Biometrika (2024), **00**, 0, *pp.* 1–20 https://doi.org/10.1093/biomet/asae011
Advance Access publication 23 February 2024

Network-adjusted covariates for community detection

By Y. HU

Department of Statistics and Data Science, Southern Methodist University, 3225 Daniel Avenue, Dallas, Texas 75205, U.S.A.
yaofangh@mail.smu.edu

and W. WANG

Department of Statistics and Data Science, National University of Singapore, S16 Science Drive 2, 117546 Singapore
wanjie.wang@nus.edu.sg

Summary

Community detection is a crucial task in network analysis that can be significantly improved by incorporating subject-level information, i.e., covariates. Existing methods have shown the effectiveness of using covariates on the low-degree nodes, but rarely discuss the case where communities have significantly different density levels, i.e., multiscale networks. In this paper, we introduce a novel method that addresses this challenge by constructing network-adjusted covariates, which leverage the network connections and covariates with a node-specific weight for each node. This weight can be calculated without tuning parameters. We present novel theoretical results on the strong consistency of our method under degree-corrected stochastic blockmodels with covariates, even in the presence of misspecification and multiple sparse communities. Additionally, we establish a general lower bound for the community detection problem when both the network and covariates are present, and it shows that our method is optimal for connection intensity up to a constant factor. Our method outperforms existing approaches in simulations and a LastFM app user network. We then compare our method with others on a statistics publication citation network where 30% of nodes are isolated, and our method produces reasonable and balanced results. Our method is implemented in the R package *NAC*.

Some key words: Community detection; Degree-corrected stochastic blockmodel; ℓ_∞ norm; Node attribute; Random matrix theory; Spectral clustering.

1. Introduction

Network data refer to the records of connections or relationships between subjects, and can be found in a large variety of scientific fields (Gil-Mendieta & Schmidt, 1996; Chen & Yuan, 2006; Deco & Corbetta, 2011; Jacob et al., 2011; Leskovec & McAuley, 2012; Sporns & Betzel, 2016; Binkiewicz et al., 2017; Ying et al., 2018). Network data are often represented as a graph $\mathcal{A} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of n nodes, i.e., subjects, and $\mathcal{E}$ denotes the set of edges or links between nodes. Mathematically, $\mathcal{A}$ with $|\mathcal{V}| = n$ can also be expressed

by an adjacency matrix $A \in \{0,1\}^{n\times n}$, where $A_{ij} = A_{ji} = 1$ if $(i,j) \in \mathcal{E}$ and $A_{ij} = A_{ji} = 0$ otherwise. Studies on the network data provide valuable insights into the structure by exploring interactions among subjects.

Among topics on network analysis, the most influential one is the community detection problem. This problem is also known as the clustering of nodes in a network. Communities refer to the groups of nodes, so that nodes within the same community are more densely connected than nodes in different communities. If we consider networks on genome data or brain images, the community structure can represent functional modules or coordination between nodes. Therefore, detecting the communities can provide insights into challenging biological problems.

There are plenty of studies of the algorithms and theoretical limits on the community detection problem, especially for sparse networks. Let d_i denote the number of neighbours of node i. It has been found in multiple works that $\min_{i\in\mathcal{V}} E(d_i) \geqslant C\log n$ for a constant $C > 0$ is required to ensure the exact recovery of community labels for each node (Bickel & Chen, 2009; Abbe, 2017). Such limits are often referred to as information theoretical lower bounds. How to handle networks where some nodes' expected degrees are bounded remains a challenging problem. Joseph & Yu (2016) suggested classifying all sparsely connected nodes as a single community, but this might be an oversimplification in some scenarios. Lei et al. (2021) discussed multiscale networks where the communities have different density levels, but $E(d_i) \geqslant C\log n$ needs to hold for all levels.

This work studies the possibility of sparse network community detection by leveraging covariate information. We consider a challenging mixture setting where the communities can be either relatively dense or extremely sparse. According to the fundamental limits in community detection (Bickel & Chen, 2009; Abbe, 2017), we require the relatively dense communities to have expected degrees larger than $c_d \log n$, where $d_i \gg \log n$ is also allowed. Meanwhile, nodes in extremely sparse communities have expected degrees no larger than $c_s \log n$, where $d_i = O(1)$ is allowed. Mathematically, a multiscale network with both relatively dense and extremely sparse communities is defined as follows.

Definition 1. *Consider a network $\mathcal{A} = (\mathcal{V}, \mathcal{E})$ and assume that constants $c_d > c_s > 0$. Then community k is called a*

(i) *(relatively) dense community if $E(d_i) \geqslant c_d \log n$ for all i so that $\ell(i) = k$;*
(ii) *(extremely) sparse community if $E(d_i) \leqslant c_s \log n$, for all i so that $\ell(i) = k$.*

Network $\mathcal{A}$ is called a multiscale network with extremely sparse communities if both kinds of community exist.

Consider the nodes in sparse communities where the labels cannot be recovered by the network. Modern datasets often include subject-level covariates other than the network. Let $x_i \in \mathbb{R}^p$ denote the covariate vector of node i. The covariate matrix is defined as $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{n\times p}$. In biology, these covariates may include demographic information, clinical or genetic data, or other relevant features. The covariates often depend on the community structure of the subjects. Therefore, integrating X with $\mathcal{A}$ will largely improve the community detection results, especially for nodes in sparse communities.

The integration of covariates with network data for community detection has become popular very recently. Newman & Clauset (2016) proposed a model that incorporates low-dimensional discrete covariates and the network based on community memberships. The maximum likelihood estimate is obtained using the belief propagation package. Yan &

Sarkar (2021) interpreted it as an optimization problem on the weighted sum of the network Laplacian and covariate kernel matrix. Binkiewicz et al. (2017) discussed the spectral clustering method, with an application on a weighted sum of the network Laplacian and covariates. Later, the issue of mismatching between covariates and community labels became a concern. To solve this problem, Yang et al. (2013) used the maximum likelihood approach where covariates and the network are considered separately, and Zhang et al. (2016) optimized a joint community detection criterion analogous to the modularity. Other approaches can be found in various studies, including those of Yan et al. (2019), Weng & Feng (2022), Huang & Feng (2023) and Xu et al. (2023).

The theoretical consistency of network community detection algorithms has also been a hot topic in recent years. For networks, studies first discuss weak consistency, i.e., the clustering error rate converges to 0 as $n \to \infty$. Later, strong consistency, i.e., exact recovery, is brought into concern, which means that the label of every node can be exactly recovered with a high probability. Without covariates, weak consistency can be achieved when $E(d_i) \to \infty$ and strong consistency requires that $E(d_i) \geqslant C \log n$ (Abbe et al., 2015; Gao et al., 2017; Abbe et al., 2020; Chen et al., 2021). With covariates, weak consistency has been found for algorithms under regular conditions. Furthermore, for the high-dimensional covariates, Deshpande et al. (2018) set up the fundamental limit of the signal-to-noise ratio to guarantee weak consistency, and Ma & Nandy (2023) generalized the results to multilayer networks. Yet there are very few works on the strong consistency results when covariates are involved. The only work is that of Abbe et al. (2022), who considered the two-community stochastic blockmodel. The upper bound of their proposed spectral method and the lower bound for strong consistency have been established.

A direct generalization of these methods to multiscale networks with covariates faces several challenges. Firstly, in multiscale networks, the usefulness of covariates depends on the density of the communities. Indeed, dense communities can be recovered with connections alone, while a successful recovery of sparse communities would depend more on the covariates. However, most existing algorithms leverage the network and covariates by putting a single weight on the whole covariate matrix X, without calibrations on the node-specific effects. Secondly, to elaborate on the performance of each node in relatively dense communities and sparse communities, we want to establish the strong consistency results of our algorithm. However, a multiscale network will induce multiscale errors, which is a challenge in theoretical analysis.

We introduce a novel approach called spectral clustering on network-adjusted covariates, which is tuning-free and efficient on multiscale networks with covariates. This approach contains two steps. We first define the network-adjusted covariate vectors

$$y_i = \alpha_i x_i + \sum_{\{j:\, A_{ij}=1\}} x_j, \qquad i \in [n].$$

The new covariate y_i combines the original covariate x_i and the network information using $\sum_{\{i:\, A_{ij}=1\}} x_j$. We design the node-specific coefficient α_i so that it effectively balances the contribution of x_i and the neighbours. Then we apply spectral clustering on the network-adjusted covariate matrix $Y = [y_1, \ldots, y_N]$. Under the degree-corrected stochastic blockmodel (Bickel & Chen, 2009; Karrer & Newman, 2011; Zhao et al., 2012), we prove novel results on the spectral properties of Y, where we control the row-wise distance between the population and empirical spectral matrices. It hence induces strong consistency of our new approach. We further set up the lower bound, which meets the upper bound induced by our algorithm up to a constant.

Our work also considers the challenging scenarios where the covariates can be misspecified. For node i, covariate x_i may not follow the common covariate distribution of nodes in this community. This misspecification may come from random error or a systematic mismatching between covariates and community labels. By our new spectral analysis results, we find that the node label can be exactly recovered, when the node is either in the relatively dense communities or its covariate is correctly specified. In other words, even with the existence of misspecification, our method still recovers the node labels as long as the information, either from $\mathcal{A}$ or X, is sufficient. Such analysis on each node is novel.

