

# Clustering of Diverse Multiplex Networks

**Marianna Pensky\***

*Department of Mathematics  
University of Central Florida  
Orlando, FL 32816, USA*

MARIANNA.PENSKY@UCF.EDU

**Yaxuan Wang**

*Department of Mathematics  
University of Central Florida  
Orlando, FL 32816, USA*

YXWANG.MATH@KNIGHTS.UCF.EDU

**Editor:**

## Abstract

The paper introduces the Diverse MultiPLEx (DIMPLE) network model where all layers of the network have the same collection of nodes and are equipped with the Stochastic Block Models (SBM). In addition, all layers can be partitioned into groups with the same community structures, although the layers in the same group may have different matrices of block connection probabilities. The DIMPLE model generalizes a multitude of papers that study multilayer networks with the same community structures in all layers (which include the tensor block model and the checker-board model as particular cases), as well as the Mixture Multilayer Stochastic Block Model (MMLSBM), where the layers in the same group have identical matrices of block connection probabilities. Since the techniques from either of the above mentioned groups cannot be applied to the DIMPLE model, we introduce novel algorithms for the between-layer and the within-layer clustering. We study the accuracy of those algorithms, both theoretically and via computer simulations. Finally, we show how our between-layer clustering algorithm can be extended to the Heterogeneous Multiplex Random Dot-Product Graph model, which generalizes the COmmon Subspace Independent Edge (COSIE) random graph model developed in Arroyo et al. (2021).

**Keywords:** Multiplex Network, Stochastic Block Model, Community Detection, Spectral Clustering

## 1. Introduction

### 1.1 Multiplex network models

Stochastic network models appear in a variety of applications, including genetics, proteomics, medical imaging, international relationships, brain science and many more. While in the early years of the field of stochastic networks, research mainly focused on studying a single network, in recent years the frontier moved to investigation of collection of networks, the so called *multilayer network*, which allows to model relationships between nodes with respect to various modalities (e.g., relationships between species based on food or space), or consists of network data collected from different individuals (e.g., brain networks). Although there are many different ways of modeling a multilayer network (see, e.g., an excellent review article of Kivela et al. (2014)), in this paper, we consider the case where all layers

have the same set of nodes, and all the edges between nodes are drawn within layers, i.e., there are no edges connecting the nodes in different layers. Many authors, who work in a variety of research fields, study this particular version of a multilayer network (see, e.g., Aleta and Moreno (2019), Durante et al. (2017), Han and Dunson (2018), Kao and Porter (2017), MacDonald et al. (2021) among others). MacDonald et al. (2021) called this type of multilayer network models the *Multiplex Network Model* and argued that it appears in a variety of real life situations.

For example, the multiplex network models include brain networks where nodes are associated with brain regions, and edges are drawn if signals in those regions exhibit some kind of similarity (Sporns (2018)). In this setting, the nodes are the same for each individual network, and there is no connection between brain regions of different individuals. Another type of multiplex networks are trade networks between a set of countries (see, e.g., De Domenico et al. (2015)), where nodes and layers represent, respectively, various countries and commodities in which they are trading. In this case, edges are drawn if countries trade specific products with each other.

In this paper, we assume that nodes in the layers of the network are partitioned into communities. Specifically, we consider a multiplex network where all layers are equipped with the Stochastic Block Models (SBM). The SBM, according to Olhede and Wolfe (2014), provides an universal tool for description of time-independent stochastic network data. It is also very common in applications. For example, Sporns (2018) argues that stochastic block models provide a powerful tool for brain studies. In fact, in the last few years, such models have been widely employed in brain research (see, e.g., Crossley et al. (2013), Faskowitz et al. (2018), Nicolini et al. (2017), among others).

In this scenario, the problems of interest include finding groups of layers that are similar in some sense, finding the communities in those groups of layers, and estimation of the tensor of connection probabilities. While the scientific community attacked all three of those problems, often in a somewhat ad-hoc manner (see e.g., Brodka et al. (2018), Kao and Porter (2017), Mercado et al. (2018) among others), the theoretically inclined papers in the field of statistics mainly have been investigating the case where communities persist throughout all layers of the network. This includes studying the so called “checker board model” in Chi et al. (2020), where the matrices of block probabilities take only finite number of values, and communities are the same in all layers. The tensor block models of Wang and Zeng (2019) and Han et al. (2021) belong to the same category. In recent years, statistics publications extended this type of research to the case where community structure is preserved in all layers of the network, but the matrices of block connection probabilities can take arbitrary values (see, e.g., Bhattacharyya and Chatterjee (2020), Lei (2020), Lei et al. (2019), Paul and Chen (2016), Paul and Chen (2020) and references therein). The authors studied precision of community detection, and provided theoretical and numerical comparisons between various techniques that can be employed in this case.

Nevertheless, there are many real life scenarios where the assumption, that all layers of the network have the same communities, is too restrictive. For example, it is known that some brain disorders are associated with changes in brain network organizations (see, e.g., Buckner and DiNicola (2019)), and that alterations in the community structure of the brain have been observed in several neuropsychiatric conditions, including Alzheimer disease (see, e.g., Chen et al. (2016)), schizophrenia (see, e.g., Stam (2014)) and epilepsy disease

(see, e.g., Munsell et al. (2015)). Hence, assessment of the brain modular organization may provide a key to understanding the relation between aberrant connectivity and brain disease. In this case, one would like to examine brains networks of the individuals with and without brain disorder to derive the differences in community structures. Similar situations occur when one examines several groups of networks, often corresponding to subjects with different biological conditions (e.g., males/females, healthy/diseased, etc.)

One of the possible approaches here is to assume that both, the community structures and the probabilities of connections in the network layers, will be identical under the same biological condition and dissimilar for different conditions. This type of setting, called the **Mixture MultiLayer Stochastic Block Model** (MMLSBM), has recently been studied in Fan et al. (2021) and Jing et al. (2020). Specifically, the MMLSBM assumes that all layers can be partitioned into a few different types, with each distinct type of layers equipped with its own community structure and a distinct matrix of block connection probabilities, and that both are identical within the same type of layers.

The assumption that all layers of the networks corresponding to the same biological condition have identical matrices of block connection probabilities may not be justified in applications. Indeed, while some psychiatric and neurological disorders are associated with the alteration of the community structures in brain networks, the matrices of block connection probabilities may differ from one individual to another. Specifically, one can encounter several groups of networks where the communities remain the same within groups but the block connection probabilities may vary. We call this type of a multiplex network the **DIVERse MultiPLEx (DIMPLE)** network. One can view the above multiplex network as a concatenation of several multiplex networks, each with an unique community structure. Alternatively, one can regard DIMPLE as the generalization of the MMLSBM described above where matrices of the block connection probabilities may vary from one layer to another. Specifically, we consider the following model.

## 1.2 DIVERse MultiPLEx (DIMPLE) network model framework

Consider an  $L$ -layer network on the same set of  $n$  vertices  $\{1, \dots, n\}$  where each layer is equipped with the Stochastic Block Model. Here, the tensor of probabilities of connections  $\mathcal{P} \in [0, 1]^{n \times n \times L}$  is formed by layers  $\mathbf{P}^{(l)}$ ,  $l = 1, \dots, L$ , that can be partitioned into  $M$  groups with the common community assignments. Without loss of generality, we assume that each of these groups has  $K$  communities.

Indeed, since the labels of the groups are interchangeable, in the case of non-identical numbers of communities, it is hard to choose, which of the values correspond to which of the groups. An alternative approach to this model is to view  $K$  as the maximum possible number of communities in each network, with a subsequent adjustment at the stage of finding community assignments in each of the groups of layers.

For any positive integers  $N$  and  $K < N$ , denote  $[N] = \{1, \dots, N\}$ , and let  $\mathcal{M}_{N,K}$  be the set of the *clustering* matrices

$$\mathcal{M}_{N,K} = \{ \mathbf{X} \in \{0, 1\}^{N \times K}, \quad \mathbf{X}\mathbf{1} = \mathbf{1}, \quad \mathbf{X}^T \mathbf{1} \neq \mathbf{0} \},$$

where  $\mathbf{X} \in \mathcal{M}_{N,K}$  are such that  $\mathbf{X}_{i,j} = 1$  if node  $i$  is in cluster  $j$  and  $\mathbf{X}_{i,j} = 0$  otherwise. In this paper, we assume that there exists a label function  $c : [L] \rightarrow [M]$  with

the corresponding clustering matrix  $\mathbf{C} \in \mathcal{M}_{L,M}$  such that the matrix of probabilities of connection in layer  $l$  can be expressed as

$$\mathbf{P}^{(l)} = \mathbf{Z}^{(m)} \mathbf{B}^{(l)} (\mathbf{Z}^{(m)})^T, \quad m = c(l), \quad l = 1, \dots, L, \quad m = 1, \dots, M, \quad (1)$$

where  $\mathbf{Z}^{(m)} \in \mathcal{M}_{n,K}$  is the clustering matrix in the layer of type  $m = c(l)$ , and  $\mathbf{B}^{(l)} = (\mathbf{B}^{(l)})^T \in [0, 1]^{K \times K}$  is a matrix of block probabilities,  $l = 1, \dots, L$ .

One observes the adjacency tensor  $\mathcal{A} \in \{0, 1\}^{n \times n \times L}$  with layers  $\mathbf{A}^{(l)}$  such that  $\mathbf{A}^{(l)}(i, j) = \mathbf{A}^{(l)}(j, i)$  and, for  $1 \leq i < j \leq n$  and  $1 \leq l \leq L$ , where  $\mathbf{A}^{(l)}(i, j)$  are the Bernoulli random variables with  $\mathbb{P}(\mathbf{A}^{(l)}(i, j) = 1) = \mathbf{P}^{(l)}(i, j)$ , and they are independent from each other. The objective is to recover the layer clustering matrix  $\mathbf{C}$  as well as the community assignment matrices  $\mathbf{Z}^{(m)}$ ,  $m = 1, \dots, M$ , on the basis of the observed tensor  $\mathcal{A}$ .

It is easy to see that the model (1) reduces to the multiplex models in Bhattacharyya and Chatterjee (2020), Lei (2020), Lei et al. (2019), Paul and Chen (2016), Paul and Chen (2020) with the persistent communities if  $M = 1$ , and it becomes the MMLSBM if  $\mathbf{B}^{(l)}$  takes only  $M$  distinct values, so that  $\mathbf{B}^{(l)} = \mathbf{B}^{(m)}$  if  $c(l) = m$ .

Since the matrices of the block connection probabilities take different values in each of the layers, techniques employed in Fan et al. (2021) and Jing et al. (2020) cannot be applied in the new environment of DIMPLE. Indeed, the TWIST algorithm of Jing et al. (2020) is based on the alternating regularized low rank approximations of the adjacency tensor. The latter relies on the fact that the tensor of connection probabilities is truly low rank in the case of MMLSBM. This, however, is not true for the DIMPLE model, where the matrices of block connection probabilities vary from layer to layer. On the other hand, the ALMA algorithm of Fan et al. (2021) exploits the fact that the matrices of connection probabilities are identical in the groups of layers with the same community structures. This is no longer the case in the environment of DIMPLE model, where matrices of connection probabilities are all different for different layers.

In our investigation, similarly to the MMLSBM setting, we assume that the groups of layers, as well as the community structures within those groups of layers, are unknown and need to be discovered. On the other hand, unlike in the MMLSBM, the block matrices of the connection probabilities take different values in each of the layers, hence, in this paper, we only study procedures for the between-layer and the within-layer clustering but do not explicitly consider estimation of those block connection probabilities in every layer. Indeed, this task can be easily accomplished by averaging over the estimated community assignments (see, e.g., Gao et al. (2015)).

Our paper makes several key contributions. First, to the best of our knowledge, our paper is the first one that considers SBM-equipped multiplex network, where both the probabilities of connections and the community structures can vary. In this sense, our paper generalizes both the model, which attracted a multitude of publications, where the community structure is identical in all layers, and the MMLSBM, where there are only  $M$  types of the matrices of the connection probabilities, so that the probability tensor has collections of identical layers. As we have mentioned above, this generalization requires new clustering techniques. Specifically, we develop a novel between-layer clustering algorithm. Subsequently, the communities in the groups of layers are found by clustering the averaged adjacency matrices (if the networks layers are strongly assortative), or adjusted averaged squares of the adjacency matrices, if assortativity is in question. We derive the expressions

for the clustering errors and show that they tend to zero at a high rate under very simple and intuitive assumptions. Our simulations confirm that the between-layer and the within-layer clustering algorithms deliver high precision in a finite parameter settings.

Finally, we note that the between-layer clustering procedure developed in the paper can be applied to an extension of the COmmon Subspace Independent Edge (COSIE) random graph model developed in Arroyo et al. (2021). In COSIE, each of the matrices  $\mathbf{P}^{(l)}$  is generated on the basis of the Random Dot-Product Graph (RDPG) model with the common subspace, specifically

$$\mathbf{P}^{(l)} = \mathbf{V}\mathbf{B}_{DP}^{(l)}\mathbf{V}^T, \quad l = 1, \dots, L, \quad (2)$$

where  $\mathbf{V}$  is a common matrix with orthonormal columns, and  $\mathbf{B}_{DP}^{(l)}$  are such that the entries of matrices  $\mathbf{P}^{(l)}$  lie between zero and one. The RDPG model is well studied (see, e.g., an extensive survey Athreya et al. (2018)), and refers to the setting where the probability matrix is generated as  $\mathbf{P} = \mathbf{X}\mathbf{X}^T$ . Here, matrix  $\mathbf{X} \in \mathbf{R}^{n \times K}$  is the matrix of latent positions. Specifically, in the formulation (2), the goal of the authors is to recover the subspace defined by  $\mathbf{V}$  and the multipliers  $\mathbf{B}_{DP}^{(l)}$ . The model loosely corresponds to the multiplex SBM network setting with the common communities across the layers discussed in, e.g., Bhattacharyya and Chatterjee (2020), Lei (2020), Lei et al. (2019), Paul and Chen (2016), Paul and Chen (2020).

In this paper, we extend the COSIE model in (2) to the DIMPLE network setting

$$\mathbf{P}^{(l)} = \mathbf{V}^{(m)}\mathbf{B}_{DP}^{(l)}(\mathbf{V}^{(m)})^T, \quad m = c(l), \quad l = 1, \dots, L, \quad m = 1, \dots, M, \quad (3)$$

where, similarly to (1),  $c : [L] \rightarrow [M]$  is the clustering function with the corresponding clustering matrix  $\mathbf{C} \in \mathcal{M}_{L,M}$ , and  $\mathbf{V}^{(m)}$  is the common basis matrix for the layers of type  $m$ . We call a network, that follows (3), the Heterogeneous Multilayer Random Dot Product Graph (HMRDPG). We show that our between-layers clustering procedure can be applied to model (3) with no modifications. Since the within-layer inference in the model (3) is very different than in the multiples SBM (1), we leave the within-layer inference in the HMRDPG for future investigation.

In this paper, for the purpose of methodological developments, we assume that the number of groups of layers  $M$  in the network as well as the number of communities  $K$  in each group of layers is known. This is a common practice in network literature, and, for example, both Fan et al. (2021) and Jing et al. (2020) make these assumptions. In practical situations, this is usually no longer true, and we discuss those issues in Remarks 1 and 2.

The rest of the paper is organized as follows. In order to analyze the model and construct the clustering algorithms, the next section introduces required notations. Section 2.1 proposes between-layer clustering algorithm while Section 2.2 provides within-layer clustering procedures in the case of assortative and, not necessarily assortative, networks. Section 3 is dedicated to theoretical developments. Specifically, Section 3.1 introduces assumptions that guarantee the between-layer and the within-layer clustering error rates, derived in Sections 3.2 and 3.3, respectively. Section 4 presents simulation studies for the DIMPLE model. Section 5 provides an extension of the DIMPLE setting to the environment of HMRDPG. Section 6 concludes the paper with the discussion of its results. Finally, Section 7 contains proofs of the statements in the paper.

### 1.3 Notations and organization of the paper

We denote tensors by calligraphy letters and matrices by bold letters. For any matrix  $\mathbf{X}$ , denote the Frobenius, the infinity and the operator norm by  $\|\mathbf{X}\|_F$ ,  $\|\mathbf{X}\|_\infty$  and  $\|\mathbf{X}\|$ , respectively, and its  $r$ -th largest singular value by  $\sigma_r(\mathbf{X})$ . The column  $j$  and the row  $i$  of a matrix  $\mathbf{Q}$  are denoted by  $\mathbf{Q}(:, j)$  and  $\mathbf{Q}(i, :)$ , respectively. Denote the identity and the zero matrix of size  $K$  by, respectively,  $\mathbf{I}_K$  and  $\mathbf{0}_K$  (where  $K$  is omitted when this does not cause ambiguity). Denote

$$\mathcal{O}_{n,K} = \{\mathbf{X} \in \mathbb{R}^{n \times K} : \mathbf{X}^T \mathbf{X} = \mathbf{I}_K\}, \quad \mathcal{O}_n = \mathcal{O}_{n,n}. \quad (4)$$

Let  $\text{vec}(\mathbf{X})$  be the vector obtained from matrix  $\mathbf{X}$  by sequentially stacking its columns. Denote by  $\mathbf{X} \otimes \mathbf{Y}$  the Kronecker product of matrices  $\mathbf{X}$  and  $\mathbf{Y}$ . Denote  $n$ -dimensional vector with unit components by  $\mathbf{1}_n$ . Denote diagonal of a matrix  $\mathbf{A}$  by  $\text{diag}(\mathbf{A})$ . Also, denote the  $M$ -dimensional diagonal matrix with  $a_1, \dots, a_M$  on the diagonal by  $\text{diag}(a_1, \dots, a_M)$ .

