Journal of the American Statistical Association

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

Fast Network Community Detection With Profile-Pseudo Likelihood Methods

Jiangzhou Wang, Jingfei Zhang, Binghui Liu, Ji Zhu & Jianhua Guo

To cite this article: Jiangzhou Wang, Jingfei Zhang, Binghui Liu, Ji Zhu & Jianhua Guo (2023) Fast Network Community Detection With Profile-Pseudo Likelihood Methods, Journal of the American Statistical Association, 118:542, 1359-1372, DOI: 10.1080/01621459.2021.1996378

To link to this article: https://doi.org/10.1080/01621459.2021.1996378

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 542, 1359–1372: Theory and Methods
https://doi.org/10.1080/01621459.2021.1996378

Fast Network Community Detection With Profile-Pseudo Likelihood Methods

Jiangzhou Wang[a,b], Jingfei Zhang[c], Binghui Liu[a], Ji Zhu[d], and Jianhua Guo[a]

[a]School of Mathematics and Statistics & KLAS, Northeast Normal University, Jilin, China; [b]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China; [c]Department of Management Science, University of Miami, Coral Gables, FL; [d]Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

The stochastic block model is one of the most studied network models for community detection, and fitting its likelihood function on large-scale networks is known to be challenging. One prominent work that overcomes this computational challenge is the fast pseudo-likelihood approach proposed by Amini et al. for fitting stochastic block models to large sparse networks. However, this approach does not have convergence guarantee, and may not be well suited for small and medium scale networks. In this article, we propose a novel likelihood based approach that decouples row and column labels in the likelihood function, enabling a fast alternating maximization. This new method is computationally efficient, performs well for both small- and large-scale networks, and has provable convergence guarantee. We show that our method provides strongly consistent estimates of communities in a stochastic block model. We further consider extensions of our proposed method to handle networks with degree heterogeneity and bipartite properties. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2020
Accepted July 2021

KEYWORDS
Network analysis; Profile likelihood; Pseudo likelihood; Stochastic block model; Strong consistency

1. Introduction

One of the fundamental problems in network data analysis is community detection which aims to divide the nodes in a network into several communities such that nodes within the same community are densely connected, and nodes from different communities are relatively sparsely connected. Identifying such communities can provide important insights on the organization of a network. For example, in social networks, communities may correspond to groups of individuals with common interests (Moody and White 2003); in protein interaction networks, communities may correspond to proteins that are involved in the same cellular functions (Spirin and Mirny 2003). There is a vast literature on network community detection contributed from different scientific communities, such as computer science, physics, social science, and statistics. We refer to Fortunato (2010), Fortunato and Hric (2016), and Zhao (2017) for comprehensive reviews on this topic.

In the statistics literature, the majority of community detection methods are model-based, which postulate and fit a probabilistic model that characterizes networks with community structures (Holland, Laskey, and Leinhardt 1983; Airoldi et al. 2008; Karrer and Newman 2011). Within this family, the stochastic block model (Holland, Laskey, and Leinhardt SBM; 1983) is perhaps the best studied and most commonly used. The SBM is a generative model, in which the nodes are divided into blocks, or communities, and the probability of an edge between two nodes only depends on which communities they belong to and is independent across edges once given the community assignment. Several extensions of the SBM have been considered, notably the mixed membership model (Airoldi et al. 2008), which allows each node to be associated with multiple clusters, and the degree corrected stochastic block model (Karrer and Newman DCSBM 2011), which accommodates degree heterogeneity by including additional degree parameters. Due to the rapidly increasing interests, the statistical literature on community detection in SBMs is fast growing with great advances on algorithmic solutions (Snijders and Nowicki 1997; Nowicki and Snijders 2001; Daudin, Picard, and Robin 2008; Karrer and Newman 2011; Decelle et al. 2011; Amini et al. 2013; Bickel et al. 2013, among others) and theoretical understandings of consistency and detection thresholds (Bickel and Chen 2009; Rohe, Chatterjee, and Yu 2011; Zhao, Levina, and Zhu 2012; Lei and Rinaldo 2015; Abbe 2017; Gao et al. 2017; Gao et al. 2018; Su, Wang, and Zhang 2019; Abbe et al. 2020, among others.)

It is well known that fitting the block model (i.e., SBM and DCSBM) likelihood functions is a nontrivial task, and in principle optimizing over all possible community assignments is a NP-hard problem (Bickel and Chen 2009). Many work have considered using spectral clustering for community detection in SBMs, which is computationally efficient and ensures weak consistency, that is, the proportion of misclassified nodes tends to zero as the network size increases, under certain regularity conditions (Rohe, Chatterjee, and Yu 2011; Lei and Rinaldo 2015; Joseph et al. 2016). As such, spectral clustering is often used to produce initializations for methods that aim to achieve strong consistency (Gao et al. 2017), that is, probability of the estimated label being equal to the true label converges to one as

CONTACT Jianhua Guo *jhguo@nenu.edu.cn* School of Mathematics and Statistics & KLAS, Northeast Normal University, Jilin 130024, China.
Ji Zhu *jizhu@umich.edu* Department of Statistics, University of Michigan, Ann Arbor, MI, USA.
The first three authors contributed equally to this work.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

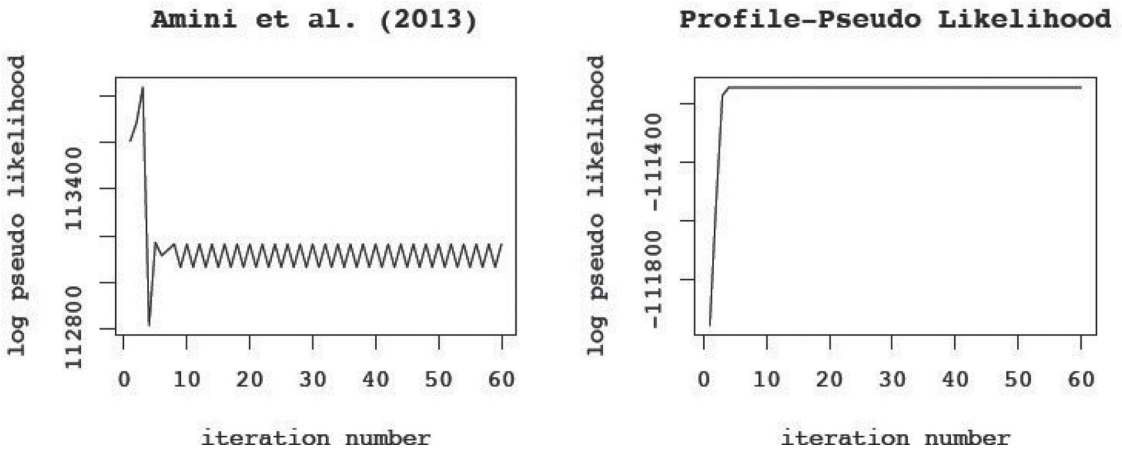

Figure 1. An illustrative example comparing the pseudolikelihood method by Amini et al. (2013) and the proposed profile-pseudo likelihood method. Details of the simulation setting are described in Section 5.1.

the network size grows, and methods that aim to directly maximize the nonconvex SBM and DCSBM likelihood functions (Amini et al. 2013; Bickel et al. 2013).

To overcome the computational challenge in fitting the SBM likelihood, Amini et al. (2013) proposed a novel pseudolikelihood approach that approximates the row rums within blocks using Poisson random variables, and simplifies the likelihood function by lifting the symmetry constraint on the adjacency matrix. This leads to a fast approximation to the block model likelihood, which subsequently enables efficient maximization that can easily handle up to millions of nodes. Additionally, it is shown that the maximum pseudolikelihood estimator achieves (weak) community detection consistency, in the case of a sparse SBM with two communities. This pioneer work makes the SBM an attractive approach for network community detection, due to its computational scalability and theoretical properties such as the community detection consistency. However, this method may have two drawbacks. First, in the examples that were presented in Amini et al. (2013), the authors found that empirically the pseudolikelihood maximization algorithm converged fast. It is, however, not guaranteed that the algorithm will converge in general (see, e.g., Figure 1). Convergence is a critical property as it guarantees that the final estimator exists, and is therefore important both computationally and statistically. Second, the pseudolikelihood approach may not be suitable for small and medium scale networks, as the Poisson approximation may have non negligible approximation errors in such cases. In the case of the DCSBM, cleverly employing the observation that the conditional distribution (on node degrees) of the Poisson variables is multinomial, Amini et al. (2013) proposed a conditional pseudolikelihood approach that permits a fast estimation and adapts to both small- and large-scale networks. However, the algorithm still does not have convergence guarantees.

Motivated by the pseudolikelihood approach, in this work, we propose a new SBM likelihood fitting method that decouples the membership labels of the rows and columns in the likelihood function, treating the row label as a vector of latent variables and the column label as a vector of unknown parameters. Correspondingly, the likelihood can be maximized in an alternating fashion over the block model parameters and over the column label, where the maximization now involves a tractable sum over the distribution of latent row label. Furthermore, we consider a profile-pseudo likelihood that adopts a hybrid framework of the profile likelihood and the pseudolikelihood, where the symmetry constraint on the adjacency matrix is also lifted. Our proposed method retains and improves on the computational efficiency of the pseudolikelihood method, performs well for both small and large scale networks and has provable convergence guarantee. We show that the community label (i.e., column label) estimated from our proposed method enjoys strong consistency, as long as the initial label has an overlap with the truth beyond that of random guessing. We further consider two extensions of the proposed method, including to the DCSBM and to the bipartite stochastic block model (Larremore, Clauset, and Jacobs BiSBM; 2014).

Our work is closely related to the recent and growing literature on strong consistency (or exact recovery) pursuit in community detection (see, e.g., Abbe, Bandeira, and Hall 2015; Lei and Zhu 2017; Gao et al. 2017; Gao et al. 2018). The strong consistency property may be more desirable than weak consistency, as it enables establishing the asymptotic normality of the SBM plug-in estimators (Amini et al. 2013) and performing goodness-of-fit tests (Lei 2016; Hu et al. 2020b). To achieve strong consistency, the above methods usually consider a refinement step after obtaining the initial label, which is assumed to obey weak consistency. For example, in Gao et al. (2017), a majority voting algorithm is applied to the clustering label obtained from spectral clustering. Similarly, our proposed profile-pseudo likelihood estimation can be viewed as a refinement on the initial label to achieve strong consistency. Similar to other refinement algorithms, the scalability of our proposed method depends on the initialization step. While spectral clustering is used to produce initial solutions in our work, other initialization methods can be considered as well (see Section 7).

The rest of the article is organized as follows. Section 2 introduces the profile-pseudo likelihood function and an efficient algorithm for its maximization. Moreover, we discuss the convergence guarantee of the algorithm. Section 3 shows the strong consistency property of the community label estimated from the proposed algorithm. Section 4 considers two important

extensions of the proposed method. Section 5 demonstrates the efficacy of the proposed method through comparative simulation studies. Section 6 presents analyses of two real-world networks with communities. A discussion section concludes the article.