Spectral information is commonly used in various statistical fields, including community detection; see Chung & Graham (1997), Rohe et al. (2011), Chaudhuri et al. (2012), Amini et al. (2013) and Jin (2015). Weak consistency can be proved by controlling the Frobenius norm with the Davis–Kahan theorem (Jin, 2015; Lei & Rinaldo, 2015). Recently, Fan et al. (2018) established the upper bound on the ℓ_∞ norm of the eigenvector perturbation. This improvement has motivated the strong consistency results of spectral methods; see Su et al. (2019) and Abbe et al. (2020). When the covariates are included, Binkiewicz et al. (2017) applied the spectral method with weak consistency and Abbe et al. (2022) considered an aggregate spectral method with the strong consistency results. Therefore, we consider spectral clustering for multiscale networks with covariates. We demonstrate that our method can achieve the exact recovery results, except for low-degree nodes that are misspecified.

To conclude, this work discusses multiscale networks with covariates where misspecification is considered. We propose the new network-adjusted covariate vectors that assign node-specific weights to covariates. Using these network-adjusted covariates, we propose a tuning-free and computationally efficient community detection method. We provide solid theoretical results for the new method. The entrywise perturbation of eigenvectors shows that the label recovery is based on $\mathcal{A}$ for nodes in dense communities and X for nodes in sparse communities. Hence, an exact recovery can be achieved when the misspecification happens only in dense communities. We further establish the lower bound for the community detection problem on networks with covariates. The lower bound matches the upper bound from our new method up to a constant factor, which suggests the optimality of our approach.

2. Methodology

2.1. *Notation and background*

We represent a network $\mathcal{A}$ with covariates as a duplex (A, X), where $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix and $X \in \mathbb{R}^{n\times p}$ is the covariate matrix. Each row in X is x_i, the covariate vector associated with node i. For the adjacency matrix A, let $d_i = \sum_{j=1}^n A_{ij}$ denote the degree of node i and $\bar{d} = \sum_{i=1}^n d_i/n$ be the average degree. Let $\mathcal{D}$ denote nodes in relatively dense communities and $\mathcal{S}$ denote nodes in sparse communities. Nodes $\mathcal{D}$ and $\mathcal{S}$ are unknown to us.

Let K be the number of communities and $\ell \in \mathbb{R}^n$ be the community label vector, where each entry $\ell(i) \in [K]$ corresponds to the community membership of node i. It can also be represented in the matrix form $\Pi \in \{0,1\}^{n\times K}$, where $\Pi(i,j) = 1$ if $\ell(i) = j$ and 0 otherwise. Our objective is to recover ℓ.

For a vector a, $\|a\|$ gives the ℓ_2 norm of a and $\|a\|_\infty = \max_i \|a_i\|$ gives the ℓ_∞ norm. Let A be a matrix, $\lambda_k(A)$ denote the kth largest singular value of A and $\|A\| = \lambda_1(A)$. For two series a_n and b_n, we say that $a_n \asymp b_n$ if there is a constant C such that $a_n \leqslant Cb_n$ and

$b_n \leqslant Ca_n$ when n is large enough. We say that $a_n \lesssim b_n$ if $\overline{\lim}_{n\to\infty} a_n/b_n \leqslant 1$. We define $a_n \gtrsim b_n$ similarly. Finally, we use the notation $[N] := \{1, \ldots, N\}$ for any integer N.

2.2. *Network-adjusted covariates*

To leverage the network and covariates, we propose the network-adjusted covariate vectors

$$y_i = \alpha_i x_i + \sum_{\{j:\, A_{ij}=1\}} x_j, \qquad i \in [n].$$

Here, the weight α_i of the node's covariate is defined as

$$\alpha_i = \frac{\bar{d}/2}{d_i/\log n + 1}. \tag{1}$$

The network-adjusted covariate vector y_i consists of two parts: the covariate vector of the node itself, $\alpha_i x_i$, and the sum of covariates of its neighbours, $\sum_{\{j:\, A_{ij}=1\}} x_j$. The former conveys the node's individual covariate information, while the latter establishes the node's network information. Similar methods utilizing neighbours can be found in Hu et al. (2022).

Here is the intuition of how α_i balances these two parts. Consider a multiscale network. When the community sizes are comparable, the average degree $\bar{d}$ is dominated by the dense communities. Let $d_{\mathcal{D}} \geqslant c_d \log n$ and $d_{\mathcal{S}} \leqslant c_s \log n$ denote the average degrees of the dense communities and sparse communities, respectively. Hence, $\bar{d} \asymp d_{\mathcal{D}}$. Suppose that nodes in the same community have expected degrees at the same asymptotic rate; then $\alpha_i \approx (\log n)/2$ for $i \in \mathcal{D}$ and $\alpha_i \approx \bar{d} \asymp d_{\mathcal{D}}$ for $i \in \mathcal{S}$ as the weight of x_i. Meanwhile, recall that the number of neighbours is $d_i \asymp d_{\mathcal{D}}$ for $i \in \mathcal{D}$ and $d_i \asymp d_{\mathcal{S}} \leqslant d_{\mathcal{D}}$ for $i \in \mathcal{S}$. Therefore, $\|y_i\| \asymp d_{\mathcal{D}}\|x_i\|$ for all nodes. Furthermore, when $c_{\mathcal{D}}$ is sufficiently large, y_i focuses on $\sum_{\{j:\, A_{ij}=1\}} x_j$ when $i \in \mathcal{D}$ and $\alpha_i x_i$ when $i \in \mathcal{S}$. Hence, this new network-adjusted covariate vector y_i can efficiently leverage the information from both the network and original covariates.

The definition of α_i in (1) includes two manually decided factors. The numerator $\bar{d}/2$ has a constant factor of $1/2$. We want to point out that this constant factor is not essential and any constant between 0 and 1 can be used instead. The strong consistency results in § 3.2 below hold for any such constants; the proofs are given in the Supplementary Material. The denominator $d_i/\log n + 1$ has a $\log n$ term. This term is decided by the definition of relatively dense communities and extremely sparse communities, which traces back to the fundamental limits of strong consistency. In our numerical tests, we found that using $1/2$ and $\log n$ yields good results.

Let $Y = (y_1, \ldots, y_n)'$ be the network-adjusted covariate matrix; then

$$Y = AX + D_\alpha X, \tag{2}$$

where D_α is a diagonal matrix with diagonals $(\alpha_1, \ldots, \alpha_n)$.

Consider the special case that $\mathcal{A}$ has a uniform scale and $d_i \geqslant c_d \log n$, i.e., $\mathcal{V} = \mathcal{D}$. By the formula, we have $D_\alpha \approx (\log n/2)I$ and $Y \approx \{A + (\log n/2)I\}X$. The left singular vectors of Y mainly depend on A. Hence, community detection based on the left singular vectors yields the same error rate based on A.

2.3. *Spectral clustering on network-adjusted covariates*

Spectral methods on community detection were first proposed by Chung & Graham (1997), and have since been developed in various directions, such as spectral methods on the graph Laplacian, regularized Laplacian, nonbacktracking matrix and more (Rohe et al., 2011; Chaudhuri et al., 2012; Krzakala et al., 2013; Joseph & Yu, 2016). Theoretical discussions about eliminating the degree effects using spectral methods have been shown under degree-corrected stochastic blockmodels by Jin (2015) and Lei & Rinaldo (2015). Binkiewicz et al. (2017) employed spectral methods for networks with covariates, on a weighted summation of the network Laplacian and covariate matrix, where the weight is a tuning parameter. Abbe et al. (2022) discussed the aggregated spectral methods on the two-community stochastic blockmodel. Here, we apply spectral clustering on the network-adjusted covariate matrix Y in (2) and propose the following algorithm.

Algorithm 1. Spectral clustering on network-adjusted covariates.

Input: adjacency matrix A, covariate matrix X, number of communities K.

1. Find $Y = AX + D_\alpha X$, where D_α is defined in (1).
2. Find the top K left singular vectors $\hat{\Xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_K]$ of Y.
3. Find $\hat{R}$ by normalizing $\hat{\Xi}$ such that each row has norm 1.
4. Perform k-mean clustering on $\hat{R}$ with K clusters, treating every row as a data point.
5. The output label vector $\hat{\ell}$ by k means in step 4 gives us the community label.

In step 4, we apply the built-in k-means function in R, which finds a local optimum by the algorithm of Hartigan & Wong (1979). To reduce errors, we use multiple random seeds.

