GLOBAL AND INDIVIDUALIZED COMMUNITY DETECTION IN INHOMOGENEOUS MULTILAYER NETWORKS

BY SHUXIAO CHEN¹,*, SIFAN LIU² AND ZONGMING MA¹,†

¹ Department of Statistics, University of Pennsylvania, *shuxiaoc@wharton.upenn.edu; †zongming@wharton.upenn.edu
² Department of Statistics, Stanford University, sfliu@stanford.edu

In network applications, it has become increasingly common to obtain datasets in the form of multiple networks observed on the same set of subjects, where each network is obtained in a related but different experiment condition or application scenario. Such datasets can be modeled by multilayer networks where each layer is a separate network itself while different layers are associated and share some common information. The present paper studies community detection in a stylized yet informative inhomogeneous multilayer network model. In our model, layers are generated by different stochastic block models, the community structures of which are (random) perturbations of a common global structure while the connecting probabilities in different layers are not related. Focusing on the symmetric two block case, we establish minimax rates for both global estimation of the common structure and individualized estimation of layer-wise community structures. Both minimax rates have sharp exponents. In addition, we provide an efficient algorithm that is simultaneously asymptotic minimax optimal for both estimation tasks under mild conditions. The optimal rates depend on the parity of the number of most informative layers, a phenomenon that is caused by inhomogeneity across layers. The method is extended to handle multiple and potentially asymmetric community cases. We demonstrate its effectiveness on both simulated examples and a real-world single-cell dataset.

1. Introduction. Network data is among the most common types of relational data. As a fundamental task in network data analysis [67, 28], community detection refers to the problem of partitioning the nodes of a network into clusters so that intra-cluster nodes are connected in a different way from inter-cluster nodes, usually more densely. Stochastic block model (SBM) [32] is a canonical model for studying community detection. In an SBM, n nodes are partitioned into k disjoint subsets. Each unordered pair of nodes are connected independently with probability p if they come from the same community and with a different probability q otherwise. The observed connection pattern is encoded in an n × n symmetric adjacency matrix A. Here the goal of community detection is to, upon observing A, estimate the partitioning of nodes. The stochastic block model, albeit simple, has found its success in many fields of science (see, e.g., [33, 46, 34, 27]). It has also undergone a plethora of theoretical investigations [1]. With joint efforts from mathematics, statistics, and computer science, we not only have a large algorithmic toolbox for detecting communities in SBMs, but also know the information-theoretic limits of this task as well as which algorithms are optimal.

Despite its popularity, stochastic block model focuses only on a single adjacency matrix. This is in sharp contrast to the widely recognized fact that real world networks are often superpositions of multiple networks (layers), each encoding a potentially different but corre-
lated interaction pattern among the same set of nodes \([37, 9]\). For example, in social network data, different layers could correspond to different types of relationship that link the social entities, and the information presented in the friendship network, compared to that in the professional network, is different, but not completely unrelated \([20]\). Another example is given by the network representation of human brains, where each layer corresponds to an individual person’s functional brain network. It is well known that the parcellation of brain regions into different functional units are different but strongly correlated among human beings \([16]\).

A natural attempt at generalizing SBMs to multilayer networks is as follows. Let us focus on the symmetric two block case where in each layer all nodes are partitioned into two blocks of roughly equal sizes. Instead of observing a single adjacency matrix, the data analyst is now presented with a collection of \(L\) adjacency matrices \(\{A^{(\ell)}\}_{\ell=1}^L\). To model that “communities in different layers are different but correlated”, we take a hierarchical modeling approach. A global community assignment vector \(z^* \in \{\pm 1\}^n\) is introduced in our model. To ensure symmetry, we require \(z^*\) to have roughly equal numbers of 1’s and -1’s. We let the individual community assignments \(\{z^{(\ell)}\}_\ell\) be independent samples from the following distribution:

\[
\forall \ell \in [L], i \in [n], \quad z^{(\ell)}_i \overset{\text{iid}}{\sim} z^* \times \left[2\text{Bern}(1 - \rho) - 1\right].
\]

Here, \([L] = \{1, \ldots, L\}, [n] = \{1, \ldots, n\}\) and \(\text{Bern}(\cdot)\) is a Bernoulli random variable. That is, in a specific layer \(\ell \in [L]\), with probability \(1 - \rho\), the community membership of the \(i\)-th node agrees with the global one, \(z^*_i\), and with probability \(\rho\), it “flips” to the opposite side \(-z^*_i\). The parameter \(\rho\) controls the level of inhomogeneity across layers. When \(\rho = 0\), all layers share the same community structure, whereas when \(\rho = 1/2\), the community structures across layers are mutually uninformative. Upon realizations of \(z^{(\ell)}\)'s, the adjacency matrices are independently generated by

\[
A^{(\ell)}_{ij} = A^{(\ell)}_{ji} \sim \text{Bern}(p_\ell) \cdot \mathbb{I}\{z^{(\ell)}_i = z^{(\ell)}_j\} + \text{Bern}(q_\ell) \cdot \mathbb{I}\{z^{(\ell)}_i \neq z^{(\ell)}_j\},
\]

\[
\forall i \neq j \in [n], \ell \in [L],
\]

where \(\mathbb{I}\{\cdot\}\) is the indicator function, and all diagonal entries are zeros. In other words, the \(\ell\)-th layer network is generated by an SBM with community partitioning specified by \(z^{(\ell)}\), intra-community connection probability \(p_\ell\) and inter-community connection probability \(q_\ell\). Connection probabilities across different layers are not linked in any way. The foregoing generalization, to the best of our knowledge, was first introduced by Paul and Chen \([57]\), which they termed as the random effects stochastic block model. In \([57]\), the random effects (1.1) can take other forms. Hence, to avoid confusion, we term the model in (1.1)–(1.2) as the inhomogeneous multilayer stochastic block model (IMLSBM). Clearly, the model can be generalized in obvious ways to include more than two communities and unequal community sizes. However, the present manuscript shall focus on the foregoing symmetric two block case as it is the simplest nontrivial model that reveals key new phenomena of community detection in inhomogeneous multilayer networks.

The goal of community detection in an IMLSBM is now two-fold—upon observing \(\{A^{(\ell)}\}_\ell\), we are interested in:

1. **Global estimation.** Estimating the global community assignment \(z^*\);
2. **Individualized estimation.** Estimating each of the individual assignments \(\{z^{(\ell)}\}_\ell\).

Global estimation needs to aggregate connection patterns across layers to better infer the global consensus structure, an instance of integrative data analysis \([15]\). On the other hand,
individualized estimation requires borrowing information from different layers to better estimate the layer-wise community structure, an example of multi-task learning [11].

Theoretical understanding of community detection in an IMLSBM is lacking. Partial results exist in the homogeneous case ($\rho = 0$) where global estimation and individualized estimation coincide. Here, by “homogeneity” we mean the layer-wise community structures are shared across all layers, and the connecting probabilities are allowed to differ. Under such a setup, it has been proved by Paul and Chen [56] that the minimax rates for expected proportion of misclustered nodes scales as

$$\exp\left\{ - \left( 1 + o(1) \right) \frac{1}{2} \sum_{\ell \in [L]} \left( \sqrt{p_{\ell}} - \sqrt{q_{\ell}} \right)^2 \right\},$$

provided that the exponent diverges to infinity as $n$ tends to infinity. They in fact established the rates for a more general setting than the symmetric two block case considered in this paper. Later, a polynomial-time algorithm that achieves this rate was proposed by Xu, Jog and Loh [68]. Nonetheless, it is unclear how to generalize their results to the inhomogeneous setting. From an algorithmic perspective, spectral clustering [8, 43, 58] and least-square estimators [44, 66] have been proposed and justified to be consistent (i.e., achieving an $o(1)$ misclustering proportion with high probability) under homogeneity. However, it is unknown whether any of these methods attains the information-theoretic limit (1.3).

Although the homogeneous case ($\rho = 0$) is interesting in its own right, it is the inhomogeneous case ($\rho > 0$) that characterizes our inductive bias — “layers are different but correlated”. In [57], a few heuristic algorithms were introduced and their performances were assessed by simulations. To the best of our limited knowledge, in the inhomogeneous regime, no algorithm with provable guarantee for either global or individualized estimation is known in the literature, let alone any optimality statement.

1.1. Main Contributions. The main contributions of the present manuscript are two-fold. First, we give precise characterization of information-theoretic limits of both global and individualized community detection in a symmetric two block IMLSBM when $\rho = o(1)$; Moreover, we provide a polynomial-time algorithm that simultaneously attains information-theoretic limits for both global and individualized estimation under mild conditions. We reiterate that results in the present manuscript are obtained under the symmetric two block setting which has already posed highly nontrivial theoretical and algorithmic challenges. We leave extensions to more general settings for future work.

To provide an overview of our main results, we start with several key information-theoretic quantities that will appear throughout this paper. For any $\ell \in [L]$ and $t \in [0, 1]$, define

$$I^t_\ell := -\log \left[ p_{\ell}^{1-t} q_{\ell} + (1 - p_{\ell})^{1-t} (1 - q_{\ell})^t \right] p_{\ell} q_{\ell}^{1-t} + (1 - p_{\ell})^t (1 - q_{\ell})^{1-t}.$$ The quantity $I^t_\ell$ can be regarded as the signal strength of the $\ell$-th layer. When $p_{\ell} \simeq q_{\ell} = o(1)$, one can show that $I^t_\ell = (1 + o(1)) \left( \sqrt{p_{\ell}} - \sqrt{q_{\ell}} \right)^2$, and hence the minimax rate for community detection (i.e., the worst-case misclustering proportion) in an SBM with community assignment $z^{(t)}$, intra-community connection probability $p_{\ell}$ and the inter-community connection probability $q_{\ell}$ derived in Zhang and Zhou [72] can be equivalently written as $e^{-(1 + o(1)) n I^t_\ell / 2}$, as long as the exponent tends to infinity. For any collection of layers $S \subseteq [L]$, let

$$\psi_S(t) := -\frac{1}{2} \sum_{\ell \in S} I^t_\ell, \quad \psi_S^*(a) := \sup_{0 \leq t \leq 1} at - \psi_S(t).$$
The function \( \psi^*_S \) characterizes the collective signal strength for layers in \( S \). Indeed, from the definition of \( I^*_t \), one readily checks that \( \psi^*_S(0) = -\psi_S(1/2) = (n/2) \sum_{\ell \in S} I^*_t, \) and thus the minimax rate (1.3) for community detection in a homogeneous multilayer SBM can be expressed equivalently as \( e^{-(1+o(1))\psi^*_S(0)} \). Intuitively, inhomogeneity (\( \rho > 0 \)) introduces additional noises. To characterize the noise level, define

\[
J_\rho := -\log 2\sqrt{\rho(1-\rho)}.
\]

Since \( J_0 = \infty \) and \( J_{1/2} = 0 \), one can effectively think of \( J_\rho \) as a measure of proximity of the individual layer community assignments \( \{z^{(t)}_L\}_1^L \) to the global assignment \( z^* \). Note that both \( I^*_t \) and \( J_\rho \) can be written as convex combinations of Rényi divergences [60] between Bernoulli distributions. Specifically, we have \( I^*_t = (1 - t)D_t(p_t || q_t) + tD_{1-t}(p_t || q_t) \), and \( J_\rho = \frac{1}{2}D_{1/2}(\rho \parallel 1 - \rho) \), where \( D_t(p \parallel q) \) is the Rényi divergence of order \( t \) between \( \text{Bern}(p) \) and \( \text{Bern}(q) \).

1.1.1. Global estimation error. With the foregoing definitions, we first show that, under certain regularity conditions, the minimax rate for global estimation, measured in terms of proportion of misclustered nodes, is given by

\[
\exp \left\{ - (1 + o(1)) \min_{S \subseteq [L]} I_S \right\},
\]

where \( I_S \) represents the signal-to-noise ratio (SNR) for global estimation for layers in \( S \), defined as

\[
I_S := \begin{cases} 
|S^c|J_\rho + \psi^*_S(0) & \text{if } |S^c| \text{ is even}, \\
(|S^c| + 1)J_\rho + \psi^*_S(-2J_\rho) & \text{if } |S^c| \text{ is odd}.
\end{cases}
\]

The minimax error rate (1.7) exhibits two intriguing properties. First, when \( \rho = 0 \), we have \( J_\rho = \infty \), and thus the only way to make \( I_S \) finite is to choose \( S = [L] \), which gives \( \min_{S \subseteq [L]} I_S = I_{[L]} = \psi^*_S(0) \). As the result, (1.7) recovers the minimax rate in a homogeneous multilayer SBM given in (1.3). Second, the SNR for layers in \( S \) takes different forms according to the parity of \( |S^c| \), a phenomenon induced by inhomogeneity across layers.

1.1.2. Individualized estimation error. In correspondence, the minimax rate for individualized estimation for the \( \ell \)-th layer, measured by proportion of misclustered nodes, is given by

\[
\exp \left\{ - (1 + o(1)) \min_{S \subseteq [L] \backslash \{\ell\}} I_{S \cup \{\ell\}} \right\} + \exp \left\{ - (1 + o(1)) J_{(\ell)} \right\},
\]

where \( J_{(\ell)} \) is a suitably defined quantity (see (2.21) for a precise definition) that measures the SNR for individualized estimation for the \( \ell \)-th layer. We briefly mention here that similar to (1.7), the last display can recover the minimax rate in a homogeneous multilayer SBM by setting \( \rho = 0 \), and it crucially depends on the parity of the “most informative” set \( S^c \) as well. We refer readers to Sections 2 and 4 for details.

