Consistency of community detection in multi-layer networks using spectral and matrix factorization methods

Subhadeep Paul and Yuguo Chen [1]

Abstract

We consider the problem of estimating a consensus community structure by combining information from multiple layers of a multi-layer network or multiple snapshots of a time-varying network. Numerous methods have been proposed in the literature for the more general problem of multi-view clustering in the past decade based on the spectral clustering or a low-rank matrix factorization. As a general theme, these "intermediate fusion" methods involve obtaining a low column rank matrix by optimizing an objective function and then using the columns of the matrix for clustering. However, the theoretical properties of these methods remain largely unexplored and most researchers have relied on the performance in synthetic and real data to assess the goodness of the procedures. In the absence of statistical guarantees on the objective functions, it is difficult to determine if the algorithms optimizing the objective will return a good community structure. We apply some of these methods for consensus community detection in multi-layer networks and investigate the consistency properties of the global optimizer of the objective functions under the multi-layer stochastic blockmodel. We derive several new asymptotic results showing consistency of the intermediate fusion techniques along with the spectral clustering of mean adjacency matrix under a high dimensional setup, where the number of nodes, the number of layers and the number of communities of the multi-layer graph grow. Our numerical study shows that in comparison to the intermediate fusion techniques, late fusion methods, namely spectral clustering on aggregate spectral kernel and module allegiance matrix, underperform in sparse networks, while the spectral clustering of mean adjacency matrix under-performs in multi-layer networks that contain layers with both homophilic and heterophilic clusters.

KEYWORDS: Community detection; consistency; Co-regularized spectral clustering; Multi-layer networks; Multi-layer stochastic block model; Orthogonal linked matrix factorization; Three way tensor factorization

1 Introduction

The study of multi-layer networks has received significant interest recently, driven by its myriad of applications in neuroscience, economics, genetics and social sciences [MRM+10, KAB+14, BBC+14, HXA15]. A multi-layer network is a powerful representation of relational data with the nodes representing the entities of interest and the network layers representing the multiple relations among those entities. While the term "multi-layer network" is often used in a more general context, we focus our attention only on a network where the nodes are connected only within a layer and there are no inter-layer edges (such networks are also called "multiplex networks" in the literature).

[1]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: *spaul10@illinois.edu*, *yuguo@illinois.edu*). Supported in part by NSF Grant DMS-1406455.

A dynamic or time-varying network represents different states of a single network over time. A dynamic network can also be represented as a multi-layer network, with the same node in consecutive time period usually being linked by an edge to respect the time ordering [MRM+10, BPW+13, GZC+16]. When appropriate for the application, e.g., in the problem of consensus community detection, we can ignore the time order and consider a time-varying network as a regular multi-layer network with no inter-layer edges [HXA15, BSW+15].

The problem of consensus community detection in multi-layer and dynamic networks has many important applications. Often in such networks one underlying community structure is in force while the different layers of interactions are merely different manifestations of the unobserved community structure. For example, in the multi-layer twitter networks in [GC13], ground truth community memberships can be attributed to the users (nodes) based on attributes more fundamental and independent of the observed twitter interactions (e.g., political views, country of origin, football clubs), whereas the interactions provide multiple sources of information about the same latent community structure. Combining information from these multiple sources would then lead to enhanced performance in the learning task. Moreover, different representations of the same phenomenon often provide complimentary information, any one of which is not sufficient to describe the underlying process (see [LWGH13] and the examples therein).

Even in situations where the hypothesis of a single constant community structure may not be true, e.g., in the analysis of dynamic brain networks, it is still often desirable to obtain a consensus partition that does not vary over time, but is a static average partition that remains in force throughout the experiment. Such an overall partition is crucial to obtain stable modules of brain regions as baseline for computing measures of local and global dynamism in the brain, e.g., "flexiblity" and "integration" in [BWP+11] and [BSW+15].

The present problem is also related to a more general class of problems that generally goes under the theme of multi-view clustering and has received considerable attention over the last decade, particularly in the computer science community. Numerous methods have been proposed to combine information from multiple views of a multi-view relational data for clustering. The goal is usually to leverage the diversity and often complimentary nature of the information in different layers to outperform simply summing the layers or using any one of the layers [LWGH13]. A great many of those methods use spectral clustering or a low rank matrix factorization as a basis [LZWY06, ZB07, SG08, TLD09, KRD11, NTK11, DFVN12, LWGH13].

The "linked matrix factorization" algorithm in [TLD09] and "RESCAL" algorithm in [NTK11] approximate the adjacency matrices in each layer of a multi-layer graph, or each slice of a three way tensor, with a low rank symmetric matrix factorization. While one of the factors is shared the other one varies across layers or slices. Although the algorithms employed in the two papers are quite different, the factorization in both cases is computed by minimizing an identical joint Frobenius norm objective function. [DFVN12] use similar common low rank matrix factorization ideas with a slightly different objective function to obtain a "joint spectrum" of a multi-layer graph which is subsequently used for clustering.

The co-regularized spectral clustering in [KRD11] with centroid based co-regularization maximizes the combined normalized cut objective function over the Laplacian matrices from all views of the data, subject to a smoothness penalty. This idea is similar to the evolutionary spectral clustering used in [CSZ+07] for clustering dynamic networks with a temporal

smoothness penalty, and is part of a general theme of co-regularization in multi-view machine learning [XTX13]. The co-regularization framework was extended to "joint non-negative matrix factorization" using a Frobenius norm based objective function in [LWGH13]. See [Sun13] and [XTX13] for surveys of multi-view learning methods.

However, there is a lack of theoretical understanding of the objective functions in these spectral and matrix factorization based methods. Researchers often rely on simulations and applications to specific datasets to compare the methods. However, this approach fails to explore different scenarios that might arise in practice. For example, in multi-layer network applications, the component layers might have very different sparsity, signal quality and node degree distributions. Hence it is important to explore the utility of the methods under different statistical models and asymptotic settings through a principled theoretical study.

In this article we investigate the consistency properties of various methods for community detection under data generated from a multi-layer network model, the multi-layer stochastic blockmodel (MLSBM) [VCMGSP16, HXA15, Pei15, PC16a, SSTM16, PC16b, WPBN16, BDLBH17]. We derive several asymptotic results to show consistency of the global optimizers of co-regularized spectral clustering and orthogonal linked matrix factorization under a high dimensional asymptotic setup where the number of nodes, the number of layers and the number of communities all grow. We use slight variations of the original algorithms to compute the solutions to the respective optimization problems. We note that both algorithms are not guaranteed to reach a global optimum. The present paper is an attempt to prove goodness of the objective functions rather than the algorithms, and is concerned with the following question: If it is possible to compute a global optimum or approximate one reasonably, will the global solution be consistent under a random graph model, namely the MLSBM?

In addition to the two methods mentioned above, we also consider two baseline methods previously used in literature. The first method is performing spectral clustering [NJW02, RCY11] on the mean of the adjacency matrices from different layers of the multi-layer network. This method has also been considered in [TLD09] and [DFVN12] as a baseline method and is generally thought to be a simple but effective procedure [KRD11]. In addition to including this method in our numerical comparisons, we also study its asymptotic consistency under MLSBM.

The second baseline method first computes a low dimensional spectral embedding (matrix of eigenvectors corresponding to top eigenvalues) and creates a spectral kernel for each layer, and then aggregates these kernel functions. A single layer community detection method is then applied to this aggregate spectral kernel [TLD09]. Another variation of this idea is to compute the community assignments in each layer independently using a single layer method (e.g., modularity or spectral clustering), and then create a "module allegiance matrix" on which a single layer community detection method can be applied to compute the consensus communities [BSW+15].

The rest of the article is organized as follows. Section 2 describes the methods and algorithms considered in the article. Section 3 describes the MLSBM, defines mis-clustering rate and proves correct recovery in the noiseless case. Section 4 describes the consistency results. Section 5 contains a simulation study to numerically evaluate the methods. Section 6 gives concluding remarks. The Appendix contains all the proofs.

2 Methods and algorithms

We define an undirected multi-layer network with M layers as a collection of graphs $\mathcal{G} = \{G^{(1)}, \ldots, G^{(M)}\}$ over a common set of n vertices. The vertices represent the entities/actors, while the layers represent different types of interactions among the entities. For the layer of the mth type, we define the adjacency matrix $A^{(m)}$ corresponding to that layer as follows: $A^{(m)}_{ij} = 1$, if there is an edge of type m between nodes i and j, and $A^{(m)}_{ij} = 0$, otherwise.

We define the vector of degrees of node i as $\mathbf{d_i} = \{d^{(m)}_i; m = 1, \ldots, M\}$, where $d^{(m)}_i = \sum_j A^{(m)}_{ij}$ is its degree of the mth type. Then the normalized graph Laplacian matrix for the mth layer can be defined as $L^{(m)} = (D^{(m)})^{-1/2} A^{(m)} (D^{(m)})^{-1/2}$, where $D^{(m)}$ is a diagonal matrix with the degrees of the mth type of the nodes as elements, i.e., $D^{(m)}_{ii} = d^{(m)}_i$. Together the M adjacency matrices create the three-way $n \times n \times M$ adjacency tensor of the multi-layer network $\mathbb{A} = \{A^{(1)}, \ldots, A^{(M)}\}$. The corresponding Laplacian tensor is defined as $\mathbb{L} = \{L^{(1)}, \ldots, L^{(M)}\}$. We denote the number of communities in the network by k. It will be assumed to be known throughout the paper. We use the notations $\|\cdot\|_2$, $\|\cdot\|_F$ and $\|\cdot\|_\Sigma$ to denote the spectral (operator) norm, the Frobenius norm and the trace norm, respectively, while $\mathrm{tr}(\cdot)$ denotes the matrix trace. We will use $\sin\Theta(U, V)$ to denote the diagonal matrix whose elements are sines of the principle angles between the subspaces $\mathcal{U}$ and $\mathcal{V}$, spanned by the columns of the matrices U and V respectively (Definition 1.5.3 in [SS90]).

We consider the following methods and algorithms for consensus clustering in multi-layer networks. The first two methods are so called "intermediate fusion" techniques whereby the multiple layers are integrated through a clustering objective function [LWGH13]. Such methods are often preferred over "early" and "late" fusion techniques due to superior performance [ŽZ15].

2.1 Linked matrix factorization

The first of the intermediate fusion methods is the linked matrix factorization (LMF) for clustering multiple graphs in [TLD09]. Our adaptation of the method is slightly different from the one described in [TLD09] in the sense that we enforce the columns of the shared factor to be strictly orthonormal and consequently drop the Frobenius norm regularization term (indeed this has been suggested in [TLD09]). In our simulations, we found the performance of both methods to be the same. To avoid confusion, we call our adaptation the orthogonal LMF (OLMF). Note that LMF has the identical objective function as the RESCAL algorithm, which is a three-way tensor factorization for learning in multi-relational data [NTK11]. However the algorithm for RESCAL is different from that of LMF.

The OLMF solves the following optimization problem on the adjacency tensor of a multi-layer network:

$$[\hat{P}, (\hat{\Lambda}^{(1)}, \ldots, \hat{\Lambda}^{(M)})] = \underset{P^T P = I}{\arg\min} \sum_{m=1}^{M} \|A^{(m)} - P\Lambda^{(m)} P^T\|_F^2, \tag{2.1}$$

where $P \in \mathbb{R}^{n \times k}$ is a common factor matrix and $\Lambda^{(m)} \in \mathbb{R}^{k \times k}$ are M layer specific symmetric

factor matrices. This is equivalent to the following optimization problem:

$$\hat{P} = \arg\max_{P^T P = I} \sum_{m=1}^{M} \|P^T A^{(m)} P\|_F^2, \quad \hat{\Lambda}^{(m)} = \hat{P}^T A^{(m)} \hat{P},\ m = \{1, \ldots, M\}. \tag{2.2}$$

We will refer the objective function in (2.2) as $F(\mathbb{A}, P)$. While we require P to have orthonormal columns, we do not put any constraint on the $\Lambda^{(m)}$ matrices, and specifically we do not require them to be diagonal matrices. Note that in general $[\hat{P}, \hat{\Lambda}^{(m)}]$ is not the solution of the problem of finding the best at most rank k approximating matrix for $L^{(m)}$. Hence in general, the matrices $\hat{\Lambda}^{(m)}$ are not the diagonal matrices of singular values. Intuitively the shared factor P is expected to capture the common characteristics of the nodes in a multi-layer network including the latent community structure, while the different $\Lambda^{(m)}$ matrices capture the layer/relation specific characteristics.

