Received: 7 June 2021 | Revised: 16 September 2021 | Accepted: 27 September 2021

DOI: 10.1002/sta4.428

SPECIAL ISSUE PAPER

WILEY

Community detection with nodal information: Likelihood and its variational approximation

Haolei Weng[1] | **Yang Feng**[2]

[1]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

[2]Department of Biostatistics, New York University, New York City, New York, USA

Correspondence
Yang Feng, Department of Biostatistics, New York University, New York, NY, USA.
Email: yang.feng@nyu.edu

Community detection is one of the fundamental problems in the study of network data. Most existing community detection approaches only consider edge information as inputs, and the output could be suboptimal when nodal information is available. In such cases, it is desirable to leverage nodal information for the improvement of community detection accuracy. Towards this goal, we propose a flexible network model incorporating nodal information and develop likelihood-based inference methods. For the proposed methods, we establish favorable asymptotic properties as well as efficient algorithms for computation. Numerical experiments show the effectiveness of our methods in utilizing nodal information across a variety of simulated and real network data sets.

KEYWORDS
community detection, consistency, maximum likelihood, multilogistic regression, networks, profile likelihood, semidefinite programming, stochastic block model, variational inference

1 | INTRODUCTION

Networked systems are ubiquitous in modern society. Examples include the worldwide web, gene regulatory networks and social networks. Network analysis has attracted a lot of research attention from social science, physics, computer science and mathematical sciences. There have been some interesting findings regarding the network structures, such as small world phenomena and power law degree distributions (Newman, 2003). One of the fundamental problems in network analysis is detecting and characterizing community structure in networks. Communities can be intuitively understood as groups of nodes which are densely connected within groups but sparsely connected between groups.[1] Identifying network communities not only helps better understand structural features of the network but also offers practical benefits. For example, communities in social networks tend to share similar interest, which could provide useful information to build recommendation systems.

Existing community detection methods can be roughly divided into algorithmic and model-based ones (Zhao et al., 2012). Algorithmic methods typically define an objective function such as modularity (Newman, 2006), which measures the goodness of a network partition and design algorithms to search for the solution of the corresponding optimization problem. See Fortunato (2010) for a thorough discussion of various algorithms. Unlike algorithmic approaches, model-based methods first construct statistical models that are assumed to generate the networks under study and then develop statistical inference tools to learn the latent communities. Some popular models include the stochastic block model (SBM) (Holland et al., 1983), degree-corrected SBM (Dasgupta et al., 2004; Karrer & Newman, 2011) and mixed membership SBM (Airoldi et al., 2009).

In recent years, there have been increasingly active researches towards understanding the theoretical performances of community detection methods under different types of models. Regarding the SBM, consistency results have been proved for likelihood-based approaches, including maximum likelihood (Celisse et al., 2012; Choi et al., 2012), profile likelihood (Bickel & Chen, 2009), pseudo likelihood (Amini et al., 2013) and variational inference (Bickel et al., 2013; Celisse et al., 2012). Some of the existing results are generalized to degree-corrected block models

[1]More rarely, one can encounter communities of the opposite meaning in disassortative mixing networks.

(Zhao et al., 2012). Another line of theoretical works focuses on methods of moments. See Joseph and Yu (2013), Jin (2015), Lei and Rinaldo (2014), Qin and Rohe (2013) and Rohe et al. (2011) on theoretical analysis of spectral clustering for detecting communities in block models. Spectral clustering (Zhang et al., 2014) and tensor spectral methods (Anandkumar et al., 2014) have also been used to detect overlapping communities under mixed membership models. Spectral clustering has also been studied for the optimal combination of multi-layer SBMs (Huang et al., 2020). In addition, carefully constructed convex programming has been shown to enjoy provable guarantees for community detection (Amini & Levina, 2014; Cai & Li, 2015; Chen et al., 2012, 2015, Guédon & Vershynin 2016). See also the interesting theoretical works of community detection under the minimax framework (Gao et al., 2015; Zhang & Zhou, 2015). Finally, there exists a different research theme focusing on detectability instead of consistency (Abbe & Sandon, 2015; Decelle et al., 2011; Krzakala et al., 2013; Saade et al., 2014).

All the aforementioned methods are based on only the observations of the edge connections in the networks. In the real world, however, networks often appear with additional nodal information. For example, social networks such as Facebook and Twitter contain users' personal profile information. A citation network has the authors' names, keywords and abstracts of papers. Since nodes in the same communities tend to share similar features, we can expect that nodal attributes are in turn indicative of community structures. Combining both sources of edge and nodal information opens the possibility for more accurate community discovery. Many efficient heuristic algorithms are proposed in recent years to accomplish this goal (Akoglu et al., 2012; Chang & Blei, 2010; Nallapati & Cohen, 2008; Newman & Clauset, 2016; Ruan et al., 2013; Yang et al., 2013). However, not much theory has been established to understand their statistical properties. Zhang et al. (2016) proposed a modularity optimization approach and proved its community detection consistency. Binkiewicz et al. (2017) introduced a covariate-assisted spectral method and derived an upper error bound. Huang and Feng (2018) studied a pair-wise covariate adjusted block model. Yan and Sarkar (2020) developed a convex optimization algorithm and obtained upper error bounds for sparse networks. See also Deshpande et al. (2018) and Stegehuis and Massoulié (2019) for phase transition analysis in some special settings. In this paper, we aim to give a thorough study of the community detection with nodal information problem. Our work first introduces a flexible modelling framework tuned for community detection when edge and nodal information coexist. Under a specific model, we then study three likelihood methods and derive their asymptotic properties. Regarding the computation of the estimators, we resort to a convex relaxation (semidefinite programming, SDP) approach to obtain a preliminary community estimate serving as a good initialization fed into "coordinate" ascent-type iterative algorithms, to help locate the global optima. Various numerical experiments demonstrate that our methods can accurately discover community structures by making efficient use of nodal information.

The rest of the paper is organized as follows. Section 2 introduces our network model with nodal information. We then propose likelihood-based methods and derive the corresponding asymptotic properties in Section 3. Section 4 is devoted to the development of practical algorithms. Simulation examples and real data analysis are presented in Section 5. We conclude the paper with a discussion in Section 6. All the technical proofs are collected in the supporting information.

2 | NETWORK MODELLING WITH NODAL INFORMATION

A network is usually represented by a graph $G(V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of nodes and E is the set of edges. Throughout the paper, we will focus on the networks in which the corresponding graphs are undirected and contain no self-edges. The observed edge information can be recorded in the adjacency matrix $A \in \{0, 1\}^{n \times n}$, where $A_{ij} = A_{ji} = 1$ if and only if $(i, j) \in E$. Suppose the network can be divided into K nonoverlapping communities. Let $\boldsymbol{c} = (c_1, \ldots, c_n)$ be the community assignment vector, with c_i denoting the community membership of node i and taking values in $\{1, 2, \ldots, K\}$. Additionally, the available nodal information is formulated in a covariate matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the i-th node's covariate vector. The goal is to estimate $\boldsymbol{c}$ from the observations A and X.

2.1 | Conditional independence

We treat A, X and $\boldsymbol{c}$ as random and posit a statistical model for them. Before introducing the model, we would like to elucidate the main motivation. For the purpose of community detection, we follow the standard two-step inference procedure: (1) Derive the parameter estimator $\hat{\boldsymbol{\theta}}$ based on $P(A, X; \boldsymbol{\theta})$; (2) Perform posterior inference according to $P(\boldsymbol{c}|A, X; \hat{\boldsymbol{\theta}})$.

Under this framework, we now make a conditional independence assumption: $A \perp X|\boldsymbol{c}$. Admittedly, the assumption imposes a strong constraint that given the community membership, what nodes are like (described by covariates X) does not affect how they are connected (encoded in A). On the other hand, this assumption is consistent with our belief that knowing nodal information can help identify the community structure $\boldsymbol{c}$. More importantly, this assumed conditional independence turns out to simplify the above two steps to a great extent. First, for the parameter estimation step, the conditional independence assumption implies that

$$P(A, X; \boldsymbol{\theta}) = \sum_{\boldsymbol{c}} P(A|\boldsymbol{c}) P(X|\boldsymbol{c}) P(\boldsymbol{c}) = P(X; \boldsymbol{\theta}_1) \sum_{\boldsymbol{c}} P(A|\boldsymbol{c}; \boldsymbol{\theta}_2) P(\boldsymbol{c}|X; \boldsymbol{\theta}_3), \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ indexes a family of generative models (not restricted to parametric forms). Regarding the second step, conditional independence leads to

$$P(\boldsymbol{c}|A,X) = \frac{P(A|\boldsymbol{c})P(X|\boldsymbol{c})P(\boldsymbol{c})}{\sum_{\boldsymbol{c}} P(A|\boldsymbol{c})P(X|\boldsymbol{c})P(\boldsymbol{c})} = \frac{P(A|\boldsymbol{c})P(\boldsymbol{c}|X)P(X)}{\sum_{\boldsymbol{c}} P(A|\boldsymbol{c})P(\boldsymbol{c}|X)P(X)} = \frac{P(A|\boldsymbol{c})P(\boldsymbol{c}|X)}{\sum_{\boldsymbol{c}} P(A|\boldsymbol{c})P(\boldsymbol{c}|X)}. \tag{2}$$

From (1) and (2), we observe that the distribution $P(X)$ is a "nuisance" in the two-step procedure. Hence, we are able to avoid modelling and estimating the marginal distribution of X. As a result, the effort can be saved for the inference of $P(A|\boldsymbol{c})$ and $P(\boldsymbol{c}|X)$.

2.2 | Node-coupled stochastic block model

The conditional independence and follow-up arguments in Section 2.1 pave the way to a flexible framework of models for networks with nodal covariates: specifying the two conditionals $P(A|\boldsymbol{c})$ and $P(\boldsymbol{c}|X)$. A similar modelling strategy was proposed in Newman and Clauset (2015), along with detailed empirical results. Unlike them, we will consider a different model and present a thorough study from both theoretical and computational perspectives. Note that the conditional distribution $P(A|\boldsymbol{c})$ only involves the edge information of the network, while $P(\boldsymbol{c}|X)$ is often encountered in the standard regression setting for i.i.d. data. This motivates us to consider the following model.