1.1.3. Algorithm. We propose an algorithm that achieves the optimal rates in (1.7) and (1.9) simultaneously under mild conditions. The idea stems from maximum a posteriori (MAP) estimation. Note that IMLSBM is a hierarchical model where individual community assignments \( \{z^{(t)}_L\}_1^L \) are drawn from the prior distribution (1.1). It is thus tempting to write out the posterior of \( \{z^{(t)}_L\}_1^L \) given the observed data \( \{A^{(t)}_L\}_1^L \) and maximize the posterior density with respect to the parameters \( (z^*, \{z^{(t)}_L\}_1^L) \). A naive implementation of this strategy is
doomed to fail, due to the fact that the MAP objective function gives rise to a combinatorial optimization problem, whose search space has cardinality $2^{n(L+1)}$. To bypass the combinatorial search, we adopt a two-stage “warm-start” MAP algorithm. In the first stage, an initial estimator of $z^*$ is obtained using spectral clustering on a trimmed version of the weighted average of layer-wise adjacency matrices. In the second stage, a refined estimator of $z^*$ and estimators of $\{z^{(\ell)}\}$ are simultaneously obtained by optimizing a “decoupled” MAP objective function, which can be computed in linear time (in $n$ and $L$). Although the formal definition of the MAP refinement step requires knowledge of $\rho$, our numerical experiments later show that the outcome is not sensitive to misspecification of $\rho$.

While “spectral clustering + refinement” procedures have appeared in community detection in SBMs and variants (e.g., [50, 25, 70]), our algorithm has novelties in both stages, especially in their technical analysis. For Stage I, compared to existing analyses of spectral clustering for homogeneous multilayer SBMs [8, 43, 58], our analysis is novel in that we establish a stability result, asserting that inhomogeneity hurts spectral clustering error rate by at most an additive factor of $\text{poly}(\rho)$. The proof is based on a new concentration inequality on the spectral norm of a weighted average of Bernoulli random matrices, which is derived via a nontrivial generalization of the graph decomposition approach in Le, Levina and Vershynin [42] to multilayer networks. The concentration inequality improves the ones used in the existing work (e.g., [8, 58]) and could be of independent interest. For Stage II, due to presence of multiple layers, devising a refinement scheme with time complexity that is polynomial in the number of layers presents new challenges. In addition, due to inhomogeneity, the analysis is considerably more involved. A key step towards establishing matching upper bounds lies in a novel application of Sion’s minimax theorem [61].

1.2. Related Work. The past decade has witnessed a venerable line of work on the theoretical development of community detection for SBMs. Optimal algorithms have been developed under various criteria, including (1) weak recovery, where the best achievable goal is to cluster the nodes better than random guess [19, 49, 52, 10, 51, 13, 54, 48, 4]; (2) exact recovery, where the requirement is to reconstruct from data the ground truth up to relabeling [50, 2, 3]; and more related to our formulation, (3) almost exact recovery, where the goal is to output a community assignment with vanishing misclustering error [69, 50, 2, 25]. The study under the minimax framework was initiated by [72, 3, 70] and was later extended to more general settings such as [26, 68]. The above list of work is by no means exhaustive and we refer the readers to the review papers [1, 45, 22] for a more systematic account.

In comparison, study of community detection in multilayer networks is still in its early stage. Initial works in this area have focused on algorithmic developments (see, e.g., [53, 39, 17, 18, 59]), and most theoretical studies are restricted to the homogeneous case where all layers share the same community structure [56, 8, 57, 58, 44, 43, 71]. There are a few exceptions, such as [62, 41, 5, 35], where consistency has been established under several inhomogeneity-aware variants of SBMs while optimality results are missing.

In addition to estimating community structures, there is a related line of work aiming at testing whether community structures across layers are indeed correlated or equivalent. We refer interested readers to recent papers [24, 21, 23] and references therein.

1.3. Paper Organization. The rest of the paper is organized as follows. In Section 2, we provide our construction of minimax lower bounds for both global and individualized estimation. We present the two-stage algorithm in Section 3, and its theoretical analysis is given in Section 4. We conduct numerical experiments to corroborate our theoretical results in Sec-
tion 5. Section 6 extends our algorithm to multi-cluster and asymmetric cases. We finally illustrate our algorithm in a multi-modal single-cell dataset in Section 7. For brevity, additional theoretical and numerical results, as well as the technical proofs are deferred to the supplementary material [12].

1.4. Notation. We conclude this section by introducing some notations that will be used throughout this paper. For a positive integer \( n \), we write \([n] := \{1, \ldots, n\}\). Given \( a, b \in \mathbb{R} \), we denote \( a \land b := \min\{a, b\} \) and \( a \lor b := \max\{a, b\} \). For a set \( S \), we let \( \mathbb{1}_S \) be its indicator function and we use \( \#S \) and \( |S| \) interchangeably to denote its cardinality. For two positive sequences \( \{a_n\} \) and \( \{b_n\} \), we write \( a_n \lesssim b_n \) or \( a_n = \Theta(b_n) \) to denote \( \limsup a_n/b_n < \infty \), and we let \( a_n \gtrsim b_n \) or \( a_n = \Omega(b_n) \) to denote \( b_n \lesssim a_n \). Meanwhile, the notation \( a_n \asymp b_n \) or \( a_n = \Theta(b_n) \) means \( a_n \lesssim b_n \) and \( a_n \gtrsim b_n \) simultaneously. Moreover, we write \( a_n \ll b_n \) to mean \( b_n/a_n \to \infty \) and \( a_n \gg b_n \) to mean \( b_n \ll a_n \). For a vector \( x \), we let \( \|x\|_p \) denote its \( \ell_p \) norm, and we write \( \|x\|_2 = \|x\| \) when there is no ambiguity. For a matrix \( A \), we let \( \|A\|_F \) be its Frobenius norm and \( \|A\|_{p \to q} \) be its \( \ell_p \) to \( \ell_q \) operator norm. We will write \( \|A\|_{2 \to 2} = \|A\|_2 = \|A\| \) when there is no ambiguity.

2. Fundamental Limits and Costs of Inhomogeneity. In this section, we present minimax lower bounds for estimating both \( z^* \) and individual \( z^{(\ell)} \)'s. To start with, let us recall that a two-block IMLSBM parameterized by \((z^*, \rho, \{p_\ell\}_{1}^{L}, \{q_\ell\}_{1}^{L})\) is a probability measure on a multilayer network, whose adjacency matrices \( \{A^{(\ell)}\}_{1}^{L} \) are generated according to (1.1)–(1.2).

Parameter space and the loss function. Let \( n_{\pm}^*(z^*) = \sum_{i \in [n]} \mathbb{1}\{z^* = \pm 1\} \) be the sizes of the positive and negative clusters of \( z^* \), respectively. We propose to consider the following collection of IMLSBMs:

\[
\mathcal{P}_n(\rho, \{p_\ell\}_{1}^{L}, \{q_\ell\}_{1}^{L}, \beta) := \left\{ \text{IMLSBM}(z^*, \rho, \{p_\ell\}_{1}^{L}, \{q_\ell\}_{1}^{L}) : \frac{n}{2 \beta} \leq n_{\pm}^*(z^*) \leq \frac{n \beta}{2}, \quad p_\ell > q_\ell \quad \forall \ell \right\}.
\]

As we focus on the symmetric case, the constant \( \beta \geq 1 \) is taken to be \( 1 + o(1) \) as \( n \to \infty \). The rest of the quantities appearing above, namely \( L, \rho, \{p_\ell\}_{1}^{L}, \{q_\ell\}_{1}^{L} \), are all allowed to scale with \( n \). For an estimator \( \hat{z}^* \) of \( z^* \), we evaluate its performance by the misclustering proportion, defined as

\[
\mathcal{L}(\hat{z}^*, z^*) := \frac{d_\mathcal{H}(\hat{z}^*, z^*) \land d_\mathcal{H}(-\hat{z}^*, z^*)}{n},
\]

where \( d_\mathcal{H}(z, z') = \sum_{i \in [n]} \mathbb{1}\{z_i \neq z'_i\} \) is the Hamming distance between \( z \) and \( z' \), and the minimum is taken because \( \hat{z}^* \) and \( -\hat{z}^* \) give rise to the same partitioning of nodes. Similarly, for an estimator \( \hat{z}^{(\ell)} \) of \( z^{(\ell)} \), we evaluate its performance by \( \mathcal{L}(\hat{z}^{(\ell)}, z^{(\ell)}) \).

An idealized setup. To characterize the information-theoretic limits in estimating \( z^* \) and \( z^{(\ell)} \)'s, we will consider an idealized setup as follows. Suppose \( n = 2m + 1 \) and the nodes are labeled as \( 0, 1, \ldots, 2m \). Consider a global assignment vector \( z^* \) whose value on \( i \in \{1, \ldots, 2m\} \) is known to us:

\[
z^*_i = \begin{cases} +1 & 1 \leq i \leq m \\ -1 & m + 1 \leq i \leq 2m. \end{cases}
\]
We further observe \( \{A^{(\ell)}\}_1^L \sim \text{IMLSBM}(\mathbf{z}^*, \rho, \{p_{\ell}\}_1^L, \{q_{\ell}\}_1^L) \). And our goal is to estimate \( z_0^* \) as well as \( z_0^{(\ell)} \)'s, a substantially simplified task compared to the original one. Such a strategy of “reducing” the task of doing inference for the whole parameter vector to doing inference for each coordinate is an instance of the celebrated Assouad’s method [6], and has been successfully used in many recent works on characterizing the fundamental limits of community detection in SBMs and variants (see, e.g., [72, 25, 26]). Our discussion in the rest of this section will largely rely on the intuitions built upon this idealized setup, and we refer the readers to Section B of the supplementary material [12] for fully rigorous proofs.

2.1. Minimax Lower Bound for Global Estimation. Consider the task of estimating \( \mathbf{z}^* \). Under the idealized setup (2.3), our goal is to differentiate between \( z_0^* = +1 \) and \( z_0^* = -1 \) based on the data \( \{A^{(\ell)}\}_1^L \), which gives rise to a binary hypothesis testing problem:

\[
H_0 : z_0^* = +1 \quad \text{v.s.} \quad H_1 : z_0^* = -1.
\]

By Neyman–Pearson lemma, in principle, we can characterize the difficulty of the above testing problem by calculating the error made by the likelihood ratio test. However, due to the complicated structure of the likelihood function, this strategy is analytically intractable, calling for further simplifications.

2.1.1. The fundamental testing problem. If \( z_0^{(\ell)} \) is actually observed by us, then since \( z_0^{(\ell)} \sim \text{Bern}(1 - \rho) - 1 \) under \( H_0 \) and \( z_0^{(\ell)} \sim \text{Bern}(\rho) - 1 \) under \( H_1 \), deciding the value of \( z_0^* \) from \( z_0^{(\ell)} \) is equivalent to the problem of differentiating between

\[
H_0 : \text{Bern}(1 - \rho) - 1 \quad \text{v.s.} \quad H_1 : \text{Bern}(\rho) - 1.
\]

In reality, we need to estimate each \( z_0^{(\ell)} \) from its corresponding \( A^{(\ell)} \). This is the community detection problem in a vanilla two-block SBM, whose fundamental difficulty is characterized by the following testing problem [25]:

\[
H_0 : \bigotimes_{i \in [m]} \text{Bern}(p_{\ell}) \otimes \bigotimes_{i \in [m]} \text{Bern}(q_{\ell}) \quad \text{v.s.} \quad H_1 : \bigotimes_{i \in [m]} \text{Bern}(q_{\ell}) \otimes \bigotimes_{i \in [m]} \text{Bern}(p_{\ell}),
\]

where \( \otimes \) denotes the product of probability measures.

Intuitively, if the signal strength in the \( \ell \)-th layer is strong enough, then we are close to the case of known \( z_0^{(\ell)} \), in which the error for testing (2.4) is mainly captured by that of testing (2.5). On the other hand, if we have barely any signal in the \( \ell \)-th layer, then we are in the unknown \( z_0^{(\ell)} \) case, in which the error for testing (2.4) mainly comes from testing (2.6). This intuition is formalized in the following lemma.

**Lemma 2.1** (The fundamental testing problem for global estimation). Assume \( \log L \ll n^c \) for some \( c \in (0, 1) \). Then for any sequence \( \delta_n = o(1) \), there exists another sequence \( \delta_n' = o(1) \) satisfying \( (1 + \delta_n') n/2 \in \mathbb{N} \), such that for any \( S \subseteq [L] \),

\[
\inf_{\hat{\mathbf{z}}^*, \mathbf{z}^*} \sup_{\mathbb{P}_n} \mathbb{E} \mathcal{L}(\hat{\mathbf{z}}^*, \mathbf{z}^*) \gtrsim \delta_n \inf_{\phi} \left( \mathbb{E}_{H_0}[\phi] + \mathbb{E}_{H_1}[1 - \phi] \right),
\]

where \( \phi \) is a testing function of the following problem:

\[
H_0 : \left( \bigotimes_{\ell \in S^c} [\text{Bern}(1 - \rho) - 1] \right) \otimes \left( \bigotimes_{\ell \in S} (1 + \delta_n') \otimes \bigotimes_{i = 1}^{(1 + \delta_n') n/2} \text{Bern}(p_{\ell}) \otimes \text{Bern}(q_{\ell}) \right).
\]
\begin{equation}
(2.8) \quad v.s. \quad H_1 : \left( \bigotimes_{\ell \in S^c} [2\text{Bern}(\rho) - 1] \right) \otimes \left( \bigotimes_{\ell \in S} \bigotimes_{i = 1}^{(1+\delta_n')n/2} \text{Bern}(q_\ell) \otimes \text{Bern}(p_\ell) \right).
\end{equation}

**Proof.** See Section B.1 of the supplementary material [12].

**Remark 2.1.** Though the lower bound (2.7) holds for an arbitrary $S \subseteq [L]$, by our previous intuition, it is the tightest if we choose $S$ such that $S^c$ corresponds to layers with high SNRs, where (2.5) kicks in, and $S$ corresponds to layers with low SNRs, where the error from (2.6) dominates.

