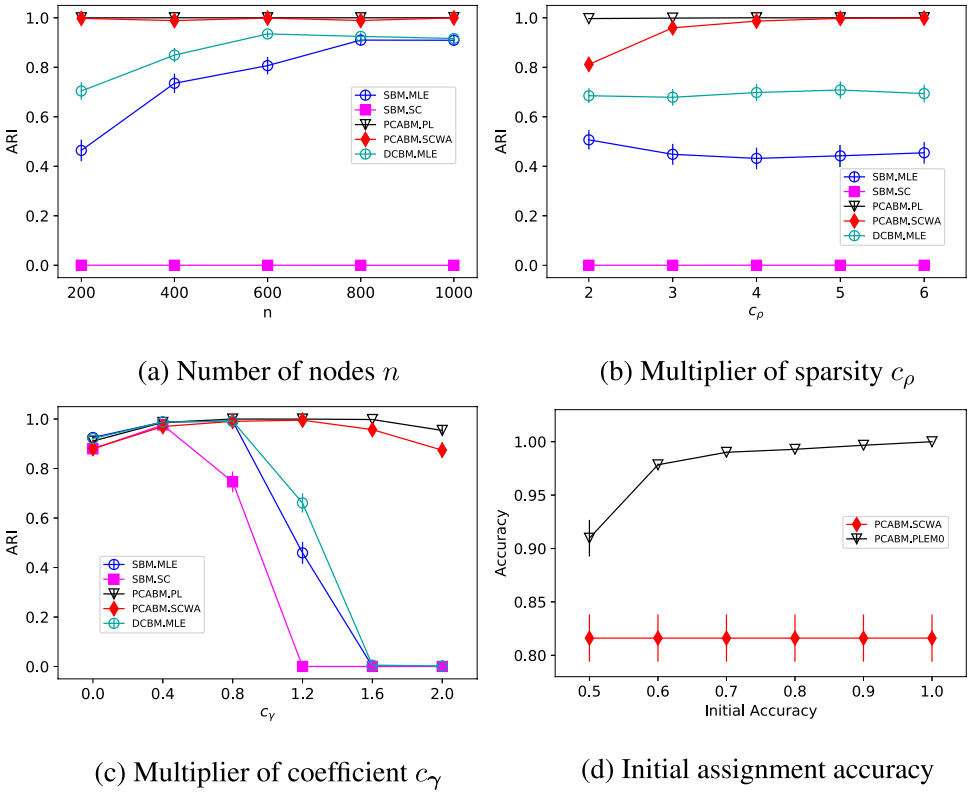


Figure 2. Simulation results for $\hat{\boldsymbol{\gamma}}$ compared with theoretical values.



(a) Number of nodes n

(b) Multiplier of sparsity c_ρ

(c) Multiplier of coefficient c_γ

(d) Initial assignment accuracy

Figure 3. Simulation results under PCABM for different parameter settings

both PCABM algorithms continue to yield good results. As we know, when $\boldsymbol{\gamma}^0 = \mathbf{0}$, our model reduces to SBM, so it is not surprising that SBM.MLE and SBM.SC both perform well when the magnitude of $\boldsymbol{\gamma}^0$ is relatively small and fail when the magnitude increases.

7.3. Impact of Initial Assignments Accuracy

The performance of the pseudo-likelihood EM (Algorithm 1) depends on the initial assignments. To further understand its influence on our algorithm, we simulate initial community assignments with different accuracy rates and examine how they affect prediction accuracy. The parameters are fixed to be $n = 200, c_\rho = 2, c_\gamma = 1.5$. We change the accuracy of initial assignments from 0.5 to 1. To make the results easier to interpret, we use accuracy rather than ARI to evaluate performance. Note that SCWA does not use class assignment initialization, and we plot its accuracy as a reference flat line in panel (d) of Figure 3. On one hand, even with completely random initial assignments, the PLEM algorithm yields satisfactory clustering accuracy. On the other hand, as the accuracy of initial assignments increases, the prediction accuracy of the PLEM method also improves. If we use the prediction of SCWA, with an accuracy of around 0.82, as the initial assignments for the PLEM method, we can enhance the prediction accuracy from around 0.9 (random initial) to almost 1. Therefore, it is preferable to use the output of SCWA as initial assignments for the PLEM method.