2. Profile-Pseudo Likelihood

Let $G(V, E)$ denote a network, where $V = \{1, 2, \ldots, n\}$ is the set of n nodes and E is the set of edges between the nodes. The network $G(V, E)$ can be uniquely represented by the corresponding $n \times n$ adjacency matrix $\boldsymbol{A}$, where $A_{ij} = 1$ if there is an edge $(i, j) \in E$ from node i to node j and $A_{ij} = 0$ otherwise. In our work, we focus on unweighted and undirected networks, and thus $\boldsymbol{A}$ is a binary symmetric matrix. Under the stochastic block model, there are K communities (or blocks) and each node belongs to only one of the communities. Let $\boldsymbol{c} = (c_1, c_2, \ldots, c_n) \in \{1, 2, \ldots, K\}^n$ denote the true community labels of the nodes, and assume that c_i's are iid categorical variables with parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where $\sum_k \pi_k = 1$. Conditional on the community labels, the edge variables A_{ij}'s are independent Bernoulli variables with $\mathbb{E}(A_{ij}|\boldsymbol{c}) = P_{c_i c_j}$, where $\boldsymbol{P} \in [0, 1]^{K \times K}$ is the symmetric edge-probability matrix with the klth entry P_{kl} characterizing the probability of connection between nodes in communities k and l. Let $\boldsymbol{\Omega} = (\boldsymbol{\pi}, \boldsymbol{P})$. Our objective is to estimate the unknown community labels $\boldsymbol{c}$ given the observed adjacency matrix $\boldsymbol{A}$.

Denote the rows of $\boldsymbol{A}$ as $\boldsymbol{a}_i = (A_{i1}, A_{i2}, \ldots, A_{in})$, $1 \le i \le n$ and let $\boldsymbol{e} = (e_1, e_2, \ldots, e_n) \in \{1, 2, \ldots, K\}^n$ denote the column labeling vector. Define the pseudolikelihood function as

$$L_{\mathrm{PL}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\}) = \prod_{i=1}^{n} \left\{ \sum_{l=1}^{K} \pi_l \prod_{j=1}^{n} P_{l e_j}^{A_{ij}} (1 - P_{l e_j})^{1 - A_{ij}} \right\}, \quad (1)$$

with its logarithm as

$$\ell_{\mathrm{PL}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\}) = \sum_{i=1}^{n} \log \left\{ \sum_{l=1}^{K} \pi_l \prod_{j=1}^{n} P_{l e_j}^{A_{ij}} (1 - P_{l e_j})^{1 - A_{ij}} \right\}.$$

We make a few remarks on the objective function defined in Equation (1). First, in Equation (1), we treat the row labels as a vector of latent variables and the column labels $\boldsymbol{e}$ as a vector of unknown model parameters. That is, given e_j, each A_{ij} is considered a mixture of K Bernoulli random variables with mean $P_{l e_j}$, $1 \le l \le K$. This formulation decouples the row and column labels, and allows us to derive a tractable sum when optimizing for the column labels $\boldsymbol{e}$ and the block model parameter $\boldsymbol{\Omega}$. Second, the objective function $L_{\mathrm{PL}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ is calculated while lifting the symmetry constraint on the adjacency matrix $\boldsymbol{A}$, or equivalently, ignoring the dependence among the rows $\boldsymbol{a}_i$'s. Hence, we refer to Equation (1) as the pseudolikelihood function, which can be considered as an approximation to the SBM likelihood function.

We consider an iterative algorithm that alternates between updating $\boldsymbol{e}$ and updating $\boldsymbol{\Omega}$. In each iteration, the estimation is carried out by first profiling out the nuisance parameter $\boldsymbol{\Omega}$ using $\max_{\boldsymbol{\Omega}} L_{\mathrm{PL}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ given the current estimate of $\boldsymbol{e}$, and then maximizing the profile likelihood with respect to $\boldsymbol{e}$. This is referred to as the *profile-pseudo likelihood* method. We show in Theorem 1 the convergence guarantee of this efficient algorithm, and establish in Theorem 2 the strong consistency of the estimated column labels $\boldsymbol{e}$.

The estimation procedure proceeds in detail as follows. First, given the current $\hat{\boldsymbol{e}}$ and treating the row labels as a vector of latent variables, $L_{\mathrm{PL}}(\boldsymbol{\Omega}, \hat{\boldsymbol{e}}; \{\boldsymbol{a}_i\})$ can be viewed as the likelihood of a mixture model with iid observations $\{\boldsymbol{a}_i\}$ and parameter $\boldsymbol{\Omega}$. Consequently, $L_{\mathrm{PL}}(\boldsymbol{\Omega}, \hat{\boldsymbol{e}}; \{\boldsymbol{a}_i\})$ can be maximized over $\boldsymbol{\Omega}$ using an expectation-maximization (EM) algorithm, where both the E-step and M-step updates have closed-form expressions. Next, given the estimated $\widehat{\boldsymbol{\Omega}}$, we update $\boldsymbol{e}$, treating $L_{\mathrm{PL}}(\widehat{\boldsymbol{\Omega}}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ as the objective function. In this step, finding the maximizer of $L_{\mathrm{PL}}(\widehat{\boldsymbol{\Omega}}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ with respect to $\boldsymbol{e}$ is a NP-hard problem since, in principle, it requires searching over all possible label assignments. As an alternative, we propose a fast updating rule that leads to a nondecreasing objective function $L_{\mathrm{PL}}(\widehat{\boldsymbol{\Omega}}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ (although not necessarily maximized), which ensures the desirable ascent property of the iterative algorithm. This algorithm is summarized in Algorithm 1.

In what follows, we discuss in details the profile-pseudo likelihood algorithm. We refer to the iterations between updating $\boldsymbol{e}$ and $\boldsymbol{\Omega}$ as the outer iterations, and the iterations in the EM algorithm used to update $\boldsymbol{\Omega}$ as the inner iterations. Specifically, in the $(t+1)$-th step of the EM (inner) iteration, given $\boldsymbol{e}^{(s)}$ and the parameter estimate from the previous EM update $\boldsymbol{\Omega}^{(s,t)} = (\boldsymbol{\pi}^{(s,t)}, \boldsymbol{P}^{(s,t)})$, we let

$$\tau_{ik}^{(s,t+1)} = \frac{\pi_k^{(s,t)} \prod_{j=1}^{n} \left\{ P_{k e_j^{(s)}}^{(s,t)} \right\}^{A_{ij}} \left\{ 1 - P_{k e_j^{(s)}}^{(s,t)} \right\}^{1 - A_{ij}}}{\sum_{l=1}^{K} \pi_l^{(s,t)} \prod_{j=1}^{n} \left\{ P_{l e_j^{(s)}}^{(s,t)} \right\}^{A_{ij}} \left\{ 1 - P_{l e_j^{(s)}}^{(s,t)} \right\}^{1 - A_{ij}}} \quad (2)$$

for each $1 \le i \le n$ and $1 \le k \le K$, which calculates the conditional probability that the row label of node i equals to k at the $(t+1)$th step of the EM iteration. Next, we define

$$Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)}) = \mathbb{E}_{\boldsymbol{z}|\{\boldsymbol{a}_i\}; \boldsymbol{\Theta}^{(s,t)}, \boldsymbol{e}^{(s)}} \left\{ \log f\left(\{\boldsymbol{a}_i\}, \boldsymbol{z}; \boldsymbol{\Omega}, \boldsymbol{e}^{(s)} \right) \right\},$$

where $\boldsymbol{z}$ denotes the latent row labels and

$$f(\{\boldsymbol{a}_i\}, \boldsymbol{z}; \boldsymbol{\Omega}, \boldsymbol{e}^{(s)}) = \prod_{i=1}^{n} \left\{ \pi_{z_i} \prod_{j=1}^{n} P_{z_i e_j^{(s)}}^{A_{ij}} (1 - P_{z_i e_j^{(s)}})^{1 - A_{ij}} \right\}.$$

In the M-step, $\boldsymbol{\Omega}^{(s,t+1)}$ is updated by

$$\boldsymbol{\Omega}^{(s,t+1)} = \arg\max_{\boldsymbol{\Omega}} Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)}),$$

which has closed-form solutions as follows:

$$\pi_k^{(s,t+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(s,t+1)}, \quad (3)$$

$$P_{kl}^{(s,t+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \tau_{ik}^{(s,t+1)} I(e_j^{(s)} = l)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \tau_{ik}^{(s,t+1)} I(e_j^{(s)} = l)},$$

for $1 \le k, l \le K$. Once the EM algorithm has converged, we let $\boldsymbol{\Omega}^{(s+1)}$ and $\left\{ \tau_{il}^{(s+1)} \right\}$ take the values from the last EM

Algorithm 1 Profile-Pseudo Likelihood Maximization Algorithm.

Step 1: Initialize $\boldsymbol{e}^{(0)}$ using spectral clustering with permutations (SCP).

Step 2: Calculate $\boldsymbol{\Omega}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{P}^{(0)})$. That is, for $1 \le l, k \le K$,

$$\pi_k^{(0)} = \frac{1}{n}\sum_{i=1}^{n} I(e_i^{(0)} = k), \quad P_{kl}^{(0)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} I(e_i^{(0)}=k) I(e_j^{(0)}=l)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I(e_i^{(0)}=k) I(e_j^{(0)}=l)}.$$

Step 3: Initialize $\boldsymbol{\Omega}^{(0,0)} = (\boldsymbol{\pi}^{(0,0)}, \boldsymbol{P}^{(0,0)}) = (\boldsymbol{\pi}^{(0)}, \boldsymbol{P}^{(0)})$.

repeat

  repeat

    Step 4: E-step: compute $\tau_{ik}^{(s,t+1)}$ using (2) for $1 \le k \le K$ and $1 \le i \le n$.

    Step 5: M-step: compute $\pi_k^{(s,t+1)}$ and $P_{kl}^{(s,t+1)}$ using (3) for $1 \le k, l \le K$.

  until the EM algorithm converges.

  Step 6: Set $\boldsymbol{\Omega}^{(s+1)}$ and $\left\{\tau_{ik}^{(s+1)}\right\}$ to be the final EM update.

  Step 7: Given $\boldsymbol{\Omega}^{(s+1)}$ and $\left\{\tau_{ik}^{(s+1)}\right\}$, update $e_j^{(s+1)}$, $1 \le j \le n$, using (4).

until the profile-pseudo likelihood converges.

update, respectively. Next, given $\boldsymbol{\Omega}^{(s+1)}$, we propose to update $\boldsymbol{e}$ as follows:

$$e_j^{(s+1)} = \arg\max_{k\in\{1,2,\ldots,K\}} \sum_{i=1}^{n}\sum_{l=1}^{K} \tau_{il}^{(s+1)} \left\{A_{ij}\log P_{lk}^{(s+1)} + (1-A_{ij})\log\left(1-P_{lk}^{(s+1)}\right)\right\}. \quad (4)$$

The update for $\boldsymbol{e}^{(s+1)}$ is obtained separately for each node, which can be carried out efficiently. As we discussed earlier, this update is not guaranteed to maximize the pseudolikelihood function $L_{\mathrm{PL}}(\boldsymbol{\Omega}^{(s+1)}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$, which in fact is an intractable problem. Nevertheless, it can be shown that the update in Equation (4) leads to a nonnegative increment in the pseudolikelihood. This gives the desirable ascent property, which we will formally state in the following theorem.

Theorem 1. For a given initial labeling vector $\boldsymbol{e}^{(0)}$, Algorithm 1 generates a sequence $\{\boldsymbol{\Omega}^{(s)}, \boldsymbol{e}^{(s)}\}$ such that

$$L_{\mathrm{PL}}(\boldsymbol{\Omega}^{(s)}, \boldsymbol{e}^{(s)}; \{\boldsymbol{a}_i\}) \le L_{\mathrm{PL}}(\boldsymbol{\Omega}^{(s+1)}, \boldsymbol{e}^{(s+1)}; \{\boldsymbol{a}_i\}).$$

The proof of Theorem 1 is provided in the supplemental material. Theorem 1 guarantees that the pseudolikelihood function is non-decreasing at each iteration in Algorithm 1. Assuming that the parameter space for $\boldsymbol{\Omega}$ is compact, we arrive at the conclusion that $L_{\mathrm{PL}}(\boldsymbol{\Omega}^{(s)}, \boldsymbol{e}^{(s)}; \{\boldsymbol{a}_i\})$ converges as the number of iterations s increases. This is a desirable property that guarantees the stability of the proposed algorithm. Since the pseudolikelihood function is not concave, Algorithm 1 is not guaranteed to converge to the global optimum. Whether it converges to a global or local solution depends on the initial value. In practice, we find that the initialization procedure in Algorithm 1 shows good performance, that is, we are able to achieve high clustering accuracy in our simulation studies. To avoid local solutions in real data applications, we recommend considering multiple random initializations in addition to the initialization in Algorithm 1.