2.1.2. **Optimal testing error and parity of $|S^c|$.** By Neyman–Pearson lemma, the test that gives the optimal Type-I plus Type-II error is the likelihood ratio test with a cutoff of $1$. For the problem (2.8), it can be shown (see Section B.1 of the supplementary material [12] for a detailed derivation) that the optimal error is given by

\begin{equation}
(2.9) \quad P \left( \sum_{\ell \in S} \log \left( \frac{q_\ell (1 - p_\ell)}{p_\ell (1 - q_\ell)} \right) \cdot \sum_{i = 1}^{(1+\delta_n')n/2} (X_i^{(\ell)} - Y_i^{(\ell)}) \geq \log \left( \frac{1 - \rho}{\rho} \right) \cdot \sum_{\ell \in S^c} Z^{(\ell)} \right),
\end{equation}

where

\begin{equation}
(2.10) \quad \{X_i^{(\ell)}\}_{i = 1}^n \overset{i.i.d.}{\sim} \text{Bern}(p_\ell), \quad \{Y_i^{(\ell)}\}_{i = 1}^n \overset{i.i.d.}{\sim} \text{Bern}(q_\ell), \quad \{Z^{(\ell)}\}_{i = 1}^L \overset{i.i.d.}{\sim} 2\text{Bern}(1 - \rho) - 1,
\end{equation}

all of which are mutually independent.

Readers with an expertise in large deviation principles may have noticed that (2.9) is the tail probability of a sum of independent random variables, and a tight characterization of this probability should involve the cumulant generating functions (CGFs) of those random variables as well as their rate functions (i.e., the Legendre transforms of CGFs). This intuition explains the appearance of the two key information-theoretic quantities, namely $\psi_S^c$ and $J_\rho$, in the definition of $\mathcal{I}_S$ in (1.8). As one can check, $\psi_S^c(t)$ is precisely the CGF of the random variable that appears to the left of the “$\geq$” sign in (2.9) (if we set $\delta_n' = 0$), and $-J_\rho$ is the CGF of $\log \left( \frac{(1 - \rho)}{\rho} \right) \cdot Z^{(\ell)}$ evaluated at $1/2$.

The following lemma gives the asymptotically optimal testing error for (2.8).

**Lemma 2.2 (Optimal testing error for global estimation).** Assume $\rho = o(1)$ and that there exist constants $C_1, C_2 > 1, c \in (0, 1)$ such that $C_1 q_\ell \leq p_\ell \leq (C_2 q_\ell) \wedge (1 - c)$ for any $\ell \in [L]$. Then there exists a sequence $\delta_n'' = o(1)$ which is independent of $S$, such that the probability (2.9) is lower bounded by

\begin{equation}
(2.11) \quad C \cdot \exp \left\{ - (1 + \delta_n'') \mathcal{I}_S \right\},
\end{equation}

where $C > 0$ is an absolute constant and $\mathcal{I}_S$ is defined in (1.8).

**Proof.** See Section B.2 of the supplementary material [12].

**Remark 2.2.** The SNR $\mathcal{I}_S$ for global estimation, which appears on the exponent in the optimal testing error (2.11), takes different forms according to the parity of $|S^c|$. There is a fundamental reason for this. It happens that the dominating term in the probability (2.9) is given by the part with $\sum_{\ell \in S^c} Z^{(\ell)}$ being non-positive and closest to zero. Since $Z^{(\ell)}$’s are $\{\pm 1\}$-valued, such a requirement translates to $\sum_{\ell \in S^c} Z^{(\ell)} = 0$ when $|S^c|$ is even, and gives $\sum_{\ell \in S^c} Z^{(\ell)} = -1$ when $|S^c|$ is odd.
2.1.3. Minimax lower bound for global estimation. Equipped with Lemmas 2.1 and 2.2, we are ready to present the main result in this subsection.

THEOREM 2.1 (Minimax lower bound for global estimation). Assume \( \rho = o(1) \). Meanwhile, assume there exist constants \( C_1, C_2 > 1 \) and \( c_1, c_2 = (0, 1) \) such that \( C_1q \leq p_\ell \leq (C_2q) \wedge (1 - c_1), \forall \ell \in [L] \) and \( \log L \ll n^{c_2} \). If \( \min_{S \subseteq [L]} I_S \to \infty \), then there exists a sequence \( \delta_n = o(1) \) such that

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq \exp \left\{ - (1 + \delta_n) \min_{S \subseteq [L]} I_S \right\}.
\]

On the other hand, if \( \min_{S \subseteq [L]} I_S = O(1) \), then there exists some \( c' > 0 \) such that

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq c'.
\]

PROOF. If \( \min_{S \subseteq [L]} I_S \to \infty \), then we can find a sequence \( \delta_n = o(1) \) such that \( \log(\delta_n^{-1}) \ll \min_{S \subseteq [L]} I_S \). Then, invoking Lemmas 2.1 and 2.2, there exists \( \delta'_n = o(1) \) such that

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq \delta_n \exp \left\{ - (1 + \delta_n') \min_{S \subseteq [L]} I_S \right\} = \exp \left\{ - (1 + \delta_n' + \delta_n'') \min_{S \subseteq [L]} I_S \right\},
\]

where \( \delta''_n = \log(\delta_n^{-1})/\min_{S \subseteq [L]} I_S = o(1) \) by construction. Choosing \( \delta_n = \delta_n' + \delta_n'' \) gives (2.12).

On the other hand, if \( \min_{S \subseteq [L]} I_S = O(1) \), then repeating the above arguments gives

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq \delta_n \text{ for any } o(1) \text{ sequence } \delta_n.
\]

If \( \inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \) is itself \( o(1) \), then we would have

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq \sqrt{\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*)},
\]

a contradiction. Hence (2.13) follows.

As an immediate corollary, we have the following result for the homogeneous case \( (\rho = 0) \).

COROLLARY 2.1 (Minimax lower bound for global estimation under homogeneity). Under the setup of Theorem 2.1, assume in addition that \( \rho = 0 \). If \( n \sum_{\ell \in [L]} I^{(\ell)}_1 \to \infty \), then there exists a sequence \( \delta_n = o(1) \) such that

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq \exp \left\{ - (1 + \delta_n) \frac{n}{2} \sum_{\ell \in [L]} I^{(\ell)}_1 \right\}.
\]

On the other hand, if \( n \sum_{\ell \in [L]} I^{(\ell)}_1 = O(1) \), then there exists some \( c' > 0 \) such that

\[
\inf_{\hat{z}^*, z^* \in \mathcal{P}_n} \mathbb{E} \mathcal{L}(\hat{z}^*, z^*) \geq c'.
\]

PROOF. This follows from the fact that if \( \rho = 0 \), then \( J_\rho = \infty \) and thus the set \( S \) that minimizes \( I_S \) is \( S = [L] \).

2.2. Minimax Lower Bound for Individualized Estimation. We now derive minimax lower bound for estimating individual \( z^{(\ell)} \)'s. Let us again consider the idealized setup (2.3).
2.2.1. Two testing problems from two sources of errors. Suppose that we additionally know the value of $z_0^*$, say $z_0^* = +1$. Since $z_0^{(t)}$ is independent of $z_0^{(-t)} := \{z_r^{(r)} : r \neq t\}$, the only information that’s useful in determining $z_0^{(t)}$ comes from the following “label sampling” model:

\begin{equation}
 z_i^{(t)} \sim z_i^* \times [2\text{Bern}(1 - \rho) - 1], \quad A^{(t)} \sim \text{SBM}(z^{(t)}; p_\ell, q_\ell).
\end{equation}

Now, for any estimator $\hat{z}_0^{(t)}$ of $z_0^{(t)}$, the error probability reads

\[ \mathbb{P}(\hat{z}_0^{(t)} \neq z_0^{(t)}) = (1 - \rho) \cdot \mathbb{P}(\hat{z}_0^{(t)} = -1 | z_0^{(t)} = +1) + \rho \cdot \mathbb{P}(\hat{z}_0^{(t)} = +1 | z_0^{(t)} = -1), \]

which can be regarded as the $(1 - \rho) \times$ type-I error $+ \rho \times$ type-II error of testing $H_0 : z_0^{(t)} = +1$ v.s. $H_1 : z_0^{(t)} = -1$ in a vanilla two-block SBM, which is almost equivalent to (2.6). This intuition is formalized by the following lemma.

**Lemma 2.3** (The fundamental testing problem for individualized estimation, Part I). Assume there exist constants $c_1, c_2 \in (0, 1)$ such that $\rho \leq 1/2 - c_1$ and $\log L \ll n^{c_2}$. Then there exists sequence $\delta_n = o(1)$ satisfying $(1 + \delta_n)n/2 \in \mathbb{N}$, such that for any $\ell \in [L]$,

\begin{equation}
 \inf \sup_{\hat{z}^{(t)} \in \mathbb{P}_n} \mathbb{E}\mathcal{L}(\hat{z}^{(t)}, z^{(t)}) \geq \inf_{\phi} \left( (1 - \rho) \cdot \mathbb{E}_{H_0}[\phi] + \rho \cdot \mathbb{E}_{H_1}[1 - \phi] \right),
\end{equation}

where $\phi$ is a testing function of the following problem:

\begin{equation}
 H_0 : \bigotimes_{i=1}^{(1+\delta_n)n/2} \text{Bern}(p_\ell) \otimes \text{Bern}(q_\ell) \quad \text{v.s.} \quad H_1 : \bigotimes_{i=1}^{(1+\delta_n)n/2} \text{Bern}(q_\ell) \otimes \text{Bern}(p_\ell).
\end{equation}

**Proof.** See Section B.3 of the supplementary material [12].

Recall that Lemma 2.3 reflects the situation when $z_0^*$ is known to us. In practice, we need to estimate $z_0^*$ from the data, giving rise to a testing problem similar to the one presented in Lemma 2.1.

**Lemma 2.4** (The fundamental testing problem for individualized estimation, Part II). Assume there exist constants $c_1, c_2 \in (0, 1)$ such that $\rho \leq 1/2 - c_1$ and $\log L \ll n^{c_2}$. Then for any sequence $\delta_n = o(1)$, there exists another sequence $\delta'_n = o(1)$ satisfying $(1 + \delta'_n)n/2 \in \mathbb{N}$, such that for any $\ell \in [L]$ and any $S \subseteq [L] \setminus \{\ell\}$, we have

\begin{equation}
 \inf \sup_{\hat{z}^{(t)} \in \mathbb{P}_n} \mathbb{E}\mathcal{L}(\hat{z}^{(t)}, z^{(t)}) \geq \delta'_n \inf_{\phi} \left( \mathbb{E}_{H_0}[\phi] + \mathbb{E}_{H_1}[1 - \phi] \right),
\end{equation}

where $\phi$ is a testing function of the following problem:

\begin{equation}
 H_0 : \bigotimes_{r \in (S \cup \{\ell\})^c} [2\text{Bern}(1 - \rho) - 1] \otimes \bigotimes_{r \in S \cup \{\ell\}} \text{Bern}(p_\ell) \otimes \text{Bern}(q_\ell) \quad \text{v.s.}\quad H_1 : \bigotimes_{r \in (S \cup \{\ell\})^c} [2\text{Bern}(\rho) - 1] \otimes \bigotimes_{r \in S \cup \{\ell\}} \text{Bern}(q_\ell) \otimes \text{Bern}(p_\ell).
\end{equation}

**Proof.** See Section B.4 of the supplementary material [12].
The testing problem in (2.20) differs from the one in (2.8) in that the layer $\ell$ is never involved in the term regarding $\text{Bern}(1-\rho)$ and $\text{Bern}(\rho)$. This makes sense, because according to our intuition in Section 2.1, this term reflects the case when $z^{(\ell)}$ is (nearly) known to us, which can never happen since $z^{(\ell)}$ itself is the estimating target.

To characterize the optimal testing error for the two testing problems given in Lemmas 2.3 and 2.4, apart from the SNR for global estimation $I_S$ defined in (1.8), we additionally define the corresponding SNR for individualized estimation:

$$J_S := \begin{cases} |S|J_\rho + \psi_\delta(0) & \text{if } |S| \text{ is even,} \\ (|S|+1)J_\rho + \psi_\delta(-2J_\rho) & \text{if } |S| \text{ is odd.} \end{cases}$$

A careful analysis on the error incurred by the likelihood ratio test gives the following result.