The high-level intuition of why applying k means on $\hat{R}$ yields a satisfactory community detection result is as follows. As we explained, the network-adjusted covariate vectors y_i are dominated by the covariates of neighbours if $i \in \mathcal{D}$ or by the covariate itself if $i \in \mathcal{S}$. Therefore, consider a relatively dense community k_d with all nodes in community k_d sharing the same distribution of neighbours. Thus, the sum of neighbours' covariates has the same distribution, up to the degree heterogeneity factor. Therefore, rows in Y corresponding to community k_d share the same centre up to a constant factor. Now we turn to a sparse community, say k_s. For i in community k_s, y_i is dominated by $\alpha_i x_i$. Since the x_i have the same distribution, up to the constant factor α_i, these y_i have the same distribution. Again, rows of Y corresponding to community k_s share the same centre up to a constant factor. With a delicate random matrix analysis, we can prove that the left singular matrix $\hat{\Xi}$ in step 2 inherits such consistency within each community, and hence rows of $\hat{R}$ in step 3 corresponding to the same community will have the same centre, eliminating the constant factor by normalization. The k-means algorithm minimizes the within-cluster sum of the squared distance to the centre, which achieves the true labels.

2.4. *Generalization with uninformative covariates*

The newly proposed network-adjusted covariate matrix Y can be seen as a product of $A + D_\alpha$ and X. By linear algebra, a meaningful $\hat{R}$ requires X to hold some information on the community structure. In other words, X cannot be uninformative; otherwise, involving X is pointless. In most cases, researchers can tell whether this is the case based on their experience.

But, for the sake of completeness, we still take this case into consideration. When it is difficult to decide whether X should be involved, we propose a slightly modified version, Algorithm 2 below. In Algorithm 2, we combine the new covariate matrix YY' and the network AA' by a weighted summation, and then apply spectral clustering on this combined matrix. The new term AA' does not rely on X. Adding it helps us to handle the extreme scenario when X is uninformative.

Algorithm 2. Spectral clustering on generalized network-adjusted covariates.

Input: adjacency matrix A, covariate matrix X, number of communities K.

1. Find $Y = AX + D_\alpha X$, where D_α is defined in (1).
2. Define $L = YY' + \beta nAA'$.
3. Find the top K left singular vectors $\hat{\Xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_K]$ of L.
4. Find $\hat{R}$ by normalizing $\hat{\Xi}$ such that each row has norm 1.
5. Perform k-mean clustering on $\hat{R}$ with K clusters, treating every row as a data point.
6. The output label vector $\hat{\ell}$ by k means in step 4 gives us the community label.

In § 3.3 below, we show that the oracle matrix of L can be written in the same YY' format, but with a generalized definition of X. Therefore, we use the term *generalized network-adjusted covariates*.

The tuning parameter β intends to balance the term nAA' and YY'. Theoretical analysis in § 3.3 below suggests that β should be a constant time $\max_i \|x_i\|^2$. For numerical analysis, we choose $\beta = \|\bar{x}\|^2$, where $\bar{x}$ is the average covariate vector. It has shown promising clustering results in data analysis. To understand this selection, consider the simplified case that X is uninformative, i.e., all nodes have the same mean covariate vector $\mu = E(x_i)$. For this case, A must be dense and $Y \approx AX$. When μ overrides the noise in x_i, it follows that $\|YY'\| \approx \|AXX'A'\| \approx \|A1_n\mu'\mu 1_n'A'\| \leqslant n\|\mu\|^2\|AA'\|$, where $1_n \in \mathcal{R}^n$ has all entries as 1. When $\beta \geqslant \|\mu\|^2$, $\|\beta nAA'\| \geqslant n\|\mu\|^2\|AA'\| \geqslant \|YY'\|$. So $L \approx \beta nAA'$, which provides community information. This motivates us to use $\beta = \|\mu\|^2$, which becomes $\|\bar{x}\|^2$ as the data version. For the special case that $\|\mu\|$ is much smaller than $\|x_i\|$, which can occur when $\mu = 0$ due to signal cancellation, quantiles of $\{\|x_i\|^2\}_{i\in[n]}$ may be a good choice for β. Further theoretical discussion of β can be found in § 3.3 below.

3. Theoretical guarantee

3.1. *Degree-corrected stochastic blockmodel with covariates*

To formulate the consistency of our proposed approach, we first model the network with covariates (A, X), under the assumption that A and X are independent given ℓ. Then we introduce the relatively dense and sparse communities and misspecification into the model.

One of the most popular network models is the degree-corrected stochastic blockmodel. The original stochastic blockmodel was first proposed in the seminal work of Holland et al. (1983) and it produced promising community detection results. Later works (Bickel & Chen, 2009; Karrer & Newman, 2011; Zhao et al., 2012; Yan et al., 2014) generalized it to the degree-corrected stochastic blockmodel, which allows for degree heterogeneity.

We follow the same line to model the network. Say $\mathcal{A}$ has n nodes in K communities, and the community membership matrix is Π. Define a symmetric matrix $P \in \mathbb{R}^{K\times K}$, where $P(k, l)$ denotes the connection intensity parameter between a node in community k and a

node in community l. To account for degree heterogeneity, we introduce a diagonal matrix Θ with diagonals $\Theta_{ii} = \theta_i$, representing the popularity of node i. The adjacency A has Bernoulli-distributed entries with parameter $P(A_{ij} = 1) = \theta_i\theta_j P\{\ell(i), \ell(j)\}$, so the probability of an edge between nodes i and j depends on their popularity and the connection intensity between the communities to which they belong. In matrix form, the adjacency matrix A can be fully identified by $E(A \mid \Pi)$, that is,

$$E(A \mid \Pi) = \Omega_A - \mathrm{diag}(\Omega_A), \qquad \Omega_A = \Theta\Pi P\Pi'\Theta. \tag{3}$$

Here, $\mathrm{diag}(\Omega_A)$ is the diagonal matrix formed by replacing all the off-diagonals of Ω_A with 0. This eliminates the possibility of self-loops in the network.

Now we model the covariates. Given label ℓ, we assume that X is independent of A. The covariates x_i are generated by a standard cluster model (Jin & Wang, 2016; Jin et al., 2017), that is, they are independently distributed as

$$x_i \mid \Pi \sim F_k, \qquad \ell(i) = k. \tag{4}$$

Here, F_k is a general distribution for community k, $k \in [K]$. We further model the misspecification issue. Let $\mathcal{M}$ denote the set of misspecified nodes, which means that x_i does not follow F_k for $i \in \mathcal{M}$ and $\ell(i) = k$. We allow x_i to follow any distribution G_i, which can be either F_k, $k \neq \ell(i)$, or other distributions.

Combining (3) and (4) gives the degree-corrected stochastic blockmodel with covariates.

DEFINITION 2 (DEGREE-CORRECTED STOCHASTIC BLOCKMODEL WITH COVARIATES). *Consider a network $\mathcal{A} = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ has a covariate vector x_i. We call the network a degree-corrected stochastic blockmodel with covariates if* (3) *and* (4) *are satisfied, with parameter set* $(\Theta, K, P, \Pi, F_{[K]}, \mathcal{M})$.

Under the degree-corrected stochastic blockmodel with covariates, we interpret the definitions of relatively dense communities and sparse communities. Recall that θ_i is the degree heterogeneity parameter of node i and that the expected degree $E(d_i) \leqslant n\theta_i\theta_{\max}$. Let $\theta_{\max} = \|\theta\|_\infty$ denote the maximum.

DEFINITION 3. *Consider a network $\mathcal{A} = (\mathcal{V}, \mathcal{E})$ that follows the degree-corrected stochastic blockmodel with parameters (Θ, K, P, Π), and let $\theta \in \mathbb{R}^n$ denote the diagonals of Θ with $\theta_{\max} = \|\theta\|_\infty$. Then community k is called a*

(i) *(relatively) dense community if there exist constants $c, c_d > 0$ such that $\theta_i \geqslant c\theta_{\max}$ and $n\theta_i\theta_{\max} \geqslant c_d \log n$ for all i, so that $\ell(i) = k$;*
(ii) *sparse community if there exists a constant $0 < c_s < c_d$ such that $n\theta_i\theta_{\max} \leqslant c_s \log n$ for all i, so that $\ell(i) = k$.*

Let $\mathcal{D}$ be the set of nodes in relatively dense communities and $\mathcal{S}$ be the set of nodes in sparse communities. We consider the node set $\mathcal{V} = \mathcal{D} \cup \mathcal{S}$. By Definition 3, all nodes in $\mathcal{D}$ have expected degree $E(d_i) \asymp n\theta_i\theta_{\max} \geqslant c_d \log n$. Hence, the diverging $E(d_i)$ indicates sufficient network information. Meanwhile, all nodes in $\mathcal{S}$ have expected degree $E(d_i) \lesssim n\theta_i\theta_{\max} \leqslant c_s \log n$. In the extremely sparse case, $E(d_i) \to 0$ for $i \in \mathcal{S}$. It is challenging and the covariates X will be leveraged for accurate clustering.

Most existing works assume that the θ_i diverge at the same rate and that $n\theta_{\max}^2 \geqslant c \log n$; see Amini et al. (2013), Krzakala et al. (2013) and Jin (2015). However, some communities in

practice tend to make more connections, while other communities have more isolated nodes. This phenomenon was put forward by Joseph & Yu (2016), who used the terms 'dense and weak clusters' to refer to communities with different connection intensities, under the stochastic blockmodel. They studied the case where the size of weak clusters does not increase with n and proved consistency on the dense clusters. By leveraging the covariates, we can achieve exact recovery results even for these sparse communities using our new method.