We propose a BFGS algorithm to solve the OLMF optimization problem, similar to the algorithm in [TLD09]. The gradients are given by

$$\frac{\partial O}{\partial P} := -\sum_m (I - PP^T) A^{(m)} P \Lambda^{(m)},$$
$$\frac{\partial O}{\partial \Lambda^{(m)}} := -P^T (A^{(m)} - P\Lambda^{(m)} P^T) P, \quad m = 1, \ldots, M,$$

where O denotes the objective fuction in (2.1). Once the algorithm converges, we cluster the rows of the matrix P using the k-means algorithm. Since each row in P corresponds to one of the nodes, this gives a community assignment for the nodes.

2.2 Co-regularized spectral clustering

The second intermediate fusion method we study is the co-regularization based approach to multi-layer spectral clustering due to [KRD11]. The idea of co-regularization has also been previously applied to various learning problems [XTX13]. We adopt the centroid based co-regularization method from [KRD11] unchanged in the context of multi-layer networks. The method, applied to the adjacency tensor, is based on solving the following optimization problem:

$$[\hat{U}^{(1)}, \ldots, \hat{U}^{(M)}, \hat{U}^*] = \underset{\substack{U^{(m)T} U^{(m)} = I, \forall m, \\ U^{*T} U^* = I}}{\arg\max} \sum_{m=1}^{M} \{ \operatorname{tr}(U^{(m)T} A^{(m)} U^{(m)}) + \gamma_m \operatorname{tr}(U^{*T} U^{(m)} U^{(m)T} U^*) \}, \tag{2.3}$$

where $U^{(1)}, \ldots, U^{(M)}$ and U^* are $n \times k$ matrices with orthonormal columns. We denote $\mathbb{U}$ as the tensor containing the matrices $\{U^{(1)}, \ldots, U^{(M)}\}$. The objective function of the optimization problem in (2.3) is denoted as $F(\mathbb{A}, \mathbb{U}, U^*)$. The optimization problem can be easily solved by alternating eigen decomposition of the matrices $A^{(m)} - \gamma_m U^* U^{*T}$ and $\sum_m \gamma_m U^{(m)} U^{(m)T}$ [KRD11]. After the algorithm converges, consensus community assignments for the nodes can be obtained by clustering the rows of the matrix $\hat{U}^*$ with the k-means algorithm.

Note that the objective function contains two parts. The first part is the usual association cut spectral clustering objective function for different layers. The second part is a penalty function that seeks to maximize the cohesion between the eigenspaces obtained from different layers. To see this, we have the following proposition that characterizes the second part in terms of $\|\sin\Theta(U^{(m)}, U^*)\|_F$, which measures the distance between the column spaces spanned by $U^{(m)}$ and U^* [SS90]. The proof of this proposition, along with all lemmas and theorems, can be found in the Appendix.

Proposition 1. *For $U^{(m)}$ and U^* as defined above, we have*

$$\mathrm{tr}(U^{*T}U^{(m)}U^{(m)T}U^*) = k - \frac{1}{2}\|U^*U^{*T} - U^{(m)}U^{(m)T}\|_F^2 = k - \|\sin\Theta(U^{(m)}, U^*)\|_F^2.$$

The penalty function alone is maximized when all the subspaces are identical, since $\|\sin\Theta(U^{(m)}, U^*)\|_F$ is 0 when the subspaces spanned by $U^{(m)}$ and U^* are identical [SS90]. Hence the objective function represents a trade-off between optimizing the community structure in each layer (which might be noisy) and maintaining similarity with the mean community structure. The weights γ_m's should be chosen to reflect both the desired trade-off between this two competing goals and the relative importance of the different layers. In particular, small values of γ_m's will prevent sharing information across layers, which will result in estimates of $U^{(m)}$ being the one that is best for its own layer and the U^* simply being the matrix of eigenvectors of $\sum_m \gamma_m U^{(m)}U^{(m)T}$. On the other hand, large values of γ_m's will ensure the $U^{(m)}$'s try to achieve similarity with a common U^* in expense of being sub-optimal for its own layer.

2.3 Spectral clustering on mean adjacency matrix

The first of the two baseline procedures we consider collapses the multi-layer network into a single layer network by taking the mean of the adjacency matrices from each of the layers. The usual single layer spectral clustering algorithm [NJW02, RCY11] is then applied to the resultant matrix. This procedure can be thought of as an "early integration" or "early fusion" technique, since data from multiple layers are aggregated before any processing is made [ŽZ15]. Spectral clustering on some form of the aggregate matrix has appeared as a "baseline procedure" in [TLD09, KRD11, DFVN12] and [TWL12]. In particular, consensus community detection proceeds through spectral clustering of the matrix $\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$. Consistency results for this method under the stochastic block model (SBM) were derived in [HXA15] in the scenario when the number of layers grows but the number of nodes does not. [CHI16] also derived phase transition results for a weighted version of this method under a model they characterize as "multi-layer signal plus noise model".

2.4 Aggregate spectral kernel and module allegiance matrix

The other baseline method we consider is a "late fusion" technique, where we first compute the eigenvector matrices $U^{(m)}$'s corresponding to the top k eigenvalues from each of the M

layers of the graph and construct the aggregate spectral kernel matrix

$$K_{n\times n} = \frac{1}{M}\sum_{m=1}^{M} U^{(m)}U^{(m)T}.$$

However, instead of using kernel k-means to cluster the resulting matrix K as in [TLD09] and [DFVN12], we apply spectral clustering to this matrix again to obtain the community assignments. We call this method "aggregate spectral kernel". This is in spirit of clustering the "module allegiance matrix" described in [BSW+15], where community assignment for each layer is first obtained using the Newman-Girvan modularity [NG04], and subsequently an $n \times n$ module allegiance matrix is formed, each of whose elements counts the number of times two nodes appear in the same module.

We use both the aggregate spectral kernel and the module allegiance matrix methods in our numerical study. It is worth pointing out that these methods are distinct from the majority voting method described in [HXA15, PC16a]. Although, much like the majority voting, these methods process each layer separately and fuse information later, one advantage is that both the aggregate spectral kernel and module allegiance matrix methods avoid the cumbersome issue of label switching ambiguity. To see this, assume we have two community assignment matrices Z_1 and Z_2 with $Z_1 = Z_2P$, where P is a permutation matrix, i.e., Z_2 gives the same community assignments as Z_1 but with its labels switched. However when we compute the module allegiance matrix, we have $Z_1Z_1^T = Z_2PP^TZ_2^T = Z_2Z_2^T$. The same is true for the aggregate spectral kernel. Intuitively, for each element they are concerned with whether two nodes belong to the same community or not, irrespective of which community that is. Hence they do not require solving a linear sum assignment problem as is required for majority voting.

3 Models and mis-clustering

The multi-layer stochastic block model (MLSBM) is a statistical model of multi-layer networks with a shared latent community structure [HLL83, HXA15, PC16a]. We define the k block, M layer, n node MLSBM as follows. Each node of the network is assigned a community label vector of length k, which takes the value of 1 at the position corresponding to its community and 0 in all other positions. Let Z denote the $n \times k$ community assignment matrix whose ith row Z_i is the community label vector for the ith node.

Given the community labels of two nodes, the edges between them in different layers are formed independently following a Bernoulli distribution with a probability that depends only on the community assignments and the relation the edge depicts. Hence within a community the nodes have "stochastic equivalence" in the sense that the probability of an edge formation (in any layer) with another node is the same for all the nodes in a community. We further assume that there is at least one node in each community which implies that there is at least one non-zero element in each column of Z.

The k block, M layer, n node MLSBM with parameters $[Z, \mathbb{B} = \{B^{(1)}, \ldots, B^{(M)}\}]$ can be written in the matrix form as

$$E(A^{(m)}) = \mathcal{A}^{(m)} = ZB^{(m)}Z^T, \quad B^{(m)} \in [0,1]^{k\times k},\ Z \in \{0,1\}^{n\times k}, \tag{3.1}$$

where the matrices $B^{(m)}$ are $k \times k$ non-negative symmetric matrices of probabilities. For our analysis we assume each of the $B^{(m)}$'s is full rank, which is a standard assumption in the analysis of spectral clustering for single layer SBM as well [RCY11, LR14]. We will refer to the matrix $\mathcal{A}^{(m)}$ as the population adjacency matrix for the mth layer and the tensor $\mathscr{A} = \{\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(m)}\}$ as the population adjacency tensor.

3.1 Correct recovery in the noiseless case

Before we can tackle consistency of the methods, the first question that needs to be answered is whether a method can correctly recover the community assignments from the true population adjacency tensor when there is no sampling noise involved. The following lemma shows that OLMF, co-regularized spectral clustering, spectral clustering of mean adjacency matrix, and aggregate spectral kernel, all can correctly identify the node community labels from the population adjacency tensor of MLSBM.

Lemma 1. *Let $\mathscr{A} = \{\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(M)}\}$ be the three-way $n \times n \times M$ population adjacency tensor for MLSBM $[Z, \mathbb{B}]$ with each of the M slices $\mathcal{A}^{(m)} \in \mathbb{R}^{n \times n}$ defined as in (3.1). Then we have the following results:*

(i) The optimization problem in (2.2) of orthogonal linked matrix factorization applied to the tensor $\mathscr{A}$ has $\bar{P} = ZQ^{-1/2}, \bar{\Lambda}^{(m)} = Q^{1/2}B^{(m)}Q^{1/2}, m = 1, \ldots, M$, as the unique solution up to an orthogonal matrix, where $Q = Z^TZ$. Further $Z_iQ^{-1/2} = Z_jQ^{-1/2}$ if and only if $Z_i = Z_j$.

(ii) The optimization problem in (2.3) of co-regularized spectral clustering applied to the tensor $\mathscr{A}$ has $\bar{U}^{(m)} = Z\mu^{(m)}$, $m = 1, \ldots, M$, $\bar{U}^* = ZQ$ as the unique solution up to an orthogonal matrix, where $\mu^{(m)}$ and Q are invertible matrices. Further $Z_iQ = Z_jQ$ if and only if $Z_i = Z_j$.

(iii) The matrix containing the eigenvectors corresponding to the k largest eigenvalues of $\bar{\mathcal{A}} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{A}^{(m)}$ is ZQ for some invertible matrix $Q \in \mathbb{R}^{k \times k}$ provided the matrix $\frac{1}{M}\sum_{i=1}^{M}B^{(m)}$ is of full rank. Further $Z_iQ = Z_jQ$ if and only if $Z_i = Z_j$.

(iv) Define $\bar{K} = \frac{1}{M}\sum_{i=1}^{M}\bar{U}^{(m)}\bar{U}^{(m)T}$, where $\bar{U}^{(m)}$ is the matrix of eigenvectors corresponding to the largest k eigenvalues of $\mathcal{A}^{(m)}$. The matrix containing the eigenvectors corresponding to the k largest eigenvalues of $\bar{K}$ is ZQ for some invertible matrix $Q \in \mathbb{R}^{K \times K}$. Further $Z_iQ = Z_jQ$ if and only if $Z_i = Z_j$.

We make two observations on the results of this lemma. First, note that in all of the above methods, the matrix whose rows are clustered using k-means algorithm for community detection has only k distinct rows. Moreover, two rows are identical if and only if they are identical in the true community assignment matrix. This ensures that k-means algorithm in each case will correctly cluster the rows. Second, the spectral clustering on mean adjacency matrix requires an additional condition on the aggregate connectivity matrix $\frac{1}{M}\sum_{i=1}^{M}B^{(m)}$, which is not needed for the intermediate and late fusion methods. This is also related to the general issues associated with aggregating a multi-layer graph with diverse layers, explored from an information theoretical point in [PC16a]. Third, the noiseless recovery in co-regularized spectral clustering does not depend on what we choose for γ_m's. This quite counter-intuitive phenomenon is true because both parts of the objective function, the

association cut and the penalty term, are separately maximized by the true communities, with the penalty term achieving its global maximum irrespective of γ_m.

3.2 Characterizing mis-clustering

Although Lemma 1 shows that the methods under consideration can perfectly recover community labels from the true population adjacency tensor, in reality we do not observe the true population tensor. Instead we observe a noisy sample version of it. Consequently, community assignment using the methods will lead to some error. For a given benchmark community assignment and an estimated community assignment, we define a mis-clustering rate as the proportion of nodes for which the assignments do not agree. Let $\bar{e}$ denote the vector of true community labels extracted from Z and $\hat{e}$ denote the vector of a candidate assignment. Then we define the mis-clustering rate

$$r = \frac{1}{n} \inf_{\Pi} d_H(\bar{e}, \Pi(\hat{e})),$$

where $\Pi(\cdot)$ is a permutation of the labels and $d_H(\cdot, \cdot)$ is the Hamming distance between two vectors.