For any matrix  $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ , denote its projection on the nearest rank  $K$  matrix by  $\Pi_K(\mathbf{X})$ , that is, if  $\sigma_k$  are the singular values, and  $u_k$  and  $v_k$  are the left and the right singular vectors of  $\mathbf{X}$ ,  $k = 1, \dots, r$ , then

$$\mathbf{X} = \sum_{k=1}^r \sigma_k u_k v_k^T \implies \Pi_K(\mathbf{X}) = \sum_{k=1}^{\min(r,K)} \sigma_k u_k v_k^T.$$

For any matrices  $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$  and  $\mathbf{U} \in \mathcal{O}_{n_1,K}$ ,  $K \leq n_1$ , projection of  $\mathbf{X}$  on the column space of  $\mathbf{U}$  and on its orthogonal space are defined, respectively, as

$$\Pi_{\mathbf{U}}(\mathbf{X}) = \mathbf{U} \mathbf{U}^T \mathbf{X}, \quad \Pi_{\mathbf{U}^\perp}(\mathbf{X}) = (\mathbf{I} - \Pi_{\mathbf{U}}) \mathbf{X}.$$

Following Kolda and Bader (2009), we define the following tensor operations. For any tensor  $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$  and a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n_3}$ , their product  $\mathcal{X} \times_3 \mathbf{A}$  along dimension 3 is a tensor in  $\mathbb{R}^{n_1 \times n_2 \times m}$  with elements

$$[\mathcal{X} \times_3 \mathbf{A}](i_1, i_2, j) = \sum_{i_3=1}^{n_3} \mathbf{A}(j, i_3) \mathcal{X}(i_1, i_2, i_3), \quad j = 1, \dots, m.$$

If  $\mathcal{Y} \in \mathbb{R}^{m \times n_2 \times n_3}$  is another tensor, the product between tensors  $\mathcal{X}$  and  $\mathcal{Y}$  along dimensions (2,3), denoted by  $\mathcal{X} \times_{2,3} \mathcal{Y}$ , is a matrix in  $\mathbb{R}^{n_1 \times m}$  with elements

$$[\mathcal{X} \times_{2,3} \mathcal{Y}](i_1, i_2) = \sum_{j_2=1}^{n_2} \sum_{j_3=1}^{n_3} \mathcal{X}(i_1, j_2, j_3) \mathcal{Y}(i_2, j_2, j_3), \quad i_1 = 1, \dots, n_1, \quad i_2 = 1, \dots, m.$$

The mode-3 matricization of tensor  $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$  is a matrix  $\mathcal{M}_3(\mathcal{X}) = \mathbf{X} \in \mathbb{R}^{n_3 \times (n_1 n_2)}$  with rows  $\mathbf{X}(i, :) = [\text{vec}(\mathcal{X}(:, :, i))]^T$ . Please, see Kolda and Bader (2009) for a more extensive discussion of tensor operations and their properties.

We use the  $\sin \Theta$  distances to measure the separation between two subspaces with orthonormal bases  $\mathbf{U} \in \mathcal{O}_{n,K}$  and  $\tilde{\mathbf{U}} \in \mathcal{O}_{n,K}$ , respectively. Suppose the singular values of  $\mathbf{U}^T \tilde{\mathbf{U}}$  are  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K > 0$ . Then

$$\Theta(\mathbf{U}, \tilde{\mathbf{U}}) = \text{diag}(\cos^{-1}(\sigma_1), \dots, \cos^{-1}(\sigma_K))$$

are the principle angles. Quantitative measures of the distance between the column spaces of  $\mathbf{U}$  and  $\tilde{\mathbf{U}}$  are then

$$\left\| \sin \Theta(\mathbf{U}, \tilde{\mathbf{U}}) \right\| = \sqrt{1 - \sigma_{\min}^2(\mathbf{U}^T \tilde{\mathbf{U}})} \quad \text{and} \quad \left\| \sin \Theta(\mathbf{U}, \tilde{\mathbf{U}}) \right\|_F = \sqrt{K - \|\mathbf{U}^T \tilde{\mathbf{U}}\|_F^2}$$

Some convenient characterizations of those distances can be found in Section 8.1 of Cai and Zhang (2018).

Below, we also introduce some notations used in this paper. We consider a multiplex network with  $L$  layers of  $M$  types, with  $L_m$  layers of type  $m$ ,  $m = 1, \dots, M$ . Let  $\mathbf{C} \in \mathcal{M}(L, M)$  be the layer clustering matrix with  $\mathbf{C}^T \mathbf{C} = \text{diag}(L_1, \dots, L_M)$ . Each layer has  $K$  communities, and  $n_{k,m}$  is the number of nodes of type  $k$  in the layer of type  $m$ ,  $k = 1, \dots, K$ ,  $m = 1, \dots, M$ . With this notation, one has in (1)

$$\mathbf{D}_z^{(m)} = (\mathbf{Z}^{(m)})^T \mathbf{Z}^{(m)} = \text{diag}(n_{1,m}, \dots, n_{K,m}). \quad (5)$$

## 2. Between and within-layer clustering

### 2.1 Between-layer clustering

In order to find the clustering matrix  $\mathbf{C}$ , denote  $\mathbf{U}_z^{(m)} = \mathbf{Z}^{(m)}(\mathbf{D}_z^{(m)})^{-1/2}$ , where matrices  $\mathbf{D}_z^{(m)}$  are defined in (5), and  $\mathbf{U}_z^{(m)} \in \mathcal{O}_{n,K}$ ,  $m = 1, \dots, M$ . Observe that matrices  $\mathbf{P}^{(l)}$  in (1) can be written as

$$\mathbf{P}^{(l)} = \mathbf{U}_z^{(m)}(\mathbf{D}_z^{(m)})^{1/2} \mathbf{B}^{(l)}(\mathbf{D}_z^{(m)})^{1/2}(\mathbf{U}_z^{(m)})^T, \quad l = 1, \dots, L. \quad (6)$$

In order to extract common information from matrices  $\mathbf{P}^{(l)}$ , we consider the singular value decomposition (SVD) of  $\mathbf{P}^{(l)}$

$$\mathbf{P}^{(l)} = \mathbf{U}_{P,l} \mathbf{\Lambda}_{P,l} (\mathbf{U}_{P,l})^T, \quad \mathbf{U}_{P,l} \in \mathcal{O}_{n,K}, \quad l = 1, \dots, L, \quad (7)$$

and relate it to expansion (6). Let

$$\mathbf{B}_D^{(l)} \equiv (\mathbf{D}_z^{(m)})^{1/2} \mathbf{B}^{(l)} (\mathbf{D}_z^{(m)})^{1/2} = \mathbf{O}_z^{(l)} \mathbf{S}_z^{(l)} (\mathbf{O}_z^{(l)})^T \quad (8)$$

be the corresponding SVDs of  $\mathbf{B}_D^{(l)}$ , where  $\mathbf{S}_z^{(l)}$  are  $K$ -dimensional diagonal matrices. If, as we assume later, matrices  $\mathbf{B}^{(l)}$  are of full rank, then  $\mathbf{O}_z^{(l)} \in \mathcal{O}_K$ , so that  $\mathbf{O}_z^{(l)} (\mathbf{O}_z^{(l)})^T = (\mathbf{O}_z^{(l)})^T \mathbf{O}_z^{(l)} = \mathbf{I}_K$ . Plugging the SVD of  $\mathbf{B}_D^{(l)}$  into (6), obtain

$$\mathbf{P}^{(l)} = \mathbf{U}_z^{(m)} \mathbf{O}_z^{(l)} \mathbf{S}_z^{(l)} (\mathbf{O}_z^{(l)})^T (\mathbf{U}_z^{(m)})^T, \quad l = 1, \dots, L. \quad (9)$$

Since  $\mathbf{U}_z^{(m)} \mathbf{O}_z^{(l)} \in \mathcal{O}_{n,K}$ , expansion (9) is another way of writing the SVD of  $\mathbf{P}^{(l)}$  and, up to permutation of columns,  $\mathbf{U}_z^{(m)} \mathbf{O}_z^{(l)}$  equals to  $\mathbf{U}_{P,l}$ ,  $l = 1, \dots, L$ . It is too computationally expensive, however, to consider every permutation of columns in  $\mathbf{U}_{P,l}$ . Fortunately, this is also unnecessary, since

$$\mathbf{U}_{P,l} (\mathbf{U}_{P,l})^T = \mathbf{U}_z^{(m)} \mathbf{O}_z^{(l)} (\mathbf{O}_z^{(l)})^T (\mathbf{U}_z^{(m)})^T = \mathbf{U}_z^{(m)} (\mathbf{U}_z^{(m)})^T, \quad m = c(l), \quad (10)$$

---

**Algorithm 1:** The between-layer clustering

---

**Input:** Adjacency tensor  $\mathcal{A} \in \{0, 1\}^{n \times n \times L}$ , number of groups of layers  $M$ , number of communities  $K$

**Output:** Estimated clustering matrix  $\hat{\mathbf{C}} \in \mathcal{M}_{L,M}$

**Steps:**

**1:** For  $l = 1, \dots, L$ , find the SVDs of  $\mathbf{A}^{(l)}$  using (13)

**2:** Form matrix  $\hat{\Theta} \in \mathbb{R}^{n^2 \times L}$  with columns  $\hat{\Theta}(:, l) = \text{vec}(\hat{\mathbf{U}}_{A,l}(\hat{\mathbf{U}}_{A,l})^T)$

**3:** Construct the SVD of  $\hat{\Theta}$  using (14) and obtain matrix  $\hat{\mathbf{W}} = \widetilde{\mathbf{W}}(:, 1 : M) \in \mathcal{O}_{L,M}$

**4:** Cluster  $L$  rows of  $\hat{\mathbf{W}}$  into  $M$  clusters using  $(1 + \epsilon)$ -approximate  $K$ -means clustering. Obtain estimated clustering matrix  $\hat{\mathbf{C}}$

---

so that matrices  $\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T$  depend on  $l$  only via  $m = c(l)$  and are uniquely defined for  $l = 1, \dots, L$ .

The latter implies that the between-layer clustering can be based on the matrices  $\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T$ ,  $l = 1, \dots, L$ , or rather on their vectorized versions. Denote

$$\mathbf{D}_c = \mathbf{C}^T \mathbf{C} = \text{diag}(L_1, \dots, L_M), \quad \mathbf{W} = \mathbf{C}(\mathbf{D}_c)^{-1/2} \in \mathcal{O}_{L,M} \quad (11)$$

Consider matrices  $\mathbf{Q} \in \mathbb{R}^{n^2 \times M}$  and  $\Theta \in \mathbb{R}^{n^2 \times L}$  with respective columns

$$\mathbf{Q}(:, m) = \text{vec}(\mathbf{U}_z^{(m)}(\mathbf{U}_z^{(m)})^T), \quad \Theta(:, l) = \text{vec}(\mathbf{U}_z^{(c(l))}(\mathbf{U}_z^{(c(l))})^T) = \text{vec}(\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T),$$

where  $m = 1, \dots, M$ ,  $l = 1, \dots, L$ . It is easy to see that

$$\Theta = \mathbf{Q}\mathbf{C}^T, \quad \mathbf{Q} = \Theta\mathbf{C}\mathbf{D}_c^{-1}, \quad (12)$$

so that clustering assignment can be recovered by spectral clustering of columns of an estimated version of matrix  $\Theta$ .

For this purpose, consider layers  $\mathbf{A}^{(l)}(i, j) = \mathcal{A}(:, :, l)$  of the adjacency tensor  $\mathcal{A}$  and construct their SVDs

$$\mathbf{A}^{(l)} = \hat{\mathbf{U}}_{A,l} \hat{\Lambda}_{P,l} (\hat{\mathbf{U}}_{A,l})^T, \quad \hat{\mathbf{U}}_{A,l} \in \mathcal{O}_{n,K}, \quad l = 1, \dots, L. \quad (13)$$

Then, replace matrix  $\Theta$  by its proxy  $\hat{\Theta}$  with columns  $\hat{\Theta}(:, l) = \text{vec}(\hat{\mathbf{U}}_{A,l}(\hat{\mathbf{U}}_{A,l})^T)$ . The major difference between  $\Theta$  and  $\hat{\Theta}$ , however, is that  $\text{rank}(\Theta) = M$  while, in general,  $\text{rank}(\hat{\Theta}) = L \gg M$ . If the SVD of  $\hat{\Theta}$  is

$$\hat{\Theta} = \tilde{\mathbf{V}} \tilde{\Lambda} \widetilde{\mathbf{W}}, \quad \tilde{\mathbf{V}} \in \mathcal{O}_{n^2,L}, \quad \widetilde{\mathbf{W}} \in \mathcal{O}_L, \quad (14)$$

then, we can form reduced matrices

$$\hat{\mathbf{V}} = \tilde{\mathbf{V}}(:, 1 : M) \in \mathcal{O}_{n^2,M}, \quad \hat{\mathbf{W}} = \widetilde{\mathbf{W}}(:, 1 : M) \in \mathcal{O}_{L,M}, \quad (15)$$

and apply clustering to the rows of  $\hat{\mathbf{W}}$  rather than to the rows of  $\widetilde{\mathbf{W}}$ . The latter results in Algorithm 1. We use  $(1 + \epsilon)$ -approximate  $K$ -means clustering to obtain the final clustering assignments. There exist efficient algorithms for solving the  $(1 + \epsilon)$ -approximate  $K$ -means problem (see, e.g., Kumar et al. (2004)).



---

**Algorithm 2:** The within-layer clustering

---

**Input:** Adjacency tensor  $\mathcal{A} \in \{0, 1\}^{n \times n \times L}$ , number of groups of layers  $M$ , number of communities  $K$ , estimated layer clustering matrix  $\hat{\mathbf{C}} \in \mathcal{M}_{L, M}$

**Output:** Estimated community assignments  $\hat{\mathbf{Z}}^{(m)} \in \mathcal{M}_{n, K}$ ,  $m = 1, \dots, M$

**Steps:**

- 1: Construct tensor  $\hat{\mathcal{H}}$  using formula (16)
  - 2: Construct the SVDs of layers  $\hat{\mathbf{H}}^{(m)} = \tilde{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)} \hat{\mathbf{\Lambda}}_{\hat{\mathbf{H}}}^{(m)} (\tilde{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)})^T$ ,  $m = 1, \dots, M$
  - 3: Find  $\hat{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)} = \tilde{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)}(:, 1 : K) = \Pi_K(\tilde{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)})$ ,  $m = 1, \dots, M$
  - 4: Cluster rows of  $\hat{\mathbf{U}}_{\hat{\mathbf{H}}}^{(m)}$  into  $K$  clusters using  $(1 + \epsilon)$ -approximate  $K$ -means clustering. Obtain clustering matrices  $\hat{\mathbf{Z}}^{(m)}$ ,  $m = 1, \dots, M$
- 

**Remark 1. Unknown number of layers.** While Algorithm 1 assumes  $M$  to be known, in many practical situations this is not true, and the value of  $M$  has to be discovered from data. Identifying the number of clusters is a common issue in data clustering, and it is a separate problem from the process of actually solving the clustering problem with a known number of clusters. A common method for finding the number of clusters is the so called “elbow” method that looks at the fraction of the variance explained as a function of the number of clusters. The method is based on the idea that one should choose the smallest number of clusters, such that adding another cluster does not significantly improve fitting of the data by a model. There are many ways to determine the “elbow”. For example, one can base its detection on evaluation of the clustering error in terms of an objective function, as in, e.g., Zhang et al. (2012). Another possibility is to monitor the eigenvalues of the non-backtracking matrix or the Bethe Hessian matrix, as it is done in Le and Levina (2015). One can also employ a simple technique of checking the eigen-gaps of the matrix  $\hat{\mathbf{A}}$  in (14), as it has been discussed in von Luxburg (2007).

## 2.2 Within-layer clustering

If we knew the true clustering matrix  $\mathbf{C}$  and the true probability tensor  $\mathcal{P} \in \mathbb{R}^{n \times n \times L}$  with layers  $\mathbf{P}^{(l)}$  given by (1), then we could average layers with identical communities. The averaging procedure, however, depends on how similar matrices  $\mathbf{B}^{(l)}$  are. Specifically, if all networks in the  $m$ -th group of layers are assortative, so that the eigenvalues of matrices  $\mathbf{B}^{(l)}$  are all positive for  $c(l) = m$ , then the lowest eigenvalues of their sum will add together, leading to successful clustering (see, e.g., Paul and Chen (2020)). The case of assortative networks can be assured by Assumption **A6** in Section 3.3.

