Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Covariate-Assisted Community Detection in Multi-Layer Networks

Shirong Xu, Yaoming Zhen & Junhui Wang

To cite this article: Shirong Xu, Yaoming Zhen & Junhui Wang (2023) Covariate-Assisted Community Detection in Multi-Layer Networks, Journal of Business & Economic Statistics, 41:3, 915-926, DOI: 10.1080/07350015.2022.2085726

To link to this article: https://doi.org/10.1080/07350015.2022.2085726

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2023, VOL. 41, NO. 3, 915–926
https://doi.org/10.1080/07350015.2022.2085726

Covariate-Assisted Community Detection in Multi-Layer Networks

Shirong Xu[a,b], Yaoming Zhen[a], and Junhui Wang[a]

[a]School of Data Science, City University of Hong Kong, Kowloon, Hong Kong; [b]Krannert School of Management, Purdue University, West Lafayette, IN

ABSTRACT

Communities in multi-layer networks consist of nodes with similar connectivity patterns across all layers. This article proposes a tensor-based community detection method in multi-layer networks, which leverages available node-wise covariates to improve community detection accuracy. This is motivated by the network homophily principle, which suggests that nodes with similar covariates tend to reside in the same community. To take advantage of the node-wise covariates, the proposed method augments the multi-layer network with an additional layer constructed from the node similarity matrix with proper scaling, and conducts a Tucker decomposition of the augmented multi-layer network, yielding the spectral embedding vector of each node for community detection. Asymptotic consistencies of the proposed method in terms of community detection are established, which are also supported by numerical experiments on various synthetic networks and two real-life multi-layer networks.

ARTICLE HISTORY
Received November 2021
Accepted May 2022

KEYWORDS
Community detection; Multi-layer network; Network homophily; Stochastic block model; Tensor decomposition

1. Introduction

Network structure has been widely applied to represent the relationship among a vast number of entities, which finds applications in various domains, ranging from social networks (Du et al. 2007; Leskovec, Lang, and Mahoney 2010), biological networks (Rahiminejad, Maurya, and Subramaniam 2019; Calderer and Kuijjer 2021), to scientific networks (Jung and Segev 2014; Gao et al. 2021). It is also interesting to note that nodes commonly interact with each other from different aspects, leading to multi-layer networks. Examples of multi-layer networks include multi-task brain connectivity patterns (Cole et al. 2013), world import and export trades with different goods (Yuan and Qu 2021; Jing et al. 2021), and social networks with different interaction channels (Borondo et al. 2015; Zhang, Xue, and Zhu 2020).

One of the key challenges in multi-layer network is to integrate information across different network layers to identify nodes with similar connectivity patterns for homogeneous communities. To circumvent this difficulty, various efforts have been devoted to developing community detection methods for multi-layer networks, including spectral clustering techniques on the aggregated network (Han, Xu, and Airoldi 2015; Paul and Chen 2020b; Lei and Lin 2022), latent-position-based likelihood estimation (Wang et al. 2019; Zhang, Xue, and Zhu 2020; Lyu, Xia, and Zhang 2022; Macdonald, Levina, and Zhu 2022), least square estimation (Lei, Chen, and Lynch 2020), and multi-layer network modularity (Wilson et al. 2017; Paul and Chen 2021). Furthermore, detecting heterogeneous community structures in different network layers has also been studied in the literature (Paul and Chen 2020a; Jing et al. 2021; Chen, Liu, and Ma 2022).

In practice, the observed networks are often accompanied by node-wise covariates, which may provide supplementary information in revealing nodes' community structure following the well-accepted network homophily principle (McPherson, Smith-Lovin, and Cook 2001; Krivitsky et al. 2009). Some recent efforts has paid to integrate node-wise covariate to help community detection in both single-layer or multi-layer networks. For single-layer networks, Zhang, Levina, and Zhu (2016) employs node-wise covariate to adjust network modularity, Binkiewicz, Vogelstein, and Rohe (2017) converts covariate information into the graph Laplacian, Yan and Sarkar (2021) proposes a covariate regularized positive semidefinite programming approach, and Zhao et al. (2022) leverages the network data to reduce the dimension of node-wise covariates for subsequent clustering. For multi-layer networks, Contisciani, Power, and De Bacco (2020) maximizes the combined likelihood of both multi-layer network and node-wise covariates, Zhang, Xu, and Zhu (2022) proposes a joint latent space model for multi-layer networks and node-wise covariates, and Ma and Nandy (2021) investigates the information-theoretic limit achieved by the Bayes optimal estimator based on the joint signal to noise ratio of multi-layer networks with covariates.

In this article, we construct a covariate-augmented multi-layer network adjacency tensor for community detection. Specifically, the multi-layer network is augmented with an additional network layer representing the nodes' similarities based on nodal covariate information. A Tucker decomposition is then implemented on the augmented network adjacency tensor to obtain the spectral embeddings of nodes, which integrate information from both multi-layer networks and node-wise covariates. This nodes' embedding matrix is subsequently used as an input for an $(1+\epsilon)$-optimal K-means algorithm to detect the community structure. In addition, a scaling factor is introduced to strike a balance between the multi-layer network

CONTACT Junhui Wang *j.h.wang@cityu.edu.hk* School of Data Science, City University of Hong Kong, Kowloon, Hong Kong.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

and node-wise covariates for better community detection accuracy. This method is particularly attractive when the multi-layer network is relatively sparse or less informative while the community structure is transparently encoded in the node-wise covariates.

Also, we have conducted a thorough analysis of the asymptotic behavior of the proposed method, leading to its several attractive theoretical advantages. First, if the node-wise covariates in different communities are asymptotically distinguishable, Corollary 1 shows that the community detection consistency in multi-layer networks can be attained with much smaller sparsity factors than the existing results in literature without considering the node-wise covariates (Paul and Chen 2016, 2020b; Lei, Chen, and Lynch 2020; Lei and Lin 2022). On the contrary, when node-wise covariates do not contain any information about the community structure, Corollary 2 still implies the community detection consistency as long as the sparsity factor $s_n \gg \frac{\log n}{nL}$ with n and L being the number of nodes and network layers, which matches up with the best existing results in literature (Jing et al. 2021) when the Tucker rank of the expected adjacency tensor is assumed to be of constant order. Second, instead of assuming a uniform network sparsity factor and asymptotically balanced community sizes (Lei, Chen, and Lynch 2020; Jing et al. 2021), Theorem 1 establishes the community detection consistency of the proposed method, allowing for varying sparsity factors across different layers and asymptotically unbalanced communities.

The rest of the article is organized as follows. After introducing some necessary notations in Section 1.1, Section 2 lays out the background of multi-layer networks with community structure and proposes a covariate-assisted community detection method for multi-layer networks. Section 3 establishes the theoretical guarantees of the proposed method and explicitly quantifies how the node-wise covariates interplay with the multi-layer network. In Section 4, we conduct extensive numerical experiments to illustrate the superior performance of the proposed method in various synthetic networks, a business relation network, and a citation network on statisticians, and a summary is provided in Section 5. All technical proofs and necessary lemmas are relegated to the Appendix.

1.1. Notations

For a positive integer n, denote $[n] = \{1, \ldots, n\}$ to be the n-set. For two sequences f_n and g_n, $f_n = O(g_n)$ if $\lim_{n\to+\infty} \sup |f_n|/g_n < +\infty$, $f_n = o(g_n)$ if $\lim_{n\to+\infty} |f_n|/g_n = 0$, $f_n = \Omega(g_n)$ if $\lim_{n\to+\infty} \sup |f_n|/g_n > 0$, $f_n = \Theta(g_n)$ if $f_n = O(g_n)$ and $f_n = \Omega(g_n)$, and $f_n \gg g_n$ if $\lim_{n\to+\infty} f_n/g_n = +\infty$. Let $\|\cdot\|$ denote the l_2-norm of a vector or the spectral norm of a matrix, $\|\cdot\|_\infty$ denote the l_∞-norm of a vector or the vectorization of the input argument, and $\|\cdot\|_F$ denote the Frobenius norm of a matrix or tensor, and $\boldsymbol{A} \otimes \boldsymbol{B}$ denote the Kronecker product between two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

For a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times I_3}$, denote $\boldsymbol{\mathcal{A}}_{i_1,:,:} \in \mathbb{R}^{I_2\times I_3}$, $\boldsymbol{\mathcal{A}}_{:,i_2,:} \in \mathbb{R}^{I_1\times I_3}$ and $\boldsymbol{\mathcal{A}}_{:,:,i_3} \in \mathbb{R}^{I_1\times I_2}$ by the i_1th horizontal, i_2th lateral, and i_3th frontal slide of $\boldsymbol{\mathcal{A}}$, respectively. In addition, denote $\boldsymbol{\mathcal{A}}_{:,i_2,i_3} \in \mathbb{R}^{I_1}$, $\boldsymbol{\mathcal{A}}_{i_1,:,i_3} \in \mathbb{R}^{I_2}$, and $\boldsymbol{\mathcal{A}}_{i_1,i_2,:} \in \mathbb{R}^{I_3}$ as the (i_2, i_3)th mode-1, (i_1, i_3)th mode-2 and (i_1, i_2)th mode-3 fiber of $\boldsymbol{\mathcal{A}}$, respectively. For $j \in [3]$, let $\boldsymbol{\mathcal{M}}_j(\boldsymbol{\mathcal{A}})$ be the mode-j matricization of $\boldsymbol{\mathcal{A}}$ (Kolda and Bader 2009), which unfolds $\boldsymbol{\mathcal{A}}$ by concatenating its mode-j fibers of $\boldsymbol{\mathcal{A}}$ horizontally. Specifically, $\boldsymbol{\mathcal{M}}_j(\boldsymbol{\mathcal{A}})$ is a matrix in $\mathbb{R}^{I_j\times\prod_{i\neq j} I_i}$ such that

$$\boldsymbol{\mathcal{A}}_{i_1,i_2,i_3} = \big(\boldsymbol{\mathcal{M}}_j(\boldsymbol{\mathcal{A}})\big)_{i_j,m}, \text{ with } m = 1 + \sum_{\substack{l=1\\ l\neq j}}^{3} (i_l - 1) \prod_{\substack{i=1\\ i\neq j}}^{l-1} I_i.$$

For some matrices $\boldsymbol{M}^{(1)} \in \mathbb{R}^{J_1\times I_1}$, $\boldsymbol{M}^{(2)} \in \mathbb{R}^{J_2\times I_2}$, $\boldsymbol{M}^{(3)} \in \mathbb{R}^{J_3\times I_3}$, the mode-1 product between $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{M}^{(1)}$ is a $J_1 \times I_2 \times I_3$ tensor which is defined as $(\boldsymbol{\mathcal{A}} \times_1 \boldsymbol{M}^{(1)})_{j_1,i_2,i_3} = \sum_{i_1=1}^{I_1} \boldsymbol{\mathcal{A}}_{i_1,i_2,i_3} \boldsymbol{M}^{(1)}_{j_1,i_1}$, for $j_1 \in [J_1]$, $i_2 \in [I_2]$, and $i_3 \in [I_3]$. The mode-2 product $\boldsymbol{\mathcal{A}} \times_2 \boldsymbol{M}^{(2)} \in \mathbb{R}^{I_1\times J_2\times I_3}$ and mode-3 product $\boldsymbol{\mathcal{A}} \times_3 \boldsymbol{M}^{(3)} \in \mathbb{R}^{I_1\times I_2\times J_3}$ are defined in a similar fashion. The Tucker rank, also known as multi-linear rank, of $\boldsymbol{\mathcal{A}}$ is defined as (r_1, r_2, r_3), where $r_1 = \text{rank}(\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}))$, $r_2 = \text{rank}(\boldsymbol{\mathcal{M}}_2(\boldsymbol{\mathcal{A}}))$ and $r_3 = \text{rank}(\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{A}}))$. Further, if $\boldsymbol{\mathcal{A}}$ has Tucker rank (r_1, r_2, r_3), then $\boldsymbol{\mathcal{A}}$ admits the following Tucker decomposition,

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{U} \times_2 \boldsymbol{V} \times_3 \boldsymbol{W},$$

where $\boldsymbol{\mathcal{C}}) \in \mathbb{R}^{r_1\times r_2\times r_3}$ is a core tensor and $\boldsymbol{U} \in \mathbb{R}^{I_1\times r_1}$, $\boldsymbol{V} \in \mathbb{R}^{I_2\times r_2}$ and $\boldsymbol{W} \in \mathbb{R}^{I_3\times r_3}$ have orthonormal columns.