Note that in each of the methods we consider, we obtain a low rank matrix with orthonormal columns whose rows are then clustered using the k-means algorithm for community detection. Hence we also need to relate this mis-clustering rate with the low rank matrices obtained from the methods. For a method under consideration, let $\hat{U}_{n\times k}$ be the low rank matrix with orthonormal columns it outputs, whose rows can subsequently be clustered to estimate community assignment $\hat{e}$. Then we have the following relationship,

$$r \leq \frac{8n_{\max}}{n} \|\hat{U} - Z(Z^T Z)^{-1/2} O\|_F^2, \tag{3.2}$$

where O is an arbitrary orthogonal matrix and $n_{\max}$ is the number of nodes in the largest true community [RCY11].

4 Consistency results

In this section we investigate the asymptotic consistency of consensus community detection using the methods outlined in Section 2. The asymptotic setup we consider is as follows. We let both n and M grow, and assume no relationship between their growth rate. However we will be most interested in the case when M grows faster than n. This framework is particularly suitable for consensus community detection in dynamic graphs, where the number of layers represents the number of temporal snapshots available to us and can potentially be exponentially larger compared to the number of nodes. We also let the number of communities k (which is assumed to be known in advance) to grow with both n and M.

Before proceeding with the main results we prove the following theorem with two results on a multi-layer graph with independent edges, the first of which extends the results contained in [CR11] to multi-layer graph settings and the second one is a new result using matrix Chernoff bound [Tro12].

Theorem 1. *Let $\mathcal{G}$ be a multi-layer graph with each edge being independent of all other edges of all types. Let $\mathbb{A}$ be its adjacency tensor and let $\mathscr{A}$ denote the expected adjacency tensor. Further, let $\Delta^{(m)}$ be the maximum expected degree for a node in layer m. Define $\Delta_{\max} = \max\{\Delta^{(1)}, \ldots, \Delta^{(M)}\}$ as the maximum expected degree of a node in any layer of the multi-layer graph. Then we have the following results:*

(i) For any $\epsilon > 0$, if $M\Delta_{\max} > \frac{4}{9}\log(2n/\epsilon)$, then with probability at least $1-\epsilon$,

$$\|\frac{1}{M}\sum_m (A^{(m)} - \mathcal{A}^{(m)})\|_2 \leq \sqrt{\frac{4\Delta_{\max}\log(2n/\epsilon)}{M}}.$$

(ii) If $\Delta_{\max} > \frac{4}{3}\log n$, then with probability at least $1 - o(1)$ we have

$$\|\frac{1}{M}\sum_m (A^{(m)} - \mathcal{A}^{(m)})^2\|_2 \leq \Delta_{\max}(\log n)^{2+\epsilon}$$

for some $\epsilon > 0$.

Next we use the results of Theorem 1 to prove consistency results for co-regularized spectral clustering, OLMF and spectral clustering in mean adjacency matrix. The common settings under which the results are proved are as follows. Let $\mathcal{G}$ be a multi-layer network with M layers generated from the MLSBM with parameters $[Z, \mathbb{B}]$. Let $\mathbb{A}$ be its adjacency tensor. Let $\lambda^{(m)}$ denote the minimum eigenvalue of the mth layer population adjacency matrix and $n_{\max}$ denote the number of nodes in the largest true community.

Theorem 2. *Let $[\hat{\mathbb{U}}, \hat{U}^*]$ be the solution that maximizes the co-regularized spectral clustering objective function in (2.3) applied to $\mathbb{A}$, and r_{coreg} be the fraction of nodes misclusterd by a k-means procedure applied to $\hat{U}^*$. Assume $M\Delta_{\max} > \frac{4}{9}\log(2n/\epsilon)$. If the conclusion (i) of Theorem 1 holds, and we choose γ_m large enough such that $\gamma_m \geq \|A^{(m)}\|_2 \left(\frac{\|\sin\Theta(\hat{U}^{(m)}, \hat{U}^*)\|_\Sigma}{\|\sin\Theta(\hat{U}^{(m)}, \hat{U}^*)\|_F^2}\right)$ for all m, then for any $\epsilon > 0$, with probability at least $1-\epsilon$,*

$$r_{coreg} \leq \frac{128 n_{\max} k (\Delta_{\max})^{3/2}}{\frac{1}{M}\sum_m (\lambda^{(m)})^2 n}\sqrt{\frac{\log(4n/\epsilon)}{M}}.$$

Several discussions on the results of the previous theorem are in order. First, although correct recovery under the noiseless case does not require any condition on γ_m's, the consistency requires γ_m's to be larger than $\|A^{(m)}\|_2$ multiplied by a factor involving the principle angles between the estimated subspaces $\hat{U}^{(m)}$ and $\hat{U}^*$. Based on this result, in our simulations in Section 5 we choose $\gamma^{(m)}$ to be a constant times $\max \|A^{(m)}\|_2$, identically in each layer. Second, since it is not immediately clear when the above bound will imply consistent community detection, we make some further assumptions to simplify the bound. In particular we interpret the bound under a multi-layer extension of the four parameter stochastic blockmodel introduced in [RCY11].

Co-regularized spectral clustering under four parameter MLSBM

We define a MLSBM of M layers and n nodes with four parameters $\mathbf{p} = \{p^{(1)}, \ldots, p^{(M)}\}, \mathbf{q} = \{q^{(1)}, \ldots, q^{(M)}\}, k, s$ as follows. In layer m, the connection probability within a community is $p^{(m)}$ and between communities is $q^{(m)}$. We assume $p^{(m)} \neq q^{(m)}$ but are of the same asymptotic order with respect to n, for all m. The number of communities is k and all communities are of the same size $s = n/k$. Hence $n_{\max} = s = n/k$. We have the following lemma on the minimum eigenvalues of the population adjacency matrices $\lambda^{(m)}$'s.

Lemma 2. *For the four parameter MLSBM, $\lambda^{(m)} = s(p^{(m)} - q^{(m)})$, for all $m = 1, \ldots, M$.*

Let $a^{(m)}\frac{\Delta_{\max}}{n} = p^{(m)}$ and $b^{(m)}\frac{\Delta_{\max}}{n} = q^{(m)}$. Then $\lambda^{(m)} = \frac{\Delta_{\max}}{k}(a^{(m)} - b^{(m)})$. Consequently, the common asymptotic order of $p^{(m)}$ and $q^{(m)}$ is captured in the $\frac{\Delta_{\max}}{n}$ term and $a^{(m)} \asymp b^{(m)} \asymp 1$. However, note that the difference $a^{(m)} - b^{(m)}$ could still be very small. Define $f(\mathbf{a}, \mathbf{b}) = \frac{1}{M}\sum_m (a^{(m)} - b^{(m)})^2$. Then Theorem 2 implies

$$r_{coreg} \lesssim \frac{\frac{n}{k} k \Delta_{\max}^{3/2}}{(\frac{\Delta_{\max}}{k})^2 n f(\mathbf{a}, \mathbf{b})} \sqrt{\frac{\log(4n/\epsilon)}{M}} \asymp \frac{k^2}{\sqrt{M\Delta_{\max}/\log(4n/\epsilon)} f(\mathbf{a}, \mathbf{b})}.$$

Hence we have consistency of community detection as long as
$k = o((M\Delta_{\max}/\log(4n/\epsilon))^{1/4}\sqrt{f(\mathbf{a}, \mathbf{b})})$.

We consider three growth regimes on the density of the component layers of the multi-layer graph. In the first regime we assume the dense graph setting where the vectors $\mathbf{p}$ and $\mathbf{q}$ do not change with n. This implies that $\Delta_{\max} \asymp n$ and consequently

$$r_{coreg} \lesssim \frac{\frac{n}{k} k n^{3/2}}{(\frac{n}{k})^2 n f(\mathbf{a}, \mathbf{b})} \sqrt{\frac{\log(4n/\epsilon)}{M}} \asymp \frac{k^2}{\sqrt{nM/\log(4n/\epsilon)} f(\mathbf{a}, \mathbf{b})}.$$

Hence as long as $k = o((nM/\log(4n/\epsilon))^{1/4}\sqrt{f(\mathbf{a,b})})$, $r_{coreg} \to 0$ with probability at least $1 - \epsilon$, and we have consistent community detection.

In the second regime, we assume a semi-sparse setting where both $p^{(m)}$ and $q^{(m)}$ are of the order of $\log n/n$ for all m. Then $\Delta_{\max} \asymp \log n$ and we have

$$r_{coreg} \lesssim \frac{\frac{n}{k} k (\log n)^{3/2}}{(\frac{\log n}{k})^2 n f(\mathbf{a}, \mathbf{b})} \sqrt{\frac{\log(4n/\epsilon)}{M}} \asymp \frac{k^2}{\sqrt{M} f(\mathbf{a}, \mathbf{b})}.$$

This implies that in this setting, as long as $k = o\left(M^{1/4}\sqrt{f(\mathbf{a}, \mathbf{b})}\right)$, $r_{coreg} \to 0$, and we have consistent community detection.

Finally in the sparse "constant degree" regime, where $p^{(m)}$ and $q^{(m)}$ are of the order of $1/n$ for all m, we have $\Delta_{\max} \asymp 1$. Note that the density condition on the layers of the network for Theorem 2(i) to hold is $M\Delta_{\max} = \omega(\log n)$, which can be satisfied even in the constant degree regime if $M = \omega(\log n)$. If this is satisfied, then we have from Theorem 2 that

$$r_{coreg} \lesssim \frac{\frac{n}{k} k}{(\frac{1}{k})^2 n f(\mathbf{a}, \mathbf{b})} \sqrt{\frac{\log(4n/\epsilon)}{M}} \asymp \frac{k^2}{\sqrt{M/\log(4n/\epsilon)} f(\mathbf{a}, \mathbf{b})}.$$

Hence consistent community detection is possible as long as $k = o((M/\log(4n/\epsilon))^{1/4}\sqrt{f(\mathbf{a},\mathbf{b})})$. Consequently, a large number of very sparse graphs can also lead to consistent community detection, whereas in single layer networks consistent recovery is not possible in the constant degree regime. This is also true for spectral clustering in mean adjacency matrix as we will see in Theorem 4, and is along the lines of the results obtained in [PC16a].

The next theorem proves a similar result for the OLMF method.

Theorem 3. *Let $[\hat{P}, (\hat{\Lambda}^{(1)}, \ldots, \hat{\Lambda}^{(M)})]$ be the solution that minimizes the OLMF objective function in (2.1) applied to $\mathbb{A}$, and r_{LMF} be the fraction of nodes misclustered by a k-means procedure applied to $\hat{P}$. If $\Delta_{\max} > \frac{4}{3}\log n$ and part (ii) of Theorem 1 holds, then with probability at least $1 - o(1)$,*

$$r_{LMF} \leq \frac{16 n_{\max} k \Delta_{\max} \log n(\sqrt{\Delta_{\max}} + \log n)}{\frac{1}{M}\sum_m (\lambda^{(m)})^2 n}.$$

This bound can also be simplified under the four parameter MLSBM defined earlier. Under the four parameter MLSBM, the bound in Theorem 3 simplifies to

$$r_{LMF} \lesssim \frac{\Delta_{\max}\log n(\sqrt{\Delta_{\max}} + \log n)}{(\frac{\Delta_{\max}}{k})^2 f(\mathbf{a},\mathbf{b})} \asymp \frac{k^2}{f(\mathbf{a},\mathbf{b})} \max\left(\frac{(\log n)^2}{\Delta_{\max}}, \frac{\log n}{\sqrt{\Delta_{\max}}}\right).$$

In the dense case where $p^{(m)}$'s and $q^{(m)}$'s do not grow with n, $\Delta_{\max} \asymp n$ and $r_{LMF} \lesssim \frac{k^2 \log n}{n^{1/2} f(\mathbf{a},\mathbf{b})}$. Hence consistent estimation is possible as long as $k = o(n^{1/4}\sqrt{f(\mathbf{a},\mathbf{b})/\log n})$.