If Assumption **A6** holds, one can average the layers with the same community structure. Specifically, one can form tensor  $\hat{\mathcal{H}} \in \mathbb{R}^{n \times n \times M}$  as

$$\hat{\mathcal{H}} = \mathcal{A} \times_3 \hat{\mathbf{W}}^T \quad \text{with} \quad \hat{\mathbf{W}} = \hat{\mathbf{C}}(\hat{\mathbf{D}}_{\hat{\mathbf{C}}})^{-1/2} \in \mathcal{O}_{L, M}, \quad \hat{\mathbf{D}}_{\hat{\mathbf{C}}} = \hat{\mathbf{C}}^T \hat{\mathbf{C}}. \quad (16)$$

---

**Algorithm 3:** The within-layer clustering

---

**Input:** Adjacency tensor  $\mathcal{A} \in \{0, 1\}^{n \times n \times L}$ , number of groups of layers  $M$ , number of communities  $K$ , estimated layer clustering matrix  $\hat{\mathbf{C}} \in \mathcal{M}_{L, M}$

**Output:** Estimated community assignments  $\hat{\mathbf{Z}}^{(m)} \in \mathcal{M}_{n, K}$ ,  $m = 1, \dots, M$

**Steps:**

- 1: Construct tensor  $\hat{\mathcal{G}}$  with layers (18),  $l = 1, \dots, L$
  - 2: Construct tensor  $\hat{\mathcal{H}}$  using formula (19)
  - 3: Follow steps 2–4 of Algorithm 2
- 

The technique relies on the fact that the population version  $\mathcal{H}$  of tensor  $\hat{\mathcal{H}}$  has layers  $\mathbf{H}^{(m)} = \mathcal{H}(:, :, m)$ ,  $m = 1, \dots, M$ , where

$$\mathbf{H}^{(m)} = \sqrt{L_m} \mathbf{U}_z^{(m)} \bar{\mathbf{B}}^{(m)} (\mathbf{U}_z^{(m)})^T \quad \text{with} \quad \bar{\mathbf{B}}^{(m)} = L_m^{-1} \sum_{c(l)=m} \mathbf{B}_D^{(l)}. \quad (17)$$

Here, by (8),  $\mathbf{B}_D^{(l)} = (\mathbf{D}_z^{(m)})^{1/2} \mathbf{B}^{(l)} (\mathbf{D}_z^{(m)})^{1/2}$  and  $\mathbf{D}_z^{(m)}$  is defined in (5). Due to Assumption **A6**, the eigenvalues of matrices  $\mathbf{B}_D^{(l)}$  add up. Subsequently, communities in layers  $\hat{\mathbf{H}}^{(m)} = \hat{\mathcal{H}}(:, :, m)$ ,  $m = 1, \dots, M$ , can be found using spectral clustering, resulting in Algorithm 2.

Assumption **A6** is not necessarily valid in every practical situations. If it does not hold, then, as simulation studies in Paul and Chen (2020) show, averaging the adjacency matrices may not lead to improved precision of community detection in groups of layers. This is due to the fact that, if some of the layers are assortative and some are not, the smallest singular value of the matrices  $\bar{\mathbf{B}}^{(m)}$  in (17) may be smaller than individual singular values of matrices  $\mathbf{B}_D^{(l)}$ , leading to poor clustering results. For this reason, in the absence of Assumption **A6**, instead of constructing tensor  $\hat{\mathcal{H}}$  using formula (16), it is advantageous to use a different procedure. Specifically, one can average adjusted squares of adjacency matrices  $(\mathbf{A}^{(l)})^2$  with  $\hat{c}(l) = m$ , as it is suggested in Lei and Lin (2021), or average matrices  $\hat{\mathbf{U}}_{A,l} (\hat{\mathbf{U}}_{A,l})^T$  with  $\hat{c}(l) = m$ . Note that in the latter case, one does not have the advantage of reduction of stochastic errors by averaging layers with similar spectral structure. Therefore, following Lei and Lin (2021), we evaluate the degree vector for each layer  $\hat{\mathbf{d}}^{(l)} = \mathbf{A}^{(l)} \mathbf{1}_n$  and form diagonal matrices  $\text{diag}(\hat{\mathbf{d}}^{(l)})$  with vectors  $\hat{\mathbf{d}}^{(l)}$  on the diagonals. We construct a tensor  $\hat{\mathcal{G}} \in \mathbb{R}^{n \times n \times L}$  with layers  $\hat{\mathbf{G}}^{(l)} = \hat{\mathcal{G}}(:, :, l)$  of the form

$$\hat{\mathbf{G}}^{(l)} = \left( \mathbf{A}^{(l)} \right)^2 - \text{diag}(\hat{\mathbf{d}}^{(l)}), \quad l = 1, \dots, L. \quad (18)$$

Subsequently, we average layers of the same types, obtaining tensor  $\hat{\mathcal{H}} \in \mathbb{R}^{n \times n \times M}$

$$\hat{\mathcal{H}} = \hat{\mathcal{G}} \times_3 \hat{\mathbf{W}}^T, \quad (19)$$

where  $\hat{\mathbf{W}}$  is defined in (16). We apply spectral clustering to layers of tensor  $\hat{\mathcal{H}}$  similarly to Algorithm 2. The alternative within-layer clustering procedure, described above, is summarized in Algorithm 3

**Remark 2. Unknown number of communities in groups of layers.** In this paper, for the purpose of methodological developments, we assume that the number of communities in each group of layers is the same and is equal to a known number  $K$ . In real life applications, this assumption may not hold. Indeed, after one determines groups of similar layers, the problem reduces to finding the number of communities in each of the individual groups of the networks. At this stage, the assumption that all layers in each group have the same number of communities becomes unnecessary, and one can find a distinct number of communities in each of the groups of layers using standard techniques. Note that having a different number of communities  $K_m$  for each of the groups of layers does not compromise our theoretical developments as long as  $K_m$  are comparable, so that  $\underline{c}K \leq K_m \leq \bar{c}K$  for some absolute constants  $\underline{c}$  and  $\bar{c}$  and  $m = 1, \dots, M$ . Incorporating those different numbers of communities into our theory will, however, cause some confusion, as we have mentioned in the Section 1.2, since labels of the groups of layers are interchangeable.

### 3. Theoretical analysis

In this section, we study the between-layer error rates for network layer clustering and the within-layer clustering error rates. Since the clustering is unique only up to a permutation of clusters, denote the set of  $K$ -dimensional permutation functions of  $[K]$  by  $\aleph(K)$  and the set of  $K \times K$  permutation matrices by  $\mathfrak{F}(K)$ . The misclassification error rate of the between-layer clustering is then given by

$$R_{BL} = (2L)^{-1} \min_{\mathcal{P} \in \mathfrak{F}(M)} \|\hat{\mathbf{C}} - \mathbf{C} \mathcal{P}\|_F^2. \quad (20)$$

Similarly, the local community detection error in the layer of type  $m$  is

$$R_{WL}(m) = (2n)^{-1} \min_{\mathcal{P}_m \in \mathfrak{F}(K)} \|\hat{\mathbf{Z}}^{(m)} - \mathbf{Z}^{(m)} \mathcal{P}_m\|_F^2, \quad m = 1, \dots, M. \quad (21)$$

Note that since the numbering of layers is defined also up to a permutation, the errors  $R_{WL}(1), \dots, R_{WL}(M)$  should be minimized over the set of permutations  $\aleph(M)$ . The average error rate of the within-layer clustering is then given by

$$R_{WL} = \frac{1}{M} \min_{\aleph(M)} \sum_{m=1}^M R_{WL}(m) = \frac{1}{2Mn} \min_{\aleph(M)} \sum_{m=1}^M \left[ \min_{\mathcal{P}_m \in \mathfrak{F}(K)} \|\hat{\mathbf{Z}}^{(m)} - \mathbf{Z}^{(m)} \mathcal{P}_m\|_F^2 \right] \quad (22)$$

#### 3.1 Assumptions

The accuracies of our algorithms depend on the success of the between-layer clustering, which, in turn, relies on the fact that matrices  $\mathbf{U}_z^{(m)}(\mathbf{U}_z^{(m)})^T$ ,  $m = 1, \dots, M$ , are not too similar to each other for different values of  $m$ . Clearly, columns of matrices  $\mathbf{U}_z^{(m)}$  are not linearly independent for different values of  $m$  since  $\mathbf{U}_z^{(m)} \mathbf{1}_K = \mathbf{1}_n$  for any  $m$ .

Consider matrix  $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times MK}$  which is obtained as horizontal concatenation of matrices  $\mathbf{U}_z^{(m)} \in \mathbb{R}^{n \times K}$ ,  $m = 1, \dots, M$ . Let the SVD of  $\bar{\mathbf{Z}}$  be

$$\bar{\mathbf{Z}} = [\mathbf{U}_z^{(1)} | \dots | \mathbf{U}_z^{(M)}] = \bar{\mathbf{U}} \bar{\mathbf{D}} \bar{\mathbf{V}}^T, \quad \bar{\mathbf{U}} \in \mathcal{O}_{n,r}, \bar{\mathbf{V}} \in \mathcal{O}_{MK,r}. \quad (23)$$

Here,  $r$  is the rank of  $\bar{\mathbf{Z}}$  and  $\bar{\mathbf{D}}$  is an  $r$ -dimensional diagonal matrix.

We impose the following assumptions.

**A1.** Clusters of layers and local communities are balanced, so that there exist absolute constants  $\underline{c}$  and  $\bar{c}$  such that

$$\underline{c}L/M \leq L_m \leq \bar{c}L/M, \quad \underline{c}n/K \leq n_{k,m} \leq \bar{c}n/K,$$

where  $L_m$  is the number of networks in the layer of type  $m$ , and  $n_{k,m}$  is the number of nodes in the  $k$ -th community in the layer of type  $m$ ,  $m = 1, \dots, M$ ,  $k = 1, \dots, K$ .

**A2.** For some absolute constant  $\kappa_0$ , one has  $\sigma_1(\bar{\mathbf{D}}) \leq \kappa_0 \sigma_r(\bar{\mathbf{D}})$  in (23).

**A3.** The block probability matrices  $\mathbf{B}^{(l)}$  in (1) are such that

$$\mathbf{B}^{(l)} = \rho_{n,l} \mathbf{B}_0^{(l)}, \quad \|\mathbf{B}_0^{(l)}\|_\infty = 1, \quad \min_{l=1, \dots, L} \rho_{n,l} \geq C_\rho n^{-1} \log n, \quad l = 1, \dots, L, \quad (24)$$

for some absolute constant  $C_\rho$ .

**A4.** For some absolute constant  $C_\lambda \in (0, 1)$ , one has

$$\min_{l=1, \dots, L} [\sigma_K(\mathbf{B}_0^{(l)}) / \sigma_1(\mathbf{B}_0^{(l)})] \geq C_\lambda. \quad (25)$$

**A5.** There exist absolute constants  $\underline{c}_\rho$  and  $\bar{c}_\rho$  such that

$$\underline{c}_\rho \rho_n \leq \rho_{n,l} \leq \bar{c}_\rho \rho_n \quad \text{with} \quad \rho_n = L^{-1} \sum_{l=1}^L \rho_{n,l} \quad (26)$$

Assumptions above are very common and are present in many other network papers. Specifically, Assumption **A1** is identical to Assumptions **A3**, **A4** in Jing et al. (2020), or Assumption **A3** in Fan et al. (2021). Assumption **A2** is identical to Assumption **A2** in Jing et al. (2020). Assumption **A3** is present in majority of papers that study community detection in individual networks (see, e.g. Lei and Rinaldo (2015)). It is required here since we rely on similarity of the sets of eigenvectors of the similar layers, and, hence, need the sample eigenvectors to converge to the true ones. Assumption **A4** is equivalent to Assumption **A1** in Jing et al. (2020), or Assumption **A4** in Fan et al. (2021). Finally, Assumption **A5** requires that the sparsity factors are of approximately the same order of magnitude. The latter guarantees that the discrepancies between the true and the sample-based eigenvectors are similar across all layers of the network. Hypothetically, Assumption **A5** can be removed, and one can trace the impact of different scales  $\rho_{n,l}$  on the clustering errors. This, however, will make clustering error bounds very complicated, so we leave this case for future investigation.

### 3.2 The between-layer clustering error

Evaluation of the between-layer clustering error relies on the Tucker decomposition of the tensor with layers  $\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T$ ,  $l = 1, \dots, L$ . Consider tensor  $\mathfrak{S} \in \mathbb{R}^{n \times n \times L}$  with layers

$$\mathfrak{S}(:, :, l) = \mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T = \mathbf{U}_z^{(m)}(\mathbf{U}_z^{(m)})^T, \quad m = c(l), \quad l = 1, \dots, L \quad (27)$$

and its clustered version  $\mathcal{G} \in \mathbb{R}^{n \times n \times M}$  of the form

$$\mathcal{G} = \mathfrak{S} \times_3 [\mathbf{C}(\mathbf{D}_c)^{-1}]^T, \quad (28)$$

where  $\mathbf{D}_c$  is defined in (11). Here, tensor  $\mathcal{G}$  has layers identical to the set of distinct layers of tensor  $\mathfrak{S}$ , so that  $\mathcal{G}(:, :, m) = \mathbf{U}_z^{(m)}(\mathbf{U}_z^{(m)})^T$ ,  $m = 1, \dots, M$ .

Recall that, according to (27) and (28), one has  $\mathfrak{S} = \mathcal{G} \times_3 \mathbf{C}$  and, using matrix  $\overline{\mathbf{Z}}$  in (23), one can rewrite  $\mathfrak{S}$  as  $\mathfrak{S} = \mathcal{B} \times_1 \overline{\mathbf{Z}} \times_2 \overline{\mathbf{Z}} \times_3 \mathbf{C}$  where  $\mathcal{B} \in \mathbb{R}^{KM \times KM \times M}$  is the core tensor with layers

$$\mathcal{B}(:, :, m) = \text{diag}(\underbrace{\mathbf{0}_K, \dots, \mathbf{0}_K}_{m-1}, \mathbf{I}_K, \underbrace{\mathbf{0}_K, \dots, \mathbf{0}_K}_{M-m}) \in \{0, 1\}^{KM \times KM}$$

Using the SVD in (23) and the definition of  $\mathbf{W}$  in (11), we obtain

$$\mathfrak{S} = \mathcal{F} \times_1 \overline{\mathbf{U}} \times_2 \overline{\mathbf{U}} \times_3 \mathbf{W}, \quad \mathcal{F} = \overline{\mathcal{R}} \times_1 \overline{\mathbf{D}} \times_2 \overline{\mathbf{D}} \times_3 \mathbf{D}_c^{1/2}, \quad \overline{\mathcal{R}} = \mathcal{B} \times_1 \overline{\mathbf{V}}^T \times_2 \overline{\mathbf{V}}^T, \quad (29)$$

where  $\mathcal{F}, \overline{\mathcal{R}} \in \mathbb{R}^{r \times r \times M}$ . Now, in order to use representation (29) for analyzing matrix  $\Theta$  in (12), note that  $\Theta$  is the transpose of mode 3 matricization of  $\mathfrak{S}$ ,  $\Theta = \mathfrak{S}_{(3)}^T$ . Using Proposition 1 of Kolda and Bader (2009), obtain

$$\Theta = (\overline{\mathbf{U}} \otimes \overline{\mathbf{U}}) \mathbf{F} \mathbf{W}^T, \quad \mathbf{F} = \mathcal{F}_{(3)}^T \in \mathbb{R}^{r^2 \times M}. \quad (30)$$

Here, by (11) and (23),  $\mathbf{W} = \mathbf{C} \mathbf{D}_c^{-1/2} \in \mathcal{O}_{L,M}$  and  $\overline{\mathbf{U}} \in \mathcal{O}_{n,r}$ . The following statement explores the structure of matrix  $\mathbf{F}$  in (30).

**Lemma 1.** *Matrix  $\mathbf{F}$  can be presented as  $\mathbf{F} = (\overline{\mathbf{D}} \otimes \overline{\mathbf{D}}) \overline{\mathbf{R}} \mathbf{D}_c^{1/2}$  where  $\overline{\mathbf{R}} = (\overline{\mathbf{V}} \otimes \overline{\mathbf{V}})^T \mathbf{R}$  and  $\mathbf{R} = \mathcal{B}_{(3)}^T$ . Here,  $\text{rank}(\mathbf{F}) = M$ , and, under Assumptions A1–A4, one has*

$$\sigma_{\min}^2(\mathbf{F}) = \sigma_M^2(\mathbf{F}) \geq \frac{\underline{c}}{\bar{c} \kappa_0^4 M} \|\mathbf{F}\|_F^2 = \frac{\underline{c} K L}{\bar{c} \kappa_0^4 M} \quad (31)$$

Let the SVD of  $\mathbf{F}$  be of the form  $\mathbf{F} = \mathbf{U}_F \mathbf{\Lambda}_F \mathbf{V}_F$ , where  $\mathbf{U}_F \in \mathcal{O}_{r^2,M}$  and  $\mathbf{V}_F \in \mathcal{O}_M$ . Then, the SVD of  $\Theta$  in (30) can be written as

$$\Theta = \mathbf{V} \mathbf{\Lambda} \mathbf{W}, \quad \mathbf{V} = (\overline{\mathbf{U}} \otimes \overline{\mathbf{U}}) \mathbf{U}_F \in \mathcal{O}_{n^2,M}, \quad \mathbf{W} = \mathbf{W} \mathbf{V}_F \in \mathcal{O}_{L,M}, \quad \mathbf{\Lambda} = \mathbf{\Lambda}_F \quad (32)$$

Representation (32) allows one to bound above the between-layer clustering error.