**Lemma 2.5** (Optimal testing error for individualized estimation). Assume $\rho = o(1)$ and that there exist constants $C_1, C_2 > 1$, $c \in (0, 1)$ such that $C_1 q_\ell \leq p_\ell \leq (C_2 q_\ell) \wedge (1-c), \forall \ell \in [L]$. Then there exists a sequence $\delta_n = o(1)$ such that for any $\ell \in [L], S \subseteq [L] \setminus \{\ell\}$, the optimal $(1-\rho) \times \text{type-I error} + \rho \times \text{type-II error}$ of the testing problem in (2.18) is lower bounded by

$$C \cdot \exp \left\{ - (1 + \delta_n)J_{(\ell)} \right\},$$

and the optimal type-I plus type-II error of the testing problem in (2.20) is lower bounded by

$$C \cdot \exp \left\{ - (1 + \delta_n)I_{S \cup \{\ell\}} \right\},$$

where $C > 0$ is an absolute constant.

**Proof.** See Section B.5 of the supplementary material [12].

For the same reason as explained in Remark 2.2, the optimal testing error for (2.20) depends on the parity of $|S \cup \{\ell\}|$.

2.2.2. Minimax lower bound for individualized estimation. We are now ready to present the main result in this subsection.

**Theorem 2.2** (Minimax lower bound for individualized estimation). Under the same setup as Theorem 2.1, if for a fixed $\ell \in [L]$, it holds that $\min_{S \subseteq [L] \setminus \{\ell\}} I_{S \cup \{\ell\}} \wedge J_{(\ell)} \rightarrow \infty$, then there exists a sequence $\delta_n = o(1)$, independent of $\ell$, such that

$$\inf_{\hat{z}} \sup_{z^* \in \mathcal{P}_n} \mathbb{E}\mathcal{L}(\hat{z}^{(\ell)}, z^{(\ell)}) \geq \exp \left\{ - (1 + \delta_n) \min_{S \subseteq [L] \setminus \{\ell\}} I_{S \cup \{\ell\}} \right\} + \exp \left\{ - (1 + \delta_n)J_{(\ell)} \right\}.$$  

On the other hand, if $\min_{S \subseteq [L] \setminus \{\ell\}} I_{S \cup \{\ell\}} \wedge J_{(\ell)} = O(1)$, then there exists $c' > 0$ such that

$$\inf_{\hat{z}} \sup_{z^* \in \mathcal{P}_n} \mathbb{E}\mathcal{L}(\hat{z}^{(\ell)}, z^{(\ell)}) \geq c'.$$

**Proof.** Given Lemma 2.3, 2.4 and 2.5, the proof is essentially the same as the proof of Theorem 2.1, and we omit the details.
Under homogeneity, we have \( z^{(\ell)} = z^* \), \( \forall \ell \in [n] \), and the lower bound in the above theorem should coincide with (2.14). Indeed, when \( \rho = 0 \), the only way to make the exponent finite is to choose \( \tilde{S} = [L] \setminus \{\ell\} \), in which case we have
\[
\min_{S \subseteq [L] \setminus \{\ell\}} I_{S \cup \{\ell\}} \wedge J_{\{\ell\}} = I_{[L]} \wedge J_{\{\ell\}} = \psi_{[L]}^*(0) \wedge \infty = \psi_{[L]}^*(0),
\]
and hence Corollary 2.1 can be alternatively derived from Theorem 2.2.

Based on the intuitions built from Lemmas 2.3 and 2.4, the interpretations of the two terms in the lower bound (2.24) should be clear: \( J_{\{\ell\}} \) is the error incurred by the label sampling model (2.16), which we cannot avoid even if we know the ground truth \( z^* \), whereas \( \min_{S \subseteq [L] \setminus \{\ell\}} I_{S \cup \{\ell\}} \) represents the error incurred by empirically estimating \( z^* \).

3. A Two-Stage Algorithm. Recall that the IMLSBM is a hierarchical model, where the individual assignments \( z^{(\ell)} \)'s are independent realizations from the “prior” distribution (1.1) which is parametrized by \( z^* \) and \( \rho \). We start by writing down the posterior density, which is proportional to
\[
\prod_{\ell \in [L]} \prod_{i \in [n]} (1 - \rho)^{1 \{z_{i}^{(\ell)} = z_{i}^*\}} \rho^{1 \{z_{i}^{(\ell)} \neq z_{i}^*\}} \times \prod_{\ell \in [L]} \prod_{i < j} \left( p_{\ell}^{A_{ij}^{(\ell)}} (1 - p_{\ell})^{1 - A_{ij}^{(\ell)}} \cdot I \{z_{i}^{(\ell)} = z_{j}^{(\ell)}\} \right) \nonumber
\]
\[
+ q_{\ell}^{A_{ij}^{(\ell)}} (1 - q_{\ell})^{1 - A_{ij}^{(\ell)}} \cdot I \{z_{i}^{(\ell)} \neq z_{j}^{(\ell)}\}.
\]

Computing the vanilla MAP estimator requires searching over a discrete set with cardinality \( 2^n n^{L+1} \), a hopeless task for even moderately-sized \( n \) and \( L \).

Now, supposed that for a fixed \( i \in [n] \), we are given a collection of estimators \( \{\tilde{z}_{j}^{(\ell)} : \ell \in [L], j \neq i\} \) for the individual assignments \( \{z_{j}^{(\ell)} : \ell \in [L], j \neq i\} \). On the event that \( \tilde{z}_{j}^{(\ell)} \)'s \( j \neq i \) agree with the ground truth parameters, the posterior density given in (3.1), as a function of \( (z_{i}^{*}, \{z_{i}^{(\ell)}\}_{\ell=1}^{L}) \), reduces to a constant multiple of
\[
\prod_{\ell \in [L]} (1 - \rho)^{1 \{z_{i}^{(\ell)} = z_{i}^*\}} \rho^{1 \{z_{i}^{(\ell)} \neq z_{i}^*\}} \times \prod_{\ell \in [L]} \prod_{j \neq i} \left( p_{\ell}^{A_{ij}^{(\ell)}} (1 - p_{\ell})^{1 - A_{ij}^{(\ell)}} \cdot I \{z_{i}^{(\ell)} = \tilde{z}_{j}^{(\ell)}\} \right) \nonumber
\]
\[
+ q_{\ell}^{A_{ij}^{(\ell)}} (1 - q_{\ell})^{1 - A_{ij}^{(\ell)}} \cdot I \{z_{i}^{(\ell)} \neq \tilde{z}_{j}^{(\ell)}\}.
\]

With some algebra, one finds that maximizing the above display over \( (z_{i}^{*}, \{z_{i}^{(\ell)}\}_{\ell=1}^{L}) \) is equivalent to maximizing the following objective function:
\[
\sum_{\ell \in [L]} \left\{ \log \left( \frac{1 - \rho}{\rho} \right) \cdot I \{z_{i}^{(\ell)} = z_{i}^*\} + \sum_{j \neq i, j \neq i} \left[ \log \left( \frac{p_{\ell}(1 - q_{j})}{q_{j}(1 - p_{j})} \right) A_{ij}^{(\ell)} + \log \left( \frac{1 - p_{j}}{1 - q_{j}} \right) \right] \right\}.
\]

This is already simpler than the original one of maximizing (3.1), because the search space now has cardinality \( 2^{L+1} \ll 2^{n(L+1)} \).
A closer look at (3.3) reveals that this function can be maximized in linear (in $L$) time. Indeed, if we fix $z_i^*$, the problem of seeking for optimal $z^{(ℓ)}$’s is decoupled into $L$ subproblems. That is, it suffices to maximize

$$
\log \left( \frac{1 - r}{r} \right) \cdot \mathbb{1}\{z_i^{(ℓ)} = z_i^{(⋆)}\} + \sum_{j \neq i, z_j^{(ℓ)} = z_i^{(ℓ)}} \left[ \log \left( \frac{p_ℓ(1 - q_ℓ)}{q_ℓ(1 - p_ℓ)} \right) A_{ij}^{(ℓ)} + \log \left( \frac{1 - p_ℓ}{1 - q_ℓ} \right) \right]
$$

for each $ℓ \in [L]$. Note that each subproblem can be efficiently solved, since one only needs to search over a space with cardinality two (i.e., $z_i^{(ℓ)} \in \{±1\}$). Thus, to obtain the global maximizer of (3.3), one can proceed as follows:

1. Solve $L$ subproblems (3.4) with $z_i^* = +1$, and record the objective value of (3.3);
2. Repeat Step 1 with $z_i^* = -1$;
3. Obtain the global maximizer of (3.3) by comparing the two objective values in the previous two steps.

The foregoing discussion shows that the MAP estimator of $(z_i^*, \{z_i^{(ℓ)}\}_{ℓ=1}^L)$ can be efficiently computed, provided the remaining parameters are given. This observation motivates the main algorithm of this paper, which is a two-stage procedure that first obtains initial estimators of $\{z^{(ℓ)}\}_{ℓ=1}^L$ via spectral clustering, and then refines the initial estimators in a node-wise fashion using MAP estimation.

### 3.1. Stage I: Initialization via Spectral Clustering

While our analysis in Section 4 reveals that any consistent initialization would work, we will focus on a specific initialization scheme in this subsection: spectral clustering.

If $ρ$ is of order $o(1)$, then the proportion of flips in $z^{(ℓ)}$ from $z^*$ will also be of order $o(1)$ with high probability. Hence, as long as a consistent estimator $\tilde{z}^*$ of the global assignment $z^*$ is given, consistent individualized estimation is automatic by setting $\tilde{z}^{(ℓ)} = \tilde{z}^*$. In the rest of this subsection, we restrict ourselves to global estimation.

Let $ω = (ω_1, \ldots, ω_L)$ be an arbitrary positive (i.e., $ω_ℓ > 0, \forall ℓ ∈ [L]$) weight vector, and let us consider the following weighted adjacency matrix

$$
A := \sum_{ℓ ∈ [L]} ω_ℓ A^{(ℓ)}.
$$

In the case of $ρ = 0$, one readily checks that all the information in $z^*$ is contained in the top two eigenvectors of $E[A]$. A natural proposal is then to take top two eigenvectors of $\tilde{A}$, and apply $k$-means clustering to them.

In the case of a small $ρ > 0$, we expect spectral clustering to continue to work well for estimating $z^*$, provided it exhibits a certain level of stability to the additional “noise” induced by $ρ$. Our later analysis in Section 4.1 shows that this is indeed the case.

The overall initialization scheme is detailed in Algorithm 1. There are two subtleties in this algorithm. First, instead of applying spectral clustering to $A$, we apply it to a trimmed version, $\tau(\tilde{A})$, which is obtained by setting the “larger-than-average” entries of $\tilde{A}$ to zero. As shown in Section 4.1, such a trimming operation can significantly improve the concentration of $\bar{A}$, especially when the signal-to-noise ratio is low. When $\{p_ℓ\}_{ℓ=1}^L$ are unknown, one can replace them with conservative estimators; see Section F.4 of the supplementary material [12] for details. Alternatively, one can replace them with the sample average connecting probabilities but with a larger $γ$. The second subtlety is a computational one: since exactly solving the
Specifically, letting $U$ be the clustering induced by $\tilde{z}$, we propose to solve the $k$-means objective (3.6) on $U$ to get $\hat{Z}$ and set

$$\hat{Z}_i^* = \mathbb{1}\{\hat{Z}_{i,1} = 1\} - \mathbb{1}\{\hat{Z}_{i,2} = 1\}$$

for any $i \in [n]$;

5. return $\hat{z}^*$;

Algorithm 1: Stage I: initialization via spectral clustering

**Input:** Adjacency matrices $\{A^{(\ell)}\}_1^L$, weight vector $\omega$, intra-cluster connecting probabilities $\{p_{\ell}\}_1^L$, trimming parameter $\gamma > 1$  

**Output:** Initial global estimator $\hat{z}^*$

1. Identify nodes $I \subseteq [n]$ such that $\sum_{j \in [n]} \tilde{A}_{ij} > \gamma n \sum_{\ell \in [L]} \omega_{\ell} p_{\ell}, \forall i \in I$;
2. For any $i \in (I \times [n]) \cup ([n] \times I)$, set $A_{ij}$ to zero and call the resulting matrix $\tau(\tilde{A})$;
3. Compute $U \in \mathbb{R}^{n \times 2}$, the first two eigenvectors of $\tau(\tilde{A})$;
4. Solve the $(1 + \varepsilon)$-approximate $k$-means objective (3.6) on $U$ to get $\hat{Z}$ and set

$$\hat{Z}_i^* = \mathbb{1}\{\hat{Z}_{i,1} = 1\} - \mathbb{1}\{\hat{Z}_{i,2} = 1\}$$

$k$-means objective is NP-hard, we instead find the solution of an $(1 + \varepsilon)$-approximation of it [40]. Specifically, letting $U \in \mathbb{R}^{n \times 2}$ be the top two eigenvectors of $\tau(\tilde{A})$, we seek for $(\hat{Z}, \hat{X})$ such that

$$\|\hat{Z}\hat{X} - U\|_F^2 \leq (1 + \varepsilon) \min_{\hat{Z}, \hat{X}} \|Z\hat{X} - U\|_F^2,$$

where the the minimum is over all $n \times 2$ assignment matrix $Z$ (i.e., each row of $Z$ is a canonical basis of $\mathbb{R}^2$) and all $2 \times 2$ matrix $X$. The initial estimator $\hat{z}^* = \hat{z}^{(\ell)}$ is then taken to be the clustering induced by $\hat{Z}$.