Note that Theorem 3 requires $\Delta_{\max}$ to be at least $O(\log n)$, and hence the theorem does not hold for sparse constant degree case. While we do not have consistency for $O(\log n)$ degree density case either, in the slightly denser case when $\Delta_{\max} \asymp O((\log n)^3)$, we have consistency as follows. Since $r_{LMF} \lesssim \frac{k^2}{(\log n)^{1/2} f(\mathbf{a},\mathbf{b})}$, consistent recovery is possible as long as $k = o((\log n)^{1/4}\sqrt{f(\mathbf{a},\mathbf{b})})$.

The final result we prove provides an upper bound on the mis-clustering rate for consensus community detection using the usual single layer spectral clustering on the mean adjacency matrix.

Theorem 4. *Define $\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$ and let $\lambda^{\bar{\mathcal{A}}}$ denote the minimum non-zero eigenvalue of the mean population adjacency matrix $\bar{\mathcal{A}} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{A}^{(m)}$. Let r_{av} be the fraction of nodes misclustered by the spectral clustering algorithm applied to $\bar{A}$. If $M\Delta_{\max} > \frac{4}{9}\log(2n\epsilon)$, the conclusion of part (i) of Theorem 1 holds, and $\bar{B} = \frac{1}{M}\sum_{m=1}^{M} B^{(m)}$ is of full rank, then with probability at least $1 - \epsilon$,*

$$r_{av} \leq \frac{256 n_{\max} k \Delta_{\max} \log(2n/\epsilon)}{(\lambda^{\bar{\mathcal{A}}})^2 n M}.$$

Note the presence of $\lambda^{\bar{\mathcal{A}}}$ in the denominator of the bound implies that the bound depends on the eigen-gap of the mean adjacency matrix, which can be heavily influenced by denser layers. To prove this result, we employ a proof technique using Theorem 1, which is different from [HXA15] and allows us to characterize the dependence of the misclustering rate on the

growth rates of various MLSBM parameters. While the concentration result in Frobenius norm of [HXA15] would imply consistent community detection through spectral clustering in mean adjacency matrix for fixed k as long as $n = o(M^{1/2})$, our technique yields a bound on the mis-clustering rate with direct dependence on the number of communities, sparsity, signal to noise ratio along with n and M. We will once again interpret the bound under the four parameter MLSBM. First we have the following lemma on $\lambda^{\bar{\mathcal{A}}}$.

Lemma 3. *For the four parameter MLSBM,* $\lambda^{\bar{\mathcal{A}}} = s\frac{1}{M}\sum_m (p^{(m)} - q^{(m)})$.

Similar to previous cases, writing the result in terms of $\Delta_{\max}, a^{(m)}$, and $b^{(m)}$ we have $\lambda^{\bar{\mathcal{A}}} = \frac{\Delta_{\max}}{k}\frac{1}{M}\sum_m (a^{(m)} - b^{(m)})$. Define $g(\mathbf{a}, \mathbf{b}) = (\frac{1}{M}\sum_m (a^{(m)} - b^{(m)}))^2$. Then from Theorem 4 we have with probability at least $1 - \epsilon$,

$$r_{av} \lesssim \frac{\frac{n}{k} k\Delta_{\max}\log(2n/\epsilon)}{(\frac{\Delta_{\max}}{k})^2 g(\mathbf{a},\mathbf{b})nM} \asymp \frac{k^2}{M\Delta_{\max} g(\mathbf{a},\mathbf{b})/\log(2n/\epsilon)}.$$

This implies that $r_{av} \to 0$ as long as $k = o(\sqrt{M\Delta_{\max} g(\mathbf{a},\mathbf{b})/\log(2n/\epsilon)})$, and we have consistent community detection. Note that this growth rate on k is improved by a factor of $\sqrt{M}$ over the rate for single layer case in [QR13] and [LR14], as we would have anticipated (The $\log n$ term does not appear in [LR14] due to tighter bound on $\|A - \mathcal{A}\|_2$). We also note that the denominator in the rate for r_{av} contains the term $g(\mathbf{a}, \mathbf{b}) = (\frac{1}{M}\sum_m (a^{(m)} - b^{(m)}))^2$ instead of $f(\mathbf{a}, \mathbf{b}) = \frac{1}{M}\sum_m (a^{(m)} - b^{(m)})^2$, which appeared earlier in the rates of OLMF and co-regularized spectral clustering. From Jensen's inequality,

$$g(\mathbf{a}, \mathbf{b}) = (\frac{1}{M}\sum_m (a^{(m)} - b^{(m)}))^2 \leq \frac{1}{M}\sum_m (a^{(m)} - b^{(m)})^2 = f(\mathbf{a}, \mathbf{b}),$$

with equality holding if and only if all the $(a^{(m)} - b^{(m)})$'s are equal. Hence equality holds if the layers are of similar signal quality, and otherwise $f(\mathbf{a}, \mathbf{b})$ is larger than $g(\mathbf{a}, \mathbf{b})$. Hence the goodness of the rate for spectral clustering in mean adjacency matrix depends on if the aggregate of the layers has good signal quality or not. In the situation where some of the layers in the multi-layer network contain heterophilic clusters while the others contain homophilic clusters, then $a^{(m)} - b^{(m)}$ is negative in some layers and positive in other layers. In that case $\lambda^{\bar{\mathcal{A}}}$ could be very small and performance guarantee on spectral clustering of mean adjacency matrix become poor. These conclusions are in line with previous conclusions from minimax rates and phase transitions of consistency thresholds in [PC16a].

In the dense regime where the vectors $\mathbf{p}$ and $\mathbf{q}$ do not change with n, we have the mis-clustering rate in spectral clustering in mean adjacency matrix is bounded by $r_{av} \lesssim \frac{k^2}{nMg(\mathbf{a},\mathbf{b})/\log(2n/\epsilon)}$. In the semi-sparse regime where both $p^{(m)}$ and $q^{(m)}$ are of the order of $\frac{\log n}{n}$ for all m, we have $\Delta_{\max} \asymp \log n$ and, $r_{av} \lesssim \frac{k^2}{Mg(\mathbf{a},\mathbf{b})}$. Finally, in the sparse constant degree regime where both $p^{(m)}$ and $q^{(m)}$ are of the order of $1/n$ for all m, we have $r_{av} \lesssim \frac{k^2}{Mg(\mathbf{a},\mathbf{b})/\log(2n/\epsilon)}$.

5 Simulation studies

In this section, we numerically compare the performance of the following methods through a principled simulation study: spectral clustering on mean adjacency matrix (Mean adj.),

OLMF, co-regularized spectral clustering (Coreg spec), spectral clustering on aggregate spectral kernel (SpecK) and the module allegiance matrix (Module alleg.). We initialize the OLMF algorithm with P being the community assignment matrix from a randomly chosen layer and $\Lambda^{(m)}$ being the matrix containing the top k eigenvalues of $A^{(m)}$ in the diagonal. For the co-regularized spectral clustering algorithm we choose $\gamma^{(m)}$ as $4 \max \|A^{(m)}\|_2$ for all m, since the theoretical results have indicated that $\gamma^{(m)}$ should be larger than $\|A^{(m)}\|_2$ for each m.

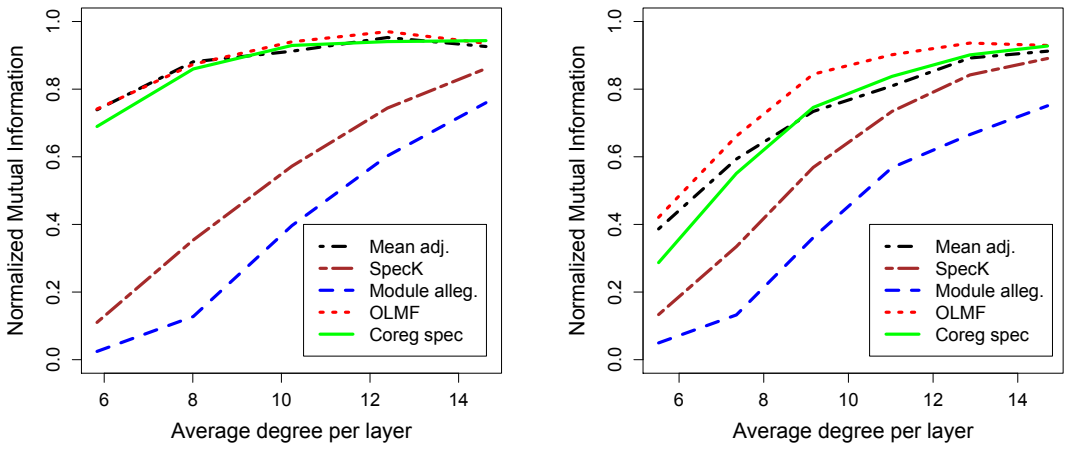
For the first three simulations, we simulate networks from the MLSBM with the number of nodes $n = 600$ and the number of layers $M = 5$, under three different scenarios on the connection probability matrices of different layers. The performances of the methods are evaluated with increasing average degree of the multi-layer network since we would expect any reasonable method to perform better as the network gets denser. The number of communities is fixed at 3 and we assume it to be known in advance. The fourth simulation involves generating networks from MLSBM with varying number of layers and testing the performance of the methods with increasing number of layers. The fifth and final simulation considers the scenario where the multi-layer network contains layers with both heterophilic and homophilic communities.

The evaluation criterion is the normalized mutual information (NMI) with the ground truth community assignments which generate the network. The NMI is an information theoretic measure of similarity between two vectors of community assignments, with 1 indicating a perfect match and 0 indicating the vectors are random with respect to each other. The first three experiments are replicated 40 times while the last two experiments are repeated 100 times, and the average performance across the repetitions is reported.

The data are generated according to MLSBM as defined in (3.1) in the following fashion. The community vector for each node is generated according to multinomial distribution with equal probability of being in any of the 3 clusters. The block model matrices in different layers are generated by the following scheme. Let δ be the vector of k diagonal elements and ϵ be the vector of $k^2 - k$ off-diagonal elements. We generate half of the elements of the ϵ vector from a uniform distribution $U(a, b)$ within a short range $[a, b]$ and the other half is a replication of the first half such that the matrix is symmetric. The elements of δ are generated from $U(\rho a, \rho b)$, where ρ is the parameter that controls the signal to noise ratio (SNR). We call an SNR of $2-3$ as "strong" signal and an SNR which is only slightly greater than 1 as "weak" signal.

Strong signals

In the first simulation from MLSBM, we make all the layers contain generally strong signals, but the exact SNR is randomly varied slightly so as to have some variations in signal quality across the layers. The performance of various methods under consideration is presented in Figure 1(a). Note that the layers are sparse at an average initially which is evident from the low average degree per layer: an average degree of 6 in a layer of 600 nodes, which is about 1% degree density. The layers then become denser gradually and reach about 2.5% degree density per layer. The performance of all the methods generally increases with increasing average degree. We note that spectral clustering on mean adjacency matrix, OLMF and co-regularized spectral clustering perform similarly throughout the range of the simula-



(a) (b)

Figure 1: Performance of various methods with increasing average degree of nodes for data generated from MLSBM with 600 nodes, 5 layers and 3 communities. (a) All layers have strong signals with some variations; (b) the layers are mixed in terms of signal quality.

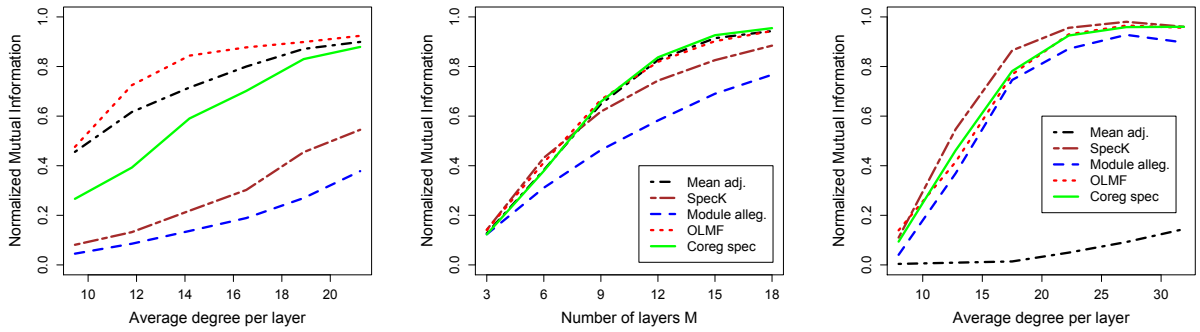
tion. The aggregate of spectral kernel and module allegiance matrix method substantially underperform, especially in sparse multi-layer networks.