**Theorem 1.** *Let Assumptions A1–A5 hold. Then, for any  $\tau > 0$ , there exists a constant  $\check{C}$  that depends only on  $\tau$ ,  $\kappa_0$ ,  $C_\rho$ ,  $C_\lambda$ ,  $\bar{c}$ ,  $\underline{c}$ ,  $\bar{c}_\rho$  and  $\underline{c}_\rho$  in Assumptions A1–A5, such that the between-layer clustering error, defined in (20), satisfies*

$$\mathbb{P} \left\{ R_{BL} \leq \frac{\check{C} K^4 M^2}{n^2 \rho_n^2} \right\} \geq 1 - L n^{-\tau} \quad (33)$$

### 3.3 The within-layer clustering error

As we have discussed in Section 2.2, our approach to the within-layer clustering differs, depending on whether layer networks are all assortative. The latter is guaranteed by the assumption below.

**A6.** All networks are assortative, so that the smallest eigenvalues of matrices  $\mathbf{B}^{(l)}$  are all positive, i.e.,  $\lambda_{\min}(\mathbf{B}^{(l)}) > 0$ ,  $l = 1, \dots, L$ .

Assumption **A6** is very different from Assumptions **A1–A5**. Indeed, Assumption **A6** is not required for our inference or analysis. If this assumption holds, one can apply Algorithm 2 to carry out within-layer clustering. If this assumption is not valid (or in question), one should use the alternative within-layer clustering in Algorithm 3.

If Assumption **A6** holds and all networks are assortative, then the lowest eigenvalue of  $\sum \mathbf{B}^{(l)} I(c(l) = m)$  is bounded below by the sum of lowest eigenvalues of  $\mathbf{B}^{(l)}$  with  $c(l) = m$ . Thus, one obtains the following bounds for the within-layer clustering error of Algorithm 2.

**Theorem 2.** *Let Assumptions **A1–A6** hold, and matrices  $\widehat{\mathbf{Z}}^{(m)}$ ,  $m = 1, \dots, M$ , be obtained using Algorithm 2. Then, for any  $\tau > 0$ , there exists a constant  $\check{C}$  that depends only on  $\tau$ ,  $\kappa_0$ ,  $C_\rho$ ,  $C_\lambda$ ,  $\bar{c}$ ,  $\underline{c}$ ,  $\bar{c}_\rho$  and  $\underline{c}_\rho$  in Assumptions **A1–A5**, such that the average within-layer clustering error, defined in (22), satisfies*

$$\mathbb{P} \left\{ R_{WL} \leq \check{C} \left( \frac{M^3 K^6}{n^2 \rho_n^2} + \frac{MK^2 \log n}{nL\rho_n} \right) \right\} \geq 1 - Ln^{-\tau} - 8Mn^{1-2\tau} \quad (34)$$

**Remark 3. Disassortative networks.** Consider the situation when all networks are disassortative, so that the probabilities of connections of nodes in the same community are lower than those, for nodes in different communities. Note that, even in this situation, the eigenvalues of  $\mathbf{B}^{(l)}$  cannot be all negative, due to the Perron-Frobenius theorem (Rao and Rao (1998), **P.15.1.14**). However, if all networks are strongly disassortative, i.e., for any node, its probability of connection within its own block is much smaller than outside it, Algorithm 2 results in a better precision than in the case when networks in the layers may be assortative or disassortative. We demonstrate this feature in our simulations study in Section 4.

If Assumption **A6** is not valid, we follow Algorithm 3 that starts with evaluating squares of individual adjacency matrices and adjusting the diagonals of the squares. The necessity for the latter is due to the fact that those diagonals have positive biases, which need to be corrected. While we are using the same bias adjustment as in Lei and Lin (2021), the details of our theory are somewhat different since the sparsity assumption in Lei and Lin (2021), that  $\rho_n n \leq C$  for some absolute constant  $C$ , contradicts our Assumption **A3**.

**Theorem 3.** *Let Assumptions **A1–A5** hold, and matrices  $\widehat{\mathbf{Z}}^{(m)}$ ,  $m = 1, \dots, M$ , be obtained using by Algorithm 3. Let  $\tau \geq 2$ . Then, for any  $\tau > 0$ , there exists a constant  $\check{C}$ , which depends only on  $\tau$ ,  $\kappa_0$ ,  $C_\rho$ ,  $C_\lambda$ ,  $\bar{c}$ ,  $\underline{c}$ ,  $\bar{c}_\rho$  and  $\underline{c}_\rho$  in Assumptions **A1–A5**, and  $C_{\tau,\epsilon}$ , which depends only on  $\tau$  and  $\epsilon$ , such that the average within-layer clustering error, defined in (22), satisfies*

$$\mathbb{P} \left\{ R_{WL} \leq \check{C} \left( \frac{M^3 K^8}{n^2 \rho_n^2} + \frac{MK^4 \log n}{nL\rho_n} \right) \right\} \geq 1 - C_{\tau,\epsilon} [Ln^{-\tau} + n^{2-\tau}] \quad (35)$$

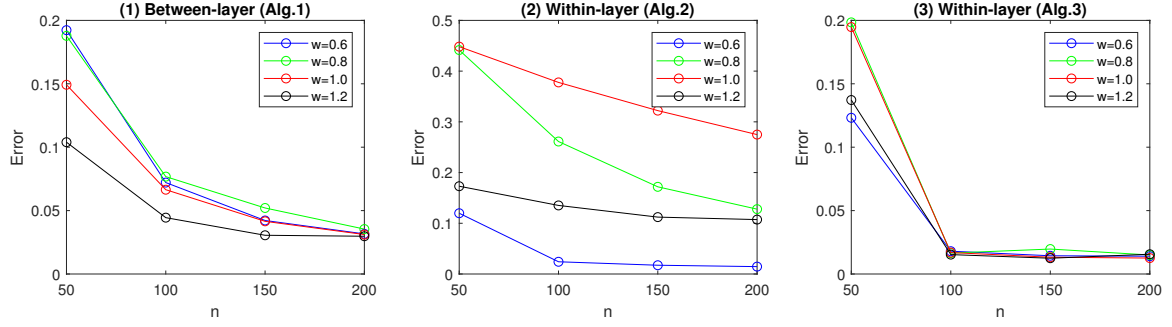


Figure 1: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 2 and 3 for  $L = 100$ ,  $c = 0$ ,  $d = 0.8$  and  $n = 50, 100, 150, 200$ .

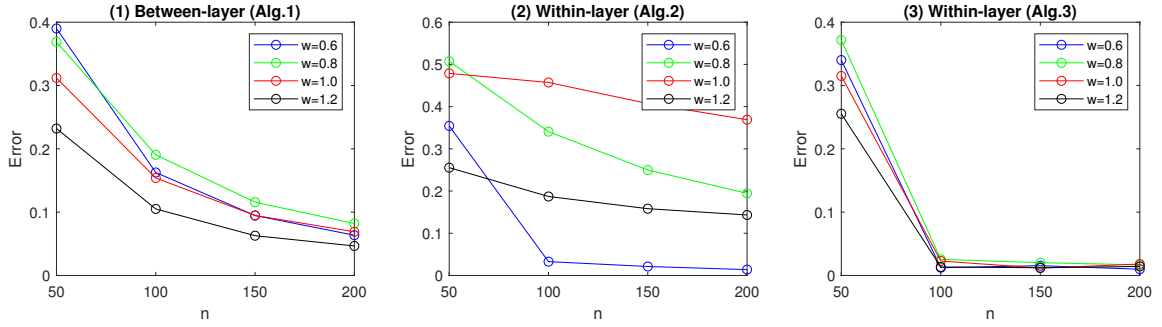


Figure 2: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 2 and 3 for  $L = 100$ ,  $c = 0$ ,  $d = 0.5$  and  $n = 50, 100, 150, 200$ .

In many practical situations, one may be unsure whether Assumption **A6** is valid or not. In this case, we suggest to apply Algorithm 3 rather than Algorithm 2. Indeed, up to factors of  $M$  and  $K$ , Algorithm 3 results in the within-layer average clustering error of the same order of magnitude as Algorithm 2 under Assumption **A6**. In addition, as simulation studies in Section 4 show, even when Assumption **A6** holds, the within-layer clustering errors of both algorithms are similar, while they are much higher for Algorithm 2, when Assumption **A6** is invalid.

#### 4. Simulation study

In order to study performances of our algorithms for various combinations of parameters, we carry out a limited simulation study. In each of the simulations scenarios, we vary the number of nodes  $n$  or the number of layers  $L$ . We fix the number of groups of layers  $M$  and generate a multilayer network such that all layers have the same number of communities

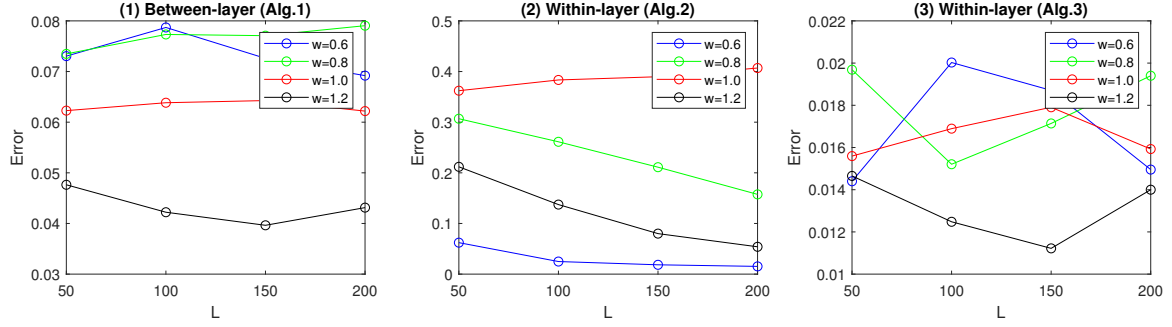


Figure 3: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 2 and 3 for  $n = 100$ ,  $c = 0$ ,  $d = 0.8$  and  $L = 50, 100, 150, 200$ .

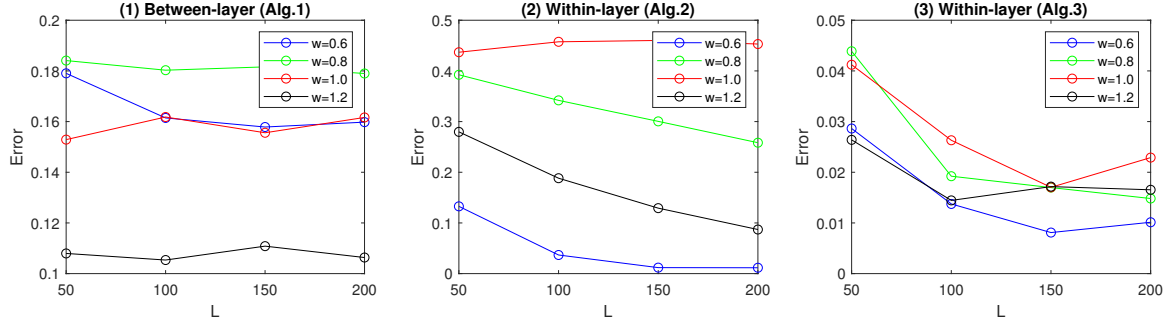


Figure 4: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 2 and 3 for  $n = 100$ ,  $c = 0$ ,  $d = 0.8$  and  $L = 50, 100, 150, 200$ .

$K$ , and that both the groups of the layers and the communities in the layers are balanced.

To obtain a network like this, we first fix  $n$ ,  $L$ ,  $M$ ,  $K$ , the sparsity parameters  $c$  and  $d$  and the assortativity parameter  $w$ . Then we generate  $M$  groups of layers using the multinomial distribution with equal probabilities  $1/M$ . Similarly, we generate  $K$  communities in each of the groups of layers using the multinomial distribution with equal probabilities  $1/K$ . In this manner, we obtain community assignment matrices  $\mathbf{Z}^{(m)}$ ,  $m = 1, \dots, M$ , in each layer  $l$  with  $c(l) = m$ , where  $c : L \rightarrow M$  is the layer assignment function.

Next, we generate the entries of  $\mathbf{B}^{(l)}$ ,  $l = 1, \dots, L$ , as uniform random numbers between  $c$  and  $d$ , and then multiply all the non-diagonal entries of those matrices by  $w$ . In this manner, if  $w < 1$  is small, then the network is strongly assortative, i.e. there is higher probability for nodes in the same community to connect. If  $w > 1$  is large, then the network is disassortative, i.e. the probability of connection for nodes in different communities is higher than for nodes in the same community. Finally, since entries of matrices  $\mathbf{B}^{(l)}$  are generated at random, when  $w$  is close to one, the networks in all layers are neither assortative or



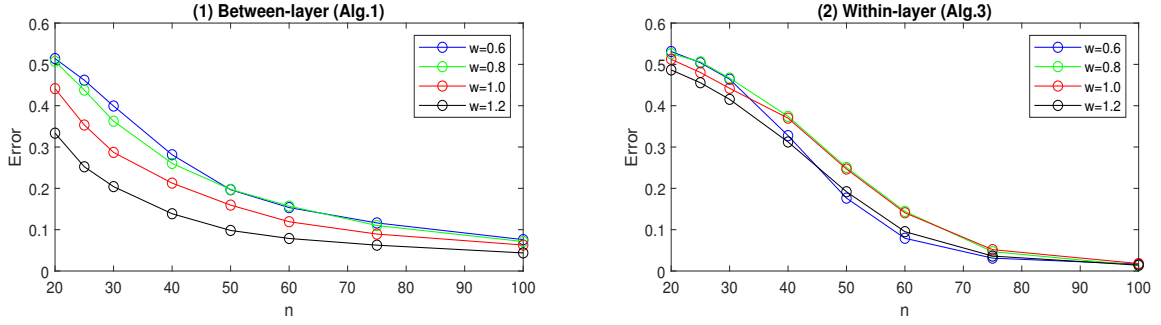


Figure 5: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 3 for  $L = 50$ ,  $c = 0$ ,  $d = 0.8$  and  $n = 20, 25, 30, 40, 50, 60, 75, 100$ .

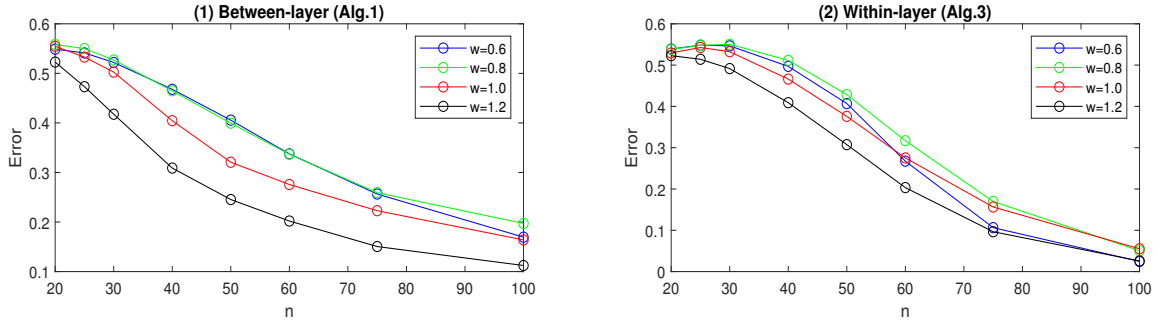


Figure 6: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 3 for  $L = 50$ ,  $c = 0$ ,  $d = 0.5$  and  $n = 20, 25, 30, 40, 50, 60, 75, 100$ .

disassortative. After the community assignment matrices  $\mathbf{Z}^{(m)}$  and the block probability matrices  $\mathbf{B}^{(l)}$  have been obtained, we construct the probability tensor  $\mathcal{P}$  with layers  $\mathcal{P}(:, :, l) = \mathbf{Z}^{(m)} \mathbf{B}^{(l)} (\mathbf{Z}^{(m)})^T$ , where  $m = c(l)$ ,  $l = 1, \dots, L$ . Subsequently, the layers  $\mathbf{A}^{(l)}$  of the adjacency tensor  $\mathcal{A}$  are obtained as symmetric matrices with zero diagonals and independent Bernoulli entries  $\mathbf{A}^{(l)}(i, j)$  for  $1 \leq i < j \leq n$ .

In this paper, we present two sets of simulations. In the first set, for relatively large values of  $n$  and  $L$ , we examine effectiveness of Algorithm 1 along with comparative performances of Algorithms 2 and 3. In the second simulations set, we explore the error rates of only Algorithms 1 and 3 when either  $n$  vary widely for a fixed  $L$ , or  $L$  vary widely for a fixed  $n$ . For both sets of simulations, we consider two sparsity scenarios,  $c = 0$ ,  $d = 0.8$  and  $c = 0$ ,  $d = 0.5$ , and four values of assortativity parameter  $w = 0.6, 0.8, 1.0$  and  $1.2$ . All simulations are carried out with  $M = 3$  and  $K = 3$ , and all results are averaged over 500 simulation runs.