Node-coupled Stochastic Block Model (NSBM):

$$P(A|\boldsymbol{c}) = \prod_{i<j} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1-A_{ij}}, \quad P(\boldsymbol{c}|X) = \prod_i \frac{\exp(\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)},$$

where $B = (B_{ab}) \in [0,1]^{K \times K}$ is symmetric, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) \in \mathbb{R}^{Kp}$. The distribution $P(A|\boldsymbol{c})$ follows the SBM, which, as the fundamental model, has been extensively studied in the literature. The SBM implies that the distribution of an edge between nodes i and j only depends on their community membership c_i and c_j. The nodes from the same community are stochastically equivalent. The element B_{ab} in the matrix B represents the probability of edge connection between a node in community a and a node in community b. The $P(\boldsymbol{c}|X)$ simply takes a multilogistic regression form, where we will assume $\boldsymbol{\beta}_K = \boldsymbol{0}$ for identifiability. Simple as it looks, we would like to point out some advantages of NSBM:

- The parameters in NSBM can be estimated by combining the estimation of B and $\boldsymbol{\beta}$, as we shall elaborate in Section 4.
- The coefficient $\boldsymbol{\beta}$ reflects the contribution of each nodal covariate for identifying community structures. This information can help us better understand the implication of the network communities.
- The probability $p(c = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_k^T \boldsymbol{x})}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x})}$ can be used to predict a new node's community membership c based on its covariates $\boldsymbol{x}$, without waiting for it to form network connections.
- Both $P(A|\boldsymbol{c})$ and $P(\boldsymbol{c}|X)$ can be readily generalized to fit more complicated structures.

Remark 1. As illustrated in Section 2.1, under the conditional independence assumption $A \perp X|\boldsymbol{c}$, it is sufficient to consider the conditional likelihood $P(A,\boldsymbol{c}|X)$ instead of the full version $P(A,\boldsymbol{c},X)$. In particular, we study the maximum likelihood estimate, maximum variational likelihood estimate and the maximum profile likelihood estimate based on the conditional likelihood in the next section. However, we emphasize that the conditional independence assumption is not part of NSBM, though it was used to motivate the model. In the next section, we will treat this assumption as working independence to derive the likelihood-based estimates. Hence, the three estimates are in fact based on pseudo likelihood. With a slight abuse of terminology and for simplicity, we still call them the aforementioned likelihood names in the rest of the paper.

3 | STATISTICAL INFERENCE UNDER NSBM

For community detection, our main goal is to find an accurate community assignment estimator $\hat{\boldsymbol{c}}$ for the underlying true communities $\boldsymbol{c}$. Theoretically, we would like to study the consistency of community detection for a given method. We adopt the notions of consistency from Bickel and Chen (2009) and Zhao et al. (2012): as $n \to \infty$,

$$\text{strong consistency}: \quad P(\hat{\boldsymbol{c}} = \boldsymbol{c}) \to 1; \quad \text{weak consistency}: \quad \forall \epsilon > 0, P\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}(\hat{c}_i \neq c_i) < \epsilon\right) \to 1.$$

As the network size increases to infinity, with probability approaching 1, strong consistency requires perfect recovery of the true community structure, while weak consistency only needs the misclassification rate to be arbitrarily small. Note that since community structure is invariant under a permutation of the community labels in {1, 2, … , K}, the consistency notations above as well as the estimators to be introduced should always be interpreted up to label permutations.

In the asymptotic setting where the network size $n \to \infty$, holding the parameter $B \in [0, 1]^{K \times K}$ unchanged implies that the total number of edges present in the network is of order $O(n^2)$. Such networks are unrealistically dense. To study under a more realistic asymptotic framework, we allow B to change with n. In particular, we consider a sequence of submodels where $B = \rho_n \overline{B}$ with $\overline{B}$ fixed and $\rho_n = P(A_{ij} = 1) \to 0$ as $n \to \infty$. The same asymptotic formulation was studied in Bickel and Chen (2009), Bickel et al. (2013) and Zhao et al. (2012). In this way, the parameter ρ_n directly represents the sparsity level of the network. For the consistency results to be derived in the subsequent sections, we will specify the sufficient conditions on the order of ρ_n.

As pointed out in Section 2.2, the parameter $\boldsymbol{\beta}$ in NSBM is associated with the contribution of each nodal covariate for discovering communities. Measuring the importance of each nodal attribute to the community structure may provide insightful information about the network. For that purpose, in addition to community detection, we will study the asymptotics of the estimators for $\boldsymbol{\beta}$ as well. Since the parameter B is not of current interest, we will skip the theoretical analysis of the corresponding estimators.

3.1 | Consistency of maximum likelihood method

In Section 2.1, we pointed out the appealing implication of the assumed conditional independence for the likelihood based inference procedure. We now evaluate this procedure under the asymptotic framework we introduced at the beginning of Section 3. Towards that end, we define the following maximum likelihood based estimators[2]:

$$(\hat{\boldsymbol{\beta}}, \hat{B}) = \underset{\substack{\boldsymbol{\beta}_K = \mathbf{0}, \boldsymbol{\beta} \in \mathbb{R}^{Kp} \\ B \in [0,1]^{K \times K}, B^T = B}}{\operatorname{argmax}} \sum_{\boldsymbol{c}} \prod_{i<j} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1 - A_{ij}} \cdot \prod_i \frac{\exp(\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)}, \tag{3}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c} \in \{1, \ldots, K\}^n}{\operatorname{argmax}} \prod_{i<j} \hat{B}_{c_i c_j}^{A_{ij}} (1 - \hat{B}_{c_i c_j})^{1 - A_{ij}} \cdot \prod_i \frac{\exp(\hat{\boldsymbol{\beta}}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i)}. \tag{4}$$

The estimators defined above are the realizations of the two-step procedure we mentioned at the beginning of Section 2.1. We are mainly interested in studying the consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{c}}$.

First, we would like to introduce several technical conditions.

Condition 1. $\overline{B}$ has no two identical columns.

If the probability matrix $\overline{B}$ has two identical columns, then there exist at least two communities unidentifiable with each other. In practice, it makes sense to combine those communities into a bigger one.

Condition 2. $(c_1, \boldsymbol{x}_1), \ldots, (c_n, \boldsymbol{x}_n) \overset{iid}{\sim} (c, \boldsymbol{x})$ with $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) \succ 0$, where $\succ 0$ represents the matrix being positive definite.

Condition 2 ensures that the coefficient vector $\boldsymbol{\beta}$ is uniquely identifiable.

Condition 3. There exist constants κ_1 and κ_2 such that for sufficiently large t, we have $P(\|\boldsymbol{x}\|_2 > t) \le \kappa_1 e^{-\kappa_2 t}$.

Condition 3 imposes a subexponential tail bound on $\|\boldsymbol{x}\|_2$, which is equivalent to a subexponential tail assumption on each component of $\boldsymbol{x}$, via a simple union bound argument. This covers many different types of covariates like discrete, Gaussian and exponential.

Theorem 1. *Assume the data* (A, X) *follows NSBM and Conditions 1,2 and 3 hold. In addition, assume* $\frac{n\rho_n}{\log n} \to \infty$ *as* $n \to \infty$. *Then, we have as* $n \to \infty$, $P(\hat{\boldsymbol{c}} = \boldsymbol{c}) \to 1$ and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta}))$, *where* $I(\boldsymbol{\beta})$ *is the Fisher information for the multilogistic regression problem of regressing* $\boldsymbol{c}$ *on* X.

[2]Recall that the likelihood formulation is the pseudo version as pointed out in Remark 1. Similar explanations hold for the other two likelihood based methods presented in the subsequent sections.

The key condition $\frac{n\rho_n}{\log n} \to \infty$ requires that the expected degree of every node to grow faster than the order of log$n$. The same condition has been used in Bickel and Chen (2009) and Zhao et al. (2012) to derive strong consistency under SBM. Under the conditions of the theorem, the maximum likelihood method gives us not only a strong consistent community assignment estimate $\hat{\boldsymbol{c}}$ but also a coefficient estimate $\hat{\boldsymbol{\beta}}$ which is as efficient as if the true label $\boldsymbol{c}$ were known.

3.2 | Consistency of variational method

The maximum likelihood method studied in Section 3.1 has been shown to have nice theoretical properties. However, the likelihood function form in (3) renders the computation of the estimators $(\hat{\boldsymbol{\beta}}, \hat{B})$ intractable. In particular, it is computationally infeasible to even evaluate the likelihood function value at a nondegenerate point (when n is not too small), due to the marginalization over all possible membership assignments. To address this computation issue, we propose a tractable variational method and demonstrate that it enjoys equally favorable asymptotic properties as the maximum likelihood approach. This is motivated by the works about variational methods under SBMs (Bickel et al., 2013; Celisse et al., 2012; Daudin et al., 2008). Throughout this section, we will use the generic symbol $P(\cdot)$ to denote joint distributions and $\boldsymbol{\theta} = (\boldsymbol{\beta}, B)$. To begin with, recall the well known identity: $\log P(A,X;\boldsymbol{\theta}) = \mathbb{E}_Q[\log P(A,X,\boldsymbol{c};\boldsymbol{\theta}) - \log Q(\boldsymbol{c})] + D[Q(\boldsymbol{c})||P(\boldsymbol{c}|A,X;\boldsymbol{\theta})]$, where $Q(\cdot)$ denotes any joint distribution of $\boldsymbol{c}$; the expectation $\mathbb{E}_Q(\cdot)$ is taken with respect to $\boldsymbol{c}$ under $Q(\boldsymbol{c})$; $D[\cdot\,||\cdot]$ is the Kullback–Leibler divergence. Since $D[Q(\boldsymbol{c})||P(\boldsymbol{c}|A,X;\boldsymbol{\theta})] \geq 0$ and the equality holds when $Q(\boldsymbol{c}) = P(\boldsymbol{c}|A,X;\boldsymbol{\theta})$, it is not hard to verify the following variational equality,

$$\max_{\boldsymbol{\theta}} \log P(A,X;\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}, Q(\cdot)} \mathbb{E}_Q[\log P(A,X,\boldsymbol{c};\boldsymbol{\theta}) - \log Q(\boldsymbol{c})]. \tag{5}$$

Hence, to compute the maximum likelihood value, we can equivalently solve the optimization problem on the right-hand side of (5). Note that iteratively optimizing over $\boldsymbol{\theta}$ and $Q(\cdot)$ leads to the EM algorithm (Dempster et al., 1977). However, the calculation of $P(\boldsymbol{c}|A,X;\boldsymbol{\theta})$ at each iteration of EM is computationally intensive. Instead of optimizing over the full distribution space of $Q(\cdot)$, variational methods aim to solve an approximate optimization problem by searching over a subset of all possible $Q(\cdot)$. In particular, we consider the mean-field variational approach (Jordan et al., 1999),

$$\max_{\boldsymbol{\theta}, Q \in \mathcal{Q}} \mathbb{E}_Q[\log P(A,X,\boldsymbol{c};\boldsymbol{\theta}) - \log Q(\boldsymbol{c})], \tag{6}$$

where $\mathcal{Q} = \{Q : Q(\boldsymbol{c}) = \prod_{i=1}^n q_{ic_i}, \sum_k q_{ik} = 1, 1 \leq i \leq n\}$. The subset $\mathcal{Q}$ contains all the distributions under which the elements of $\boldsymbol{c}$ are mutually independent. The independence structure turns out to make the computation in (6) manageable. We postpone the detailed calculations to Section 4, and focus on the asymptotic analysis in this section. Denote the maximizer in (6) by $(\check{\boldsymbol{\beta}}, \check{B})$ and

$$\check{\boldsymbol{c}} = \underset{\boldsymbol{c} \in \{1,\ldots,K\}^n}{\operatorname{argmax}} \prod_{i<j} \check{B}_{c_i c_j}^{A_{ij}} (1 - \check{B}_{c_i c_j})^{1-A_{ij}} \cdot \prod_i \frac{\exp(\check{\boldsymbol{\beta}}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\check{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i)}. \tag{7}$$

Theorem 2. *Suppose the conditions in Theorem 1 hold. Then as* $n \to \infty, P(\check{\boldsymbol{c}} = \boldsymbol{c}) \to 1$ *and* $\sqrt{n}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta}))$, *where* $I(\boldsymbol{\beta})$ *is the Fisher information for the multi-logistic regression problem of regressing* $\boldsymbol{c}$ *on* X.