Mixed signals

In this simulation, the component layers are mixed in signal quality in the following manner. We have three layers with strong signals and two layers with weak signals. This scenario is very useful to test the robustness of methods against possible variation or absence of community patterns in some of the layers. The results are presented in Figure 1(b). The OLMF method performs the best over the entire range of values of average degree, followed by co-regularized spectral clustering and spectral clustering of mean adjacency matrix. The aggregate of spectral kernel and module allegiance matrix methods once again perform poorly when the average density in the layers is low, but recover subsequently as the layers become denser. The spectral kernel method performs better than the module allegiance matrix method in both the strong signals and mixed signals scenarios.

Complementary information

The third scenario considers the so-called "complementary" principle of multiple views in multi-view learning [LWGH13]. In our case, this is equivalent to the following: none of the layers alone is sufficient to describe the community structure properly, but the layers can complement each other and together describe the community patterns. For our simulation, we generate data from MLSBM with 600 nodes, 5 layers and 3 communities with the following setting. In each of the first 3 layers, two of the communities are difficult to distinguish



(a) (b)

Figure 2: Performance of various methods with (a) increasing average degree of nodes for data generated from MLSBM with 600 nodes, 5 layers and 3 communities, (b) increasing number of layers with 300 nodes and 6 communities, (c) increasing average degree of nodes with 600 nodes, 5 layers and 3 communities, where 3 layers contain homophilic clusters and the other 2 contain heterophilic clusters. The labels in (b) and (c) are shared for all figures.

from noise while the third community has a SNR of 3. The fourth layer has two of the communities with high SNR and the fifth layer has the same two communities with low SNR. The performance of the competing methods are presented in Figure 2(a). We observed the same trend as before that both aggregate of spectral kernel and module allegiance matrix method perform poorly when the layers are sparse. The OLMF appears to have a clear advantage in this scenario over both co-regularized spectral clustering and spectral clustering of the mean adjacency matrix.

Increasing number of layers

This simulation setup tests the abilities of the methods to recover the community structure with a small fixed number of nodes, but increasing number of layers (and consequently more data). However, as is the case with many real world multi-layer networks, not all of the layers are strongly informative of the community structure. We fix n at 300, k at 6 and increase M from 3 to 18 in steps of 3. At every step, we add 3 layers to the multi-layer network, two of which have weak signal quality, while the third one has a strong signal. The performance of the competing methods in this simulation with 100 repetitions is depicted in Figure 2(b). We observe that the accuracy of consensus community detection in all the methods generally increases with increasing number of layers. As with the previous scenarios, we observe that OLMF, co-regularized spectral clustering, and spectral clustering of the mean adjacency matrix have more improvement in performance as compared to aggregate of spectral kernel and module allegiance matrix methods.

Heterophilic clusters

Finally, we consider the scenario where some layers contain homophilic clusters while others contain heterophilic clusters. The layers with heterophilic clusters have less density within the blocks as opposed to inter-block densities. As indicated in our theoretical results, we expect the spectral clustering in mean adjacency matrix to perform poorly in this setup. We need to modify some of the methods slightly for this scenario. For aggregate spectral kernel and module allegiance methods, we choose the eigenvectors corresponding to the top k eigenvalues in *absolute value* to form the $\hat{U}^{(m)}$ matrix in each layer. For co-regularized spectral clustering, we update the $\hat{U}^{(m)}$ matrix during the alternating eigen-decomposition by selecting the vectors corresponding to the top k eigenvalues in *absolute value*. The mean adjacency matrix and OLMF methods do not require any change to be made, however, we make the following optional modifications. For mean adjacency matrix during the eigen-decomposition, we choose eigenvectors corresponding to the top k eigenvalues in *absolute value* of the mean adjacency matrix, while for OLMF we only change the initialization of $\Lambda^{(m)}$ matrices to include the k largest eigenvalues in *absolute value* as its diagonal. We fix n at 600, k at 3, M at 5, and increase the average degree per layer from 8 to 32 (from about 1% to 4.5% in degree density). We make 3 of the 5 layers contain homophilic clusters by setting the ρ parameter (SNR) at 3, while we make the other two layers contain heterophilic clusters by setting $\rho = 1/3$ so that the elements of δ are smaller than that of ϵ. The results are presented in Figure 2(c). We observe that the performance of spectral clustering in mean adjacency matrix completely breaks down and is susbtantially worse than the competing methods in this scenario. The other four methods behave similarly with increasing accuracy of community detection with increasing degree density. We also note that the aggregate of spectral kernel method performs slightly better compared to the remaining three methods throughout the range of the simulation.

Discussion on the simulation results

Our simulations clearly show that in sparse networks the intermediate fusion of information based methods, OLMF and co-regularized spectral clustering, perform better than late fusion methods, aggregate of spectral kernel and module allegiance matrix method. We believe sharing information across layers while computing individual layer wise spectral embeddings increases the accuracy in each of them, and hence the centroid is a more effective combination than aggregate spectral kernel or module allegiance matrix type of combination. We also observe in our simulations that aggregate spectral kernel performs better than module allegiance matrix. We think the performance in module allegiance suffers because of additional noise introduced in discrete community assignments. The spectral clustering of mean adjacency matrix performs well in our simulations except the last scenario where the multi-layer network contains layers with both homophilic and heteophilic clusters, in which case its performance is extremely poor.

6 Conclusion

In this paper we have analyzed a number of previously proposed spectral and matrix factorization based techniques for multi-view clustering in terms of their asymptotic consistency properties for community detection in multi-layer graphs generated from the MLSBM. We have considered a high dimensional asymptotic framework where both the number of layers (M) and the number of nodes (n) of the multi-layer graph grow. We have shown that the spectral clustering on mean adjacency matrix and global optimum solutions of both co-regularized spectral clustering and OLMF enjoy consistency guarantees under some conditions on the number of communities k, the maximum expected degree of a layer $\Delta_{\max}$ and signal to noise ratios of the layers. The spectral clustering on mean adjacency matrix requires additional conditions for consistency that limits the allowable signal heterogeneity of the layers of the multi-layer network.

We have also compared five methods in terms of finite sample performance under data generated from the MLSBM through a simulation study. We found both the co-regularized spectral clustering and OLMF to be robust under varied scenarios. We also note from the simulations that widely popular methods where each layer is dealt separately and the results are fused at a later state, such as aggregating spectral kernels or module allegiance matrix, do not perform well in sparse networks when the individual layers do not contain sufficient information to recover the community structure efficiently. However, the OLMF and co-regularized spectral clustering perform well in those scenarios due to sharing information across layers while computing the community structure solution at each layer.

Appendix: Proofs

Proof of Proposition 1

Proof. Note that,

$$\begin{aligned}
\|\sin\Theta(\mathbb{U}^{(m)},\mathbb{U}^*)\|_F^2 &= \frac{1}{2}\|U^*U^{*T}-U^{(m)}U^{(m)T}\|_F^2 \qquad \text{[Theorem 1.5.5 of [SS90]]}\\
&= \frac{1}{2}\{\text{tr}(U^*U^{*T}U^*U^{*T})+\text{tr}(U^{(m)}U^{(m)T}U^{(m)}U^{(m)T})-2\ \text{tr}(U^{*T}U^{(m)}U^{(m)T}U^*)\}\\
&= k-\text{tr}(U^{*T}U^{(m)}U^{(m)T}U^*).
\end{aligned}$$

Rearranging the terms we have the proposition. □

Proof of Lemma 1

Proof. To prove part (i) of the lemma, note that for the case of OLMF, it is evident that $[P = Z(Z^TZ)^{-1/2}, \Lambda^{(m)} = (Z^TZ)^{1/2}B^{(m)}(Z^TZ)^{1/2}]$ is a solution to the optimization problem on the population adjacency tensor $\mathscr{A}$. Indeed the value of the minimization objective function in

(2.1) is 0, which is its minimum possible value and $P^TP = I$. This solution however, is unique only up to an orthogonal matrix, since for any orthogonal matrix $O \in \mathbb{R}^{k\times k}$, PO and $O^T\Lambda^{(m)}O$ for all m is also a solution. Moreover, since $Q^{-1/2} = (Z^TZ)^{-1/2}$ is a diagonal matrix with positive elements and hence invertible, we have $Z_iQ^{-1/2} = Z_jQ^{-1/2} \iff Z_i = Z_j$.

For co-regularized spectral clustering, note that for each m, maximizing $\mathrm{tr}(U^{(m)T}\mathcal{A}^{(m)}U^{(m)})$ under the given constraints is the usual spectral clustering association cut objective function and hence is maximized by the matrix containing the top k eigenvectors of $\mathcal{A}^{(m)}$ [NJW02, KCS11]. In this case the matrix is $\bar{U}^{(m)} = Z\mu^{(m)}$, where $\mu^{(m)} = (Z^TZ)^{-1/2}V^{(m)}$ for some orthogonal matrix $V^{(m)}$ [RCY11]. Moreover by Proposition 1, it is clear that the second term has an absolute maximum value of k irrespective of the first term. This maximum value is also attained with the same $\bar{U}^{(m)}$'s along with $\bar{U}^* = Z(Z^TZ)^{-1/2}O$, where O is an orthogonal matrix. This is so because $\mathrm{tr}(\bar{U}^{*T}\bar{U}^{(m)}\bar{U}^{(m)T}\bar{U}^*) = \mathrm{tr}(O(Z^TZ)^{-1/2}Z^TZ(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1/2}O) = k$ for all m. Hence, $\bar{U}^{(m)} = Z(Z^TZ)^{-1/2}V^{(m)}$ for $m = 1,\ldots,M$ and $\bar{U}^* = Z(Z^TZ)^{-1/2}O$ are solutions up to the ambiguity of orthogonal matrices to the optimization problem. Since the matrix $Q = (Z^TZ)^{-1/2}O$ is invertible, we have $Z_iQ = Z_jQ \iff Z_i = Z_j$. This gives us part (ii) of the lemma.

Next we prove part (iii) of the lemma concerning spectral clustering applied to the mean population adjacency matrix. Note that the population version is

$$\bar{\mathcal{A}} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{A}^{(m)} = \frac{1}{M}\sum_{m=1}^{M}ZB^{(m)}Z^T = Z\left(\frac{1}{M}\sum_{m=1}^{M}B^{(m)}\right)Z^T = Z\bar{B}Z^T,$$

with $\bar{B} \in \mathbb{R}^{k\times k}$ and is full rank as mentioned in the statement of the lemma. Then by Lemma 3.1 of [RCY11], there exists an invertible matrix $\mu \in \mathbb{R}^{k\times k}$ such that columns of $Z\mu$ are the eigenvectors of $\bar{\mathcal{A}}$, corresponding to the non-zero eigenvalues and $Z_i\mu = Z_j\mu \iff Z_i = Z_j$.

Finally, for part (iv) note that if spectral clustering on aggregate spectral kernel is applied to the population adjacency tensor, each of the spectral kernels would be $Z\mu^{(m)}\mu^{(m)T}Z^T = Z(Z^TZ)^{-1}Z^T$. Note that the spectral kernels do not depend on m. Clearly, $Z(Z^TZ)^{-1/2}O$ for some orthogonal matrix O is the matrix containing eigenvectors corresponding to top k eigenvalues of $\bar{K}$. Denoting $(Z^TZ)^{-1/2}O$ as Q we note that $Q \in \mathbb{R}^{k\times k}$ is an invertible matrix and hence $Z_iQ = Z_jQ \iff Z_i = Z_j$.

□

Proof of Theorem 1

Proof. Our main tool to prove the result (i) will be the matrix Bernstein inequality in Theorem 5 of [CR11], which we reproduce below.