Simulations results are summarized in Figures 1–8. In particular, Figure 1 ( $c = 0$ ,  $d = 0.8$ ) and Figure 2 ( $c = 0$ ,  $d = 0.5$ ) exhibit error rates of all three algorithms when

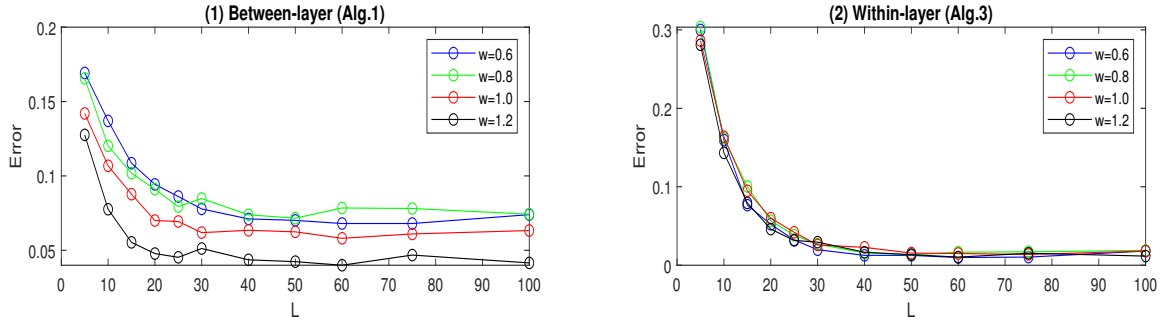


Figure 7: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 3 for  $n = 100$ ,  $c = 0$ ,  $d = 0.8$  and  $L = 5, 10, 15, 20, 25, 30, 40, 50, 60, 75, 100$ .

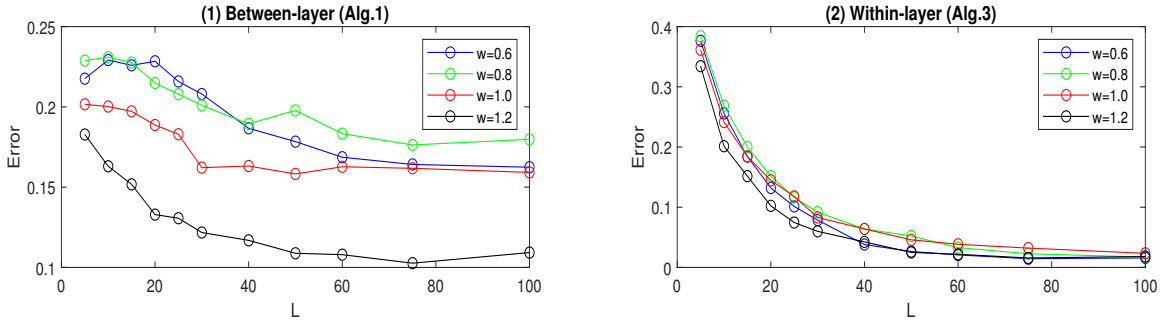


Figure 8: The between-layer clustering error rates of Algorithm 1 and the within-layer error rates of Algorithms 3 for  $n = 100$ ,  $c = 0$ ,  $d = 0.5$  and  $L = 5, 10, 15, 20, 25, 30, 40, 50, 60, 75, 100$ .

$L = 100$  and  $n = 50, 100, 150, 200$ . Figure 3 ( $c = 0$ ,  $d = 0.8$ ) and Figure 4 ( $c = 0$ ,  $d = 0.5$ ) display error rates of all three algorithms when  $n = 100$  and  $L = 50, 100, 150, 200$ . Figure 5 ( $c = 0$ ,  $d = 0.8$ ) and Figure 6 ( $c = 0$ ,  $d = 0.5$ ) show error rates of Algorithms 1 and 3 when  $L = 50$  and  $n = 20, 25, 30, 40, 50, 60, 75, 100$ . Figure 7 ( $c = 0$ ,  $d = 0.8$ ) and Figure 8 ( $c = 0$ ,  $d = 0.5$ ) demonstrate error rates of Algorithms 1 and 3 when  $n = 100$  and  $L = 5, 10, 15, 20, 25, 30, 40, 50, 60, 75, 100$ . All those figures present four graphs corresponding to  $w = 0.6, 0.8, 1, 1.2$ .

As the middle and the right panels in Figures 1–4 show, if the layers of the network are not all strongly assortative ( $w = 0.6$ ) or disassortative ( $w = 1.2$ ), Algorithm 2 has much higher errors than Algorithm 3. On the other hand, even in the best case scenario for Algorithm 2, when  $w = 0.6$ , it does not outperform Algorithm 3 by a high margin. For this reason, the second set of simulations employs only Algorithms 1 and 3. Note that the case of  $w = 1.2$  is beneficial for the between-layer clustering since larger  $w$  leads to a denser graph. The latter translates into smaller within-layer clustering error rates.

One can see from Figures 1–2 and 5–6 that, when  $n$  grows, all clustering errors decrease. The influence of  $L$  on the error rates is more complex. As Theorem 1 implies, the between-layer clustering errors are of the order  $(n\rho_n)^{-2}$  for fixed values of  $M$  and  $K$ . This agrees with the left panels in Figures 3 and 4, where curves exhibit constant behavior for all values of  $L$  (small fluctuations are just due to random errors). This, however, happens only when  $L$  is relatively large. Indeed, the left panel of Figure 7 shows steady decrease of the between-layer clustering errors for  $L \leq 40$  and after that the errors become almost constant and very small. The curves in the left panel of Figure 8 show similar behavior, although the decline is less prominent.

The within-layer clustering error has two components which, for fixed  $K$  and  $M$ , are of respective orders of  $(n\rho_n)^{-2}$  and  $(nL\rho_n)^{-1}$ . The second component of the error dominates the first one if  $L < C(K, M)n\rho_n$  where the factor  $C(K, M)$  depends only on  $K$  and  $M$ . While the second term is larger than the first, the within-layer clustering errors exhibit steady decline as  $L$  grows, as the right panels of Figures 7 and 8 demonstrate. On the other hand, as Figures 3 and 4 reveal, when  $L$  is large, the within-layer clustering rates do not reduce when  $L$  grows.

## 5. Extension of DIMPLE model to the Heterogeneous Multilayer Random Dot Product Graph (HMRDPG) setting

The between-layer clustering procedure in our paper can be extended to the Heterogeneous Multilayer Random Dot Product Graph (HMRDPG) settings. Specifically, as it has been mentioned in Section 1.2, we consider an extension of the COSIE model in (2) to the case, where groups of layers correspond to different subspaces. The probability tensor  $\mathcal{P}$  for such model has layers  $\mathbf{P}^{(l)}$ ,  $l = 1, \dots, L$ , where  $\mathbf{P}^{(l)}$  are described by equation (3). The adjacency tensor  $\mathcal{A}$  in this case, same as in the DIMPLE model, has layers  $\mathbf{A}^{(l)}$  such that  $\mathbf{A}^{(l)}(i, j) = \mathbf{A}^{(l)}(j, i)$  and, for  $1 \leq i < j \leq n$  and  $1 \leq l \leq L$ ,  $\mathbf{A}^{(l)}(i, j)$  are the Bernoulli random variables with  $\mathbb{P}(\mathbf{A}^{(l)}(i, j) = 1) = \mathbf{P}^{(l)}(i, j)$ , and they are independent from each other.

Note that, similarly to the case of the DIMPLE model, if the SVD of  $\mathbf{P}^{(l)}$  is  $\mathbf{P}^{(l)} = \mathbf{U}_{P,l}\mathbf{\Lambda}_{P,l}(\mathbf{U}_{P,l})^T$ , then, similarly to (7),  $\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T = \mathbf{V}^{(m)}(\mathbf{V}^{(m)})^T$  for  $m = c(l)$ ,  $l = 1, \dots, L$ ,  $m = 1, \dots, M$ . Therefore, one can base the between-layer clustering on the matrices  $\widehat{\mathbf{U}}_{A,l}$  in the SVD of the layers of the adjacency tensor  $\mathcal{A}$ . The Algorithm 1 can be applied with no modification, except that  $K$  is a common dimension of matrices  $\mathbf{V}^{(m)}$  rather than the common number of communities in the groups of layers.

Theoretical assessment of the precision of Algorithm 1 for the HMRDPG is very similar to the case of the DIMPLE network. The main difference between the two models is scaling. Specifically, in the case of the DIMPLE model with balanced community sizes in the groups of layers, one has  $\mathbf{B}_{DP}^{(l)} \asymp (n/K)^2 \mathbf{B}^{(l)}$  for  $\mathbf{B}^{(l)}$  in (1) and  $\mathbf{B}_{DP}^{(l)}$  in (3). Moreover, in the case of the DIMPLE network, one has  $\|\mathbf{P}^{(l)}\|_\infty = \|\mathbf{B}^{(l)}\|_\infty$ , and hence, the sparsity assumptions rely on scaling of matrices  $\mathbf{B}^{(l)}$ ,  $l = 1, \dots, L$ . In the case of the HMRDPG, in general,  $\|\mathbf{P}^{(l)}\|_\infty$  is not equal to  $\|\mathbf{B}_{DP}^{(l)}\|_\infty$ , so one needs to formulate slightly different assumptions.

**A1\*.** Clusters of layers and local communities are balanced, so that there exist absolute constants  $\underline{c}$  and  $\bar{c}$  such that  $\underline{c}L/M \leq L_m \leq \bar{c}L/M$ , where  $L_m$  is the number of networks in the layer of type  $m$ ,  $m = 1, \dots, M$ .

**A2\*.** Let  $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times MK}$  be a matrix, which is obtained as horizontal concatenation of matrices  $\mathbf{V}^{(m)} \in \mathbb{R}^{n \times K}$ ,  $m = 1, \dots, M$ . Let the SVD of  $\bar{\mathbf{Z}}$  be  $\bar{\mathbf{Z}} = [\mathbf{V}^{(1)} | \dots | \mathbf{V}^{(M)}] = \bar{\mathbf{U}} \bar{\mathbf{D}} \bar{\mathbf{V}}^T$ , where  $\bar{\mathbf{U}} \in \mathcal{O}_{n,r}$ ,  $\bar{\mathbf{V}} \in \mathcal{O}_{MK,r}$ ,  $r$  is the rank of  $\bar{\mathbf{Z}}$  and  $\bar{\mathbf{D}}$  is an  $r$ -dimensional diagonal matrix. Then, for some absolute constant  $\kappa_0$ , one has  $\sigma_1(\bar{\mathbf{D}}) \leq \kappa_0 \sigma_r(\bar{\mathbf{D}})$ .

**A3\*.** The layers  $\mathbf{P}^{(l)}$  of the probability tensor  $\mathcal{P}$  are such that  $\mathbf{P}^{(l)} = \rho_{n,l} \mathbf{P}_0^{(l)}$ ,  $\|\mathbf{P}_0^{(l)}\|_\infty = 1$ ,  $l = 1, \dots, L$ , and  $\min_{l=1, \dots, L} \rho_{n,l} \geq C_\rho n^{-1} \log n$ , for some absolute constant  $C_\rho$ .

**A4\*.** Matrices  $\mathbf{B}_{DP}^{(l)}$  in (3) are such that, for some absolute constant  $C_\lambda \in (0, 1)$ , one has  $\min_{l=1, \dots, L} \left[ \sigma_K(\mathbf{B}_{DP}^{(l)}) / \sigma_1(\mathbf{B}_{DP}^{(l)}) \right] \geq C_\lambda$ .

**A5\*.** There exist absolute constants  $\underline{c}_\rho$  and  $\bar{c}_\rho$  such that  $\underline{c}_\rho \rho_n \leq \rho_{n,l} \leq \bar{c}_\rho \rho_n$  with  $\rho_n = L^{-1} \sum \rho_{n,l}$ .

**A6\*.** For some absolute constant  $C_{0,P}$  one has

$$\|\mathbf{P}^{(l)}\|_F^2 \geq C_{0,P} K^{-1} \rho_{n,l}^2 n^2 \quad (36)$$

Assumptions **A1\***–**A5\*** are very similar to the Assumptions **A1**–**A5** for the DIMPLE model. Specifically, Assumption **A1\*** coincides with the first part of Assumption **A1**. Assumption **A2\*** mirrors Assumption **A2**, just for a somewhat different matrix  $\bar{\mathbf{Z}}$ . The only difference between the definitions of  $\rho_{n,l}$  in Assumptions **A3\*** and **A3** is that  $\rho_{n,l}$  is based on matrices  $\mathbf{B}^{(l)}$  in the first case and  $\mathbf{P}^{(l)}$  in the second. Since for the SBM,  $\|\mathbf{B}^{(l)}\|_\infty = \|\mathbf{P}^{(l)}\|_\infty$ , it would have been possible to use  $\mathbf{P}^{(l)}$  for the definition of  $\rho_{n,l}$  even for the SBM; the use of  $\mathbf{B}^{(l)}$  was motivated merely by convenience. Assumptions **A4\*** and **A5\*** are identical to the corresponding Assumptions **A4** and **A5**. Assumption **A6\*** is the only one which differs from assumptions in the case of the DIMPLE model. While it holds naturally in the case of the DIMPLE setting, it requires postulating in the case of the HMRDPG. Indeed, for a network equipped with balanced SBMs, such that the lowest and the highest eigenvalues of the block probability matrices are of the same order of magnitude, then, one has  $\|\mathbf{P}^{(l)}\|_F^2 \geq C \rho_{n,l}^2 n^2 / K$ , where the lower bound corresponds to a scenario where the not-diagonal blocks have zero probability of connection. Under the assumptions above, one obtains the following result.

**Theorem 4.** *Let the layers of the network be clustered by Algorithm 1 where  $K$  is the common dimension of the matrices  $\mathbf{V}^{(m)} \in \mathcal{O}_{n,K}$ . Let Assumptions **A1\***–**A6\*** hold. Then, for any  $\tau > 0$ , there exists a constant  $\check{C}$  that depends only on  $\tau$ ,  $\kappa_0$ ,  $C_\rho$ ,  $C_\lambda$ ,  $\bar{c}$ ,  $\underline{c}$ ,  $\bar{c}_\rho$ ,  $\underline{c}_\rho$  and  $C_{0,P}$  in Assumptions **A1\***–**A6\***, such that the  $n$  between-layer clustering error, defined in*

(20), *satisfies*

$$\mathbb{P} \left\{ R_{BL} \leq \frac{\ddot{C} K^4 M^2}{n^2 \rho_n^2} \right\} \geq 1 - L n^{-\tau} \quad (37)$$

## 6. Conclusions

To the best of our knowledge, our paper is the first one to consider the case, where both the community affiliations and the matrices of block probabilities of connections vary from one layer to another. We formulate the efficient clustering algorithms for the within-layer and the between-layer clustering and examine the clustering precision, both theoretically and in simulations. Our theory is developed under a set of natural, easy to interpret assumptions, majority of which are rather standard in stochastic networks models.

As our theory (Theorems 1 and 3) and the simulation results imply, when  $L$  grows, the clustering precision in both algorithms cease to decrease for a given number of nodes  $n$ :

$$R_{BL}^{DIMPLE} \lesssim C \rho_n^{-2} n^{-2}, \quad R_{WL}^{DIMPLE} \lesssim C (\rho_n^{-2} n^{-2} + n^{-1} L^{-1} \rho_n^{-1})$$

We believe that this is not caused by the deficiency of our methodology but is rather due to the fact, that the number of parameters in the model grows linearly in  $L$  for a fixed  $n$ . Indeed, the total number of independent parameters in the model is  $O(K^2 L + M n \log K + L \log M)$ , since we have  $L$  matrices  $\mathbf{B}^{(l)}$ ,  $M$  clustering matrices for the SBMs in the groups of layers, and a clustering matrix of the layers, while the total number of observations is  $O(n^2 L)$ . The latter implies that, while for small values of  $L$ , the term  $(M n \log K)/(n^2 L)$  may dominate the error, eventually, as  $L$  grows, the term  $L(K^2 + \log M)/(n^2 L)$  becomes larger for a fixed  $n$ .

Incidentally, we observe that a similar phenomenon holds in the MMLSBM, where the block probability matrices are the same in all layers of each of the groups. Specifically, if we look at the ALMA clustering error rates

$$\begin{aligned} R_{BL}^{ALMA} &\lesssim C (\rho_n^{-1} n^{-2} + \rho_n^{-2} n^{-2} [\min(n, L)]^{-1}), \\ R_{WL}^{ALMA} &\lesssim C (n^{-1} L^{-1} \rho_n^{-1} + \rho_n^{-1} n^{-2} + \rho_n^{-2} n^{-2} [\min(n, L)]^{-1}), \end{aligned}$$

we see that, for given  $n$  and  $\rho_n$ , as  $L$  grows, the clustering errors flatten. Here, we used ALMA rather than TWIST for comparison, since ALMA allows  $L$  to be larger than  $n$ . Also, ALMA is based on identification of clusters of similar layers, similarly to Algorithm 1.