As we can see, under the same conditions as the maximum likelihood method, the variational approach can deliver equally good estimators, at least in the asymptotic sense. In other words, the approximation made by the variational method does not degrade the asymptotic performance. This should be attributed to the condition $\frac{n\rho_n}{\log n} \to \infty$, which guarantees the network has sufficient edge information for doing approximate inference.

3.3 | Consistency of maximum profile likelihood method

The two methods presented in Sections 3.1 and 3.2 are implementations of the two-step procedure we discussed in Section 2.1: first estimating the parameters based on the likelihood function and then doing posterior inference using the estimated distribution. In this section, we introduce a one-step method that outputs the parameter and community assignment estimates simultaneously. The method solves the following problem,

$$(\tilde{\boldsymbol{\beta}},\tilde{B},\tilde{\boldsymbol{c}}) = \arg\ \max_{\substack{\boldsymbol{\beta}_K=\mathbf{0},\boldsymbol{\beta}\in\mathbb{R}^{Kp} \\ \boldsymbol{c}\in\{1,\dots,K\}^n \\ B\in[0,1]^{K\times K},B^T=B}} \prod_{i<j} B_{c_i c_j}^{A_{ij}}(1-B_{c_i c_j})^{1-A_{ij}} \cdot \prod_i \frac{\exp(\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)}. \tag{8}$$

In the above formulation, we treat the latent variables $\boldsymbol{c}$ as parameters and obtain the estimators as the maximizer of the joint likelihood function. This enables us to avoid the cumbersome marginalization encountered in the maximum likelihood method. This approach is known as the maximum profile likelihood (Bickel & Chen, 2009; Zhao et al., 2012). Bickel and Chen (2009) showed strong consistency under the SBM, and Zhao et al. (2012) generalized the results to degree-corrected block models. Following similar ideas, we will investigate this method in the NSBM. For theoretical convenience, we consider a slightly different formulation:

$$(\tilde{\boldsymbol{\beta}},\tilde{B},\tilde{\boldsymbol{c}}) = \arg\ \max_{\substack{\boldsymbol{\beta}_K=\mathbf{0},\boldsymbol{\beta}\in\mathbb{R}^{Kp} \\ \boldsymbol{c}\in\{1,\dots,K\}^n \\ B\in[0,1]^{K\times K},B^T=B}} \prod_{i<j} e^{-B_{c_i c_j}} B_{c_i c_j}^{A_{ij}} \cdot \prod_i \frac{\exp(\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)}, \tag{9}$$

where the Bernoulli distribution in (8) is replaced by the Poisson distribution. In our asymptotic setting $\rho_n \to 0$, the difference becomes negligible.

Theorem 3. *Assume the data* (A, X) *follow the NSBM and Conditions 1 and 2 hold.*

(i) *If* $n\rho_n \to \infty$ *and* $\mathbb{E}\|\boldsymbol{x}\|_2^\alpha < \infty (\alpha > 1)$, *then there exists a constant* $\gamma > 0$ *such that, as* $n \to \infty$

$$P\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}(\tilde{c}_i \neq c_i) \le \gamma (n\rho_n)^{-1/2}\right) \to 1,\ \|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 = O_p((n\rho_n)^{\frac{1-\alpha}{2\alpha}}).$$

(ii) *Assume Condition 3 is satisfied. If* $\frac{n\rho_n}{\log n} \to \infty$, *then as* $n \to \infty, P(\tilde{\boldsymbol{c}} = \boldsymbol{c}) \to 1$ and $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta}))$, *where* $I(\boldsymbol{\beta})$ *is the Fisher information for the multilogistic regression problem of regressing* $\boldsymbol{c}$ *on* X.

We see that part (ii) in Theorem 3 is identical to Theorems 1 and 2. Hence, the maximum profile likelihood method is equivalently good as the previous two in certain sense. The conclusions in part (i) shed lights on how the network edges and nodal covariates affect the consistency results. Under the scaling $n\rho_n \to \infty, \tilde{\boldsymbol{c}}$ is only weak consistent. And the higher moment $\|\boldsymbol{x}\|_2$ has the faster convergence rate $\check{\boldsymbol{\beta}}$ can achieve. Suppose all moments of $\|\boldsymbol{x}\|_2$ exist, then we would have $\sqrt{n\rho_n}\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 = O_p(1)$. Since $\rho_n \to 0$, this convergence rate is slower than and may be arbitrarily close to the one in part (ii) when $\frac{n\rho_n}{\log n} \to \infty$.

4 | PRACTICAL ALGORITHMS

In Section 3, we have studied three likelihood based community detection methods and shown their superb asymptotic performances. In this section, we design specialized algorithms for computing the variational estimators defined by (6), (7) and the maximum profile likelihood estimators in (9). As discussed in Section 3.2, the maximum likelihood estimators are computationally infeasible, hence omitted here. The key challenge lies on the fact that the likelihood-based functions in (6), (7) and (9) are all nonconvex. Multiple local optima may exist, and the global solution is often impossible to allocate accurately. To address this issue, we first obtain a "well-behaved" preliminary estimator via convex optimization and then feed it as an initialization into "coordinate" ascent-type iterative schemes. The idea is that the carefully chosen initialization may help the followed-up iterations to escape "bad" local optima and arrive "closer" (better approximation) to the ideal global solution. As we shall see in the numerical studies, the results with a "well-behaved" initial estimator are significantly better than those with a random initialization. These two steps will be discussed in detail in Sections 4.1 and 4.2, respectively.

4.1 | Initialization via convex optimization

The convex optimization we consider in this section is SDP. Different formulations of SDP have been shown to yield good community detection performances in Amini and Levina (2014), Chen et al. (2012), Cai and Li (2015), Guédon and Vershynin (2016), Montanari and Sen (2015), among others. One illuminating interpretation of SDP is to think of it as a convex relaxation of the maximum likelihood method. For example, starting from a specialized SBM, one can derive SDP by approximating the corresponding likelihood function. See Amini and

Levina (2014), Chen et al. (2012) and Cai and Li (2015) for the detailed arguments. However, under the NSBM, because of the nodal covariate term, it is not straightforward to generalize the convex relaxation arguments. We hence resort to a different understanding of SDP elaborated in Guédon and Vershynin (2016). The key idea is to construct SDP based on the observations directly, with the goal of having the true community assignment $\boldsymbol{c}$ to be the solution of a "population" version of the SDP under construction. In particular, we consider the following SDP problem,

$$\begin{aligned} &\hat{Z} = \arg\max_Z \ \langle A + \gamma_n XX^T, Z\rangle \\ &\text{subject to} \quad Z \succeq 0, Z \in \mathbb{R}^{n\times n}, 0 \le Z_{ij} \le 1, 1 \le i,j \le n, \sum_{ij} Z_{ij} = \lambda_n \end{aligned} \tag{10}$$

where $\gamma_n, \lambda_n > 0$ are two tuning parameters; $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices. We then obtain the communities by running K-means on $\hat{Z}$ (treating each row of $\hat{Z}$ as a data point in $\mathbb{R}^n$). Under some mild conditions on $\overline{B}$, we can show that this approach produces a consistent community assignment estimate, as $n\rho_n \to \infty$. But since we propose the SDP method mainly for the algorithmic purpose, we do not detail out the asymptotic analysis to avoid potential digression. Instead, we present some insights to justify the design of SDP in (10). The argument follows closely Guédon and Vershynin (2016). Recall $\boldsymbol{c} = (c_1, \ldots, c_n)$ is the underlying true community membership vector. Denote

$$\mathcal{M}_{\boldsymbol{c}} = \left\{ Z \in \mathbb{R}^{n\times n} : Z \succeq 0, 0 \le Z_{ij} \le 1, \sum_{ij} Z_{ij} = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \mathbb{1}(c_i = k) \right)^2 \right\},$$

and $S(Z) = \langle \mathbb{E}(A|\boldsymbol{c}) + \gamma_n \mathbb{E}(X|\boldsymbol{c})\mathbb{E}(X^T|\boldsymbol{c}), Z\rangle$. We can construct a "population" version of (10) as follows:

$$\arg\max_{Z \in \mathcal{M}_{\boldsymbol{c}}} S(Z). \tag{11}$$

Define $Z(\boldsymbol{c}) = M(\boldsymbol{c})M^T(\boldsymbol{c})$ where $M(\boldsymbol{c}) \in \mathbb{R}^{n\times K}$ with $M_{ik}(\boldsymbol{c})$ equals 1 if $c_i = k$ and 0 otherwise for $1 \le i \le n, 1 \le k \le K$. Hence, the matrix $Z(\boldsymbol{c}) \in \mathbb{R}^{n\times n}$ encodes the true community structure. We show below that $Z(\boldsymbol{c})$ is in fact the unique solution of (11). This is because for any $Z \in \mathcal{M}_{\boldsymbol{c}}$,

$$\begin{aligned} & S(Z(\boldsymbol{c})) - S(Z) = \langle \mathbb{E}(A|\boldsymbol{c}) + \gamma_n \mathbb{E}(X|\boldsymbol{c})\mathbb{E}(X^T|\boldsymbol{c}), Z(\boldsymbol{c}) - Z\rangle \\ = & \sum_{ij} (\rho_n \overline{B}_{c_i c_j} + \gamma_n \mathbb{E}(\boldsymbol{x}_i^T|c_i)\mathbb{E}(\boldsymbol{x}_j|c_j)) \cdot (Z_{ij}(\boldsymbol{c}) - Z_{ij}) \cdot \mathbb{1}(c_i = c_j) - \sum_{ij} (\rho_n \overline{B}_{c_i c_j} + \gamma_n \mathbb{E}(\boldsymbol{x}_i^T|c_i)\mathbb{E}(\boldsymbol{x}_j|c_j)) \cdot (Z_{ij} - Z_{ij}(\boldsymbol{c})) \cdot \mathbb{1}(c_i \ne c_j) \\ \overset{(a)}{\ge} & U \cdot \sum_{ij} (Z_{ij}(\boldsymbol{c}) - Z_{ij}) \cdot \mathbb{1}(c_i = c_j) - L \cdot \sum_{ij} (Z_{ij} - Z_{ij}(\boldsymbol{c})) \cdot \mathbb{1}(c_i \ne c_j) \overset{(b)}{\ge} \frac{U - L}{2} \cdot \|Z(\boldsymbol{c}) - Z\|_1, \end{aligned} \tag{12}$$

where

$$U = \min_{1 \le k \le K} \{\rho_n \overline{B}_{kk} + \gamma_n \mathbb{E}(\boldsymbol{x}^T|c = k)\mathbb{E}(\boldsymbol{x}|c = k)\} \text{ and } L = \max_{1 \le a \ne b \le K} \{\rho_n \overline{B}_{ab} + \gamma_n \mathbb{E}(\boldsymbol{x}^T|c = a)\mathbb{E}(\boldsymbol{x}|c = b)\}.$$