2. Proposed Method

2.1. Multi-Layer Stochastic Block Model

Consider an L-layer network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [n]$ denotes the set of n nodes, $\mathcal{E} = \{\boldsymbol{E}^{(l)}\}_{l=1}^{L}$ denotes the edge sets for all L layers, and $\boldsymbol{E}^{(l)} \subset \mathcal{V}\times\mathcal{V}$ can be equivalently represented by an adjacency matrix $\boldsymbol{A}^{(l)} \subset \{0,1\}^{n\times n}$ with $\boldsymbol{A}^{(l)}_{i,j} = \boldsymbol{A}^{(l)}_{j,i} = 1$ if $(i,j) \in \boldsymbol{E}^{(l)}$ and $\boldsymbol{A}^{(l)}_{i,j} = \boldsymbol{A}^{(l)}_{j,i} = 0$ otherwise, for $i \le j$. Denote $\boldsymbol{P}^{(l)}$ as the probability matrix in the lth layer such that $\boldsymbol{P}^{(l)}_{i,j} = P(\boldsymbol{A}^{(l)}_{i,j} = 1)$ is the probability that there exists an edge between nodes i and j in the lth layer. We further assume that $\boldsymbol{P}^{(l)}$ follows the multi-layer stochastic block model (MSBM; Paul and Chen 2016) that

$$\boldsymbol{P}^{(l)}_{i,j} = \boldsymbol{B}^{(l)}_{c_i,c_j}, \text{ for } i,j \in [n] \text{ and } l \in [L],$$

where $c_i \in [K]$ denotes the community assignment of node i, and $\boldsymbol{B}^{(l)}_{c_i,c_j}$ is the edge probability between communities c_i and c_j in the lth layer. It is important to remark that community assignment c_i is homogeneous across all layers, whereas the linking probability matrix $\boldsymbol{B}^{(l)}$ may vary from one layer to another.

Let $\boldsymbol{Z} \in \{0,1\}^{n\times K}$ be the nodes' community membership matrix. Each row of $\boldsymbol{Z}$ has exactly one element being 1, indicating its associated community membership. Therefore, MSBM can be reformulated in a matrix form that $\boldsymbol{P}^{(l)} = \boldsymbol{Z}\boldsymbol{B}^{(l)}\boldsymbol{Z}^T$, for $l \in [L]$.

Moreover, the multi-layer network $\mathcal{G}$ can be represented by a third order adjacency tensor $\boldsymbol{\mathcal{A}} \in \{0,1\}^{n\times n\times L}$ with $\boldsymbol{\mathcal{A}}_{:,:,l} = \boldsymbol{A}^{(l)}$, for $l \in [L]$, whose associated underlying probability tensor can be denoted as

$$\boldsymbol{\mathcal{P}}) = \boldsymbol{\mathcal{B}}) \times_1 \boldsymbol{Z} \times_2 \boldsymbol{Z},$$

where $\boldsymbol{\mathcal{P}})_{:,:,l} = \boldsymbol{P}^{(l)} = \boldsymbol{Z}\boldsymbol{B}^{(l)}\boldsymbol{Z}^T$, and $\boldsymbol{\mathcal{B}}) \in [0,1]^{K\times K\times L}$ with $\boldsymbol{\mathcal{B}})_{:,:,l} = \boldsymbol{B}^{(l)}$, for $l \in [L]$.

2.2. Covariate-Assisted MSBM

In multi-layer network data, the node-wise covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are usually available, where $\boldsymbol{x}_i \in \mathbb{R}^p$ denotes the attributes of node i. The network homophily principle implies that nodes sharing similar covariates tend to reside in the same community. It leads to the following generative model for the node-wise covariates,

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{M} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$, each row of $\boldsymbol{M}$ denotes the mean vector of the node-wise covariates within a community, and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n]^T \in \mathbb{R}^{n\times p}$ denotes the zero-mean noise of the covariate matrix $\boldsymbol{X}$. We further assume each $\boldsymbol{\epsilon}_i$ are independently generated from a certain distribution with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p} \in \mathbb{R}^{p\times p}$. This generative model is ubiquitous in the literature of covariate-assisted community detection in network data (Binkiewicz, Vogelstein, and Rohe 2017; Yan and Sarkar 2021).

To couple MSBM with the node-wise covariates, we consider an augmented adjacency tensor $\widetilde{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{n\times n\times(L+1)}$, with

$$\widetilde{\boldsymbol{\mathcal{A}}}_{:,:,l} = \begin{cases} \boldsymbol{\mathcal{A}}_{:,:,l}, \text{ for } l \in [L]; \\ \alpha_n \boldsymbol{X}\boldsymbol{X}^T, \text{ for } l = L+1, \end{cases}$$

where α_n is a tuning parameter controlling the tradeoff between the multi-layer network and node-wise covariates. Note that $\mathbb{E}(\boldsymbol{\mathcal{A}}_{:,:,l}) = \boldsymbol{Z}\boldsymbol{B}_{:,:,l}\boldsymbol{Z}^T$ for $l \in [L]$ and $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T) = \boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T + (\sum_{i=1}^p \sigma_{ii})\boldsymbol{I}_n$ with $\boldsymbol{I}_n$ being an n-dimensional identity matrix, so the underlying probability tensor for $\widetilde{\boldsymbol{\mathcal{A}}}$ also extends to $\widetilde{\boldsymbol{\mathcal{P}}} \in \mathbb{R}^{n\times n\times(L+1)}$, with

$$\widetilde{\boldsymbol{\mathcal{P}}}_{:,:,l} = \begin{cases} \boldsymbol{Z}\boldsymbol{B}_{:,:,l}\boldsymbol{Z}^T, \text{ for } l \in [L]; \\ \alpha_n \boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T, \text{ for } l = L+1. \end{cases}$$

It is clear that $\widetilde{\boldsymbol{\mathcal{P}}}$ can be formulated as $\widetilde{\boldsymbol{\mathcal{P}}} = \widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{Z} \times_2 \boldsymbol{Z}$, with $\widetilde{\boldsymbol{\mathcal{B}}}_{:,:,l} = \boldsymbol{\mathcal{B}}_{:,:,l}$ if $l \in [L]$ and $\alpha_n \boldsymbol{M}\boldsymbol{M}^T$ otherwise. We then term the proposed model as covariate-assisted MSBM (CAMSBM). Essentially, the proposed CAMSBM integrates the signal strength of both multi-layer network and node-wise covariates, which is of great interest for the downstream community detection analysis, especially when the community structures encoded in the multi-layer network and the node-wise covariates are complementary to each other.

It is important to point out that $\widetilde{\boldsymbol{\mathcal{P}}}$ admits a standard Tucker decomposition with orthonormal factor matrices. Specifically, let (K, K, L_0) be the Tucker rank of $\widetilde{\boldsymbol{\mathcal{P}}}$ and $\boldsymbol{\Gamma} = \text{diag}(\sqrt{n_1}, \ldots, \sqrt{n_K})$ with $n_k = \sum_{i=1}^n \boldsymbol{Z}_{ik}$ being the size of the kth community, for $k \in [K]$. The definition of the Tucker rank immediately implies that $L_0 \le \min\{K(K+1)/2, L+1\}$. Simple algebra yields that

$$\begin{aligned} \widetilde{\boldsymbol{\mathcal{P}}} &= \left(\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma}\right) \times_1 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1} \times_2 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1} \\ &= (\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O} \times_3 \boldsymbol{V}) \times_1 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1} \times_2 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1} \\ &= \boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{O} \times_2 \boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{O} \times_3 \boldsymbol{V}, \end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O} \times_3 \boldsymbol{V}$ is the Tucker decomposition of $\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma}$, with a core tensor $\boldsymbol{\mathcal{C}}) \in \mathbb{R}^{K\times K\times L_0}$ and orthonormal factor matrices $\boldsymbol{O}$ and $\boldsymbol{V}$. It thus clear that $\boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{O}$ also has orthonormal columns, and (1) gives the standard Tucker decomposition of $\widetilde{\boldsymbol{\mathcal{P}}}$.

To estimate $\boldsymbol{Z}$ from $\widetilde{\boldsymbol{\mathcal{A}}}$ that is generated from the proposed CAMSBM model, we implements a Tucker approximation on $\widetilde{\boldsymbol{\mathcal{A}}}$ with Tucker rank (K, K, L_0),

$$\widetilde{\boldsymbol{\mathcal{A}}} \approx \widehat{\boldsymbol{\mathcal{C}}} \times_1 \widehat{\boldsymbol{U}} \times_2 \widehat{\boldsymbol{U}} \times_3 \widehat{\boldsymbol{V}},$$

where $\widehat{\boldsymbol{U}} \in \mathbb{R}^{n\times K}$ is any matrix having orthonormal columns and its column space is spanned by the first K leading left singular vectors of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}})$, $\widehat{\boldsymbol{V}} \in \mathbb{R}^{(L+1)\times L_0}$ is any matrix having orthonormal columns and its column space is spanned by the first L_0 leading left singular vectors of $\boldsymbol{\mathcal{M}}_3(\widetilde{\boldsymbol{\mathcal{A}}})$, and $\widehat{\boldsymbol{\mathcal{C}}} = \widetilde{\boldsymbol{\mathcal{A}}} \times_1 \widehat{\boldsymbol{U}}^T \times_2 \widehat{\boldsymbol{U}}^T \times_3 \widehat{\boldsymbol{V}}^T$. The Tucker approximation on $\widetilde{\boldsymbol{\mathcal{A}}}$ can be obtained by various algorithms, including the higher order singular value decomposition (HOSVD; Kolda and Bader 2009), the higher order orthogonal iteration (HOOI) (De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009) and the regularized HOOI (Jing et al. 2021) algorithm. In our algorithm, the Tucker decomposition is implemented by the HOOI algorithm with HOSVD as an initializer in the Python package `Tensorly` (Kossaifi et al. 2019), which guarantees a fast convergence rate of $\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^T - \boldsymbol{U}\boldsymbol{U}^T\|$ (Ke, Shi, and Xia 2019).

Subsequently, an $(1+\epsilon)$-optimal K-means algorithm is applied to $\widehat{\boldsymbol{U}}$ to estimate the nodes' community memberships. Specifically, it searches for a solution $(\widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{W}})$ such that

$$\|\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{W}} - \widehat{\boldsymbol{U}}\|_F^2 \le (1+\epsilon) \min_{\boldsymbol{Z}\in\boldsymbol{\Delta}, \boldsymbol{W}\in\mathbb{R}^{K\times K}} \|\boldsymbol{Z}\boldsymbol{W} - \widehat{\boldsymbol{U}}\|_F^2,$$

where $\boldsymbol{\Delta} \subset \{0,1\}^{n\times K}$ is the set of all possible community membership matrices with exactly one 1 in each row, and ϵ serves as the statistical-computational gap from the estimator $(\widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{W}})$ to the global optimizer of the K-means algorithm. Clearly, the $(1+\epsilon)$-optimal K-means algorithm provides a feasible approximation to the K-means algorithm in that it reduces to K-means algorithm as ϵ approaches 0. Similar treatment has also been employed in Lei and Rinaldo (2015).