We remark that, unlike the ALMA methodology in Fan et al. (2021) or the TWIST algorithm in Jing et al. (2020), all three algorithms in this paper are not iterative. It is known, that if one needs to recover a low rank tensor, then the power iterations can improve precision guarantees. This has been shown in the context of estimation of a low rank tensor in, e.g., Zhang and Xia (2018), and in the context of the clustering in the tensor block model in Han et al. (2021). While both ALMA and TWIST are designed for the MMLSBM, which results in a low rank probability tensor, the DIMPLE model does not lead to a low rank probability tensor. Therefore, it is not clear whether iterative techniques are advantageous in the DIMPLE setting. Our very limited experimentation with iterative algorithms did not lead to significant improvement of clustering precision. Investigation of this issue is a matter of future research.

## 7. Proofs

### 7.1 Proof of Theorem 1

Consider matrix

$$\Xi = \hat{\Theta} - \Theta, \quad \Xi(:, l) = \text{vec}(\hat{\mathbf{U}}_{A,l}(\hat{\mathbf{U}}_{A,l})^T - \mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T), \quad l = 1, \dots, L, \quad (38)$$

where  $\hat{\mathbf{U}}_{A,l}$  and  $\mathbf{U}_{P,l}$  are defined in (13) and (27), respectively. Assume, without loss of generality, that  $n^2 \geq L$  and let

$$\hat{\Theta} = \tilde{\mathbf{V}}\tilde{\Lambda}\tilde{\mathbf{W}}, \quad \tilde{\mathbf{V}} \in \mathcal{O}_{n^2, L}, \quad \tilde{\mathbf{W}} \in \mathcal{O}_L \quad (39)$$

be the SVD of  $\hat{\Theta}$ . Partition  $\tilde{\mathbf{V}}$  and  $\tilde{\mathbf{W}}$  as  $\tilde{\mathbf{V}} = [\hat{\mathbf{V}}|\hat{\mathbf{V}}_\perp]$  and  $\tilde{\mathbf{W}} = [\hat{\mathbf{W}}|\hat{\mathbf{W}}_\perp]$ , where  $\hat{\mathbf{V}} \in \mathcal{O}_{n^2, M}$  and  $\hat{\mathbf{W}} \in \mathcal{O}_{L, M}$  correspond to the first  $M$  columns in  $\tilde{\mathbf{V}}$  and  $\tilde{\mathbf{W}}$ . Observe that  $\hat{\Theta}$  can be written as  $\hat{\Theta} = \hat{\mathbf{V}}\hat{\Lambda}\hat{\mathbf{W}}^T + \hat{\mathbf{V}}_\perp\hat{\Lambda}_\perp\hat{\mathbf{W}}_\perp^T$ , where  $\hat{\Lambda}$  and  $\hat{\Lambda}_\perp$  are corresponding parts of  $\tilde{\Lambda}$ .

By Proposition 1 of Cai and Zhang (2018), if  $\sigma_M(\hat{\Theta}\mathbf{W}) > \sigma_{M+1}(\hat{\Theta})$ , then one has the following inequality

$$\|\sin \Theta(\hat{\mathbf{W}}, \mathbf{W})\| \leq \frac{\sigma_{M+1}(\hat{\Theta}) \|\Pi_{\hat{\Theta}\mathbf{W}}(\hat{\Theta}\mathbf{W}_\perp)\|}{\sigma_M^2(\hat{\Theta}\mathbf{W}) - \sigma_{M+1}^2(\hat{\Theta})} \quad (40)$$

We start with analysis of quantities in (40). By Lemma 2 of Cai and Zhang (2018), obtain that

$$\sigma_M^2(\hat{\Theta}\mathbf{W}) = \sigma_M^2(\mathbf{V}^T\hat{\Theta}\mathbf{W} + \mathbf{V}_\perp^T\hat{\Theta}\mathbf{W}) \geq \sigma_M^2(\mathbf{V}^T\Theta\mathbf{W} + \mathbf{V}_\perp^T\Xi\mathbf{W}),$$

so that  $\sigma_M^2(\hat{\Theta}\mathbf{W}) \geq [\sigma_M(\mathbf{V}^T\Theta\mathbf{W}) - \|\Xi\|]^2$ . Take into account that, due to (32), one has  $\sigma_M(\mathbf{V}^T\Theta\mathbf{W}) = \sigma_M(\Lambda) = \sigma_{\min}(\mathbf{F})$ . Hence, combination of Lemma 1 and (30) yield that  $\sigma_M^2(\hat{\Theta}\mathbf{W}) \geq [\sigma_{\min}(\mathbf{F}) - \|\Xi\|]^2$ . Since, by (30) and (31)

$$\sigma_M(\Theta) = \sigma_{\min}(\mathbf{F}) \geq \sqrt{\underline{c}/\bar{c}}\kappa_0^{-2} \|\mathbf{F}\| = \sqrt{\underline{c}/\bar{c}}\kappa_0^{-2} \|\Theta\|, \quad (41)$$

infer that

$$\sigma_M(\hat{\Theta}\mathbf{W}) \geq \sqrt{\underline{c}/\bar{c}}\kappa_0^{-2} \|\Theta\| - \|\Xi\| \quad (42)$$

By part 1 of Lemma 2 of Cai and Zhang (2018), since  $\sigma_{M+1}(\Theta) = 0$ , one has

$$\sigma_{M+1}(\hat{\Theta}) \leq \min \left( \|\mathbf{V}_\perp\mathbf{V}_\perp^T\hat{\Theta}\|, \|\hat{\Theta}\mathbf{W}_\perp\mathbf{W}_\perp^T\| \right) \leq \|\Xi\| \quad (43)$$

Now, examine  $\|\Pi_{\hat{\Theta}\mathbf{W}}(\hat{\Theta}\mathbf{W}_\perp)\| = \|\hat{\Theta}\mathbf{W}(\mathbf{W}^T\hat{\Theta}^T\hat{\Theta}\mathbf{W})^{-1}(\hat{\Theta}\mathbf{W})^T\hat{\Theta}\mathbf{W}_\perp\|$ . Since  $\hat{\Theta}\mathbf{W}_\perp = \Xi\mathbf{W}_\perp$ , derive  $\|\Pi_{\hat{\Theta}\mathbf{W}}(\hat{\Theta}\mathbf{W}_\perp)\| \leq [\sigma_M(\hat{\Theta}\mathbf{W})]^{-1}\|\hat{\Theta}\mathbf{W}\|\|\Xi\mathbf{W}_\perp\|$ . Here,  $\|\hat{\Theta}\mathbf{W}\| \leq \|\Theta\| + \|\Xi\|$ ,  $\|\Xi\mathbf{W}_\perp\| \leq \|\Xi\|$  and the lower bound on  $\sigma_M(\hat{\Theta}\mathbf{W})$  is given by (42). Combining all terms together, we arrive at

$$\|\Pi_{\hat{\Theta}\mathbf{W}}(\hat{\Theta}\mathbf{W}_\perp)\| \leq \frac{\|\Xi\|(\|\Theta\| + \|\Xi\|)}{\sqrt{\underline{c}/\bar{c}}\kappa_0^{-2} \|\Theta\| - \|\Xi\|} \quad (44)$$

Denote  $C_{\kappa_0} = \sqrt{\bar{c}/\underline{c}} \kappa_0^2$  and  $\delta = \|\Xi\|/\|\Theta\|$ . Then, plugging (42)–(44) into (40), obtain

$$\|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\| \leq \frac{C_{\kappa_0}^3 \delta^2 (1 + \delta)}{(1 - C_{\kappa_0} \delta) [(1 - C_{\kappa_0} \delta)^2 - C_{\kappa_0}^2 \delta^2]} \quad (45)$$

Next, we need to establish an upper bound on  $\delta$ . Note that

$$\|\Xi(:, l)\|^2 = \left\| \widehat{\mathbf{U}}_{A,l} (\widehat{\mathbf{U}}_{A,l})^T - \mathbf{U}_{P,l} (\mathbf{U}_{P,l})^T \right\|_F^2 = 2 \left\| \sin \Theta(\mathbf{U}_{P,l}, \widehat{\mathbf{U}}_{A,l}) \right\|_F^2$$

Since  $\text{rank}(\widehat{\mathbf{U}}_{A,l} (\widehat{\mathbf{U}}_{A,l})^T - \mathbf{U}_{P,l} (\mathbf{U}_{P,l})^T) \leq 2K$ , by Davis-Kahan theorem (see, e.g., Theorem 2 of Yu et al. (2014)) obtain that

$$\|\Xi(:, l)\|^2 \leq 8K \left[ \sigma_{\min}(\mathbf{P}^{(l)}) \right]^{-2} \left\| \mathbf{A}^{(l)} - \mathbf{P}^{(l)} \right\|^2 \quad (46)$$

Using Theorem 5.2 of Lei and Rinaldo (2015) and applying the union bound, obtain that, for any  $\tau > 0$  there exists a constant  $C_{\tau, \rho}$  that depends on  $\tau$ ,  $C_\rho$ ,  $\underline{c}_\rho$  and  $\bar{c}_\rho$  only, and a set  $\Omega_\tau$ , such that

$$\max_{l=1, \dots, L} \left\| \mathbf{A}^{(l)} - \mathbf{P}^{(l)} \right\|^2 \leq C_{\tau, \rho} n \rho_n \quad \text{for } \omega \in \Omega_\tau, \quad \mathbb{P}(\Omega_\tau) \geq 1 - Ln^{-\tau} \quad (47)$$

In order to construct a lower bound for  $\sigma_{\min}(\mathbf{P}^{(l)})$ , note that under Assumptions **A1**–**A5**, one has  $\sigma_{\min}(\mathbf{B}_0^{(l)}) \geq C_\lambda \|\mathbf{B}_0^{(l)}\| \geq C_\lambda$  since, by definition,  $\|\mathbf{B}_0^{(l)}\| \geq 1$ . Then,

$$\sigma_{\min}(\mathbf{P}^{(l)}) = \rho_{n,l} \sigma_{\min} \left( \sqrt{\mathbf{D}_z^{(c(l))}} \mathbf{B}_0^{(l)} \sqrt{\mathbf{D}_z^{(c(l))}} \right) \geq \underline{c} \underline{c}_\rho C_\lambda \rho_n n K^{-1} \quad (48)$$

Plugging the latter into (46) together with (47), arrive at

$$\|\Xi\|^2 \leq \|\Xi\|_F^2 \leq \frac{\tilde{C} K^3 L}{n \rho_n}, \quad \tilde{C} = \frac{8 C_{\tau, \rho}}{\underline{c}^2 \underline{c}_\rho^2 C_\lambda^2}, \quad \omega \in \Omega_\tau \quad (49)$$

To find an upper bound for  $\delta = \|\Xi\|/\|\Theta\|$ , note that, by (72),  $\|\Theta\|^2 \geq \|\Theta\|_F^2/M = KL/M$ , so that, for  $\omega \in \Omega_\tau$ ,  $\delta^2 \leq \tilde{C} K^2 M/(n \rho_n)$ . Since  $\text{rank}(\widehat{\mathbf{W}}) = \text{rank}(\mathbf{W}) = M$ , if  $n$  is large enough to guarantee that  $C_{\kappa_0} \delta \leq 1/4$ , the latter, together with (45), imply that

$$\|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_F \leq \frac{10}{3} C_{\kappa_0}^3 \sqrt{M} \delta^2 \leq \frac{10 C_{\kappa_0}^3 \tilde{C} K^2 M \sqrt{M}}{3 n \rho_n}, \quad \omega \in \Omega_\tau \quad (50)$$

Now, to relate the distance  $\|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_F$  and the number of clustering errors, we use a version of Lemma 8 of Lei and Lin (2021):

**Lemma 2. ( Lemma 8 of Lei and Lin (2021)).** *Let  $\mathbf{W}$  be an  $N \times d$  matrix with  $K$  distinct rows and minimum pairwise Euclidean norm separation  $\gamma$ . Let  $\widehat{\mathbf{W}}$  be another  $N \times d$  matrix and  $\widehat{\mathbf{C}} \in \mathcal{M}_{N, K}$  and  $\widehat{\mathbf{X}} \in \mathbb{R}^{K \times d}$  be an  $(1 + \epsilon)$ -appropriate solution to  $K$ -means problem with  $K$  clusters:*

$$\|\widehat{\mathbf{W}} - \widehat{\mathbf{C}} \widehat{\mathbf{X}}\|_F^2 \leq (1 + \epsilon) \min_{\mathbf{C}, \mathbf{X}} \|\widehat{\mathbf{W}} - \mathbf{C} \mathbf{X}\|_F^2$$

*Then, the number of errors in  $\widehat{\mathbf{C}}$  as an estimate of the row clusters of  $\mathbf{W}$  is no larger than*

$$\Delta = C_\epsilon \gamma^{-2} \|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_F^2.$$

In order to apply Lemma 2, we need to find

$$\gamma^2 = \min_{i \neq j} \|\mathbf{W}(i, :) - \mathbf{W}(j, :)\|^2, \quad i, j = 1, \dots, L.$$

Recall that, by (32),  $\mathbf{W} = \mathbf{W}\mathbf{V}_F$  where  $\mathbf{V}_F \in \mathcal{O}_M$  and  $\mathbf{W} = \mathbf{C}\mathbf{D}_c^{-1/2}$  where  $\mathbf{C} \in \mathcal{M}_{L,M}$  is a clustering matrix, so that  $\gamma^2 \geq 2(\bar{c}L/M)^{-1}$ . Hence,  $R_{BL} \leq 2\bar{c}M^{-1}\|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_F^2$ , where  $R_{BL}$  is defined in (20) and  $\|\sin \Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_F$  is given by (50). Combination of all the constants into  $\bar{C}$  completes the proof.

## 7.2 Proof of Theorem 2

The proof requires the following lemma.

**Lemma 3.** *Let  $\mathbf{W}$  and  $\widehat{\mathbf{W}}$  be defined as in (11) and (16), respectively. Then,*

$$\min_{\mathcal{P} \in \mathfrak{F}(M)} \|\widehat{\mathbf{W}} - \mathbf{W} \mathcal{P}\|_F^2 \leq \frac{6M}{\underline{c}} R_{BL}, \quad (51)$$

where  $R_{BL}$  is defined in (20).

Let  $\mathcal{H} = \mathcal{P} \times_3 \mathbf{W}^T$  be the reduced probability tensor with layers  $\mathbf{H}^{(m)} = \mathcal{H}(:, :, m)$ , and let  $\mathbf{H}^{(m)} = \mathbf{U}_{\mathbf{H}}^{(m)} \mathbf{\Lambda}_{\mathbf{H}}^{(m)} (\mathbf{U}_{\mathbf{H}}^{(m)})^T$  be the SVD of  $\mathbf{H}^{(m)}$ . If  $\widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}$  is defined in steps 2 and 3 of Algorithm 2, then, according to Lemma 2, the number of clustering errors for layer  $m$  is bounded by

$$\Delta_m = \tilde{C}_\epsilon \gamma_m^{-2} \left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2, \quad (52)$$

where  $\gamma_m$  is the Euclidean separation of rows of  $\mathbf{U}_{\mathbf{H}}^{(m)}$ . Recalling calculations in Section 2.1, obtain that

$$\mathbf{H}^{(m)} = \sqrt{L_m} \mathbf{U}_z^{(m)} \bar{\mathbf{B}}^{(m)} (\mathbf{U}_z^{(m)})^T, \quad \bar{\mathbf{B}}^{(m)} = L_m^{-1} \sum_{c(l)=m} \mathbf{B}_D^{(l)},$$

where  $\mathbf{U}_z^{(m)}$  and  $\mathbf{B}_D^{(l)}$  are defined in (6) and (8), respectively. If  $\bar{\mathbf{B}}^{(m)} = \bar{\mathbf{O}}^{(m)} \bar{\mathbf{\Lambda}}^{(m)} (\bar{\mathbf{O}}^{(m)})^T$  with  $\bar{\mathbf{O}}^{(m)} \in \mathcal{O}_K$  is the SVD of  $\bar{\mathbf{B}}^{(m)}$ , then  $\mathbf{U}_{\mathbf{H}}^{(m)} = \mathbf{U}_z^{(m)} \bar{\mathbf{O}}^{(m)}$  and  $\mathbf{\Lambda}_{\mathbf{H}}^{(m)} = L_m^{1/2} \bar{\mathbf{\Lambda}}$ . Therefore,  $\gamma_m$  in (52) is the Euclidean separation of rows of  $\mathbf{U}_z^{(m)}$ , and

$$\gamma_m^2 \geq 2 \min(n_{k,m}^{-1}) \geq 2K/(\bar{c}n) \quad (53)$$

In order to construct an upper bound for  $\left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2$ , recall that by Davis-Kahan theorem (Yu et al. (2014), Theorem 2)

$$\left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2 \leq \frac{4K \|\widehat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2}{\sigma_K^2(\mathbf{H}^{(m)})} \quad (54)$$



To obtain a lower bound for  $\sigma_K^2(\mathbf{H}^{(m)})$  observe that  $\sigma_K^2(\mathbf{H}^{(m)}) = L_m \sigma_K^2(\bar{\mathbf{B}}^{(m)})$  and

$$\begin{aligned} \sigma_K(\bar{\mathbf{B}}^{(m)}) &\geq \frac{\min_k n_{k,m}}{L_m} \sigma_K \left( \sum_{c(l)=m} \mathbf{B}_D^{(l)} \right) \\ &\geq \frac{\underline{c} n}{L_m K} \sum_{c(l)=m} \rho_{n,l} \sigma_K(\mathbf{B}_0^{(l)}) \geq \frac{\underline{c} C_\lambda \underline{c}_\rho \rho_n n}{K}, \end{aligned} \quad (55)$$

where (55) follows from Assumptions **A1–A6** and the Theorem in Complement 10.1.2 on page 327 of Rao and Rao (1998). Therefore, for some absolute constant  $\tilde{C}$ , one has  $\sigma_K^2(\mathbf{H}^{(m)}) \geq \tilde{C} \rho_n^2 n^2 L / (MK^2)$ . Consequently, (52)–(55) yield that the total number of clustering errors in all  $M$  layers is bounded by

$$\Delta = \sum_{m=1}^M \Delta_m \leq C_H \frac{M K^2}{n \rho_n^2 L} \Delta_H \quad \text{with} \quad \Delta_H = \sum_{m=1}^M \|\hat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2, \quad (56)$$

where  $C_H$  is an absolute constant.