Finally, we summarize the differences between our proposal and the method in Amini et al. (2013). Both our method and Amini et al. (2013) considered algorithms that iterate through two parameter updating steps, namely the step that updates the block model parameter $\boldsymbol{\Omega}$ using EM and the step that updates the membership label. However, the likelihood function is treated very differently in these two methods. As the row and column labels are enforced to be the same in Amini et al. (2013), a Poisson approximation is needed in the pseudolikelihood calculation. The label $\boldsymbol{e}$ in Amini et al. (2013) is treated as an initial in the EM estimation, and its value is assigned heuristically in each iteration. As such, the resulting procedure is not guaranteed to converge, as seen in Figure 1. In comparison, our method decouples the row and column labels (i.e., $\boldsymbol{z}$ and $\boldsymbol{e}$), and does not require a Poisson approximation in the pseudo likelihood calculation. When updating the column labels $\boldsymbol{e}$, we use $L_{\mathrm{PL}}(\widehat{\boldsymbol{\Omega}}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ as the objective function that guides our updating routine. The proposed node-wise update enjoys the ascent property, which subsequently guarantees the convergence of the algorithm (see Theorem 1). We also remark that due to the differences in our problem formulation, our theoretical analysis is nontrivial and new technical tools are needed.

3. Consistency Results

In this section, we investigate the strong consistency of the estimator obtained from one outer loop iteration (i.e., updating the column labels $\boldsymbol{e}$) of Algorithm 1, denoted as $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$, where $\boldsymbol{e}^{(0)}$ is an initial of Algorithm 1. We first consider strong consistency in the case of SBMs with two balanced communities, and then extend our strong consistency result to SBMs with K communities.

We first present the consistency result for directed SBMs with two communities, fitted to directed networks, and then modify the result to handle the more challenging case of undirected SBMs, fitted to undirected networks. To separate the cases of directed and undirected SBMs, we adopt different notations for the corresponding adjacency matrices and edge-probability matrices. First, for a directed SBM, we denote the adjacency matrix as $\widetilde{\boldsymbol{A}}$ and assume that its entries $\widetilde{A}_{ij}$'s are mutually independent given $\boldsymbol{c}$, that is,

$$\text{(directed) } \widetilde{A}_{ij}|\boldsymbol{c} \sim \text{Bernoulli}(\widetilde{P}_{c_i c_j}), \qquad \text{for} \quad 1 \le i, j \le n. \quad (5)$$

For an undirected SBM, we denote the adjacency matrix as $\boldsymbol{A}$ and assume its entries A_{ij}'s, $i \le j$, are mutually independent given $\boldsymbol{c}$, that is,

$$\text{(undirected) } A_{ij}|\boldsymbol{c} \sim \text{Bernoulli}(P_{c_i c_j}) \quad \text{and} \quad A_{ij} = A_{ji}, \quad \text{for} \quad 1 \le i \le j \le n. \quad (6)$$

Furthermore, we assume that the edge-probability matrix of the directed SBM has the form

$$\widetilde{\boldsymbol{P}} = \frac{1}{m}\begin{pmatrix} a & b \\ b & a \end{pmatrix}, \quad (7)$$

while that of the undirected SBM has the form

$$\boldsymbol{P} = \frac{2}{m}\begin{pmatrix} a & b \\ b & a \end{pmatrix} - \frac{1}{m^2}\begin{pmatrix} a^2 & b^2 \\ b^2 & a^2 \end{pmatrix}. \tag{8}$$

Such a coupling between the directed and undirected models makes it possible to extend the consistency result of the directed SBM to the undirected case.

Given an initial labeling vector $\boldsymbol{e}^{(0)}$, estimates $\hat{a}$, $\hat{b}$, and $(\hat{\pi}_1, \hat{\pi}_2)$, the estimator $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$ can be written as follows:

$$\hat{c}_j\{\boldsymbol{e}^{(0)}\} = \arg\max_{k\in\{1,2\}} \sum_{i=1}^{n}\sum_{l=1}^{2} \hat{\tau}_{il}\left\{A_{ij}\log(\widehat{P}_{lk}) + (1-A_{ij})\log(1-\widehat{P}_{lk})\right\}, \tag{9}$$

where $\hat{\tau}_{il}$ is defined as in Equation (2), $\widehat{\boldsymbol{P}}$ is defined as in Equation (7) for directed SBMs and as in Equation (8) for undirected SBMs, with a and b replaced by $\hat{a}$ and $\hat{b}$, respectively. Here, the estimates $\hat{a}$, $\hat{b}$ and $(\hat{\pi}_1, \hat{\pi}_2)$ are outputs from the inner loop (i.e., EM) iterations, and are in effect initials for the outer loop calculation. Consistency of the inner loop (i.e., EM) outputs $\hat{a}$, $\hat{b}$ and $(\hat{\pi}_1, \hat{\pi}_2)$ can be established using the result in Amini et al. (2013). In our theoretical analysis, we focus our efforts on establishing strong consistency of the column labels $\boldsymbol{e}$ estimated in the outer loop, given that the outer loop initials satisfy $(\hat{a}, \hat{b}) \in \mathcal{P}^{\delta}_{a,b}$ in Equation (10) and $\hat{\pi}_1 = \hat{\pi}_2 = 1/2$.

For SBMs with two balanced communities, we make the following assumption:

1. Assume that each community contains $m = n/2$ nodes and $\hat{\pi}_1 = \hat{\pi}_2 = 1/2$.

The assumption that $\hat{\pi}_1 = \hat{\pi}_2 = 1/2$ is reasonable as the inner loop outputs $(\hat{\pi}_1, \hat{\pi}_2)$ are consistent estimators of $(\pi_1, \pi_2) = (1/2, 1/2)$, as shown in Amini et al. (2013). Without loss of generality, let $c_i = 1$ for $i \in \{1, \ldots, m\}$, and $c_i = 2$ for $i \in \{m+1, \ldots, n\}$. Assume that $\boldsymbol{e}^{(0)} \in \{1,2\}^n$ assigns equal numbers of nodes to the two communities, that is, the initial labeling vector is balanced. Let $\boldsymbol{e}^{(0)}$ match with the truth on γm labels in each of the two communities for some $\gamma \in (0,1)$. We assume γm to be an integer. Next, let $\mathcal{E}^{\gamma}$ denote the set that collects all such initial labeling vectors, that is,

$$\mathcal{E}^{\gamma} = \left\{\boldsymbol{e}^{(0)} \in \{1,2\}^n : \sum_{i=1}^{m} I(e_i^{(0)} = 1) = \gamma m, \sum_{i=m+1}^{n} I(e_i^{(0)} = 2) = \gamma m\right\}.$$

Note that $\gamma = 1/2$ corresponds to "no correlation" between $\boldsymbol{e}^{(0)}$ and $\boldsymbol{c}$, whereas $\gamma = 0$ and $\gamma = 1$ both correspond to perfect correlation. In our analysis, we do not require knowing the value of γ, or knowing which labels are matched. In Theorem 2, we show that the amount of overlap γ can be any value, as long as $\gamma \neq 1/2$. Our goal is to establish strong consistency for $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$. For a constant $\delta > 1$, we define $\mathcal{P}^{\delta}_{a,b}$ as follows:

$$\mathcal{P}^{\delta}_{a,b} = \left\{(\hat{a}, \hat{b}) : \frac{\hat{a}}{\hat{b}} I(a > b) + \frac{\hat{b}}{\hat{a}} I(a < b) \geq \delta\right\}. \tag{10}$$

The set $\mathcal{P}^{\delta}_{a,b}$ specifies that $(\hat{a}, \hat{b})$ has the same ordering as (a, b), and the relative difference between the estimates $\hat{a}$ and $\hat{b}$ is lower bounded. Our next theorem considers the collection of estimates $(\hat{a}, \hat{b})$ in $\mathcal{P}^{\delta}_{a,b}$.

Theorem 2. Assume (A) holds, $\delta > 1$, $\gamma \in (0,1)\backslash\{\frac{1}{2}\}$ and $\frac{(a-b)^2}{(a+b)} \geq C\log n$ for a sufficiently large constant $C > 0$. For a directed SBM in Equation (5) with the edge-probabilities given by Equation (7) with $a \neq b$, we have that for any $\epsilon > 0$, there exists $N > 0$ such that for all $n \geq N$, the following holds

$$\mathbb{P}\left\{\bigcap_{(\hat{a},\hat{b})\in\mathcal{P}^{\delta}_{a,b}} \hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}\right\} \geq 1 - \left\{ne^{-\frac{(a-b)^2 - 4(a-b)\epsilon + 4\epsilon^2}{4(a+b)}} + n(n+2)e^{-\frac{(2\gamma-1)^2(a-b)^2}{8(a+b)}}\right\},$$

for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^{\gamma}$, where $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}$ means that they belong to the same equivalent class of label permutations.

The proof of Theorem 2 is provided in the supplemental material. It can be seen from Theorem 2 that the one-step estimate $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$ for a directed SBM is a strongly consistent estimate of $\boldsymbol{c}$ for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^{\gamma}$. Note that weak consistency was established in Amini et al. (2013) under the assumption that $\frac{(a-b)^2}{(a+b)} \to \infty$. In comparison, our result requires $\frac{(a-b)^2}{(a+b)} \geq C\log n$ to establish strong consistency. In the existing literature on strong consistency, the condition $\frac{\lambda_n}{\log n} \to \infty$ is often commonly imposed (Bickel and Chen 2009; Zhao, Levina, and Zhu 2012), where λ_n denotes the average network degree. Specifically, under the SBM setting considered in Bickel and Chen (2009) and Zhao, Levina, and Zhu (2012), we have that $a - b \asymp \lambda_n$ and $a + b \asymp \lambda_n$, where $\asymp$ denotes that the two quantities on both sides are of the same order. In this case, $\frac{\lambda_n}{\log n} \to \infty$ implies $\frac{(a-b)^2}{(a+b)} \geq C\log n$ for any constant $C > 0$.

Theorem 2 guarantees strong consistency for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^{\gamma}$. In comparison, the weak consistency in Amini et al. (2013) holds uniformly for all $\boldsymbol{e}^{(0)} \in \mathcal{E}^{\gamma}$, even if it is derived from the data. Indeed, $\boldsymbol{e}^{(0)}$ is usually derived from data using initialization procedures such as the spectral clustering. For the strong consistency result to apply, one may consider a data splitting strategy following the method in Li, Levina, and Zhu (2020). Specifically, we may sample a proportion of the node pairs to produce an initial value $\boldsymbol{e}^{(0)}$ and estimate $\hat{\boldsymbol{c}}(\boldsymbol{e}^{(0)})$ using the rest of the node pairs. In this case, $\boldsymbol{e}^{(0)}$ is independent of the data used for community detection and the result in Theorem 2 can be used to ensure strong consistency of $\hat{\boldsymbol{c}}(\boldsymbol{e}^{(0)})$. In our numerical studies, for simplicity we did not use data splitting, while the simulation results show that the proposed method still performs well. We also note that Theorem 2 can be adapted to hold uniformly for all $\boldsymbol{e}^{(0)} \in \mathcal{E}^{\gamma}$, if stronger conditions are placed on γ and a, b. Specifically, if the misclassification ratio of $\boldsymbol{e}^{(0)}$ is, for example, $O(1/(a+b))$ and the condition on a, b is strengthen to $(a-b) \gtrsim \sqrt{n\log n}$ (i.e., average degree is at least of order $\sqrt{n\log n}$), then strong consistency in Theorem 2 holds uniformly for all such $\boldsymbol{e}^{(0)}$, even if it is derived from the data. This can be shown by combining the union bound

argument and a Stirling approximation that gives $\log \binom{n}{n_\gamma} \le n\log(en/n_\gamma)$, where n_γ is the number of misclassified nodes. The misclassification ratio of $O(1/(a+b))$ imposed above is known to hold with high probability for spectral clustering (see, e.g., Corollary 3.2 in Lei and Rinaldo (2015)).