The performance of the proposed method relies on the appropriate choice of α_n, which balances information of the multi-layer network and node-wise covariates. It can be done by following the guideline from the theoretical order in Section 3. Throughout the article, we assume K and L_0 are prespecified, otherwise the Eigen-gaps of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}})$ and $\boldsymbol{\mathcal{M}}_3(\widetilde{\boldsymbol{\mathcal{A}}})$ can be employed to estimate the unknown K and L_0 as in Ke, Shi, and Xia (2019).

3. Theory

In this section, we establish the consistency of the proposed method in terms of community detection, which also explicitly quantifies the impact of the node-wise covariates on the multi-layer network as well as network sparsity.

To assess the community detection performance, we employ the Hamming error of $\hat{\boldsymbol{c}}$ (Jin 2015; Jing et al. 2021; Zhen and Wang 2022), which is defined as the minimum scaled Hamming distance between $\hat{\boldsymbol{c}}$ and $\boldsymbol{c}^*$ under permutations. Formally,

$$\text{Err}(\widehat{\boldsymbol{c}}, \boldsymbol{c}^*) = \min_{\boldsymbol{\pi}\in\mathbf{S}_K} \frac{1}{n}\sum_{i=1}^n I(c_i^* = \boldsymbol{\pi}(\widehat{c}_i)), \tag{2}$$

where $\boldsymbol{c}^* = (c_1^*, c_2^*, \ldots, c_n^*)$ denotes the true community membership, $\widehat{\boldsymbol{c}} = (\widehat{c}_1, \ldots, \widehat{c}_n)$ denotes the estimated community membership, $I(\cdot)$ is the indicator function, and $\mathbf{S}_K$ is the symmetric group of degree K. Clearly, (2) measures the minimum fraction of nodes with inconsistent assignments between $\widehat{\boldsymbol{c}}$ and $\boldsymbol{c}^*$ over all possible permutations of community assignments.

Denote $s_n^{(l)}$ as the sparsity factor for the lth network layer such that $\boldsymbol{B}^{(l)} = s_n^{(l)} \boldsymbol{B}^{(l,0)}$ with $\boldsymbol{B}^{(l,0)}_{k_1,k_2} = \Theta(1)$, for $k_1, k_2 \in [K]$ and $l \in [L]$. Also, denote $s_n^{(L+1)} = \alpha_n$, and denote $s_n^{[L_0]}$ as the L_0th largest value among $\{s_n^{(l)}\}_{l=1}^{L+1}$. In the sequel, we will use $s_n^{(L+1)}$ and α_n interchangeably where no confusion is caused.

Assumption A. Assume that the core tensor $\widetilde{\boldsymbol{\mathcal{B}}}$ in the CAMSBM model satisfies that

$$\sigma_{\min}\Big(\boldsymbol{\mathcal{M}}_3\big(\widetilde{\boldsymbol{\mathcal{B}}}\big)\Big) = \Omega\big(\sqrt{L} s_n^{[L_0]}\big),$$

where $\sigma_{\min}(\cdot)$ denotes the smallest nonzero singular value of a matrix.

Assumption A characterizes the required signal strength for $\widetilde{\boldsymbol{\mathcal{B}}}$, which plays an important role in deriving the incoherence property of the factor matrices obtained from the Tucker decomposition of $\boldsymbol{\mathcal{P}}$). Note that $\boldsymbol{\mathcal{M}}_3\big(\widetilde{\boldsymbol{\mathcal{B}}}\big)$ can be factorized as $\mathrm{diag}(s_n^{(1)}, \ldots, s_n^{(L+1)})\boldsymbol{H}$, where $\boldsymbol{H} \in \mathbb{R}^{(L+1)\times K^2}$ is a matrix made up from $\boldsymbol{B}^{(1,0)}, \ldots, \boldsymbol{B}^{(L,0)}$ and $\boldsymbol{M}\boldsymbol{M}^T$. Essentially, Assumption A assumes the smallest nonzero singular value of $\boldsymbol{H}$ is of an asymptotic order $\Omega(\sqrt{L})$. This assumption is mild and can be satisfied with high probability if the entries of $\boldsymbol{H}$ are independently and identically distributed from some sub-Gaussian distributions (Rudelson and Vershynin 2009). Moreover, since the rank of $\boldsymbol{\mathcal{M}}_3\big(\widetilde{\boldsymbol{\mathcal{B}}}\big)$ is L_0, $\sigma_{\min}\Big(\boldsymbol{\mathcal{M}}_3\big(\widetilde{\boldsymbol{\mathcal{B}}}\big)\Big)$ is the same as the L_0th largest singular value of $\boldsymbol{\mathcal{M}}_3\big(\widetilde{\boldsymbol{\mathcal{B}}}\big)$, which scales linearly with $s_n^{[L_0]}$ when $\boldsymbol{H}$ is well-conditioned. Similar assumptions have also been made in Jing et al. (2021) and Lyu, Xia, and Zhang (2022) for multi-layer networks without covariates.

Denote by $n_{\min}$, $n_{\max}$, and $n'_{\max}$ the minimal, maximal and second maximal community sizes, respectively. The following lemma establishes some properties for the factor matrices of $\boldsymbol{\mathcal{P}}$) under the proposed CAMSBM.

Lemma 1. Denote $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{O}$ in (1). Under Assumption A, we have $\boldsymbol{U}_{i,:} = \boldsymbol{U}_{j,:}$ if $c_i^* = c_j^*$ and $\|\boldsymbol{U}_{i,:} - \boldsymbol{U}_{j,:}\| = \sqrt{1/n_{c_i^*} + 1/n_{c_j^*}}$ otherwise, for $i, j \in [n]$, and $\|\boldsymbol{V}_{l,:}\| = O\Big(\frac{n s_n^{(l)}}{n_{\min}\sqrt{L} s_n^{[L_0]}}\Big)$, for $l \in [L+1]$.

Lemma 1 justifies the incoherence property of the factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ that the magnitudes of rows of $\boldsymbol{U}$ and $\boldsymbol{V}$ scale at orders not faster than $n^{-1/2}$ and $L^{-1/2}$, respectively. These two incoherence properties play a key role in deriving a sharp upper bound for the principle angle between subspaces (Liu and Moitra 2020) spanned by the columns of $\widehat{\boldsymbol{U}}$ and $\boldsymbol{U}$, which essentially governs the community detection error. Furthermore, we make the following technical assumptions to establish the asymptotic consistency of community detection.

Assumption B. Assume that $\boldsymbol{\epsilon}_i \in \mathbb{R}^p, i \in [n]$ are independently and identically distributed p-dimensional random vectors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, and that there exists an absolute constant C_1 such that

$$\|\boldsymbol{\epsilon}_i^T \boldsymbol{x}\|_{\psi_2} \le C_1 \sqrt{\mathbb{E}(\boldsymbol{\epsilon}_i^T \boldsymbol{x})^2}, \text{ for any } \boldsymbol{x} \in \mathbb{R}^p,$$

where $\|x\|_{\psi_2} = \inf\{a \ge 0 : \mathbb{E}(\exp(\lambda x)) \le \exp\big(\frac{a^2\lambda^2}{2}\big), \text{ for any } \lambda \in \mathbb{R}\}$.

Assumption C. Assume that $\boldsymbol{M}$ is of full row rank; that is $\mathrm{rank}(\boldsymbol{M}) = K$.

Assumption B imposes a distributional constraint on the random noise $\boldsymbol{\epsilon}_i$ for $i \in [n]$, which is satisfied by the sub-Gaussian distribution as well as many other distributions. Assumption C ensures that the underlying community centers of covariates are distinguishable.

Theorem 1. Under Assumptions A–C, there exists an absolute constant C such that the Hamming error of $\widehat{\boldsymbol{c}}$ satisfies that

$$\mathrm{Err}(\widehat{\boldsymbol{c}}, \boldsymbol{c}^*) \le \frac{C(2+\epsilon) n'_{\max}\big(2n_{\max}\sqrt{\sum_{l=1}^L (s_n^{(l)})^2 + \alpha_n^2} + \delta_n\big)^2 \delta_n^2}{n n_{\min}^4 \big(\sum_{l=1}^L (s_n^{(l)})^2 + \alpha_n^2\big)^2},$$

where

$$\delta_n = O_p\Big(\frac{n^{3/2}\sqrt{\sum_{l=1}^L s_n^{(l)} \log n}\max_l s_n^{(l)} + \alpha_n^2 n^2(\|\Sigma\| + \|\Sigma\|^{1/2})}{n_{\min}\sqrt{L} s_n^{[L_0]}}\Big).$$

Theorem 1 establishes the asymptotic consistency of the proposed method in detecting communities under the proposed CAMSBM, whose convergence rate is governed by a number of factors, including the network size n, the number of layers L, the minimal and maximal community sizes, the magnitude of covariate noise, as well as the sparsity factors $\{s_n^{(l)}\}_{l=1}^{L+1}$. In sharp contrast to most existing results in literature (Lei, Chen, and Lynch 2020; Jing et al. 2021). Theorem 1 holds true for unbalanced community structures, allowing for the divergence of $n_{\max}/n_{\min}$. Furthermore, instead of considering a uniform sparsity factor on all network layers, Theorem 1 permits varying network sparsity factors $s_n^{(l)}$, which leads to some explicit rate of convergence for the proposed method.

Theorem 1 also provides some important guidelines to determine the weighting parameter α_n. Consider a simple case with $s_n^{(1)} = \cdots = s_n^{(L)}$ and $n_{\max}/n_{\min} = O(1)$, $\mathrm{Err}(\hat{\boldsymbol{c}}, \boldsymbol{c}^*)$ will vanish with high probability if and only if $\delta_n = o_p(n\sqrt{L s_n^2 + \alpha_n^2})$. Some simple algebra further implies that α_n should be chosen appropriately. For instances, when the network signal is sufficiently large such that $s_n \gg \frac{\log n}{nL}$, we need to set $\alpha_n < \sqrt{L} s_n$ with $\alpha_n = o\big(\frac{\sqrt{L} s_n}{(||\Sigma||+||\Sigma||^{1/2})^{1/2}}\big)$, or $\alpha_n \ge \sqrt{L} s_n$ with $\alpha_n = o\big(\frac{\sqrt{L} s_n}{||\Sigma||+||\Sigma||^{1/2}}\big)$. When the network is relatively sparse such that $s_n = O(\frac{\log n}{nL})$, we need to set $\alpha_n \ge \sqrt{L} s_n$ with $\alpha_n = o\big(\frac{\sqrt{L} s_n}{||\Sigma||+||\Sigma||^{1/2}}\big)$, which is feasible only when the noise of the covariates $||\Sigma||$ vanishes. The following corollaries further characterize the precise interplay of the Hamming error with the signal from the multi-layer network, noise level of the covariates, and the weighting parameter.

Corollary 1. Under Assumptions A–C, if we further assume $s_n^{(1)} = \cdots = s_n^{(L)} = s_n$, $n_{\max} = O(n_{\min})$ and regard ϵ as a constant. Then, we have

$$\mathrm{Err}(\hat{\boldsymbol{c}}, \boldsymbol{c}^*) = O_p\Big(\frac{s_n \log n + \alpha_n^4 n L^{-1} s_n^{-2}(||\boldsymbol{\Sigma}||^2 + ||\boldsymbol{\Sigma}||)}{n(L s_n^2 + \alpha_n^2)}\Big),$$

provided that α_n is in its feasible region satisfying:

1. $\alpha_n \geq \sqrt{L} s_n$ with $\alpha_n = o\big(\frac{\sqrt{L} s_n}{||\Sigma||+||\Sigma||^{1/2}}\big)$; or
2. $\alpha_n < \sqrt{L} s_n$ with $\alpha_n = o\big(\frac{\sqrt{L} s_n}{(||\Sigma||+||\Sigma||^{1/2})^{1/2}}\big)$ if $s_n \gg \frac{\log n}{nL}$.