By the high-level analysis in § 2.3, the label recovery of nodes in $\mathcal{D}$ relies on the network information and that of nodes in $\mathcal{S}$ relies on the covariates. Ideally, only the misspecified nodes in $\mathcal{S}$ should be affected and misclustered, i.e., the set $\mathcal{M} \cap \mathcal{S}$. Thus we define a parameter as the proportion of these nodes:

$$\epsilon = |\mathcal{M} \cap \mathcal{S}|/n. \tag{5}$$

This parameter captures the proportion of the 'essential errors' under the model. Nodes in $\mathcal{M} \cap \mathcal{S}$ cannot be labelled correctly, no matter the method applied. We give a theorem to rigorously demonstrate this intuition in § 3.4 below. One interesting fact is that nodes in $\mathcal{M} \cap \mathcal{D}$ are not included, even though they are misspecified.

3.2. *Consistency of spectral clustering on network-adjusted covariates*

Under the degree-corrected stochastic blockmodel with covariates, we demonstrate strong consistency of spectral clustering on network-adjusted covariates.

To derive the consistency, we first consider the oracle case where model parameters are known. With the population version of Y, we derive the left singular matrix Ξ in Lemma 1 below. It shows that the rows of Ξ are closely related to the community labels of corresponding nodes. We then introduce noise into the model and find $\hat{\Xi}$ in Algorithm 1. In Theorem 1 below, we bound the ℓ_2 norm between each row of Ξ and $\hat{\Xi}$ up to a rotation. This row-wise bound leads to the exact recovery of the underlying community memberships on the good nodes, i.e., nodes in $(\mathcal{M} \cap \mathcal{S})^c$, as shown in Theorem 2 below. General forms of the theorems and proofs are given in the Supplementary Material.

Consider the oracle case where the parameters are known. We define the population version of A, X and D_α. By (3), $E(A) = \Theta \Pi P \Pi' \Theta$. The oracle matrix for X is that all the nodes display the correct information, which means that $\tilde{E}(x_i) = E(F_{\ell(i)})$ for all $i \in [n]$. Note that $\tilde{E}(x_i)$ and $E(x_i)$ may not be the same if x_i is misspecified. We define $\tilde{E}(X)$ as the matrix formed by these mean vectors. Finally, define

$$\alpha_i^* = E(\bar{d}) \log n / [2\{E(d_i) + \log n\}], \qquad i \in [n],$$

and let D_{α^*} be a diagonal matrix formed by the α_i^*.

Now we set up an oracle matrix for the network-adjusted covariate matrix $Y = AX + D_\alpha X$. Instead of reverting every part to the population version, we consider the dominating terms only. Consider the ith row in Y, i.e., y_i. If $i \in \mathcal{D}$, the dominating term is $\sum_{\{j:\, A_{ij}=1\}} x_j$ and the population version is the ith row of $E(A)\tilde{E}(X)$. Meanwhile, if $i \in \mathcal{S}$ then the dominating term is $\alpha_i x_i$ and the population version is the ith row of $D_{\alpha^*}\tilde{E}(X)$. To express the oracle matrix, let $I_{\mathcal{D}} \in \mathbb{R}^{n \times n}$ denote the identity matrix where only diagonals on $\mathcal{D}$ are preserved and others are set as 0. We define $I_{\mathcal{S}}$ similarly. The oracle matrix is defined as

$$\Omega = \{I_{\mathcal{D}} E(A) I_{\mathcal{D}}\}\tilde{E}(X) + I_{\mathcal{S}}\{D_{\alpha^*}\tilde{E}(X)\} = \{I_{\mathcal{D}} E(A) I_{\mathcal{D}} + I_{\mathcal{S}} D_{\alpha^*}\}\tilde{E}(X). \tag{6}$$

The nodes can be decomposed into three disjoint sets: $\mathcal{D}$, where nodes are in relatively dense communities; $\mathcal{S} \cap \mathcal{M}^c$, where nodes are in sparse communities with correct covariate distribution; and $\mathcal{S} \cap \mathcal{M}$, where nodes are in sparse communities and the covariates are misspecified. Only nodes in the first two sets are possible to cluster correctly. Theorem 4 below rigorously states the impossibility for nodes in $\mathcal{S} \cap \mathcal{M}$. Hence, we define the set of good nodes as

$$\mathcal{G} = \mathcal{D} \cup (\mathcal{S} \cap \mathcal{M}^c).$$

Comparing this to (5), we see that $\epsilon = 1 - |\mathcal{G}|/n$. We then define $I_{\mathcal{G}}$ in a similar way to $I_{\mathcal{D}}$.

Lemma 1 (Spectral analysis of the oracle matrix). *Consider the oracle matrix Ω with a good node set $\mathcal{G}$. Let $\Omega_{\mathcal{G}} = I_{\mathcal{G}}\Omega$ be the oracle matrix with rows restricted on $\mathcal{G}$. Denote the singular value decomposition of $\Omega_{\mathcal{G}}$ as $\Omega_{\mathcal{G}} = \Xi\Lambda U'$, where $\Xi \in \mathbb{R}^{n\times K}$, $U \in \mathbb{R}^{p\times K}$ and $\Lambda \in \mathbb{R}^{K\times K}$.*

Under the degree-corrected stochastic blockmodel with covariates,

$$\Xi_i = \begin{cases} \theta_i v_{\ell(i)}, & i \in \mathcal{D}, \\ \alpha_i^* u_{\ell(i)}, & i \in \mathcal{S} \cap \mathcal{M}^c, \\ 0, & i \in \mathcal{S} \cap \mathcal{M} = \mathcal{G}^c, \end{cases}$$

where the v_k and u_k are K-dimensional vectors.

Lemma 1 shows that nodes in the same community share the same rows in Ξ, up to a constant factor. This constant factor is the degree heterogeneity parameter θ_i for $i \in \mathcal{D}$ or the weightage $\alpha^* \approx E(\bar{d})/2$ for $i \in \mathcal{S} \cap \mathcal{M}^c$. The explicit formulae for u_k and v_k can be found in the Supplementary Material. Normalizing Ξ_i to be of unit length removes the constant factor. When the centres, the normalized u_k and v_k, are well separated, the labels of nodes in $\mathcal{G}$ can be exactly recovered.

Theorem 1 (Row-wise empirical and oracle singular matrix distance). *Consider the degree-corrected stochastic blockmodel with covariates and parameters $(\Theta, K, P, \Pi, F_{[K]}, \mathcal{M})$, where $p > 0$ is a constant, $\mathcal{G}$ is the set of good nodes and $\epsilon = |\mathcal{G}^c|/n$. Let Ω be the oracle matrix defined in (6), Ξ be the left singular matrix of $\Omega_{\mathcal{G}}$ and let $\hat{\Xi}$ consist of the top K left singular vectors of Y.*

Let $c, C > 0$ be constants that vary case by case. We assume that

(i) *the submatrix of P that is restricted to dense communities $P_{\mathcal{D}}$ is full rank;*
(ii) *$\|x_i\| \leqslant R$ almost surely, and, for $i \in \mathcal{S} \cap \mathcal{G}$, with high probability, $\|x_i - \tilde{E}(x_i)\| \leqslant \delta_X R$;*
(iii) *$\lambda_K\{\tilde{E}(X)\} \geqslant c\sqrt{n}R$; and*
(iv) *the number of nodes in any community $n_k/n \geqslant c > 0$.*

Then there exist threshold constants $C_\theta, \epsilon_0, n_0$ and δ_0 such that if $\delta_X \leqslant \delta_0, \epsilon \leqslant \epsilon_0, n \geqslant n_0, n\theta_{\max}^2 \geqslant C_\theta \log n$, there exists an orthogonal matrix O and a constant $C > 0$ with probability $1 - O(1/n)$ such that

$$\max_{i\in\mathcal{G}} \|\hat{\Xi}_i - O\Xi_i\| \leqslant C(\delta_X + \sqrt{\epsilon} + 1/\sqrt{C_\theta})/\sqrt{n},$$

where $\hat{\Xi}_i$ and Ξ_i are vectors formed from the ith row of $\hat{\Xi}$ and Ξ.

Assumptions (i) and (iv) are regular conditions on networks and assumptions (ii) and (iii) impose regularity conditions on the covariates. Assumption (i) requires linearly independent rows/columns of $P_{\mathcal{D}}$ for different dense communities, a common requirement in community detection on P (Jin, 2015; Weng & Feng, 2022). We use the network information only for $\mathcal{D}$, so the requirement is on $P_{\mathcal{D}}$. Assumption (iv) is a standard requirement in stochastic blockmodels to ensure comparable community sizes. Assumption (ii) states the range of covariates x_i and the concentration. Bounded covariates are preferred, as in Binkiewicz et al. (2017), although the results can be extended to unbounded cases, with more complicated interpretations. Nodes in sparse communities are labelled by their covariates, so for those nodes, the noise or deviation $\|x_i - \tilde{E}(x_i)\|$ must be small. The noise level is denoted by δ_X. Assumption (iii) requires the smallest singular value of $\tilde{E}(X)$ to be $O(\sqrt{nR})$, the same order as the largest singular value. It can be verified if the $K \times p$ matrix formed by the $E(F_k)$ is nonsingular. On the other hand, if $\tilde{E}(X)$ has rank $< K$ then even the oracle matrix Ω has rank $< K$ and Ξ can be insufficient for label recovery. Finally, the results hold when the relatively dense communities have degrees exceeding $C_\theta \log n$; otherwise, the network does not contribute any information.