Next, we consider the case of undirected SBMs. Let $a_\gamma = \big[(1-\gamma)a+\gamma b\big]I(\gamma > \frac{1}{2}) + \big[\gamma a+(1-\gamma)b\big]I(\gamma < \frac{1}{2})$. We have the following result on the strong consistency of $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$.

Theorem 3. Assume (A) holds, $\delta > 1$, $\gamma \in (0,1)\backslash\{\frac{1}{2}\}$ and $\frac{(a-b)^2}{(a+b)} \ge C\log n$ for a sufficiently large constant $C > 0$. For an undirected SBM in Equation (6) with the edge-probabilities given by Equation (8) with $2(1+\epsilon)a_\gamma \le \epsilon|(1-2\gamma)(a-b)|$ for some $\epsilon \in (0,1)$, there exist $\rho \in (0,1)$ and $N > 0$, such that for all $n \ge N$, the following holds:

$$\mathbb{P}\Bigg\{\bigcap_{(\hat{a},\hat{b})\in\mathcal{P}^\delta_{a,b}} \hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}\Bigg\} \ge 1 - \Bigg[3ne^{-\frac{(\frac{1-\rho}{4})^2(a-b)^2}{4(a+b)}} + n(n+2)\left\{e^{-\frac{(\frac{1-\epsilon}{2})^2(2\gamma-1)^2(a-b)^2}{4(a+b)}} + 2e^{-\frac{\epsilon^2/2}{1+\epsilon/2}a_\gamma}\right\}\Bigg],$$

for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^\gamma$, where $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}$ means that they belong to the same equivalent class of label permutations.

The proof of Theorem 3 is provided in the supplemental material. It can be seen that the one-step estimate $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\}$ for an undirected SBM is a strongly consistent estimate of $\boldsymbol{c}$, for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^\gamma$. Given ϵ and γ, the condition $2(1+\epsilon)a_\gamma \le \epsilon|(1-2\gamma)(a-b)|$ places an upper bound on b/a. For example, for $\epsilon = \frac{1}{3}$ and $\gamma < \frac{1}{10}$, the above condition is satisfied if $b/a \le (1-10\gamma)/(9-10\gamma)$.

Strong consistency can be more desirable than weak consistency, as it enables normal distribution based inference and goodness-of-fit tests (see numerical studies in Section 5.2). For example, consider a SBM with $K = 2$, $\boldsymbol{\pi} = (\pi_1, \pi_2)$ and true community labels $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$. Suppose we can construct a label vector $\hat{\boldsymbol{c}}^{(\mathrm{w})}$ such that $\{\hat{c}_i^{(\mathrm{w})}\}_{i=1}^n$ are independent with $\mathbb{P}(\hat{c}_i^{(\mathrm{w})} \ne c_i) = 2p_n$ for $c_i = 1$ and $\mathbb{P}(\hat{c}_i^{(\mathrm{w})} \ne c_i) = p_n$ for $c_i = 2$, where $p_n = 1/\log n$. Then it can be shown that $\hat{\boldsymbol{c}}^{(\mathrm{w})}$ is weakly consistent, with a misclassification ratio of $O_p(1/\log n)$, but not strongly consistent to $\boldsymbol{c}$. Let $\hat{\pi}_1^{\mathrm{w}} = \sum_{i=1}^n I(\hat{c}_i^{(\mathrm{w})} = 1)/n$. It holds that $\sqrt{n}\left\{\hat{\pi}_1^{\mathrm{w}} - \left(\pi_1 + \frac{1-3\pi_1}{\log n}\right)\right\} \xrightarrow{d} N\{0, \pi_1(1-\pi_1)\}$ (See the proof in the supplemental material). Thus, the bias term of $\hat{\pi}_1^{\mathrm{w}}$ is $O(1/\log n)$, which can be non negligible for inference. On the other hand, for a strongly consistent estimator $\hat{\boldsymbol{c}}^{(\mathrm{s})} = \left(\hat{c}_1^{(\mathrm{s})}, \hat{c}_2^{(\mathrm{s})}, \ldots, \hat{c}_n^{(\mathrm{s})}\right)$, letting $\hat{\pi}_1^{\mathrm{s}} = \sum_{i=1}^n I(\hat{c}_i^{(\mathrm{s})} = 1)/n$, it holds that $\sqrt{n}\{\hat{\pi}_1^{\mathrm{s}} - \pi_1\} \xrightarrow{d} N\{0, \pi_1(1-\pi_1)\}$.

Next, we consider the more general case of directed and undirected SBMs with K communities. Similar to Assumption (A), we make the following assumption:

1. Assume that each community contains $m = n/K$ nodes and $\hat{\pi}_k = 1/K$.

Let the edge-probability matrix of the directed SBM be

$$\widetilde{P}_{kl} = \frac{a}{m}1(k=l) + \frac{b}{m}1(k\ne l), \tag{11}$$

and that of the undirected SBM be

$$P_{kl} = \left(\frac{2a}{m} - \frac{a^2}{m^2}\right)1(k=l) + \left(\frac{2b}{m} - \frac{b^2}{m^2}\right)1(k\ne l), \tag{12}$$

for $k, l = 1, \ldots, K$. Without loss of generality, let $c_i = k$ for $i \in \{(k-1)m+1, \ldots, km\}$ for $k = 1, \ldots, K$. Let $\mathcal{E}^\gamma$ denote the set that collects all initial labeling vectors such that

$$\mathcal{E}^\gamma = \Bigg\{\boldsymbol{e}^{(0)} \in \{1,\ldots,K\}^n : \sum_{i=(k-1)m+1}^{km} I(e_i^{(0)} = k) = \gamma_k m, \sum_{i=1}^n I(e_i^{(0)} = k) = m,\ \ k = 1,\ldots,K\Bigg\},$$

where $\gamma = (\gamma_1, \ldots, \gamma_K)$. Corollaries 1 and 2 establish the strong consistency of profile-pseudo likelihood estimators for directed and undirected SBMs, respectively.

Corollary 1. Assume (B) holds, $\delta > 1$, $\min\{\gamma_1, \gamma_2, \ldots, \gamma_K\} \in (\frac{1}{2}, 1)$ and $\frac{(a-b)^2}{(a+b)} \ge C\log n$ for a sufficiently large constant $C > 0$. For a directed SBM in Equation (5) with the edge-probabilities given by Equation (11) with $a \ne b$, we have that for each $\epsilon > 0$, there exists $N > 0$ such that for all $n \ge N$, the following holds:

$$\mathbb{P}\Bigg\{\bigcap_{(\hat{a},\hat{b})\in\mathcal{P}^\delta_{a,b}} \hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}\Bigg\} \ge 1 - \Bigg\{(K-1)ne^{-\frac{(a-b)^2-4(a-b)\epsilon+4\epsilon^2}{4(a+b)}} + \frac{(10K-8)n^2}{K}\sum_{k=1}^K\sum_{l=1}^K e^{-\frac{(\gamma_k+\gamma_l-1)^2(a-b)^2}{8(a+b)}}\Bigg\},$$

for any $\boldsymbol{e}^{(0)} \in \mathcal{E}^\gamma$, where $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}$ means that they belong to the same equivalent class of label permutations.

Corollary 2. Assume (B) holds, $\delta > 1$, $\min\{\gamma_1, \gamma_2, \ldots, \gamma_K\} \in (\frac{1}{2}, 1)$ and $\frac{(a-b)^2}{(a+b)} \ge C\log n$ for a sufficiently large constant $C > 0$. For an undirected SBM in Equation (6) with the edge-probabilities given by (12) with $2(1+\epsilon)a_{\gamma_k} \le \epsilon(\gamma_k+\gamma_l-1)(a-b)$ for all $1 \le k, l \le K$ and some $\epsilon \in (0,1)$, where $a_{\gamma_k} = (1-\gamma_k)a + \gamma_k b$, there exist $\rho \in (0,1)$ and $N > 0$, such that for all $n \ge N$, the following holds

$$\mathbb{P}\Bigg\{\bigcap_{(\hat{a},\hat{b})\in\mathcal{P}^\delta_{a,b}} \hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}\Bigg\} \ge 1 - \Bigg[3(K-1)ne^{-\frac{(\frac{1-\rho}{4})^2(a-b)^2}{2(a+b)}} + \frac{(10K-8)n^2}{K}\sum_{k=1}^K\sum_{l=1}^K\left\{e^{-\frac{(\frac{1-\epsilon}{2})^2(\gamma_k+\gamma_l-1)^2(a-b)^2}{6(a+b)}} + 2e^{-\frac{3\epsilon^2 a_{\gamma_k}}{8(4+\epsilon)}}\right\}\Bigg],$$

for any $\boldsymbol{e}^{(0)} \in \varepsilon_n^\gamma$, where $\hat{\boldsymbol{c}}\{\boldsymbol{e}^{(0)}\} = \boldsymbol{c}$ means that they belong to the same equivalent class of label permutations.

The proofs of Corollaries 1 and 2 follow very similar steps as in the proofs of Theorems 2 and 3, respectively. We omit presenting the details.

4. Extensions

In this section, we study two useful extensions of the proposed method. First, we consider the case of fitting the degree corrected stochastic block model using the proposed profile-pseudo likelihood method. Second, we consider the case of fitting the bipartite stochastic block model using the proposed profile-pseudo likelihood method (see Section A5 in the supplemental material).

It has often been observed that real-world networks exhibit high degree heterogeneity, with a few nodes having a large number of connections and the majority of the rest having a small number of connections. The stochastic block model, however, cannot accommodate such degree heterogeneity. To incorporate the degree heterogeneity in community detection, Karrer and Newman (2011) proposed the degree-corrected SBM. Specifically, conditional on the label vector $\boldsymbol{c}$, it is assumed that the edge variables A_{ij} for all $i \le j$ are mutually independent Poisson variables with

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = \theta_i\theta_j\lambda_{c_ic_j},$$

where $\boldsymbol{\Lambda} = [\lambda_{kl}]$ is a $K \times K$ symmetric matrix and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ is a degree parameter vector, with the additional constraint $\sum_{i=1}^{n} \theta_i/n = 1$ that ensures identifiability (Zhao, Levina, and Zhu 2012).