Denote  $\mathcal{X} = \mathcal{A} - \mathcal{P}$  and  $\mathbf{X}^{(l)} = \mathcal{X}(:, :, l)$ ,  $l = 1, \dots, L$ . It is easy to see that

$$\begin{aligned} \|\hat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\| &= \left\| [\mathcal{A} \times_3 \hat{\mathbf{W}}^T - \mathcal{P} \times_3 \mathbf{W}^T](:, :, m) \right\| \\ &\leq \left\| [\mathcal{X} \times_3 \mathbf{W}^T](:, :, m) \right\| + \left\| [\mathcal{P} \times_3 (\hat{\mathbf{W}} - \mathbf{W})^T](:, :, m) \right\| + \left\| [\mathcal{X} \times_3 (\hat{\mathbf{W}} - \mathbf{W})^T](:, :, m) \right\|, \end{aligned} \quad (57)$$

so that  $\Delta_H \leq 3(\Delta_{H,1} + \Delta_{H,2} + \Delta_{H,3})$ , where the three terms correspond to the three terms in the expansion above.

Note that, by Theorem 2 of Lei and Lin (2021) with  $v_1 = 2\bar{c}_\rho \rho_n$ ,  $R_1 = 1$ ,  $H_l = \mathbf{I}_n$  and  $t = 8\sqrt{\tau \bar{c}_\rho \bar{c} L \rho_n n \log n / M}$ , one has

$$\mathbb{P} \left\{ \left\| \sum_{c(l)=m} \mathbf{X}^{(l)} \right\|^2 \leq 64 \tau \bar{c}_\rho L_m \rho_n n \log n \right\} \geq 1 - 8n^{1-2\tau}$$

Since  $\mathbf{W}(l, m) = L_m^{-1/2}$  if  $c(l) = m$  and  $\mathbf{W}(l, m) = 0$  otherwise, taking the union bound obtain, for  $\omega \in \tilde{\Omega}_\tau$ :

$$\Delta_{H,1} = \sum_{m=1}^M \left\| \sum_{c(l)=m} \frac{\mathbf{X}^{(l)}}{\sqrt{L_m}} \right\|^2 \leq 64 \tau \bar{c}_\rho M \rho_n n \log n, \quad \text{where } \mathbb{P}(\tilde{\Omega}_\tau) \geq 1 - 8n^{1-2\tau} M \quad (58)$$

In order to derive an upper bound for  $\Delta_{H,2}$ , denote  $\tilde{\mathbf{P}} = (\mathcal{M}_3(\mathcal{P}))^T \in \mathbb{R}^{n^2 \times L}$ , the transpose of the mode 3 matricization of tensor  $\mathcal{P}$ . Then,

$$\begin{aligned} \Delta_{H,2} &\leq \sum_{m=1}^M \left\| [\mathcal{P} \times_3 (\hat{\mathbf{W}} - \mathbf{W})^T](:, :, m) \right\|_F^2 = \sum_{m=1}^M \left\| \tilde{\mathbf{P}} (\hat{\mathbf{W}} - \mathbf{W})^T \right\|^2 \\ &= \|\tilde{\mathbf{P}} (\hat{\mathbf{W}} - \mathbf{W})\|_F^2 \leq \|\tilde{\mathbf{P}}\|_F^2 \|\hat{\mathbf{W}} - \mathbf{W}\|_F^2 \leq 6 \underline{c}^{-1} M n^2 \rho_n^2 L R_{BL}, \end{aligned} \quad (59)$$

by Lemma 3. Similarly,

$$\begin{aligned}\Delta_{H,3} &\leq \sum_{m=1}^M \left[ \sum_{l=1}^L \|\mathbf{X}^{(l)}\| |\widehat{\mathbf{W}}(l, m) - \mathbf{W}(l, m)| \right]^2 \leq L \left[ \max_l \|\mathbf{X}^{(l)}\|^2 \right] \|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2 \\ &\leq C_{\tau, \rho} L n \rho_n \|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2 \leq 6 \underline{c}^{-1} C_{\tau, \rho} M L n \rho_n R_{BL} \quad \text{for } \omega \in \Omega_\tau\end{aligned}\quad (60)$$

It follows from (58)–(60) and Theorem 1 that, for  $\omega \in \Omega_\tau \cap \tilde{\Omega}_\tau$

$$\Delta_H \leq C \left( M L n^2 \rho_n^2 \frac{K^4 M^2}{n^2 \rho_n^2} + M n \rho_n \log n \right) \quad (61)$$

Since  $R_{WL} \leq \Delta/(M n)$ , combination of (56) and (61) complete the proof.

### 7.3 Proof of Theorem 3

Consider tensors  $\mathcal{G} \in \mathbb{R}^{n \times n \times L}$  and  $\mathcal{H} = \mathcal{G} \times_3 \mathbf{W}^T \in \mathbb{R}^{n \times n \times M}$  with layers, respectively,  $\mathbf{G}^{(l)} = \mathcal{G}(:, :, l)$  and  $\mathbf{H}^{(m)} = \mathcal{H}(:, :, m)$  of the forms

$$\mathbf{G}^{(l)} = (\mathbf{P}^{(l)})^2, \quad \mathbf{H}^{(m)} = L_m^{-1/2} \sum_{c(l)=m} \mathbf{G}^{(l)}, \quad l = 1, \dots, L, \quad m = 1, \dots, M \quad (62)$$

In order to assess  $R_{WL}$ , one needs to examine the spectral structure of matrices  $\mathbf{H}^{(m)}$  and their deviation from the sample-based versions  $\widehat{\mathbf{H}}^{(m)} = \widehat{\mathcal{H}}(:, :, m)$ . We start with the first task.

It follows from (6) and (8) that

$$\mathbf{H}^{(m)} = \mathbf{U}_z^{(m)} \overline{\mathbf{Q}}_D^{(m)} (\mathbf{U}_z^{(m)})^T \quad \text{with} \quad \overline{\mathbf{Q}}_D^{(m)} = L_m^{-1/2} \sum_{c(l)=m} \left( \mathbf{B}_D^{(l)} \right)^2 \quad (63)$$

Here, by (8), one has  $(\mathbf{B}_D^{(l)})^2 = \mathbf{O}_z^{(l)} (\mathbf{S}_z^{(l)})^2 (\mathbf{O}_z^{(l)})^T$ , so that all eigenvalues of  $(\mathbf{B}_D^{(l)})^2$  are positive. Again applying the Theorem in Complement 10.1.2 on page 327 of Rao and Rao (1998) and Assumptions **A1**–**A5**, obtain (due to  $\|\mathbf{B}_0^{(l)}\| \geq 1$ )

$$\begin{aligned}\sigma_{\min}(\mathbf{H}^{(m)}) &= \sigma_K \left( \overline{\mathbf{Q}}_D^{(m)} \right) \geq L_m^{-1/2} \sum_{c(l)=m} \sigma_K \left( (\mathbf{B}_D^{(l)})^2 \right) \\ &\geq L_m^{-1/2} \min_k (n_{k,m}^2) \sum_{c(l)=m} \rho_{n,l}^2 \sigma_K \left( (\mathbf{B}_0^{(l)})^2 \right) \geq L_m^{-1/2} \underline{c}^2 n^2 K^{-2} \underline{c}_\rho^2 \rho_n^2 L_m C_\lambda^2 \\ &\geq \underline{c}^{5/2} \underline{c}_\rho^2 C_\lambda^2 n^2 \rho_n^2 K^{-2} \sqrt{LM}^{-1}\end{aligned}\quad (64)$$

In order to assess the row separation of matrices  $\mathbf{U}_H^{(m)}$  in the SVD's of  $\mathbf{H}^{(m)}$ ,  $m = 1, \dots, M$ , write the SVD of  $\overline{\mathbf{Q}}_D^{(m)}$  as

$$\overline{\mathbf{Q}}_D^{(m)} = \mathbf{O}_{Q,D}^{(m)} \mathbf{S}_{Q,D}^{(m)} (\mathbf{O}_{Q,D}^{(m)})^T, \quad \mathbf{O}_{Q,D}^{(m)} \in \mathcal{O}_K.$$

Therefore, up to the column order, the matrix  $\mathbf{U}_{\mathbf{H}}^{(m)}$  of singular vectors of  $\mathbf{H}^{(m)}$  is of the form  $\mathbf{U}_{\mathbf{H}}^{(m)} = \mathbf{U}_z^{(m)} \mathbf{O}_{Q,D}^{(m)}$ , and the Euclidean separation  $\gamma_m$  of rows of  $\mathbf{U}_{\mathbf{H}}^{(m)}$  is the same as the Euclidean separation of rows of  $\mathbf{U}_z^{(m)}$  given by (53). Consequently, it follows from Lemma 2 that the total number of errors in the layer of type  $m$  is bounded above by  $C_\epsilon \gamma_m^{-2} \left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2$ . Using the lower bound on  $\gamma_m^2$  in (53), obtain that the total number of clustering errors  $\Delta$  within all layers is bounded as

$$\Delta \leq \frac{C_\epsilon \bar{c}}{2} \frac{n}{K} \sum_{m=1}^M \left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2$$

Using Davis-Kahan theorem and formula (64), obtain

$$\left\| \sin \Theta \left( \widehat{\mathbf{U}}_{\widehat{\mathbf{H}}}^{(m)}, \mathbf{U}_{\mathbf{H}}^{(m)} \right) \right\|_F^2 \leq \frac{4K \|\widehat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2}{\sigma_{\min}^2(\mathbf{H}^{(m)})} \leq C \frac{K^5 M \|\widehat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2}{n^4 \rho_n^4 L}$$

where we use  $C$  for different constants that depend on the constants in Assumptions **A1**–**A5**. Combination of the last two inequalities yields that the total number of clustering errors within all layers is bounded as

$$\Delta \leq C \frac{K^4 M}{n^3 \rho_n^4 L} \sum_{m=1}^M \|\widehat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2 \quad (65)$$

Recall that  $\mathbf{H}^{(m)} = [\mathcal{G} \times_3 \mathbf{W}^T](:, :, m)$  and  $\widehat{\mathbf{H}}^{(m)} = [\widehat{\mathcal{G}} \times_3 \widehat{\mathbf{W}}^T](:, :, m)$ . Denote

$$\overline{\mathbf{G}}^{(m)} = \sum_{c(l)=m} \mathbf{G}^{(l)} = \sqrt{L_m} \mathbf{H}^{(m)}, \quad \overline{\widehat{\mathbf{G}}}^{(m)} = \sqrt{L_m} [\widehat{\mathcal{G}} \times_3 \mathbf{W}^T](:, :, m) = \sum_{c(l)=m} \widehat{\mathbf{G}}^{(l)}$$

Then, repeating the steps in (57) with  $\mathcal{A}$  and  $\mathcal{P}$  replaced by, respectively,  $\widehat{\mathcal{G}}$  and  $\mathcal{G}$ , obtain that

$$\begin{aligned} \|\widehat{\mathbf{H}}^{(m)} - \mathbf{H}^{(m)}\|^2 &\leq 3 \left[ \Delta_{H,1}^{(m)} + \Delta_{H,2}^{(m)} + \Delta_{H,3}^{(m)} \right], \quad \Delta_{H,1}^{(m)} = L_m^{-1} \left\| \overline{\widehat{\mathbf{G}}}^{(m)} - \overline{\mathbf{G}}^{(m)} \right\|^2 \\ \Delta_{H,2}^{(m)} &= \left\| [\mathcal{G} \times_3 (\widehat{\mathbf{W}} - \mathbf{W})^T](:, :, m) \right\|^2, \quad \Delta_{H,3}^{(m)} = \left\| [(\widehat{\mathcal{G}} - \mathcal{G}) \times_3 (\widehat{\mathbf{W}}(:, m) - \mathbf{W})^T](:, :, m) \right\|^2 \end{aligned} \quad (66)$$

Therefore,

$$\Delta \leq 3(\Delta_1 + \Delta_2 + \Delta_3) \quad \text{with} \quad \Delta_i = C \frac{K^4 M}{n^3 \rho_n^4 L} \sum_{m=1}^M \Delta_{H,i}^{(m)}, \quad i = 1, 2, 3. \quad (67)$$

To upperbound  $\Delta_{H,i}^{(m)}$ ,  $i = 1, 2, 3$ , we use the following lemma that modifies upper bounds in Lei and Lin (2021) in the absence of the sparsity assumption  $\rho_n n \leq C$ :

**Lemma 4.** *Let Assumptions **A1**–**A5** hold,  $\mathbf{G}^{(l)} = (\mathbf{P}^{(l)})^2$  and  $\widehat{\mathbf{G}}^{(l)} = (\mathbf{A}^{(l)})^2 - \text{diag}(\mathbf{A}^{(l)}\mathbf{1})$ , where  $c(l) = m$ ,  $l = 1, \dots, \tilde{L}$ . Let*

$$\mathbf{G} = \sum_{l=1}^{\tilde{L}} \mathbf{G}^{(l)}, \quad \widehat{\mathbf{G}} = \sum_{l=1}^{\tilde{L}} \widehat{\mathbf{G}}^{(l)}$$

*Then, for any  $\tau > 0$ , there exists a constant  $\tilde{C}$  that depends only on  $\tau, \kappa_0, C_\rho, C_\lambda, \bar{c}, \underline{c}, \bar{c}_\rho$  and  $\underline{c}_\rho$  in Assumptions **A1**–**A5**, and  $\tilde{C}_{\tau,\epsilon}$  which depends only on  $\tau$  and  $\epsilon$ , such that one has*

$$\mathbb{P} \left\{ \|\widehat{\mathbf{G}} - \mathbf{G}\|^2 \leq \tilde{C} \left[ \rho_n^3 n^3 \tilde{L} \log(\tilde{L} + n) + \rho_n^4 n^2 \tilde{L}^2 \right] \right\} \geq 1 - \tilde{C}_{\tau,\epsilon} (\tilde{L} + n)^{1-\tau} \quad (68)$$

Applying Lemma 4 with  $\tilde{L} = L_m$ , obtain that, with probability at least  $1 - \tilde{C}_{\tau,\epsilon} n^{1-\tau}$ , one has  $\Delta_{H,1}^{(m)} \leq \tilde{C} [\rho_n^3 n^3 \log(L + n) + \rho_n^4 n^2 L_m]$ , so that

$$\mathbb{P} \left\{ \Delta_1 \leq \tilde{C}_1 [\rho_n^{-1} L^{-1} K^4 M^2 \log(L + n) + n^{-1} K^4 M] \right\} \geq 1 - \tilde{C}_{\tau,\epsilon} M n^{1-\tau} \quad (69)$$

In order to obtain an upper bound for  $\Delta_2$ , note that for  $\tilde{\mathbf{G}} = (\mathcal{M}_3(\mathcal{G}))^T \in \mathbb{R}^{n^2 \times L}$ , the transpose of mode 3 matricization of  $\mathbf{G}$ , one has

$$\|\tilde{\mathbf{G}}\|_F^2 = \sum_{l=1}^L \left\| (\mathbf{P}^{(l)})^2 \right\|_F^2 \leq \sum_{l=1}^L \left\| \mathbf{P}^{(l)} \right\|_F^4 \leq n^4 \rho_n^4 L,$$

so that  $\sum_m \Delta_{H,2}^{(m)} \leq n^4 \rho_n^4 L \|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2$ , similarly to (59). Applying Lemma 3, obtain

$$\mathbb{P} \left\{ \Delta_2 \leq \tilde{C}_2 n^{-1} \rho_n^{-2} K^8 M^4 \right\} \geq 1 - L n^{-\tau} \quad (70)$$

For  $\Delta_3$ , note that, similarly to the case of  $\Delta_2$ , one has

$$\sum_m \Delta_{H,3}^{(m)} \leq L \|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2 \max_l \|\widehat{\mathbf{G}}^{(l)} - \mathbf{G}^{(l)}\|^2,$$

where, by Lemma 4 with  $\tilde{L} = 1$ , one has  $\|\widehat{\mathbf{G}}^{(l)} - \mathbf{G}^{(l)}\|^2 \leq \tilde{C} [\rho_n^3 n^3 \log n + \rho_n^4 n^2]$ . Hence, application of Lemma 3 and the union bound over  $l = 1, \dots, L$ , yield

$$\mathbb{P} \left\{ \Delta_3 \leq \tilde{C}_3 n^{-3} \rho_n^{-3} K^8 M^4 \log n \right\} \geq 1 - C_{\tau,\epsilon} [L n^{-\tau} + n^{2-\tau}] \quad (71)$$

To complete the proof, combine formulas (67) and (69)–(71), and recall that  $R_{WL} = \Delta/(Mn)$  and  $n\rho_n \geq C_\rho \log n$ .