Clearly, when the noise level of the covariates, measured by $\|\boldsymbol{\Sigma}\|$, is sufficiently small, the probabilistic upper bound for the Hamming error decreases as α_n increases in its feasible region, even for very sparse networks. This confirms the benefit of incorporating the node-wise covariates to assist community detection in multi-layer networks, most notably when the community structure is clearly presented in the node-wise covariates.

Corollary 2. Under Assumptions A–C, suppose that $\alpha_n = \epsilon = 0$ and $s_n^{(1)} = \cdots = s_n^{(L)} = s_n = \Omega(\frac{n \log n}{n_{\min}^2 L})$, and then $\mathrm{Err}(\hat{\boldsymbol{c}}, \boldsymbol{c}^*) = O_p(\frac{n'_{\max} n^4 \log n}{n_{\min}^6 L s_n})$.

Corollary 2 shows that the asymptotic consistency of community detection of MSBM with a uniform sparsity factor can be regarded as a special case of Theorem 1, matching up with the fastest convergence rate of community detection in multi-layer networks in literature (Paul and Chen 2020b; Jing et al. 2021), provided that $n_{\max} = O(n_{\min})$ and K and L_0 are fixed at a constant order.

4. Simulation Studies

In this section, we conduct a series of simulation studies to illustrate the superiority of the proposed method in community detection with the assistance of covariate information. Specifically, the objective of the simulations is 2-fold. First, we intend to highlight the necessity of incorporating node-wise covariates into multi-layer network for community detection, which brings significant improvement of the accuracy for community detection, especially when the networks are relatively sparse. To this end, we compare the proposed method against two baselines, MSBM and K-means, which uses either network data or node-wise covariates for community detection. Second, we compare the proposed method against several competitors to illustrate the effectiveness of the proposed method in blending information from both multi-layer network and node-wise covariates, including least square estimation (LSE; Lei, Chen, and Lynch 2020), mean adjacency spectral embedding (MAS; Han, Xu, and Airoldi 2015), and joint embedding of graphs (JEG; Wang et al. 2019). Specifically, for LSE, MAS, and JEG, similarity matrix based on covariates will be treated as additional layer as in the proposed method, since these methods can be applied to networks with weighted edges. The value of tuning parameter α_n is chosen according to our theoretical findings in Corollary 1.

Table 1. The averaged Hamming errors of various methods with their standard errors in parentheses in Scenario 1. The best performer in each case is bold-faced.

s_n	n	CAMSBM	MSBM	Kmeans	MAS	LSE	JEG
0.05	300	0.4233	0.6888	0.3938	**0.3783**	0.4851	0.3810
		(0.0066)	(0.0022)	(0.0104)	(0.0050)	(0.0069)	(0.0045)
	400	0.3766	0.6818	0.3730	**0.3602**	0.4809	0.3708
		(0.0054)	(0.0027)	(0.0041)	(0.0037)	(0.0066)	(0.0036)
	500	**0.3457**	0.6600	0.3693	0.3579	0.4719	0.3673
		(0.0040)	(0.0043)	(0.0033)	(0.0032)	(0.0069)	(0.0031)
0.1	300	0.3841	0.5482	0.3873	**0.3597**	0.4862	0.3828
		(0.0126)	(0.0102)	(0.0090)	(0.0044)	(0.0067)	(0.0047)
	400	**0.1266**	0.2051	0.3780	0.3447	0.4790	0.3730
		(0.0118)	(0.0143)	(0.0044)	(0.0034)	(0.0072)	(0.0034)
	500	**0.0055**	0.0122	0.3699	0.3404	0.4756	0.3668
		(0.0004)	(0.0026)	(0.0032)	(0.0027)	(0.0069)	(0.0032)

The synthetic multi-layer networks $\boldsymbol{\mathcal{A}} \in \{0,1\}^{n\times n\times L}$ are generated as follows. First, the probability tensor $\boldsymbol{\mathcal{B}} \in [0,1]^{K\times K\times L}$ is generated as $\boldsymbol{\mathcal{B}}_{k_1,k_2,l} = s_n^{(l)} b_{k_1,k_2,l}$ with $b_{k_1,k_2,l} \sim \mathrm{Unif}(0.2 + 0.2 * I(k_1 = k_2), 0.5)$, for $k_1, k_2 \in [K]$. Second, $\boldsymbol{c} = (c_1, \ldots, c_n)$ are randomly drawn from $[K]$ with equal probability, leading to the community assignment matrix $\boldsymbol{Z}$. Third, we set $\boldsymbol{\mathcal{P}}) = \boldsymbol{\mathcal{B}} \times_1 \boldsymbol{Z} \times_2 \boldsymbol{Z}$, and $\boldsymbol{\mathcal{A}}$ is generated via $\boldsymbol{\mathcal{A}}_{i,j,l} \sim \mathrm{Bernoulli}(\boldsymbol{\mathcal{P}})_{i,j,l})$ independently for $1 \leq i \leq j \leq n$ and $l \in [L]$. Fourth, the mean matrix $\boldsymbol{M} = (\boldsymbol{I}_K, \boldsymbol{0}_{K\times(p-K)})$, and we generate $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{M} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}_{i,j} \sim N(0, \sigma^2)$ for $i \in [n]$ and $j \in [p]$. We consider three scenarios to illustrate the numerical performance.

Scenario 1. This scenario intends to characterize the convergence of the Hamming errors of all the competing methods with respect to the number of nodes under sparse and dense networks. To this end, we fix $(K, \sigma, L, p) = (4, 0.75, 20, 6)$ and consider cases $(n, s_n^{(l)}) \in \{300, 400, 500\} \times \{0.05, 0.1\}$. The averaged Hamming errors and their standard errors over 100 independent replications are reported in Table 1.

As shown in Table 1, CAMSBM shows its advantage in combining the multi-layer network and node-wise covariates to yield better community detection performances compared with two baselines. As n increase, the performance of the proposed method improves significantly, which demonstrates the importance of the chosen weighting parameter in balancing two sources of information. Conversely, MAS, LSE, and JES relies on covariates, and hence their performances stay less affected when network information gets richer.

Scenario 2. This scenario intends to demonstrate the effect of number of layers under a fixed effect of covariates. To this end, we fix $(n, K, s_n, p) = (300, 4, 0.08, 6)$ and consider cases $L \in \{10, 15, 20, 25, 30, 35, 40\}$ with the same $\boldsymbol{\mathcal{B}}$, $\boldsymbol{M}$ and $\boldsymbol{c}$ as in Scenario 1. The averaged Hamming errors and their standard errors over 100 independent replications are reported in Figure 1.

As can be seen in Figure 1, as number of layers increases, the network information strengthens and hence all the competing methods except K-means observe obvious improvement in the performance of community detection. Additionally, it is interesting to note that the decreasing patterns of JEG and

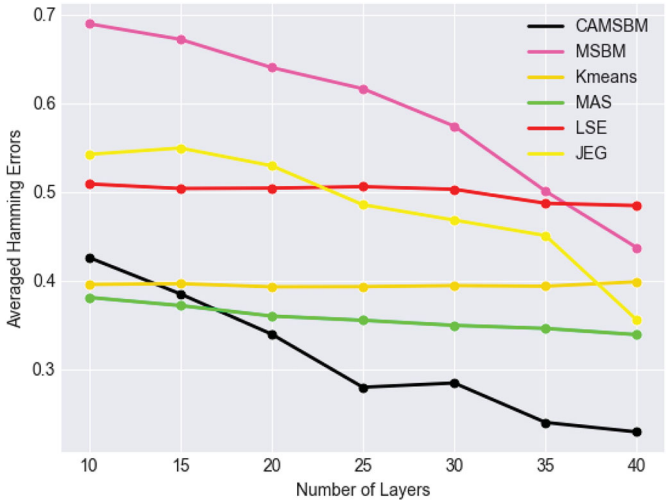

Figure 1. The averaged Hamming errors of various methods versus the number of layers.

Table 2. The averaged Hamming errors of various methods with their standard errors in parentheses in Scenario 3. The best performer in each case is bold-faced.

σ	CAMSBM	MSBM	Kmeans	MAS	LSE	JEG
0.5	**0.1347**	0.5177	0.185	0.1484	0.2044	0.3612
	(0.0028)	(0.0107)	(0.0023)	(0.0022)	(0.0038)	(0.0086)
0.75	**0.2585**	0.5175	0.3963	0.357	0.4818	0.5038
	(0.0072)	(0.0108)	(0.0055)	(0.0044)	(0.0066)	(0.011)
1	**0.3386**	0.5175	0.539	0.5162	0.5711	0.5210
	(0.0088)	(0.0108)	(0.0053)	(0.0057)	(0.0046)	(0.0093)
1.25	**0.3986**	0.5172	0.6005	0.5877	0.6192	0.5383
	(0.0096)	(0.0108)	(0.0038)	(0.004)	(0.0031)	(0.0093)
1.5	**0.4420**	0.5175	0.6239	0.6251	0.6417	0.5617
	(0.0108)	(0.0108)	(0.0032)	(0.003)	(0.0027)	(0.0078)
3	**0.5163**	0.5171	0.6829	0.6789	0.6837	0.6101
	(0.0098)	(0.0107)	(0.0016)	(0.0021)	(0.0018)	(0.0068)

CAMSBM resemble that of MSBM compared with the other methods, indicating that these two methods perform better in blending the information of network and covariates.

Scenario 3. This scenario intends to explore the effect of covariates on the Hamming errors of all competing methods. To this end, we set $(K, n, L, s_n^{(l)}, p) = (4, 300, 20, 0.1, 6)$ and consider cases with $\sigma \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 3\}$. To control the information from the networks, all generated networks are set to be identical under different values of σ. The hamming errors versus different variances of the covariates are reported in Table 2.

As shown in Table 2, the Hamming errors of those covariate-assisted methods deteriorate expectedly as the variance of covariates increases. CAMSBM appears to be the only method outperforming two baselines simultaneously under all settings, which essentially credits to the appropriate scaling on the covariate layer. Moreover, when $\sigma = 3$, the Hamming error of CAMSBM becomes not much different from that of MSBM, as covariates with large variable provide very little information on the community structure.

5. Real Applications

5.1. Business Relation Network

We apply the proposed CAMSBM method to a business relation network to showcase its practical advantages in terms of community detection against its competitors. Specifically, we show that the proposed method is capable of combining information from both multi-layer network and node-wise covariates effectively, which results in superior performance of community detection than its existing competitors.

The business relation network is collected by *Relato* and publicly available at *https://data.world/datasyndrome/relato-business-graph-database*. The dataset contains 373,663 links of five kinds of pairwise relationships among 51,222 companies, including "partnership," "customer," "competitor," "investment," and "supplier." For illustration, we only consider those companies forming the Standard & Poor's 500, and the dataset is then processed into a 5-layer network with each layer corresponding to a kind of pairwise relationship. We further remove those companies without any link in the analysis and retain the five largest sectors, including "health," "information technology," "financials," "consumer discretionary," and "industrials," which finally leads to a $312 \times 312 \times 5$ multi-layer network. In the constructed multi-layer network, the associated sectors of each company will be treated as its community label. Figure 2 displays the competitor network, in which two companies has a link if there exists a competitor relationship between them, and it is clear that companies within the same sector tend to have competitor relationships.