7.4. DCBM

Considering that PCABM includes DCBM as a special case in terms of having the same profile likelihood, we are curious about the performance of Algorithms 2 and 3 on networks generated by DCBM. The degree parameter for each node is chosen from $\{1, 4\}$ with equal probability, $\bar{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, and $\rho_n = c_\rho \log n/n$. For covariates, we take $z_{ij} = \log d_i + \log d_j$, where d_i is the degree of node i. As a comparison, we also implemented the likelihood method in Zhao, Levina, and Zhu (2012) (DCBM.MLE) and the SCORE method in Jin (2015). As in Section 7.2, we vary one parameter while fixing the remaining one in each experiment. The detailed parameter settings for the two experiments are as follows, with results presented in Figure 4.

(a) $n \in \{200, 400, 600, 800, 1000\}$, with $c_\rho = 3$.

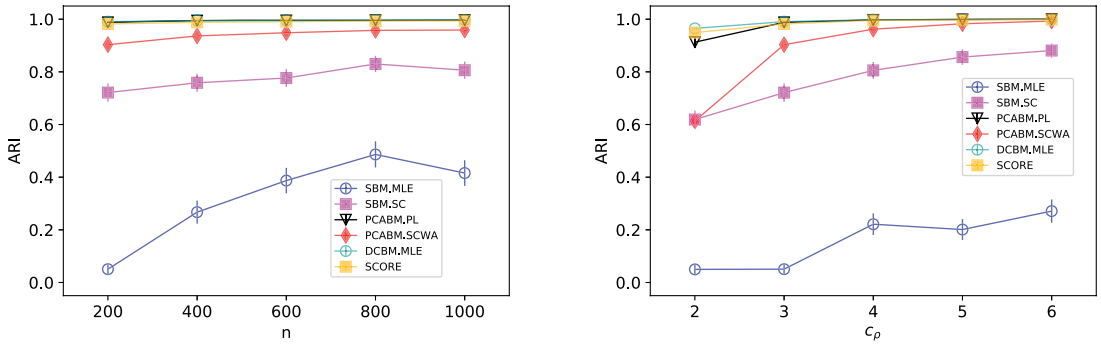
(b) $c_\rho \in \{2, 3, 4, 5, 6\}$, with $n = 200$.

From the results, we observe that, except for SBM.MLE and SBM.SC, all the other methods work well, with the ARI being almost 1 when n or c_ρ is large. The flexibility of PCABM allows us to model any factors that may contribute to the network structure in addition to the underlying communities.

7.5. Estimation of the Number of Communities K

In this section, we study the performance of our approach for selecting the number of communities K, that is, Algorithm 4. We set $\bar{B}$ to have diagonal elements of 2 and off-diagonal elements of 1. Additionally, we set $n = 1000$ and $\rho_n = 5 \log n/n$. Coefficient $\boldsymbol{\gamma}^0$ and covariates Z are generated in the same way as in Section 7.1. The ECV parameters are set as $p = 0.9$ and $N_{rep} = 5$. We consider cases where the true underlying K is 2, 3, or 4, and let $K_{\max} = 6$, that is, selecting $\hat{K}$ from $\{1, 2, \ldots, 6\}$. The simulation results are presented in Table 2.

The results show that Algorithm 4 selects the correct K with a high probability.



(a) Different n with $c_\rho = 3$ (b) Different c_ρ with $n = 200$

Figure 4. Simulation results under DCBM for different parameter settings

Table 2. Accuracy of ECV estimation of community number K in 100 realizations under scaled negative log-likelihood (snll) and scaled L_2 loss.

Loss	snll loss		scaled L_2 loss	
$\hat{K}$	$\Pr(\hat{K} = K)$	$\Pr(\hat{K} \geq K)$	$\Pr(\hat{K} = K)$	$\Pr(\hat{K} \geq K)$
$K = 2$	97%	100%	97%	100%
$K = 3$	100%	100%	100%	100%
$K = 4$	92%	100%	93%	93%

Table 3. Performance comparison on political blogs data.

	DCBM.MLE	DCBM.RSC	DCBM.CMM	SCORE	PCABM.PL
ARI	0.819	–	–	0.819	0.813
NMI	0.72	–	–	0.725	0.725
Errors	–	–	62	58	60
Accuracy	–	95%	94.9%	95.3%	95.1%

NOTE: The performance of DCBM.MLE is taken from Karrer and Newman (2011) and Zhao, Levina, and Zhu (2012); the performance of SCORE is from Jin (2015); the performance of regularized spectral clustering (RSC) based on DCBM is reported in Joseph and Yu (2016); the performance of convexified modularity maximization (CMM) for DCBM is from Chen, Li, and Xu (2018).