3.2. Stage II: Node-Wise Refinement via MAP Estimation. According to our discussion in the previous subsection, once an initial global estimator $\hat{z}^*$ is given, we can also take that to be the initial individualized estimator. Now, in view of the MAP objective functions (3.1)—(3.4), we propose to solve

$$(\hat{z}_i^*, \hat{z}_i^{(1)}, \ldots, \hat{z}_i^{(L)}) = \arg \max_{s_\ell \in \{\pm 1\}, \forall \ell \in [L]} \sum_{s_\ell \in \{\pm 1\}} f_i^{(\ell)}(s_\ell),$$

where

$$f_i^{(\ell)}(s_\ell, \hat{z}_i^*) = \log \left(\frac{1 - \rho}{\rho}\right) \cdot \mathbb{1}\{s_\ell = s_\nu\} + \sum_{j \neq i : \hat{z}_j^* = s_\nu} \left[ \log \left(\frac{p_{\ell}(1 - q_{\ell})}{q_{\ell}(1 - p_{\ell})}\right) A_{ij}^{(\ell)} + \log \left(\frac{1 - p_{\ell}}{1 - q_{\ell}}\right) \right],$$

By our discussion at the beginning of this section, the above optimization problem can be solved in linear time. A detailed description is given in Algorithm 2. When $(p_{\ell}, q_{\ell})_{\ell=1}^L$ are unknown, one can replace them with their estimators; see Section A.3 of the supplementary material [12] for details.

We conclude this section by remarking that our proposed algorithm is naturally a distributed one: the two for loops in Algorithm 2 can be easily parallelized.

4. Performance of the Two-Stage Algorithm. In this section, we present theoretical results on the two-stage algorithm introduced in Section 3. Specifically, the performance of spectral clustering is presented in Section 4.1, followed by an analysis of MAP-based refinement in Section 4.2. The minimax optimality of the two-stage algorithm is proved in Section 4.3. Throughout this section, the high probability error bounds are uniform with
respect to probability measures defined in the parameter space (2.1). In particular, the "$P$" symbol represents the probability after marginalizing over the realizations of the $z^{(\ell)}$’s.

4.1. Performance of Spectral Clustering. In this subsection, we analyze theoretical properties of Algorithm 1. An important degree of freedom in Algorithm 1 is the choice of the weight vector $\omega$, and it is restricted by the following assumption.

**ASSUMPTION A (Balanced weights across layers).** Assume $\omega_{\ell} > 0, \forall \ell \in [L]$ and $\sum_{\ell \in [L]} \omega_{\ell} = 1$. Moreover, assume that there exist two absolute constants $c_0 > 0$ and $c_1 \geq 1$ such that the following two inequalities hold:

$$\max_{\ell \in [L]}\{\omega_{\ell}\} \cdot \sum_{\ell \in [L]} \omega_{\ell} p_{\ell} \leq c_0 \sum_{\ell \in [L]} \omega_{\ell}^2 p_{\ell}, \quad \max_{\ell \in [L]}\{\omega_{\ell}\} \cdot \sum_{\ell \in [L]} p_{\ell} \leq c_1 \sum_{\ell} \omega_{\ell} q_{\ell}. \tag{4.1}$$

The above assumption essentially states that $\omega$ should be relatively balanced across layers. In particular, if $\omega_{\ell} \approx 1/L, \forall \ell \in [L]$, then this assumption holds.

We are now ready to state the main theorem of this subsection. We emphasize that the following theorem does not require $\rho = o(1)$ or $\beta = 1 + o(1)$.

**THEOREM 4.1 (Performance of spectral clustering).** Let Assumption A hold with $c_0 > 0, c_1 \geq 1$. Let the input to Algorithm 1 be an instance generated by an IMLSBM $\mathcal{P}_n(\rho, \{p_{\ell}\}_1^L, \{q_{\ell}\}_1^L, \beta)$ and assume

$$\frac{\bar{p}}{\bar{p} - \bar{q}} \leq c \cdot \frac{2(1 - 2\rho)^2}{\beta - \beta^{-1} + 4n^{-1} + 2(\rho - \rho^2)}, \tag{4.2}$$

where $c \in [0, 1)$ is an absolute constant and $\bar{p} = \sum_{\ell \in [L]} \omega_{\ell} p_{\ell}, \bar{q} = \sum_{\ell \in [L]} \omega_{\ell} q_{\ell}$ are the weighted averages of connecting probabilities. Fix any $r \geq 1$ and choose the regularization parameter to be $\gamma = c^{c_1}$. Then, there exist constants $c_2 = c_2(\gamma)$ and $C = C(c, c_0, c_1, r)$ such that with probability at least $1 - 11n^{-r} - c_2^{-n}$, the output $\hat{z}^*$ of Algorithm 1 satisfies

$$\mathcal{L}(\hat{z}^*, z^*) \leq \frac{C(2 + \varepsilon)(\Delta_1 + \Delta_2)}{n^2(1 - 2\rho)^4(\bar{p} - \bar{q})^2}, \tag{4.3}$$

**Algorithm 2: Stage II: Node-wise refinement via MAP estimation**

**Input:** Initial global estimator $\hat{z}^*$, adjacency matrices $\{A^{(\ell)}\}_1^L$, flipping probability $\rho$

**Output:** Global estimator $\hat{z}^*$, individualized estimator $\{\hat{z}^{(\ell)}\}_1^L$

1. for $i = 1, \ldots, n$
2.    for $\ell = 1, \ldots, L$
3.        $z(\ell, +1) \leftarrow \arg\max_{s \in \{\pm 1\}} f_i^{(\ell)}(+1, s, \hat{z}^*); \quad // \ f_i^{(\ell)}$ is defined in (3.8)
4.        $z(\ell, -1) \leftarrow \arg\max_{s \in \{\pm 1\}} f_i^{(\ell)}(-1, s, \hat{z}^*);$
5.        $\hat{z}_i^{(\ell)} \leftarrow \arg\max_{s \in \{\pm 1\}} \sum_{\ell \in [L]} f_i^{(\ell)}(s, z(\ell, s), \hat{z}^*); \quad // \ \text{final global estimator}$
6.    for $\ell = 1, \ldots, L$
7.        $\hat{z}_i^{(\ell)} \leftarrow z(\ell, \hat{z}^*); \quad // \ \text{final individualized estimators}$
8. return $\hat{z}^*, \{\hat{z}^{(\ell)}\}_1^L$
where
\begin{equation}
\Delta_1 = n \sum_{\ell \in [L]} \omega_{\ell}^2 p_{\ell},
\end{equation}
\begin{equation}
\Delta_2 = \max_{\ell \in [L]} \{ \omega_{\ell}^2 (p_{\ell} - q_{\ell})^2 \} \cdot \left[ L^2 \rho^2 + n^2 L \rho + n \log n + (\log L)^2 \right] \\
+ \rho^2 (1 - \rho)^2 (\bar{p} - \bar{q})^2.
\end{equation}

**Proof.** See Section C.1 of the supplementary material [12].

**Remark 4.1.** Under our working assumption that \( \beta = 1 + o(1) \), the first term on the right-hand side of (4.2) tends to infinity, and so the inequality holds if \( p_{\ell} / (p_{\ell} - q_{\ell}) \) is uniformly bounded for all \( \ell \in [L] \).

In the upper bound (4.3), the two terms \( \Delta_1 \) and \( \Delta_2 \) come from the fact that in our proof, we relate the misclustering error to the deviation (in spectral norm) of the trimmed weighted adjacency matrix \( \tau(\bar{A}) \) from the expectation of \( \bar{A} \). More explicitly, \( \Delta_1 \) is induced by the concentration of \( \tau(\bar{A}) \) around \( \mathbb{E}[\bar{A} \mid \{z^{(\ell)}\}_1^L] \), the conditional mean of \( \bar{A} \), conditioning on the realization of \( z^{(\ell)} \)'s, whereas \( \Delta_2 \) is induced by the concentration of \( \mathbb{E}[\bar{A} \mid \{z^{(\ell)}\}_1^L] \) around the marginal mean \( \mathbb{E}[\bar{A}] \).

It turns out that bounding \( \| \tau(\bar{A}) - \mathbb{E}[\bar{A}] \| \) is closely related to bounding \( \| \tau(\bar{B}) - \mathbb{E}[\bar{B}] \| \), where \( \bar{B} = \sum_{\ell \in [L]} \omega_{\ell} B^{(\ell)} \) and \( B^{(\ell)} \)'s are \( n \times n \) independent Bernoulli random matrices with independent \( \text{Bern}(p_{ij}^{(\ell)}) \) entries. In order to have a tight control of \( \| \bar{B} - \mathbb{E}[\bar{B}] \| \), we give a non-trivial generalization of the results in [42] to the multilayer setup in Appendix D, which roughly states the following: if the weight vector is sufficiently “balanced”, then with high probability, for the trimmed version of \( \bar{B} \), we have
\begin{equation}
\| \tau(\bar{B}) - \mathbb{E}[\bar{B}] \| \lesssim \sqrt{n \sum_{\ell \in [L]} \omega_{\ell}^2 \max_{i,j} p_{ij}^{(\ell)}},
\end{equation}
and without any trimming operation, we have
\begin{equation}
\| \bar{B} - \mathbb{E}[\bar{B}] \| \lesssim \sqrt{n \sum_{\ell \in [L]} \omega_{\ell}^2 \max_{i,j} p_{ij}^{(\ell)} + \max_{\ell \in [L]} \{ \omega_{\ell} \}} \cdot \sqrt{\log n}.
\end{equation}

If \( \rho = 0 \), then the conditional mean \( \mathbb{E}[\bar{A} \mid \{z^{(\ell)}\}_1^L] \) coincides with the marginal mean \( \mathbb{E}[\bar{A}] \), and thus (4.3) holds with \( \Delta_2 = 0 \).

Curious readers may wonder why the expression of \( \Delta_2 \) given in (4.5) does not vanish as \( \rho \) tends to zero. In particular, there is an additive term of \( n \log n \). This is related to an interesting phenomenon regarding the concentration of Bernoulli random matrices. It happens that the problem of bounding \( \| \mathbb{E}[\bar{A} \mid \{z^{(\ell)}\}_1^L] - \mathbb{E}[\bar{A}] \| \) can be related to bounding \( \| B - \mathbb{E}[B] \| \), where \( B \in \{0,1\}^{n \times n} \) has i.i.d. \( \text{Bern}(\rho) \) entries. The expression of \( \Delta_2 \) in (4.5) is based on that \( \| B - \mathbb{E}[B] \| \lesssim \sqrt{n \rho} + \sqrt{\log n} \) with high probability, which does not vanish as \( \rho \) tends to zero. In fact, such a “discontinuity at zero” is unavoidable: it has been shown in [38] that if \( \rho \lesssim 1/n \), then with probability tending to one, \( \| B - \mathbb{E}[B] \| = (1 + o(1)) \sqrt{\log n / \log \log n} \), which diverges as \( n \) tends to infinity.

To have a better understanding on the magnitude of the bound (4.3), let us choose \( \omega_{\ell} = 1/L, \forall \ell \in [L] \) and consider the following scaling of the connecting probabilities:
\begin{equation}
p_{\ell} = \frac{a \log n}{nL}, \quad q_{\ell} = \frac{b \log n}{nL}, \quad \forall \ell \in [L],
\end{equation}

where
where \( a > b > 0 \) are two constants. With some algebra, it follows that

\[
\Delta_1 = \frac{a \log n}{L^2}, \quad \Delta_2 = \frac{(a - b)^2 (\log n)^2}{n^2 L^2} \cdot \left( \rho^2 + \frac{n^2 \rho}{L} + \frac{n \log n}{L} + \frac{(\log L)^2}{L^2} \right).
\]

It is clear that \( \Delta_1 + \Delta_2 \ll \frac{\log n}{L^2} \cdot (1 + \frac{\rho \log n}{L}) \), and from (4.3) we arrive at

\[
(4.9) \quad \mathcal{L}(\tilde{z}^*, z^*) \leq \frac{1}{(1 - 2\rho)^3(a - b)^2} \cdot \left( \frac{1}{\log n} + \frac{\rho}{L} \right).
\]

In summary, consistent estimation of \( z^* \) by spectral clustering is possible when the connecting probabilities are as small as \( \Omega\left(\frac{\log n}{L^2}\right) \).

Since \( \mathcal{L}(\tilde{z}^*, z^{(\ell)}) \leq \mathcal{L}(\tilde{z}^*, z^*) + \mathcal{L}(z^*, z^{(\ell)}) \), we can obtain performance guarantees of Algorithm 1 for individualized estimation by bounding the number of flips at each layer, as detailed in the following corollary.

**Corollary 4.1 (Spectral clustering for individualized estimation).** Under the setup of Theorem 4.1, for any \( \rho' \in (0, 1 - \rho) \), with probability at least \( 1 - 11n^{-r} - c_2^{-n} - Le^{-nD_{KL}(\rho\|\rho')}, \) we have

\[
(4.10) \quad \max_{\ell \in [L]} \mathcal{L}(\tilde{z}^*, z^{(\ell)}) \leq \frac{C(2 + \varepsilon)(\Delta_1 + \Delta_2)}{n^2(1 - 2\rho)^3(\bar{p} - \bar{q})^2} + \rho + \rho'.
\]

where \( D_{KL}(p\|q) \) is the Kullback-Leibler divergence between Bernoulli and Bernoulli random variables.