Theorem 1 provides a row-wise bound for the closeness between the rows of $\hat{\Xi}$ and Ξ, up to a rotation. It supports the row-wise operations, normalization and k means in Algorithm 1. The most notable aspect of the theorem is that the bound is row-wise, instead of the standard Frobenius norm bound on $\hat{\Xi} - \Xi O$ using the Davis–Kahan approach. To the best of our knowledge, this result is not proved in the context of degree-corrected stochastic blockmodels with covariates, and it cannot be obtained using existing analysis tools under our challenging misspecified setting. A new singular vector stacking result is developed in the Supplementary Material to handle misspecification. This row-wise analysis enables the separation of good nodes $\mathcal{G}$ and other nodes $\mathcal{G}^c$, leading to exact recovery on $\mathcal{G}$, even in the presence of misspecification.

The bound given in Theorem 1 contains three parts: $1/\sqrt{C_\theta}$ due to density of the network, δ_X due to the randomness in X and $\sqrt{\epsilon}$ due to the misspecified nodes in sparse communities, i.e., $\mathcal{G}^c$. Although only nodes in $\mathcal{G}$ are considered, perturbation from $\mathcal{G}^c$ is unavoidable since the singular vectors are obtained from the whole matrix. When all three terms are well bounded, we achieve exact recovery of the entire network.

THEOREM 2 (STRONG CONSISTENCY OF ALGORITHM 1). *Suppose that conditions (i)–(iv) in Theorem 1 hold. Let $\hat{\ell}$ be the estimated labels by the spectral clustering method on Y. Then there exists a constant C_θ independent of n such that if $n\theta_{\max}^2 \geqslant C_\theta \log n$, there exists a permutation π so that, with probability $1 - O(1/n)$,*

$$\pi\{\hat{\ell}(i)\} = \ell(i) \quad \textit{all } i \in \mathcal{G}.$$

Therefore, the community detection error rate is bounded by $|\mathcal{G}^c|/n = \epsilon$.

In summary, under reasonable conditions, with a high probability, our method can exactly recover the label of every node in $\mathcal{G}$.

3.3. *Consistency of spectral clustering on generalized network-adjusted covariates*

To achieve strong consistency, we require $\tilde{E}(X)$ to have rank K. In Algorithm 2, we consider the case that $\text{rank}\{\tilde{E}(X)\} < K$ and add AA' to YY' to achieve good clustering results. In this section, we demonstrate the consistency of this algorithm.

Consider the multiscale network with covariates. Nodes in $\mathcal{D}$ are labelled by the network information, for which we do not expect conditions on X. Nodes in $\mathcal{S}$ are labeled by the covariate information, and distinction is required on these x_i. Hence, we relax Theorem 1(iii) as follows on only the sparse communities.

(iii') Let $K_{\mathcal{S}}$ be the number of sparse communities. There exists a constant $c > 0$ such that $\lambda_{K_{\mathcal{S}}}\{I_{\mathcal{S}}\tilde{E}(X)\} \geqslant c\sqrt{nR}$.

To demonstrate consistency of Algorithm 2, we first introduce $\tilde{\Omega}$ as the counterpart of Ω in the generalized case. The contribution from AA' is summarized in a diagonal matrix $T = (\Pi'\Theta\Pi)^{-1}(n\Pi'\Theta^2\Pi)^{1/2} \in \mathbb{R}^{K_{\mathcal{D}}\times K_{\mathcal{D}}}$, where each entry is the ℓ_2 norm divided by ℓ_1 of the degree heterogeneity vector restricted to one community. Let T_k denote the row of T corresponding to community k. Define new extended covariates as

$$\tilde{x}_i = \begin{cases} (x_i, \sqrt{\beta}T_{\ell(i)}), & i \in \mathcal{D}, \\ (x_i, 0), & i \notin \mathcal{D}. \end{cases}$$

Denote the matrix of extended covariates as $\tilde{X}$, and its expectation as $\tilde{E}(\tilde{X})$. Therefore, we have

$$\tilde{\Omega} = \{I_{\mathcal{D}}E(A)I_{\mathcal{D}} + I_{\mathcal{S}}D_{\alpha^*}\}\tilde{E}(\tilde{X}).$$

The definition of $\tilde{\Omega}$ is almost the same as Ω, except we replace $\tilde{E}(X)$ with the extended covariates $\tilde{E}(\tilde{X})$. For $\tilde{\Omega}$, we have $\tilde{\Omega}\tilde{\Omega}' = \Omega\Omega' + \beta n\{I_{\mathcal{D}}E(A)I_{\mathcal{D}}\}^2$, which concludes the oracle matrix corresponding to YY' and the dense communities in $\beta nAA'$.

Following the same proof, we have a generalized version of Lemma 1 for $\tilde{\Omega}$, where the spectral matrix of $\tilde{\Omega}\tilde{\Omega}'$ has identical rows for nodes in the same community, up to a constant factor. We further investigate the difference between $\hat{\Xi}$ and Ξ in terms of the Frobenius norm. Based on it, we prove the consistency of community detection by using $\hat{\Xi}$.

THEOREM 3 (CONSISTENCY OF ALGORITHM 2). *Consider the degree-corrected stochastic blockmodel with covariates and parameters* $(\Theta, K, P, \Pi, F_{[K]}, \mathcal{M})$. *Let* $\hat{\Xi}$ *be the matrix formed by the top K eigenvectors of* $YY' + \beta nAA'$ *in Algorithm 2 and* Ξ *be the eigenvectors of* $\tilde{\Omega}\tilde{\Omega}'$. *Suppose that conditions (i), (ii), (iv) of Theorem 1 and condition (iii') hold. Then there exist threshold constants* C_θ, ϵ_0, n_0 *and* δ_0 *such that if* $\delta_X \leqslant \delta_0, \epsilon \leqslant \epsilon_0, n \geqslant n_0, n\theta^2_{\max} \geqslant C_\theta \log n$, *there exist constants* β_0 *and* β_1 *such that, when* $\beta_0 < \beta < \beta_1$, *with probability* $1 - O(1/n)$, *there exists an orthogonal matrix O and a constant* $C > 0$ *such that*

$$\|\hat{\Xi} - O\Xi\|_F \leqslant C(\delta_X + \sqrt{\epsilon} + 1/\sqrt{C_\theta}).$$

Let

$$\delta_{\text{net}} = \frac{\max_{i\in\mathcal{S}} n\theta_i\theta_{\max}}{\min_{i\in\mathcal{D}} n\theta_i\theta_{\max}}, \qquad \text{Err}_n = n^{-1} \min_{\{\pi:\,[K]\to[K]\}} |\{i:\,\pi[\hat{\ell}(i)] \neq \ell(i)\}|.$$

Then there exists a permutation π *such that, with probability* $1 - O(1/n)$, *the clustering error rate by Algorithm 2 follows*

$$\text{Err}_n \leqslant C(\delta_X + 1/\sqrt{C_\theta} + \delta_{\text{net}} + \sqrt{\epsilon}).$$

Remark 1. Theorem 3 gives the weak consistency results of Algorithm 2. While both algorithms share a similar format for the oracle matrix, the noise caused by AA' is relatively large. It arises from the nature of the Bernoulli distribution. The large noise causes row-wise control of the empirical singular matrix to be hardly available, and hence strong consistency not achievable. It further provides the importance of introducing covariates into community detection.

Remark 2. The error control contains three parts: $\delta_X + 1/\sqrt{C_\theta}$ from the covariates, δ_{net} from the network and $\sqrt{\epsilon}$ from the misspecification. For a consistent signal recovery, we require the covariates to have relatively small deviations, the network to be sufficiently dense and the sparse communities to be well separated from the dense communities, with no misspecifications.

Remark 3. The tuning parameter β is a fixed constant. The requirement on β depends on $K_{\mathcal{S}}$, i.e., the number of sparse communities. When $K_{\mathcal{S}} = 0$, the information is in AA', so larger β is always preferred. Delicate analysis shows that $\beta \gtrsim cR^2$ is required, where c is a small constant based on $\delta_X, \delta_{\text{net}}, C_\theta, \epsilon$ and P. When $K_{\mathcal{S}} > 0$, β must be in an interval (β_0, β_1), so that $\beta nAA'$ is sufficiently large to detect dense communities, but not too large so that Y still works for the sparse communities. In the Supplementary Material, we find that the detailed requirement on β_0 and β_1, where both can be represented as $c\lambda^2_{K_{\mathcal{S}}}[I_{\mathcal{S}}\{\tilde{E}(X)\}]/n \approx cR^2$ by (iii'). For both cases, $\beta \approx cR^2$ works.