To construct the covariates for each company, we collect their closing prices in the stock market during the period from 2021-01-01 to 2021-06-01, which are then standardized to have zero mean and unit variance. Therefore, each company has a 102-dimensional covariate vector that characterizes its stock price flowing trend. Intuitively, companies in the same sector tend to have a similar stock price trend and stay close in the covariate space.

To evaluate the numerical performance in terms of community detection, we repeat all experiments 100 times and report their averaged Hamming errors and standard errors in Table 3. Clearly, CAMSBM outperforms K-means and MSBM simultaneously with percentages of improvements 8.8% and 16.9%, respectively. This result confirms that node-wise covariates indeed help improve the community detection accuracy in this multi-layer business relation network. Moreover, CAMSBM also outperforms MAS, LSE, and JEG, showcasing its effectiveness in incorporating covariate information for community detection.

5.2. Statisticians' Citation Network

In this section, we apply CAMSBM to analyze the statisticians' citation network (Ji and Jin 2016), which consists of a total of 3248 research papers by 3607 researchers, published in four leading statistics journals from 2003 to 2012. It also contains the citation relationship among all papers, as well as covariates such as DOI, year, title, citation count and abstract. The dataset is publicly available at *https://www.stat.uga.edu/directory/people/pengsheng-ji*.

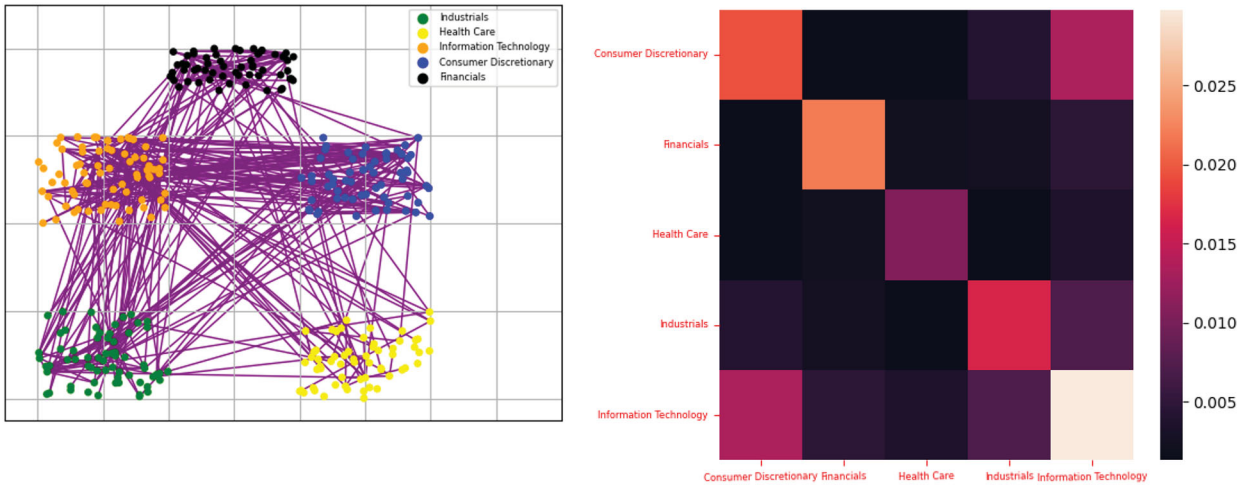

Figure 2. The edge pattern of the competitor network (Left) and the frequency of edges among communities in the layer of competitor relationship (Right).

Table 3. The averaged Hamming errors of various methods with their standard errors in parentheses for the business relation network.

CAMSBM	Kmeans	MSBM	MAS	LSE	JEG
0.5992 (0.0004)	0.6571 (0.0011)	0.7210 (0.0002)	0.6014 (0.0011)	0.6628 (0.0022)	0.6415 (0.0025)

The dataset is preprocessed into an 8-layer network, where nodes in different layer represent the same set of 3607 researchers and edges in different layers capture different interactions between two researchers. Specifically, for the first four layers, an edge between two researchers in a layer indicates there exists a citation between their papers in the corresponding journal, while an edge in the last four layers indicates these two researchers coauthored at least one paper in the corresponding journal. The covariates for each researcher are constructed based on averaging the bag-of-word representation of his or her papers, which is obtained by the term frequency–inverse document frequency (tf-idf) of the 1000 most significant terms from each abstract, including 1-gram to 4-gram. The bag-of-word representation is also standardized for the downstream analysis.

Note that the true community structure in the citation network is unavailable, and thus the Hamming error is no longer applicable. As remedy, we use the averaged internal density (Yang and Leskovec 2015) illustrate the performance of community detection. Formally, let $R_{ij}^{(l)} = 1$ denote there exists citation or co-authorship between nodes i and j, and 0 otherwise for $1 \le i < j \le n$ and $l \in [8]$. Then the employed performance metric is given as

$$\mathrm{Acc}(\widehat{\boldsymbol{c}}) = \sum_{k=1}^{K} \frac{\widehat{n}_k}{nL} \sum_{(i,j)\in N_k(\widehat{\boldsymbol{c}})} \sum_{l=1}^{L} \frac{R_{ij}^{(l)}}{|N_k(\widehat{\boldsymbol{c}})|}, \tag{3}$$

where $\widehat{n}_k = \sum_{i=1}^{n} I(\widehat{c}_i = k)$ denotes the estimation of the size of the kth community by $\widehat{\boldsymbol{c}}$ and $N_k(\widehat{\boldsymbol{c}}) = \{(i,j) : \widehat{c}_i = \widehat{c}_j = k, i \ne j\}$ for $k \in [K]$. The evaluation metric in (3) measures the averaged linking intensity within the communities, and the intuition behind is that nodes within a community ought to be densely connected.

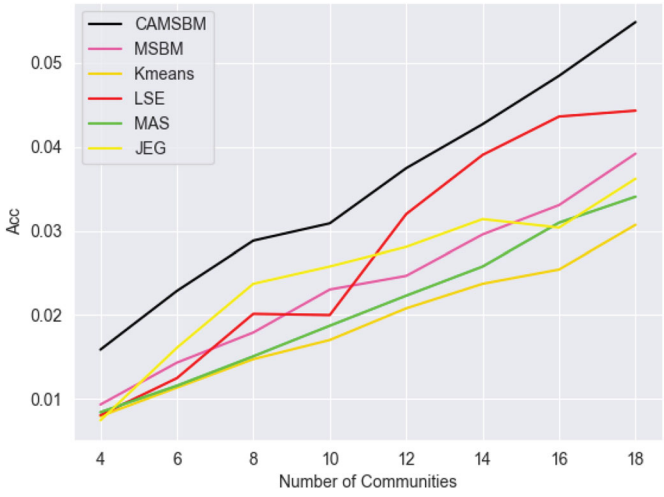

Figure 3. The averaged internal densities of all the methods for $K = 4, 6, \ldots, 18$.

In this application, we only compare CAMSBM with its two baselines MSBM and K-means to illustrate the effectiveness in integrating two sources of information from network and covariates to uncover better community structure, since the superiority of the proposed method is already illustrated in synthetic networks. To this end, we set the number of communities as $K = \{2 + 2 * i, i \in [9]\}$ and display the performance metrics of all the competing methods in Figure 3.

As shown in Figure 3, the proposed method yield better performance than its two baselines MSBM and Kmeans for $K = 4, \ldots, 18$ with improvements ranging from 34.08% to 101.4%, suggesting that the proposed framework is effective in combining information from network and covariates to improve the community structure.

6. Summary

In this article, we propose a covariate-assisted community detection method for multi-layer network. It first constructs a similarity matrix based on the node-wise covariates as an additional layer with appropriate scaling, and implements the Tucker decomposition on the augmented network adjacency tensor to

obtain the spectral embedding vectors for nodes, which is then cast into the K-means algorithm to obtain an $(1+\epsilon)$-optimal solution for community detection. One of the key advantages of the proposed method is that it incorporates the node-wise covariates into multi-layer network for better community detection accuracy, which is supported by theoretical analysis and extensive numerical experiments. Particularly, when the node-wise covariates are not available, our theoretical result matches up with the best existing results in literature when the Tucker rank of the expected adjacency tensor is of constant order; and when the node-wise covariates provide clear information about the community structure, the proposed method can outperform existing methods by allowing very sparse multi-layer networks.

Appendix

Proof of Lemma 1: By the definition of $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{O}$, it is clear that $\boldsymbol{U}$ has only K distinct rows, each corresponding to one community. It immediately follows that $\boldsymbol{U}_{i,:} = \boldsymbol{U}_{j,:}$ if $c_i^* = c_j^*$. Further, as $\boldsymbol{O}$ is an orthogonal matrix, $\boldsymbol{U}_{i,:}$ and $\boldsymbol{U}_{j,:}$ are perpendicular to each other if $c_i^* \neq c_j^*$ and $\|\boldsymbol{U}_{i,:}\| = \sqrt{1/n_{c_i^*}}$. We thus conclude that $\|\boldsymbol{U}_{i,:} - \boldsymbol{U}_{j,:}\| = \sqrt{1/n_{c_i^*} + 1/n_{c_j^*}}$ if $c_i^* \neq c_j^*$, for $i, j \in [n]$.

Next, we turn to bound $\|\boldsymbol{V}_{l,:}\|$. Note that

$$\begin{aligned}\sigma_{L_0}\Big(\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\Big) &= \sigma_{L_0}\Big(\boldsymbol{V}\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\Big) \\ &= \sigma_{L_0}\Big(\boldsymbol{\mathcal{M}}_3(\widetilde{\boldsymbol{B}})(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\Big) \\ &= \Omega\big(n_{\min}\sqrt{L}s_n^{[L_0]}\big),\end{aligned}$$

where the last equality follows from Assumption A. Therefore, for any $l \in [L+1]$, we have

$$\begin{aligned}\|\boldsymbol{V}_{l,:}\|\sigma_{L_0}\Big(\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\Big) &\le \|(\boldsymbol{V}_{l,:})^T\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\| \\ &= \|(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O} \times_3 \boldsymbol{V})_{:,:,l}\|_F \\ &= \|(\widetilde{\boldsymbol{B}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma})_{:,:,l}\|_F = O(ns_n^{(l)}).\end{aligned}$$

Combining the above two bounds implies that

$$\|\boldsymbol{V}_{l,:}\| = \frac{\|\boldsymbol{V}_{l,:}\|\sigma_{L_0}\Big(\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\Big)}{\sigma_{L_0}\Big(\boldsymbol{\mathcal{M}}_3(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{O} \times_2 \boldsymbol{O})\Big)} = O\Big(\frac{ns_n^{(l)}}{n_{\min}\sqrt{L}s_n^{[L_0]}}\Big).$$

This completes the proof of Lemma 1. □

Lemma 2. Under the conditions of Lemma 1, there exists an orthogonal matrix $\boldsymbol{O}^{(1)} \in \mathbb{R}^{K\times K}$, such that the left singular vectors corresponding to the nonzero singular values of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})$ is $\boldsymbol{U}\boldsymbol{O}^{(1)}$.

Proof of Lemma 2: Since $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})$ and $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})$ share the same left singular space, it suffices to construct the left singular vectors corresponding to the nonzero singular values of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})$.