8. Real Data Examples

8.1. Example 1: Political Blogs

The first real-world dataset we used is the network of political blogs created by Adamic and Glance (2005). The nodes represent blogs about U.S. politics, and the edges indicate hyperlinks between them. We treated the network as undirected and focused only on the largest connected component of the network, resulting in a subnetwork with 1222 nodes and 16,714 edges.

Since there are no other nodal covariates available in this dataset, we created one pairwise covariate by aggregating degree information. We set $z_{ij} = \log(d_i \times d_j)$, where d_i is the degree for the ith node. The coefficient estimate for the covariate $\hat{\boldsymbol{\gamma}}$ is 1.0005 with a 95% confidence interval of $(0.9898, 1.0111)$. Table 3 summarizes the performance comparison of PCABM with some existing results on this dataset. In addition to ARI, we also evaluated normalized mutual information (NMI) (Danon et al. 2005), which is a measure of mutual dependence.

We observed that the performance of our model is on par with previous methods designed specifically for DCBM, and the error rate is very close to the ideal results mentioned in Jin (2015), which is 55/1222. This demonstrates that PCABM provides an alternative approach to DCBM by incorporating degree information into a specific pairwise covariate. As a more flexible model, PCABM also suggests that DCBM is indeed a suitable model for this dataset since the coefficient estimate is close to 1. This is consistent with our argument that PCABM includes DCBM from a profile likelihood perspective. Lastly, PCABM offers a significant improvement over the vanilla SBM, whose NMI is only 0.0001, as reported in Karrer and Newman (2011).

8.2. Example 2: School Friendship

In real networks, people often use specific nodal covariates as the ground "truth" for community labels to evaluate the performance of various community detection methods. However, there could be different "true" community assignments based on different nodal covariates (e.g., gender, job, and age). Peel, Larremore, and Clauset (2017) mentioned that communities and covariates might capture various aspects of the network, which is in line with the idea presented in this article. To examine whether PCABM can discover different community structures, in our second example, we treat one covariate as the indicator for the unknown "true" community assignments while using the remaining covariates to construct the pairwise covariates in our PCABM model.

The dataset is a friendship network of school students from the National Longitudinal Study of Adolescent to Adult Health (Add Health). It contains 795 students from a high school (Grades 9–12) and its feeder middle school (Grades 7–8). The nodal covariates include grade, gender, ethnicity, and the number of friends nominated (up to 10). We focused on the largest connected component with at least one non-missing covariate and treated the network as undirected, resulting in a network with 777 nodes and 4124 edges. For the nodes without gender, we assigned them to the female group, which is the smaller group. For those without grades, we generated a random grade within their schools.

Unlike traditional community detection methods that can only detect one underlying community structure, PCABM provides us with more flexibility to uncover different community structures by controlling different covariates. Our intuition is that social network is usually determined by multiple underlying structures and cannot be simply explained by one covariate.

Sometimes one community structure seems to dominate the network, but if we adjust the covariate associated with that structure, we may discover other interesting community structures.

In this example, we conducted two community detection experiments. In each experiment, out of the two nodal covariates, school and ethnicity, one was viewed as the proxy for the "true" underlying community, and community detection was carried out using the pairwise covariates constructed with other covariates. For school and ethnicity, we created indicator variables to represent whether the corresponding covariate values were the same for the pair of nodes. For example, if two students come from the same school, the corresponding pairwise covariate equals 1; if they have different genders, the corresponding pairwise covariate equals 0. We also considered the number of nominated friends in all experiments and grades for predicting ethnicity and gender. For the number of nominated friends, we used $\log(n_i + 1) + \log(n_j + 1)$ as one pairwise covariate, where n_i is the number of nominated friends by the ith student. We added "+1" because some students did not nominate anyone. For grades, we used the absolute difference to form a pairwise covariate. Using random initial community labels, we computed the estimates $\hat{\boldsymbol{\gamma}}$ in each experiment. In Tables 4 and 5, we show respectively the estimates when school and ethnicity are taken as the targeted community.