Proposition 2. *([CR11]) Let $X_1,\ldots,X_p$ be independent random $n\times n$ Hermitian matrices. Moreover, assume that $\|X_i - E(X_i)\|_2 \le L$ for all i, and put $v^2 = \|\sum_i var(X_i)\|_2$. Let $X = \sum_i X_i$. Then for any $a > 0$,*

$$P(\|X - E(X)\|_2 \ge a) \le 2n\exp\left(-\frac{a^2}{2v^2 + 2La/3}\right).$$

Let E^{ij} be a (deterministic) matrix with 1 in the (i,j)th and (j,i)th position and 0 everywhere else. Let

$$X_{ij}^{(m)} = (A_{ij}^{(m)} - \mathcal{A}_{ij}^{(m)})E^{ij}.$$

Hence $X_{ij}^{(m)}$ is an $n \times n$ symmetric matrix for all m, i, j. Moreover, since each of $A_{ij}^{(m)}$ is an independent random variable for all m, i, j, the matrices $X_{ij}^{(m)}$ are also independent. Now $\sum_m (A^{(m)} - \mathcal{A}^{(m)}) = \sum_m \sum_{ij} X_{ij}^{(m)}$. Then following the arguments in [CR11], we have

$$\|X_{ij}^{(m)}\|_2 \leq 1, \quad \forall\, m, i, j,$$

and

$$\begin{aligned} v^2 = \|\sum_m \sum_{ij} var(X_{ij}^{(m)})\|_2 &= \|\sum_m \sum_{ij} E[(X_{ij}^{(m)})^2]\|_2 \\ &= \|\sum_m \sum_{ij} (\mathcal{A}_{ij}^{(m)} - (\mathcal{A}_{ij}^{(m)})^2)E^{ii}\|_2 \\ &= \max_i \left(\sum_m \sum_j (\mathcal{A}_{ij}^{(m)} - (\mathcal{A}_{ij}^{(m)})^2) \right) \\ &\leq \sum_m \max_i \sum_j \mathcal{A}_{ij}^{(m)} = \sum_m \Delta_m \leq M\Delta_{\max}. \end{aligned}$$

The third line follows since $\sum_m \sum_j (\mathcal{A}_{ij}^{(m)} - (\mathcal{A}_{ij}^{(m)})^2)E^{ii}$ is a diagonal matrix and hence the eigenvalues are the same as the elements.

Now we can apply the matrix concentration result in Proposition 2 to the set of independent $n \times n$ Hermitian matrices $X_{ij}^{(m)}$. Take $a = \sqrt{4M \log(2n/\epsilon)\Delta_{\max}}$. The assumption $M\Delta_{\max} > \frac{4}{9}\log(2n/\epsilon)$ implies that $a < 3M\Delta_{\max}$. Then applying Proposition 2 we have,

$$\begin{aligned} P(\|\sum_m (A^{(m)} - \mathcal{A}^{(m)})\|_2 \geq a) &\leq 2n \exp\left(-\frac{4M\Delta_{\max}\log(2n/\epsilon)}{2M\Delta_{\max} + 2a/3} \right) \\ &\leq 2n \exp\left(-\frac{4M\Delta_{\max}\log(2n/\epsilon)}{4M\Delta_{\max}} \right) \\ &\leq \epsilon. \end{aligned}$$

To prove part (ii) we use the following matrix Chernoff bound [Tro12].

Proposition 3. *(Corollary 5.2 of [Tro12]) For a finite sequence $\{X_k\}$ of independent, random, self-adjoint, positive semi-definite matrices of common dimension n that satisfy $\lambda_{\max}(X_k) \leq R$ almost surely, let $\mu_{\max} := \lambda_{\max}(\sum_k E[X_k])$ be the maximum eigenvalue of the sum of expectations, then*

$$P\left\{ \lambda_{\max}\left(\sum_k X_k \right) \geq (1+\delta)\mu_{\max} \right\} \leq n \left[\frac{e^\delta}{(1+\delta)^{1+\delta}} \right]^{\mu_{\max}/R} \leq e^{\log n - \delta\mu_{\max}/3R}, \quad \delta \geq 1.$$

To apply this bound in our case, we first note that $\{(A^{(m)} - \mathcal{A}^{(m)})^2\}$ is a sequence of independent, random, self-adjoint (Hermitian) and positive semidefinite matrices. Now, for $i \neq j$, we have

$$E[(A^{(m)}A^{(m)})_{ij}] = E[\sum_k A^{(m)}_{ik}A^{(m)}_{kj}] = \sum_k \mathcal{A}^{(m)}_{ik}\mathcal{A}^{(m)}_{jk} = (\mathcal{A}^{(m)}\mathcal{A}^{(m)})_{ij}, \quad \forall m,$$

and $E[(A^{(m)}A^{(m)})_{ii}] = E[\sum_k A^{(m)2}_{ik}] = E[\sum_k A^{(m)}_{ik}] = \sum_k \mathcal{A}^{(m)}_{ik}$ for all m. Also, we have

$$\begin{aligned} E[\sum_m (A^{(m)} - \mathcal{A}^{(m)})^2] &= E[\sum_m \{A^{(m)}A^{(m)} - A^{(m)}\mathcal{A}^{(m)} - \mathcal{A}^{(m)}A^{(m)} + \mathcal{A}^{(m)}\mathcal{A}^{(m)}\}] \\ &= \sum_m E[A^{(m)}A^{(m)} - \mathcal{A}^{(m)}\mathcal{A}^{(m)}]. \end{aligned}$$

Hence the matrix $E[\sum_m (A^{(m)} - \mathcal{A}^{(m)})^2]$ has 0's in all its off diagonal elements and its n diagonal elements are $\sum_m \sum_k (\mathcal{A}^{(m)}_{ik} - (\mathcal{A}^{(m)}_{ik})^2)$. Then we have

$$\lambda_{\max}\left(E[\sum_m (A^{(m)} - \mathcal{A}^{(m)})^2]\right) = \max_{i=1,\ldots,n} \sum_m \sum_k (\mathcal{A}^{(m)}_{ik} - (\mathcal{A}^{(m)}_{ik})^2) \leq M\Delta_{\max}.$$

Hence we have $\mu_{\max} \leq M\Delta_{\max}$.

Next to find almost sure upper bound for $\lambda_{\max}((A^{(m)} - \mathcal{A}^{(m)})^2)$ we use Theorem 1 of [CR11] with the choice of $\epsilon = 2/n^2$. Since $\Delta_{\max} > \frac{4}{3}\log n$ by assumption, applying a union bound to the result of Theorem 1 of [CR11] we have with probability at least $1 - 2/n^2$,

$$\lambda_{\max}(A^{(m)} - \mathcal{A}^{(m)}) \leq \sqrt{4\Delta_{\max}\log(Mn^3)}$$

uniformly for all m. Next using Borel-Cantelli Lemma, we have almost surely,

$$\lambda_{\max}((A^{(m)} - \mathcal{A}^{(m)})^2) = (\lambda_{\max}(A^{(m)} - \mathcal{A}^{(m)}))^2 \leq 4\Delta_{\max}\log(Mn^3).$$

Then from Proposition 3 we have,

$$\begin{aligned} &P\{\lambda_{\max}(\sum_m (A^{(m)} - \mathcal{A}^{(m)})^2) \geq (\log n)^{2+\epsilon} M\Delta_{\max}\} \\ &\quad \leq \exp\left(\log n - \frac{((\log n)^{2+\epsilon} - 1)M\Delta_{\max}}{12\Delta_{\max}\log(Mn^3)}\right) \\ &\quad = \exp\left(\log n - \frac{((\log n)^{2+\epsilon} - 1)M}{12(\log M + 3\log n)}\right) \\ &\quad \leq \exp\left(\log n - \frac{(\log n)^{2+\epsilon} M}{24(\log M + 3\log n)}\right). \end{aligned}$$

If $M > n$, the last term becomes

$$\exp\left(\log n - \frac{(\log n)^{2+\epsilon} M}{24(\log M + 3\log n)}\right) \leq \exp\left(\log n - \frac{(\log n)^{2+\epsilon} M}{96\log M}\right) = o(1).$$

If $M \leq n$, the last term becomes

$$\exp\left(\log n - \frac{(\log n)^{2+\epsilon} M}{24(\log M + 3\log n)}\right) \leq \exp\left(\log n - (\log n)^{1+\epsilon} M/96\right) = o(1).$$

The result follows by noting that the sum of symmetric positive semi-definite matrices is a symmetric positive semi-definite matrix, and the spectral norm of a symmetric positive semi-definite matrix is the same as its largest eigenvalue.

□

Proof of Theorem 2

Proof. The proof consists of three steps.

1. The first step is to show that it is possible to recover the communities by maximizing the population version of the objective function.

2. In the second step we show that for any feasible set of solutions $[\mathbb{U}, U^*]$, the sample version of the objective function is "close" to the population version of the objective function provided γ_m's are large.

3. Finally, in the last step we will relate the misclustering rate with the difference between $\hat{U}^*$ and $\bar{U}^*$, and then relate this difference with the difference between the maximized sample and the population versions of the objective function.

The result of Lemma 1 shows that $\bar{U}^{(m)} = Z\mu^{(m)}$, $m = 1, \ldots, M$, $\bar{U}^* = Z(Z^TZ)^{-1/2}O$ is the solution up to the ambiguity of (several different) orthogonal matrices obtained by optimizing the population version of the objective function $F(\mathscr{A}, \mathbb{U}, U^*)$. We call the tensor containing the layer-wise low rank matrices, $\bar{U}^{(m)}$, as $\bar{\mathbb{U}}$. Note that $\bar{U}^{(m)} = \bar{U}^* V^{(m)}$, for some orhtogonal matrix $V^{(m)}$. Lemma 1 further shows that the true community assignments Z can be recovered by applying k-means algorithm to the columns of $\bar{U}^*$.

Let $[\mathbb{U}, U^*]$ be a feasible set of solutions. Then we have with probability at least $1 - \epsilon/2$,

$$\begin{aligned}
|\sum_m \mathrm{tr}(U^{*T}(\mathcal{A}^{(m)} - A^{(m)})U^*)| &\leq k\|\sum_m U^{*T}(\mathcal{A}^{(m)} - A^{(m)})U^*\|_2 \\
&\leq k\|U^*\|_2^2\|\sum_m (\mathcal{A}^{(m)} - A^{(m)})\|_2 \\
&\leq k\sqrt{4M\Delta_{\max}\log(4n/\epsilon)},
\end{aligned}$$

where the first inequality is true since $U^{*T}(\mathcal{A}^{(m)} - A^{(m)})U^*$ is a $k \times k$ matrix, the second line follows since $\|AB\|_2 \leq \|A\|_2\|B\|_2$ for any two matrices A and B, while the third inequality follows from Theorem 1.

We define two square symmetric $k \times k$ matrices, $\bar{S}^{(m)} = \bar{U}^{*T}\mathcal{A}^{(m)}\bar{U}^*$ and $\hat{S}^{(m)} = \hat{U}^{*T}\mathcal{A}^{(m)}\hat{U}^*$. Since $\bar{U}^* V^{(m)}$ is the matrix of eigenvectors corresponding to the non-zero eigenvalues of $\mathcal{A}^{(m)}$, we also have the eigenvalue decomposition, $\mathcal{A}^{(m)} = \bar{U}^*\bar{S}^{(m)}\bar{U}^{*T}$. We define a new quantity

$\mathcal{A}_1^{(m)} = \hat{U}^* \hat{S}^{(m)} \hat{U}^{*T} = \hat{U}^* \hat{U}^{*T} \mathcal{A}^{(m)} \hat{U}^* \hat{U}^{*T}$. Then $\hat{U}^*$ is an invariant subspace of $\mathcal{A}_1^{(m)}$ [PC16c]. A couple of lines of algebra show that (see [PC16c] for a proof)

$$\|\mathcal{A}^{(m)} - \mathcal{A}_1^{(m)}\|_F^2 = \|\bar{S}^{(m)}\|_F^2 - \|\hat{S}^{(m)}\|_F^2. \tag{6.1}$$