Note that $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}}) = \boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}) \times_1 \boldsymbol{U} \times_2 \boldsymbol{U} \times_3 \boldsymbol{V})$ and $\widetilde{\boldsymbol{\mathcal{P}}}$ has Tucker rank (K, K, L_0), we have $\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))$ is of full row rank K. Let $\boldsymbol{O}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\boldsymbol{V}^T_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}$ be the singular value decomposition of $\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))$, it then follows that

$$\begin{aligned}\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}}) &= \boldsymbol{U}\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))(\boldsymbol{V} \otimes \boldsymbol{U})^T \\ &= (\boldsymbol{U}\boldsymbol{O}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))})\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big((\boldsymbol{V} \otimes \boldsymbol{U})\boldsymbol{V}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big)^T.\end{aligned}$$

It can be verified that $(\boldsymbol{U}\boldsymbol{O}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))})^T(\boldsymbol{U}\boldsymbol{O}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}) = \boldsymbol{I}$ and $\big((\boldsymbol{V} \otimes \boldsymbol{U})\boldsymbol{V}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big)^T\big((\boldsymbol{V} \otimes \boldsymbol{U})\boldsymbol{V}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big) = \boldsymbol{I}$. Therefore, $(\boldsymbol{U}\boldsymbol{O}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))})\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big((\boldsymbol{V} \otimes \boldsymbol{U})\boldsymbol{V}_{\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))}\big)^T$ is the singular value decomposition of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})$, and thus there exists an orthogonal matrix $\boldsymbol{O}^{(1)}$ such that the left singular vector of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})$ is $\boldsymbol{U}\boldsymbol{O}^{(1)}$. □

Lemma 3. Under the conditions in Lemma 1 and Assumptions C, it holds true that

$$\sigma_k\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma})\big) = \Theta\Big(n_k\sqrt{\sum_{l=1}^{L+1}(s_n^{(l)})^2}\Big), \text{ for } k \in [K].$$

Proof of Lemma 3: Let $\boldsymbol{F}^{(l)} = \boldsymbol{\Gamma}\widetilde{\boldsymbol{\mathcal{B}}}_{:,:,l}\boldsymbol{\Gamma}$, for $l \in [L+1]$. By the definition of $\widetilde{\boldsymbol{\mathcal{B}}}_{:,:,l}$, we have

$$\boldsymbol{F}^{(l)} = \begin{cases}\boldsymbol{\Gamma}\boldsymbol{B}_{:,:,l}\boldsymbol{\Gamma}, \text{ for } l \in [L], \\ \alpha_n\boldsymbol{\Gamma}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{\Gamma}, \text{ for } l = L+1.\end{cases}$$

Note that $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma}) = (\boldsymbol{F}^{(1)}, \dots \boldsymbol{F}^{(L+1)})$, and then

$$\begin{aligned}\sigma_k\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma})\big) &= \sigma_k(\boldsymbol{F}^{(1)}, ..., \boldsymbol{F}^{(L+1)}) \\ &= \Big(\lambda_k\Big(\sum_{l=1}^{L+1}\boldsymbol{F}^{(l)}(\boldsymbol{F}^{(l)})^T\Big)\Big)^{1/2},\end{aligned}$$

where $\lambda_k(\cdot)$ denotes the kth largest eigenvalue of a symmetric matrix.

By Assumptions C, we have

$$\lambda_k\big(\boldsymbol{F}^{(l)}(\boldsymbol{F}^{(l)})^T\big) = \Theta\big((n_k s_n^{(l)})^2\big), \text{ for } l \in [L+1].$$

Further, it follows from Weyl's inequality that

$$\lambda_k\Big(\sum_{l=1}^{L+1}\boldsymbol{F}^{(l)}(\boldsymbol{F}^{(l)})^T\Big) = \Theta\Big(n_k^2\sum_{l=1}^{L+1}(s_n^{(l)})^2\Big), \text{ for } k \in [K].$$

The desired result then follows immediately. □

Lemma 4. Under the conditions in Lemma 1 and Assumptions C, it holds true that

$$\sigma_k\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})\big) = \Theta\Big(n_k\sqrt{\sum_{l=1}^{L+1}(s_n^{(l)})^2}\Big), \text{ for } k \in [K].$$

Proof of Lemma 4: The Tucker decomposition of $\widetilde{\boldsymbol{\mathcal{P}}}$ implies that

$$\begin{aligned}\sigma_k\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})\big) &= \sigma_k\big(\boldsymbol{U}\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))(\boldsymbol{V}^T \otimes \boldsymbol{U}^T)(\boldsymbol{V} \otimes \boldsymbol{U})\big) \\ &= \sigma_k\big(\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))\big),\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{U}$ has orthonormal columns and hence does not affect the singular values. It also follows from Lemma 3 and the Tucker decomposition of $\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma}$ that

$$\begin{aligned}\sigma_k\big(\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))\big) &= \sigma_k\big(\boldsymbol{O}\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{C}}))(\boldsymbol{V} \otimes \boldsymbol{O})^T\big) \\ &= \sigma_k\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{B}}} \times_1 \boldsymbol{\Gamma} \times_2 \boldsymbol{\Gamma})\big) \\ &= \Theta\Big(n_k\sqrt{\sum_{l=1}^{L+1}(s_n^{(l)})^2}\Big).\end{aligned}$$

This completes the proof of Lemma 4. □

Lemma 5. Denote $\delta_n = \|\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}})(\boldsymbol{V} \otimes \boldsymbol{U}) - \boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})\|$. Then there exists an orthogonal matrix $\boldsymbol{O}^{(2)}$ such that

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{O}^{(2)}\|_F \le \frac{2\sqrt{2}\Big(2\sigma_1\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})\big) + \delta_n\Big)\delta_n}{\sigma_K^2\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V} \otimes \boldsymbol{U})\big)}.$$

Proof of Lemma 5: First, it follows from Lemma 2 that the columns of $\boldsymbol{U}\boldsymbol{O}^{(1)}$ are the left singular vectors corresponding to the nonzero singular values of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})$. Following the similar argument in Lemma 2, there exists an orthogonal matrix $\boldsymbol{O}^{(3)}$ such that $\widehat{\boldsymbol{U}}\boldsymbol{O}^{(3)}$ are the singular vectors corresponding to the first K leading singular values of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}})(\boldsymbol{V}\otimes\boldsymbol{U})$. By Theorem 4 in Yu, Wang, and Samworth (2015), there exists an orthogonal matrix $\boldsymbol{O}^{(4)}$ such that

$$\|\widehat{\boldsymbol{U}}-\boldsymbol{U}\boldsymbol{O}^{(1)}\boldsymbol{O}^{(4)}(\boldsymbol{O}^{(3)})^T\|_F=\|\widehat{\boldsymbol{U}}\boldsymbol{O}^{(3)}-\boldsymbol{U}\boldsymbol{O}^{(1)}\boldsymbol{O}^{(4)}\|_F \le \frac{2\sqrt{2}\big(2\sigma_1\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})\big)+\delta_n\big)\delta_n}{\sigma_K^2\big(\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})\big)}.$$

The desired result then follows immediately by taking $\boldsymbol{O}^{(2)}=\boldsymbol{O}^{(1)}\boldsymbol{O}^{(4)}(\boldsymbol{O}^{(3)})^T$. □

Lemma 6. Let δ_n be defined as in Lemma 5. Under Assumptions A–C, it holds true that

$$\delta_n=O_p\Big(\frac{\sqrt{n\sum_{l=1}^L s_n^{(l)}\log n}\max_l s_n^{(l)}+n\alpha_n^2(\|\Sigma\|+\|\Sigma\|^{1/2})}{\sqrt{L}s_n^{[L_0]}}\Big).$$

Proof of Lemma 6: Since $\boldsymbol{\mathcal{M}}_1(\cdot)$ is a linear operator, we have $\delta_n=\|\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}}-\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})\|$, and that

$$\begin{aligned}&\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{\mathcal{A}}}-\widetilde{\boldsymbol{\mathcal{P}}})(\boldsymbol{V}\otimes\boldsymbol{U})\\&\quad=[\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}})),\alpha_n(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})](\boldsymbol{V}\otimes\boldsymbol{U}).\end{aligned}$$

The triangle inequality then implies that

$$\delta_n\le\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L_:}\otimes\boldsymbol{U})\|+\alpha_n\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})(\boldsymbol{V}_{L+1,:}^T\otimes\boldsymbol{U})\|.$$

Thus, it suffices to bound $\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L_:}\otimes\boldsymbol{U})\|$ and $\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})(\boldsymbol{V}_{L+1,:}^T\otimes\boldsymbol{U})\|$, respectively.

(I). To bound $\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L_:}\otimes\boldsymbol{U})\|$, we note that

$$\Big\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L_:}\otimes\boldsymbol{U})\Big\|=\Big\|\sum_{l=1}^L\sum_{i\le j}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\Big\|,$$

where $\boldsymbol{E}^{(i,j)}$ is the square matrix with 1 in the (i,j)th and (j,i)th entries and 0 otherwise. Clearly, $(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})$ are independent zero-mean random matrices, for $i,j\in[n]$, $l\in[L]$.

To apply the matrix Bernstein inequality in Theorem 1.6 of Tropp (2012), we need to verify the required conditions for $(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})$. For each $i,j\in[n]$ and $l\in[L]$, we have

$$\begin{aligned}\Big\|(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\Big\|&=\Big\|(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\big(\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{j,:},\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{i,:}\big)^T\Big\|\\&\le\Big\|\big(\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{j,:},\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{i,:}\big)\Big\|.\end{aligned}\tag{4}$$

When $c_i^*\ne c_j^*$, the right-hand side of (4) can be upper bounded as

$$\Big\|\big(\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{j,:},\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{i,:}\big)\Big\|\le\|\boldsymbol{V}_{l,:}\|\sqrt{\frac{1}{n_{\min}}}.$$

When $c_i^*=c_j^*$, the right-hand side of (4) can be upper bounded as

$$\Big\|\big(\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{j,:},\boldsymbol{V}_{l,:}\otimes\boldsymbol{U}_{i,:}\big)\Big\|\le\|\boldsymbol{V}_{l,:}\|\sqrt{\frac{2}{n_{\min}}}.$$

Thus, it follows that

$$\Big\|(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\Big\|\le\|\boldsymbol{V}_{l,:}\|\sqrt{\frac{2}{n_{\min}}}\le\frac{\sqrt{2}n\max_{l\in[L]}s_n^{(l)}}{n_{\min}^{3/2}\sqrt{L}s_n^{[L_0]}}.$$

Next, we proceed to bound the second order moment. Denote

$$\begin{aligned}\mathrm{var}_1&=\sum_{l=1}^L\sum_{i\le j}\mathbb{E}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})^2\boldsymbol{E}^{(i,j)}\big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\big)\big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\big)^T\boldsymbol{E}^{(i,j)},\\\mathrm{var}_2&=\sum_{l=1}^L\sum_{i\le j}\mathbb{E}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})^2\Big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\Big)^T\boldsymbol{E}^{(i,j)}\boldsymbol{E}^{(i,j)}\big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\big).\end{aligned}$$

In what follows, we proceed to bound $\|\mathrm{var}_1\|$ and $\|\mathrm{var}_2\|$, separately, and then obtain a upper bound for $\max\{\|\mathrm{var}_1\|,\|\mathrm{var}_2\|\}$.