In both tables, the standard error is calculated using Theorem 1, with the theoretical values replaced by the estimated counterparts. Thus, we can calculate the t value for each coefficient and perform the corresponding statistical tests. We can see that in both experiments, the coefficients for gender and the number of nominations are positive and significant in the creation of the friendship network. The significant positive coefficient of nominations shows that students with a large number of nominations have a higher chance to be friends with each other, which is intuitive. The positive coefficients of gender and school indicate that students of the same gender and school are more likely to be friends with each other, which aligns with our expectations. The negative coefficient of grade means that students with closer grades are more likely to be friends. If we examine the coefficients of different ethnic groups in Table 4, we find that only those corresponding to white and black are significant. This is understandable, as we observe that among 777 students, 476 are white, and 221 are black. As for school and grade, students in the same school or grade tend to be friends with each other, as expected.

Table 4. Inference results when school is targeted community.

Covariate	Estimate	t value	Pr(> |t|)
White	1.251	29.002	< 0.001***
Black	1.999	38.886	< 0.001***
Hispanic	0.048	0.091	0.927
Others	0.019	0.035	0.972
Gender	0.192	5.620	< 0.001***
Nomination	0.438	18.584	< 0.001***

Table 5. Inference results when ethnicity is targeted community.

Covariate	Estimate	t value	Pr(> |t|)
School	1.005	13.168	< 0.001***
Grade	−1.100	−39.182	< 0.001***
Gender	0.198	5.813	< 0.001***
Nomination	0.498	21.679	< 0.001***

Table 6. ARI comparison on school friendship data.

	PCABM.PL	SBM.MLE	SBM.SC	DCBM.MLE	SCORE
School	0.924	0.048	0.043	0.909	0.799
Ethnicity	0.909	0.138	−0.024	0.001	0.012

The network is divided into two communities each time (we only look at white and black students in the second experiment because the sizes of other ethnicities are very small). We apply our algorithm PCABM.PL, as well as some classic methods on SBM and DCBM, to cluster the network in both experiments. The results in terms of ARI are shown in Table 6. It can be seen that while DCBM can capture one main structure of the network, "School," which is likely the dominating structure, our method can not only capture "School" but also capture "Ethnicity" when adjusting for the covariate "School". Note that for all methods other than ours, we would obtain only one community structure, whose performance is bound to be suboptimal for capturing different community structures. Compared to the SBM or DCBM, the proposed PCABM can adjust some known structures in the network to discover additional latent structures.

9. Discussion

In this article, we extend the classical stochastic block model to allow the connection rate between nodes to depend on not only the community memberships but also the pairwise covariates. We prove consistency in terms of both coefficient estimates and community label assignments for MLE under PCABM, and provide an efficient algorithm to solve an approximate MLE. Additionally, we introduce a fast spectral method, SCWA, with theoretical justification, which could serve as a good initial solution for the likelihood-based method. Furthermore, we propose cross-validation-based algorithms for estimating the number of communities and feature selection.

There are many interesting future research directions on PCABM. In our article, we assume the entries in the adjacency matrix are nonnegative integers. However, this can be relaxed to be any nonnegative numbers, and we expect similar theoretical results to hold. It would also be interesting to consider highly imbalanced community sizes, where $n_{\min}/n_{\max} = o(1)$. Moreover, when we have high-dimensional pairwise covariates, adding a penalty term to conduct variable selection is worth investigating. For instance, in the estimation of $\boldsymbol{\gamma}$, we can regularize (2) with an L_1 penalty $\hat{\boldsymbol{\gamma}}_\lambda(\mathbf{e}_0) = \arg\max_{\boldsymbol{\gamma}} \left\{\ell_{\mathbf{e}_0}(\boldsymbol{\gamma}) - \lambda\|\boldsymbol{\gamma}\|_1\right\}$ to estimate a sparse high-dimensional $\boldsymbol{\gamma}$.

One model assumption in PCABM is the independence among edges conditional on observed covariates. However, the assumption might be inappropriate if there are unobserved covariates. To address this, one possible extension is a degree-corrected PCABM, which can incorporate unobserved nodal covariates. The adjacency matrix could be modeled as, for example, $A_{ij}|\mathbf{c}, Z, \boldsymbol{\theta} \sim \text{Poisson}(B_{c_i c_j}\theta_i\theta_j \exp(\mathbf{z}_{ij}^\top \boldsymbol{\gamma}))$, where $\boldsymbol{\theta}$ represents degree correction parameters. From a modeling

perspective, the $\boldsymbol{\theta}$ term could be one way of incorporating unobserved nodal covariates or random effects. From a model fitting point of view, the first question to ask about this model is whether it is, in some sense, equivalent to PCABM by adding the covariate $\log(d_i d_j)$, where d_i is the degree of the ith node, or the degree after scaling by the covariate effect.