**Proof.** Classical Chernoff–Hoeffding bound [14, 31] gives that for each fixed \( \ell \in [L] \),

\[
(4.11) \quad \mathbb{P}\left( \mathcal{L}(z^*, z^{(\ell)}) \geq \rho + \rho' \right) \leq e^{-nD_{KL}(\rho\|\rho')},
\]

and the desired result follows by taking a union bound over \( \ell \in [L] \). \(\square\)

It is well-known that (4.11), which is obtained by computing the rate function of Bernoulli random variables, is asymptotically tight (see, e.g., [65]). If the goal is merely to ensure consistency (i.e., \( \rho' = o(1) \)), then we can use the following standard weakening of (4.11):

\[
\mathbb{P}\left( \mathcal{L}(z^*, z^{(\ell)}) \geq \rho + \rho' \right) \leq e^{-2n(\rho')^2}.
\]

If \( \log L \ll n^c \) for some \( c \in (0, 1) \), then we can choose \( \rho' = n^{-(1-c)/2} \), so that (4.10) holds with \( \rho' = o(1) \) with probability at least \( 1 - 11n^{-r} - c_2^{-n} - e^{-2n^c + \log L} \), and one can bound \( e^{-2n^c + \log L} \leq e^{-cn^c} \) for some absolute constant \( c_3 > 0 \) because \( \log L \ll n^c \).

### 4.2. Performance of MAP-Based Refinement.

The refinement procedure as introduced in Section 3.2, in its current form (Algorithm 2), is highly flexible in that no assumption is imposed on the initial estimator \( \tilde{z}^* \) other than consistency. While such a flexibility is favored in practice, it brings some unnecessary complications to its theoretical analysis. In addition, for a fixed \( i \), the initial estimators \( \tilde{z}^* \) may have arbitrary dependence structures with \( \{A^{(\ell)}_{ij} : j \neq i, \ell \in [L]\} \), which makes the analysis intractable.

To facilitate the analysis, we propose a modified version as shown in Algorithm 3. Instead of taking an arbitrary initial estimator as input (as done in Algorithm 2), we consider a leave-one-out initialization scheme. In Stage I, for each fixed \( i \), the initial estimator \( \tilde{z}^{(*,i)} \) of \( z^*_i \) are computed using only \( \{A^{(\ell)}_{-i,j} \} : j \neq i, \ell \in [L]\} \), which ensures the conditional
Algorithm 3: A provable version of Algorithm 2

Input: Adjacency matrices \( \{A^{(\ell)}\}_1^L \), connection probabilities \( \{(p_{\ell}, q_{\ell})\}_1^L \), flipping probability \( \rho \)

Output: Global estimator \( \hat{z}^* \), individualized estimator \( \{\hat{z}^{(\ell)}\}_1^L \)

1 /* Stage I: Leave-one-out initialization */
2 for \( i = 1, \ldots, n \) do
3 \( \hat{z}_{i}^{(s,-i)} \leftarrow 0 \); 
4 \( \hat{z}_{i}^{(s,-i)} \leftarrow \text{Initialize}(\{A_{i-i}^{(\ell)}, p_{\ell}, q_{\ell}\}_{\ell=1}^L); /\text{ initial \ estimator \ of} \ \hat{z}_{i}^* \)
5 Set \( \hat{z}_{i}^{(\ell,-i)} \leftarrow \hat{z}_{i}^{(s,-i)}, \forall \ell \in [L]; 
6 /* Stage II: MAP-based refinement */
7 for \( i = 1, \ldots, n \) do
8 \( \text{for } \ell = 1, \ldots, [L] \text{ do} \)
9 \( z(\ell, +1) \leftarrow \text{argmax}_{s \in \{\pm 1\}} f_i^{(\ell)} (+1, s, \hat{z}_{i}^{(s,-i)}); \)
10 \( z(\ell, -1) \leftarrow \text{argmax}_{s \in \{\pm 1\}} f_i^{(\ell)} (-1, s, \hat{z}_{i}^{(s,-i)}); \)
11 // final estimator of \( \hat{z}_{i}^* \), not aligned
12 \( \hat{z}_{i}^{(s,-i)} \leftarrow \text{argmax}_{s \in \{\pm 1\}} \sum_{\ell \in [L]} f_i^{(\ell)} (s, z(\ell, s), \hat{z}_{i}^{(s,-i)}); \)
13 \( \text{for } \ell = 1, \ldots, [L] \text{ do} \)
14 \( \hat{z}_{i}^{(\ell,-i)} \leftarrow z(\ell, \hat{z}_{i}^{(s,-i)}); // \text{ final \ estimator \ of} \ z_{i}^{(\ell)}, \text{ not \ aligned} \)
15 /* Stage III: Alignment */
16 Set \( \hat{z}_{1}^* = \hat{z}_{i}^{(s,-i)} \) and \( \hat{z}_{1}^{(\ell)} = \hat{z}_{i}^{(\ell,-i)}, \forall \ell \in [L]; 
17 \text{for } i = 2, \ldots, n \text{ do} \)
18 \( \hat{z}_{i}^* = \text{argmax}_{s \in \{\pm 1\}} \# \left\{ \{j \in [n]: \hat{z}_{j}^{(s,-i)} = s \} \cap \{j \in [n]: \hat{z}_{j}^{(s,-i)} = \hat{z}_{i}^{(s,-i)} \} \right\} ; \)
19 \( \text{for } \ell = 1, \ldots, L \text{ do} \)
20 \( \hat{z}_{i}^{(\ell)} = \text{argmax}_{s \in \{\pm 1\}} \# \left\{ \{j \in [n]: \hat{z}_{j}^{(\ell,-i)} = s \} \cap \{j \in [n]: \hat{z}_{j}^{(\ell,-i)} = \hat{z}_{i}^{(\ell,-i)} \} \right\} ; \)
21 return \( \hat{z}^*, \{\hat{z}^{(\ell)}\}_1^L \);

(on the realization of \( \{z^{(\ell)}\}) \) independence between \( \{z^{(\ell)}\} \) and \( \{A_{j-i}^{(\ell)}: j \neq i, \ell \in [L] \} \), thus simplifying the analysis, though the final analysis still turns out to be highly nontrivial.

In Stage II, for each \( i \in [n] \), we conduct MAP-based refinement using the initial estimators \( \{\hat{z}_{i-i}^{(\ell,-i)}\}_{\ell=1}^L \) (which are all equal to \( \hat{z}_{i-i}^{(s,-i)} \) of \( z_{i-i}^{(\ell)} \)) and the “diagonal slots” \( \{\hat{z}_{i-i}^{(s,-i)}, \{\hat{z}_{i-i}^{(\ell,-i)}\}_{\ell=1}^L \) are all zeros before Stage II by our construction. These “diagonal slots” are then filled in by the refined estimators of \( (z_{i-i}^{*}, \{z_{i-i}^{(\ell)}\}_{\ell=1}^L) \).

After Stage II, it is tempting to directly output \( \hat{z}_{j}^{*} = \hat{z}_{j}^{(s,-i)} \) and \( \hat{z}_{j}^{(\ell)} = \hat{z}_{j}^{(\ell,-i)} \) as the final estimators. However, a subtlety arises due to the leave-one-out initialization. Since the initial estimators \( \{\hat{z}_{i-i}^{(s,-i)}\}_{i=1}^n \) are not necessarily aligned, the refined estimators \( \{\hat{z}_{i-i}^{(s,-i)}\}_{i=1}^n \) and \( \{\hat{z}_{i-i}^{(\ell,-i)}: i \in [n], \ell \in [L] \} \) can have different orientations. For example, it could happen that \( \hat{z}_{1}^{(s,-i)} \) is estimating \( z_{1}^{*} \), but \( \hat{z}_{2}^{(s,-2)} \) is estimating \( -z_{2}^{*} \). This is where the extra Stage III of Algorithm 3 comes into play. By using an alignment procedure, all coordinates of \( \hat{z}^* \) and \( \{\hat{z}^{(\ell)}\} \) will have the same orientation with high probability.

We shall remark that Algorithm 3 is mostly of theoretical interest, and similar strategies have appeared in [25, 26]. Our simulation in Section 5 indicates that the estimation accuracy
of Algorithm 2 is indistinguishable from that of Algorithm 3, while Algorithm 2 is much faster in speed. Such a near perfect match in accuracy between the two algorithms is itself an interesting phenomenon, which we leave for future work. A promising approach for analyzing Algorithm 2 is the “leave-one-out” analysis such as that used in [47].

Before we present the main result of this subsection, we introduce the following assumption on consistent initialization.

**ASSUMPTION B (Consistent initialization).** Assume the Initialize procedure used in Algorithm 3 takes an instance generated by an IMLSBM \( P_n(\rho, \{p_l\}_1^L, \{q_l\}_1^L, \beta) \) as its input and outputs a \( \tilde{z}^* \) satisfying

\[
\mathbb{P}(\mathcal{L}(\tilde{z}^*, z^*) \geq \delta_{\text{init},n}) \lesssim n^{-(1+\varepsilon_{\text{init}})}
\]

for some \( \delta_{\text{init},n} = o(1) \) and \( \varepsilon_{\text{init}} > 0 \).

4.2.1. **Performance for global estimation.** The performance of Algorithm 3 for global estimation is given by the following theorem.

**THEOREM 4.2 (Performance of MAP-based refinement for global estimation).** Let the input to Algorithm 3 be an instance generated by an IMLSBM \( P_n(\rho, \{p_l\}_1^L, \{q_l\}_1^L, \beta) \) satisfying \( \rho = o(1), q_\ell < p_\ell \leq (C q_\ell) \land (1 - c), \forall \ell \in [L], \beta = 1 + o(1) \) and \( \log L \ll n^{c'} \), where \( C > 1 \) and \( c, c' \in (0, 1) \) are absolute constants. Let Assumption B hold and assume that for any \( \delta_n = o(1) \), the following holds:

\[
\lim_{n \to \infty} \sum_{S \subseteq [L]} e^{-(1-\delta_n)\mathcal{I}_S} = 0,
\]

where \( \mathcal{I}_S \) is defined in (1.8). Then, there exist two sequences \( \bar{\delta}_n, \tilde{\delta}_n = o(1) \) such that

\[
\lim_{n \to \infty} \inf_{\tilde{z}^* \in \tilde{P}_n} \mathbb{P} \left( \mathcal{L}(\tilde{z}^*, z^*) \leq \left( \sum_{S \subseteq [L]} e^{-(1-\bar{\delta}_n)\mathcal{I}_S} \right)^{1-\tilde{\delta}_n} \right) = 1.
\]

**PROOF.** See Section C.2 of the supplementary material [12].

Note that the lower bound given in (2.12) takes the form of the maximum of \( 2^L \) terms indexed by \( S \subseteq [L] \), whereas the upper bound given in (4.14) is a summation of \( 2^L \) terms. Our later analysis in Section 4.3 shows that under slightly stronger conditions on the SNR, the upper and lower bounds match asymptotically.

Under homogeneity \( (\rho = 0) \), we have \( \mathcal{I}_S = \infty \) for every \( S \) but \( S = [L] \). So we have

\[
\sum_{S \subseteq [L]} e^{-(1-\delta_n)\mathcal{I}_S} = e^{-(1-\delta_n)\mathcal{I}_{[L]}} = \exp \left\{ - (1 - \delta_n) \frac{n}{2} \sum_{\ell \in [L]} I_{1/2}^{(\ell)} \right\},
\]

and the upper bound in (4.14) matches the lower bound provided by Corollary 2.1.

However, the derivation of (4.15) is not fully rigorous, because the layer-wise objective function \( f_\ell^{(\ell)} \) defined in (3.8) becomes infinity when \( \rho = 0 \), which makes the optimization problem in (3.7) ill-defined. To address this issue, let us note that when \( \rho = 0 \), the “regularization term” in \( f_\ell^{(\ell)} \), namely \( \log((1-\rho)/\rho) \cdot \mathbb{1}\{s_\ell = s^*_\ell\} \), essentially requires \( s_\ell \) to exactly agree with \( s^*_\ell \). Thus, we can shift from solving (3.7) to solving the following problem:

\[
\tilde{z}^*_\ell = \argmax_{s_\ell \in \{\pm 1\}} \sum_{\ell \in [L]} \sum_{j \neq i} \log \left( \frac{p_\ell(1-q_\ell)}{q_\ell(1-p_\ell)} A_{ij}^{(\ell)} \right) + \log \left( \frac{1-p_\ell}{1-q_\ell} \right).
\]
With the above optimization formulation, Algorithm 3 can be modified in a *mutatis mutandis* fashion, and the upper bound in (4.15) can be made rigorous, as detailed in the following corollary.