3.4. *Statistical lower bound on networks with covariates*

In this section, we demonstrate the information bound of the community detection problem on multiscale networks with covariates. We focus on the exact recovery problem. Our goal is to find the region of θ_i and F_k in which all estimators will fail.

To capture the effects of θ_i and F_k for a multiscale network, we need at least three communities: one of them is relatively dense and the other two are extremely sparse. If there is only one sparse community and one dense community then the labels can be recovered by the degree distribution, which is not of interest.

Consider a simplified model $\text{SM}(\theta_0, \theta_{\max}, P, \mu_{[K]}, \sigma)$ with $K = 3$ communities. Under this model, nodes fall into each community equally likely. Furthermore, there are only two possible values of $\theta_i \in \{\theta_0, \theta_{\max}\}$. Nodes in communities 1 and 2 have $\theta_i = \theta_0$ and nodes in community 3 have $\theta_i = \theta_{\max}$. Therefore, community 3 will be a dense community and communities 1 and 2 have the flexibility to be either dense or sparse. The covariates follow $x_i \sim \mathcal{N}(\mu_{\ell(i)}, \sigma^2 I_p)$. Here, σ^2 is to capture the deviation from the mean vector. We find the upper bound decided by our new algorithm and the statistical lower bound under model $\text{SM}(\theta_0, \theta_{\max}, P, \mu_{[3]}, \sigma_n)$.

THEOREM 4 (STATISTICAL LOWER BOUND). *Consider the simplified model with $K = 3$ communities, denoted* $\text{SM}(\theta_0, \theta_{\max}, P, \mu_{[3]}, \sigma_n)$. *There exists a constant $C > 0$ and a constant c_p on P such that if*

(i) $n\theta_0\theta_{\max} < Cc_p \log n$ *and*
(ii) $\|\mu_2 - \mu_1\|/\sigma_n < \sqrt{C \log n}$

then, for any estimator $\hat{\ell}$,

$$E\{\text{Err}_n(\hat{\ell}, \ell)\} \geqslant 1/n,$$

i.e., the exact recovery cannot be achieved.

By Theorem 4, to achieve the exact recovery, the communities either have a diverging degree as (i) or have well-separated covariates as (ii). When neither is met, all estimators will fail.

Remark 4 (*Connection to the upper bound*). Compare the lower bounds in Theorem 4 with the upper bounds in Theorem 2. The exact recovery can be guaranteed when all communities are dense or all nodes in the sparse communities have covariates $\|x_i - \tilde{E}(x_i)\| \leqslant \delta_X R$ for a small constant δ_X. Under $\mathrm{SM}(\theta_0, \theta_{\max}, P, \mu_{[3]}, \sigma_n)$, this means that either $n\theta_0\theta_{\max} \geqslant C \log n$ or $\|\mu_2 - \mu_1\|/\sigma_n \geqslant C\sqrt{\log n}$. Therefore, the upper bound in Theorem 2 meets the lower bound in Theorem 4, up to a constant factor. It supports the optimality of our spectral clustering on network-adjusted covariate approach.

4. Simulation

Consider a large network with $n = 1200$ nodes where the number of covariates can be either small ($p = 20$) or large ($p = 600$). We conduct three sets of simulation studies that focus on the effects on the estimation when (i) the network changes, (ii) the signal strength changes and (iii) the proportion of misspecified nodes changes.

We set up the baseline degree-corrected stochastic blockmodel with covariates as follows. The network has $K = 4$ communities and $\ell(i) = k$ with equal probability for $k = 1, 2, 3, 4$. The first two are dense communities with $\theta_i \sim \mathrm{Un}(0.3, 0.5)$ and the last two are sparse communities with $\theta_i \sim \mathrm{Un}(0.03, 0.05)$. The intensity matrix

$$P = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}.$$

The covariates follow a mixture of five distributions F_k, $k = 1, 2, 3, 4, 5$. Here F_5 only appears for some of the misspecified nodes. In detail, the covariates of node i follow the mixture distribution $x_i \sim (1-\gamma)F_{\ell(i)} + \sum_{k=1, k\neq\ell(i)}^{5}(\gamma/4)F_k$. Hence, γ represents the fraction of misspecified nodes. Half of these misspecified nodes belong to dense communities 1 and 2, so the bad nodes' proportion is about $\epsilon = |\mathcal{G}^c|/n = \gamma/2$.

We set $F_k \sim \mathcal{N}(m_k, I_p)$ for the covariates and discuss the setting of mean vectors m_k. In the large covariate case, $p = 600$, most covariates are noise with mean 0. We first select 5% of the p covariates as the useful covariates and then set the mean for those covariates as $m_k(j) \sim \mu_1 \times \mathrm{Ber}(1/2)$ independently. Therefore, $m_k(j)$ is different across communities k, which carries the label information. The leftover covariates have mean $m_k(j) = 0$. In the small covariate case, $p = 20$. Let $m_k(j) \sim \mu_2 + 0.1 \times \mathrm{Ber}(0.5)$ when $j \in \{5k-1, 5k-2, 5k-3, 5k-4\}$ and $m_k(j) \sim 0.1 \times \mathrm{Ber}(0.5)$ otherwise.

For each simulation study, we compare six different methods. The first four are popular community detection methods based on both the network and covariates: our newly proposed spectral clustering on network-adjusted covariate method, i.e., Algorithm 1, our new algorithm on generalized covariates, i.e., Algorithm 2, covariate-assisted spectral clustering (Binkiewicz et al., 2017) and semidefinite programming (Yan & Sarkar, 2021). We also consider the spectral method on the network only, in particular the spectral clustering with regularized Laplacian method of Joseph & Yu (2016). To examine the effects using only

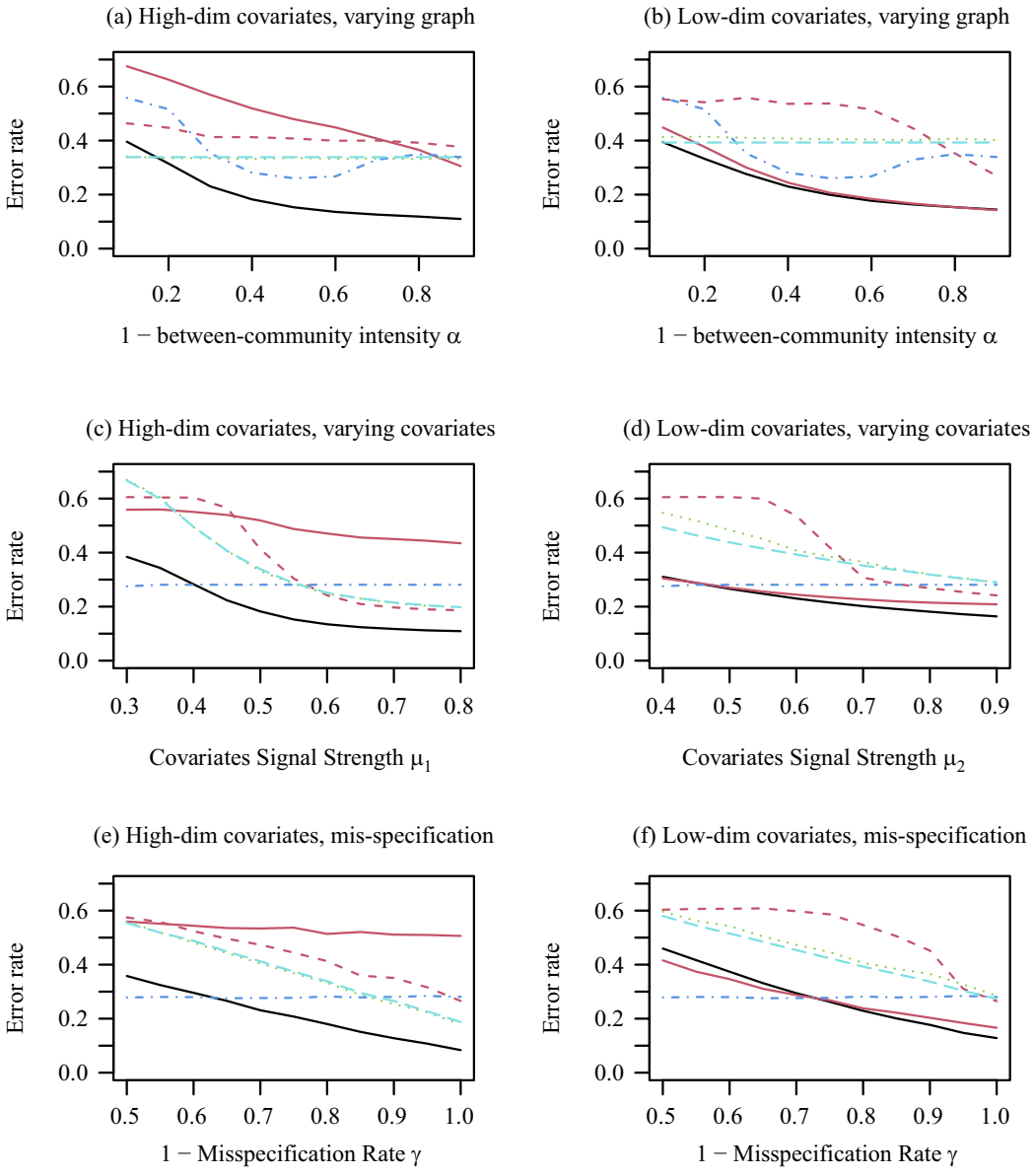

Fig. 1. The average clustering error rate among 50 repetitions of 6 community detection methods: (1) our method (black solid); (2) our method on generalized covariates (long dash); (3) covariate-assisted Laplacian (dashed); (4) semidefinite programming (dotted); (5) spectral clustering on regularized Laplacian (dotdash); and (6) spectral clustering on the covariate matrix (dot-dot-dot-dash). The fixed parameters are $n = 1200$, $\alpha = 0.4$, $\mu_1 = 0.5$, $\mu_2 = 0.8$ and $\gamma = 0.2$.

the covariate matrix, the last method is the spectral clustering method on XX' of Lee et al. (2010).