The Python and Matlab codes for implementing the proposed algorithms are available on GitHub at *https://github.com/sihanhuang/pcabm_code*.

Supplementary Materials

The supplementary material contains the proofs of the theoretical results, presents some technical lemmas, and shows additional simulation results.

Acknowledgments

We thank the editor, the AE, and anonymous reviewers for their insightful comments which have greatly improved the scope and quality of the article.

Funding

This work was supported by NIH grant 1R21AG074205-01, NYU University Research Challenge Fund, and a grant from NYU School of Global Public Health.

ORCID

Yang Feng http://orcid.org/0000-0001-7746-7598

Disclosure Statement

The authors report there are no competing interests to declare.

References

Adamic, L. A., and Glance, N. (2005), "The Political Blogosphere and the 2004 US Election: Divided They Blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, ACM. [2101]

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009), "Np-hardness of Euclidean Sum-of-Squares Clustering," *Machine Learning*, 75, 245–248. [2096]

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E., et al. (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks," *The Annals of Statistics*, 41, 2097–2122. [2096]

Beasley, J. E. (1998), *Heuristic Algorithms for the Unconstrained Binary Quadratic Programming Problem* (Vol. 4), London, UK: Management School, Imperial College. [2095]

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic Normality of Maximum Likelihood and its Variational Approximation for Stochastic Blockmodels," *The Annals of Statistics*, 41, 1922–1943. [2092]

Bickel, P. J., and Chen, A. (2009), "A Nonparametric View of Network Models and Newman–Girvan and Other Modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073. [2092]

Chen, Y., Li, X., and Xu, J. (2018), "Convexified Modularity Maximization for Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 46, 1573–1602. [2101]

Chen, K., and Lei, J. (2018), "Network Cross-Validation for Determining the Number of Communities in Network Data," *Journal of the American Statistical Association*, 113, 241–251. [2098]

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012), "Stochastic Blockmodels with a Growing Number of Classes," *Biometrika*, 99, 273–284. [2092]

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005), "Comparing Community Structure Identification," *Journal of Statistical Mechanics: Theory and Experiment*, 2005, P09008. [2101]

Deshpande, Y., Sen, S., Montanari, A., and Mossel, E. (2018), "Contextual Stochastic Block Models," in *Advances in Neural Information Processing Systems* (Vol. 31). [2093]

Fortunato, S. (2010), "Community Detection in Graphs," *Physics Reports*, 486, 75–174. [2092]

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017), "Achieving Optimal Misclassification Proportion in Stochastic Block Models," *The Journal of Machine Learning Research*, 18, 1980–2024. [2092,2096]

Getoor, L., and Diehl, C. P. (2005), "Link Mining: A Survey," *ACM SIGKDD Explorations Newsletter*, 7, 3–12. [2092]

Ghoshdastidar, D., and Dukkipati, A. (2014), "Consistency of Spectral Partitioning of Uniform Hypergraphs under Planted Partition Model," in *Advances in Neural Information Processing Systems* (Vol. 27). [2092]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends® in Machine Learning*, 2, 129–233. [2092]

Graham, B. S. (2014), "An Econometric Model of Link Formation with Degree Heterogeneity," Technical Report, National Bureau of Economic Research. [2092]

Greenland, S., Robins, J. M., and Pearl, J. (1999), "Confounding and Collapsibility in Causal Inference," *Statistical Science*, 14, 29–46. [2093]

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-based Clustering for Social Networks," *Journal of the Royal Statistical Society*, Series A, 170, 301–354. [2092]

Hoff, P. (2008), "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data," in *Advances in Neural Information Processing Systems*. [2092]

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [2092]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [2092]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [2099]

Jin, J. (2015), "Fast Community Detection by Score," *The Annals of Statistics*, 43, 57–89. [2100,2101]

Joseph, A., and Yu, B. (2016), "Impact of Regularization on Spectral Clustering," *The Annals of Statistics*, 44, 1765–1791. [2096,2101]

Karrer, B., and Newman, M. E. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [2092,2101]