**Corollary 4.2 (Performance of MAP-based refinement for global estimation under homogeneity).** Consider again Algorithm 3, except that we change its Stage II from MAP-based refinement (3.7) to maximum-likelihood-based refinement (4.16). Let the input be an instance generated by an IMLSBM $P_n(\rho, \{p_\ell\}^L_{\ell=1}, \{q_\ell\}^L_{\ell=1}, \beta)$ satisfying $\rho = o(1)$, $q_\ell < p_\ell \leq (Cq_\ell) \wedge (1 - c), \forall \ell \in [L]$, $\beta = 1 + o(1)$ and $\log L \ll n^{c'}$, where $C > 1$ and $c, c' \in (0, 1)$ are absolute constants. Let Assumption B hold and assume that for a sequence $\delta_n = o(1)$ such that

\[
\lim_{n \to \infty} \inf_{z^* \in P_n} \mathbb{P} \left( \mathcal{L}(\hat{z}^*, z^*) \leq \exp \left\{ - \frac{(1 - \delta_n)n}{2} \sum_{\ell \in [L]} I_{1/2}^{(\ell)} \right\} \right) = 1.
\]

**Proof.** The proof is a straightforward adaptation of the proof of Theorem 4.2, and we omit the details. \(\square\)

Minimax optimal algorithms for community detection in a homogeneous MLSBM have appeared in the literature \([56, 68]\). The procedure in \([56]\) is based on exactly solving the maximum likelihood objective, which is computationally infeasible. The algorithm in \([68]\) is computable in polynomial-time and it operates on a variant of SBM, called weighted SBM, of which the homogeneous multilayer SBM is a special case. The corollary above gives another polynomial-time minimax optimal algorithm for community detection in homogeneous multilayer SBMs.

**4.2.2. Performance for individualized estimation.** The performance guarantee of Algorithm 3 for individualized estimation is given by the following theorem.

**Theorem 4.3 (Performance of MAP-based refinement for individualized estimation).** Let the input to Algorithm 3 be an instance generated by an IMLSBM $P_n(\rho, \{p_\ell\}^L_{\ell=1}, \{q_\ell\}^L_{\ell=1}, \beta)$ satisfying $\rho = o(1)$, $q_\ell < p_\ell \leq (Cq_\ell) \wedge (1 - c), \forall \ell \in [L]$, $\beta = 1 + o(1)$ and $\log L \ll n^{c'}$, where $C > 1$ and $c, c' \in (0, 1)$ are absolute constants. Let Assumption B hold and assume that for a fixed $\ell \in [L]$ and for any $\delta_n = o(1)$, the following holds:

\[
\lim_{n \to \infty} \inf_{S \subseteq [L] \setminus \{\ell\}} \sum_{S \subseteq [L] \setminus \{\ell\}} \left( e^{-(1 - \delta_n)I_{S \cup \{\ell\}}} + e^{-(1 - \delta_n)J_{S \cup \{\ell\}}} \right) = 0,
\]

where $I_{S \cup \{\ell\}}$ and $J_{S \cup \{\ell\}}$ are defined in (1.8) and (2.21), respectively. Then, there exist two sequences $\delta_n, \tilde{\delta}_n = o(1)$, independent of $\ell$, such that

\[
\lim_{n \to \infty} \inf_{z^* \in P_n} \mathbb{P} \left[ \mathcal{L}(\hat{z}^{(\ell)}, z^{(\ell)}) \leq \left( \sum_{S \subseteq [L] \setminus \{\ell\}} \left( e^{-(1 - \delta_n)I_{S \cup \{\ell\}}} + e^{-(1 - \delta_n)J_{S \cup \{\ell\}}} \right) \right)^{1 - \tilde{\delta}_n} \right] = 1.
\]

**Proof.** See Section C.3 of the supplementary material \([12]\). \(\square\)

In this upper bound, the terms involving $I_{S \cup \{\ell\}}$’s come from estimating $z^{(\ell)}$ given the knowledge of $z^*$ (i.e., error from the label sampling model defined in (2.16)), whereas the terms involving $J_{S \cup \{\ell\}}$’s come from empirically estimating $z^*$. 
Similar to Theorem 4.2, the bound given in this theorem is a summation of $2 \times 2^{L-1}$ terms, whereas the corresponding lower bound in (2.24) is the maximum of $2^{L-1} + 1$ terms. We will show in Section 4.3 that the two bounds asymptotically coincide under slightly stronger assumptions on both global and individualized SNRs.

If $\rho = 0$, then $I_{S \cup \{\ell\}}$ is infinity for every $S$ but $S = [L] \setminus \{\ell\}$. On the other hand, $I_{S \cup \{\ell\}}$ is infinity for any $S \subseteq [L] \setminus \{\ell\}$. It follows that the upper bound in (4.19) becomes

$$e^{-\left(1-o(1)\right)I_{[L]}} = \exp\left\{-\left(1-o(1)\right)\sum_{\ell \in [L]} I_{\ell}^{(L)}\right\},$$

which agrees with the upper bound in (4.17). This makes sense as global and individualized estimation coincide under homogeneity.

4.3. Minimax Optimality. Recall that the two upper bounds in Theorems 4.2 and 4.3 are both summations of exponentially many (in $L$) terms indexed by some subset $S \subseteq [L]$, whereas the corresponding lower bounds in Theorems 2.1 and 2.2 are both maxima of that many terms. Thus, a priori, there is no reason to believe that the upper and lower bounds should match, especially when $L$ tends to infinity with $n$. However, in this subsection, we show that this is indeed the case under mild regularity conditions, establishing asymptotic minimaxity of Algorithm 3 for both global and individualized estimation.

4.3.1. Minimax optimality for global estimation. Based on (4.14), a naive argument would upper bound $\sum_{S \subseteq [L]} e^{-I_{S}}$ by $2^L \exp\left\{-\left(1-n\right) \min_{S \subseteq [L]} I_{S}\right\}$. In order to match the lower bound (2.12), we need to assume that $L \ll \min_{S \subseteq [L]} I_{S}$. It turns out such a requirement on the growth rate of $L$ can be substantially relaxed, as detailed in the next theorem.

**Theorem 4.4** (Minimax optimality of MAP-based refinement for global estimation). Consider the setup of Theorem 4.2, except that instead of assuming (4.13), we now assume

$$\log L + L e^{-c'J_{\rho}} \ll \min_{S \subseteq [L]} I_{S} \to \infty \quad \text{as } n \to \infty$$

for some $c' \in (0,1)$. Then, there exists a sequence $\delta_n = o(1)$ such that

$$\lim_{n \to \infty} \inf_{z^* \in P_n} \mathbb{P}\left(\mathcal{L}(\hat{z}, z^*) \leq \exp\left\{-\left(1-\delta_n\right) \min_{S \subseteq [L]} I_{S}\right\}\right) = 1.$$

**Proof.** See Section C.4 of the supplementary material [12].

Consider the case where $\rho = n^{-c''}$ for some constant $c'' > 0$. In this case, we have $J_{\rho} = (1 + o(1)) \frac{1}{2} \log \frac{1}{\rho} = (1 + o(1)) \frac{c''}{2} \log n$, and thus the requirement in (4.20) becomes $\log L + L n^{-\left(1+o(1)\right)c''/2} \ll \min_{S \subseteq [L]} I_{S}$, a vast improvement over the naive requirement of $L \ll \min_{S \subseteq [L]} I_{S}$. In the homogeneous case of $\rho = 0$, the requirement in (4.20) becomes $\log L \ll I_{[L]} = \frac{n}{2} \sum_{\ell \in [L]} I_{\ell}^{(L)}$.

In the proof, we need to identify the optimal $S \subseteq [L]$ such that $I_{S}$ is minimized. While it is easy to do so when $\rho = 0$, this task turns out to be challenging in the presence of inhomogeneity, and our proof is based on a nontrivial application of a generalization of Von Neumann’s minimax theorem [55] due to Sion [61].
4.3.2. Minimax optimality for individualized estimation. Similar to the case of global estimation, in order to prove the tightness of (4.19), a naive argument would require $L \ll \left( \min_{S \subseteq [L] \setminus \{t\}} I_{S \cup \{t\}} \right) \wedge \left( \min_{S \subseteq [L] \setminus \{t\}} J_{S \cup \{t\}} \right)$, and this is relaxed in the following theorem.

**Theorem 4.5** (Minimax optimality of MAP-based refinement for individualized estimation). Consider the setup of Theorem 4.3, except that instead of assuming (4.18), we now assume
\[
\log L + L e^{-c' J_n} \ll \min_{S \subseteq [L] \setminus \{t\}} I_{S \cup \{t\}} \wedge J_{\{t\}} \to \infty \quad \text{as } n \to \infty
\]
for some $c' \in (0, 1)$. Then, there exists a sequence $\delta_n = o(1)$ such that
\[
\liminf_{n \to \infty} \mathbb{P} \left( L(\hat{z}^{(\ell)}, z^{(\ell)}) \leq \exp \left\{ - (1 - \delta_n) \min_{S \subseteq [L] \setminus \{t\}} I_{S \cup \{t\}} \right\} + \exp \left\{ - (1 - \delta_n) J_{\{t\}} \right\} \right) = 1.
\]

**Proof.** See Section C.5 of the supplementary material [12].

In the proof, we identify that the optimal $S$ that minimizes $J_{S \cup \{t\}}$ is precisely the empty set. On the other hand, the identification of the set that minimizes $I_{S \cup \{t\}}$ is done in a similar fashion as in the proof of Theorem 4.4.

5. Numerical Experiments. In this section, we conduct simulation studies to corroborate our theoretical results. Since we focus on the symmetric case, we can without loss of generality assume $z^*_i = +1$ for $1 \leq i \leq \lfloor n/2 \rfloor$ and $z^*_i = -1$ for $\lceil n/2 \rceil + 1 \leq i \leq n$. The $L$ layers are divided into three disjoint groups:

1. Weak layers. For $1 \leq \ell \leq \lfloor 0.3L \rfloor$, we let $p_\ell = c/(nL)$, $q_\ell = 1/(nL)$ for some constant $c$ that controls the amount of information. Since $I_{1/2}^{(\ell)} \asymp 1/(nL)$, in view of the lower bound (2.12), these layers, even when pooled together, cannot consistently estimate $z^*$.

2. Intermediate layers. For $\lfloor 0.3L \rfloor + 1 \leq \ell \leq \lfloor 0.95L \rfloor$, we let $p_\ell = c(\log n)/(nL)$, $q_\ell = (\log n)/(nL)$ where $c$ is the same constant as that appears in the weak layers. Note that $I_{1/2}^{(\ell)} \asymp \log(n)/(nL)$. Thus, while each individual layer does not contain sufficient information for consistent estimation of its own $z^{(\ell)}$, consistent estimation of $z^*$ becomes possible if information is aggregated across these layers.

3. Strong layers. For $\lfloor 0.95L \rfloor + 1 \leq \ell \leq L$, we let $p_\ell = c\log(n)/n$, $q_\ell = \log(n)/n$ again for the same $c$ as above. As $I_{1/2}^{(\ell)} \asymp \log(n)/n$, these layers are capable of consistently estimating their $z^{(\ell)}$'s, even when treated individually without aggregation.

The rationale behind the above partition is to simulate the behaviors that are likely to appear in real world multilayer networks. For example, let us consider the case where layers are distinct “participants” collaborating with each other, with the hope that they can borrow information from others to better estimate their own $z^{(\ell)}$'s. Such a setup is also known as “federated learning” in the machine learning literature [36]. The intermediate layers are participants with the most incentive in the collaboration, as “united they stand, divided they fall”. In comparison, the weak layers may not be as incentivized as the intermediate layers, because they would “fall even when united”. Nevertheless, they may still want to participate as “hitch-hikers”. Finally, the strong layers are participants that would “stand even when divided”, and the only reason for them to participate is the hope for even more accurate estimation of $z^{(\ell)}$'s.
5.1. Comparison between Algorithms 2 and 3. Recall that we have developed two versions of the same algorithm: Algorithm 2 is fast but we were not able to establish any theoretical guarantees, whereas Algorithm 3 is slower but provably optimal. We have argued that these two versions should perform similarly in terms of estimation accuracy, and we now empirically justify this claim. We set \( n = 200, L = 100, \rho = 0.1 \) and let \( c \) be either 2 or 5. We then run the two algorithms over 500 instances of the model (assuming \( \{p_{\ell}\}, \{q_{\ell}\}, \rho \) are known) and record the misclustering proportions for global estimation (Layer=Global), individualized estimation in weak layers (Layer=Weak), intermediate layers (Layer=Intermediate) and strong layers (Layer=Strong).

The results are presented in Figure 1. We see that the performances of the two versions are indeed similar, and they even become indistinguishable when \( c = 5 \). Thus, in the rest of this section, we always use Algorithm 2.