For a $k \times k$ matrix B, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ be the eigenvalues of B sorted in decreasing order. Since the eigenvalues of $\bar{S}^{(m)}$ are the non-zero eigenvalues of $\mathcal{A}^{(m)}$, we also have the following eigenvalue interlacing property (Theorem 2.1 of [Hae95]),

$$\lambda_i(\bar{S}^{(m)}) = \lambda_i(\bar{U}^{*T} \mathcal{A}^{(m)} \bar{U}^*) \geq \lambda_i(\hat{U}^{*T} \mathcal{A}^{(m)} \hat{U}^*) = \lambda_i(\hat{S}^{(m)}), \tag{6.2}$$

for all $1 \leq i \leq k$. Then we have the following bound on the Frobenius norm of difference between $\mathcal{A}^{(m)}$ and $\mathcal{A}_1^{(m)}$ in terms of the traces of $\bar{S}^{(m)}$ and $\hat{S}^{(m)}$:

$$\begin{aligned}
\|\mathcal{A}^{(m)} - \mathcal{A}_1^{(m)}\|_F^2 &= \sum_{i=1}^{k} \lambda_i^2(\bar{S}^{(m)}) - \sum_{i=1}^{k} \lambda_i^2(\hat{S}^{(m)}) \qquad \text{[Equation (6.1)]} \\
&\leq \sum_{i=1}^{k} |\lambda_i(\bar{S}^{(m)}) - \lambda_i(\hat{S}^{(m)})||\lambda_i(\bar{S}^{(m)}) + \lambda_i(\hat{S}^{(m)})| \\
&\leq \sum_{i=1}^{k} |\lambda_i(\bar{S}^{(m)}) - \lambda_i(\hat{S}^{(m)})| \cdot 2|\lambda_i(\bar{S}^{(m)})| \qquad \text{[Property (6.2)]} \\
&\leq 2|\lambda_1(\bar{S}^{(m)})| \sum_{i=1}^{k} (\lambda_i(\bar{S}^{(m)}) - \lambda_i(\hat{S}^{(m)})) \qquad \text{[Property (6.2)]} \\
&\leq 2\Delta_{\max}(\mathrm{tr}(\bar{S}^{(m)}) - \mathrm{tr}(\hat{S}^{(m)})) \qquad \text{[Since } \lambda_1(\bar{S}^{(m)}) \leq \Delta_{\max}\text{]}.
\end{aligned} \tag{6.3}$$

Finally we use this result to prove the following bound which then leads to a bound on misclustering rate:

$$\begin{aligned}
&\frac{1}{2\Delta_{\max}} \sum_m (\lambda^{(m)})^2 \|\hat{U}^* - \bar{U}^* O\|_F^2 \\
&\quad \leq \frac{1}{2\Delta_{\max}} \sum_m \|\mathcal{A}^{(m)} - \mathcal{A}_1^{(m)}\|_F^2 \quad \text{[Davis-Kahan Theorem in [SS90]]} \\
&\quad \leq \sum_m \mathrm{tr}(\bar{U}^{*T} \mathcal{A}^{(m)} \bar{U}^* - \hat{U}^{*T} \mathcal{A}^{(m)} \hat{U}^*) \quad \text{[Equation (6.3)]} \\
&\quad \leq \sum_m \{\mathrm{tr}(\bar{U}^{*T} \mathcal{A}^{(m)} \bar{U}^* - \hat{U}^{*T} \mathcal{A}^{(m)} \hat{U}^*) \\
&\qquad + \mathrm{tr}(\hat{U}^{(m)T} A^{(m)} \hat{U}^{(m)} - \bar{U}^{(m)T} A^{(m)} \bar{U}^{(m)}) + \gamma_m (k - \frac{1}{2} \|\hat{U}^{(m)} \hat{U}^{(m)T} - \hat{U}^* \hat{U}^{*T}\|_F^2 - k)\} \\
&\quad = \sum_m \{\mathrm{tr}(\bar{U}^{*T} \mathcal{A}^{(m)} \bar{U}^* - \bar{U}^{*T} A^{(m)} \bar{U}^*) + \mathrm{tr}(\hat{U}^{*T} A^{(m)} \hat{U}^* - \hat{U}^{*T} \mathcal{A}^{(m)} \hat{U}^*)\} \\
&\qquad + \sum_m \{\mathrm{tr}(\hat{U}^{(m)T} A^{(m)} \hat{U}^{(m)} - \hat{U}^{*T} A^{(m)} \hat{U}^*) - \gamma_m \frac{1}{2} \|\hat{U}^{(m)} \hat{U}^{(m)T} - \hat{U}^* \hat{U}^{*T}\|_F^2\}
\end{aligned}$$

$$\leq 2k\sqrt{4M\Delta_{\max}\log(4N/\epsilon)} + \sum_m \{\sum_j |\lambda_j(\hat{U}^{(m)T}A^{(m)}\hat{U}^{(m)}) - \lambda_j(\hat{U}^{*T}A^{(m)}\hat{U}^*)|$$
$$- \gamma_m \|\sin\Theta(\hat{U}^{(m)}, \hat{U}^*)\|_F^2\}$$
$$\leq 2k\sqrt{4M\Delta_{\max}\log(4N/\epsilon)} + \sum_m (\|A^{(m)}\|_2 \|\sin\Theta(\hat{U}^{(m)}, \hat{U}^*)\|_\Sigma - \gamma_m \|\sin\Theta(\hat{U}^{(m)}, \hat{U}^*)\|_F^2),$$

with probability at least $1-\epsilon$. The third inequality follows from the fact that $F(\mathbb{A}, \hat{\mathbb{U}}, \hat{U}^*) \geq F(\mathbb{A}, \bar{\mathbb{U}}, \bar{U}^*)$ and Proposition 1. The first term in the fourth equality has used the fact: since $V^{(m)}$'s are orthogonal matrices, $\mathrm{tr}(V^{(m)T}\bar{U}^{*T}A^{(m)}\bar{U}^*V^{(m)}) = \mathrm{tr}(\bar{U}^{*T}A^{(m)}\bar{U}^*)$. The last line follows from Theorem 2.1 of [KA10] which states that $\sum_j |\lambda_j(\hat{U}^{(m)T}A^{(m)}\hat{U}^{(m)}) - \lambda_j(\hat{U}^{*T}A^{(m)}\hat{U}^*)| \leq \sum_j \|A^{(m)}\|_2 \sin\theta_j(\hat{U}^{(m)}, \hat{U}^*)$ and Proposition 1. Since $\sin\Theta$ is a diagonal matrix with non-negative elements, we represent $\sum_j \sin\theta_j$ as $\|\sin\Theta\|_\Sigma$. Hence if we choose γ_m to be large enough such that $\gamma_m \geq \|A^{(m)}\|_2 \frac{\|\sin\Theta(\hat{U}^{(m)},\hat{U}^*)\|_\Sigma}{\|\sin\Theta(\hat{U}^{(m)},\hat{U}^*)\|_F^2}$, then we have with probability at least $1-\epsilon$,

$$\|\hat{U}^* - \bar{U}^*O\|_F^2 \leq \frac{4k\Delta_{\max}}{\sum_m (\lambda^{(m)})^2}\sqrt{4M\Delta_{\max}\log(2n/\epsilon)}.$$

In addition, as n goes to infinity, each of the $\hat{U}^{(m)}$ is closer to being an invariant subspace of $A^{(m)}$, and then the bound improves to $\sum_j |\lambda_j(\hat{U}^{(m)T}A^{(m)}\hat{U}^{(m)}) - \lambda_j(\hat{U}^{*T}A^{(m)}\hat{U}^*)| \leq \sum_j \|A^{(m)}\|_2 \sin^2\theta_j(\hat{U}^{(m)}, \hat{U}^*)$ [KA10]. Hence asymptotically the condition on $\gamma^{(m)}$ is required to be simply larger than a constant times $\|A^{(m)}\|_2$. The bound on misclustering rate follows:

$$r_{coreg} \leq \frac{8n_{\max}}{n}\|\hat{U} - \bar{U}^*O\|_F^2 \leq \frac{64n_{\max}k(\Delta_{\max})^{3/2}}{\frac{1}{M}\sum_m (\lambda^{(m)})^2 n}\sqrt{\frac{\log(2n/\epsilon)}{M}},$$

with probability at least $1-\epsilon$. □

Proof of Lemma 2

Proof. We note that for the four parameters MLSBM,

$$\mathcal{A}^{(m)} = Z(Z^TZ)^{-1/2}(Z^TZ)^{1/2}B^{(m)}(Z^TZ)^{1/2}(Z^TZ)^{-1/2}Z^T = HS^{(m)}H^T,$$

where $H = Z(Z^TZ)^{-1/2}$ and $S^{(m)} = (Z^TZ)^{1/2}B^{(m)}(Z^TZ)^{1/2}$. Clearly $\mathcal{A}^{(m)}H = HS^{(m)}$, and hence columns of H span a k dimensional invariant subspace of $\mathcal{A}^{(m)}$. Moreover since $\mathrm{rank}(\mathcal{A}^{(m)}) = \mathrm{rank}(S^{(m)})$, all non-zero eigenvalues of $\mathcal{A}^{(m)}$ are also eigenvalues of $S^{(m)}$. This implies the smallest non-zero eigenvalue of $\mathcal{A}^{(m)}$ is also the smallest eigenvalue of $S^{(m)}$. To determine the smallest eigenvalue we proceed as in [RCY11]. Note that we have $S^{(m)} = \sqrt{s}I_kB^{(m)}\sqrt{s}I_k = sB^{(m)}$, and $B^{(m)}$ can be written as $B^{(m)} = (p^{(m)}-q^{(m)})I_k + q^{(m)}1_k1_k^T$. Then 1_k is an eigenvector of $S^{(m)}$ since $sB^{(m)}1_k = (s(p^{(m)} - q^{(m)}) + sq^{(m)}k)1_k = (s(p^{(m)} - q^{(m)}) + nq^{(m)})1_k$. Let u be another eigenvector of $sB^{(m)}$. Then $\|u\|_2 = 1$ and $u^T1_k = 0$. Hence we have $sB^{(m)}u = s(p^{(m)} - q^{(m)})u$. This implies all the remaining eigenvalues of $sB^{(m)}$ are $s(p^{(m)} - q^{(m)})$. Since $nq^{(m)} > 0$, we conclude the smallest eigenvalue of $sB^{(m)}$ is $s(p^{(m)} - q^{(m)})$. This is also the smallest non-zero eigenvalue of $\mathcal{A}^{(m)}$. □

Proof of Theorem 3

Proof. Similar to the proof of Theorem 2, the proof for this theorem also consists of three steps. The first step was addressed in Lemma 1, where it was shown that true community labels can be recovered from the solution $\bar{P}$ of the objective function applied to the population adjacency tensor. Next we show the second step. For any feasible solution of P we have,

$$\begin{aligned}
|F(\mathscr{A}, P) - F(\mathbb{A}, P)| &= |\sum_m \{\|P^T\mathcal{A}^{(m)}P\|_F^2 - \|P^TA^{(m)}P\|_F^2\}| \\
&= \sum_m \{(\|P^TA^{(m)}P\|_F - \|P^T\mathcal{A}^{(m)}P\|_F)^2 \\
&\quad + |(\|P^TA^{(m)}P\|_F - \|P^T\mathcal{A}^{(m)}P\|_F) \cdot 2\|P^T\mathcal{A}^{(m)}P\|_F|\} \\
&= \sum_m \{(x^{(m)} - y^{(m)})^2 + |2y^{(m)}(x^{(m)} - y^{(m)})|\} \\
&= \sum_m \{(x^{(m)} - y^{(m)})^2 + 2|y^{(m)}|\,|(x^{(m)} - y^{(m)})|\}
\end{aligned}$$

where $x^{(m)} = \|P^TA^{(m)}P\|_F$ and $y^{(m)} = \|P^T\mathcal{A}^{(m)}P\|_F$.

First, for the $2|y^{(m)}|$ term we have,

$$\begin{aligned}
|y^{(m)}| &= \|P^T\mathcal{A}^{(m)}P\|_F \\
&\le \sqrt{k}\|P^T\mathcal{A}^{(m)}P\|_2 \quad [\text{equivalence of norm since } P^T\mathcal{A}^{(m)}P \text{ is } k \times k] \\
&\le \sqrt{k}\|P\|_2^2\|\mathcal{A}\|_2 \quad [\text{property of spectral norm, } \|AB\|_2 \le \|A\|_2\|B\|_2] \\
&\le \sqrt{k}\Delta_{\max} \quad [\text{since } \|\mathcal{A}^{(m)}\| \le \Delta_{\max}].
\end{aligned}$$

Now, since $\|A\|_F - \|B\|_F \le \|A - B\|_F$, we have

$$\sum_m 2|y^{(m)}||(x^{(m)} - y^{(m)})| \le \sqrt{k}\Delta_{\max}\sum_m \|P^T(A^{(m)} - \mathcal{A}^{(m)})P\|_F$$

Then using Cauchy-Schwartz inequality we have the following result,

$$\begin{aligned}
&\sum_m \|P^T(A^{(m)} - \mathcal{A}^{(m)})P\|_F \\
&\quad \le \sum_m \sqrt{\operatorname{tr}((P^T(A^{(m)} - \mathcal{A}^{(m)})PP^T(A^{(m)} - \mathcal{A}^{(m)})P))} \\
&\quad \le \sqrt{M}\sqrt{\operatorname{tr}((\sum_m (A^{(m)} - \mathcal{A}^{(m)})PP^T(A^{(m)} - \mathcal{A}^{(m)}))PP^T)} \\
&\quad \le \sqrt{M}\sqrt{\|PP^T\|_2 \operatorname{tr}(\sum_m (A^{(m)} - \mathcal{A}^{(m)})PP^T(A^{(m)} - \mathcal{A}^{(m)}))} \\
&\quad = \sqrt{M}\sqrt{\operatorname{tr}(P^T\sum_m (A^{(m)} - \mathcal{A}^{(m)})(A^{(m)} - \mathcal{A}^{(m)})P)}
\end{aligned}$$

$$\leq \sqrt{M}\sqrt{k\|P\|_2^2\|\sum_m (A^{(m)}-\mathcal{A}^{(m)})(A^{(m)}-\mathcal{A}^{(m)})\|_2}$$
$$\leq \sqrt{M}\sqrt{kM\Delta_{\max}(\log n)^2} \leq \sqrt{k}M\sqrt{\Delta_{\max}}\log n,$$

with probability at least $1-o(1)$. In the above result, the inequality in line 3 is due to Cauchy-Schwartz inequality and line 4 follows from the inequality on trace of product of a positive semi-definite matrix $((A^{(m)}-\mathcal{A}^{(m)})PP^T(A^{(m)}-\mathcal{A}^{(m)}))$ with a Hermitian matrix (PP^T) due to [WKH86] (See also [FLF94]). The inequality in line 6 follows from the relations $\mathrm{tr}(XY) \leq k\|XY\|_2 \leq k\|X\|_2\|Y\|_2$. Finally the inequality in line 7 follows from Theorem 1 part (ii).