Define $\boldsymbol{\Omega} = (\boldsymbol{\pi}, \boldsymbol{\Lambda}, \boldsymbol{\theta})$. To fit the DCSBM to an observed adjacency matrix $\boldsymbol{A}$, we define the following log-pseudolikelihood function:

$$\ell_{\mathrm{PL}}^{\mathrm{DC}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\}) = \sum_{i=1}^{n} \log \left\{ \sum_{l=1}^{K} \pi_l \prod_{j=1}^{n} e^{-\theta_i\theta_j\lambda_{le_j}} (\theta_i\theta_j\lambda_{le_j})^{A_{ij}} \right\}.$$

Let $d_i = \sum_{j=1}^{n} A_{ij}$, $1 \le i \le n$. A profile-pseudo likelihood algorithm that maximizes $\ell_{\mathrm{PL}}^{\mathrm{DC}}(\boldsymbol{\Omega}, \boldsymbol{e}; \{\boldsymbol{a}_i\})$ is described in Algorithm 2. At step 4, we update the conditional probabilities for the row labels by

$$\tau_{ik}^{(s,t+1)} = \frac{\prod_{j=1}^{n} \pi_k^{(s,t)} e^{-\theta_i^{(s,t)}\theta_j^{(s,t)}\lambda_{ke_j^{(s)}}^{(s,t)}} \left\{ \theta_i^{(s,t)}\theta_j^{(s,t)}\lambda_{ke_j^{(s)}}^{(s,t)} \right\}^{A_{ij}}}{\sum_{l=1}^{K} \prod_{j=1}^{n} \pi_l^{(s,t)} e^{-\theta_i^{(s)}\theta_j^{(s,t)}\lambda_{le_j^{(s)}}^{(s,t)}} \left\{ \theta_i^{(s,t)}\theta_j^{(s,t)}\lambda_{le_j^{(s)}}^{(s,t)} \right\}^{A_{ij}}}. \quad (13)$$

At step 5, we update the parameters by sequentially solving the following optimization problems:

$$(\boldsymbol{\pi}^{(s,t+1)}, \boldsymbol{\Lambda}^{(s,t+1)}) = \arg\max_{(\boldsymbol{\pi},\boldsymbol{\Lambda})} Q(\boldsymbol{\pi}, \boldsymbol{\Lambda}, \boldsymbol{\theta}^{(s,t)}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)}),$$

$$\theta_i^{(s,t+1)} = \arg\max_{\theta_i} Q(\boldsymbol{\pi}^{(s,t+1)}, \boldsymbol{\Lambda}^{(s,t+1)}, \theta_1^{(s,t+1)}, \ldots, \theta_{i-1}^{(s,t+1)}, \theta_i, \theta_{i+1}^{(s,t)}, \ldots, \theta_n^{(s,t)}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)}).$$

Here, the objective function $Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)})$ is defined as

$$Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)}) = \mathbb{E}_{\boldsymbol{z}|\{\boldsymbol{a}_i\};\boldsymbol{\Theta}^{(s,t)},\boldsymbol{e}^{(s)}} \left\{ \log f\left(\{\boldsymbol{a}_i\}, \boldsymbol{z}; \boldsymbol{\Omega}, \boldsymbol{e}^{(s)} \right) \right\},$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ denotes the row label vector and

$$f(\{\boldsymbol{a}_i\}, \boldsymbol{z}; \boldsymbol{\Omega}, \boldsymbol{e}^{(s)}) = \prod_{i=1}^{n} \left[\pi_{z_i} \prod_{j=1}^{n} e^{-\theta_i\theta_j\lambda_{z_ie_j^{(s)}}} \frac{\left\{ \theta_i\theta_j\lambda_{z_ie_j^{(s)}} \right\}^{A_{ij}}}{A_{ij}!} \right].$$

Algorithm 2 DCSBM Profile-Pseudo Likelihood Maximization Algorithm.

Step 1: Initialize $\boldsymbol{e}^{(0)}$ using spectral clustering with permutations (SCP).

Step 2: Calculate $\boldsymbol{\Omega}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\Lambda}^{(0)}, \boldsymbol{\theta}^{(0)})$. That is, for $1 \le l, k \le K$, $1 \le i \le n$,

$$\pi_k^{(0)} = \frac{1}{n} \sum_{i=1}^{n} I(e_i^{(0)} = k), \quad \theta_i^{(0)} \propto d_i,$$

$$\lambda_{kl}^{(0)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} I(e_i^{(0)} = k) I(e_j^{(0)} = l)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I(e_i^{(0)} = k) I(e_j^{(0)} = l)\theta_i^{(0)}\theta_j^{(0)}}.$$

Step 3: Initialize $\boldsymbol{\Omega}^{(0,0)} = (\boldsymbol{\pi}^{(0,0)}, \boldsymbol{\Lambda}^{(0,0)}, \boldsymbol{\theta}^{(0,0)}) = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\Lambda}^{(0)}, \boldsymbol{\theta}^{(0)})$.

repeat

repeat

Step 4: E-step: compute $\tau_{ik}^{(s,t+1)}$ using (13) for $1 \le k \le K$ and $1 \le i \le n$.

Step 5: CM-step: compute $\boldsymbol{\pi}^{(s,t+1)}$, $\boldsymbol{\Lambda}^{(s,t+1)}$, $\boldsymbol{\theta}^{(s,t+1)}$. For $1 \le k, l \le K$, set

$$\pi_k^{(s,t+1)} = \sum_{i=1}^{n} \tau_{ik}^{(s,t+1)}/n,$$

$$\lambda_{kl}^{(s,t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_{ik}^{(s,t+1)} I(e_j^{(s)} = l) A_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_{ik}^{(s,t+1)} I(e_j^{(s)} = l)\theta_i^{(s,t)}\theta_j^{(s,t)}},$$

Letting $g_{ij}^{(s,t+1)} = \sum_{k,l=1}^{K} \tau_{ik}^{(s,t+1)} I(e_j^{(s)} = l)\lambda_{kl}^{(s,t+1)}$, for $1 \le i \le n$, set

$$\theta_i^{(s,t+1)} = \left(-h_i^{(s,t+1)} + \sqrt{h_i^{(s,t+1)2} + 8d_i g_{ii}^{(s,t+1)}} \right) \Big/ 4g_{ii}^{(s,t+1)},$$

where $h_i^{(s,t+1)} = \sum_{j=1}^{i-1} \theta_j^{(s,t+1)} g_{ij}^{(s,t+1)} + \sum_{j=i+1}^{n} \theta_j^{(s,t)} g_{ij}^{(s,t+1)}$.

until the ECM algorithm converges.

Step 6: Set $\boldsymbol{\Omega}^{(s+1)}$ to be the final ECM update.

Step 7: Given $\boldsymbol{\Omega}^{(s+1)}$, update $e_j^{(s+1)}$, $1 \le j \le n$, using

$$e_j^{(s+1)} = \arg\max_{k \in \{1,2,\ldots,K\}} \sum_{i=1}^{n}\sum_{l=1}^{K} \left\{ -\theta_i^{(s+1)}\theta_j^{(s+1)}\lambda_{lk}^{(s+1)} + A_{ij}\log(\lambda_{lk}^{(s+1)}) \right\} \tau_{il}^{(s+1)}.$$

until the profile-pseudo likelihood converges.

The inner loop of Algorithm 2, that is, Steps 4 and 5, is different from that in Algorithm 1, as it considers a conditional EM (ECM) update. Specifically, the objective function $Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(s,t)}, \boldsymbol{e}^{(s)})$ in the M-step, that is, Step 5, which solves for

block parameters λ_{kl}'s and degree parameters θ_i's, is nonconvex and does not have closed form solutions. Hence, directly optimizing it using numerical techniques can be computationally costly and is not ensured to find the global optimum. The ECM algorithm replaces the challenging optimization problem in the M-step with a sequence of alternating updates, each of which has a closed-form solution. It is easy to implement and enjoys the desirable ascent property (Meng and Rubin 1993). Consequently, Algorithm 2 has convergence guarantees, which improves over Amini et al. (2013).

We also note that in our profile-pseudo likelihood approach, while the conditional distribution (on node degrees) of the Poisson variables is multinomial, the multinomial coefficient (i.e., the $\frac{d_i!}{b_{i1}!b_{i2}!\cdots b_{iK}!}$ factorial term) in the density function involves the column labels (in b_{ik}'s). As such, optimizing for the column labels in the outer loop becomes highly challenging. In Algorithm 2, we work with the pseudolikelihood without conditioning on node degrees and it requires estimating the degree parameters in the M-step. This is different from that in Amini et al. (2013).

5. Simulation Studies

In this section, we carry out simulation studies to investigate the finite sample performance of our proposed profile-pseudo likelihood method (referred to as PPL), and to compare with existing solutions including the spectral clustering with permutations (referred to as SCP) and the pseudolikelihood method (referred to as PL) proposed in Amini et al. (2013). Both SCP and PL are implemented using the code provided by Amini et al. (2013). We also compare with the strongly consistent majority voting method proposed in Gao et al. (2017) (see Section A6 in the supplemental material).

We consider two evaluation criteria. The first one is the normalized mutual information (NMI), which measures the distance between the true labeling vector and an estimated labeling vector. The NMI takes values between 0 and 1, and a larger value implies a higher accuracy. The second one is the CPU running time, which measures the computational cost. Note the reported running time does not include the initialization step (see Section A6 in the supplemental material and discussions in Section 7). All methods are implemented in Matlab and run on a single processor of an Intel(R) Core(TM) i7-4790 CPU 3.60 GHz PC.

5.1. SBM

In this section, we simulate networks from SBMs. Three different settings are considered. In Setting 1, we evaluate the convergence of PPL and PL; in Setting 2, we compare the performance of PPL, SCP, and PL when the networks are small and dense; in Setting 3, we compare the three methods when the networks are large and sparse.

Setting 1: In this simulation, we evaluate the convergence performance of PPL and PL with varying initial labeling vectors. We simulate from SBMs with $n = 500$ nodes that are divided into K equal sized communities, and the within/between community connecting probabilities are $P_{kl} = p_1 + p_2 \times 1(k = l), k, l = 1, \ldots, K$. We consider $(K, p_1, p_2) = (2, 0.13, 0.07)$, and $(K, p_1, p_2) = (5, 0.10, 0.13)$. Both the PPL and PL algorithms are considered to have converged if the change of the latest update (relative to the previous one) is less than 10^{-6} or if the number of outer iterations exceeds 60. We let the NMI of the initial labeling vector vary from 0.1 to 0.5. All simulations are repeated 100 times. The proportion of convergence for PPL and PL are presented in Figure 2. It is seen that the PL does not have a satisfactory convergence performance. One example (in the case of $K = 2$) of the convergence of PPL and nonconvergence of PL is shown in Figure 1, where it is observed that the PL algorithm did not converge, and the final estimate has a smaller log pseudolikelihood when compared to the initial value.

Setting 2: In this simulation, we compare the performance of SCP, PL, and PPL on small-scale and dense networks. The PL method is not expected to perform well in this setting due to the relatively large Poisson approximation error. We acknowledge that many networks in real applications are large and/or sparse, and we note that here we use simulated examples to investigate a limitation of the PL method. We simulate from SBMs with n

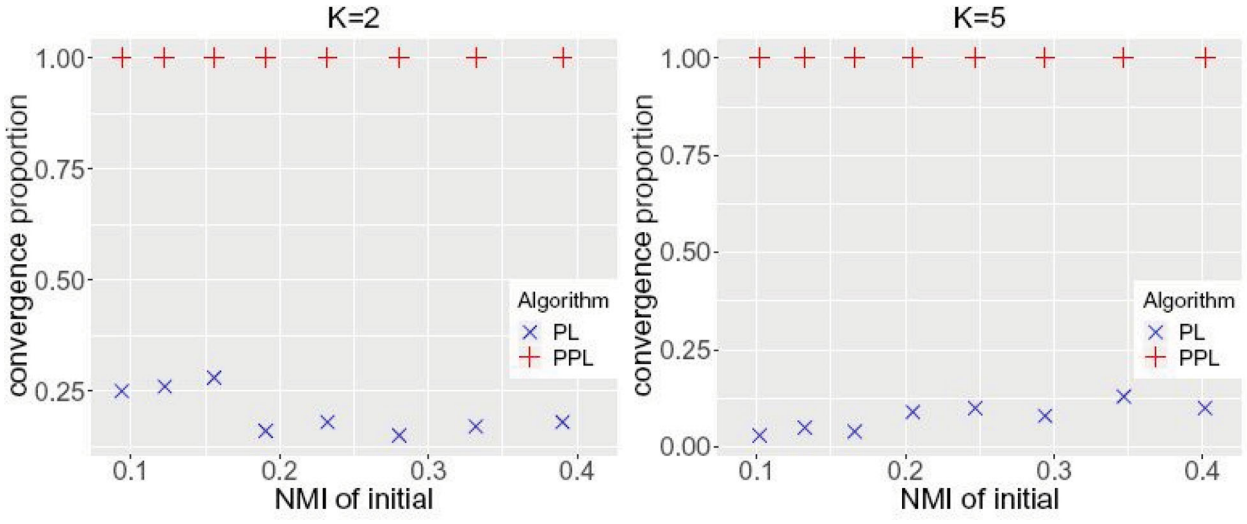

Figure 2. Proportion of convergence of PPL and PL with initial labels of varying NMI.