## 7.4 Proof of Theorem 4

The proof of Theorem 4 is almost identical to the proof of Theorem 1. The only difference is the construction of the lower bound for  $\sigma_{\min}(\mathbf{P}^{(l)})$ . Specifically, under Assumptions **A1\***–**A6\***, one has

$$\begin{aligned} \sigma_{\min}(\mathbf{P}^{(l)}) &= \sigma_K(\mathbf{B}_{DP}^{(l)}) \geq C_\lambda \sigma_1(\mathbf{B}_{DP}^{(l)}) \geq C_\lambda K^{-1/2} \|\mathbf{B}_{DP}^{(l)}\|_F \\ &= C_\lambda K^{-1/2} \|\mathbf{P}^{(l)}\|_F \geq C_\lambda \sqrt{C_{0,P}} K^{-1} \rho_{n,l} n, \end{aligned}$$

which differs from (48) only by a constant. The rest of the proof follows the proof of Theorem 1.

## 7.5 Proof of supplementary lemmas

**Proof of Lemma 1** Note that

$$\sigma_1(\mathbf{F}) \leq \sigma_1^2(\overline{\mathbf{D}}) \sqrt{\max_{m=1,\dots,M} L_m}, \quad \sigma_M(\mathbf{F}) \geq \sigma_M^2(\overline{\mathbf{D}}) \sqrt{\min_{m=1,\dots,M} L_m},$$

and, by Assumptions **A1** and **A4**,  $\sigma_1^2(\mathbf{F}) \leq \kappa_0^4 \sigma_M^2(\mathbf{F}) \bar{c}/\underline{c}$ . Therefore, the first inequality in (31) holds. To prove the second inequality, observe that

$$\|\Theta\|_F^2 = \text{Tr}(\mathbf{F}\mathbf{F}^T (\overline{\mathbf{U}}^T \overline{\mathbf{U}} \otimes \overline{\mathbf{U}}^T \overline{\mathbf{U}})) = \|\mathbf{F}\|_F^2$$

and, on the other hand,

$$\|\Theta\|_F^2 = \sum_{l=1}^L \|\mathbf{U}_{P,l}(\mathbf{U}_{P,l})^T\|_F^2 = \sum_{l=1}^L \text{Tr}(\mathbf{I}_K) = KL, \quad (72)$$

which together complete the proof.

**Proof of Lemma 3** Recall that formulas (11) and (16) imply that

$$\begin{aligned} \|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2 &\leq \left\| \widehat{\mathbf{C}}(\widehat{\mathbf{D}}_{\hat{c}})^{-1/2} - \mathbf{C}(\mathbf{D}_c)^{-1/2} \right\|_F^2 \\ &\leq 2 \left\| \widehat{\mathbf{C}}(\widehat{\mathbf{D}}_{\hat{c}})^{-1/2} \right\|_F^2 \left\| \mathbf{I}_M - (\widehat{\mathbf{D}}_{\hat{c}})^{1/2}(\mathbf{D}_c)^{-1/2} \right\|_F^2 + 2 \left\| \widehat{\mathbf{C}} - \mathbf{C} \right\|_F^2 \left\| (\mathbf{D}_c)^{-1/2} \right\|_F^2 \end{aligned} \quad (73)$$

where  $\mathbf{D}_c = \text{diag}(L_1, \dots, L_M)$  and  $\widehat{\mathbf{D}}_{\hat{c}} = \text{diag}(\widehat{L}_1, \dots, \widehat{L}_M)$ . It is easy to see that  $\|\widehat{\mathbf{C}}(\widehat{\mathbf{D}}_{\hat{c}})^{-1/2}\| = 1$  in (73), and that, by Assumption **A1**,  $\|(\mathbf{D}_c)^{-1/2}\|^2 \leq (\min L_m)^{-1} \leq M/(\underline{c}L)$ . Also,  $\|\widehat{\mathbf{C}} - \mathbf{C}\|_F^2 \leq 2L R_{BL}$ , and

$$\begin{aligned} \|\mathbf{I}_M - (\widehat{\mathbf{D}}_{\hat{c}})^{1/2}(\mathbf{D}_c)^{-1/2}\|_F^2 &= \text{Tr}(\mathbf{I}_M + \widehat{\mathbf{D}}_{\hat{c}}\mathbf{D}_c^{-1} - 2(\widehat{\mathbf{D}}_{\hat{c}})^{1/2}(\mathbf{D}_c)^{-1/2}) \\ &= \sum_{m=1}^M \frac{(\widehat{L}_m^{1/2} - L_m^{1/2})^2}{L_m} \leq \sum_{m=1}^M \frac{|\widehat{L}_m - L_m|}{L_m} \leq \frac{M}{\underline{c}L} \sum_{m=1}^M |\widehat{L}_m - L_m|, \end{aligned}$$

due to Assumption **A1**, and since for any  $a, b > 0$  one has  $(\sqrt{a} - \sqrt{b})^2 \leq |a - b|$ . Since  $\sum |\widehat{L}_m - L_m|$  is dominated by the number of clustering errors  $L R_{BL}$ , plugging all components into (73), obtain (51).

**Proof of Lemma 4** Let  $\mathbf{X}^{(l)} = \mathbf{A}^{(l)} - \mathbf{P}^{(l)}$ ,  $l = 1, \dots, \widetilde{L}$ . With some abuse of notations, for any square matrix  $\mathbf{Q}$ , let  $\text{diag}(\mathbf{Q})$  be the diagonal matrix which diagonal entries are equal

to the diagonal entries of  $\mathbf{Q}$ , while for any vector  $\mathbf{q}$ , let  $\text{diag}(\mathbf{q})$  be the diagonal matrix with the vector  $\mathbf{q}$  on the diagonal. Then,  $\widehat{\mathbf{G}} - \mathbf{G} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3$  where

$$\begin{aligned}\mathbf{S}_1 &= \sum_{l=1}^{\tilde{L}} (\mathbf{P}^{(l)} \mathbf{X}^{(l)} + \mathbf{X}^{(l)} \mathbf{P}^{(l)}), \quad \mathbf{S}_2 = \sum_{l=1}^{\tilde{L}} \left[ (\mathbf{X}^{(l)})^2 - \text{diag}((\mathbf{X}^{(l)})^2) \right], \\ \mathbf{S}_3 &= \sum_{l=1}^{\tilde{L}} \left[ \text{diag}((\mathbf{X}^{(l)})^2) - \text{diag}(\mathbf{A}^{(l)} \mathbf{1}) \right]\end{aligned}$$

Therefore,  $\|\widehat{\mathbf{G}} - \mathbf{G}\|^2 \leq 3(\|\mathbf{S}_1\|^2 + \|\mathbf{S}_2\|^2 + \|\mathbf{S}_3\|^2)$ .

To bound above  $\|\mathbf{S}_1\|^2$ ,  $\|\mathbf{S}_2\|^2$  and  $\|\mathbf{S}_3\|^2$ , apply Theorems 2 and 3 of Lei and Lin (2021) with  $v_1 = v_2 = 2\bar{c}_\rho \rho_n$ ,  $R_1 = R_2 = R'_2 = 1$  and  $v'_2 = 2\bar{c}_\rho^2 \rho_n^2$ . Using Theorems 2 with  $m = r = n$  and  $t^2 = \tau \bar{c}_\rho^2 C_\rho \rho_n^3 n^3 \tilde{L} \log n$ , obtain

$$\mathbb{P} \left\{ \|\mathbf{S}_1\|^2 \leq \tilde{C} \rho_n^3 n^3 \tilde{L} \log n \right\} \geq 1 - 4n^{1-\tau}$$

The first part of Theorem 3 yields that, due to Assumption **A3**,

$$\mathbb{P} \left\{ \|\mathbf{S}_2\|^2 \leq \tilde{C} \rho_n^2 n^2 \tilde{L} \log^2(n + \tilde{L}) \right\} \geq 1 - C(n + \tilde{L})^{1-\tau}$$

Now,  $\|\mathbf{S}_3\| \leq \|\mathbf{S}_3 - \mathbb{E}(\mathbf{S}_3)\| + \max_i |(\mathbb{E}\mathbf{S}_3)(i, i)|$ , since  $\mathbf{S}_3$  is a diagonal matrix. Applying second part of Theorem 3 with  $\sigma_2 = 1$  and  $\sigma'_2 = \sqrt{\tilde{L}n}$ , obtain

$$\mathbb{P} \left\{ \|\mathbf{S}_3 - \mathbb{E}(\mathbf{S}_3)\|^2 \leq \tilde{C} \rho_n n \tilde{L} \log^2(n + \tilde{L}) \right\} \geq 1 - C(n + \tilde{L})^{1-\tau}$$

Finally,

$$|(\mathbb{E}\mathbf{S}_3)(i, i)| = \left| \sum_{l=1}^{\tilde{L}} \left[ \mathbb{E} \sum_{j=1}^n [\mathbf{X}^{(l)}(i, j)]^2 - \sum_{j=1}^n \mathbf{P}^{(l)}(i, j) \right] \right| = \sum_{l=1}^{\tilde{L}} \sum_{j=1}^n [\mathbf{P}^{(l)}(i, j)]^2 \leq \rho_n^2 n \tilde{L},$$

which completes the proof.

## Acknowledgments

Both authors of the paper were partially supported by National Science Foundation (NSF) grant DMS-2014928.

## References

Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10(1):45–62, Mar 2019. doi: 10.1146/annurev-conmatphys-031218-013259. URL <http://dx.doi.org/10.1146/annurev-conmatphys-031218-013259>.

- Jesus Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E. Priebe, and Joshua T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *Journal of Machine Learning Research*, 22(142):1–49, 2021. URL <http://jmlr.org/papers/v22/19-558.html>.
- Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018. URL <http://jmlr.org/papers/v18/17-448.html>.
- Sharmodeep Bhattacharyya and Shirshendu Chatterjee. General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers. *ArXiv:2004.03480*, 2020.
- Piotr Brodka, Anna Chmiel, Matteo Magnani, and Giancarlo Ragozini. Quantifying layer similarity in multiplex networks: a systematic study. *Royal Society Open Science*, 5(8):171747, 2018. doi: 10.1098/rsos.171747. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rsos.171747>.
- Randy L. Buckner and Lauren M. DiNicola. The brains default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, pages 1–16, 2019.
- T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89, 2018. doi: 10.1214/17-AOS1541. URL <https://doi.org/10.1214/17-AOS1541>.
- Xiaobo Chen, Han Zhang, Yue Gao, Chong-Yaw Wee, Gang Li, Dinggang Shen, and the Alzheimer’s Disease Neuroimaging Initiative. High-order resting-state functional connectivity network for mci classification. *Human Brain Mapping*, 37(9):3282–3296, 2016. doi: 10.1002/hbm.23240. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23240>.
- Eric C. Chi, Brian J. Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58, 2020. URL <http://jmlr.org/papers/v21/18-155.html>.
- Nicolas A Crossley, Andrea Mechelli, Petra E Vértés, Toby T Winton-Brown, Ameera X Patel, Cedric E Ginestet, Philip McGuire, and Edward T Bullmore. Cognitive relevance of the community structure of the human brain functional coactivation network. *Proceedings of the National Academy of Sciences*, 110(28):11583–11588, 2013.
- Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. Structural reducibility of multilayer networks. *Nature Communications*, 6(6864), 2015. doi: 10.1038/ncomms7864.
- Daniele Durante, Nabanita Mukherjee, and Rebecca C. Steorts. Bayesian learning of dynamic multilayer networks. *Journal of Machine Learning Research*, 18(43):1–29, 2017. URL <http://jmlr.org/papers/v18/16-391.html>.

- Xing Fan, Marianna Pensky, Feng Yu, and Teng Zhang. Alma: Alternating minimization algorithm for clustering mixture multilayer network. *ArXiv:2102.10226*, 2021.
- Joshua Faskowitz, Xiaoran Yan, Xi-Nian Zuo, and Olaf Sporns. Weighted stochastic block models of the human connectome across the life span. *Scientific Reports*, 8(1):12997, 2018. doi: 10.1038/s41598-018-31202-1. URL <https://app.dimensions.ai/details/publication/pub.1106343698> and <https://www.nature.com/articles/s41598-018-31202-1.pdf>.
- Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624 – 2652, 2015. doi: 10.1214/15-AOS1354. URL <https://doi.org/10.1214/15-AOS1354>.
- Rungang Han, Yuetian Luo, Miaoyan Wang, and Anru R. Zhang. Exact clustering in tensor block model: Statistical optimality and computational limit. *ArXiv:2012.09996*, 2021.
- Shaobo Han and David B. Dunson. Multiresolution tensor decomposition for multiple spatial passing networks. *ArXiv:1803.01203*, 2018.
- Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multi-layer networks via regularized tensor decomposition. *ArXiv:2002.04457*, 2020.
- Ta-Chu Kao and Mason A. Porter. Layer communities in multiplex networks. *Journal of Statistical Physics*, 173(3-4):1286–1302, Aug 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1858-z. URL <http://dx.doi.org/10.1007/s10955-017-1858-z>.
- Mikko Kivela, Alex Arenas, Marc Barthélemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 07 2014. ISSN 2051-1329. doi: 10.1093/comnet/cnu016. URL <https://doi.org/10.1093/comnet/cnu016>.
- Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM REVIEW*, 51(3):455–500, 2009.
- A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time  $(1 + \epsilon)$ -approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462, Oct 2004. doi: 10.1109/FOCS.2004.7.
- Can M. Le and E. Levina. Estimating the number of communities in networks by spectral methods. *ArXiv:1507.00827*, 2015.
- Jing Lei. Tail bounds for matrix quadratic forms and bias adjusted spectral clustering in multi-layer stochastic block models. *ArXiv:2003.08222*, 2020.
- Jing Lei and Kevin Z. Lin. Bias-adjusted spectral clustering in multi-layer stochastic block models. *ArXiv:2003.08222*, 2021.
- Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015. doi: 10.1214/14-AOS1274.



- Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 12 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz068. URL <https://doi.org/10.1093/biomet/asz068>.
- Peter W. MacDonald, Elizaveta Levina, and Ji Zhu. Latent space models for multiplex networks with shared structure. *ArXiv:2012.14409*, 2021.
- Pedro Mercado, Antoine Gautier, Francesco Tudisco, and Matthias Hein. The power mean laplacian for multilayer graph clustering. *ArXiv:1803.00491*, 2018.
- B.C. Munsell, C.-Y. Wee, S.S. Keller, B. Weber, C. Elger, L.A.T. da Silva, T. Nesland, M. Styner, D. Shen, and L. Bonilha. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *NeuroImage*, 118:219–230, 2015.
- Carlo Nicolini, Cécile Bordier, and Angelo Bifone. Community detection in weighted brain connectivity networks beyond the resolution limit. *Neuroimage*, 146:28–39, 2017.
- Sofia C. Olhede and Patrick J. Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1400374111. URL <https://www.pnas.org/content/111/41/14722>.
- Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Statist.*, 10(2):3807–3870, 2016. doi: 10.1214/16-EJS1211. URL <https://doi.org/10.1214/16-EJS1211>.
- Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.*, 48(1):230–250, 02 2020. doi: 10.1214/18-AOS1800. URL <https://doi.org/10.1214/18-AOS1800>.
- C.R. Rao and M.B. Rao. *Matrix Algebra and its Applications to Statistics and Econometrics*. World Scientific Publishing Co., 1st edition, 1998.
- Olaf Sporns. Graph theory methods: applications in brain networks. *Dialogues in Clinical Neuroscience*, 20(2):111–121, 2018.
- Cornelis J. Stam. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695, 2014. doi: 10.1038/nrn3801. URL <https://app.dimensions.ai/details/publication/pub.1037745277>.
- Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. URL <https://doi.org/10.1007/s11222-007-9033-z>.
- Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/9be40cee5b0eee1462c82c6964087ff9-Paper.pdf>.

- Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis-kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014. ISSN 0006-3444. doi: 10.1093/biomet/asv008. URL <https://doi.org/10.1093/biomet/asv008>.
- Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018. doi: 10.1109/TIT.2018.2841377.
- Teng Zhang, Arthur Szlam, Yi Wang, and Gilad Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100(3):217–240, Dec 2012. ISSN 1573-1405. doi: 10.1007/s11263-012-0535-6. URL <https://doi.org/10.1007/s11263-012-0535-6>.