Similarly, we can derive

$$\begin{aligned}\sum_m (x^{(m)}-y^{(m)})^2 &\leq \sum_m \|P^T(A^{(m)}-\mathcal{A}^{(m)})P\|_F^2 \\ &= \mathrm{tr}((P^T(A^{(m)}-\mathcal{A}^{(m)})PP^T(A^{(m)}-\mathcal{A}^{(m)})P)) \\ &\leq kM\Delta_{\max}(\log n)^2,\end{aligned}$$

with probability at least $1-o(1)$.

Finally, combining the results together we have with probability at least $1-2\cdot o(1)$,

$$|F(\mathscr{A},P)-F(\mathbb{A},P)| \leq kM\Delta_{\max}(\log n)^2 + kM(\Delta_{\max})^{3/2}\log n.$$

Let $\hat{P}$ be the solution of the optimization problem in OLMF, i.e., $\hat{P}$ maximizes $F(\mathbb{A},P)$. Further let $\bar{P}$ maximizes the population version of the objective function $F(\mathscr{A},P)$. Then $F(\mathbb{A},\hat{P}) \geq F(\mathbb{A},\bar{P})$, and $F(\mathscr{A},\bar{P}) \geq F(\mathscr{A},\hat{P})$. Consequently, we have with probability at least $1-4\cdot o(1)$,

$$\begin{aligned}F(\mathscr{A},\bar{P})-F(\mathscr{A},\hat{P}) &\leq F(\mathscr{A},\bar{P})-F(\mathscr{A},\hat{P})+F(\mathbb{A},\hat{P})-F(\mathbb{A},\bar{P}) \\ &\leq |F(\mathscr{A},\bar{P})-F(\mathbb{A},\bar{P})|+|F(\mathscr{A},\hat{P})-F(\mathbb{A},\hat{P})| \\ &\leq 2kM\Delta_{\max}\log n(\log n+\sqrt{\Delta_{\max}}).\end{aligned}$$

Now define $\bar{\Lambda}^{(m)} = \bar{P}^T\mathcal{A}^{(m)}\bar{P}$ and $\Lambda_1^{(m)} = \hat{P}^T\mathcal{A}^{(m)}\hat{P}$. Note that since $\bar{P}$ is an invariant subspace of $\mathcal{A}^{(m)}$, we have $\mathcal{A}^{(m)} = \bar{P}^T\bar{\Lambda}^{(m)}\bar{P}$. We define $\mathcal{A}_1^{(m)} = \hat{P}\Lambda_1^{(m)}\hat{P}^T = \hat{P}\hat{P}^T\mathcal{A}^{(m)}\hat{P}\hat{P}^T$. Then $\hat{P}$ is an invariant subspace of $\mathcal{A}_1^{(m)}$. Further, we have for all m,

$$\|\bar{P}^T\mathcal{A}^{(m)}\bar{P}\|_F^2 - \|\hat{P}^T\mathcal{A}^{(m)}\hat{P}\|_F^2 = \|\mathcal{A}^{(m)}-\mathcal{A}_1^{(m)}\|_F^2. \quad [[\text{PC16c}]]$$

This result along with (3.2) imply,

$$\begin{aligned}F(\mathscr{A},\bar{P})-F(\mathscr{A},\hat{P}) &= \sum_m \|\bar{P}^T\mathcal{A}^{(m)}\bar{P}\|_F^2 - \|\hat{P}^T\mathcal{A}^{(m)}\hat{P}\|_F^2 \\ &= \sum_m \|\mathcal{A}^{(m)}-\hat{P}\hat{P}^T\mathcal{A}^{(m)}\hat{P}\hat{P}^T\|_F^2 \geq \sum_m (\lambda^{(m)})^2\|\hat{P}-\bar{P}O\|_F^2 \geq \frac{nr_{LMF}}{8n_{\max}}\sum_m(\lambda^{(m)})^2.\end{aligned}$$

Hence we have with probability at least $1-o(1)$

$$r_{LMF} \leq \frac{16n_{\max}k\Delta_{\max}\log n(\sqrt{\Delta_{\max}}+\log n)}{\frac{1}{M}\sum_m(\lambda^{(m)})^2 n}.$$

□

Proof of Theorem 4

Proof. We use the bound on the quantity $\|\bar{A} - \bar{\mathcal{A}}\|_2$ obtained in Theorem 1 part (i), Lemma 5.1 in [LR14] and the Davis-Kahan Theorem [DK70] to obtaint the following bound:

$$\|\hat{U} - \bar{U}O\|_F \leq \frac{2\sqrt{2}\sqrt{k}\|\bar{A} - \bar{\mathcal{A}}\|_2}{\lambda^{\bar{\mathcal{A}}}} \leq \frac{2\sqrt{2}\sqrt{k}}{\lambda^{\bar{\mathcal{A}}}}\sqrt{\frac{4\Delta_{\max}\log(2n/\epsilon)}{M}} = \frac{4\sqrt{2}}{\lambda^{\bar{\mathcal{A}}}}\sqrt{\frac{k\Delta_{\max}\log(2n/\epsilon)}{M}},$$

with probability at least $1 - \epsilon$ for any $\epsilon > 0$. Hence using (3.2) the misclustering rate is bounded as

$$r_{av} \leq \frac{8n_{\max}}{n}\|\hat{U} - \bar{U}O\|_F^2 \leq \frac{256n_{\max}k\Delta_{\max}\log(2n/\epsilon)}{(\lambda^{\bar{\mathcal{A}}})^2 nM},$$

with probability at least $1 - \epsilon$. □

Proof of Lemma 3

Proof. From the arguments in the proof of Lemma 2 we have,

$$\bar{B} = \frac{1}{M}\sum_m B^{(m)} = \frac{1}{M}\sum_m \{(p^{(m)} - q^{(m)})I_k + q^{(m)}1_k 1_k^T\}.$$

Hence 1_k is an eigenvector of $s\bar{B}$ corresponding to the largest eigenvalue $s\frac{1}{M}\sum_m\{(p^{(m)} - q^{(m)}) + nq^{(m)}\}$. All other eigenvectors correspond to the eigenvalue $s\frac{1}{M}\sum_m(p^{(m)} - q^{(m)})$. Hence $\lambda^{\bar{\mathcal{A}}} = s\frac{1}{M}\sum_m(p^{(m)} - q^{(m)})$. □

References

[BBC+14] Stefano Boccaletti, G Bianconi, R Criado, Charo I Del Genio, J Gómez-Gardeñes, M Romance, I Sendina-Nadal, Z Wang, and M Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[BDLBH17] Pierre Barbillon, Sophie Donnet, Emmanuel Lazega, and Avner Bar-Hen. Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A*, 180(1):295–314, 2017.

[BPW+13] Danielle S Bassett, Mason A Porter, Nicholas F Wymbs, Scott T Grafton, Jean M Carlson, and Peter J Mucha. Robust detection of dynamic community structure in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(1):013142, 2013.

[BSW+15] Urs Braun, Axel Schäfer, Henrik Walter, Susanne Erk, Nina Romanczuk-Seiferth, Leila Haddad, Janina I Schweiger, Oliver Grimm, Andreas Heinz,

and Heike Tost. Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proceedings of the National Academy of Sciences*, 112(37):11678–11683, 2015.

[BWP+11] Danielle S Bassett, Nicholas F Wymbs, Mason A Porter, Peter J Mucha, Jean M Carlson, and Scott T Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

[CHI16] Pin-Yu Chen and Alfred O Hero III. Multilayer spectral graph clustering via convex layer aggregation. *arXiv preprint arXiv:1609.07200*, 2016.

[CR11] Fan Chung and Mary Radcliffe. On the spectra of general random graphs. *The Electronic Journal of Combinatorics*, 18(1):P215, 2011.

[CSZ+07] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 153–162, 2007.

[DFVN12] Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov. Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing*, 60(11):5820–5831, 2012.

[DK70] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[FLF94] Yuguang Fang, Kenneth A Loparo, and Xiangbo Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

[GC13] D. Greene and P. Cunningham. Producing a unified graph representation from multiple social network views. *ACM Web Science*, 2:129–233, 2013.

[GZC+16] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Physical Review X*, 6(3):031005, 2016.

[Hae95] Willem H Haemers. Interlacing eigenvalues and graphs. *Linear Algebra and Its Applications*, 226:593–616, 1995.

[HLL83] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: some first steps. *Social Networks*, 5:109–137, 1983.

[HXA15] Qiuyi Han, Kevin S Xu, and Edoardo M Airoldi. Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1511–1520, 2015.

[KA10] Andrew V Knyazev and Merico E Argentati. Rayleigh-ritz majorization error bounds with applications to fem. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1521–1537, 2010.

[KAB+14] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[KCS11] Effrosini Kokiopoulou, Jie Chen, and Yousef Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011.

[KRD11] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1413–1421, 2011.

[LR14] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2014.

[LWGH13] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260, 2013.

[LZWY06] Bo Long, Zhongfei Mark Zhang, Xiaoyun Wu, and Philip S Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 585–592, 2006.

[MRM+10] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

[NTK11] M. Nickel, V. Tresp, and H. P. Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of 28th International Conference on Machine Learning*, pages 809–816, 2011.

[PC16a] Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2):3807–3870, 2016.

[PC16b] Subhadeep Paul and Yuguo Chen. Null models and modularity based community detection in multi-layer networks. *arXiv preprint arXiv:1608.00623*, 2016.

[PC16c] Subhadeep Paul and Yuguo Chen. Orthogonal symmetric non-negative matrix factorization under the stochastic block model. *arXiv preprint arXiv:1605.05349*, 2016.

[Pei15] Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

[QR13] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

[RCY11] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist*, 39(4):1878–1915, 2011.

[SG08] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658, 2008.

[SS90] Gilbert W Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, Boston, MA., 1990.

[SSTM16] Natalie Stanley, Saray Shai, Dane Taylor, and Peter J Mucha. Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering*, 3(2):95–105, 2016.

[Sun13] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[TLD09] Wei Tang, Zhengdong Lu, and Inderjit S Dhillon. Clustering with multiple graphs. In *Proceedings of the Ninth IEEE International Conference on Data Mining,*, pages 1016–1021, 2009.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[TWL12] Lei Tang, Xufei Wang, and Huan Liu. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33, 2012.

[VCMGSP16] Toni Valles-Catala, Francesco A Massucci, Roger Guimera, and Marta Sales-Pardo. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Physical Review X*, 6(1):011036, 2016.

[WKH86] Sheng-De Wang, Te-Son Kuo, and Chen-Fa Hsu. Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation. *IEEE Transactions on Automatic Control*, 31(7):654–656, 1986.

[WPBN16] James D Wilson, John Palowitch, Shankar Bhamidi, and Andrew B Nobel. Community extraction in multilayer networks with heterogeneous community structure. *arXiv preprint arXiv:1610.06511*, 2016.

[XTX13] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[ZB07] Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1159–1166, 2007.

[ŽZ15] Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):41–53, 2015.