In (*a*) we have used the fact that $Z_{ij}(\boldsymbol{c}) \ge Z_{ij}$ if $c_i = c_j$ and $Z_{ij}(\boldsymbol{c}) \le Z_{ij}$ otherwise, and (*b*) holds because $Z, Z(\boldsymbol{c}) \in \mathcal{M}_{\boldsymbol{c}}$ leads to

$$\sum_{ij} (Z_{ij}(\boldsymbol{c}) - Z_{ij}) \cdot \mathbb{1}(c_i = c_j) = \sum_{ij} (Z_{ij} - Z_{ij}(\boldsymbol{c})) \cdot \mathbb{1}(c_i \ne c_j) = \frac{1}{2} \|Z(\boldsymbol{c}) - Z\|_1.$$

We may interpret the gap $U - L$ as the signal strength of the community structure. As long as the gap is positive, we see from (12) that the true community structure $Z(\boldsymbol{c})$ can be recovered by the "population" version SDP (11). As a result, if the "sample" version SDP (10) is close to the "population" one (it holds for large samples under mild conditions), we can expect $\hat{Z}$, the solution of (10), to be close to the truth $Z(\boldsymbol{c})$ as well.

We note that there are two tuning parameters in (10). The tuning γ_n trades off the information from two different sources: network edges and nodal covariates. The way we incorporate nodal covariates has the same spirit as Binkiewicz et al. (2017) does in spectral clustering. From the simulation studies in the next section, we shall see that a flexible choice of γ_n can lead to satisfactory results. Regarding the parameter λ_n, from the preceding theoretical justifications, we observe that $\lambda_n = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \mathbb{1}(c_i = k) \right)^2$ is a desirable choice, which depends on the unknown truth in a seemingly restrictive way. However, we will demonstrate through simulations that the community detection results are quite robust to the choice of λ_n.

The convex optimization problem (10) can be readily solved by standard SDP solvers such as SDPT3 (Tütüncü et al., 2003). However, those solvers are based on interior-point methods and computationally expensive when the network size n is more than a few hundred. To overcome this limit, we apply the alternating direction method of multipliers (ADMM) to develop a more scalable algorithm for solving (10). We start by a brief description of the generic ADMM algorithm with the details available in the excellent tutorial by Boyd et al. (2011). In general, ADMM solves problems in the form

$$\text{minimize } f(\boldsymbol{y})+h(\boldsymbol{z}) \quad \text{subject to } B\boldsymbol{y}+D\boldsymbol{z}=\boldsymbol{w}, \tag{13}$$

where $\boldsymbol{y}\in\mathbb{R}^m, \boldsymbol{z}\in\mathbb{R}^n, B\in\mathbb{R}^{q\times m}, D\in\mathbb{R}^{q\times n}, \boldsymbol{w}\in\mathbb{R}^q$ and $f(\boldsymbol{y}), h(\boldsymbol{z})$ are two convex functions. The algorithm takes the following iterations at step t.

$$\begin{aligned}\boldsymbol{y}^{t+1} &= \arg\min_{\boldsymbol{y}}\Big(f(\boldsymbol{y})+(\xi/2)\|B\boldsymbol{y}+D\boldsymbol{z}^t-\boldsymbol{w}+\boldsymbol{u}^t\|_2^2\Big),\\ \boldsymbol{z}^{t+1} &= \arg\min_{\boldsymbol{z}}\Big(h(\boldsymbol{z})+(\xi/2)\|B\boldsymbol{y}^{t+1}+D\boldsymbol{z}-\boldsymbol{w}+\boldsymbol{u}^t\|_2^2\Big), \quad \boldsymbol{u}^{t+1}=\boldsymbol{u}^t+B\boldsymbol{y}^{t+1}+D\boldsymbol{z}^{t+1}-\boldsymbol{w}\end{aligned}$$

with $\xi > 0$ being a step size constant.

To use this framework, we reformulate (10) as follows:

$$\begin{aligned}&\text{minimize} \quad I(Z\succeq 0)+I(0\le Y_{ij}\le 1, 1\le i,j\le n)+I\Big(\sum_{ij}W_{ij}=\lambda_n\Big)-\langle A+\gamma_n XX^T, Z\rangle\\ &\text{subject to} \quad Y=Z, Y=W,\end{aligned}$$

where $Z,Y,W\in\mathbb{R}^{n\times n}$; $I(Z\succeq 0)$ equals 0 if $Z\succeq 0$ and $+\infty$ otherwise; similar definitions hold for other $I(\cdot)$. If we set $\boldsymbol{y}=(\text{vec}(Y),\text{vec}(Y))^T\in\mathbb{R}^{2n^2}, \boldsymbol{z}=(\text{vec}(W),\text{vec}(Z))^T\in\mathbb{R}^{2n^2}, B=-D=I_{2n^2}\in\mathbb{R}^{2n^2\times 2n^2}, \boldsymbol{w}=\boldsymbol{0}\in\mathbb{R}^{2n^2}$, where vec$(\cdot)$ denotes the vectorized version of a matrix, then the problem above becomes an instance of (13). The corresponding iterations have the following expressions:

$$\begin{aligned}Y^{t+1} &= \operatorname*{argmin}_{0\le Y_{ij}\le 1}\Big(\|Y-W^t+U^t\|_F^2+\|Y-Z^t+V^t\|_F^2\Big), \quad W^{t+1}=\operatorname*{argmin}_{\sum_{ij}W_{ij}=\lambda_n}\|Y^{t+1}-W+U^t\|_F^2,\\ Z^{t+1} &= \operatorname*{argmin}_{Z\succeq 0}\Big(-\langle A+\gamma_n XX^T, Z\rangle+(\xi/2)\|Y^{t+1}-Z+V^t\|_F^2\Big), \quad U^{t+1}=U^t+Y^{t+1}-W^{t+1}, V^{t+1}=V^t+Y^{t+1}-Z^{t+1},\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. It is not hard to see that each iteration above has a closed form update with the details summarized in Algorithm 1.[3]

Algorithm 1 Solving (10) via ADMM

Input: initialize $Z^0=A+\gamma_n XX^T, W^0=Y^0=U^0=V^0=0$, number of iterations $\mathcal{T}$, step size ξ.

For $t=0,\ldots,\mathcal{T}-1$

(a) $Y^{t+1}=\min\{\max\{0,\frac{1}{2}(W^t+Z^t-U^t-V^t)\},1\}$. (b) $W^{t+1}=Y^{t+1}+U^t+n^{-2}[\lambda_n-\sum_{ij}(Y_{ij}^{t+1}+U_{ij}^t)]\mathbf{1}\mathbf{1}^T$.

(c) $Z^{t+1}=P\Lambda_+P^T$, where $Y^{t+1}+V^t+\xi^{-1}\cdot(A+\gamma_n XX^T)=P\Lambda P^T$. (d) $U^{t+1}=U^t+Y^{t+1}-W^{t+1}, V^{t+1}=V^t+Y^{t+1}-Z^{t+1}$.

Output Z^T.

4.2 | Coordinate ascent scheme

As we may see, the problem formulations in (6) and (9) are not suitable for gradient or Hessian-based iterative algorithms, because they either involve discrete variables or have nontrivial constraints. The variables involved in those optimization problems can be divided into community assignment related and others. Naturally, we will adopt the iterative scheme that alternates between these two types of variables.

[3] In Step (c), $P\Lambda P^T$ denotes the spectral decomposition; Λ_+ represents the truncated (keep positive elements) version of Λ.

4.2.1 | Computing variational estimates

To compute the variational estimates in (6), we follow the EM style iterative fashion by maximizing the objective function in (6) with respect to $\boldsymbol{\theta}$ and $Q \in \mathcal{Q}$ alternatively. Specifically, we are solving

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pK}, \boldsymbol{\beta}_K = \mathbf{0}}{\arg\min} \sum_i \left[\left(\sum_k q_{ik}^t \boldsymbol{\beta}_k \right)^T \boldsymbol{x}_i - \log\left(\sum_k e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i} \right) \right], \tag{14}$$

$$B^{t+1} = \underset{B \in \mathbb{R}^{K \times K}, B^T = B}{\arg\min} \sum_{ab} \left[\log B_{ab} \cdot \sum_{i<j} A_{ij} q_{ia}^t q_{jb}^t + \log(1 - B_{ab}) \cdot \sum_{i<j} (1 - A_{ij}) q_{ia}^t q_{jb}^t \right], \tag{15}$$

$$\{q_{ik}^{t+1}\} = \underset{\{q_{ik}\}}{\arg\min} \sum_{ab} \left[\log B_{ab}^{t+1} \cdot \sum_{i<j} A_{ij} q_{ia} q_{jb} + \log(1 - B_{ab}^{t+1}) \cdot \sum_{i<j} (1 - A_{ij}) q_{ia} q_{jb} \right] + \sum_i \sum_k q_{ik} (\boldsymbol{\beta}_k^{t+1})^T \boldsymbol{x}_i - \sum_i \sum_k q_{ik} \log q_{ik}. \tag{16}$$