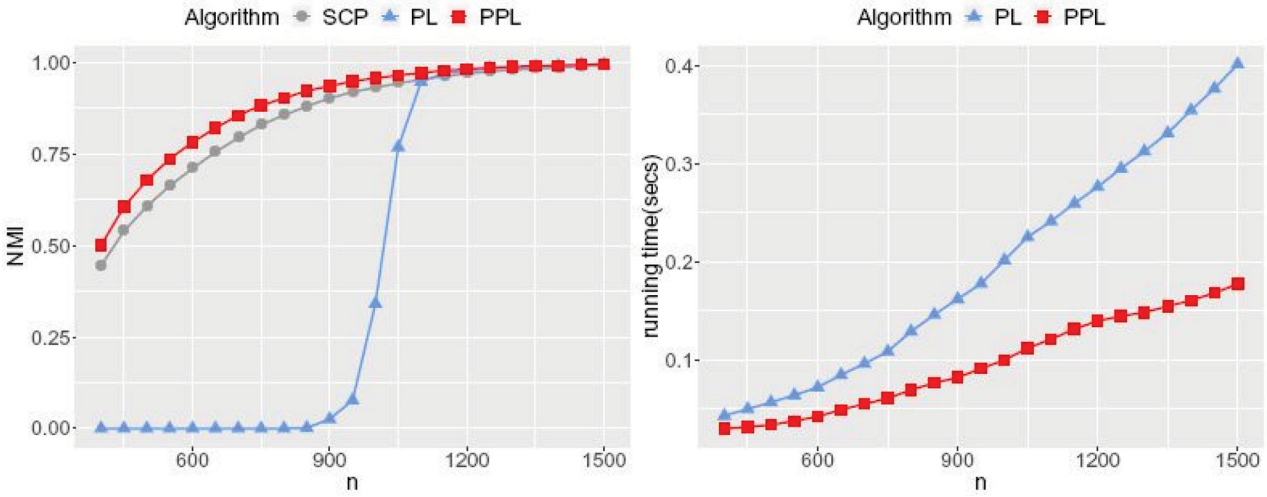

Figure 3. NMI and computing time of PPL and PL with varying network size n.

nodes that are divided into $K = 2$ equal sized communities, and the within/between community connecting probabilities are $P_{kl} = p_1 + p_2 \times 1(k = l), k, l = 1, \ldots, K$. We consider $(p_1, p_2) = (0.84, 0.06)$. Both PPL and PL are initialized by SCP. Figure 3 reports the NMI from the three methods based on 100 replications. It is seen that PPL outperforms the PL in terms of both community detection accuracy (when $n < 1000$) and computational efficiency. The unsatisfactory performance of the PL method when $n < 1000$ is due to the errors from approximating binomial random variables with Poisson random variables. This approximation is not expected to work well when p_1 (or p_2) is large and when n is small (Hodges and Le Cam 1960). Also note that the PL method may perform worse than the initial labels, as its iterations do not enjoy the ascent property. It can also be seen that as n increases, the performance of PL improves notably.

Setting 3: In this simulation, we compare the performance of SCP, PL, and PPL on large-scale and sparse networks. We consider similar simulation settings as in Amini et al. (2013). As in Decelle et al. (2011), the edge-probability matrix $\boldsymbol{P}$ is controlled by the following two parameters: the "out-in-ratio" β, varying from 0 to 0.2, and the weight vector ω, determining the relative degrees within communities. We set $\omega = (1, 1, 1)$. Once $\beta = 0$, $\boldsymbol{P}^*$ is set to be a diagonal matrix diag(ω), while otherwise we set the diagonal elements of $\boldsymbol{P}^*$ to be $\beta^{-1}\omega$ and set all the off-diagonal ones to 1. Then, the overall expected network degree is set to be λ, which varies from 3 to 5. Finally, we rescale $\boldsymbol{P}^*$ to obtain this expected degree, giving the resulting $\boldsymbol{P}$ as follows:

$$\boldsymbol{P} = \frac{\lambda}{(n-1)(\boldsymbol{\pi}^T \boldsymbol{P}^* \boldsymbol{\pi})} \boldsymbol{P}^*, \tag{14}$$

which generates sparse networks, since $P_{kl} = O(1/n)$. In this simulation study, both PL and PPL are initialized by SCP. We let $K = 3$ and $\boldsymbol{\pi} = (0.2, 0.3, 0.5)$. We consider three scenarios: 1) varying β while setting $\lambda = 5$ and $n = 4000$, 2) varying λ while setting $\beta = 0.05$, and $n = 4000$, and 3) varying n while setting $\lambda = 5$ and $\beta = 0.05$. Figure 4 reports the NMI from the three methods and the computing time from PPL and PL, based on 100 replications. We note the reported running times for PPL and PL do not include the initialization step. For comparison, when $\lambda = 5$, $\beta = 0.05$, and $n = 10^6$, the SCP initialization step takes less than 100 sec (see Section A6 in the supplemental material). It is seen that PPL outperforms both SCP and PL in terms of community detection accuracy. Moreover, PPL consistently outperforms PL in terms of computational efficiency.

5.2. Goodness-of-Fit Test and Normality of Plug-in Estimators

To evaluate goodness of fit, we consider the maximum entry-wise deviation-based testing procedure in Hu et al. (2020b). The authors showed that the distribution of the test statistic, denoted by T_n and calculated with a strongly consistent community label, converges to a Gumbel distribution. In this simulation study, we consider a SBM with $K = 3$, $\boldsymbol{\pi} = (0.2, 0.3, 0.5)$, and $P_{kl} = 0.12 + 0.08 \times I(k = l)$, and investigate the distribution of T_n calculated using estimates from PPL and SCP, respectively. The results over 1000 replications are shown in Figure 5. It is seen that the sample null distribution of T_n calculated with PPL is very close to the limiting distribution while that calculated with SCP deviates from the limit considerably. This is due to that T_n in Hu et al. (2020b) is calculated based on maximum entry-wise deviation and as such, the misclassified nodes in SCP, albeit not many, may much inflate the test statistic. With the refinement of PPL, the test statistic is seen to have a sample null distribution close to the theoretical limit, ensuring a well-controlled test size.

To examine normality of plug-in estimators, we consider a SBM with $K = 3$, $\boldsymbol{\pi} = (0.2, 0.3, 0.5)$, $P_{kl} = 0.12 + 0.08 \times I(k = l)$, and $n = 800$. We consider the empirical distribution of $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ calculated using labels produced by PPL and SCP, respectively. The results over 1000 replications are shown in Figure 6. It is seen that the empirical distributions calculated with PPL are very close to the limiting distributions while those calculated with SCP deviate, especially for $\hat{\pi}_1$ and $\hat{\pi}_3$, from the theoretical limits.

5.3. DCSBM

In this section, we evaluate the performance of the profile-pseudo likelihood method under the DCSBM, referred to as DC-PPL. We fix $K = 3$, $n = 1200$, $\boldsymbol{\pi} = (0.2, 0.3, 0.5)$ and let

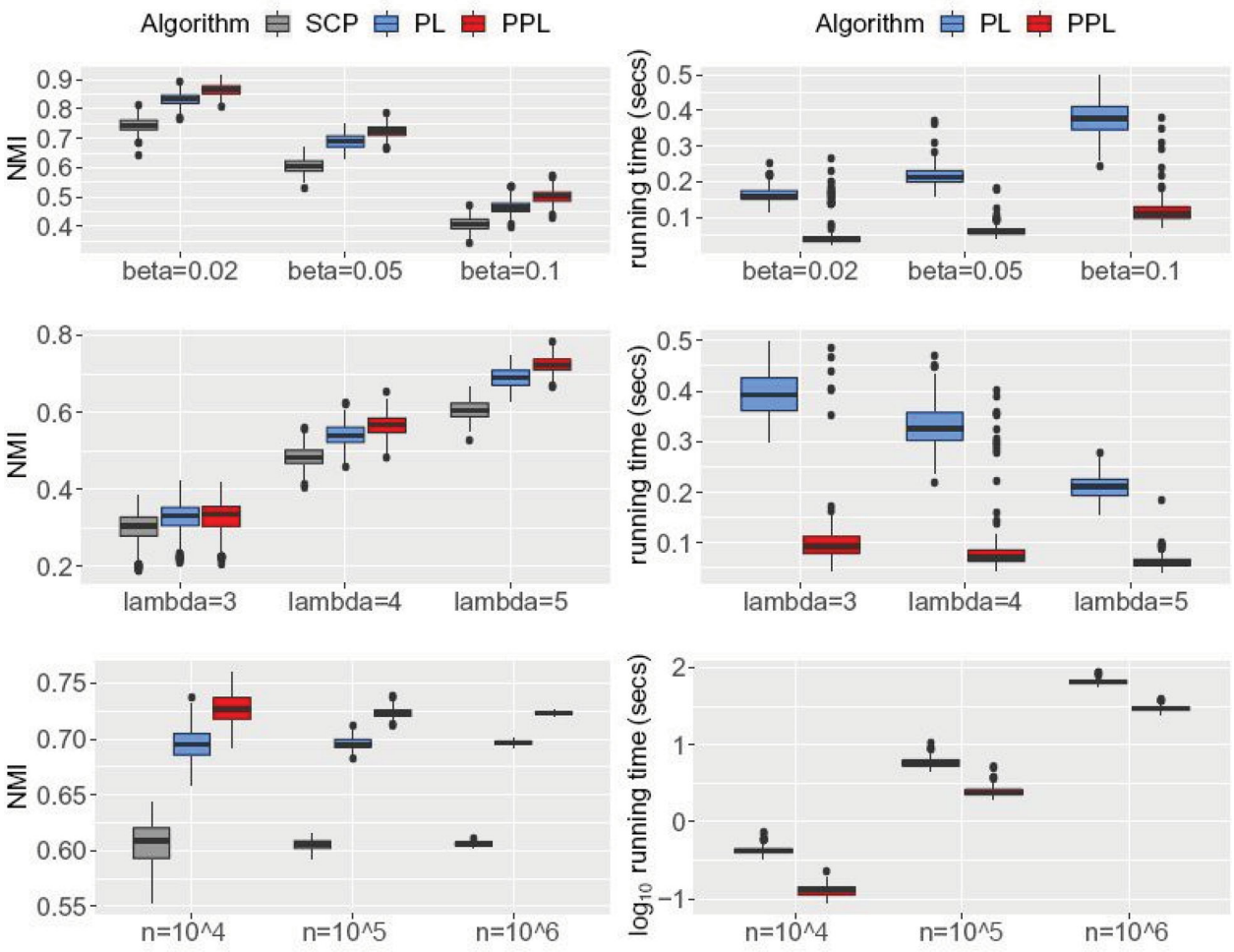

Figure 4. Comparisons of the NMI and computing time from SCP, PL, and PPL under different settings. The three rows correspond to the following three scenarios respectively: (1) varying β while setting $\lambda = 5$ and $n = 4000$, (2) varying λ while setting $\beta = 0.05$, and $n = 4000$, and (3) varying n while setting $\lambda = 5$ and $\beta = 0.05$.