Kumar, A., Sabharwal, Y., and Sen, S. (2004), "A Simple Linear Time $(1+\epsilon)$-approximation Algorithm for k-means Clustering in Any Dimensions," in *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*. IEEE. [2096]

Le, C. M., Levina, E., and Vershynin, R. (2017), "Concentration and Regularization of Random Graphs," *Random Structures and Algorithms*, 51, 538–561. [2096]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [2092,2096,2097]

Li, T., Levina, E., and Zhu, J. (2020), "Network Cross-Validation by Edge Sampling," *Biometrika*, 107, 257–276. [2093,2098]

Lu, Y., and Zhou, H. H. (2016), "Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants," arXiv preprint arXiv:1612.02099. [2096]

Ma, Z., Ma, Z., and Yuan, H. (2020), "Universal Latent Space Model Fitting for Large Networks with Edge Covariates," *J. Mach. Learn. Res.* 21, 4–1. [2093]

Newman, A. (1963), "Introduction," *Camden Third Series*, 94, vii–xiv. [2092]

Newman, M. E. (2006), "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences*, 103, 8577–8582. [2092]

Newman, M. E., and Clauset, A. (2016), "Structure and Inference in Annotated Networks," *Nature Communications*, 7, 11863. [2093]

Newman, M. E., and Leicht, E. A. (2007), "Mixture Models and Exploratory Analysis in Networks," *Proceedings of the National Academy of Sciences*, 104, 9564–9569. [2092]

Peel, L., Larremore, D. B., and Clauset, A. (2017), "The Ground Truth about Metadata and Community Detection in Networks," *Science Advances*, 3, e1602548. [2101]

Qin, T., and Rohe, K. (2013), "Regularized Spectral Clustering Under the Degree-Corrected Stochastic Blockmodel," in *Advances in Neural Information Processing Systems* (Vol. 26). [2096]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [2092,2096]

Saldana, D., Yu, Y., and Feng, Y. (2017), "How Many Communities Are There?" *Journal of Computational and Graphical Statistics*, 26, 171–181. DOI:10.1080/10618600.2015.1096790 [2098]

Tan, L. S., Chan, A. H., and Zheng, T. (2016), "Topic-Adjusted Visibility Metric for Scientific Articles," *The Annals of Applied Statistics*, 10, 1–31. [2093]

Wang, Y. X. R., and Bickel, P. J. (2017), "Likelihood-based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. DOI:10.1214/16-AOS1457 [2098]

Weng, H., and Feng, Y. (2022), "Community Detection with Nodal Information: Likelihood and its Variational Approximation," *Stat*, 11, e428. [2093]

Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2017), "Community Extraction in Multilayer Networks with Heterogeneous Community Structure," *Journal of Machine Learning Research*, 18, 1–49. [2092]

Wilson, J. D., Wang, S., Mucha, P. J., Bhamidi, S., and Nobel, A. B. (2014), "A Testing based Extraction Algorithm for Identifying Significant Communities in Networks," *The Annals of Applied Statistics*, 8, 1853–1891. [2092]

Wu, Y.-J., Levina, E., and Zhu, J. (2017), "Generalized Linear Models with Low Rank Effects for Network Data," arXiv preprint arXiv:1705.06772. [2093]

Xu, S., Zhen, Y., and Wang, J. (2022), "Covariate-Assisted Community Detection in Multi-Layer Networks," *Journal of Business and Economic Statistics*, 41, 915–926. [2093]

Yan, B., and Sarkar, P. (2021), "Covariate Regularized Community Detection in Sparse Graphs," *Journal of the American Statistical Association*, 116, 734–745. [2093]

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019), "Statistical Inference in a Directed Network Model with Covariates," *Journal of the American Statistical Association*, 114, 857–868. [2093,2095]

Yuan, M., Liu, R., Feng, Y., and Shang, Z. (2022), "Testing Community Structure for Hypergraphs," *The Annals of Statistics*, 50, 147–169. [2092]

Zhang, Y., Levina, E., and Zhu, J. (2016), "Community Detection in Networks with Node Features," *Electronic Journal of Statistics*, 10, 3153–3178. [2093]

Zhao, Y., Levina, E., and Zhu, J. (2011), "Community Extraction for Social Networks," *Proceedings of the National Academy of Sciences*, 108, 7321–7326. [2092]

——— (2012), "Consistency of Community Detection in Networks under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [2094,2095,2099,2100,2101]