Note that the objective function in (14) takes a similar form as the log-likelihood function of the multilogistic regression model. We hence use the Newton–Raphson algorithm in the same way as we fit the multilogistic regression model to compute the update in (14). This corresponds to Step (a) of Algorithm 2, in which we have used the name "FitMultiLogistic" there to denote the full step with a bit abuse of notation. In addition, the update in (15) has an explicit solution, which corresponds to Step (b) in Algorithm 2. Regarding the update for $\{q_{ik}\}_{ik}$ in (16), unfortunately, the optimization is nonconvex and does not have analytical solutions. We then implement an inner blockwise coordinate ascent loop to solve it. In particular, we update $\{q_{ik}\}_{k=1}^K$ one at a time:

$$\{q_{ik}\}_k = \underset{\{q_{ik}\}_k}{\arg\min} \sum_k q_{ik} \cdot \left[\sum_b \sum_{j \neq i} \left(A_{ij} q_{jb} \cdot \log B_{kb} + (1 - A_{ij}) q_{jb} \cdot \log(1 - B_{kb}) \right) \right] + \sum_k q_{ik} \boldsymbol{\beta}_k^T \boldsymbol{x}_i - \sum_k q_{ik} \log q_{ik}.$$

It is straightforward to show that the update above has closed forms:

$$q_{ik} = \frac{e^{a_k}}{\sum_{k=1}^K e^{a_k}}, \; a_k = \boldsymbol{\beta}_k^T \boldsymbol{x}_i + \sum_b \sum_{j \neq i} q_{jb} \cdot \left(A_{ij} \log B_{kb} + (1 - A_{ij}) \log(1 - B_{kb}) \right).$$

This yields Step (c) for Algorithm 2. After computing $\{q_{ik}^{\mathcal{T}}\}$, $\boldsymbol{\beta}^{\mathcal{T}}$, $B^{\mathcal{T}}$ via Algorithm 2, we calculate the community assignment estimate $\check{\boldsymbol{c}}$ based on (7). This could be done by coordinate ascent iterations, like Step (c) in Algorithm 3 (to be introduced in Section 4.2.2). Alternatively, we can use the following approximated posterior distribution $\{q_{ik}^{\mathcal{T}}\}$: $\check{c}_i = \underset{1 \le k \le K}{\arg\min}\, q_{ik}^{\mathcal{T}}, 1 \le i \le n$, which is used in our numerical studies since it is computationally more efficient.

Algorithm 2 Solving (6) via iterating between $(\boldsymbol{\beta}, B)$ and Q.

Input: initialize $\{q_{ik}^0\}$, number of iterations $\mathcal{T}$

For $t = 0, \ldots, \mathcal{T} - 1$

(a) $\boldsymbol{\beta}^{t+1} = \text{FitMultiLogistic}(X, \{q_{ik}^t\})$

(b) $B_{ab}^{t+1} = \frac{\sum_{i<j} A_{ij} q_{ia}^t q_{jb}^t}{\sum_{i<j} q_{ia}^t q_{jb}^t}$

(c) Update $\{q_{ik}^{t+1}\}$ via Repeating

For $i = 1, \ldots, n$, $\quad \log q_{ik} \propto (\boldsymbol{\beta}_k^{t+1})^T \boldsymbol{x}_i + \sum_b \sum_{j \neq i} q_{jb} \cdot \left(A_{ij} \log B_{kb}^{t+1} + (1 - A_{ij}) \log(1 - B_{kb}^{t+1}) \right)$

Output $\{q_{ik}^{\mathcal{T}}\}, \boldsymbol{\beta}^{\mathcal{T}}, B^{\mathcal{T}}$.

4.2.2 | Computing maximum profile likelihood estimates

Similar to the variational estimates, we maximize the likelihood function in (9) with respect to $(\boldsymbol{\beta}, B)$ and $\boldsymbol{c}$ iteratively. In other words, we solve

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pK}, \boldsymbol{\beta}_K = \mathbf{0}}{\arg\min} \sum_i \left[\boldsymbol{\beta}_{c_i^t}^T \mathbf{x}_i - \log\left(\sum_k e^{\boldsymbol{\beta}_k^T \mathbf{x}_i} \right) \right], \tag{17}$$

$$B_{ab}^{t+1} = \underset{B_{ab}}{\arg\min}\ \log B_{ab} \cdot \sum_{i<j} A_{ij} \mathbb{1}(c_i^t = a, c_j^t = b) - B_{ab} \cdot \sum_{i<j} \mathbb{1}(c_i^t = a, c_j^t = b), \tag{18}$$

$$\boldsymbol{c}^{t+1} = \underset{\boldsymbol{c} \in \{1,\ldots,K\}^n}{\arg\min} \sum_{ab} \left[\log B_{ab}^{t+1} \cdot \sum_{i<j} A_{ij} \mathbb{1}(c_i = a, c_j = b) - B_{ab}^{t+1} \cdot \sum_{i<j} \mathbb{1}(c_i = a, c_j = b) \right] + \sum_i (\boldsymbol{\beta}_{c_i}^{t+1})^T \mathbf{x}_i. \tag{19}$$

Here, solving (17) is equivalent to computing the maximum likelihood estimate of multilogistic regression. This is carried out in Step (a) of Algorithm 3. In addition, it is straightforward to see that Step (b) in Algorithm 3 is the solution to (18). For computing $\boldsymbol{c}^{t+1}$ in (19), we update its element one by one, as shown by Step (c) in Algorithm 3.

Algorithm 3 Solving (9) via iterating between $(\boldsymbol{\beta}, B)$ and $\boldsymbol{c}$.

Input: initialize $\{c_i^0\}$, number of iterations $\mathcal{T}$

For $t = 0, \ldots, \mathcal{T} - 1$

(a) $\boldsymbol{\beta}^{t+1} = \text{FitMultiLogistic}(X, \boldsymbol{c}^t)$

(b) $B_{ab}^{t+1} = \frac{\sum_{i<j} A_{ij} \mathbb{1}(c_i^t = a, c_j^t = b)}{\sum_{i<j} \mathbb{1}(c_i^t = a, c_j^t = b)}$

(c) Update $\{c_i^{t+1}\}$ via Repeating

For $i = 1, \ldots, n$, $\quad c_i = \underset{1 \le k \le K}{\text{argmin}} (\boldsymbol{\beta}_k^{t+1})^T \mathbf{x}_i + \sum_b \sum_{j \neq i} \mathbb{1}(c_j = b) \cdot (A_{ij} \log B_{kb}^{t+1} - B_{kb}^{t+1})$

Output $\{c_i^T\}, \boldsymbol{\beta}^T$.

4.2.3 | Variational estimates versus maximum profile likelihood estimates

So far we have studied the theoretical properties of the variational and maximum profile likelihood estimates and developed algorithms to compute them. The results in Sections 3.2 and 3.3 demonstrate that they have the same asymptotic performance under $\frac{n\rho_n}{\log n} \to \infty$. We now compare the corresponding algorithms. By taking a close look at Algorithms 2 and 3, we observe that the three steps in the two algorithms share a lot of similarities. Algorithm 2 is essentially a "soft" version of Algorithm 3 in the following sense: Instead of using the community assignment c_i in Algorithm 3, the steps in Algorithm 2 involve the probability of belonging to every possible community. This might remind us of the comparison between the EM algorithm and K-means under Gaussian mixture models. As we will see in Section 5, variational and maximum profile likelihood methods usually lead to similar numerical results.

5 | NUMERICAL EXPERIMENTS

In this section, we conduct a detailed experimental study of the SDP defined in (10), variational and maximum profile likelihood methods on both simulated and real datasets. We use two quantitative measures for evaluating their community detection performance: *Normalized Mutual Information* (NMI) (Ana & Jain, 2003) and *Adjusted Rand Index* (ARI) (Hubert & Arabie, 1985).

$$\text{NMI} = \frac{-2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij} \cdot n}{n_{i\cdot} n_{\cdot j}}\right)}{\sum_i n_{i\cdot} \log\left(\frac{n_{i\cdot}}{n}\right) + \sum_j n_{\cdot j} \log\left(\frac{n_{\cdot j}}{n}\right)}, \quad \text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\sum_i \binom{n_{i\cdot}}{2} + \frac{1}{2}\sum_j \binom{n_{\cdot j}}{2} - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}}.$$

In the above expressions, $n_{i\cdot}$ denotes the true number of nodes in community i, $n_{\cdot j}$ represents the number of nodes in the estimated community j and n_{ij} is the number of nodes belonging to community i but estimated to be in community j. Both NMI and ARI are bounded by 1, with the value

of 1 indicating perfect recovery while 0 implying the estimation is no better than random guess. See Steinhaeuser and Chawla (2010) for a detailed discussion.

5.1 | Simulation studies

We set $K=2, \rho_n=\frac{3[\log(n)]^{1.5}}{4n}, P(c=1)=P(c=2)=0.5, \overline{B}=\begin{pmatrix}1.6 & 0.4\\ 0.4 & 1.6\end{pmatrix}$. We consider the following two different scenarios.

(A). $p=4, \boldsymbol{x}|c=1\sim N(\boldsymbol{u}, I_4), \boldsymbol{x}|c=2\sim N(-\boldsymbol{u}, I_4), \boldsymbol{u}=(0,0.4,0.6,0.8)^T$, where $I_4\in\mathbb{R}^{4\times 4}$ is the identity matrix.

(B). $p=4, (x_1,x_2)|c=1\sim N(\boldsymbol{u},\Sigma), (x_1,x_2)|c=2\sim N(-\boldsymbol{u},\Sigma), \boldsymbol{u}=(0.5,0.5)^T, \Sigma_{11}=\Sigma_{22}=1, \Sigma_{12}=0.3, x_3|c=1\sim$ Bernoulli$(0.6), x_3|c=2\sim$ Bernoulli$(0.4), x_4|c=1\sim$ Uniform$(-0.2,0.5), x_4|c=2\sim$ Uniform$(-0.5,0.2)$; and $(x_1,x_2), x_3$ and x_4 are mutually independent.

Note that in Scenario (A), the NSBM is the correct model and the first nodal variable is independent of the community assignment; In Scenario (B), the NSBM is no longer correct. Under both correct and misspecified models, we would like to (i) investigate the impacts of the two tuning parameters (γ_n, λ_n) in the SDP (10), (ii) examine the effectiveness of the SDP as initialization and (iii) check the performances of the variational and maximum profile likelihood methods for utilizing nodal information.

5.1.1 | Tuning parameters in SDP

For both simulation settings, we solve SDP defined in (10) with different tuning parameters via Algorithm 1, with the number of iterations $\mathcal{T}=100$ and the step size $\xi=1$. We then calculate the NMI of its community detection estimates. Since the ARI gives similar results, we do not show them here for simplicity. The full procedure is repeated 500 times.