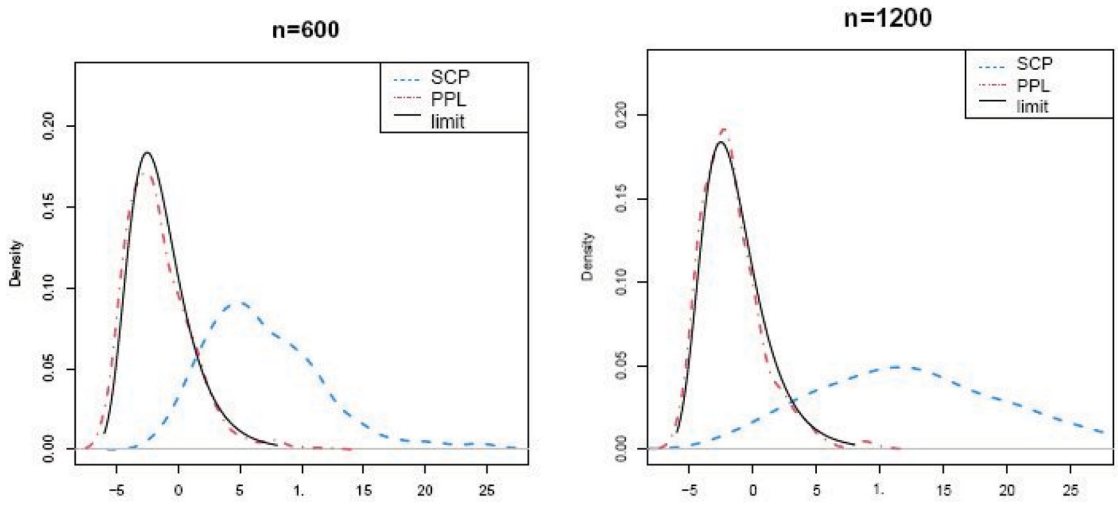

Figure 5. Null densities of the test statistic with $n = 600$ (left plot) and $n = 1200$ (right plot). The blue dashed lines, red dash-dotted lines, and black solid lines show the densities under SCP, PPL, and the theoretical limit, respectively.

$\boldsymbol{P} = 10^{-2} \times \left[J_{K,K} + \text{diag}(2,3,4)\right]$, where $J_{K,K}$ is a K by K matrix where every element is equal to one. The degree parameters $\{\theta_i\}_{i=1}^n$ are generated from (Zhao, Levina, and Zhu 2012), that is,

$$\mathbb{P}(\theta_i = mx) = \mathbb{P}(\theta_i = x) = 1/2 \quad \text{with } x = \frac{2}{m+1},$$

which ensures that $E(\theta_i) = 1$. We consider $m = 2, 4, 6$. Given $\boldsymbol{c}$ and $\boldsymbol{\theta}$, the edge variables A_{ij}'s are independently generated from a Bernoulli distribution with parameters $\theta_i\theta_j P_{c_i c_j}$, $1 \le i \le j \le n$.

We compare DC-PPL with SCP as well as CPL, an extension of PL proposed for networks with degree heterogeneity in Amini et al. (2013). The results are summarized in Figure 7, based on

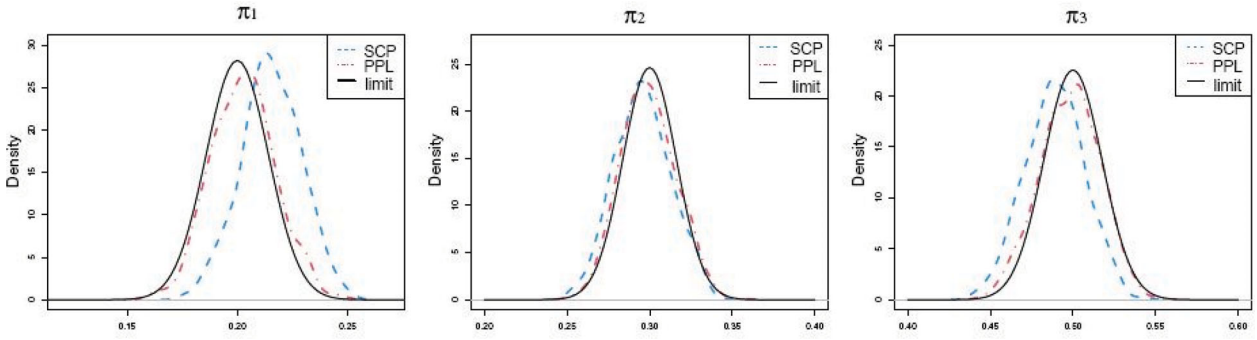

Figure 6. Empirical distributions of $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$. The blue dashed lines, red dash-dotted lines, and black solid lines show the densities under SCP, PPL, and the theoretical limit, respectively.

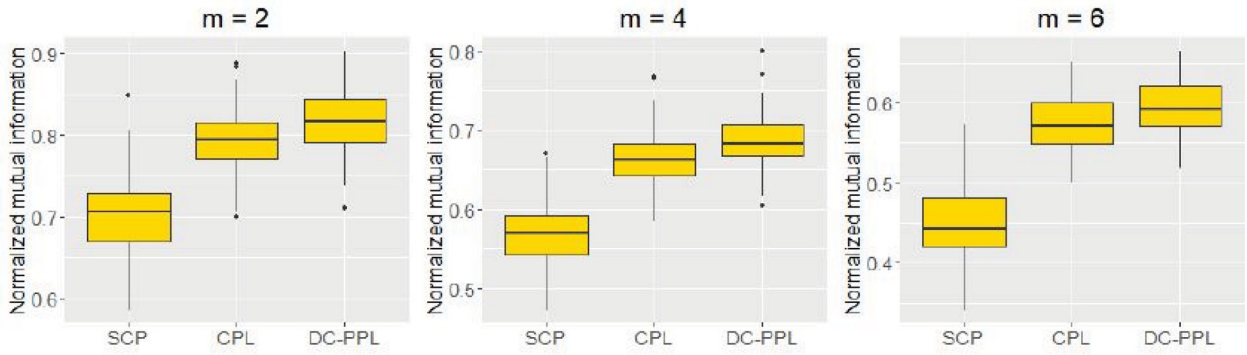

Figure 7. Comparison of SCP, CPL, DC-PPL under DCSBM with varying m.

100 replications. We can see both DC-PPL and CPL outperform SCP, and DC-PPL performs better than CPL in terms community detection accuracy.

6. Real-World Data Examples

6.1. Political Blogs Data

In this subsection, we apply our proposed method to the network of political blogs collected by Adamic and Glance (2005). The nodes in this network are blogs on the U.S. politics and the edges are hyper-links between these blogs with directions removed. This dataset was collected right after the 2004 presidential election and demonstrates strong divisions. In Adamic and Glance (2005), all the blogs were manually labeled as liberal or conservative, and we take these labels as the ground truth. As in Zhao, Levina, and Zhu (2012), we focus on the largest connected component of the original network, which contains 1222 nodes, 16,714 edges and has the average degree of approximately 27.

To perform community detection, we consider five different methods, namely, PL, PPL, SCP, CPL, and DC-PPL. We compute the NMI between the estimated community labels with the so-called ground truth labels. Figure 8 shows the community detection results from the five different methods. It is seen that PPL and PL divide the nodes into two communities, with low degree and high degree nodes, respectively. Both the PPL and PL estimates have NMI close to zero as neither of these two methods take into consideration the degree heterogeneity. The partition obtained using SCP has NMI=0.653, while that from the CPL has NMI=0.722 and that from the DC-PPL has NMI=0.727. Both CPL and DC-PPL achieve good performance in this application.

6.2. International Trade Data

In this subsection, we apply our proposed method to the network of international trades. The data contain yearly international trades among $n = 58$ countries from 1981 to 2000 (Westveld and Hoff 2011). Each node in the network corresponds to a country and an edge (i, j) measures the amount of exports from country i to country j for a given year; see Westveld and Hoff (2011) for details. Following Saldana, Yu, and Feng (2017), we focus on the international trade network in 1995 and transform the directed and weighted adjacency network to an undirected binary network. Specifically, let $W_{ij} = \text{Trade}_{ij} + \text{Trade}_{ji}$, and set $A_{ij} = 1$ if $W_{ij} \geq W_{0.5}$, and $A_{ij} = 0$ otherwise. Here, Trade_{ij} records the amount of exports from country i to country j and $W_{0.5}$ denotes the 50th percentile of $\{W_{ij}\}_{1\leq i<j\leq n}$. Using different model selection procedures, both Saldana, Yu, and Feng (2017) and Hu et al. (2020a) selected the number of SBM communities to be $K = 3$ for this dataset. Saldana, Yu, and Feng (2017) suggested that larger community numbers such as $K = 7$ are also reasonable and they tended to provide finer solutions. We apply PPL to this network with $K = 3$ and the community detection result is summarized in Table 1. It is seen that the three communities mostly correspond to developing countries in South America with low GDPs, countries with high GDPs and industrialized European and Asian countries with medium-level GDPs, respectively.

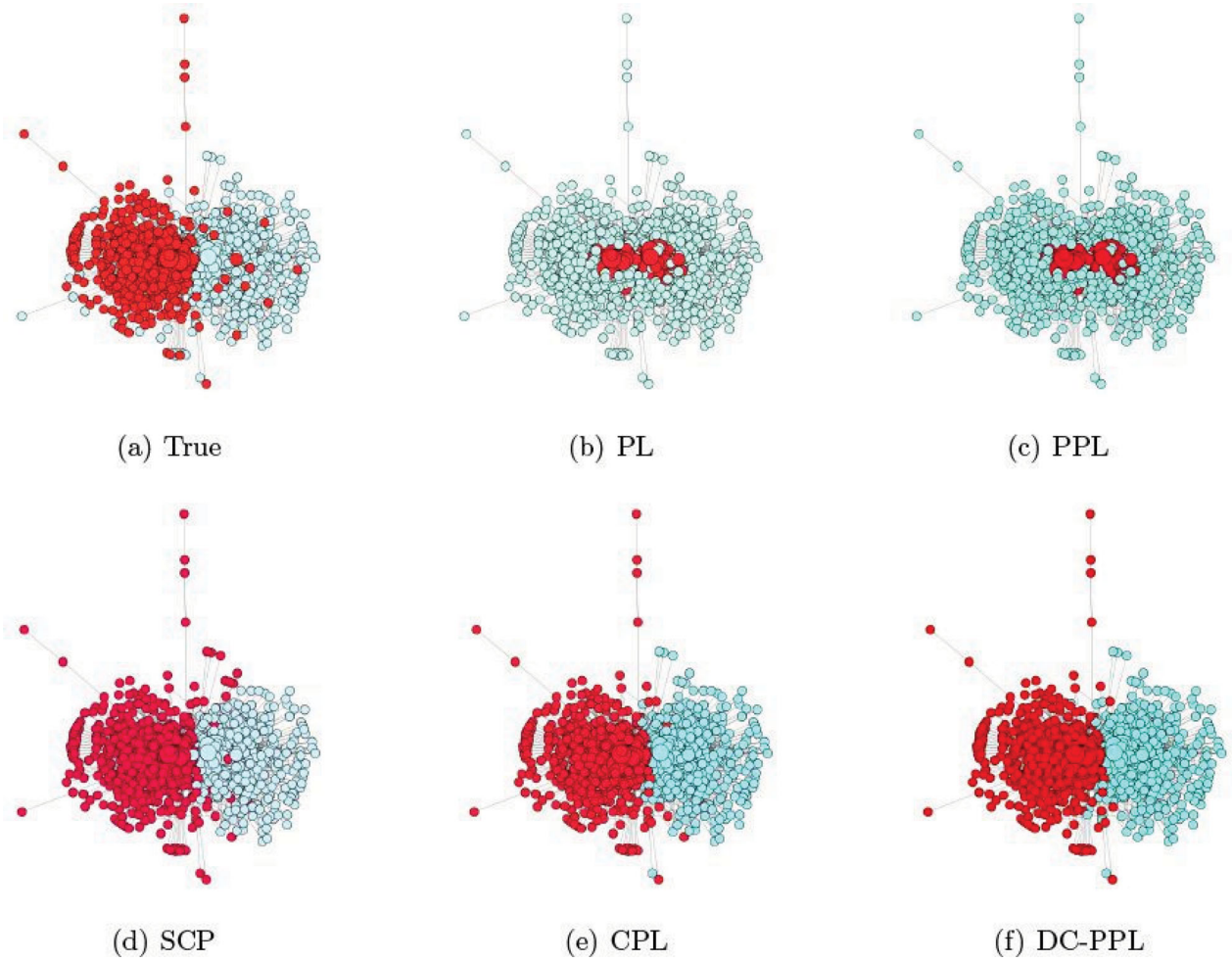

Figure 8. Community detection on the political blogs data using PL, PPL, SCP, CPL, and DC-PPL, respectively. The sizes of nodes are proportional to the their degree, and the color corresponds to different community labels.