Note that $\big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\big)\big(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}\big)^T=\|\boldsymbol{V}_{l,:}\|^2\boldsymbol{Z}\boldsymbol{\Gamma}^{-2}\boldsymbol{Z}^T$. Thus,

$$\begin{aligned}\mathrm{var}_1&=\sum_{l=1}^L\|\boldsymbol{V}_{l,:}\|^2\sum_{i\le j}\mathbb{E}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})^2\boldsymbol{E}^{(i,j)}\boldsymbol{Z}\boldsymbol{\Gamma}^{-2}\boldsymbol{Z}^T\boldsymbol{E}^{(i,j)}\\&\preceq\sum_{l=1}^L\|\boldsymbol{V}_{l,:}\|^2\max_{k_1,k_2}\boldsymbol{B}_{k_1,k_2}^{(l)}(1-\boldsymbol{B}_{k_1,k_2}^{(l)})\sum_{i\le j}\boldsymbol{E}^{(i,j)}\boldsymbol{Z}\boldsymbol{\Gamma}^{-2}\boldsymbol{Z}^T\boldsymbol{E}^{(i,j)}\\&\preceq\sum_{l=1}^L\|\boldsymbol{V}_{l,:}\|^2s_n^{(l)}\sum_{i\le j}\boldsymbol{E}^{(i,j)}\boldsymbol{Z}\boldsymbol{\Gamma}^{-2}\boldsymbol{Z}^T\boldsymbol{E}^{(i,j)}\\&=\sum_{l=1}^L s_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2(K\boldsymbol{I}_n+\boldsymbol{G}_{\boldsymbol{c}}),\end{aligned}$$

where $\boldsymbol{I}_n$ is the n-dimensional identity matrix, and $\boldsymbol{G}_{\boldsymbol{c}}$ is a $n\times n$ matrix such that $(\boldsymbol{G}_{\boldsymbol{c}})_{ii}=0$ for $i\in[n]$, $(\boldsymbol{G}_{\boldsymbol{c}})_{ij}=n_{c_i^*}^{-1}$ if $c_i^*=c_j^*$ and 0 otherwise for $i\ne j$. Herein, the partial order $\preceq$ between two matrix $\boldsymbol{M}^{(1)}$ and $\boldsymbol{M}^{(2)}$ is defined as $\boldsymbol{M}^{(1)}\preceq\boldsymbol{M}^{(2)}$ if and only if $\boldsymbol{M}^{(2)}-\boldsymbol{M}^{(1)}$ is positive semidefinite. This leads to

$$\|\mathrm{var}_1\|\le\Big\|\sum_{l=1}^L s_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2(K\boldsymbol{I}_n+\boldsymbol{G}_{\boldsymbol{c}})\Big\|\le\sum_{l=1}^L s_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2(K+1),$$

where the last inequality follows from the triangle inequality and Gershgorin circle theorem.

Next, we turn to bound the spectral norm of var_2. Note that

$$\begin{aligned}\mathrm{var}_2&=\sum_{l=1}^L\sum_{i\le j}\mathbb{E}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})^2(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})^T\boldsymbol{I}^{(i,j)}(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})\\&=\sum_{l=1}^L(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})^T\Big(\sum_{i\le j}\mathbb{E}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})^2\boldsymbol{I}^{(i,j)}\Big)(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})\\&\preceq\sum_{l=1}^L(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})^T\boldsymbol{I}_n(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})\max_{k\in[K]}\sum_{k_1=1}^K n_{k_1}\boldsymbol{B}_{k,k_1}^{(l)}(1-\boldsymbol{B}_{k,k_1}^{(l)})\\&\preceq\sum_{l=1}^L ns_n^{(l)}(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})^T(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U}),\end{aligned}$$

where $\boldsymbol{I}^{(i,j)}$ is the diagonal matrix with the (i,i)th and (j,j)th entries being 1 and all other entries being zero. Hence,

$$\|\mathrm{var}_2\|\le\Big\|\sum_{l=1}^L ns_n^{(l)}(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})^T(\boldsymbol{V}_{l,:}^T\otimes\boldsymbol{U})\Big\|\le\sum_{l=1}^L ns_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2.$$

Since $n\gg(K+1)$, the variance condition can be met by the fact that $\sigma^2=\max\{\|\mathrm{var}_1\|,\|\mathrm{var}_2\|\}\le\sum_{l=1}^L ns_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2$. With this, Theorem 1.6 in Tropp (2012) yields that, for any $t>0$,

$$\mathbb{P}\Big(\Big\|\sum_{l=1}^{L}\sum_{i\le j}(\boldsymbol{A}_{i,j}^{(l)}-\boldsymbol{P}_{i,j}^{(l)})\boldsymbol{E}^{(i,j)}(\boldsymbol{V}_{l,:}^{T}\otimes\boldsymbol{U})\Big\|\ge t\Big)$$
$$\le (n+KL_0)\exp\Big\{-\frac{3t^2}{6\sum_{l=1}^{L}ns_n^{(l)}\|\boldsymbol{V}_{l,:}\|^2+\frac{2\sqrt{2}n\max_l s_n^{(l)}}{n_{\min}^{3/2}\sqrt{L}s_n^{[L_0]}}t}\Big\}.$$

Denote $\bar{s}_n = L^{-1}\sum_{l=1}^{L}s_n^{(l)}$. By the incoherence property of $\|\boldsymbol{V}_{l,:}\|$, there exists an absolute constant C_2, such that for any $t>0$,

$$\mathbb{P}\Big(\Big\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L,:}\otimes\boldsymbol{U})\Big\|\ge t\Big)$$
$$\le \exp\Big\{\log n-\frac{C_2t^2}{(\frac{\max_l s_n^{(l)}}{s_n^{[L_0]}})^2\frac{n^3\bar{s}_n}{n_{\min}^2}+\frac{n\max_l s_n^{(l)}}{n_{\min}^{3/2}\sqrt{L}s_n^{[L_0]}}t}\Big\}.$$

Taking $t=\sqrt{\frac{6}{C_2}}\frac{n\max_l s_n^{(l)}}{n_{\min}s_n^{[L_0]}}\sqrt{n\bar{s}_n\log n}$, then with probability at least $1-n^{-2}$, we have

$$\Big\|\boldsymbol{\mathcal{M}}_1(\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{P}}))(\boldsymbol{V}_{1:L,:}\otimes\boldsymbol{U})\Big\|<\sqrt{\frac{6}{C_2}}\frac{n\max_l s_n^{(l)}}{n_{\min}s_n^{[L_0]}}\sqrt{n\bar{s}_n\log n}. \quad (5)$$

(II). We now turn to bound $\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})(\boldsymbol{V}_{L+1:}\otimes\boldsymbol{U})\|$. By the incoherence property of $\boldsymbol{V}_{L+1,:}$,

$$\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})(\boldsymbol{V}_{L+1:}\otimes\boldsymbol{U})\|\le\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})\|\|\boldsymbol{V}_{L+1:}\otimes\boldsymbol{U}\|$$
$$\le\frac{n\alpha_n}{n_{\min}\sqrt{L}s_n^{[L_0]}}\|(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z})\|,$$

It thus suffices to bound $\sigma_1(\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T)$. Note that

$$\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T=(\boldsymbol{Z}\boldsymbol{M}+\boldsymbol{\epsilon})(\boldsymbol{Z}\boldsymbol{M}+\boldsymbol{\epsilon})^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T$$
$$=\boldsymbol{\epsilon}\boldsymbol{M}^T\boldsymbol{Z}^T+\boldsymbol{Z}\boldsymbol{M}\boldsymbol{\epsilon}^T+\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T.$$

Then, by triangle inequality, we have

$$\|\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T\|\le\|\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\|+2\|\boldsymbol{Z}\boldsymbol{M}\boldsymbol{\epsilon}^T\|$$
$$=n\|\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\|+2\sqrt{n}\sqrt{\|\boldsymbol{Z}\boldsymbol{M}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\boldsymbol{M}^T\boldsymbol{Z}^T\|}.$$

Notice that $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ is exactly the sample covariance matrix for the noise term. By the result of 4.7.3 in Vershynin (2018), there exists an absolute constant C_3 such that for any $\xi>0$,

$$\|\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\|\le\|\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n-\boldsymbol{\Sigma}\|+\|\boldsymbol{\Sigma}\|$$
$$\le C_3C_1^2\Big(\sqrt{\frac{p+\xi}{n}}+\frac{p+\xi}{n}+\frac{1}{C_3C_1^2}\Big)\|\boldsymbol{\Sigma}\|, \quad (6)$$

with probability at least $1-2\exp(-\xi)$, where C_1 is defined in Assumption B. Conditional on (6), the second term satisfies that

$$\|\boldsymbol{Z}\boldsymbol{M}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\boldsymbol{M}^T\boldsymbol{Z}^T\|=\|\boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}\boldsymbol{M}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\boldsymbol{M}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^T\|$$
$$=\|\boldsymbol{\Gamma}\boldsymbol{M}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n\boldsymbol{M}^T\boldsymbol{\Gamma}\|$$
$$\le\|\boldsymbol{\Gamma}\boldsymbol{M}\big(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n-\boldsymbol{\Sigma}\big)\boldsymbol{M}^T\boldsymbol{\Gamma}\|+\|\boldsymbol{\Gamma}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{\Gamma}\|$$
$$\le\|\boldsymbol{\Gamma}\boldsymbol{M}\|^2\Big(\|\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n-\boldsymbol{\Sigma}\|+\|\boldsymbol{\Sigma}\|\Big)$$
$$\le\max_{k=1,\dots,K}n_k\|\boldsymbol{M}\|^2C_3C_1^2$$
$$\Big(\sqrt{\frac{p+\xi}{n}}+\frac{p+\xi}{n}+\frac{1}{C_3C_1^2}\Big)\|\boldsymbol{\Sigma}\|, \quad (7)$$

where the second equality follows from the fact that $\boldsymbol{Z}\boldsymbol{\Gamma}^{-1}$ has orthonormal columns. Setting $\xi=n$, combining (6) and (7) yields that

$$\|\boldsymbol{X}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{Z}^T\|\le n\Big(C_4\|\boldsymbol{\Sigma}\|+2\|\boldsymbol{M}\|\sqrt{C_4\|\boldsymbol{\Sigma}\|}\Big)$$
$$\le C_5(||\Sigma||+||\Sigma||^{1/2})n, \quad (8)$$

where $C_4=C_3C_1^2\Big(2+\sqrt{2}+\frac{1}{C_3C_1^2}\Big)$ and C_5 is a positive constant.

Combining (5) and (8), it immediately follows that

$$\delta_n=\|\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{A}}-\widetilde{\boldsymbol{P}})(\boldsymbol{V}\otimes\boldsymbol{U})\|$$
$$\le\frac{\sqrt{6\sum_{l=1}^{L}s_n^{(l)}\log n/C_2}\max_l s_n^{(l)}n^{3/2}+C_5\alpha_n^2(||\Sigma||+||\Sigma||^{1/2})n^2}{n_{\min}\sqrt{L}s_n^{[L_0]}},$$

with probability at least $1-n^{-2}-2\exp(-n)$. □

Proof of Theorem 1: Let $\widehat{\boldsymbol{U}}\in\mathbb{R}^{n\times K}$ be any matrix having orthonormal columns such that the column space of $\widehat{\boldsymbol{U}}$ is the same as the one that spanned by the first K leading left singular vectors of $\boldsymbol{\mathcal{M}}_1(\widetilde{\boldsymbol{A}})$. Then the $(1+\epsilon)$-optimal approximation K-means algorithm is applied to estimate community assignment matrix and spectral embedding centers, which finds a pair of solution $(\widehat{\boldsymbol{Z}},\widehat{\boldsymbol{W}})$ such that

$$\|\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{W}}-\widehat{\boldsymbol{U}}\|^2\le(1+\epsilon)\min_{\boldsymbol{Z}\in\boldsymbol{\Delta}_{n,K},\boldsymbol{W}\in\mathbb{R}^{K\times K}}\|\boldsymbol{Z}\boldsymbol{W}-\widehat{\boldsymbol{U}}\|_F^2.$$

Let $\xi_k=\min_{l\neq k}\sqrt{1/n_l+1/n_k}$. Define $\mathcal{S}_k=\{i:c_i^*=k,\|\widehat{\boldsymbol{Z}}_{i,:}\widehat{\boldsymbol{W}}-\boldsymbol{U}_{i,:}\boldsymbol{O}^{(2)}\|_2\ge\xi_k/2\}$, where $\boldsymbol{O}^{(2)}$ is defined as in Lemma 5. It is easy to show that

$$\sum_{k=1}^{K}|S_k|\xi_k^2/4\le\|\boldsymbol{U}\boldsymbol{O}^{(2)}-\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{W}}\|_F^2\le2\|\boldsymbol{U}\boldsymbol{O}^{(2)}-\widehat{\boldsymbol{U}}\|_F^2+2\|\widehat{\boldsymbol{U}}-\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{W}}\|_F^2$$
$$\le2(2+\epsilon)\|\boldsymbol{U}\boldsymbol{O}^{(2)}-\widehat{\boldsymbol{U}}\|_F^2,$$

where the last inequality follows from the $(1+\epsilon)$-optimality of $\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{W}}$.