Figure 1 demonstrates the joint impact of the tuning parameters on the SDP performance under Scenario (A). First of all, the comparison of NMI between $\gamma_n=0$ and $\gamma_n>0$ indicates the effectiveness of SDP (10) for leveraging nodal information. We can also see that neither small or large values of γ_n lead to optimal performances, verifying the point we discussed in Section 4.1 that γ_n plays the role of balancing the edge and nodal information. An appropriate choice, as suggested by the four plots, is $\gamma_n=\frac{[\log(n)]^{0.5}}{n}$. Regarding the parameter λ_n, we know from Section 4.1 that $\lambda_n=\frac{n^2}{2}$ is the desired choice. Interestingly, Figure 1 shows that a wide range of λ_n can give competitive results, as long as the corresponding γ_n is properly chosen. For Scenario (B), similar phenomena can be observed in Figure 2. Note that since the nodal covariates are not as informative as in Scenario (A), the optimal $\gamma_n\approx\frac{0.8[\log(n)]^{0.5}}{n}$ tends to give more weights to the adjacency matrix. The results in these two different settings imply that SDP (10) can work beyond the NSBM.

5.1.2 | Community detection performance via variational and maximum profile likelihood methods

We implement the variational and maximum profile likelihood methods via Algorithms 2 and 3. respectively, taking the outputs from Algorithm 1 as initialization (called VEM-C and MPL-C, respectively). We do not predefine the number of iterations $\mathcal{T}$ in both algorithms and instead keep iterating until convergence. To investigate the impact of SDP as an initialization, we have additionally implemented both methods with random initialization (called VEM-B and MPL-B, respectively): run Algorithms 2 and 3 with random initialization independently multiple times and choose the outputs that give the largest objective function value (e.g., the profile likelihood function). We have also applied both methods to the simulated datasets with nodal attributes removed (called VEM-A and MPL-A, respectively). This will be used as a comparison to check the effect of the two methods in incorporating nodal information. We set $\lambda_n=\frac{1}{2}n^2, \gamma_n=\frac{[\log(n)]^{0.5}}{n}$ for all the implementations of SDP under Scenario (A); and $\lambda_n=\frac{1}{2}n^2, \gamma_n=\frac{0.8[\log(n)]^{0.5}}{n}$ under Scenario (B). Finally, to shed more light on the impact of initialization on the two likelihood-based methods, we have further included the assortative covariate-assisted spectral clustering from Binkiewicz et al. (2017) (called SPEC) and both methods using it as initialization (called VEM-D and MPL-D, respectively).

Figure 3 shows the community detection results of both methods under Scenario (A). By comparing the curves in each plot, we can make a list of interesting observations: (1) SDP is a good initialization (MPL-C vs. MPL-B, VEM-C vs. VEM-B), (2) SDP itself already gives reasonable outputs, but the follow-up iterations further improve the performance (SDP vs. MPL-C, SDP vs. VEM-C), (3) the nodal covariates are helpful for detecting communities, and the two methods have made effective use of it (MPL-A vs. MPL-C, VEM-A vs. VEM-C), (4) the two methods have similar performances when initialized with SDP (MPL-A vs. VEM-A, MPL-C vs. VEM-C); and (5) the two methods may not have a stringent requirement on the quality of initialization, since initialization by the spectral clustering and SDP give competitive results (MPL-C vs. MPL-D, VEM-C vs. VEM-D) while SDP itself outperformed spectral clustering by a significant margin (SDP vs. SPEC). Moreover, we would like to point out the different behavior of the two methods with random initialization. The comparison between the two purple curves (MPL-B vs. VEM-B) implies that

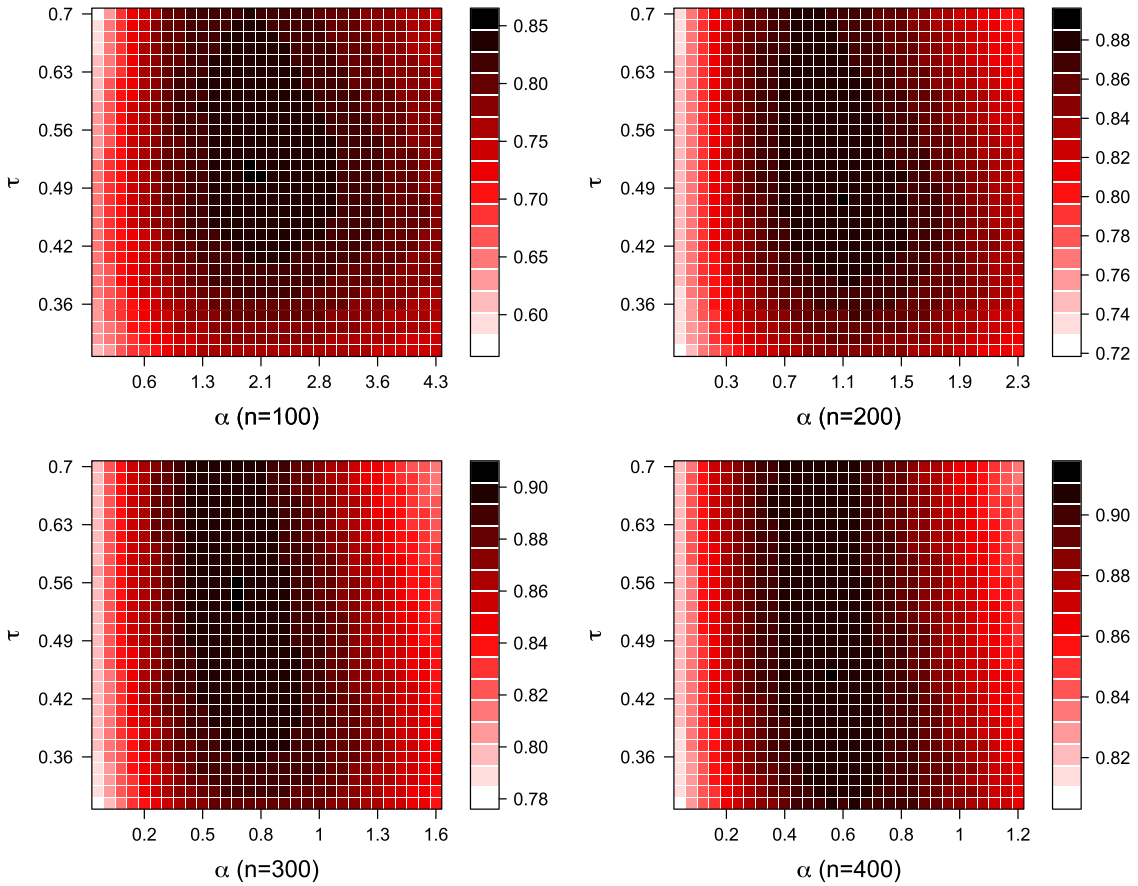

FIGURE 1 The community detection performance of semidefinite programming (SDP) (measured by Normalized Mutual Information [NMI]), under Scenario (A), with different tuning parameters (λ_n, γ_n); NMI is averaged over 500 repetitions; we have used the scaled version of the tuning parameters: $\tau = \frac{\lambda_n}{n^2}, \alpha = 100\gamma_n$

compared with the variational method, the maximum profile likelihood method has the potential of exploring the parameter space more efficiently, especially when the sample size is large. One possible explanation is that the update of the "soft" community labels (the distribution $\{q_{ik}\}$) in the variational algorithm may cause it to move very slowly in the parameter space and hence it may take many steps to change a label assignment. Furthermore, note that we can use the asymptotic normality property of the estimators for $\boldsymbol{\beta}$ in Theorems 2 and 3 to perform variable selection. The results of the Wald test regarding each component of $\boldsymbol{\beta}$ are presented in Figure 4. We see that our methods are able to identify relevant (the last three) and irrelevant (the first) nodal variables. Regarding Scenario (B), similar observations on the community detection performance can be made from Figure 5. We thus omit the details. As a final remark, the performances in Scenario (B) indicate that both methods can work to a certain extent of model misspecification.

5.2 | Real data analysis

The dataset is about a research team consisting of 77 employees in a manufacturing company (Cross & Parker, 2004). The data is openly available at https://opsahl.co.uk/tnet/datasets/Cross_Parker-Manufacturing_info.txt. The edges among the researchers are differentiated in terms of advice ("Please indicate the extent to which the people listed below provide you with information you use to accomplish your work"). The weight of an edge is based on the following scale: 0 (I do not know this person/I have never met this person); 1 (Very infrequently); 2 (Infrequently); 3 (Somewhat infrequently); 4 (Somewhat frequently); 5 (Frequently); 6 (Very frequently). A weight w_{ij} is assigned to the directed edge from employee i to employee j, according to the level of advice (measured on the preceding scale) that employee i provides to employee j for accomplishing employee j's work. In addition to the edge information, the dataset also contains several attributes of each employee: location (1: Paris, 2: Frankfurt, 3: Warsaw, 4: Geneva); tenure (1: 1-12 months, 2: 13-36 months, 3: 37-60 months, 4: 61+ months); the organizational level (1: Global Dept Manager, 2: Local Dept Manager, 3: Project Leader, 4: Researcher). Since the network is a weighted and directed network, we first convert it to a binary network such that there exists an edge from i to j if and only if $w_{ij} > 3$. This corresponds to whether the information is