Table 1. Community detection result on the international trade data using PPL with $K = 3$.

Group	Countries
1	Algeria, Barbados, Bolivia, Costa Rica, Cyprus, Ecuador, El Salvador, Guatemala, Honduras, Iceland, Jamaica, Mauritius, Nepal, Oman, Panama, Paraguay, Peru, Trinidad and Tobago, Tunisia, Uruguay, Venezuela
2	Belgium, Brazil, Canada, France, Germany, Italy, Japan, South Korea, Mexico, Netherlands, Spain, Switzerland, United Kingdom, United States
3	Argentina, Australia, Austria, Chile, Colombia, Denmark, Egypt, Finland, Greece, India, Indonesia, Ireland, Israel, Malaysia, Morocco, New Zealand, Norway, Philippines, Portugal, Singapore, Sweden, Thailand, Turkey

To evaluate goodness of fit, we consider the maximum entry-wise deviation based testing procedure (Hu et al. 2020b) that we investigated in Section 5.2. The community labels identified using SCP under $K = 3$ gives a test statistic value of 52.13 with a p-value less than 10^{-10}, suggesting a lack of fit. On the other hand, the community labels identified by PPL, initialized using SCP under $K = 3$, gives a test statistic of 4.59 with a p-value of 0.03. Therefore, the goodness-of-fit test for PPL under $K = 3$ is not rejected at the significance level of 0.01. It is also worth noting that when $K = 4$, PPL gives a test statistic of 2.38 with a p-value of 0.08 while SCP gives a p-value less than 10^{-3}. It is seen through this data example that refinement of the initial clustering solution can be useful in inferential tasks such as the goodness-of-fit test.

7. Discussion

In this article, we propose a new profile-pseudo likelihood method for fitting SBMs to large networks. Specifically, we consider a novel approach that decouples the membership labels of the rows and columns in the likelihood function, and treat the row labels as a vector of latent variables. Correspondingly, the likelihood can be maximized in an alternating fashion over the block model parameters and over the column community labels. Our proposed method retains and improves on the computational efficiency of the pseudolikelihood method, performs well for both small- and large-scale networks, and has provable convergence guarantee. We show that the community labels (i.e., column labels) estimated from our proposed method enjoy strong consistency, as long as the initial labels have an overlap with the truth beyond that of random guessing.

In our approach, we consider spectral clustering as the initialization method, which requires computing K leading eigenvectors. In real-world applications, many implementations of eigen-decomposition are scalable, such as the PageRank

algorithm adopted in Google search (Page et al. 1999). We also note that our method needs not to limit the initialization algorithm to spectral clustering. For large-scale networks, one may consider the FastGreedy method by Clauset, Newman, and Moore (2004), which has a complexity of $O(n\log^2 n)$ or the Louvain algorithm by Blondel et al. (2008), which has a complexity of $O(n\log n)$ (Yang, Algesheimer, and Tessone 2016). These fast algorithms, to our best knowledge, may not have theoretical guarantees on their performances. However, they have been validated empirically by many across various fields (Yang, Algesheimer, and Tessone 2016) and can be considered as an initialization method when spectral clustering is not feasible.

Although we focus on SBMs and DCSBMs in this work, we envision the idea of simplifying the block model likelihoods by decoupling the membership labels of rows and columns can be applied to other network block model problems, such as mixed membership SBMs (Airoldi et al. 2008), block models with additional node features (Zhang, Levina, and Zhu 2016) and SBMs with dependent edges (Yuan and Qu 2018). We plan to investigate these directions in our future work.

The code is publicly available at Github (*https://github.com/WangJiangzhou/Fast-Network-Community-Detection-with-Profile-Pseudo-Likelihood-Methods*).

Supplementary Materials

The supplementary materials collect all technical proofs, additional computational details and simulation results.

Funding

Wang, Liu and Guo's research are supported by National Key R&D Program of China (no. 2020YFA0714102), NSFC (grants nos. 11690012 and 12171079), China Postdoctoral Science Foundation 2021M701588, the Special Fund for Key Laboratories of Jilin Province, China (grant no. 20190201285JC). Zhang's research is supported by NSF (grant no. DMS-2015190) and Zhu's research is supported by NSF (grant no. DMS-1821243).

ORCID

Jiangzhou Wang http://orcid.org/0000-0003-3873-9647

References

Abbe, E. (2017), "Community Detection and Stochastic Block Models: Recent Developments," *The Journal of Machine Learning Research*, 18, 6446–6531. [1359]

Abbe, E., Bandeira, A. S., and Hall, G. (2015), "Exact Recovery in the Stochastic Block Model," *IEEE Transactions on Information Theory*, 62, 471–487. [1360]

Abbe, E., Fan, J., Wang, K., Zhong, Y. (2020), "Entrywise Eigenvector Analysis of Random Matrices With Low Expected Rank," *Annals of Statistics*, 48, 1452–1474. [1359]

Adamic, L. A., and Glance, N. (2005), "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," in *International Workshop on Link Discovery*, pp. 36–43. [1369]

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Block Models," *Journal of Machine Learning Research*, 9, 1981–2014. [1359,1371]

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks," *The Annals of Statistics*, 41, 2097–2122. [1359,1360,1362,1363,1366,1367,1368]

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic Normality of Maximum Likelihood and Its Variational Approximation for Stochastic Block Models," *The Annals of Statistics*, 1922–1943. [1359,1360]

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. [1359,1363]

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008), "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008. [1371]

Clauset, A., Newman, M. E., and Moore, C. (2004), "Finding Community Structure in Very Large Networks," *Physical Review E*, 70, 066111. [1371]

Daudin, J.-J., Picard, F., and Robin, S. (2008), "A Mixture Model for Random Graphs," *Statistics and Computing*, 18, 173–183. [1359]

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011), "Asymptotic Analysis of the Stochastic Block Model for Modular Networks and Its Algorithmic Applications," *Physical Review E*, 84, 066106. [1359,1367]

Fortunato, S. (2010), "Community Detection in Graphs," *Physics Reports*, 486, 75–174. [1359]

Fortunato, S., and Hric, D. (2016), "Community Detection in Networks: A User Guide," *Physics Reports*, 659, 1–44. [1359]

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017), "Achieving Optimal Misclassification Proportion in Stochastic Block Models," *Journal of Machine Learning Research*, 18, 1–45. [1359,1360,1366]

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018), "Community Detection in Degree-Corrected Block Models," *Annals of Statistics*, 46, 2153–2185. [1359,1360]

Hodges, J. L., and Le Cam, L. (1960), "The Poisson Approximation to the Poisson Binomial Distribution," *The Annals of Mathematical Statistics*, 31, 737–740. [1367]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Block Models: First Steps," *Social Networks*, 5, 109–137. [1359]

Hu, J., Qin, H., Yan, T., and Zhao, Y. (2020a), "Corrected Bayesian Information Criterion for Stochastic Block Models," *Journal of the American Statistical Association*, 115, 1771–1783. [1369]

Hu, J., Zhang, J., Qin, H., Yan, T., and Zhu, J. (2020b), "Using Maximum Entry-Wise Deviation to Test the Goodness of Fit for Stochastic Block Models," *Journal of the American Statistical Association*, 1–10. [1360,1367,1370]

Joseph, A., Yu, B. et al. (2016), "Impact of Regularization on Spectral Clustering," *Annals of Statistics*, 44, 1765–1791. [1359]

Karrer, B., and Newman, M. E. (2011), "Stochastic Block Models and Community Structure in Networks," *Physical Review E*, 83, 016107. [1359,1365]

Larremore, D. B., Clauset, A., and Jacobs, A. Z. (2014), "Efficiently Inferring Community Structure in Bipartite Networks," *Physical Review E*, 90, 012805. [1360]

Lei, J. (2016), "A Goodness-of-fit Test for Stochastic Block Models," *The Annals of Statistics*, 44, 401–424. [1360]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [1359,1364]

Lei, J., and Zhu, L. (2017), "Generic Sample Splitting for Refined Community Recovery in Degree Corrected Stochastic Block Models," *Statistica Sinica*, 1639–1659. [1360]

Li, T., Levina, E., and Zhu, J. (2020), "Network Cross-Validation by Edge Sampling," *Biometrika*, 107, 257–276. [1363]

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278. [1366]

Moody, J., and White, D. R. (2003), "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups," *American Sociological Review*, 103–127. [1359]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [1359]

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999), "The PageRank Citation Ranking: Bringing Order to the Web." Tech. rep., Stanford InfoLab. [1371]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [1359]

Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How Many Communities Are There?" *Journal of Computational and Graphical Statistics*, 26, 171–181. [1369]

Snijders, T. A., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Block Models for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [1359]

Spirin, V., and Mirny, L. A. (2003), "Protein Complexes and Functional Modules in Molecular Networks," *Proceedings of the National Academy of Sciences*, 100, 12123–12128. [1359]

Su, L., Wang, W., and Zhang, Y. (2019), "Strong Consistency of Spectral Clustering for Stochastic Block Models," *IEEE Transactions on Information Theory*, 66, 324–338. [1359]

Westveld, A. H., and Hoff, P. D. (2011), "A Mixed Effects Model for Longitudinal Relational and Network Data, With Applications to International Trade and Conflict," *The Annals of Applied Statistics*, 5, 843–872. [1369]

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016), "A Comparative Analysis of Community Detection Algorithms on Artificial Networks," *Scientific Reports*, 6, 1–18. [1371]

Yuan, Y., and Qu, A. (2018), "Community Detection with Dependent Connectivity," arXiv:1812.06406. [1371]

Zhang, Y., Levina, E., and Zhu, J. (2016), "Community Detection in Networks With Node Features," *Electronic Journal of Statistics*, 10, 3153–3178. [1371]

Zhao, Y. (2017), "A Survey on Theoretical Advances of Community Detection in Networks," *Wiley Interdisciplinary Reviews: Computational Statistics*, 9, e1403. [1359]

Zhao, Y., Levina, E., and Zhu, J. (2012), "Consistency of Community Detection in Networks Under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [1359,1363,1365,1368,1369]