The first simulation study is on the community-by-community connection intensity matrix P. Let the between-community intensity $\alpha \in [0.1, 0.9]$. A smaller α will cause larger differences between the within-community connections and between-community connections, so that the community detection is easier. On the other hand, a very small α indicates many low-degree nodes in the sparse communities. Hence, there is a sweet spot of α when the community detection is purely based on the network, as shown in Figs. 1(a) and 1(b). Our

approach outperforms all other methods in most cases, except the semidefinite programming method and covariate-based method when the network information is very sparse.

The second simulation study is on the signal strength in covariates. We set the signal strength $\mu_1 \in [0.3, 0.8]$ for $p = 600$ and $\mu_2 \in [0.5, 1.0]$ for $p = 20$. Figures 1(c) and 1(d) record the results for $p = 600$ and $p = 20$, respectively. In general, increasing the signal strength can improve the community detection results for all methods including covariates. Among all methods, the new approach always performs the best, while the error rates of all other methods on covariates are at least 0.19 that may come from the 20% misspecified nodes.

In the third simulation study, we focus on the misspecification rate γ where $\gamma \in [0, 0.5]$; see the results in Figs. 1(e) and 1(f). When γ is large and the network adjacency conveys strong information, the low-dimensional case, the network-based method and covariate-assisted Laplacian work better. This is because a large misspecification will cause centres of community covariates to be close to each other. When γ decreases, our method works best. In particular, the error rate of our method decreases at rate $\gamma/2$, which is roughly $\epsilon = |\mathcal{G}^c|/n$. It provides numerical support that our method fails on the misspecified nodes with low degrees.

5. Real-world networks

5.1. *Error rates on the LastFM network*

The LastFM Asian dataset is a social network of LastFM app users. This dataset was collected and cleaned by Rozemberczki & Sarkar (2020) from the public API in March 2020. The nodes are the app users from 18 unspecified Asian countries and the connections between them are identified by the mutual friendship. In pre-processing, a small country with only 17 users is removed because of insufficient information and 17 countries are left. Each user has a list of liked artists as covariates. The goal is to estimate the country membership for each node.

We consider four subdatasets from the original dataset: a small country dataset (each with less than 100 users), a medium country dataset (each with users between 100 and 300), a large-sized country dataset (each with users between 300 and 1000) and a huge country dataset (each with more than 1000 users). Each dataset has communities with similar sizes, i.e., nondegenerate communities that are required by most existing community detection methods. For each dataset, we first select the regional popular artists by examining the artists with the largest proportion of fans in the countries of interest, and then we apply the five community detection methods.

The sizes of networks and covariates are summarized in Table 1, together with the average degree and community detection error rates of the methods. Our new method outperforms all other methods on three out of the four datasets. For the covariate-assisted Laplacian method, we selected the optimal tuning parameter among five choices according to Binkiewicz et al. (2017). Our network-adjusted covariate-based community detection method does not need any tuning parameter, with comparative clustering error rates.

In Table 1, the average degree is stable when the network size changes from 343 to 3691. This is often seen in real networks, and it motivates us to investigate sparse networks. When the network sizes are small and the average degree is relatively large, the net-based method performs well except for the low-degree nodes. Combining it with the covariates further reduces such errors; see our new methods and the covariate-assisted Laplacian method.

Table 1. *Community detection error rates of LastFM networks*

	n	p	K	$\bar{d}$	New	New (generalized)	CAL	SDP	Net based	Cov based
Small countries	343	194	6	4.9	0.236	0.262	0.178	0.510	0.350	0.557
Medium countries	612	324	3	7.2	0.041	0.031	0.044	0.342	0.165	0.234
Large countries	2488	600	5	6.2	0.249	0.424	0.371	0.519	0.482	0.468
Huge countries	3691	600	3	7.1	0.019	0.022	0.019	0.416	0.240	0.328

n, number of nodes; p, number of covariates; K, number of communities; $\bar{d}$, the average degree. New, our new approach; New (generalized), our new approach on generalized covariates; CAL, covariate-assisted Laplacian method; SDP, semidefinite programming method; Net based, spectral clustering on regularized Laplacian matrix; Cov based, spectral clustering on the covariate matrix.

Here, the covariate-assisted Laplacian method outperforms ours for the small country dataset because this dataset has severe degree heterogeneity within the same community and uninformative covariates. When it comes to large/huge countries, the network is relatively sparse. Our new method performs the best on these two datasets.

5.2. *Community detection on the statisticians' citation network*

The citation network was published in Ji & Jin (2016). It contains 3232 papers published in the *Annals of Statistics*, *Journal of American Statistical Association*, *Journal of Royal Statistical Society (Series B)* and *Biometrika* from 2003 to the first half of 2012. Each paper is a node and two papers are connected if they both cite the same paper or are both cited by another paper. The covariate of each paper is its abstract. This citation network has $n = 3232$ nodes and $p = 4095$ covariates.

In this network, there is a giant component of 2179 nodes. Among the leftover 1053 nodes, 957 are isolated and 96 nodes are in small components with the largest size being 9. When it was introduced by Ji & Jin (2016), the authors applied spectral clustering on ratios of eigenvectors (Jin, 2015) to the network. This method requires the network to be connected; hence, only the giant component was analysed. It suggests interpretable results on the high-degree nodes, yet the 1053 nodes outside the giant component were not classified.

We apply our new method, the variant of our new method on the generalized covariates, covariate-assisted Laplacian (Binkiewicz et al., 2017), the semidefinite programming method (Yan & Sarkar, 2021) and the network-based method (Joseph & Yu, 2016). For all the methods, we take $K = 5$ and record the community sizes in Table 2. For the low-degree nodes, the network-based method classifies all nodes into the largest community, which is unrealistic. This issue still exists, but is slightly resolved for our method on the generalized covariates and the covariate-assisted Laplacian method, because of their dependency on the network. Our original method, Algorithm 1, and the semidefinite programming method have a more reasonable and balanced splitting.

We then investigate the splitting of the giant component. We measure the agreement of each pair of clustering results by the normalized mutual information of Meilă (2007). A larger normalized mutual information score means that the two clustering results are more coherent. The heatmap of pairwise normalized mutual information scores is given in Fig. 2. Considering the giant component only, our new method agrees with the network-based method, while the semidefinite programming method does not agree with the network-based method at all. Combining the giant component and the low-degree nodes, our method provides the best splitting.

Table 2. *Estimated community sizes in the citation network and the number of the leftover 1053 nodes in each community (in brackets)*

	Community 1	Community 2	Community 3	Community 4	Community 5
New method	1105 (345)	1062 (386)	483 (181)	325 (120)	257 (21)
New (Algorithm 2)	2433 (1032)	228 (8)	225 (1)	180 (12)	166 (0)
SDP	783 (271)	720 (226)	665 (200)	660 (217)	404 (139)
CAL	1892 (1047)	471 (4)	357 (0)	319 (1)	193 (1)
Net based	2280 (1053)	297 (0)	283 (0)	221 (0)	151 (0)

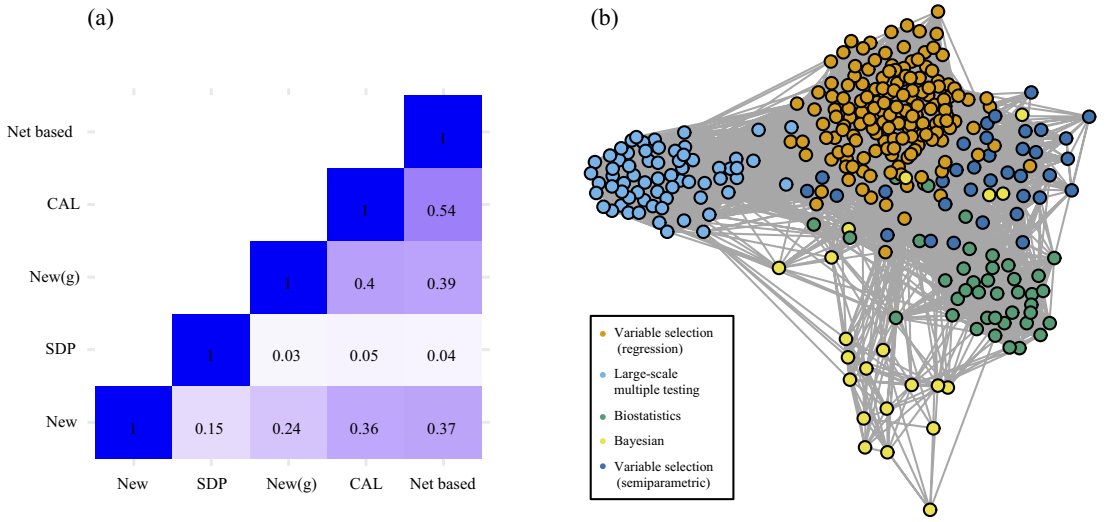

Fig. 2. (a) The heatmap of normalized mutual information scores between the five community detection methods on the giant component. (b) A subnetwork consisting of nodes with at least 50 neighbours. Each shape denotes one estimated community.