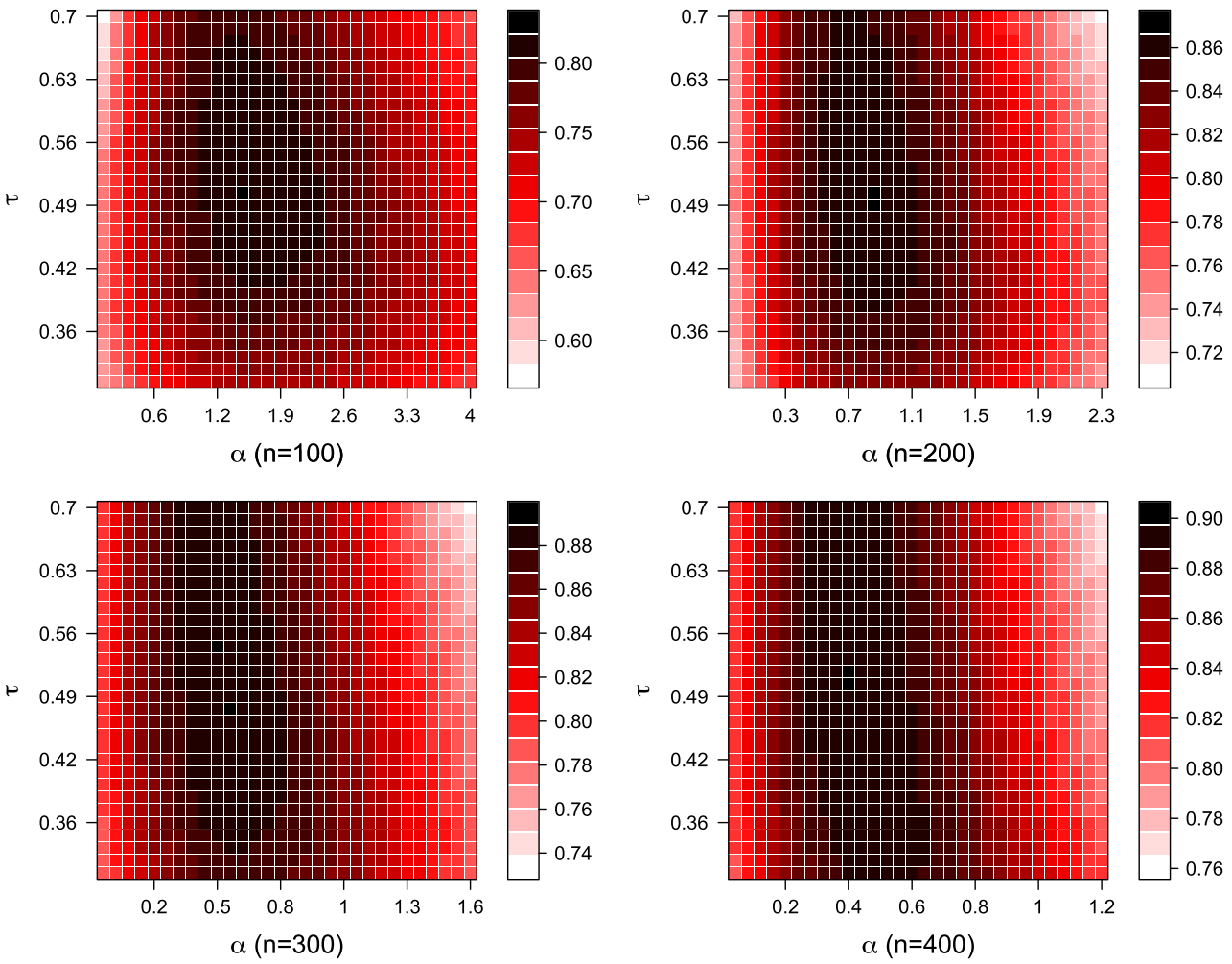

FIGURE 2 The community detection performance of semidefinite programming (SDP) (measured by Normalized Mutual Information [NMI]), under Scenario (B), with different tuning parameters (λ_n, γ_n); Normalized Mutual Information [NMI] is averaged over 500 repetitions; we have used the scaled version of the tuning parameters: $\tau = \frac{\lambda_n}{n^2}, \alpha = 100\gamma_n$

provided frequently or not. We then further convert it into an undirected network in the way that the edge between *i* and *j* exists if and only if both directed edges from *i* to *j* and *j* to *i* are present. Finally, we remove three isolated nodes from the network. To explore the intro-organizational community structure, we re-order the adjacency matrix based on random permutation and the attributes. As can be seen from Figure 6, the attribute "location" is a very informative indicator of the network's community structure. This should not come as a big surprise, since the same office location usually promotes communication and collaboration between team members. We now use the "location" as the ground truth for the community assignment and examine the performances of SDP (10), maximum profile likelihood and variational methods based on the rest of the data. For SDP (10), we first use spectral clustering on the adjacency matrix (Lei & Rinaldo, 2014) to estimate the size of the communities and plug the estimates in the formula $\lambda_n = \sum_{k=1}^{K}\left(\sum_{i=1}^{n}\mathbb{1}(c_i = k)\right)^2$. Regarding γ_n, motivated from the simulation results, we choose $\gamma_n = \frac{\hat{\rho}_n}{\log n}$, where $\hat{\rho}_n = \frac{2\times\text{number of edges}}{n^2}$. The maximum profile likelihood and variational methods are initialized by the output from SDP. We can see from Table 1 that, by incorporating the nodal information, community detection accuracy has been improved. It is interesting to observe that SDP performs as well as the two likelihood based methods, when nodal covariates are available. Note that we can calculate the mutual information between the "ground truth" variable "location" and the other two to see how much community information they contain. Given that both mutual information (0.11 & 0.03) are pretty small, the magnitude of improvement in Table 1 is reasonable.

6 | DISCUSSION

In this paper, we present a systematic study of the community detection with the nodal information problem. We propose a flexible network modelling framework, and analyze three likelihood based methods under a specialized model. Both asymptotic and algorithmic aspects have been

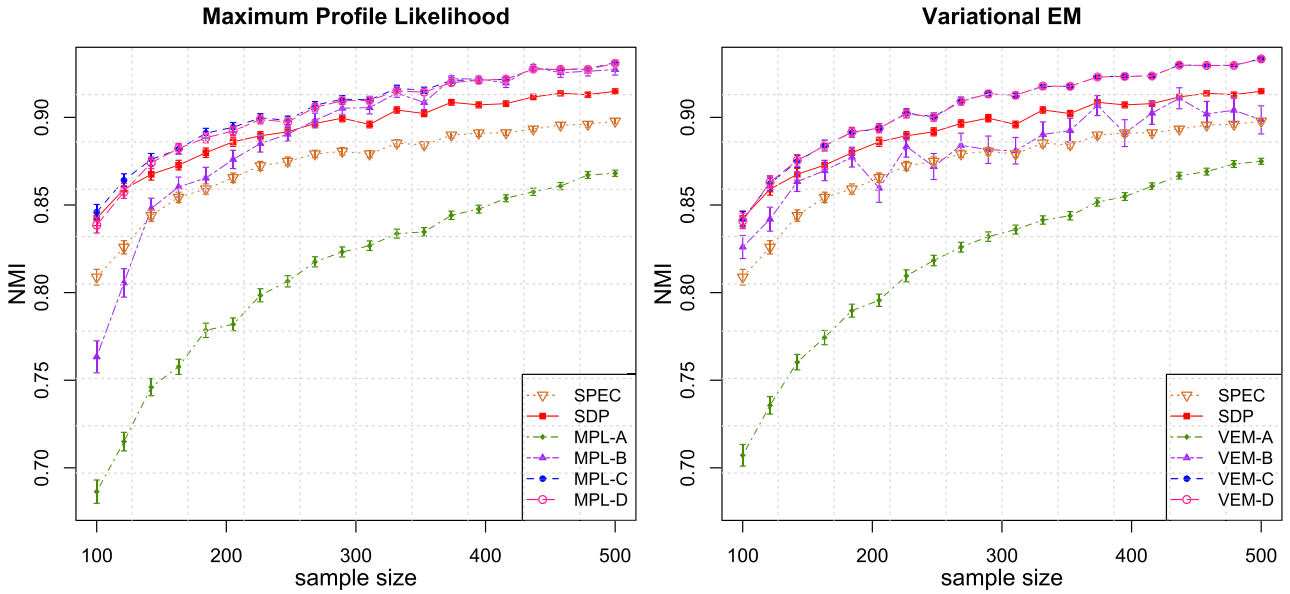

FIGURE 3 The community detection performances under Scenario (A); the average Normalized Mutual Information (NMI) is calculated over 500 repetitions along with its standard error bar; semidefinite programming (SDP) denotes the proposed semidefinite programming approach, and SPEC denotes the assortative covariate-assisted spectral clustering; MPL-A, MPL-B, MPL-C and MPL-D represent the maximum profile likelihood methods with no nodal covariates being used, random initialization, initialization from SDP and initialization from the assortative covariate-assisted spectral clustering, respectively; similar notations are used for the variational method. We have used 15 independent random initializations for the maximum profile likelihood method across all the sample sizes; for the variational method, the number of random initializations used starts from 15 for $n = 100$ and consecutively increases by 1 for the subsequent sample sizes

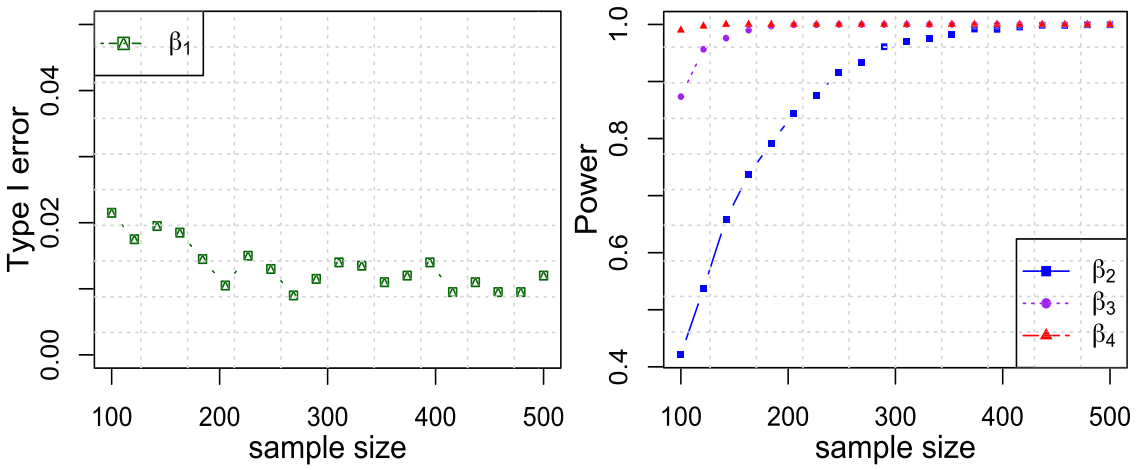

FIGURE 4 Wald test for each component of β. The calculations are averaged over 2000 repetitions. The significance level is set to 0.01. Since both variational and maximum profile likelihood methods give similar results, we only present the result of the variational method for simplicity

thoroughly discussed. The superiority of variational and maximum profile likelihood methods are verified through a variety of numerical experiments. Finally, we would like to highlight several potential extensions and open problems for future work.

1. The modelling of both the network and nodal covariates can be readily extended to more general families, such as degree-corrected SBMs, nonparametric regression, and hypergraphs (Yuan et al., 2021). The corresponding asymptotic results might be derived accordingly.
2. In the setting with high dimensional covariates, penalized likelihood methods are more appealing for both community detection and variable selection. Theoretical analysis of community detection and variable selection consistency will be necessary.