Further, by the definition of S_k, it can be verified that the Hamming error of $\hat{\boldsymbol{c}}$ can be bounded by

$$\mathrm{Err}(\hat{\boldsymbol{c}},\boldsymbol{c}^*)\le\frac{1}{n}\sum_{k=1}^{K}|S_k|.$$

Consequently,

$$\mathrm{Err}(\hat{\boldsymbol{c}},\boldsymbol{c}^*)\le\frac{1}{n}\sum_{k=1}^{K}|S_k|\le\frac{n'_{\max}}{n}\sum_{k=1}^{K}|S_k|\xi_k^2$$
$$\le\frac{8(2+\epsilon)n'_{\max}}{n}\|\boldsymbol{U}\boldsymbol{O}^{(2)}-\widehat{\boldsymbol{U}}\|_F^2,$$

where $n'_{\max}$ is second largest community size. Finally, it follows from Lemmas 4–6 that

$$\mathrm{Err}(\hat{\boldsymbol{c}},\boldsymbol{c}^*)\le\frac{C_6(2+\epsilon)n'_{\max}}{n}\Big(\frac{(2n_{\max}\sqrt{\sum_{l=1}^{L+1}(s_n^{(l)})^2}+\delta_n)\delta_n}{n_{\min}^2(\sum_{l=1}^{L+1}(s_n^{(l)})^2)}\Big)^2,$$

for some constant C_6, where

$$\delta_n=O_p\Big(\frac{\sqrt{\sum_{l=1}^{L}s_n^{(l)}\log n}\max_l s_n^{(l)}n^{3/2}+\alpha_n^2(\|\Sigma\|+\|\Sigma\|^{1/2})n^2}{n_{\min}\sqrt{L}s_n^{[L_0]}}\Big).$$

This completes the proof of Theorem 1. □

Proof of Corollary 1: Under the assumption of Corollary 1, $\delta_n = O_p(n^{1/2}\sqrt{s_n \log n} + \frac{\alpha_n^2 n(||\mathbf{\Sigma}||_+ ||\mathbf{\Sigma}||^{1/2})}{\sqrt{L s_n}})$, and Condition 1 or 2 implies that $\delta_n = o(n\sqrt{Ls_n^2 + \alpha_n^2})$. The desired result follows from Theorem 1 immediately.

Proof of Corollary 2: By the result of Theorem 1, the condition $s_n^{(1)} = \cdots = s_n^{(L)} = s_n = \Omega(\frac{n \log n}{n_{\min}^2 L})$ implies that $\delta_n = O_p\Big(\frac{n^{3/2}\sqrt{L s_n \log n} s_n}{n_{\min}\sqrt{L s_n}}\Big) = O_p\Big(\frac{n^{3/2}\sqrt{s_n \log n}}{n_{\min}}\Big) = O_p\Big(n s_n \sqrt{L}\Big)$. It then follows that

$$\mathrm{Err}(\hat{\boldsymbol{c}}, \boldsymbol{c}^*) = O_p\Big(\frac{n'_{\max} \cdot n^2 s_n^2 L \cdot n^3 s_n \log n}{n \cdot n_{\min}^2 \cdot n_{\min}^4 L^2 s_n^4}\Big) = O_p\Big(\frac{n'_{\max} n^4 \log n}{n_{\min}^6 s_n L}\Big). \quad \square$$

Supplementary Materials

All technical proofs, codes and datasets are included in the supplementary materials.

Acknowledgments

We thank the editor, the associate editor, and two anonymous referees for their constructive comments and suggestions.

Funding

This research is supported in part by HK RGC grants GRF-11300919, GRF-11304520, and GRF-11301521.

Disclosure Statement

The authors report there are no competing interests to declare.

ORCID

Yaoming Zhen https://orcid.org/0000-0002-3724-9200

References

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017), "Covariate-Assisted Spectral Clustering," *Biometrika*, 104, 361–377. [915,917]

Borondo, J., Morales, A., Benito, R., and Losada, J. (2015), "Multiple Leaders on a Multilayer Social Media," *Chaos, Solitons & Fractals*, 72, 90–98. [915]

Calderer, G., and Kuijjer, M. L. (2021, "Community Detection in Large-Scale Bipartite Biological Networks," *Frontiers in Genetics*, 12, 649440. [915]

Chen, S., Liu, S., and Ma, Z. (2022), "Global and Individualized Community Detection in Inhomogeneous Multilayer Networks," arXiv preprint arXiv:2012.00933. [915]

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Braver, T. S. (2013), "Multi-task Connectivity Reveals Flexible Hubs for Adaptive Task Control," *Nature Neuroscience*, 16, 1348–1355. [915]

Contisciani, M., Power, E. A., and De Bacco, C. (2020), "Community Detection with Node Attributes in Multilayer Networks," *Scientific Reports*, 10, 1–16. [915]

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "On the Best Rank-1 and Rank-(r_1, r_2, …,r_n) Approximation of Higher-Order Tensors," *SIAM Journal on Matrix Analysis and Applications*, 21, 1324–1342. [917]

Du, N., Wu, B., Pei, X., Wang, B., and Xu, L. (2007), "Community Detection in Large-Scale Social Networks," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pp. 16–25. [915]

Gao, T., Pan, R., Wang, S., Yang, Y., and Zhang, Y. (2021), "Community Detection for Statistical Citation Network by d-score," *Statistics and Its Interface*, 14, 279–294. [915]

Han, Q., Xu, K., and Airoldi, E. (2015), "Consistent Estimation of Dynamic and Multi-Layer Block Models," in *International Conference on Machine Learning*, pp. 1511–1520. PMLR. [915,919]

Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," *The Annals of Applied Statistics*, 10, 1779–1812. [920]

Jin, J. (2015), "Fast Community Detection by Score," *The Annals of Statistics*, 43, 57–89. [917]

Jing, B.-Y., Li, T., Lyu, Z., and Xia, D. (2021), "Community Detection on Mixture Multilayer Networks via Regularized Tensor Decomposition," *The Annals of Statistics*, 49, 3181–3205. [915,916,917,918,919]

Jung, S., and Segev, A. (2014), "Analyzing Future Communities in Growing Citation Networks," *Knowledge-Based Systems*, 69, 34–44. [915]

Ke, Z. T., Shi, F., and Xia, D. (2019), "Community Detection for Hypergraph Networks via Regularized Tensor Power Iteration," arXiv preprint arXiv:1909.06503. [917]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [916,917]

Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019), "Tensorly: Tensor Learning in Python," *Journal of Machine Learning Resarch*, 20, 1–6. [917]

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009), "Representing Degree Distributions, Clustering, and Homophily in Social Networks with Latent Cluster Random Effects Models," *Social Networks*, 31, 204–213. [915]

Lei, J., Chen, K., and Lynch, B. (2020), "Consistent Community Detection in Multi-Layer Network Data," *Biometrika*, 107, 61–73. [915,916,918,919]

Lei, J., and Lin, K. Z. (2022), "Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2022.2054817. [915,916]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [917]

Leskovec, J., Lang, K. J., and Mahoney, M. (2010), "Empirical Comparison of Algorithms for Network Community Detection," in *Proceedings of the 19th International Conference on World Wide Web*, pp. 631–640. [915]

Liu, A., and Moitra, A. (2020), "Tensor Completion Made Practical," *Advances in Neural Information Processing Systems*, 33, 18905–18916. [918]

Lyu, Z., Xia, D., and Zhang, Y. (2022), "Latent Space Model for Higher-Order Networks and Generalized Tensor Decomposition," arXiv preprint arXiv:2106.16042. [915,918]

Ma, Z., and Nandy, S. (2021), "Community Detection with Contextual Multilayer Networks," arXiv preprint arXiv:2104.02960. [915]

Macdonald, P., Levina, E., and Zhu, J. (2022), "Latent Space Models for Multiplex Networks with Shared Structure," *Biometrika*, (just-accepted), 1–24. [915]

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001), "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444. [915]

Paul, S., and Chen, Y. (2016), "Consistent Community Detection in Multi-Relational Data Through Restricted Multi-layer Stochastic Blockmodel," *Electronic Journal of Statistics*, 10, 3807–3870. [916]

Paul, S., and Chen, Y. (2020a), "A Random Effects Stochastic Block Model for Joint Community Detection in Multiple Networks with Applications to Neuroimaging," *The Annals of Applied Statistics*, 14, 993–1029. [915]

——— (2020b), "Spectral and Matrix Factorization Methods for Consistent Community Detection in Multi-layer Networks," *The Annals of Statistics*, 48, 230–250. [915,916,919]

Paul, S., and Chen, Y. (2021), "Null Models and Community Detection in Multi-layer Networks," *Sankhya A*, 84, 163–217. [915]

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019), "Topological and Functional Comparison of Community Detection Algorithms in Biological Networks," *BMC Bioinformatics*, 20, 1–25. [915]

Rudelson, M., and Vershynin, R. (2009), "Smallest Singular Value of a Random Rectangular Matrix," *Communications on Pure and Applied*

Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 62, 1707–1739. [918]

Tropp, J. A. (2012), "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434. [923]

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science* (Vol. 47), Cambridge: Cambridge University Press. [924]

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2019), "Joint Embedding of Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 1324–1336. [915,919]

Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2017), "Community Extraction in Multilayer Networks with Heterogeneous Community Structure," *The Journal of Machine Learning Research*, 18, 5458–5506. [915]

Yan, B., and Sarkar, P. (2021), "Covariate Regularized Community Detection in Sparse Graphs," *Journal of the American Statistical Association*, 116, 734–745. [915,917]

Yang, J., and Leskovec, J. (2015), "Defining and Evaluating Network Communities based on Ground-Truth," *Knowledge and Information Systems*, 42, 181–213. [921]

Yu, Y., Wang, T., and Samworth, R. J. (2015), "A Useful Variant of the Davis–Kahan Theorem for Statisticians," *Biometrika*, 102, 315–323. [923]

Yuan, Y., and Qu, A. (2021), "Community Detection with Dependent Connectivity," *The Annals of Statistics*, 49, 2378–2428. [915]

Zhang, X., Xu, G., and Zhu, J. (2022), "Joint Latent Space Models for Network Data with High-Dimensional Node Variables," *Biometrika*, (just-accepted) 1–14. [915]

Zhang, X., Xue, S., and Zhu, J. (2020), "A Flexible Latent Space Model for Multilayer Networks," in *International Conference on Machine Learning*, pp. 11288–11297. PMLR. [915]

Zhang, Y., Levina, E., and Zhu, J. (2016), "Community Detection in Networks with Node Features," *Electronic Journal of Statistics*, 10, 3153–3178. [915]

Zhao, J., Liu, X., Wang, H., and Leng, C. (2022), "Dimension Reduction for Covariates in Network Data," *Biometrika*, 109, 85–102. [915]

Zhen, Y., and Wang, J. (2022), "Community Detection in General Hypergraph via Graph Embedding," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2021.2002157. [917]