5.2. Effects of SNRs and comparison with co-regularized spectral clustering. Recall that the minimax rates for global estimation and individualized estimation both rely on two information theoretic quantities: \( J_{\rho} = (1 + o(1)) \cdot \frac{1}{L} \log(1/\rho) \) and \( I_{\ell}(t) = (1 + o(1)) \cdot (c^t - 1)(c^{1-t} - 1) \times \) (scaling of \( q_{\ell} \)). We now experiment on how these two quantities influence the performance of our proposed algorithm. We set \( n = 1000, L = 100, \) and we either fix \( \rho = 0.1 \) and vary \( c \), or fix \( c = 2 \) and vary \( \log(1/\rho) \). We run Algorithm 2 over 500 instances of the model (again assuming \( \{p_{\ell}\}, \{q_{\ell}\}, \rho \) are known) and record the misclustering proportions for both global estimation and individualized estimation. As a comparison, we implement the co-regularized spectral clustering algorithm, a popular algorithm for clustering in multilayer networks originally proposed by [39] and later shown to be consistent in the \( \rho = 0 \) case by [58]. Since co-regularized spectral clustering requires running multiple “coordinate ascent” steps, each involving computing the eigen-decomposition of an \( n \times n \) matrix, it is
Figure 2 shows the results of this simulation. We see that our method significantly outperforms co-regularized spectral clustering in all scenarios considered. By the top-left plot in Figure 2, the misclustering proportions of our method for both global estimation and individualized estimation tend to zero if we fix $c = 2$ and increase the value of $J_\rho$. In contrast, by the bottom-left plot in Figure 2, for fixed $\rho = 0.1$, misclustering proportions for $z^\star$ tend to zero as $c$ becomes large, which is as expected. However, for individualized estimation, while errors of strong layers still tend to zero, errors of intermediate and weak layers both tend to $\rho = 0.1$. This behavior actually is well explained by our theory. Note that the minimax rate (4.23) for individualized estimation consists of two terms, where the first term represents the error from label sampling (2.16) and scales as $e^{-1+o(1))J_\rho} = e^{1+o(1)} \times e^{-1+o(1)\psi_\{\ell\}(-2J_\rho)}$, whereas the second term comes from empirically estimating $z^\star$, which tends to zero much faster than the first term in the current setting. Different behaviors of individualized estimation errors in strong layers and in intermediate/weak layers occur since $\psi_\{\ell\}(-2J_\rho)$ is quite large when the $\ell$-th layer is strong while it is nearly zero when the layer only has intermediate or weak signal.

5.3. *Sensitivity to inexact parameter specifications.* The optimality of our proposed algorithm has been established assuming knowledge of the true $\{p_\ell\}, \{q_\ell\}$ and $\rho$. In practice,
COMMUNITY DETECTION IN INHOMOGENEOUS MULTILAYER NETWORKS

Fig 3: Average misclustering proportions against $\log(1/\rho_{\text{input}})$ when $\{p_\ell\}, \{q_\ell\}$ are estimated from data. The black dashed vertical line represents the location of the true $\rho$ that generates the data, and the horizontal lines represent the errors made by Algorithm 2 when $\{p_\ell\}, \{q_\ell\}, \rho$ are all known.

we need to estimate them from data. Estimating $\{p_\ell\}$ and $\{q_\ell\}$ is relatively easy — we could first obtain a crude estimate of $p_\ell$’s using method of moment (see Section F.4 of the supplement material [12] for details), then input that to Algorithm 1 to obtain an initial estimator $\tilde{z}$ of $z^*$, and finally compute the intra-cluster and inter-cluster average of edges in each layer, which will be our final estimator of $\{p_\ell\}$ and $\{q_\ell\}$.

Estimating $\rho$ is, however, a nontrivial task. Alternatively, we could treat the input $\rho$ (denoted as $\rho_{\text{input}}$) to Algorithm 2 as a hyperparameter. In this simulation, we examine the sensitivity of the algorithm to the estimated (hence inexact) $\{p_\ell\}, \{q_\ell\}$ and $\rho$. We again set $n = 1000$, $L = 100$, and we fix $c = 2, \rho = 0.1$. We run Algorithm 2 with estimated $\{p_\ell\}$ and $\{q_\ell\}$ over 500 instances of the model with different (misspecified) input values of $\rho$ and plot misclustering proportions for both global estimation and individualized estimation in Figure 3.

From Figure 3, we see that our algorithm is robust to inexact parameters. As long as we do not set $\rho_{\text{input}}$ to be too large, the performance of our algorithm with estimated $\{p_\ell\}$ and $\{q_\ell\}$ only slightly degrades compared to when true parameter values are used.

6. Extension to Multi-Cluster and Asymmetric Cases. In this section, we adapt Algorithm 2 to accommodate cases where the number of communities could be more than two and the community sizes could be considerably different.

We start by describing a canonical generalization of two-block IMLSBM (i.e., the model described by (1.1) and (1.2)) below. Suppose the global community assignment vector is $z^* \in \{1, 2, \ldots, K\}^n$, where $K \geq 2$ is the number of communities. For each layer $\ell \in [L]$ and each node $i \in [L]$, the individual community assignment $z_i^{(\ell)}$ independently follows the multinomial distribution

$$
P\left(z_i^{(\ell)} = k\right) = (1 - \rho) \cdot 1\{z_i^* = k\} + \frac{\rho}{K - 1} \cdot 1\{z_i^* \neq k\}, \quad 1 \leq k \leq K$$

for some $\rho \in (0, 1)$. That is, $z_i^{(\ell)}$ agrees with $z_i^*$ with probability $1 - \rho$, and flips to other communities with equal probabilities. The layer-wise adjacency matrices are still generated according to (1.2).

Algorithm 2 seamlessly generalizes to the current setting — one can simply write down the MAP objective function and perform node-wise refinement. All we need to do is to modify
**Algorithm 4:** Node-wise refinement via MAP estimation for multi-class IMLSBM

**Input:** Number of communities $K$, initial global estimator $\tilde{z}^\star$, adjacency matrices $\{A^{(\ell)}\}_1^L$, connecting probabilities $\{(p_\ell, q_\ell)\}_1^L$, flipping probability $\rho$

**Output:** Global estimator $\hat{z}^\star$, individualized estimator $\{\hat{z}^{(\ell)}\}_1^L$

1. for $i = 1, \ldots, n$ do
2.  for $\ell = 1, \ldots, L$ do
3.   for $k = 1, \ldots, K$ do
4.    $z^{(\ell,k)} \leftarrow \arg\max_{s \in [K]} f_i^{(\ell)}(k, s, \tilde{z}^\star)$; // $f_i^{(\ell)}$ is defined in (6.1)
5.  $\hat{z}^\star_i \leftarrow \arg\max_{s^\star \in [K]} \sum_{\ell \in [L]} f_i^{(\ell)}(s^\star, z^{(\ell)}(s^\star), \tilde{z}^\star)$; // final global estimator
6. for $\ell = 1, \ldots, L$ do
7.   $\hat{z}^{(\ell)}_i \leftarrow z(\ell, \hat{z}^\star_i)$; // final individualized estimators
8. return $\hat{z}^\star, \{\hat{z}^{(\ell)}\}_1^L$

Fig 4: Performance of Algorithm 4 for $K = 5$ under different SNRs. The top two figures fix $c = 5$ and vary $\log(1/\rho)$ from 1 to 6. The bottom two figures fix $\rho = 0.1$ and vary $c$ from 1 to 10. The left two figures have symmetric communities, meaning that all 5 communities have the same size in $z^\star$. The right two figures have asymmetric communities, where the proportions of the 5 communities are $\frac{1}{2}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}$.

(3.7) to

$$f_i^{(\ell)}(s^\star, s_\ell, \tilde{z}^\star) = \log \left( \frac{1 - \rho}{\rho/(K-1)} \right) \cdot 1\{s_\ell = s^\star\}$$
See Algorithm 4 for a detailed description.

We conduct a small scale simulation study to verify the effectiveness of Algorithm 4. In the simulation shown in Figure 4, we take $n = 1000$, $L = 100$, and $K = 5$. We still divide the layers into weak, intermediate and strong according to the description in Section 5, except that for intermediate layers, we set $p_t = 35c \log n/nL$ and $q_t = 35 \log n/(nL)$. We either fix $c = 5$ and vary $\rho$ (top), or fix $\rho = 0.1$ and vary $c$ (bottom). We consider both symmetric and asymmetric community sizes. Let $n_k = \sum_{i=1}^n I\{z_i^* = k\}$ denote the size of the $k$-th community in the global assignment. For the symmetric case, all 5 communities in $z^*$ have the same size (i.e., $n_1 = \ldots = n_5 = 200$). For the asymmetric case, we take $n_1 = 500$, $n_2 = 250$, and the other three communities have sizes 83, 83 and 84.

In this simulation, we observe similar patterns as those in Figure 2 which focused on the symmetric two-block case. When we fix $c$ and send $\rho$ to zero, both the loss in the global estimator and the loss in the individualized estimators tend to zero. When $\rho$ is fixed at 0.1 and $c$ grows, the loss in the individualized estimators in the weak layers approaches 0.1. Our algorithm also performs well in the asymmetric case.

7. A Real Data Example. Development of biotechnologies have enabled biologists to simultaneously profile multiple modalities at single-cell resolution. Here we illustrate our algorithm on a multi-modal single cell data [30] obtained via the CITE-seq technology [63].

This dataset contains 211000 human peripheral blood mononuclear cells (PBMCs). For each cell, 20729 RNA markers and 224 protein antibody panels were measured and their expression levels were recorded. The dataset also contains expert-annotated labels that categorize each cell into one of eight cell types: Monocyte (Mono), CD4 T cell, CD8 T cell, other T cell, Natural killer cell (NK), B cell, Dendritic cell (DC), and other (undefined).

We randomly subsampled 5000 cells and removed the two smallest clusters: DC (size=127) and other (size=104). Among the remaining 4769 cells, we aggregated CD4 T, CD8 T, and other T cells as a single T cell group. Thus, we obtained two data matrices $X^{(1)} \in \mathbb{R}^{4769 \times 20729}$, $X^{(2)} \in \mathbb{R}^{4769 \times 224}$, with RNA and protein information, respectively. The rows (cells) are classified into four clusters: Mono (size=1547), T (size=2224), NK (size=581), and B (size=417). For RNA data, we retained the top 50% variable columns (according to standard deviation) and computed a pairwise distance matrix $D^{(1)} \in \mathbb{R}^{4769 \times 4769}$, where $D_{i,j}^{(1)}$ is one minus the Pearson correlation coefficient between the $i$-th row and the $j$-th row of $X^{(1)}$. We then obtained a matrix $A_{row}^{(1)}$ by dichotomizing rows of $D^{(1)}$: setting the smallest 3% entries of each row to be one and the rest zero. Similarly, we obtain $A_{col}^{(1)}$ by dichotomizing columns of $D^{(1)}$. The final adjacency matrix for RNA information is given by $A^{(1)} = (A_{ij}^{(1)})$ where $A_{ii}^{(1)} = 0$ and for any $i$ and $A_{ij}^{(1)} = \max\{(A_{row}^{(1)} \odot A_{col}^{(1)})_{ij}, (A_{row}^{(1)} \odot A_{col}^{(1)})_{ji}\}$ for any $i \neq j$. Here $\odot$ stands for element-wise product. The adjacency matrix $A^{(2)}$ for protein information was similarly constructed, except that we picked the top 95% variable columns due to limited number of protein panels.

We applied the multi-class version of our algorithm (i.e., Algorithm 1+4) to the two-layer network $(A^{(1)}, A^{(2)})$ with $K = 4$ and $\rho = 0$. Note that $\rho = 0$ reveals our prior knowledge that the two individual label vectors should exactly agree. The misclustering proportion of the resulting global estimator is $5.89\%$. In contrast, spectral clustering with local likelihood refinement applied to RNA network alone and protein network alone gave estimators whose
misclustering proportions are 8.68\% and 10.99\%, respectively. The lower misclustering proportion achieved by applying our algorithm to the two-layer network reveals that our algorithm successfully integrated the two sources of information.

To illustrate our algorithm with non-zero \( \rho \), we randomly sampled an index set \( S \) with \(|S| = 1906 \) among the 4769 cells. We then randomly partitioned \( S \) into two equally-sized disjoint subsets \( S_1 \cup S_2 \) (i.e., \(|S_1| = |S_2| = 953\)), and let \( \tilde{X}^{(1)} = X^{(1)}_{S_1 \cup S_2} \in \mathbb{R}^{3816 \times 20,729} \) and \( \tilde{X}^{(2)} = X^{(2)}_{S_1 \cup S_2} \in \mathbb{R}^{3816 \times 224} \). After the operation, about 20\% of the cells have different cluster labels between the two modalities. Such a mismatch could happen, for example, when two modalities are not row-wise aligned a priori and a matching algorithm has been applied to partially align the cells [29, 7, 64, 73]. We then applied the same dichotomization procedure to \( (\tilde{X}^{(1)}, \tilde{X}^{(2)}) \) and obtain a two-layer network \( (\tilde{A}^{(1)}, \tilde{A}^{(2)}) \), to which we applied our algorithm with \( K = 4 \) and \( \rho = 0.2 \). The misclustering proportions of the two individualized estimators (for RNA labels and protein labels) are 8.12\% and 5.71\%, respectively. In contrast, the misclustering proportions of the “spectral clustering + local refinement” estimators obtained by separately working with RNA network and protein network are 8.49\% and 11.45\%, respectively. Such a noticeable difference suggests that our algorithm produces individualized estimators that not only borrow information across multiple layers, but also retain modality-specific characteristics.

SUPPLEMENTARY MATERIAL

Supplement to “Global and Individualized Community Detection in Inhomogeneous Multilayer Networks”

In Section A of the supplementary material, we provide additional theoretical results, including upper and lower bounds when \( \rho \) is of constant order (Section A.1), upper bounds when the community structure is asymmetric (Section A.2), upper bounds when \( \rho \) is mis-specified and \((p_\ell, q_\ell)_{\ell=1}^L \) are estimated from the data (Section A.3), and upper bounds when \( n = O(1), L \to \infty \) (Section A.4). In Sections B, C, D, and E of the supplementary material, we provide proofs of all the theoretical results. In Section F of the supplementary material, we provide additional details in the numerical experiments.

REFERENCES

COMMUNITY DETECTION IN INHOMOGENEOUS MULTILAYER NETWORKS