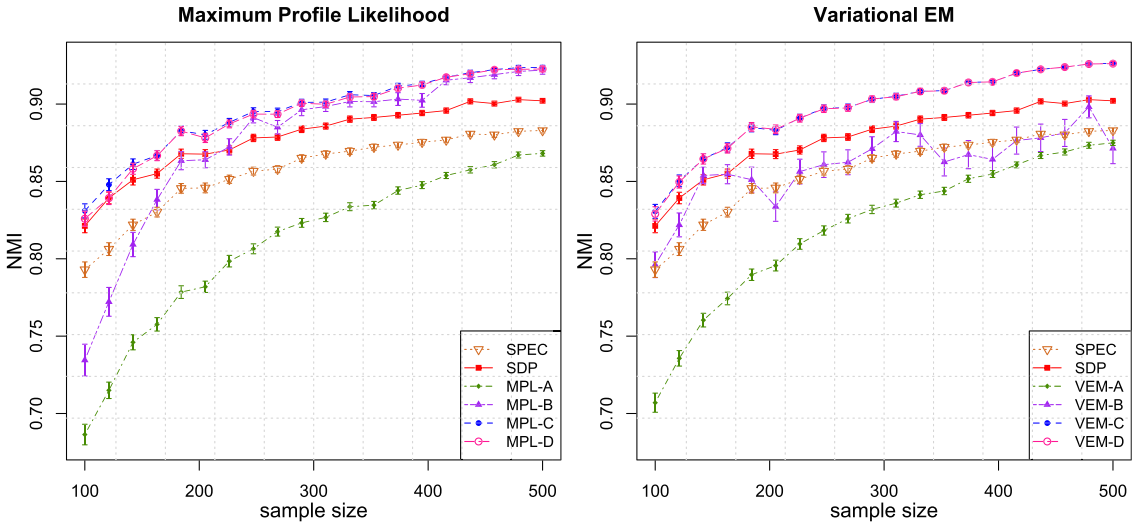

FIGURE 5 The community detection performances under Scenario (B). All the relevant descriptions are the same as in Figure 3

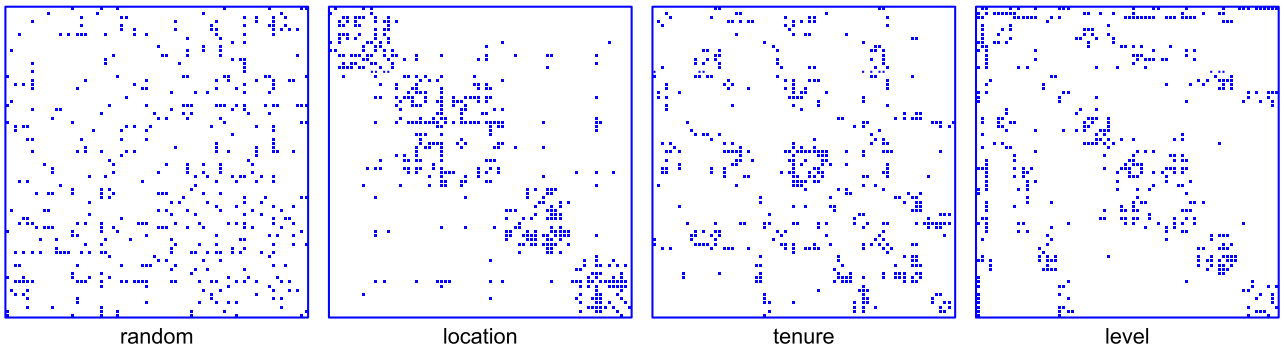

FIGURE 6 From left to right are the re-ordered adjacency matrices based on random permutation, location, tenure and organizational level

TABLE 1 The community detection results of SDP, maximum profile likelihood, and variational methods

	SDP	MPL	VEM	SDP	MPL	VEM
edge	0.881	0.894	0.894	0.882	0.883	0.883
edge + nodal	0.920	0.920	0.920	0.921	0.921	0.921

Note: MPL and VEM denote maximum profile likelihood and variational methods, respectively. NMI is computed on the left part of the table, and ARI on the right. The row indexed by "edge" shows the results based on the network without nodal information, while the other one "edge+nodal" contains the results of making use of the two attributes available.

3. For very sparse networks, considering $n\rho_n = O(1)$ seems to be a more realistic asymptotic framework. Under such asymptotic setting, community detection consistency is impossible. The effect of nodal covariates becomes more critical. It is of great interest to characterize the impact of the nodal information on community detection.
4. In this work, we assume the number of communities K is known. How to select K is an important problem in community detec tion. Some recent efforts towards this direction include (Le & Levina, 2015; Saldana et al., 2017; Wang & Bickel, 2017; Lei, 2016).
5. Running on a laptop equivalent, the proposed algorithms in the paper can efficiently handle network data with a size up to a few thousand. However, to deal with very large-scale networks, it is necessary to develop scalable versions. One promising way is to leverage some well-known randomized algorithms such as scalable SDP (Yurtsever et al., 2021), random coordinate descent (Nesterov, 2012) and stochastic variational inference (Hoffman et al., 2013).

DATA AVAILABILITY STATEMENT

The data is openly available at http://opsahl.co.uk/tnet/datasets/Cross_Parker-Manufacturing_info.txt.

ORCID

Haolei Weng https://orcid.org/0000-0002-9879-7841
Yang Feng https://orcid.org/0000-0001-7746-7598

REFERENCES

Abbe, E., & Sandon, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient recovery algorithms. arXiv: 1503.00609.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2009). Mixed membership stochastic blockmodels. In *Advances in neural information processing systems*, pp. 33–40.

Akoglu, L., Tong, H., Meeder, B., & Faloutsos, C. (2012). Pics: Parameter-free identification of cohesive subgroups in large attributed graphs. In *Sdm*, Citeseer, pp. 439–450.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, *41*(4), 2097–2122.

Amini, A. A., & Levina, E. (2014). On semidefinite relaxations for the block model. arXiv:1406.5647.

Ana, L. N. F., & Jain, A. K. (2003). Robust data clustering. In *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on*, *2*, IEEE, pp. II–128.

Anandkumar, A., Ge, R., Hsu, D., & Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, *15*(1), 2239–2312.

Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073.

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

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

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Cai, T. T., & Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, *43*(3), 1027–1059.

Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, *6*, 1847–1899.

Chang, J., & Blei, D. M. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, *4*, 124–150.

Chen, Y., Li, X., & Xu, J. (2015). Convexified modularity maximization for degree-corrected stochastic block models. arXiv:1512.08425.

Chen, Y., Sanghavi, S., & Xu, H. (2012). Clustering sparse graphs, Advances in neural information processing systems, pp. 2204–2212.

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

Cross, R. L., & Parker, A. (2004). *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*: Harvard Business Review Press.

Dasgupta, A., Hopcroft, J. E., & McSherry, F. (2004). Spectral analysis of random graphs with skewed degree distributions, Foundations of computer science, 2004. Proceedings. 45th annual IEEE symposium on, pp. 602–610.

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

Decelle, A., Krzakala, F., Moore, C., & Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, *84*(6), 066106.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Deshpande, Y., Montanari, A., Mossel, E., & Sen, S. (2018). Contextual stochastic block models, Advances in neural information processing systems, pp. 8581–8593.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174.

Gao, C., Ma, Z., Zhang, A. Y., & Zhou, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. arXiv:1505.03772.

Guédon, O., & Vershynin, R. (2016). Community detection in sparse networks via Grothendieck inequality. *Probability Theory and Related Fields*, *165*, 1025–1049.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(5), 1303–1347.

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

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

Huang, S., Weng, H., & Feng, Y. (2020). Spectral clustering via adaptive layer aggregation for multi-layer networks. arXiv preprint arXiv:2012.04646.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, *43*(1), 57–89.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Joseph, A., & Yu, B. (2013). Impact of regularization on spectral clustering. arXiv:1312.1733.

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

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., & Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, *110*(52), 20935–20940.

Le, C. M., & Levina, E. (2015). Estimating the number of communities in networks by spectral methods. arXiv preprint arXiv:1507.00827.
Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, *44*(1), 401–424.
Lei, J., & Rinaldo, A. (2014). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, *43*(1), 215–237.
Montanari, A., & Sen, S. (2015). Semidefinite programs on sparse random graphs and their application to community detection. arXiv:1504.05910.
Nallapati, R., & Cohen, W. W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence of blogs., Icwsm.
Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, *22*(2), 341–362.
Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256.
Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.
Newman, M. E. J., & Clauset, A. (2015). Structure and inference in annotated networks. arXiv:1507.04001.
Newman, M. E. J., & Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications*, *7*(1), 1–11.
Qin, T., & Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel, Advances in neural information processing systems, pp. 3120–3128.
Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, *39*(4), 1878–1915.
Ruan, Y., Fuhry, D., & Parthasarathy, S. (2013). Efficient community detection in large networks using content and links, Proceedings of the 22nd international conference on world wide web, pp. 1089–1098.
Saade, A., Krzakala, F., & Zdeborová, L. (2014). Spectral clustering of graphs with the bethe hessian, Advances in neural information processing systems, pp. 406–414.
Saldana, D. F., Yu, Y., & Feng, Y. (2017). How many communities are there? *Journal of Computational and Graphical Statistics*, *26*(1), 171–181.
Stegehuis, C., & Massoulié, L. (2019). Efficient inference in stochastic block models with vertex labels. *IEEE Transactions on Network Science and Engineering*, *7*(3), 1215–1225.
Steinhaeuser, K., & Chawla, N. V. (2010). Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, *31*(5), 413–421.
Tütüncü, R. H., Toh, K. C., & Todd, M. J. (2003). Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming*, *95*(2), 189–217.
Wang, Y. X. R. achel, & Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, *45*(2), 500–528.
Yan, B., & Sarkar, P. (2020). Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, *116*, 1–12.
Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes, Data mining (icdm), 2013 ieee 13th international conference on, pp. 1151–1156.
Yuan, M., Liu, R., Feng, Y., & Shang, Z. (2021). Testing community structures for hypergraphs. The Annals of Statistics, to appear.
Yurtsever, A., Tropp, J. A., Fercoq, O., Udell, M., & Cevher, V. (2021). Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, *3*(1), 171–200.
Zhang, Y., Levina, E., & Zhu, J. (2014). Detecting overlapping communities in networks with spectral methods. arXiv:1412.3432.
Zhang, Y., Levina, E., & Zhu, J. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, *10*(2), 3153–3178.
Zhang, A. Y., & Zhu, H. H. (2015). Minimax rates of community detection in stochastic block models. arXiv preprint arXiv:1507.05313.
Zhao, Y., Levina, E., & Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, *40*(4), 2266–2292.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Weng, H., & Feng, Y. (2022). Community detection with nodal information: Likelihood and its variational approximation. *Stat*, *11*(1), e428. https://doi.org/10.1002/sta4.428