We compare the communities found using our spectral clustering on network-adjusted covariate method with the communities found in Ji & Jin (2016). For each community, we check the top 10 popular papers and corpus. The results can be interpreted as the variable selection (regression) community, the variable selection (semiparametric) community, the large-scale multiple testing community, the biostatistics community and the Bayesian community. Compared to the estimated communities of statisticians in Ji & Jin (2016), three communities are coherent: the large-scale multiple testing community, the biostatistics community and the Bayesian community (non-parametric community of Ji & Jin (2016)). The variable selection community in their work has been decomposed into two communities by our new method: one about regression and one about the semiparametric models. In Fig. 2, we can see that the regression community and the semiparametric community are densely connected, but have an even denser connection within communities.

The interpretation of communities works on the isolated nodes. We randomly checked two papers without any connections. Node 2893 titled 'Bayesian pseudo-empirical-likelihood intervals for complex surveys (Rao & Wu, 2010)' is classified into the Bayesian community. Node 2481 titled 'Testing dependence among serially correlated multicategory variables (Pesaran & Timmermann, 2009)' is classified into the multiple testing community. These two examples suggest that our method reasonably clustered isolated nodes.

Acknowledgement

This research was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (A-8001451-00-00). The authors thank Purnamrita Sarkar and Bowei Yan for generously sharing their code and Xin T. Tong, Yang Feng, Ramon van Handel, Jiashun Jin, Liza Levina, Anru Zhang and the referees for their valuable comments and discussions.

Supplementary material

The Supplementary material includes additional results on the statistician citation network, theoretical proofs of theorems and additional theoretical results. The R code is available at https://github.com/YaofangHuYaofang/NAC.

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* **18**, 6446–531.

Abbe, E., Bandeira, A. S. & Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Trans. Info. Theory* **62**, 471–87.

Abbe, E., Fan, J. & Wang, K. (2022). An ℓ_p theory of pca and spectral clustering. *Ann. Statist.* **50**, 2359–85.

Abbe, E., Fan, J., Wang, K. & Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48**, 1452–74.

Amini, A. A., Chen, A., Bickel, P. J. & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41**, 2097–122.

Bickel, P. J. & Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106**, 21068–73.

Binkiewicz, N., Vogelstein, J. T. & Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104**, 361–77.

Chaudhuri, K., Chung, F. & Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proc. 25th Ann. Conf. Learn. Theory*, vol. 23, pp. 35.1–35.23. Cambridge, MA: Proceedings of Machine Learning Research.

Chen, J. & Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics* **22**, 2283–90.

Chen, Y., Chi, Y., Fan, J. & Ma, C. (2021). Spectral methods for data science: a statistical perspective. *Foundat. Trends Mach. Learn.* **14**, 566–806.

Chung, F. R. & Graham, F. C. (1997). *Spectral Graph Theory*, vol. 92. Providence, RI: American Mathematical Society.

Deco, G. & Corbetta, M. (2011). The dynamical balance of the brain at rest. *Neuroscientist* **17**, 107–23.

Deshpande, Y., Sen, S., Montanari, A. & Mossel, E. (2018). Contextual stochastic block models. In *Proc. 32nd Int. Conf. Neural Info. Proces. Syst.*, pp. 8590–602. Red Hook, NY: Curran Associates.

Fan, J., Wang, W. & Zhong, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18**, 1–42.

Gao, C., Ma, Z., Zhang, A. Y. & Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18**, 1980–2024.

Gil-Mendieta, J. & Schmidt, S. (1996). The political network in Mexico. *Social Networks* **18**, 355–81.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm as 136: a *k*-means clustering algorithm. *Appl. Statist.* **28**, 100–8.

Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5**, 109–37.

Hu, Y., Wang, W. & Yu, Y. (2022). Graph matching beyond perfectly-overlapping Erdős–Rényi random graphs. *Statist. Comp.* **32**, 1–16.

Huang, S. & Feng, Y. (2023). Pairwise covariates-adjusted block model for community detection. *arXiv:* 1807.03469v5.

Jacob, U., Thierry, A., Brose, U., Arntz, W. E., Berg, S., Brey, T., Fetzer, I., Jonsson, T., Mintenbeck, K., Möllmann, C. et al. (2011). The role of body size in complex food webs: a cold case. *Adv. Ecol. Res.* **45**, 181–223.

Ji, P. & Jin, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Statist.* **10**, 1779–812.

Jin, J. (2015). Fast community detection by score. *Ann. Statist.* **43**, 57–89.

Jin, J., Ke, Z. T. & Wang, W. (2017). Phase transitions for high dimensional clustering and related problems. *Ann. Statist.* **45**, 2151–89.
Jin, J. & Wang, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44**, 2323–59.
Joseph, A. & Yu, B. (2016). Impact of regularization on spectral clustering. *Ann. Statist.* **44**, 1765–91.
Karrer, B. & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.
Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L. & Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Nat. Acad. Sci.* **110**, 20935–40.
Lee, A. B., Luca, D., Klei, L., Devlin, B. & Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. *Genet. Epidem.* **34**, 51–9.
Lei, J. & Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43**, 215–37.
Lei, L., Li, X. & Lou, X. (2021). Consistency of spectral clustering on hierarchical stochastic block models. *arXiv:* 2004.14531v2.
Leskovec, J. & McAuley, J. (2012). Learning to discover social circles in ego networks. In *Proc. 25th Int. Conf. Neural Info. Proces. Syst.*, pp. 539–47. Red Hook, NY: Curran Associate.
Ma, Z. & Nandy, S. (2023). Community detection with contextual multilayer networks. *IEEE Trans. Info. Theory* **69**, 3203–39.
Meilă, M. (2007). Comparing clusterings - an information based distance. *J. Mult. Anal.* **98**, 873–95.
Newman, M. E. & Clauset, A. (2016). Structure and inference in annotated networks. *Nature Commun.* **7**, 1–11.
Pesaran, M. H. & Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *J. Am. Statist. Assoc.* **104**, 325–37.
Rao, J. & Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *J. R. Statist. Soc* B **72**, 533–44.
Rohe, K., Chatterjee, S. & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878–915.
Rozemberczki, B. & Sarkar, R. (2020). Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In *Proc. 29th ACM Int. Conf. Info. Know. Manag.*, pp. 1325–34. New York: Association for Computing Machinery.
Sporns, O. & Betzel, R. F. (2016). Modular brain networks. *Ann. Rev. Psychol.* **67**, 613–40.
Su, L., Wang, W. & Zhang, Y. (2019). Strong consistency of spectral clustering for stochastic block models. *IEEE Trans. Info. Theory* **66**, 324–38.
Weng, H. & Feng, Y. (2022). Community detection with nodal information: likelihood and its variational approximation. *Stat* **11**, e428.
Xu, S., Zhen, Y. & Wang, J. (2023). Covariate-assisted community detection in multi-layer networks. *J. Bus. Econ. Statist.* **41**, 915–26.
Yan, B. & Sarkar, P. (2021). Covariate regularized community detection in sparse graphs. *J. Am. Statist. Assoc.* **116**, 734–45.
Yan, T., Jiang, B., Fienberg, S. E. & Leng, C. (2019). Statistical inference in a directed network model with covariates. *J. Am. Statist. Assoc.* **114**, 857–68.
Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P. & Zhu, Y. (2014). Model selection for degree-corrected block models. *J. Statist. Mech.* **2014**, P05007.
Yang, J., McAuley, J. & Leskovec, J. (2013). Community detection in networks with node attributes. In *2013 IEEE Int. Conf. Data Mining*, pp. 1151–6. Piscataway, NJ: IEEE Press.
Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L. & Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proc. 24th ACM SIGKDD Int. Conf. Know. Disc. Data Mining*, pp. 974–83. New York: Association for Computing Machinery.
Zhang, Y., Levina, E. & Zhu, J. (2016). Community detection in networks with node features. *Electron. J. Statist.* **10**, 3153–78.
Zhao, Y., Levina, E. & Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40**, 2266–92.

[*Received on* 26 *April* 2023. *Editorial decision on* 14 *February* 2024]